สำนักหอสมุด

อภินันทนาการ

การปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมาย

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ศิริจรรยา จันทร์มี

สำนักหอสมุด มหาวิทยาลัยนเรศวร
วันลงทะเบียน 30 มค. 2566
เลขทะเบียน.....
เลขเรียกหนังสือ.....

วิทยานิพนธ์เสนอบัณฑิตวิทยาลัย มหาวิทยาลัยนเรศวร
เพื่อเป็นส่วนหนึ่งของการศึกษา หลักสูตรปรัชญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
ปีการศึกษา 2565
ลิขสิทธิ์เป็นของมหาวิทยาลัยนเรศวร

วิทยานิพนธ์ เรื่อง "การปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมาย"

ของ ศิริจรรยา จันทร์มี

ได้รับการพิจารณาให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการสอบวิทยานิพนธ์

(ดร.มารุต บูรณรัช)

.....ประธานที่ปรึกษาวิทยานิพนธ์

(รองศาสตราจารย์ ดร.ไกรศักดิ์ เกษร)

.....กรรมการผู้ทรงคุณวุฒิภายใน

(ผู้ช่วยศาสตราจารย์ ดร.สุธาสินี จิตต์อนันต์)

.....กรรมการผู้ทรงคุณวุฒิภายใน

(ผู้ช่วยศาสตราจารย์ ดร.ดวงเดือน อัศวสุธีรกุล)

.....กรรมการผู้ทรงคุณวุฒิภายใน

(ผู้ช่วยศาสตราจารย์ ดร.อนงค์พร ไศลวรากุล)

อนุมัติ

.....

(รองศาสตราจารย์ ดร.กรองกาญจน์ ชูทิพย์)

คณบดีบัณฑิตวิทยาลัย

24 พ.ย. 2565

3568545173 NU iThesis 60031257 thesis / recv: 15072565 16:56:58 / seq: 24

ชื่อเรื่อง การปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมาย

ผู้วิจัย ศิริจรรยา จันทร์มี

ประธานที่ปรึกษา รองศาสตราจารย์ ดร.ไกรศักดิ์ เกษร

ประเภทสารนิพนธ์ วิทยานิพนธ์ ปร.ด. เทคโนโลยีสารสนเทศ, มหาวิทยาลัยนเรศวร, 2565

คำสำคัญ การจำแนกข้อมูล, ID3, เกนสารสนเทศ, วิธีการเชิงความหมาย, ฐานความรู้

บทคัดย่อ

เทคนิคต้นไม้ตัดสินใจเป็นอัลกอริทึมสำหรับการจำแนกข้อมูลที่ได้รับความนิยม ซึ่งประสิทธิภาพในการจำแนกข้อมูลของเทคนิคนี้จะขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ในการเรียนรู้รวมถึงประสิทธิภาพของกระบวนการในการพิจารณาแอตทริบิวต์สำหรับทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจด้วยค่าเกนสารสนเทศ อย่างไรก็ตามการใช้ค่าเกนสารสนเทศในการพิจารณาโหนดสำหรับต้นไม้ตัดสินใจยังคงมีข้อจำกัดเรื่องความลำเอียงในการพิจารณาโหนดสำหรับต้นไม้ตัดสินใจโดยแอตทริบิวต์ที่มีค่าข้อมูลที่หลากหลายจะมีโอกาสถูกเลือกเป็นโหนดสำหรับต้นไม้ตัดสินใจมากกว่าแอตทริบิวต์อื่น ๆ เพื่อลดปัญหาดังกล่าวในการวิจัยนี้จึงได้นำเสนออัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย ซึ่งจะใช้อัลกอริทึม ID3 (Iterative Dichotomiser 3) เป็นพื้นฐาน อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายที่นำเสนอจะนำองค์ความรู้ในออนโทโลยีมาใช้ช่วยสนับสนุนกระบวนการสร้างต้นไม้ตัดสินใจ โดยแนวความคิดและความสัมพันธ์ของแนวความคิดในออนโทโลยีจะถูกนำมาใช้ในการระบุค่าระดับความสำคัญของแอตทริบิวต์ในชุดข้อมูล และนำค่าระดับความสำคัญที่ได้ไปใช้ในการปรับปรุงค่าเกนสารสนเทศเพื่อให้สามารถพิจารณาแอตทริบิวต์สำหรับเป็นโหนดของต้นไม้ตัดสินใจได้อย่างเหมาะสมมากยิ่งขึ้น รวมถึงนำองค์ความรู้ในออนโทโลยีมาใช้ในการสนับสนุนการจัดเตรียมข้อมูลเพื่อช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลอีกด้วย ในการวิจัยครั้งนี้ได้ทำการทดสอบวิธีการที่นำเสนอกับชุดข้อมูลจำนวน 4 ชุดข้อมูล ได้แก่ ชุดข้อมูลการเกิดโรคของถั่วเหลือง ชุดข้อมูลผู้ป่วยโรคหัวใจ ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออก ผลการวิจัยพบว่าการนำองค์ความรู้ที่อยู่ในรูปแบบของออนโทโลยีสามารถช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลของต้นไม้ตัดสินใจ โดยช่วยให้ความถูกต้องในการจำแนกข้อมูลมีค่าเพิ่มขึ้น และมีค่าความถูกต้องในการจำแนกข้อมูลมากกว่าอัลกอริทึมต้นไม้ตัดสินใจอื่น ๆ เช่น ID3 CART

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

(Classification and Regression Tree) C4.5 และMIDT (Mutual Information Decision Tree)

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

Title	DATA CLASSIFICATION IMPROVEMENT USING SEMANTIC DECISION TREE
Author	Sirichanya Chanmee
Advisor	Associate Professor Kraisak Kesorn, Ph.D.
Academic Paper	Ph.D. Dissertation in Information Technology - (Type 2.1), Naresuan University, 2022
Keywords	Classification, ID3, Information Gain, Semantic, Knowledge-base

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ABSTRACT

Decision trees are a well-known algorithm for classification tasks. The performance of a decision tree depends on the quality of the learning data and the efficiency of the decision construction process with information gain. With the multi-valued bias that is introduced by the use of information gain, the algorithm favors selecting the attribute with multiple values as a node of the decision tree rather than selecting the attributes with fewer values, although the selected attributes may be less important. To deal with this problem, we proposed a new decision tree algorithm which we titled "Semantic Decision Tree (SDT)", which is based on the Iterative Dichotomiser 3 (ID3) algorithm. The proposed algorithm exploits knowledge in an ontology to assist the decision tree construction process. The concepts and relationships between concepts are used to determine the attribute importance values. These values are used to adjust the information gain to revise the decision tree. The knowledge in the ontology is also applied during the data preparation process to improve the data quality, which enhances the classification performance. Four publicly available datasets: Soybean, Heart Disease, COVID-19 and Dengue fever, were applied to evaluate the proposed algorithm. The experimental results demonstrated that using the knowledge in the ontology enhances the decision tree construction performance. The proposed algorithm also achieved better accuracy than other decision tree algorithms, e.g., the ID3, CART, C4.5 and the Mutual

Information Decision Tree (MIDT).

4918889183 NU iThesis 600312257 thesis / recv: 31102565 16:06:32 / seq: 39

ประกาศคุณูปการ

ผู้วิจัยขอกราบขอบพระคุณอย่างสูงในความกรุณาของ รองศาสตราจารย์ ดร.ไกรศักดิ์ เกษร ประธานที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำปรึกษาแนะนำตลอดจนแก้ไขข้อบกพร่องต่าง ๆ ตลอดระยะเวลาในการทำวิทยานิพนธ์ฉบับนี้ด้วยความเอาใจใส่เป็นอย่างยิ่ง และขอขอบพระคุณ ดร.มารุต บูรณรัช ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.สุธาสินี จิตต์อนันต์ ผู้ช่วยศาสตราจารย์ ดร.ดวงเดือน อัศวสุธีรกุล และผู้ช่วยศาสตราจารย์ ดร.อนงค์พร ไศลวรากุล กรรมการผู้ทรงคุณวุฒิภายใน ที่ได้กรุณาให้คำแนะนำตลอดจนแก้ไขข้อบกพร่องของวิทยานิพนธ์จนทำให้วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์

ขอขอบพระคุณคณาจารย์หลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศทุกท่านที่ได้กรุณาถ่ายทอดความรู้และประสบการณ์ให้แก่ผู้วิจัย และขอบคุณคณาจารย์และเจ้าหน้าที่ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศที่ให้การสนับสนุนในด้านต่าง ๆ เพื่อให้การดำเนินการจัดทำวิทยานิพนธ์เป็นไปอย่างเรียบร้อย

ขอขอบพระคุณกระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม และมหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ที่ให้ทุนสนับสนุนตลอดระยะเวลาการศึกษา

เหนือสิ่งอื่นใดขอกราบขอบพระคุณบิดา มารดา ของผู้วิจัยที่ให้กำลังใจและให้การสนับสนุนในทุก ๆ ด้านเสมอมา

คุณค่าและคุณประโยชน์อันพึงจะมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบและอุทิศแด่ผู้มีพระคุณทุกท่าน ผู้วิจัยหวังเป็นอย่างยิ่งว่างานวิจัยนี้จะเป็นประโยชน์ต่อผู้ที่สนใจไม่มากก็น้อย

ศิริจรรยา จันทร์มี

4918891B3 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ค

บทคัดย่อภาษาอังกฤษ ..... จ

ประกาศคุณูปการ ..... ช

สารบัญ ..... ซ

สารบัญตาราง ..... ฎ

สารบัญภาพ ..... ฑ

บทที่ 1 บทนำ 1

ความเป็นมาและความสำคัญของปัญหา 1

จุดมุ่งหมายของการวิจัย 6

ขอบเขตการวิจัย 7

สมมุติฐานของการวิจัย 7

นิยามศัพท์เฉพาะ 8

บทสรุป 8

บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง 10

เหมืองข้อมูล (Data Mining) 10

เหมืองข้อมูลเชิงความหมาย (Semantic Data Mining) 14

อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) 16

การประเมินประสิทธิภาพแบบจำลองการจำแนกข้อมูล 22

ออนโทโลยี (Ontology) 23

49188918 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

วิธีการสรุปภาพรวมออนโทโลยี.....25
อัลกอริทึม PageRank.....27
งานวิจัยที่เกี่ยวข้อง.....28
บทสรุป.....37
บทที่ 3 วิธีดำเนินการวิจัย.....38
ข้อมูลที่ใช้ในการวิจัย.....38
เครื่องมือที่ใช้ในการวิจัย.....42
วิธีการดำเนินการวิจัย.....43
การออกแบบกรอบแนวคิดการวิจัย.....43
การวางแผนการทดลอง.....62
บทสรุป.....68
บทที่ 4 การประยุกต์ใช้ออนโทโลยีในการเตรียมข้อมูลสำหรับเทคนิคต้นไม้ตัดสินใจ.....69
ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูล.....69
ผลการอ้างอิงแนวความคิดพื้นฐานจากออนโทโลยี.....73
ผลการปรับปรุงข้อมูลด้วยแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยี.....74
ผลการจำแนกข้อมูลที่มีการนำแนวความคิดพื้นฐานมาใช้งาน.....79
สรุปผลการวิจัย.....81
บทที่ 5 การปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจโดยการประยุกต์ใช้ออนโทโลยี.....83
ผลการคำนวณค่าระดับความสำคัญจากออนโทโลยี.....83
ผลการทดสอบประสิทธิภาพการจำแนกข้อมูลของต้นไม้ตัดสินใจเชิงความหมาย.....87
ผลการทดสอบความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting).....93
ผลการทดสอบการทนทานต่อข้อมูลที่ผิดปกติต่อต้นไม้ตัดสินใจเชิงความหมาย..... 100

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ผลการทดสอบปรับพารามิเตอร์ที่เหมาะสมสำหรับเพิ่มประสิทธิภาพการจำแนกข้อมูล 103
ผลการพิจารณาโครงสร้างของต้นไม้ตัดสินใจเชิงความหมาย 107
ผลการประมาณค่าระดับความสำคัญเมื่อไม่ปรากฏองค์ความรู้ในออนโทโลยี 112
ผลการเปรียบเทียบประสิทธิภาพของต้นไม้ตัดสินใจเชิงความหมายกับอัลกอริทึมอื่น ๆ 120
สรุปผลการวิจัย 122
บทที่ 6 สรุปผลการวิจัย 124
สรุปผลการวิจัย 124
ข้อค้นพบที่ได้และการบรรลุวัตถุประสงค์การวิจัย 126
ข้อจำกัดและแนวทางในการวิจัย 127
การนำไปใช้ประโยชน์ 128
บทความทางวิชาการจากงานวิจัย 128
บรรณานุกรม 130
ภาคผนวก 138
ภาคผนวก ก ตัวอย่างแนวความคิดในออนโทโลยีโรคของถั่วเหลือง 139
ภาคผนวก ข ตัวอย่างแนวความคิดในออนโทโลยีโรคหัวใจ 140
ภาคผนวก ค ตัวอย่างแนวความคิดในออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019 141
ภาคผนวก ง ตัวอย่างแนวความคิดในออนโทโลยีโรคไข้เลือดออก 142
ประวัติผู้วิจัย 143

4918889183
NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

สารบัญตาราง

หน้า

ตาราง 1 แอตทริบิวต์ภายในชุดข้อมูลการเกิดโรคของถั่วเหลือง 39

ตาราง 2 แอตทริบิวต์ภายในชุดข้อมูลผู้ป่วยโรคหัวใจ 40

ตาราง 3 แอตทริบิวต์ภายในชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 41

ตาราง 4 แอตทริบิวต์ภายในชุดข้อมูลผู้ป่วยโรคไข้เลือดออก 42

ตาราง 5 สรุปข้อมูลออนโทโลยีที่ใช้ในการวิจัย 47

ตาราง 6 สรุปข้อมูลสถิติที่ใช้ในการทดสอบความสัมพันธ์ระหว่างข้อมูล.....50

ตาราง 7 ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์สำหรับชุดข้อมูลการเกิดโรคของถั่วเหลือง.....70

ตาราง 8 ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลสำหรับชุดข้อมูลผู้ป่วยโรคหัวใจ 71

ตาราง 9 ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์สำหรับชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019.....71

ตาราง 10 ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลสำหรับชุดข้อมูลผู้ป่วยโรคไข้เลือดออก 72

ตาราง 11 แนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีโรคของถั่วเหลือง.....73

ตาราง 12 ผลการจำแนกข้อมูลเมื่อมีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีในการปรับปรุงข้อมูล.....79

ตาราง 13 ผลการจำแนกข้อมูลเมื่อมีการใช้ค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสม 80

ตาราง 14 ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคของถั่วเหลืองซึ่งมีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูล.....84

4918891B3 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 15 ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคหัวใจซึ่งมีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูล....85

ตาราง 16 ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019 ซึ่งมีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูล....85

ตาราง 17 ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคไข้เลือดออกซึ่งมีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูล....86

ตาราง 18 ความถูกต้องในการจำแนกข้อมูลด้วยอัลกริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย....88

ตาราง 19 ค่าเกนสารสนเทศของแอตทริบิวต์ในชุดข้อมูลผู้ป่วยโรคหัวใจ....91

ตาราง 20 ค่าความซับซ้อนของออนโทโลยี....92

ตาราง 21 ผลลัพธ์การทดสอบวิลคอกซันของชุดข้อมูลการเกิดโรคของถั่วเหลือง....96

ตาราง 22 ผลลัพธ์การทดสอบวิลคอกซันของชุดข้อมูลผู้ป่วยโรคหัวใจ....97

ตาราง 23 ผลลัพธ์การทดสอบวิลคอกซันของชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 98

ตาราง 24 ผลลัพธ์การทดสอบวิลคอกซันของชุดข้อมูลผู้ป่วยโรคไข้เลือดออก....99

ตาราง 25 ผลการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเมื่อมีการใช้ค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสม....104

ตาราง 26 ตัวอย่างผลการพิจารณาคลาสแม่และแนวความคิดทั่วไปของแอตทริบิวต์ต่าง ๆ ในชุดข้อมูลผู้ป่วยโรคไข้เลือดออก....114

ตาราง 27 ค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ยจากการประมาณค่าระดับความสำคัญของแอตทริบิวต์....115

ตาราง 28 จำนวนแอตทริบิวต์ที่ใช้สำหรับทดสอบการประมาณค่าระดับความสำคัญที่สูญหาย....116

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 29 ผลการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 10 116
ตาราง 30 ผลการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 20 117
ตาราง 31 ผลการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 30 118

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

สารบัญภาพ

หน้า

ภาพ 1 ตัวอย่างขั้นตอนการสร้างต้นไม้ตัดสินใจสำหรับการวินิจฉัยโรคหวัด 3

ภาพ 2 ตัวอย่างต้นไม้ตัดสินใจในกรณีที่ชุดข้อมูลปรากฏแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาส 5

ภาพ 3 ตัวอย่างต้นไม้ตัดสินใจในกรณีแอตทริบิวต์มีค่าข้อมูลแตกต่างกัน 5

ภาพ 4 กระบวนการทำเหมืองข้อมูลตามแบบจำลอง CRISP-DM 11

ภาพ 5 องค์ประกอบภายในต้นไม้ตัดสินใจ 17

ภาพ 6 ขั้นตอนการสร้างต้นไม้ตัดสินใจ 18

ภาพ 7 เอนโทรปีของการสุ่มด้านของเหรียญ โดย X= 1 หมายถึงเหรียญด้านหัว และ P(X) หมายถึงความน่าจะเป็นของการสุ่มด้านของเหรียญ 20

ภาพ 8 ตัวอย่างแนวคิดการทำงานของอัลกอริทึม PageRank 27

ภาพ 9 ตัวอย่างแนวความคิดในออนโทโลยีโรคของถั่วเหลือง 36

ภาพ 10 กรอบแนวคิดการวิจัย 44

ภาพ 11 ตัวอย่างออนโทโลยีโรคของถั่วเหลือง 46

ภาพ 12 กระบวนการจัดเตรียมข้อมูล 47

ภาพ 13 ขั้นตอนการประยุกต์ใช้ออนโทโลยีในการปรับปรุงข้อมูล 50

ภาพ 14 รหัสเทียม (Pseudo Code) สำหรับการระบุแนวความคิดพื้นฐานที่สัมพันธ์กับข้อมูล 51

ภาพ 15 รหัสเทียม (Pseudo Code) สำหรับการแปลงข้อมูลเดิมด้วยแนวความคิดพื้นฐานที่สัมพันธ์กับข้อมูล 52

ภาพ 16 ตัวอย่างการอ้างอิงแนวความคิดพื้นฐานจากออนโทโลยี 53

4918891B3 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ภาพ 17 กระบวนการสร้างแบบจำลองโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยี 54

ภาพ 18 รหัสเทียม (Pseudo Code) สำหรับการคำนวณค่าส่วนกลับของความถี่ของความสัมพันธ์ในออนโทโลยี 56

ภาพ 19 รหัสเทียม (Pseudo Code) การคำนวณค่าน้ำหนักของแต่ละความสัมพันธ์ในออนโทโลยี 57

ภาพ 20 รหัสเทียม (Pseudo Code) สำหรับการคำนวณค่าระดับความสำคัญของแนวความคิดในออนโทโลยีด้วย Weighted Semantic PageRank 59

ภาพ 21 รหัสเทียม (Pseudo Code) สำหรับการสร้างต้นไม้ตัดสินใจที่มีการประยุกต์ใช้องค์ความรู้จากออนโทโลยี หรือ Semantic Decision Tree 61

ภาพ 22 แผนการทดลองสำหรับการปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมาย.....63

ภาพ 23 จำนวนข้อมูลในแอตทริบิวต์ stem-canker โดย (ก) เมื่อใช้ข้อมูลเดิม และ(ข) เมื่อใช้ค่าแนวความคิดพื้นฐาน 75

ภาพ 24 จำนวนข้อมูลในแอตทริบิวต์ fruit-pods โดย (ก) เมื่อใช้ข้อมูลเดิม และ(ข) เมื่อใช้ค่าแนวความคิดพื้นฐาน.....76

ภาพ 25 จำนวนข้อมูลในแอตทริบิวต์ fruit-spots โดย (ก) เมื่อใช้ข้อมูลเดิม และ(ข) เมื่อใช้ค่าแนวความคิดพื้นฐาน.....77

ภาพ 26 ผลการตรวจสอบข้อมูลที่ผิดปกติด้วย Isolation forest.....78

ภาพ 27 ตัวอย่างต้นไม้ย่อย (subtree) ที่มีความซ้ำซ้อนเมื่อสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม ID3.....89

ภาพ 28 ผลการทดสอบความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting) ของชุดข้อมูลต่าง ๆ โดย (ก) ชุดข้อมูลการเกิดโรคถั่วเหลือง (ข) ชุดข้อมูลผู้ป่วยโรคหัวใจ (ค) ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และ (ง) ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก.....94

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ภาพ 29 ผลการจำแนกข้อมูลที่มีปริมาณข้อมูลรบกวนแตกต่างกัน โดย (ก) ชุดข้อมูลการเกิดโรคของถั่วเหลือง (ข) ชุดข้อมูลผู้ป่วยโรคหัวใจ (ค) ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และ (ง) ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก 101

ภาพ 30 ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย เมื่อ ชุดข้อมูลปรากฏข้อมูลรบกวนและเมื่อกำจัดข้อมูลรบกวน..... 103

ภาพ 31 ตัวอย่างต้นไม้ย่อยในต้นไม้ตัดสินใจเชิงความหมาย โดย (ก) ต้นไม้ย่อยที่มีความสูงมากที่สุด และ (ข) ต้นไม้ย่อยเมื่อมีการปรับความสูงของต้นไม้ตัดสินใจเท่ากับ 7 105

ภาพ 32 ตัวอย่างโครงสร้างต้นไม้ตัดสินใจสำหรับชุดข้อมูลการเกิดโรคของถั่วเหลือง โดย (ก) อัลกอริทึม ID3 และ (ข) อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย 107

ภาพ 33 ค่าเอนโทรปีของแอตทริบิวต์เมื่อมีจำนวนค่าข้อมูลในแอตทริบิวต์แตกต่างกัน 108

ภาพ 34 ตัวอย่างโครงสร้างต้นไม้ตัดสินใจสำหรับชุดข้อมูลผู้ป่วยโรคไข้เลือดออก โดย (ก) อัลกอริทึม ID3 และ(ข) อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย 110

ภาพ 35 ขั้นตอนการประมาณค่าระดับความสำคัญสำหรับแอตทริบิวต์ที่ไม่ปรากฏในออนโทโลยี 112

ภาพ 36 ตัวอย่างการพิจารณาแนวความคิดในออนโทโลยีโรคไข้เลือดออกสำหรับการประมาณค่าระดับความสำคัญของแอตทริบิวต์..... 113

ภาพ 37 ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย เมื่อมีการประมาณค่าระดับความสำคัญของแอตทริบิวต์ 119

ภาพ 38 ผลการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจ เชิงความหมายและอัลกอริทึมอื่น ๆ ในกลุ่มของอัลกอริทึมต้นไม้ตัดสินใจ..... 121

ภาพ 39 ตัวอย่างแนวความคิดในออนโทโลยีโรคของถั่วเหลือง 139

ภาพ 40 ตัวอย่างแนวความคิดในออนโทโลยีโรคหัวใจ..... 140

ภาพ 41 ตัวอย่างแนวความคิดในออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019..... 141

ภาพ 42 ตัวอย่างแนวความคิดในออนโทโลยีโรคไข้เลือดออก 142

4918839183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39
491889183

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

เหมืองข้อมูล หรือ Data Mining (Hand, 2007) คือ วิธีการในการค้นหารูปแบบที่แฝงอยู่ในชุดข้อมูลหรือองค์ความรู้ที่เป็นประโยชน์จากข้อมูลขนาดใหญ่ด้วยการประยุกต์ใช้เทคโนโลยีต่าง ๆ ไม่ว่าจะเป็น วิธีการทางสถิติ เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) และเทคนิคในการวิเคราะห์ข้อมูล เป็นต้น ซึ่งการค้นหาองค์ความรู้และวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลนี้ยังคงมีข้อจำกัด เนื่องจากเครื่องคอมพิวเตอร์ไม่สามารถเข้าใจความหมายที่แท้จริงของข้อมูล ดังนั้นในกระบวนการวิเคราะห์ข้อมูล ข้อมูลจะถูกพิจารณาเป็นเพียงค่าตัวเลขและใช้วิธีการทางสถิติในการพิจารณาองค์ความรู้ที่แฝงอยู่ในข้อมูลเหล่านั้นโดยไม่นำความหมายของข้อมูลและความสัมพันธ์ระหว่างข้อมูลมาใช้ประโยชน์ในการวิเคราะห์เพื่อประสิทธิภาพการวิเคราะห์ข้อมูลที่ต้องมากขึ้น เพื่อให้คอมพิวเตอร์สามารถเข้าใจความหมายที่แท้จริงของข้อมูล นักวิจัยจึงได้นำเสนอวิธีการทำเหมืองข้อมูลเชิงความหมาย หรือ Semantic Data Mining (Dou et al., 2015) ซึ่งเป็นกระบวนการทำเหมืองข้อมูลที่มีการนำองค์ความรู้เฉพาะด้าน (Domain Knowledge) มาช่วยสนับสนุนในการวิเคราะห์ข้อมูล ไม่ว่าจะเป็นช่วยจำกัดขอบเขตหรือจำนวนของข้อมูลที่ต้องพิจารณาเพื่อค้นหาองค์ความรู้ที่แฝงอยู่ภายในข้อมูลนั้น แสดงรูปแบบของข้อมูลที่ชัดเจน รวมถึงช่วยแสดงความสัมพันธ์ของข้อมูลอีกด้วย (Anand et al., 1995) ตัวอย่างเช่น Kuo et al. (2007) ได้ประยุกต์ใช้องค์ความรู้ทางการแพทย์ที่แสดงอยู่ในรูปแบบของออนโทโลยี (Ontology) เพื่อจัดกลุ่มตัวแปรที่เกี่ยวข้องกับการเกิดโรคหลอดเลือดหัวใจจำนวน 85 ตัวแปร ออกเป็น 7 กลุ่ม และนำไปใช้ในการหากฎความสัมพันธ์ของตัวแปรต่าง ๆ ที่ส่งผลต่อการเสียชีวิตของผู้ป่วย ซึ่งการจัดกลุ่มตัวแปรโดยใช้องค์ความรู้ทางการแพทย์นี้ช่วยให้กฎที่ได้มีความชัดเจน และสามารถนำไปช่วยสนับสนุนการตัดสินใจได้ดียิ่งขึ้น นอกจากนี้ Sinha และ Zhao (2008) ได้ทำการศึกษาการประยุกต์ใช้องค์ความรู้ร่วมกับเทคนิคในการจำแนกข้อมูลหลาย ๆ เทคนิค ได้แก่ เทคนิคการเรียนรู้แบบเบย์ (Naive Bayes) การวิเคราะห์การถดถอยโลจิสติค (Logistic Regression) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคการตัดสินใจแบบตาราง (Decision Table) เทคนิคโครงข่ายประสาทเทียม (Neural Network) เทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbor) และ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) โดยผลการศึกษาพบว่าการนำองค์ความรู้มาใช้นั้น ช่วยให้การจำแนกข้อมูลมีประสิทธิภาพมากขึ้น โดยสามารถลดความผิดพลาดในการจำแนกข้อมูลได้

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

การจำแนกข้อมูล (Classification) คือ เทคนิคหนึ่งในการทำเหมืองข้อมูลที่มีวัตถุประสงค์เพื่อทำการจำแนกข้อมูลออกเป็นประเภทต่าง ๆ ที่เรียกว่า คลาส (Class) หรือ ลาเบล (Label) โดยจะทำการพิจารณาจากรูปแบบของข้อมูลที่พบในชุดข้อมูลที่ทำการศึกษา การจำแนกข้อมูลนี้จัดอยู่ในกลุ่มของเทคนิคการเรียนรู้แบบมีผู้สอน (Supervise Learning) ที่จะมีการสร้างสมการหรือแบบจำลองโดยการเรียนรู้จากข้อมูลในอดีต ในปัจจุบันมีหลายเทคนิคที่นำมาใช้ในการจำแนกข้อมูล ไม่ว่าจะเป็น เทคนิคโครงข่ายประสาทเทียม เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคเพื่อนบ้านใกล้ที่สุด และ เทคนิคต้นไม้ตัดสินใจ เป็นต้น เทคนิคต้นไม้ตัดสินใจ หรือ Decision Tree นั้นเป็นเทคนิคหนึ่งที่ได้รับความนิยมนำมาใช้ในการจำแนกข้อมูลเนื่องจากเป็นวิธีการที่รองรับทั้งข้อมูลที่เป็นแบบช่วง (Interval Data) และ ข้อมูลแบบกลุ่ม (Categorical Data) รวมทั้งมีโครงสร้างแบบต้นไม้ที่ง่ายต่อการทำความเข้าใจ โดยในการสร้างต้นไม้ตัดสินใจนั้นจะมีการพิจารณาโหนดภายในต้นไม้ตัดสินใจด้วยเกณฑ์ต่าง ๆ ซึ่งค่าเกนสารสนเทศ (Information Gain) เป็นเกณฑ์หนึ่งที่ได้รับความนิยมนำมาใช้ในการพิจารณาโหนดภายในต้นไม้ตัดสินใจ อย่างไรก็ตามการพิจารณาเลือกโหนดของต้นไม้ตัดสินใจโดยการใช้ค่าเกนสารสนเทศยังคงมีข้อจำกัด โดยแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายมีแนวโน้มที่จะถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจมากกว่าแอตทริบิวต์ที่มีข้อมูลไม่หลากหลาย (White & Liu, 1994) ซึ่งหากแอตทริบิวต์ที่ถูกเลือกนั้นเป็นแอตทริบิวต์ที่ไม่มีความสำคัญกับการจำแนกข้อมูลแล้วอาจส่งผลให้ต้นไม้ตัดสินใจที่ได้นั้นมีประสิทธิภาพในการจำแนกข้อมูลลดลง ตัวอย่างเช่นในการศึกษาของ Kononenko (1984) ได้ทำการศึกษาการสร้างกฎการจำแนกข้อมูลทางด้านการแพทย์ โดยใช้ค่าเกนสารสนเทศในการพิจารณาแอตทริบิวต์ที่จะใช้งาน พบว่าแอตทริบิวต์อายุของผู้ป่วยซึ่งประกอบไปด้วยข้อมูลช่วงอายุที่แตกต่างกันจำนวน 9 ช่วงอายุนั้นจะถูกเลือกไปใช้ในการสร้างกฎการจำแนกข้อมูล ในขณะที่แอตทริบิวต์อื่น ๆ ซึ่งมีจำนวนข้อมูลของแอตทริบิวต์น้อยกว่าแต่เป็นแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสคำตอบมากกว่าจะไม่ถูกเลือกไปใช้งาน ซึ่งการเลือกแอตทริบิวต์ที่ไม่เหมาะสมนี้ส่งผลให้ยากต่อการค้นหารูปแบบขององค์ความรู้ที่แฝงในชุดข้อมูล รวมทั้งส่งผลต่อประสิทธิภาพในการจำแนกข้อมูลอีกด้วย

นอกจากนี้คุณภาพของข้อมูล เช่น จำนวนข้อมูลที่สูญหาย (Missing value) จำนวนของแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับข้อมูลที่ต้องการจำแนก และจำนวนของข้อมูลที่ใช้ในการสร้างแบบจำลองเป็นอีกปัจจัยที่มีผลต่อประสิทธิภาพในการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ เนื่องจากเทคนิคต้นไม้ตัดสินใจนั้นเป็นวิธีการที่ถูกจัดอยู่ในกลุ่มของอัลกอริทึมเชิงละโมบ (Greedy Algorithm) รวมถึงใช้วิธีการแบ่งแยกและเอาชนะ (Divide and Conquer) ในการแบ่งชุดข้อมูลเพื่อใช้ในการสร้างต้นไม้ตัดสินใจ ซึ่งวิธีการเหล่านี้จะมีประสิทธิภาพในการทำงานลดลงเมื่อมีข้อมูลรบกวน (Noise) หรือมีแอตทริบิวต์ที่ไม่มีความเกี่ยวข้องปะปนอยู่ในชุดข้อมูล หากนำชุดข้อมูลที่มีความผิดปกติมาใช้ในการสร้างแบบจำลองอัลกอริทึมจะทำการสร้างต้นไม้ตัดสินใจเพื่ออธิบายข้อมูล

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

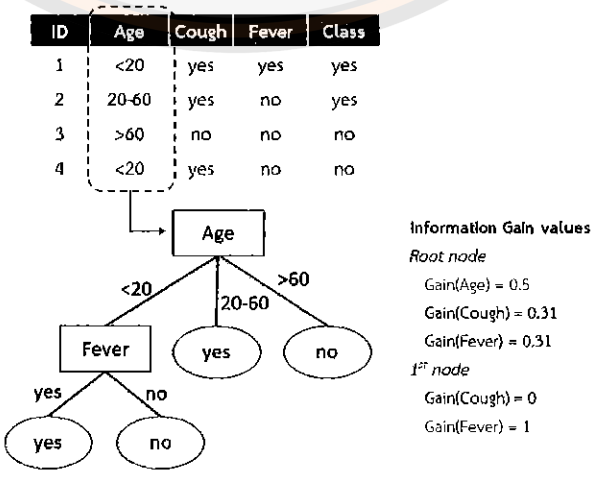
ที่ผิดปกตินั้น (Maimon & Rokach, 2014) นอกจากนี้เมื่อชุดข้อมูลมีแอตทริบิวต์จำนวนมากหรือมีค่าข้อมูลจำนวนมาก เทคนิคต้นไม้ตัดสินใจจะใช้ระยะเวลานานในการพิจารณาแอตทริบิวต์สำหรับเป็นโหนดภายในต้นไม้ตัดสินใจ (Ali & Rajamani, 2012) และสร้างต้นไม้ตัดสินใจที่มีความซับซ้อน เช่น ต้นไม้ตัดสินใจมีโหนดจำนวนมาก หรือต้นไม้ตัดสินใจมีความลึกมาก ซึ่งอาจทำให้เกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting) ได้ (Amjad et al., 2019)

ถึงแม้ว่าในปัจจุบันเทคนิคการเรียนรู้เชิงลึก (Deep Learning) เป็นเทคนิคที่ได้รับความสนใจและนำมาใช้ในการจำแนกข้อมูลอย่างแพร่หลาย แต่อย่างไรก็ตามเทคนิคนี้ยังคงมีข้อจำกัดในหลายประการ เช่น เป็นเทคนิคที่ต้องใช้ข้อมูลปริมาณมากในการสร้างแบบจำลองที่มีประสิทธิภาพ ใช้ทรัพยากรและระยะเวลานานในการประมวลผล รวมถึงผลลัพธ์ที่ได้ยากต่อการแปลผลและทำความเข้าใจ (Dargan et al., 2020) ดังนั้นการพัฒนาอัลกอริทึมต้นไม้ตัดสินใจซึ่งมีจุดเด่นในเรื่องของผลลัพธ์ที่สามารถทำความเข้าใจได้ง่ายและใช้ระยะเวลาในการประมวลผลที่รวดเร็วจึงเป็นอีกแนวทางหนึ่งสำหรับการเพิ่มประสิทธิภาพการจำแนกข้อมูลได้

จากการทบทวนวรรณกรรมข้างต้นจะพบว่าเทคนิคต้นไม้ตัดสินใจยังคงมีข้อจำกัด ดังนี้

1. **ปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลาย**

เมื่อใช้ค่าเกนสารสนเทศเป็นเกณฑ์ในการเลือกโหนดของต้นไม้ตัดสินใจจะทำให้แอตทริบิวต์ที่มีค่าข้อมูลหลากหลายมีโอกาสถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากกว่าแอตทริบิวต์อื่น ๆ ซึ่งหากแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายเป็นแอตทริบิวต์ที่มีความสำคัญกับคลาสที่ต้องการจำแนกน้อยกว่าแอตทริบิวต์อื่น ๆ แล้ว อาจทำให้ประสิทธิภาพในการจำแนกข้อมูลลดลง รวมถึงต้นไม้ตัดสินใจที่ได้อาจมีความผิดปกติ ดังตัวอย่างในภาพ 1 ซึ่งเป็นการสร้างต้นไม้ตัดสินใจสำหรับการวินิจฉัยโรคจากข้อมูลผู้ป่วยและอาการป่วยเบื้องต้น ซึ่งประกอบไปด้วยแอตทริบิวต์ต่าง ๆ ดังนี้ อายุ (Age) อาการไอ (cough) และอาการไข้ (Fever)



ภาพ 1 ตัวอย่างขั้นตอนการสร้างต้นไม้ตัดสินใจสำหรับการวินิจฉัยโรคหวัด

4618891838 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

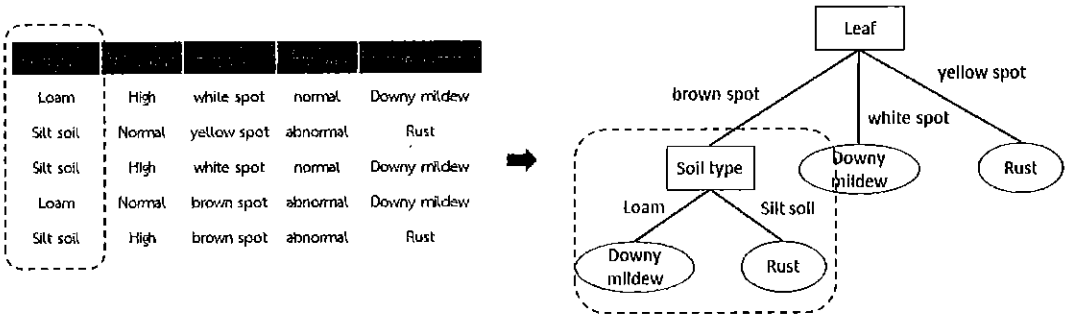
จากภาพ 1 จะพบว่าแอตทริบิวต์อายุเป็นแอตทริบิวต์ที่มีค่าข้อมูลที่แตกต่างกัน จำนวน 3 ค่า คือ <20, 20-60 และ >60 ในขณะที่แอตทริบิวต์อาการไอและอาการไข้เป็นแอตทริบิวต์ที่มีค่าข้อมูลเพียง 2 ค่า คือ yes และ no เมื่อทำการคำนวณค่าเกนสารสนเทศเพื่อพิจารณาโหนดรากของต้นไม้ตัดสินใจจะพบว่าแอตทริบิวต์อายุซึ่งเป็นแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายมากกว่าแอตทริบิวต์อื่น ๆ จะถูกเลือกเป็นโหนดรากของต้นไม้ตัดสินใจเนื่องจากมีค่าเกนสารสนเทศสูงที่สุด

โดยทั่วไปในการวินิจฉัยโรคผู้เชี่ยวชาญมักจะพิจารณาอาการผิดปกติที่เกิดขึ้นกับผู้ป่วยเพื่อระบุการเกิดโรค ในขณะที่อายุของผู้ป่วยจะใช้ในการพิจารณาความเสี่ยงของการเกิดโรคและความรุนแรงของโรค ดังนั้นแอตทริบิวต์ที่เกี่ยวข้องกับอาการผิดปกติของผู้ป่วยจึงมีความสำคัญต่อการวินิจฉัยโรคมากกว่าแอตทริบิวต์อายุ จากกระบวนการวินิจฉัยโรคนี้อาจกล่าวได้ว่าต้นไม้ตัดสินใจสำหรับการวินิจฉัยโรคไข้หวัดในภาพ 1 เป็นต้นไม้ตัดสินใจที่เกิดความผิดปกติ เนื่องจากเกิดปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลาย โดยมีการนำแอตทริบิวต์อายุมาใช้เป็นโหนดภายในต้นไม้ตัดสินใจ ในขณะที่แอตทริบิวต์อื่น ๆ ที่มีจำนวนค่าข้อมูลน้อยกว่าจะถูกพิจารณาเป็นอันดับถัดไปถึงแม้จะเป็นแอตทริบิวต์ที่มีความสำคัญมากกว่าก็ตาม

2. ปัญหาคุณภาพข้อมูล

หากในชุดข้อมูลที่นำมาวิเคราะห์ปรากฏข้อมูลรบกวน หรือมีแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสที่ต้องการจำแนกปะปนในชุดข้อมูล จะส่งผลให้ประสิทธิภาพในการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจลดลง ตัวอย่างดังภาพ 2 ซึ่งแสดงต้นไม้ตัดสินใจที่สร้างจากชุดข้อมูลที่ปรากฏแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสที่ต้องการจำแนก ซึ่งชุดข้อมูลนี้เป็นข้อมูลที่เกี่ยวข้องกับการจำแนกโรคของพืชที่ประกอบไปด้วยแอตทริบิวต์ชนิดของดินที่ปลูกพืช (Soil type) ความชื้นสัมพัทธ์ในอากาศ (Humidity) ลักษณะของใบ (Leaf) และลักษณะของลำต้น (Stem) เป็นต้น โดยแอตทริบิวต์ชนิดของดินที่ปลูกพืชเป็นแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสที่ต้องการจำแนก เนื่องจากการเกิดโรคของพืชนั้นไม่มีความเกี่ยวข้องกับชนิดของดินที่ใช้ในการเพาะปลูก ดังนั้นเมื่อนำชุดข้อมูลนี้มาทำการสร้างต้นไม้ตัดสินใจจึงทำแอตทริบิวต์นี้ปรากฏเป็นส่วนหนึ่งในต้นไม้ตัดสินใจ ทำให้ต้นไม้ตัดสินใจมีกฎการจำแนกข้อมูลที่ผิดปกติซึ่งอาจส่งผลต่อความถูกต้องในการจำแนกข้อมูลได้ เช่น จากภาพ 2 กฎการตัดสินใจสำหรับโรคราน้ำค้าง (Downny mildew) คือ เมื่อใบไม้ปรากฏจุดสีน้ำตาล (brown spot) และปลูกด้วยดินร่วน (loam) เป็นต้น

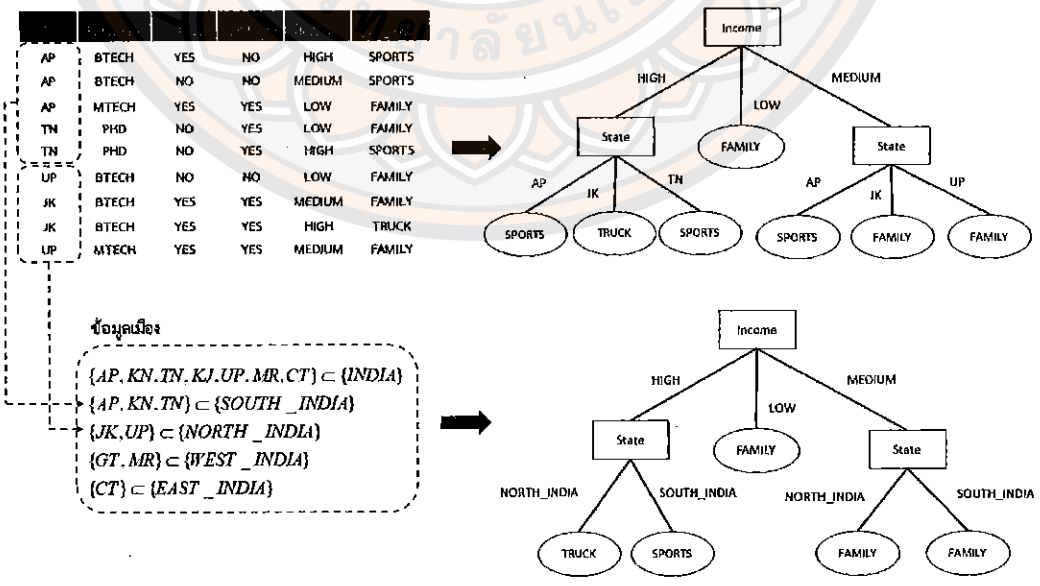
4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39



ภาพ 2 ตัวอย่างต้นไม้ตัดสินใจในกรณีที่ชุดข้อมูลปรากฏแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาส

3. ปัญหาประสิทธิภาพของอัลกอริทึมเมื่อทำงานกับชุดข้อมูลที่มีแอตทริบิวต์จำนวนมาก หรือมีค่าข้อมูลจำนวนมาก

หากชุดข้อมูลที่นำมาวิเคราะห์เป็นชุดข้อมูลที่มีแอตทริบิวต์ที่เกี่ยวข้องจำนวนมาก หรือในแอตทริบิวต์มีค่าข้อมูลจำนวนมากจะส่งผลให้ต้นไม้ตัดสินใจที่ได้มีความซับซ้อน โดยอาจเป็นต้นไม้ตัดสินใจที่มีความลึกมาก หรือมีโหนดจำนวนมาก รวมถึงใช้ระยะเวลานานในการประมวลผล ดังตัวอย่างในภาพ 3 ซึ่งเป็นต้นไม้ตัดสินใจสำหรับการจำแนกข้อมูลประเภทรถจากข้อมูลประชากรในประเทศอินเดียที่มีคุณลักษณะแตกต่างกัน โดยภาพจะแสดงให้เห็นถึงโครงสร้างของต้นไม้ตัดสินใจเมื่อมีการสร้างจากแอตทริบิวต์ที่มีจำนวนของข้อมูลที่แตกต่างกัน



ภาพ 3 ตัวอย่างต้นไม้ตัดสินใจในกรณีแอตทริบิวต์มีค่าข้อมูลแตกต่างกัน

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากภาพ 3 เมื่อแอตทริบิวต์ตำแหน่งที่ประชากรอาศัย (state) ประกอบไปด้วยชื่อเมืองจำนวน 4 เมือง ได้แก่ AP, TN, UP และ JK จะสามารถสร้างต้นไม้ตัดสินใจที่มีจำนวนโหนดทั้งสิ้น 10 โหนด ในขณะที่เมื่อแอตทริบิวต์ตำแหน่งที่ประชากรอาศัยมีการจัดเก็บข้อมูลเป็นภูมิภาคตามข้อมูลของเมืองที่ประชากรนั้นอาศัยอยู่ ซึ่งจะประกอบไปด้วย 2 ภูมิภาค คือ SOUTH_INDIA และ NORTH_INDIA จะทำให้ต้นไม้ตัดสินใจที่ได้มีจำนวนโหนดทั้งสิ้น 8 โหนด ซึ่งจะเห็นได้ว่าเมื่อนำชุดข้อมูลที่มีค่าข้อมูลจำนวนมากมาทำการสร้างแบบจำลองการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจจะส่งผลให้ได้ต้นไม้ตัดสินใจที่ได้มีความซับซ้อน

จากข้อจำกัดของเทคนิคต้นไม้ตัดสินใจผู้วิจัยจึงมีแนวความคิดที่จะนำเสนอวิธีการใหม่ในการปรับปรุงประสิทธิภาพของเทคนิคต้นไม้ตัดสินใจด้วยการประยุกต์ใช้องค์ความรู้และความสัมพันธ์ระหว่างข้อมูลซึ่งอยู่ในรูปแบบของออนโทโลยี โดย

- การปรับปรุงวิธีการคำนวณค่าเกนสารสนเทศด้วยการประยุกต์ใช้ออนโทโลยีร่วมกับอัลกอริทึม Weighted Semantic PageRank (Jun et al., 2016) เพื่อแก้ปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลาย
- การปรับปรุงคุณภาพของข้อมูลโดยการพิจารณาความสัมพันธ์ระหว่างข้อมูลโดยใช้สัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล (Point biserial correlation) และสถิติไคสแควร์ (Chi-square) เพื่อนำแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสที่ต้องการจำแนกออกจากชุดข้อมูล
- การนำองค์ความรู้และความสัมพันธ์ระหว่างข้อมูลในออนโทโลยีมาใช้ในกระบวนการแปลงข้อมูลเพื่อลดจำนวนค่าข้อมูลในแต่ละแอตทริบิวต์ ซึ่งช่วยในการแก้ปัญหาประสิทธิภาพของอัลกอริทึมเมื่อทำงานกับแอตทริบิวต์ที่มีค่าข้อมูลจำนวนมาก

การสร้างต้นไม้ตัดสินใจด้วยวิธีการใหม่ช่วยให้สามารถหลีกเลี่ยงการสร้างต้นไม้ตัดสินใจที่มีความผิดปกติ ลดความความซับซ้อนของต้นไม้ตัดสินใจ รวมทั้งช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลอีกด้วย

จุดมุ่งหมายของการวิจัย

1. เพื่อนำเสนอวิธีการประยุกต์ใช้ออนโทโลยีในการสนับสนุนการเตรียมข้อมูลสำหรับการวิเคราะห์ข้อมูล

2. เพื่อพัฒนาวิธีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีร่วมกับเทคนิคต้นไม้ตัดสินใจในการปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจ โดยการประยุกต์ใช้องค์ความรู้ที่เกี่ยวข้องมาช่วยคำนวณหาค่าเกนสารสนเทศ ให้มีประสิทธิภาพมากยิ่งขึ้น

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ขอบเขตการวิจัย

การวิจัยนี้เป็นการวิจัยเพื่อประยุกต์ใช้ออนโทโลยีในการปรับปรุงประสิทธิภาพในการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจ ซึ่งมีขอบเขตในการดำเนินงานดังนี้

1. การวิจัยในครั้งนี้จะใช้ชุดข้อมูลมาตรฐานซึ่งเผยแพร่โดยมหาวิทยาลัยแคลิฟอร์เนีย (University of California) (Dua & Graff, 2017) และชุดข้อมูลซึ่งได้รับการเผยแพร่แหล่งข้อมูลต่าง ๆ ในการดำเนินงานจำนวน 2 ด้าน ดังนี้

- ข้อมูลทางด้านการเกษตร คือ ชุดข้อมูลการเกิดโรคของถั่วเหลือง
- ข้อมูลทางด้านการแพทย์ คือ ชุดข้อมูลข้อมูลผู้ป่วยโรคหัวใจ ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

2. การปรับปรุงประสิทธิภาพการจำแนกข้อมูลจะทำการปรับปรุงเทคนิคต้นไม้ตัดสินใจโดยใช้อัลกอริทึม ID3 ซึ่งมีการใช้ค่าเกนสารสนเทศเป็นเกณฑ์ในการพิจารณาโหนดภายในต้นไม้ตัดสินใจ

สมมุติฐานของการวิจัย

- **สมมติฐานการวิจัยที่** 1 การแปลงข้อมูลโดยการประยุกต์ใช้ออนโทโลยีสามารถช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลได้

เนื่องจากในการสร้างต้นไม้ตัดสินใจนั้นจะทำการพิจารณาข้อมูลในแต่ละแอตทริบิวต์เพื่อทำการแตกกิ่งของต้นไม้ตัดสินใจ หากมีแอตทริบิวต์จำนวนมากหรือข้อมูลในแอตทริบิวต์มีความหลากหลายจะทำให้ต้นไม้ตัดสินใจมีความลึกมากและส่งผลต่อประสิทธิภาพในการจำแนกข้อมูล เนื่องจากเทคนิคต้นไม้ตัดสินใจจะทำการแตกกิ่งของต้นไม้ตัดสินใจออกไปโดยการแบ่งข้อมูลออกเป็นชุดจนกระทั่งข้อมูลถูกจัดอยู่ในคลาสเดียวกันทั้งหมด หรือจนกระทั่งไม่สามารถแตกกิ่งออกไปได้อีก ซึ่งข้อมูลที่ใช้ในการพิจารณาการแตกกิ่งเพื่อสร้างโหนดเมื่อต้นไม้ตัดสินใจเมื่อมีระดับความลึกอาจมีจำนวนน้อยมากจนทำให้โหนดนั้นไม่มีนัยสำคัญกับการจำแนกข้อมูล แต่ทำให้ต้นไม้ตัดสินใจที่ได้มีความซับซ้อน และเมื่อนำต้นไม้ตัดสินใจนั้นไปใช้ในการจำแนกข้อมูลที่ไม่เคยเรียนรู้มาก่อนจึงทำให้ประสิทธิภาพในการจำแนกข้อมูลลดลง ซึ่งการประยุกต์ใช้ออนโทโลยีในการอ้างอิงข้อมูลที่มีความสัมพันธ์กับข้อมูลในแต่ละแอตทริบิวต์ และทำการแปลงข้อมูลนั้นด้วยข้อมูลที่มีความสัมพันธ์กันจะช่วยลดจำนวนข้อมูลที่ต้องพิจารณาในการสร้างต้นไม้ตัดสินใจ และส่งผลให้ต้นไม้ตัดสินใจที่ได้มีความลึกลดลง เวลาในการสร้างต้นไม้ตัดสินใจลดลง รวมทั้งช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูล

- **สมมติฐานการวิจัยที่** 2 การปรับปรุงกระบวนการคำนวณค่าเกนสารสนเทศใหม่โดยใช้องค์ความรู้ในออนโทโลยี จะช่วยให้การสร้างต้นไม้ตัดสินใจมีประสิทธิภาพมากยิ่งขึ้น

ในการพิจารณาโหนดสำหรับต้นไม้ตัดสินใจนั้น โดยปกติจะพิจารณาจากค่าเกนสารสนเทศที่มีค่ามากที่สุด โดยค่าเกนสารสนเทศนี้จะคำนวณจากชุดข้อมูลและมีข้อจำกัดในเรื่องการลำเอียงไปยัง

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

แอตทริบิวต์ที่มีค่าหลากหลายซึ่งอาจเป็นแอตทริบิวต์ไม่มีความสำคัญในการจำแนกข้อมูล การนำออนโทโลยีที่มีโครงสร้างแบบลำดับชั้นที่สามารถแสดงความสัมพันธ์ระหว่างข้อมูลมาใช้ในการพิจารณาค่าความสำคัญของแต่ละแอตทริบิวต์ และนำค่าความสำคัญนั้นไปใช้ในการปรับปรุงค่าเกนสารสนเทศจะทำให้แอตทริบิวต์ที่มีความสำคัญมากแต่มีค่าเกนสารสนเทศน้อยมีโอกาสในการถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากขึ้น ในขณะที่แอตทริบิวต์ที่มีค่าเกนสารสนเทศมากแต่มีความสำคัญน้อยมีโอกาสเป็นโหนดของต้นไม้ตัดสินใจน้อยลง ซึ่งสามารถช่วยลดความลำเอียงไปยังแอตทริบิวต์ที่มีความหลากหลาย และช่วยให้สามารถเลือกแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดในต้นไม้ตัดสินใจได้เหมาะสมมากขึ้น การพิจารณาความสำคัญของแอตทริบิวต์ที่ได้จากออนโทโลยีร่วมกับค่าเกนสารสนเทศนี้จะเป็นเกณฑ์ใหม่สำหรับการพิจารณาโหนดของต้นไม้ตัดสินใจ

นิยามศัพท์เฉพาะ

ความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting) คือ เหตุการณ์ที่แบบจำลองในการจำแนกข้อมูลนั้นสามารถทำการจำแนกข้อมูลที่ใช้สำหรับฝึกสอนได้ถูกต้องแต่จะมีประสิทธิภาพในการจำแนกข้อมูลลดลงเมื่อทำการจำแนกข้อมูลที่ไม่เคยเรียนรู้มาก่อน

แนวความคิดพื้นฐาน (Abstract Value) คือ ข้อมูลที่เป็นแนวความคิดทั่วไปของแนวความคิดที่มีความสัมพันธ์ เช่น ในกรณีที่อธิบายถึงประเทศต่าง ๆ ไม่ว่าจะเป็น ประเทศจีน ประเทศอินเดีย หรือ ประเทศไทย จะมีแนวความคิดพื้นฐานเป็นประเทศในทวีปเอเชีย หรือในกรณีที่อธิบายถึงโรคพืช โรคแอนแทรคโนส โรคราแป้ง หรือโรคราน้ำค้าง จะมีแนวความคิดพื้นฐานเป็นโรคพืชที่มีสาเหตุจากเชื้อรา เป็นต้น

แอตทริบิวต์ที่มีค่าข้อมูลหลากหลาย คือ แอตทริบิวต์ที่มีชนิดข้อมูลเป็นข้อมูลแบบกลุ่ม (Categorical Data) ซึ่งค่าข้อมูลหลายค่า เช่น ในชุดข้อมูลผู้ป่วยโรคหัวใจ แอตทริบิวต์ sex จะมีค่าข้อมูล 2 ค่า คือ เพศหญิง และ เพศชาย แอตทริบิวต์ Cp จะมีค่าแสดงรูปแบบของอาการเจ็บหน้าอกทั้งหมด 4 รูปแบบ คือ 1) Typical Angina 2) Atypical Angina 3) Non-anginal Pain และ 4) Asymptomatic ดังนั้น แอตทริบิวต์ Cp จะเป็นแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายมากกว่าแอตทริบิวต์ sex

บทสรุป

ในบทนี้ได้กล่าวถึงปัญหาต่าง ๆ ที่เกิดขึ้นเมื่อมีการนำเทคนิคต้นไม้ตัดสินใจมาใช้ในการสร้างแบบจำลองการจำแนกข้อมูล ซึ่งประกอบไปด้วย 1) ปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายเมื่อมีการใช้ค่าเกนสารสนเทศเป็นเกณฑ์ในการพิจารณาโหนดของต้นไม้ตัดสินใจ 2) ปัญหาคุณภาพข้อมูลซึ่งภายในชุดข้อมูลปรากฏข้อมูลรบกวน หรือมีแอตทริบิวต์ที่ไม่มีความสัมพันธ์

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

กับคลาสที่ต้องการจำแนกปะปนในชุดข้อมูล และ 3) ปัญหาประสิทธิภาพของอัลกอริทึมเมื่อทำงานกับเมื่อชุดข้อมูลที่มีแอตทริบิวต์จำนวนมากหรือมีค่าข้อมูลจำนวนมาก ซึ่งปัญหาเหล่านี้ส่งผลต่อประสิทธิภาพในการจำแนกข้อมูล เนื่องจากการเลือกแอตทริบิวต์ที่ไม่เหมาะสมเป็นโหนดภายในต้นไม้ตัดสินใจ การสร้างต้นไม้ตัดสินใจที่มีความซับซ้อน เช่น มีความลึกมาก หรือจำนวนโหนดภายในต้นไม้ตัดสินใจจำนวนมากโดยอาจทำให้เกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ได้ ซึ่งการประยุกต์ใช้องค์ความรู้ซึ่งอยู่ในรูปแบบออนโทโลยีมาช่วยในกระบวนการจัดเตรียมข้อมูลเพื่อลดจำนวนและความหลากหลายของข้อมูล รวมถึงการพิจารณาความสำคัญของแอตทริบิวต์จากองค์ความรู้ที่เกี่ยวข้องในออนโทโลยีเพื่อปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจเป็นแนวทางหนึ่งที่ช่วยลดปัญหาที่เกิดขึ้นได้

ในบทถัดไปจะนำเสนอทฤษฎีที่เกี่ยวข้อง เช่น กระบวนการสร้างต้นไม้ตัดสินใจ ออนโทโลยี รวมถึงงานวิจัยต่าง ๆ ที่ใช้สำหรับการดำเนินการปรับปรุงประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยี

4918891B3 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยในครั้งนี้ผู้วิจัยได้ทำการศึกษาแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องกับการปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจเพื่อประยุกต์ใช้ในการดำเนินงานวิจัย ประกอบด้วยหัวข้อต่าง ๆ ดังนี้

- เหมืองข้อมูล
- เหมืองข้อมูลเชิงความหมาย
- อัลกอริทึมต้นไม้ตัดสินใจ
- การประเมินประสิทธิภาพแบบจำลองการจำแนกข้อมูล
- ออนโทโลยี
- วิธีการสรุปภาพรวมออนโทโลยี
- อัลกอริทึม PageRank
- งานวิจัยที่เกี่ยวข้อง

เหมืองข้อมูล (Data Mining)

เหมืองข้อมูล (Data Mining) คือ กระบวนการในการวิเคราะห์ข้อมูลเพื่อค้นหาองค์ความรู้ที่แฝงอยู่ภายในข้อมูลนั้นโดยมีการประยุกต์ใช้เทคโนโลยีต่าง ๆ เช่น สถิติ เทคนิคการเรียนรู้ของเครื่อง (Machine learning) การทำเหมืองข้อมูลสามารถนำมาใช้วิเคราะห์ข้อมูลได้หลากหลายรูปแบบ เช่น ข้อมูลที่จัดเก็บในฐานข้อมูล คลังข้อมูล ข้อมูลเว็บไซต์ ข้อมูลสื่อประสม รวมถึงข้อมูลสตรีมมิ่ง (Streaming data) เป็นต้น (Han et al., 2011; Hand, 2007)

การวิเคราะห์ข้อมูลโดยวิธีการทำเหมืองข้อมูลนั้นสามารถแบ่งออกได้เป็น 2 ประเภท ได้แก่

- **เทคนิคการเรียนรู้แบบไม่มีผู้สอน** (Unsupervised learning) ซึ่งจะเน้นการพิจารณาข้อมูลเป็นหลัก เช่น พิจารณาว่าข้อมูลมีความสัมพันธ์กันอย่างไร วิธีการทำเหมืองข้อมูลที่อยู่ในกลุ่มนี้ ได้แก่ การหากฎความสัมพันธ์ (Association rules) และการแบ่งกลุ่มข้อมูล (Clustering)

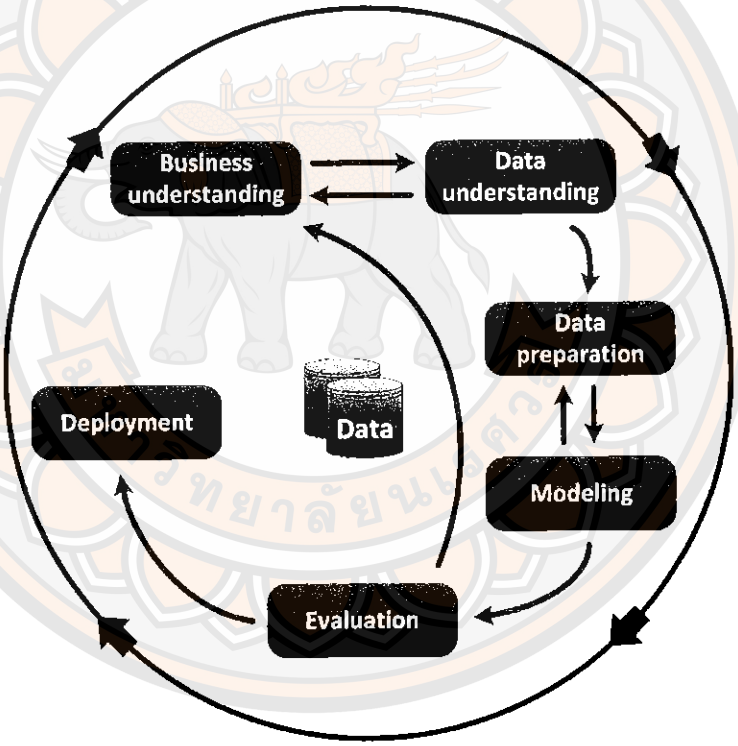
- **เทคนิคการเรียนรู้แบบมีผู้สอน** (Supervised learning) จะเป็นวิธีการที่นำข้อมูลที่มีอยู่ในอดีตมาใช้ในการสร้างแบบจำลองเพื่อคาดการณ์สิ่งที่เกิดขึ้นในอนาคต วิธีการทำเหมืองข้อมูลที่จัดอยู่ในกลุ่มนี้ได้แก่ การจำแนกข้อมูล (Classification) ซึ่งการจำแนกข้อมูลมีหลายเทคนิคที่ได้รับ

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ความนิยม เช่น เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors) และ เทคนิคโครงข่ายประสาทเทียม (Neural Network) เป็นต้น

1. กระบวนการของเหมืองข้อมูล (Data Mining Process)

กระบวนการวิเคราะห์ข้อมูลโดยใช้เทคนิคเหมืองข้อมูลนั้นสามารถอธิบายขั้นตอนได้โดยใช้แบบจำลอง CRISP-DM (Cross-Industry Standard Process for Data Mining) ซึ่งประกอบไปด้วยการทำงานทั้งสิ้น 6 ขั้นตอน ได้แก่ การทำความเข้าใจปัญหา (Business understanding) การทำความเข้าใจข้อมูล (Data understanding) การเตรียมข้อมูล (Data preparation) การสร้างแบบจำลอง (Modeling) การประเมินประสิทธิภาพแบบจำลอง (Evaluation) และ การนำไปใช้งาน (Deployment) (Witten et al., 2017) ดังแสดงในภาพ 4



ภาพ 4 กระบวนการทำเหมืองข้อมูลตามแบบจำลอง CRISP-DM

- **การทำความเข้าใจปัญหา** (Business understanding) คือ ขั้นตอนสำหรับการทำความเข้าใจเป้าหมายและความต้องการทางธุรกิจรวมถึงปัญหาที่เกิดขึ้น เพื่อทำการพิจารณาว่าการทำเหมืองข้อมูลจะสามารถนำไปใช้ในการแก้ปัญหาหรือตอบสนองความต้องการใดขององค์การ ในขั้นตอนนี้ยังรวมถึงการพิจารณาถึงข้อมูลที่จำเป็นสำหรับการสร้างแบบจำลองที่ใช้ในการแก้ปัญหาเหล่านั้น

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

- **การทำความเข้าใจข้อมูล** (Data understanding) คือ ขั้นตอนของการเก็บรวบรวมข้อมูล การตรวจสอบความถูกต้องของข้อมูลที่รวบรวมได้ รวมถึงพิจารณาว่ามีปริมาณข้อมูลที่เพียงพอต่อการนำไปใช้งานหรือไม่ ซึ่งการพิจารณาข้อมูลในขั้นตอนนี้อาจเป็นส่วนหนึ่งที่ทำให้ทราบถึงบริบทต่าง ๆ ที่เกี่ยวข้องกับการดำเนินงานขององค์กร ซึ่งอาจส่งผลให้มีการพิจารณาถึงเป้าหมายของการทำเหมืองข้อมูลอีกครั้ง

- **การเตรียมข้อมูล** (Data preparation) คือ ขั้นตอนของการจัดเตรียมข้อมูลที่ได้รวบรวมมาให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำไปใช้ในการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล ซึ่งกระบวนการจัดเตรียมข้อมูลจะเกี่ยวข้องกับการทำงานต่าง ๆ เช่น การคัดเลือกข้อมูล (Data selection) การทำความสะอาดข้อมูล (Data cleansing) การแปลงรูปแบบของข้อมูล (Data transformation) เป็นต้น

- **การสร้างแบบจำลอง** (Modeling) คือ ขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล ซึ่งจะเกี่ยวข้องกับการพิจารณาเทคนิคที่เหมาะสมในการนำมาวิเคราะห์ข้อมูล การสร้างแบบจำลองและการทดสอบแบบจำลองที่ใช้ในการวิเคราะห์ข้อมูล โดยการพิจารณาว่าจะนำเทคนิคใดมาใช้ในการวิเคราะห์ข้อมูลนั้นจะขึ้นอยู่กับลักษณะของปัญหาและลักษณะของข้อมูลที่นำมาวิเคราะห์ ในการสร้างแบบจำลองนั้นจะต้องมีการพิจารณาและปรับปรุงค่าพารามิเตอร์ที่เกี่ยวข้องของแต่ละเทคนิคเพื่อให้ได้แบบจำลองการวิเคราะห์ข้อมูลที่มีประสิทธิภาพมากที่สุด

- **การประเมินประสิทธิภาพแบบจำลอง** (Evaluation) คือ ขั้นตอนการตรวจสอบประสิทธิภาพของแบบจำลอง โดยการพิจารณาผลลัพธ์ที่ได้ว่าถูกต้อง มีความน่าเชื่อถือมากน้อยเพียงใด รวมทั้งสอดคล้องตามเป้าหมายที่กำหนดได้หรือไม่ ซึ่งอาจมีการดำเนินการย้อนกลับไปเพื่อปรับปรุงให้ผลลัพธ์ที่ได้ตรงตามเป้าหมายที่วางไว้

- **การนำไปใช้งาน** (Deployment) คือ ขั้นตอนของการนำแบบจำลองการวิเคราะห์ข้อมูลที่ได้ไปใช้งานจริง ซึ่งอาจมีการจัดทำคู่มือการใช้งาน การระบุความต้องการขั้นพื้นฐานสำหรับซอฟต์แวร์ที่เกี่ยวข้อง แผนการติดตามและปรับปรุงระบบ เป็นต้น (Schröer et al., 2021; Witten et al., 2017)

2. **การเตรียมข้อมูล (Data preparation)**

การเตรียมข้อมูลเป็นขั้นตอนหนึ่งของกระบวนการทำเหมืองข้อมูล โดยการเตรียมข้อมูลนี้จะเป็นกระบวนการที่ใช้ระยะเวลานานในการดำเนินการเพื่อให้ได้ข้อมูลที่มีคุณภาพเพียงพอและพร้อมสำหรับนำไปใช้ในการวิเคราะห์ข้อมูลด้วยเทคนิคต่าง ๆ ซึ่งการนำข้อมูลที่มีคุณภาพน้อยไปวิเคราะห์นั้นอาจทำให้ได้ผลลัพธ์ที่ไม่ถูกต้องหรือมีความคลาดเคลื่อน โดยในการพิจารณาว่าข้อมูลมีคุณภาพมากน้อยเพียงใดจะพิจารณาจากองค์ประกอบต่าง ๆ ดังนี้

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

- ความถูกต้องของข้อมูล (accuracy) คือ การพิจารณาว่าภายในชุดข้อมูลมีการจัดเก็บข้อมูลที่ผิดพลาดหรือมีค่าคลาดเคลื่อนไปจากค่าที่ควรจะเป็นหรือไม่ เช่น การจัดเก็บข้อมูลลูกค้าคนหนึ่ง โดยระบุว่าลูกค้ามีอายุ 400 ปี ซึ่งสามารถพิจารณาได้ว่าเป็นค่าที่ผิดปกติ

- ความสมบูรณ์ของข้อมูล (completeness) เช่น การพิจารณาว่าภายในชุดข้อมูลมีข้อมูลสูญหายหรือไม่ มีแอตทริบิวต์ที่สำคัญขาดหายไปหรือไม่ เป็นต้น

- ความสอดคล้องกันของข้อมูล (consistency) เช่น การจัดเก็บข้อมูลวันที่ซึ่งมีการจัดเก็บในรูปแบบ วัน/เดือน/ปี แต่มีข้อมูลบางรายการจัดเก็บข้อมูลในรูปแบบ เดือน/วัน/ปี ซึ่งเป็นการจัดเก็บข้อมูลที่ไม่สอดคล้องกัน เป็นต้น

- ข้อมูลทันต่อการใช้งาน (timeliness) ซึ่งหมายถึง มีข้อมูลที่ทันสมัยเพียงพอ หรือมีข้อมูลที่ทันต่อการใช้งานของผู้ใช้

- ความน่าเชื่อถือของข้อมูล (believability) ซึ่งหมายถึง ข้อมูลนั้นเป็นข้อมูลจริง

- ความสามารถในการแปลความหมาย (interpretability) หมายถึง ข้อมูลนั้นมีความหมายชัดเจน ใช้ภาษา สัญลักษณ์ หรือหน่วยวัดปริมาณที่เหมาะสม สามารถทำความเข้าใจได้ง่าย

การเตรียมข้อมูลเพื่อปรับปรุงคุณภาพของข้อมูลนั้นจะประกอบไปด้วยการทำงานที่สำคัญ คือ การทำความสะอาดข้อมูล (data cleansing) การผสานข้อมูล (data integration) การลดรูปข้อมูล (data reduction) และการแปลงข้อมูล (data transformation) โดยสามารถอธิบายรายละเอียดได้ดังนี้

- **การทำความสะอาดข้อมูล** (data cleansing) เป็นกระบวนสำหรับการจัดการกับข้อมูลที่สูญหาย (missing data) ข้อมูลรบกวน (noise) หรือข้อมูลที่ผิดปกติ (outlier) ซึ่งสามารถดำเนินการได้หลายวิธี เช่น การแก้ปัญหาข้อมูลสูญหายด้วยการเติมค่าที่เหมาะสม หรือการลบข้อมูลที่ผิดปกติออกจากชุดข้อมูล เป็นต้น

- **การผสานข้อมูล** (data integration) ในการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลนั้นมักจะมีการรวบรวมข้อมูลจากหลาย ๆ แหล่ง ดังนั้นในกระบวนการสำหรับการรวมข้อมูลจากแหล่งต่าง ๆ นั้นจึงเป็นกระบวนการที่มีการดำเนินการอย่างระมัดระวัง เพื่อหลีกเลี่ยงปัญหาความซ้ำซ้อนและความไม่สอดคล้องกันของข้อมูล

- **การลดรูปข้อมูล** (data reduction) คือ กระบวนการในการลดขนาดของชุดข้อมูลให้มีขนาดลดลง โดยที่ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลที่ทำการปรับปรุงนี้มีค่าไม่แตกต่างจากการใช้ข้อมูลเดิมหรือมีผลผลัพธ์ที่ดีขึ้น ซึ่งการลดรูปข้อมูลนี้จะเกี่ยวข้องกับการลดมิติของข้อมูล (dimensionality reduction) เช่น การคัดเลือกเฉพาะแอตทริบิวต์ที่มีความสำคัญไปใช้งาน การสร้างแอตทริบิวต์ใหม่โดยการยุบรวมแอตทริบิวต์ที่มีความสัมพันธ์เข้าด้วยกัน เป็นต้น

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

- **การแปลงข้อมูล** (data transformation) ในกระบวนการนี้ข้อมูลจะถูกแปลงให้อยู่ในรูปแบบที่สามารถช่วยเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูลโดยการใช้เทคนิคต่าง ๆ เช่น นอร์มัลไลเซชัน (Normalization) ซึ่งจะทำการแปลงข้อมูลที่เป็นค่าตัวเลขให้มีค่าข้อมูลอยู่ในช่วงข้อมูลที่กำหนด เช่น [-1, 1] หรือ [0, 1] การแปลงข้อมูลตัวเลขออกเป็นกลุ่มข้อมูลต่าง ๆ เช่น การแบ่งข้อมูลอายุที่เป็นตัวเลขออกเป็นช่วงอายุ 0-10 ปี อายุ 11- 20 ปี หรือช่วงอายุอื่น ๆ ตามที่กำหนด เป็นต้น (Han et al., 2011)

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

เหมืองข้อมูลเชิงความหมาย (Semantic Data Mining)

ในปัจจุบันข้อมูลมีความหลากหลายทั้งรูปแบบโครงสร้างของข้อมูล และแหล่งที่มาของข้อมูล ทำให้ในการทำเหมืองข้อมูลจะต้องมีการดำเนินการกับข้อมูลเหล่านั้นเพื่อให้พร้อมกับการนำไปวิเคราะห์ข้อมูล ซึ่งเครื่องมือในการทำเหมืองข้อมูลในปัจจุบันยังมีข้อจำกัด คือ สามารถทำงานได้ข้อมูลซึ่งอยู่ในรูปแบบของตาราง และไม่รองรับข้อมูลที่ไม่มีโครงสร้างและความสัมพันธ์ระหว่างข้อมูลที่ซับซ้อน (Ławrynowicz, 2017) ดังนั้นจึงได้มีนักวิจัยพัฒนาเทคนิคการทำเหมืองข้อมูลเชิงสัมพันธ์ (Relational Data Mining) ซึ่งสามารถทำการวิเคราะห์ข้อมูลที่อยู่ในรูปแบบของฐานข้อมูลเชิงสัมพันธ์ กราฟ หรือ เซตของข้อมูล ซึ่งวิธีการนี้จะนำความสัมพันธ์ของข้อมูลมาใช้ในการวิเคราะห์ข้อมูลซึ่งช่วยให้สามารถวิเคราะห์ข้อมูลได้มีประสิทธิภาพมากขึ้น อย่างไรก็ตามการวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูลเชิงสัมพันธ์นั้นยังมีข้อจำกัดเมื่อนำมาวิเคราะห์ข้อมูลที่มีปริมาณมาก ๆ เนื่องจากข้อมูลเหล่านั้นจะมีการเชื่อมโยงกันระหว่างข้อมูลจึงทำให้ข้อมูลมีความซับซ้อน ส่งผลให้ต้องใช้ทรัพยากรและระยะเวลานานในการค้นหาองค์ความรู้ที่อยู่ในข้อมูลเหล่านั้น ดังนั้นนักวิจัยจึงได้นำเสนอวิธีการทำเหมืองข้อมูลอีกวิธีหนึ่งที่มีการนำความรู้และความสัมพันธ์ระหว่างข้อมูลมาช่วยสนับสนุนการวิเคราะห์ข้อมูล ซึ่งวิธีการนี้เรียกว่าเหมืองข้อมูลเชิงความหมาย (Semantic Data Mining)

เหมืองข้อมูลเชิงความหมาย (Semantic Data Mining) คือ วิธีการวิเคราะห์ข้อมูลโดยใช้เทคนิคเหมืองข้อมูลที่มีการประยุกต์ใช้องค์ความรู้ที่อยู่ในรูปแบบของออนโทโลยีเป็นความรู้พื้นฐานในการวิเคราะห์ข้อมูล รวมไปถึงการวิเคราะห์ข้อมูลซึ่งอยู่ในรูปแบบของออนโทโลยีและองค์ความรู้ที่แสดงอยู่ในรูปแบบของกราฟ (Knowledge Graph) อีกด้วย (Ławrynowicz, 2017)

1. **บทบาทของออนโทโลยีในการทำเหมืองข้อมูลเชิงความหมาย**

การนำองค์ความรู้ที่อยู่ในรูปแบบออนโทโลยีมาสนับสนุนการทำเหมืองข้อมูลนั้น ออนโทโลยีจะมีส่วนช่วยในประเด็นต่าง ๆ เช่น

- **การลดปัญหาช่องว่างทางความหมาย** (Semantic Gap) นักวิจัยทางด้านเหมืองข้อมูลส่วนใหญ่ระบุว่าปัญหาช่องว่างทางความหมายซึ่งหมายถึงการตีความข้อมูลที่แตกต่างกันระหว่าง

เครื่องคอมพิวเตอร์และมนุษย์ (Koopman et al., 2016) ซึ่งปัญหานี้สามารถเกิดขึ้นได้ในทุกขั้นตอนของการทำเหมืองข้อมูล ตัวอย่างเช่น ในขั้นตอนการเตรียมข้อมูล ในชุดข้อมูลที่มีแอตทริบิวต์จำนวนมาก การพิจารณาความสัมพันธ์ระหว่างข้อมูลสำหรับการรวมแอตทริบิวต์ที่มีความสัมพันธ์กันสูงเพื่อลดจำนวนแอตทริบิวต์ที่ใช้ในการวิเคราะห์ข้อมูลนั้นจำเป็นต้องใช้ผู้เชี่ยวชาญในการพิจารณา ซึ่งการนำออนโทโลยีมาช่วยในการพิจารณาความสัมพันธ์ระหว่างข้อมูลนี้เป็นวิธีการหนึ่งที่สามารถช่วยแก้ปัญหาที่เกิดขึ้นได้ โดยในการวิจัยของ B. Zhou et al. (2020) ได้ประยุกต์ใช้ออนโทโลยีในการจัดเตรียมข้อมูลสำหรับการวิเคราะห์ข้อมูลด้านการผลิตที่ใช้ในโรงงานอุตสาหกรรม โดยออนโทโลยีจะช่วยในการพิจารณากลุ่มของข้อมูลที่มีความสัมพันธ์กัน เช่น ข้อมูลกระแสไฟฟ้า (current) ข้อมูลค่าแรงดันไฟฟ้า (voltage) และ ข้อมูลค่าความต้านทาน (resistance) จะถูกจัดอยู่ในกลุ่มข้อมูลอนุกรมเวลา (Time series) ซึ่งสามารถนำมาประมวลผลร่วมกันเพื่อได้เป็นแอตทริบิวต์ใหม่ที่ใช้ในการวิเคราะห์ข้อมูล เป็นต้น

สำหรับตัวอย่างของการแก้ปัญหาช่องว่างความหมายที่เกิดขึ้นกับผู้ใช้งานและผลลัพธ์ที่ได้จากเทคนิคการค้นหากฎความสัมพันธ์นั้น โดยเทคนิคการค้นหากฎความสัมพันธ์นั้นอาจสร้างกฎที่มีความซ้ำซ้อนกันทำให้เกิดกฎความสัมพันธ์จำนวนมาก ซึ่งการพิจารณาว่ากฎความสัมพันธ์ใดเป็นกฎที่มีความซ้ำซ้อนกันนั้นจะดำเนินการโดยผู้ใช้งาน ซึ่งออนโทโลยีสามารถช่วยในการคัดกรองกฎความสัมพันธ์ที่มีความซ้ำซ้อนกันนี้ได้โดยการพิจารณาความสัมพันธ์ระหว่างข้อมูล เช่น Gahar et al. (2018) ได้ประยุกต์ใช้ออนโทโลยีในการพิจารณากฎความสัมพันธ์ของข้อมูล โดยพิจารณาว่ารายการข้อมูล (item) ในแต่ละกฎความสัมพันธ์ปรากฏอยู่ภายใต้แนวความคิดเดียวกันในออนโทโลยีหรือไม่ หากรายการข้อมูลเหล่านั้นอยู่ภายใต้แนวความคิดเดียวกันจะหมายถึง กฎความสัมพันธ์นั้นเป็นกฎซ้ำซ้อนและสามารถตัดกฎความสัมพันธ์ที่ซ้ำซ้อนนั้นออกได้ ซึ่งจะการพิจารณาความสัมพันธ์ระหว่างข้อมูลสามารถช่วยลดจำนวนกฎความสัมพันธ์ที่ต้องพิจารณา และส่งผลให้ผู้ใช้งานสามารถทำความเข้าใจผลลัพธ์ได้ดีขึ้น

- **การขยายผลลัพธ์จากการทำเหมืองข้อมูล** องค์ความรู้หรือรูปแบบของข้อมูลที่แฝงอยู่ในข้อมูลซึ่งเป็นผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลนั้นมักจะต้องทำการแปลความหมายหรือตีความตามองค์ความรู้ในแต่ละโดเมน ซึ่งการนำออนโทโลยีซึ่งมีการแสดงความสัมพันธ์ระหว่างข้อมูลมาช่วยในการตีความหรือแปลความหมาย รวมถึงขยายความผลลัพธ์ที่เกิดขึ้นให้สอดคล้องตามองค์ความรู้ในโดเมนนั้น ๆ จะมีส่วนช่วยให้ผู้ใช้งานเกิดความเข้าใจและนำผลลัพธ์นั้นไปใช้งานได้ดียิ่งขึ้น เช่น Damak et al. (2014) ได้นำเสนอวิธีจำแนกวัตถุจากรูปภาพโดยใช้เทคนิคต้นไม้ตัดสินใจและเทคนิคซัพพอร์ตเวกเตอร์แมชชีนเพื่อระบุชนิดของเครื่องใช้ในครัวเรือนสำหรับผู้พิการทางสายตา ซึ่งผลของการจำแนกข้อมูลที่ได้จะถูกนำไปประมวลผลร่วมกับออนโทโลยีซึ่งมีองค์ความรู้ที่เกี่ยวข้อง

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

กับเครื่องใช้ภายในบ้านและตำแหน่งที่ตั้งของเครื่องใช้นั้น ๆ เพื่อใช้ในการระบุถึงตำแหน่งหรือห้องที่เครื่องใช้นั้นติดตั้งอยู่ ซึ่งช่วยให้ผู้พิการทางสายตาทราบว่าขณะนั้นเขาอยู่ในบริเวณใด

อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Decision Tree) หรือในบางครั้งอาจเรียกว่า ต้นไม้พยากรณ์ (Prediction Tree) คือ อัลกอริทึมที่ใช้สำหรับการจำแนกข้อมูลออกเป็นประเภทต่าง ๆ ตามที่กำหนด โดยมีโครงสร้างแบบต้นไม้ที่แสดงลำดับขั้นตอนในการพิจารณาค่าข้อมูลเพื่อทำการจำแนกข้อมูลออกเป็นประเภทต่าง ๆ วิธีการนี้สามารถทำงานได้กับทั้งข้อมูลแบบกลุ่มข้อมูล (Categorical) และ ข้อมูลที่เป็นค่าต่อเนื่อง (Continuous) การจำแนกข้อมูลหรือการทำนายข้อมูลนั้นจะเริ่มจากการพิจารณาข้อมูลตามโหนดภายในต้นไม้ตัดสินใจ (Internal Node) และไต่ลงไปในแต่ละกิ่งของต้นไม้ที่สอดคล้องตามค่าของข้อมูลที่ต้องการจำแนก โดยจุดสุดท้ายที่ไปถึงจะหมายถึงประเภทของข้อมูลที่จำแนกได้หรือค่าที่ทำนายได้นั่นเอง ซึ่งต้นไม้ตัดสินใจนี้ยังสามารถแปลงให้อยู่ในรูปแบบของกฎการตัดสินใจ (Decision Rules) เพื่อนำไปใช้สนับสนุนการตัดสินใจได้อีกด้วย

ต้นไม้ตัดสินใจสามารถแบ่งออกได้เป็น 2 ลักษณะ คือ

- Classification Tree คือ ต้นไม้ตัดสินใจที่ให้ผลลัพธ์ของการจำแนกข้อมูลหรือการทำนายข้อมูลเป็นประเภทข้อมูล เช่น ซื้อสินค้า/ไม่ซื้อสินค้า ใช่/ไม่ใช่ รายได้สูง/รายได้ปานกลาง/รายได้น้อย เป็นต้น ซึ่งชนิดข้อมูลของตัวแปรตามของวิธีการนี้จะเป็นข้อมูลแบบกลุ่ม

- Regression Tree คือ ต้นไม้ตัดสินใจที่ให้ผลลัพธ์ของการจำแนกข้อมูลหรือการทำนายข้อมูลเป็นค่าตัวเลข เช่น การทำนายยอดขายสินค้าที่ลูกค้าจะทำการสั่งซื้อ เป็นต้น (Dietrich et al., 2015) ซึ่งชนิดข้อมูลของตัวแปรตามของวิธีการนี้จะเป็นค่าต่อเนื่อง

ต้นไม้ตัดสินใจเป็นอัลกอริทึมสำหรับการจำแนกข้อมูลที่ได้รับความนิยมเนื่องจากผลลัพธ์ที่ได้สามารถนำมาแสดงผลในลักษณะรูปภาพเพื่อช่วยให้เห็นขั้นตอนในการตัดสินใจได้อย่างชัดเจนและง่ายต่อการเข้าใจ การทำงานกับต้นไม้ตัดสินใจนั้นผู้ใช้งานไม่จำเป็นต้องทำการกำหนดค่าพารามิเตอร์ต่าง ๆ ที่เกี่ยวข้อง รวมทั้งเป็นอัลกอริทึมที่สามารถทำการจำแนกข้อมูลได้อย่างรวดเร็ว อย่างไรก็ตามประสิทธิภาพในการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจนั้นจะขึ้นอยู่กับคุณภาพของข้อมูลที่ใช้ด้วยเช่นกัน (Han et al., 2011)

1. องค์ประกอบของต้นไม้ตัดสินใจ

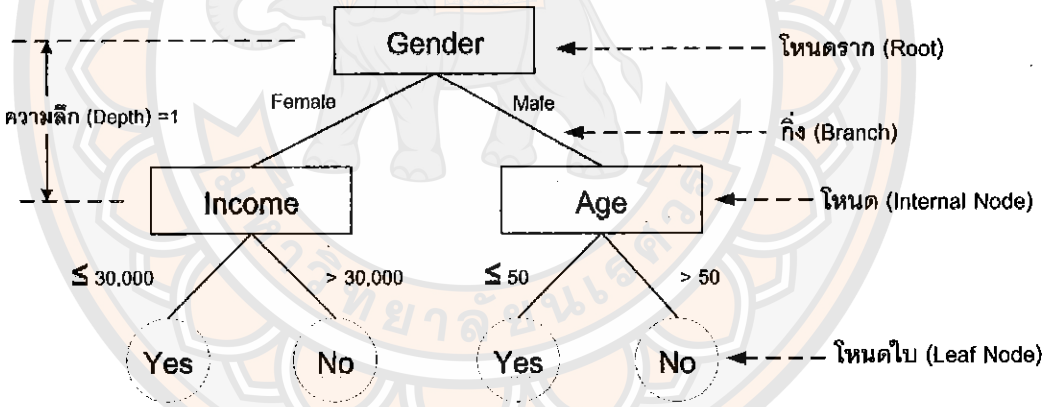
ต้นไม้ตัดสินใจมีองค์ประกอบต่าง ๆ ดังแสดงในภาพ 5

- **โหนดราก** (Root) คือ โหนดที่อยู่บนสุดของต้นไม้ตัดสินใจ
- **กิ่ง** (Branch) จะแทนค่าที่ใช้สำหรับการตัดสินใจรวมทั้งแสดงการเชื่อมต่อระหว่างโหนดสองโหนด หากค่าที่ใช้สำหรับการตัดสินใจเป็นค่าต่อเนื่อง

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

กิ่งทางด้านซ้ายจะหมายถึงค่าที่น้อยกว่าหรือเท่ากับค่าที่ใช้สำหรับตัดสินใจ และกิ่งที่อยู่ทางด้านขวามักจะหมายถึงค่าที่มากกว่าค่าที่ใช้สำหรับตัดสินใจ

- **โหนด** (Internal Node) คือ โหนดที่ใช้สำหรับการตัดสินใจ ซึ่งแต่ละโหนดจะหมายถึงตัวแปรหรือแอตทริบิวต์ในชุดข้อมูล จากภาพ 5 จะเป็นต้นไม้ทวิภาค (Binary Tree) ที่แต่ละโหนดจะแตกกิ่ง (Split) ออกเป็น 2 กิ่ง
- **ความลึกของต้นไม้ตัดสินใจ** (Depth) จะหมายถึงจำนวนลำดับชั้นที่น้อยที่สุดในการเข้าถึงโหนดใด ๆ โดยเริ่มต้นจากโหนดราก เช่น ในภาพ 5 โหนด Income และ โหนด Age จะมีความลึกเท่ากับ 1 ในขณะที่โหนดที่อยู่ล่างสุดทั้ง 4 โหนดนั้นจะมีความลึกเท่ากับ 2 เป็นต้น
- **โหนดใบ** (Leaf Node) คือ โหนดที่อยู่ปลายสุดของต้นไม้ตัดสินใจ โหนดใบนี้จะเป็นโหนดที่ใช้แทนค่าคำตอบของการจำแนกข้อมูล หรือ ค่าที่ทำนายได้ (Dietrich et al., 2015)



ภาพ 5 องค์ประกอบภายในต้นไม้ตัดสินใจ

2. การทำงานของอัลกอริทึมต้นไม้ตัดสินใจ

โดยทั่วไปแล้วต้นไม้ตัดสินใจจะถูกสร้างจากชุดข้อมูลที่ใช้สำหรับการฝึกสอน (Training Set) ที่ประกอบไปด้วยแอตทริบิวต์ข้อมูล และแอตทริบิวต์ประเภทของข้อมูลที่จะจำแนกหรือที่เรียกว่าคลาส (Class) ในที่นี้จะแทนข้อมูลที่ใช้สำหรับการฝึกสอนด้วย $\mathbb{S}$ โดยขั้นตอนในการสร้างต้นไม้ตัดสินใจจะมีการดำเนินการดังภาพ 6 ซึ่งมีรายละเอียดดังนี้

1. หากทุกแถวข้อมูลใน $\mathbb{S}$ ถูกจัดอยู่ในคลาสใดคลาสหนึ่งหรือทุกแถวข้อมูลใน $\mathbb{S}$ มีการปะปนกันของข้อมูลในคลาสใด ๆ ไม่เกินกว่าเกณฑ์ที่กำหนดแล้ว คลาสนั้นจะถูกกำหนดเป็นโหนดใบ หรือเป็นคำตอบของการจำแนกข้อมูล

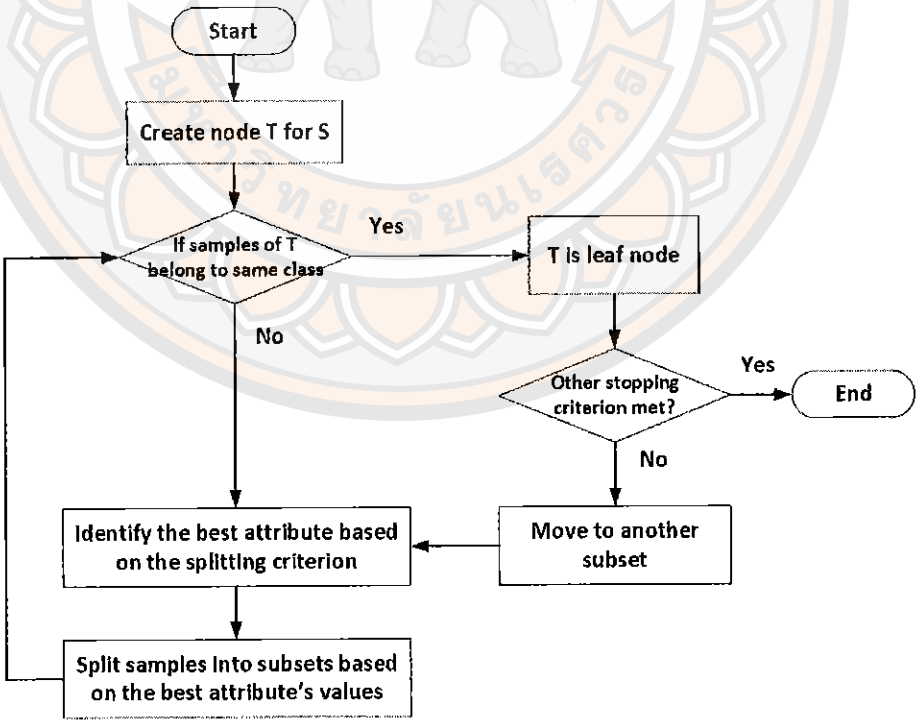
4818889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

2. ในทางกลับกันหากทุกแถวข้อมูลใน S ไม่ถูกจัดอยู่ในคลาสเดียวกันหรือมีการปะปนกันของข้อมูลในหลาย ๆ คลาสแล้ว อัลกอริทึมจะทำการพิจารณาแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสนั้น ๆ มากที่สุดมาเป็นโหนดสำหรับการตัดสินใจ และทำการแบ่งข้อมูล S นั้นออกเป็นชุดย่อย ๆ ตามค่าของแอตทริบิวต์ที่ใช้ในการตัดสินใจ ซึ่งการเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสและการแบ่งข้อมูลออกเป็นชุดย่อย ๆ นี้ คือ กระบวนการในการแตกกิ่งของต้นไม้ตัดสินใจ

3. ชุดข้อมูลย่อยแต่ละชุดจะถูกนำไปพิจารณาว่าแต่ละแถวข้อมูลของชุดข้อมูลย่อยนั้นถูกจัดอยู่ในคลาสใด โดยจะเป็นการทำงานซ้ำตามกระบวนการในข้อที่ 1 และ ข้อที่ 2

ในการสร้างต้นไม้ตัดสินใจนั้นอัลกอริทึมจะหยุดการแตกกิ่งเมื่อมีการทำงานที่ตรงตามเงื่อนไขดังนี้

- โหนดใบทุกโหนดในต้นไม้ตัดสินใจมีการปะปนกันของคลาสคำตอบไม่เกินกว่าค่าที่กำหนด
- ต้นไม้ตัดสินใจไม่สามารถแตกกิ่งออกไปได้อีก
- มีการดำเนินงานตามเงื่อนไขที่ผู้ใช้กำหนด เช่น มีการสร้างต้นไม้ตัดสินใจจนถึงระดับความลึกที่ผู้ใช้ระบุ (Dietrich et al., 2015)



ภาพ 6 ขั้นตอนการสร้างต้นไม้ตัดสินใจ

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

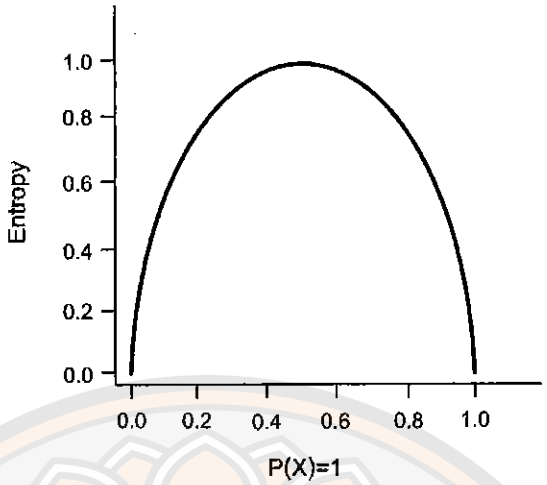
3. ค่าเกนสารสนเทศ

ในขั้นตอนการสร้างต้นไม้ตัดสินใจนั้นจะมีการพิจารณาแอตทริบิวต์ที่เหมาะสมที่สุดเพื่อนำมาใช้ในการแบ่งข้อมูลออกเป็นชุดข้อมูลย่อยเพื่อให้ข้อมูลในแต่ละชุดนั้นมีความบริสุทธิ์ของข้อมูลมากที่สุด หรือ มีการปะปนกันของข้อมูลในคลาสใด ๆ น้อยที่สุด ซึ่งค่าเกนสารสนเทศ (Information Gain) เป็นเกณฑ์หนึ่งที่ได้รับความนิยมและนำมาใช้ในการพิจารณาแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดภายในต้นไม้ตัดสินใจ โดยแอตทริบิวต์ที่มีค่าเกนสารสนเทศสูงสุดจะถูกเลือกเป็นแอตทริบิวต์สำหรับการแบ่งข้อมูล เนื่องจากแอตทริบิวต์นี้เป็นแอตทริบิวต์ที่ต้องการข้อมูลน้อยที่สุดในการแบ่งข้อมูลออกเป็นชุดย่อย ๆ หรือ เป็นแอตทริบิวต์ที่เมื่อใช้ทำการแบ่งข้อมูลแล้วข้อมูลในแต่ละชุดนั้นจะมีการปะปนกันของข้อมูล (Impurity) น้อยที่สุด

ในที่นี้จะแสดงการทำงานของการพิจารณาโหนดสำหรับการตัดสินใจโดยใช้ค่าเกนสารสนเทศโดยมีข้อกำหนดเบื้องต้นดังต่อไปนี้ กำหนดให้ D คือชุดข้อมูลสำหรับการฝึกสอนซึ่งประกอบไปด้วยแอตทริบิวต์ต่าง ๆ ที่ใช้สำหรับอธิบายคุณลักษณะของข้อมูลและแอตทริบิวต์ที่บอกถึงประเภทของข้อมูลหรือคลาสของข้อมูล สำหรับแอตทริบิวต์ที่เป็นคลาสคำตอบจะมีทั้งหมด m ค่า ดังนั้น ข้อมูลแต่ละรายการจะจัดอยู่ใคลาส C_i $(i=1,2,...,m)$ กำหนดให้ $C_{i,D}$ คือ เซตของข้อมูลที่อยู่ในคลาส C_i แล้ว $|D|$ จะหมายถึง จำนวนข้อมูลทั้งหมดในชุดข้อมูล D และ $|C_{i,D}|$ จะหมายถึงจำนวนข้อมูลทั้งหมดที่อยู่ในชุดข้อมูล $C_{i,D}$ (Han et al., 2011)

การคำนวณหาค่าเกนสารสนเทศนั้นจะเริ่มจากการหาค่าเอนโทรปี (Entropy) ซึ่งเป็นค่าที่ใช้ในการพิจารณาความไม่มีแบบแผนในข้อมูล (randomness) โดยหากเอนโทรปีมีค่าสูงจะหมายถึงข้อมูลนั้นมีค่าปะปนกันหลายค่าทำให้ยากต่อการหาข้อสรุปจากข้อมูลนั้น ตัวอย่างเช่น จากภาพ 7 ข้อมูลการสุ่มด้านของเหรียญ โดยค่าเอนโทรปีจะเป็น 0 หากมีการสุ่มได้ด้านของเหรียญเป็นด้านเดียวกันทุกครั้ง ในขณะที่เอนโทรปีจะมีค่าเท่ากับ 1 เมื่อทำการสุ่มได้ด้านของเหรียญได้จำนวนครั้งเท่ากันทั้งสองด้าน ซึ่งหมายถึงข้อมูลมีการปะปนกันจนไม่สามารถหาข้อสรุปที่เหมาะสมจากข้อมูลได้ (Dietrich et al., 2015)

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39
49188183



ภาพ 7 เอนโทรปีของการสุ่มด้านของเหรียญ โดย X= 1 หมายถึงเหรียญด้านหัว และ P(X) หมายถึงความน่าจะเป็นของการสุ่มด้านของเหรียญ

ค่าเอนโทรปี สามารถคำนวณได้จากสมการ (1)

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (1)$$

โดย p_i คือ ค่าความน่าจะเป็นที่ข้อมูลจะมีคลาสของข้อมูลเป็นคลาส C_i ซึ่งสามารถคำนวณได้จาก $\left|\frac{C_{i,D}}{D}\right|$

หากต้องการแบ่งข้อมูล D ออกเป็นชุดย่อย ๆ โดยใช้แอตทริบิวต์ A ซึ่งมีค่าข้อมูลที่แตกต่างกันจำนวน v ค่า ได้แก่ $(a_1, a_2, ..., a_v)$ และค่าข้อมูลของแอตทริวต์ A นั้นเป็นข้อมูลประเภทค่าไม่ต่อเนื่องแล้ว เราจะสามารถทำการแบ่งข้อมูลเป็นชุดย่อย ๆ ได้จำนวน v ชุด คือ $(D_1, D_2, ..., D_v)$ โดย D_j จะประกอบไปด้วยชุดของข้อมูลที่แอตทริบิวต์ A มีค่าเป็น a_j ซึ่งชุดข้อมูลย่อยเหล่านี้จะสอดคล้องกับกิ่งของต้นไม้ที่ได้ทำการแตกจากโหนดที่กำลังทำการพิจารณาอยู่ ดังนั้นในการแบ่งข้อมูลออกเป็นชุดข้อมูลย่อยจึงต้องพยายามทำให้ข้อมูลที่ได้เป็นข้อมูลที่อยู่ในคลาสเดียวกันทั้งหมด อย่างไรก็ตามการแบ่งข้อมูลออกเป็นชุดย่อย ๆ นั้นมักจะมีการปะปนกันของข้อมูลจากหลายคลาส ดังนั้นจึงจำเป็นต้องมีการพิจารณาว่าแต่ละชุดข้อมูลที่ถูกแบ่งออกมานั้นมีการปะปนกันของข้อมูลในแต่ละคลาสอย่างไรโดยการคำนวณค่าเอนโทรปีของแอตทริบิวต์ A ซึ่งหากค่าที่คำนวณได้มีค่าน้อยจะหมายถึงชุดข้อมูลที่มีการแบ่งโดยใช้แอตทริบิวต์ A นี้มีการปะปนกันของคลาสต่าง ๆ น้อยนั่นเอง โดยเราสามารถคำนวณค่าเอนโทรปีของแอตทริบิวต์ A ได้จากสมการ (2)

4681891783 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (2)$$

โดย $|D_j|$ คือ จำนวนของข้อมูลซึ่งแอตทริบิวต์ A มีค่าเท่ากับ a_j

หลังจากที่ทำการคำนวณหาค่าเอนโทรปีของชุดข้อมูล ($Info(D)$) และ ค่าเอนโทปีของแอตทริบิวต์ A ($Info_A(D)$) แล้วนั้น จะสามารถนำค่าดังกล่าวไปคำนวณหาค่าเกนสารสนเทศ ซึ่งหมายถึงค่าทางสถิติที่ใช้ในการวัดความเหมาะสมของแอตทริบิวต์ในการแบ่งชุดข้อมูลที่ใช้ในการฝึกสอนออกเป็นกลุ่มตามคลาส โดยแอตทริบิวต์ที่มีค่าเกนสารสนเทศมากที่สุดจะถูกเลือกเป็นโหนดตัดสินใจภายในต้นไม้ตัดสินใจ ซึ่งค่าเกนสารสนเทศสามารถคำนวณได้จากสมการ (3)

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

4. ข้อดีและข้อจำกัดของเทคนิคต้นไม้ตัดสินใจ

การจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจนั้นมีข้อดีหลายประการดังนี้

1. ต้นไม้ตัดสินใจสามารถแสดงอยู่ในรูปแบบโครงสร้างของต้นไม้ซึ่งง่ายต่อการทำความเข้าใจ รวมทั้งสามารถแปลงออกมาให้อยู่ในรูปแบบของกฎซึ่งช่วยให้ผู้ใช้สามารถนำไปใช้ในการตัดสินใจได้สะดวกยิ่งขึ้น

2. เทคนิคต้นไม้ตัดสินใจเป็นวิธีการที่สามารถทำงานได้กับข้อมูลที่เป็นค่าต่อเนื่องและข้อมูลแบบกลุ่ม

3. เทคนิคต้นไม้ตัดสินใจนั้นสามารถทำงานได้กับชุดข้อมูลที่มีความผิดปกติ เช่น มีข้อมูลที่สูญหาย เป็นต้น

4. เทคนิคต้นไม้ตัดสินใจจัดอยู่ในกลุ่มวิธีการที่เป็นนอนพาราเมตริก (Non-Parametric) ซึ่งหมายถึง เป็นวิธีการที่ไม่จำเป็นต้องทดสอบคุณลักษณะการกระจายของข้อมูลก่อนที่จะนำข้อมูลนั้นมาทำการวิเคราะห์

ถึงแม้ว่าการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจจะมีข้อดีหลายประการแต่วิธีการนี้ยังคงมีข้อจำกัดเช่นกัน ได้แก่

1. อัลกอริทึมต้นไม้ตัดสินใจบางอัลกอร์ทึมต้องการคลาสคำตอบที่เป็นค่าข้อมูลแบบกลุ่ม เท่านั้น เช่น ID3 และ C4.5 เป็นต้น

2. เทคนิคต้นไม้ตัดสินใจนั้นใช้วิธีการแบ่งแยกและเอาชนะ (Divide and Conquer) ในการสร้างต้นไม้ตัดสินใจ ซึ่งวิธีการนี้ทำงานได้ดีเมื่อแอตทริบิวต์ในชุดข้อมูลนั้นมีความสัมพันธ์กันสูง โดยประสิทธิภาพในการทำงานจะลดลงเมื่อปรากฏแอตทริบิวต์ที่ไม่มีความสัมพันธ์กันปะปนอยู่ในชุดข้อมูลที่ทำการศึกษา

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

3. เทคนิคต้นไม้ตัดสินใจจัดอยู่ในกลุ่มของอัลกอริทึมเชิงละโมบ (Greedy Algorithm) จึงมักจะมีความผกผันเมื่อใช้งานกับชุดข้อมูลที่มีแอตทริบิวต์ที่ไม่เกี่ยวข้องหรือมีข้อมูลรบกวน ส่งผลให้ต้นไม้ตัดสินใจที่ได้จะมีผลลัพธ์ไม่คงที่ หากมีการเปลี่ยนแปลงของข้อมูลเพียงเล็กน้อย ต้นไม้ตัดสินใจก็จะมีลักษณะเปลี่ยนแปลงไป และหากชุดข้อมูลที่ใช้มีจำนวนข้อมูลน้อยจะมีโอกาสที่อัลกอริทึมจะเลือกแอตทริบิวต์ที่ไม่เหมาะสมมาใช้ในการสร้างต้นไม้ตัดสินใจ (Maimon & Rokach, 2014)

จากทฤษฎีที่เกี่ยวข้องกับการทำเหมืองข้อมูลและการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ จะพบว่าในการจำแนกข้อมูลด้วยเทคนิคนี้เป็นวิธีการที่รองรับข้อมูลทั้งประเภทกลุ่มข้อมูลและข้อมูลที่เป็นค่าต่อเนื่อง และสามารถจำแนกข้อมูลได้อย่างมีประสิทธิภาพ อย่างไรก็ตามในปัจจุบันเทคนิคการเรียนรู้เชิงลึกจะได้รับความนิยมมาใช้ในการจำแนกข้อมูล แต่การจำแนกข้อมูลด้วยเทคนิคการเรียนรู้เชิงลึกนั้นจะใช้ระยะเวลานานในการประมวลผล ต้องใช้ข้อมูลปริมาณมากเพื่อให้ได้ผลลัพธ์ที่มีความแม่นยำ รวมทั้งผลลัพธ์ที่ได้ยากต่อการทำความเข้าใจ ดังนั้นในงานวิจัยนี้ผู้วิจัยจึงได้ประยุกต์ใช้หลักการของเทคนิคต้นไม้ตัดสินใจที่สามารถทำงานได้อย่างรวดเร็ว และมีจุดเด่นในด้านของผลลัพธ์ที่สามารถทำความเข้าใจได้ง่ายมาใช้ในการพัฒนาแบบจำลองในการจำแนกข้อมูลที่มีประสิทธิภาพ ซึ่งส่งผลให้สามารถทราบถึงความรู้ที่แฝงอยู่ในชุดข้อมูลและสามารถนำความรู้นั้นไปใช้ในการสนับสนุนการตัดสินใจในเรื่องที่เกี่ยวข้องได้

การประเมินประสิทธิภาพแบบจำลองการจำแนกข้อมูล

ในการนำแบบจำลองการจำแนกข้อมูลไปใช้งานนั้นจะต้องมีการประเมินประสิทธิภาพของแบบจำลองที่ได้พัฒนาขึ้นว่ามีความถูกต้องในการจำแนกข้อมูลเพียงใด โดยตัวชี้วัดที่นิยมใช้ในการประเมินประสิทธิภาพของแบบจำลองที่ได้รับความนิยมนั้นจะประกอบไปด้วยค่าต่าง ๆ ดังนี้

ค่าความถูกต้อง (Accuracy) คือ ค่าความถูกต้องในการจำแนกข้อมูลของแบบจำลอง ซึ่งสามารถคำนวณได้จากสมการ (4)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

ค่าความแม่นยำ (Precision) คือ ค่าความแม่นยำในการจำแนกข้อมูลของแบบจำลอง โดยจะหมายถึงค่าความน่าจะเป็นที่แบบจำลองสามารถจำแนกข้อมูลได้ถูกต้องเมื่อพิจารณาจากจำนวนการจำแนกข้อมูล ซึ่งสามารถหาได้จากสมการ (5)

$$\Pr ecision = \frac{TP}{TP + FP} \tag{5}$$

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ค่าความระลึก (Recall) คือ ค่าความครบถ้วนในการจำแนกข้อมูล โดยจะหมายถึงค่าความน่าจะเป็นที่แบบจำลองจะสามารถจำแนกข้อมูลได้ถูกต้องเมื่อพิจารณาจากข้อมูลที่ถูกต้องทั้งหมด โดยสามารถคำนวณได้ดังสมการ (6)

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{6}$$

ค่าประสิทธิภาพโดยรวม (F-measure) คือ การวัดประสิทธิภาพของแบบจำลองโดยพิจารณาค่าความแม่นยำร่วมกับค่าความระลึกซึ่งสามารถคำนวณได้จากสมการ (7)

$$F - measure = 2 \times \frac{\mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \tag{7}$$

โดยมีตัวแปรที่เกี่ยวข้องดังนี้

- TP (True Positive) คือ จำนวนข้อมูลที่มีการจำแนกคลาสที่สนใจได้ถูกต้อง
- TN (True Negative) คือ จำนวนข้อมูลที่มีการจำแนกคลาสที่ไม่สนใจได้ถูกต้อง
- FP (False Positive) คือ จำนวนข้อมูลที่มีการจำแนกข้อมูลผิด โดยทำการจำแนกคลาสที่ไม่สนใจเป็นคลาสที่กำลังสนใจ
- FN (False Negative) คือ จำนวนข้อมูลที่มีการจำแนกข้อมูลผิด โดยทำการจำแนกคลาสที่สนใจเป็นคลาสที่ไม่สนใจ

ออนโทโลยี (Ontology)

ออนโทโลยี (Ontology) เป็นคำศัพท์มีที่มาจากคำศัพท์ในภาษากรีก โดย ontos หมายถึง การมีอยู่ และ logos ที่มีความหมายว่า คำ โดยออนโทโลยีหรือภววิทยานั้นเป็นสาขาหนึ่งของวิชาอภิปรัชญาซึ่งเกี่ยวข้องกับสิ่งต่าง ๆ ที่มีอยู่ในธรรมชาติและความสัมพันธ์ของสิ่งเหล่านั้น โดยมีการนำออนโทโลยีมาใช้เป็นครั้งแรกเมื่อปี ค.ศ. 1980 ในสาขาปัญญาประดิษฐ์ และถูกนำมาใช้ในสาขาอื่น ๆ เพิ่มเติม เช่น สาขาวิศวกรรมความรู้ (Knowledge Engineering) การนำเสนอองค์ความรู้ (Knowledge Representation) การประมวลผลภาษาธรรมชาติ (Natural Language Processing) การจัดการความรู้ (Knowledge Management) และ การค้นคืนสารสนเทศ (Information Retrieval) เป็นต้น

มีผู้ให้นิยามออนโทโลยีในหลากหลายลักษณะซึ่งสำหรับการประมวลผลทางคอมพิวเตอร์นั้นได้นิยามความหมายของออนโทโลยีไว้ว่า ออนโทโลยี คือ การแสดงองค์ความรู้ซึ่งเป็นข้อกำหนดที่ชัดเจนของแนวความคิดใน (Concept) ในเรื่องต่าง ๆ ที่สนใจในรูปแบบของโครงสร้างแบบลำดับขั้น (Hierarchical Data Structure) เพื่อให้เป็นมาตรฐานและเกิดความเข้าใจที่สอดคล้องกัน

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

(Staab & Studer, 2009; มาลี กาบมาลา, 2549) ออนโทโลยีจะมีโครงสร้างของความสัมพันธ์ที่ชัดเจนโดยใช้แนวความคิด หรือ คลาส (Class) ความสัมพันธ์ระหว่างแต่ละคลาส และคุณสมบัติของคลาส (Property) ในการแสดงถึงองค์ความรู้ในเรื่องนั้น ๆ

ออนโทโลยีมีบทบาทสำคัญในการประมวลผลเชิงความหมายโดยออนโทโลยีจะช่วยทำให้เกิดความเข้าใจร่วมกันในขอบเขตของความรู้เรื่องใดเรื่องหนึ่ง โดยใช้แนวความคิดเดียวเพื่อลดจำนวนแนวความคิดอื่น ๆ ที่ไม่เกี่ยวข้อง นอกจากนี้การใช้ออนโทโลยียังช่วยให้เกิดการใช้องค์ความรู้ร่วมกัน รวมถึงสามารถนำองค์ความรู้มาใช้ได้ใหม่ซึ่งมีความสำคัญต่อการพัฒนาระบบฐานความรู้ (Knowledge Based Systems) (มาลี กาบมาลา, 2549)

1. โครงสร้างของออนโทโลยี

ออนโทโลยีมีการกำหนดโครงสร้างแบบลำดับชั้นที่ชัดเจนเพื่อแสดงถึงสิ่งที่เกี่ยวข้องกับแนวความคิดที่ทำการศึกษา โดยออนโทโลยีจะประกอบไปด้วย 5 องค์ประกอบสำคัญ ดังนี้

- **แนวความคิด** (Concept) หมายถึง แนวความคิดที่เกี่ยวข้องกับองค์ความรู้เรื่องใดเรื่องหนึ่ง ซึ่งสามารถเป็นอะไรก็ได้ในองค์ความรู้ที่ศึกษาซึ่งสามารถอธิบายรายละเอียดได้

- **คุณสมบัติ** (Property) หรือ Slots หรือ Roles หรือ Functions หมายถึง คุณสมบัติต่าง ๆ ที่เกี่ยวข้องกับแนวความคิด และนำมาใช้ในการอธิบายแนวความคิดที่ศึกษา เช่น แนวความคิดเกี่ยวกับรายได้บุคคล จะเกี่ยวข้องกับลักษณะของงาน ระดับการศึกษา ระยะเวลาประสบการณ์ทำงาน เป็นต้น

- **ความสัมพันธ์** (Relations) คือ การแสดงถึงประเภทความสัมพันธ์ (Interaction) ระหว่างแนวความคิดแต่ละแนวความคิดในองค์ความรู้ที่สนใจ โดยจะมีการกำหนดลักษณะของความสัมพันธ์ในรูปแบบต่าง ๆ เช่น ความสัมพันธ์แบบลำดับชั้น เช่น Subclass-of หรือ Is-a ซึ่งอธิบายถึงแนวความคิดหนึ่งเป็นสมาชิกย่อยของอีกแนวความคิดหนึ่ง และความสัมพันธ์แบบการเป็นส่วนหนึ่ง หรือ Part-of ที่อธิบายถึงแนวความคิดหนึ่งเป็นองค์ประกอบของอีกแนวความคิดหนึ่ง เป็นต้น

- **แอกเซียม** (Axiom) หมายถึง เงื่อนไขหรือตรรกะของการแปลความสัมพันธ์ระหว่างแนวความคิดกับคุณสมบัติที่เกี่ยวข้อง หรือการแปลความสัมพันธ์ระหว่างแต่ละแนวความคิดเพื่อการแปลความหมายที่ถูกต้อง

- **ตัวอย่างข้อมูล** (Instance) คือ ตัวอย่างข้อมูลหรือคำศัพท์ที่เกี่ยวข้องในองค์ความรู้นั้น ๆ โดยมีการกำหนดไว้ในออนโทโลยี

4911889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

สำนักหอสมุด
30 มี.ค. 2566

2. เครื่องมือที่ใช้ในการพัฒนาออนโทโลยี

ในการพัฒนาออนโทโลยีนั้นมหาวิทยาลัยสแตนฟอร์ด (Stanford University School of Medicine) ได้พัฒนาเครื่องมือที่ใช้สำหรับการออกแบบและสร้างออนโทโลยี (Ontology Editor) ที่มีส่วนติดต่อกับผู้ใช้งานแบบกราฟิก (Graphic User Interface) ชื่อว่าโปรเทเจ (protégé) (Knublauch et al., 2004) เป็นโปรแกรมแบบโอเพ่นซอส (Open Source) ซึ่งไม่เสียค่าใช้จ่ายเมื่อนำมาใช้งาน

สถาปัตยกรรมของโปรแกรมโปรเทเจสามารถแบ่งออกเป็น 2 ส่วน คือ model และ view

- **ส่วน model** นั้นเป็นส่วนที่แสดงองค์ประกอบภายในของออนโทโลยีไม่ว่าจะเป็นคลาส (Class) คุณสมบัติต่าง ๆ (Property) ข้อกำหนดภายในออนโทโลยี (Constraints) และตัวอย่างข้อมูล (Instances) โดยโปรแกรมโปรเทเจจะเตรียม Java API เพื่อใช้ในการดำเนินการกับข้อมูลรวมถึงการสอบถามข้อมูล (Query) ภายในออนโทโลยี

- **ส่วน view** จะเป็นส่วนติดต่อกับผู้ใช้งานซึ่งแสดงในรูปแบบของกราฟิกที่ช่วยให้ผู้ใช้งานสามารถทำการสร้างคลาส กำหนดคุณสมบัติและข้อกำหนดต่าง ๆ ที่เกี่ยวข้อง รวมถึงการระบุข้อมูลตัวอย่างของแต่ละแนวความคิดที่ศึกษาได้ นอกจากนี้นักพัฒนาระบบสามารถพัฒนาฟังก์ชันเพื่อสนับสนุนการทำงานเพิ่มเติมเข้ากับโปรแกรมโปเทเจได้อีกด้วย

ปัจจุบันโปรแกรมโปเทเจสามารถนำมาช่วยสนับสนุนผู้ใช้งานการเรียกดูและปรับปรุงออนโทโลยีได้สะดวกมากยิ่งขึ้น โดยโปรแกรมโปรเทเจนั้นสามารถรองรับออนโทโลยีในหลายรูปแบบไม่ว่าจะเป็น RDF, XML, UML และ OWL (Knublauch et al., 2004)

วิธีการสรุปภาพรวมออนโทโลยี

ปัจจุบันมีการรวบรวมองค์ความรู้และนำเสนอในรูปแบบของออนโทโลยีอย่างแพร่หลายส่งผลให้ออนโทโลยีมีความซับซ้อนและมีจำนวนเพิ่มขึ้น ซึ่งการเลือกใช้ออนโทโลยีให้สอดคล้องตามวัตถุประสงค์ของงานและสามารถนำองค์ความรู้ในออนโทโลยีไปใช้เกิดประโยชน์สูงสุดได้นั้น ผู้ใช้จำเป็นต้องมีความเข้าใจในองค์ความรู้ในออนโทโลยีนั้น การสรุปภาพรวมออนโทโลยี (Ontology Summarization) เป็นวิธีการหนึ่งที่นำมาใช้เพื่อช่วยให้ผู้ใช้งานสามารถทำความเข้าใจออนโทโลยีได้อย่างรวดเร็ว

การสรุปภาพรวมออนโทโลยี (Ontology Summarization) คือ วิธีการในการสกัดองค์ความรู้ที่มีความสำคัญในออนโทโลยีและแสดงอยู่ในลักษณะออนโทโลยีที่มีขนาดเล็กลงเพื่อให้ผู้ใช้สามารถทำความเข้าใจภาพรวมของออนโทโลยีที่สนใจได้อย่างรวดเร็ว (Zhang et al., 2007) โดยการสรุปภาพรวมออนโทโลยีในปัจจุบันนักวิจัยได้นำเสนอวิธีการที่แตกต่างกัน การประยุกต์ใช้วิธีการสำหรับการประมวลผลข้อมูลในรูปแบบกราฟ (Graph-based Method) เพื่อทำการสรุป

4918891B3
NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ภาพรวมของออนโทโลยีเป็นวิธีการหนึ่งที่ได้รับความนิยม ซึ่งออนโทโลยีจะถูกพิจารณาในรูปแบบของกราฟเพื่อระบุองค์ความรู้ที่มีความสำคัญโดยใช้หลักการความเป็นศูนย์กลาง (Centrality-based Measure) ซึ่งมีเกณฑ์การพิจารณา ดังนี้

- **Degree Centrality** (DC) เป็นเกณฑ์ที่ง่ายที่สุดในการพิจารณาองค์ความรู้ / แนวความคิด ที่สำคัญในออนโทโลยี โดยการคำนวณจำนวนเส้น (Edge) ที่เชื่อมโยงแต่ละแนวความคิดในออนโทโลยี ซึ่งแนวความคิดที่มีค่า DC สูงที่สุดจะเป็นแนวความคิดที่มีความสำคัญมากที่สุด ข้อดีของวิธีการนี้คือ ระยะเวลาในการประมวลผลที่สามารถทำงานได้อย่างรวดเร็วแม้จะทำการพิจารณาออนโทโลยีที่มีขนาดใหญ่ อย่างไรก็ตามวิธีการพิจารณาความสำคัญของแนวความคิดวิธีนี้ยังมีข้อจำกัดที่การพิจารณาความสำคัญของแนวความคิดจะพิจารณาจากการเชื่อมโยงข้อมูลของแนวความคิดนั้น ๆ (Local Structure) โดยไม่ได้พิจารณาถึงการเชื่อมโยงของแนวความคิดในภาพรวมของออนโทโลยี (Global Structure) ซึ่งส่งผลลต่อประสิทธิภาพของการพิจารณาแนวความคิดที่สำคัญ

- **Path-based Centrality** (PC) เป็นเกณฑ์ที่ทำการพิจารณาแนวความคิดที่สำคัญจากจำนวนเส้นทางระหว่างแนวความคิดใด ๆ ที่มีการเชื่อมโยงผ่านแนวความคิดที่สนใจ อีกนัยหนึ่งคือ แนวความคิดที่พิจารณาเป็นจุดคั่นกลางระหว่างแนวความคิดอื่น ๆ แนวความคิดที่มีค่า PC สูงที่สุดจะเป็นแนวความคิดที่มีความสำคัญที่สุดในออนโทโลยี วิธีการนี้จะใช้ประโยชน์จากการเชื่อมโยงของแนวความคิดในภาพรวมของออนโทโลยีในการระบุแนวความคิดที่มีความสำคัญ วิธีการนี้มีข้อจำกัด คือ เป็นวิธีการที่ใช้ระยะเวลานานในการประมวลผล

- **Closeness Centrality** (CC) เป็นวิธีการที่พิจารณาแนวความคิดที่มีความสำคัญจากความใกล้ชิดกับแนวความคิดอื่น ๆ ด้วยระยะทางที่สั้นที่สุด ซึ่งจะพิจารณาจากจำนวนเส้นทางที่เชื่อมโยงระหว่างแนวความคิดหนึ่งไปยังอีกแนวความคิดหนึ่งซึ่งมีเส้นทางผ่านแนวความคิดอื่น ๆ ด้วยเส้นทางที่สั้นที่สุด

- **Eigenvector Centrality** (EC) เป็นวิธีการพิจารณาความสำคัญของแนวความคิดที่ได้รับความนิยมวิธีหนึ่ง โดยแนวความคิดที่มีความสำคัญนั้นจะพิจารณาจากการเชื่อมโยงกับแนวความคิดอื่น ๆ ที่มีความสำคัญ โดยหากแนวความคิดหนึ่งมีการเชื่อมโยงกับแนวความคิดอื่น ๆ ที่มีความสำคัญสูงแล้วแนวความคิดนี้ก็จะมีความสำคัญมากกว่าแนวความคิดอื่น ๆ ที่เชื่อมโยงอยู่ ซึ่งอัลกอริทึมที่นิยมใช้ในการพิจารณาความสำคัญของแนวความคิดในออนโทโลยีด้วยวิธีนี้ ได้แก่ PageRank, Weighted PageRank และ Weighted HITS (Hyperlink-Induced Topic Search) เป็นต้น (Pouriyeh et al., 2018)

ในงานวิจัยนี้ผู้วิจัยได้ประยุกต์ใช้วิธีการสรุปภาพรวมออนโทโลยีแบบ Eigenvector Centrality ในการพิจารณาระดับความสำคัญของแนวความคิดในออนโทโลยี เนื่องจากเป็นวิธีที่มี

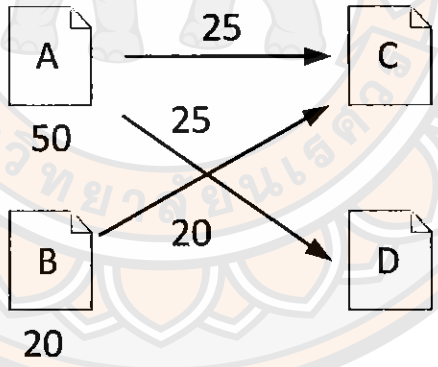
49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

การพิจารณาความสำคัญของแนวความคิดในระดับภาพรวมของออนโทโลยี (Global Structure) รวมทั้งเป็นวิธีการที่ได้รับความนิยมอีกด้วย

อัลกอริทึม PageRank

อัลกอริทึม PageRank (Brin & Page, 2012) คือ อัลกอริทึมสำหรับการจัดอันดับความสำคัญหรือคุณภาพของเว็บเพจจาก Google โดยอันดับความสำคัญของเว็บเพจจะพิจารณาจากจำนวนการเชื่อมโยงจากเว็บเพจต่าง ๆ มายังเว็บเพจนั้น ๆ (Inbound Link)

แนวคิดการทำงานของอัลกอริทึม PageRank สามารถอธิบายได้ดังภาพ 8 โดยเว็บเพจ A ซึ่งมีค่าความสำคัญของเว็บเพจเท่ากับ 50 มีการเชื่อมโยงออกไปยังเว็บเพจ C และ เว็บเพจ D ดังนั้นค่าความสำคัญของเว็บเพจ A ถ่ายทอดไปยังเว็บเพจ C และ D ในจำนวนเท่ากัน ในขณะที่เว็บเพจ B มีค่าความสำคัญของเว็บเพจเท่ากับ 20 มีการเชื่อมโยงไปยังเว็บเพจ C ทำให้ค่าความสำคัญของเว็บเพจ B มีการถ่ายทอดไปยังเว็บเพจ C เพียงเว็บเพจเดียว จากข้อมูลการเชื่อมโยงระหว่างเว็บเพจนี้จึงส่งผลให้เว็บเพจ C มีค่าความสำคัญจากผลรวมของค่าความสำคัญที่ถ่ายทอดจากเว็บเพจ A และ B คือ 45 ในขณะที่เว็บเพจ D จะมีค่าความสำคัญที่ถ่ายทอดจากเว็บเพจ A เพียงเว็บเพจเดียว คือ 25



ภาพ 8 ตัวอย่างแนวคิดการทำงานของอัลกอริทึม PageRank

การคำนวณค่าของความสำคัญของเว็บเพจด้วยอัลกอริทึม PageRank นั้นสามารถคำนวณได้จากสมการ (8)

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + ... + PR(T_n)/C(T_n)) \quad (8)$$

โดย $PR(A)$ คือ ค่าความสำคัญของเว็บเพจ A

$PR(T_i)$ คือ ค่าความสำคัญของเว็บเพจ T_i ซึ่งมีการเชื่อมโยงไปหาเว็บเพจ A

$C(T_i)$ คือ จำนวนการเชื่อมโยงที่ออกจากเว็บเพจ T_i ไปยังเว็บเพจใด ๆ

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

d คือ ค่า Damping Factor ซึ่งมีค่าอยู่ระหว่าง 0 ถึง 1 โดยนิยมกำหนดให้มีค่า 0.85

ซึ่งค่า Damping Factor จะเป็นค่าที่ใช้ในการป้องกันปัญหาค่าความสำคัญไปสะสมอยู่ที่เว็บเพจใดเว็บเพจหนึ่ง (Rank Sink) ในกรณีที่เว็บเพจมีการเชื่อมโยงวนเข้าหาเว็บเพจเดิม (Brin & Page, 2012)

ในการวิจัยเพื่อปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจนี้ จะทำการประยุกต์ใช้องค์ความรู้และความสัมพันธ์ของข้อมูลที่สามารถอ้างอิงได้จากออนโทโลยีและอัลกอริทึม PageRank ในการพิจารณาระดับความสำคัญของแต่ละแนวความคิดในออนโทโลยี เพื่อปรับปรุงกระบวนการทำงานของอัลกอริทึมต้นไม้ตัดสินใจซึ่งสามารถช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูล

งานวิจัยที่เกี่ยวข้อง

การวิจัยเพื่อปรับปรุงประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจนั้น นักวิจัยได้นำเสนอวิธีการต่าง ๆ ในการปรับปรุงประสิทธิภาพของเทคนิคนี้ เช่น การปรับปรุงคุณภาพของข้อมูลเพื่อใช้สำหรับเทคนิคต้นไม้ตัดสินใจ เนื่องจากเทคนิคต้นไม้ตัดสินใจถูกจัดอยู่ในกลุ่มของอัลกอริทึมเชิงละโมบ (Greedy algorithm) และใช้วิธีการแบ่งแยกและเอาชนะ (Divide and conquer) ในการสร้างแบบจำลองจำแนกข้อมูล ซึ่งวิธีการนี้จะมีประสิทธิภาพการทำงานลดลงเมื่อภายในชุดข้อมูลปรากฏข้อมูลที่ผิดปกติหรือไม่มีความเกี่ยวข้องกับคลาสที่ต้องการจำแนก รวมถึงเมื่อชุดข้อมูลมีแอตทริบิวต์หรือข้อมูลจำนวนมากจะทำให้ต้นไม้ตัดสินใจมีความลึกมาก ส่งผลให้ต้นไม้ตัดสินใจมีความซับซ้อนและยากต่อการทำความเข้าใจ

นอกจากนี้การปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจเป็นอีกแนวทางหนึ่งในการเพิ่มประสิทธิภาพของการจำแนกข้อมูล โดยในการสร้างต้นไม้ตัดสินใจนั้นการพิจารณาโหนดของต้นไม้ตัดสินใจหรือเกณฑ์ที่ใช้ในการพิจารณาโหนดของต้นไม้ตัดสินใจนั้นเป็นขั้นตอนที่มีความสำคัญที่ส่งผลต่อประสิทธิภาพของการจำแนกข้อมูล หากเลือกแอตทริบิวต์ที่ไม่เหมาะสมเป็นโหนดของต้นไม้ตัดสินใจแล้วจะทำให้ต้นไม้ตัดสินใจมีความผิดปกติ และส่งผลต่อความถูกต้องในการจำแนกข้อมูล ดังนั้นนักวิจัยจึงได้นำเสนอวิธีการในการพิจารณาโหนดของต้นไม้ตัดสินใจ เช่น การปรับปรุงเกณฑ์การพิจารณาโหนดของต้นไม้ตัดสินใจ รวมถึงการนำเสนอเกณฑ์ใหม่ที่ใช้ในการเลือกโหนดของต้นไม้ตัดสินใจ เป็นต้น เพื่อให้สามารถพิจารณาแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดของต้นไม้ได้อย่างเหมาะสมมากขึ้น

ตัวอย่างงานวิจัยที่เกี่ยวข้องกับการปรับปรุงประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจด้วยวิธีการต่าง ๆ มีรายละเอียดดังนี้

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

1. การปรับปรุงคุณภาพของข้อมูลเพื่อใช้สำหรับเทคนิคต้นไม้ตัดสินใจ

การเตรียมข้อมูล (Data preparation) คือ กระบวนการในการจัดการกับข้อมูลเพื่อให้พร้อมสำหรับการนำไปใช้ในการสร้างแบบจำลอง โดยขั้นตอนการเตรียมข้อมูลจะเกี่ยวข้องกับการทำความสะอาดข้อมูล (Data Cleansing) การคัดเลือกข้อมูล (Data Selection) รวมทั้งการแปลงข้อมูล (Data Transformation) ซึ่งคุณภาพของข้อมูลเป็นปัจจัยหนึ่งที่ส่งผลกระทบต่อประสิทธิภาพของการจำแนกข้อมูล โดยหากข้อมูลที่นำมาวิเคราะห์มีความถูกต้องและครบถ้วนจะช่วยให้การจำแนกข้อมูลมีประสิทธิภาพมากยิ่งขึ้น (Blake & Mangiameli, 2011) ด้วยเหตุนี้นักวิจัยหลายท่านจึงได้นำเสนอเทคนิคต่าง ๆ ที่ช่วยในการปรับปรุงคุณภาพของข้อมูลนำเข้า ก่อนที่จะสร้างแบบจำลองต้นไม้ตัดสินใจ เช่น Kudoh et al. (2003) ได้นำเสนอวิธีการที่เรียกว่า Information Theoretical Abstraction (ITA) ซึ่งจะนำข้อมูลที่เป็นแนวความคิดพื้นฐาน (Abstract Value) ของข้อมูลที่ทำการศึกษามาใช้ในการสร้างต้นไม้ตัดสินใจ โดยการพิจารณาค่าข้อมูลแนวคิดพื้นฐานที่เกี่ยวข้องกับข้อมูลที่ทำการศึกษานั้นจะอ้างอิงจากคำศัพท์ที่เกี่ยวข้องใน WordNet (Miller, 1995) ซึ่งเป็นคลังข้อมูลคำศัพท์ภาษาอังกฤษที่มีการใช้งานอย่างแพร่หลาย ข้อมูลแนวความคิดพื้นฐานที่อ้างอิงได้จะถูกนำไปแทนค่าข้อมูลเดิมที่อยู่ภายในฐานข้อมูลในกรณีที่นำค่าข้อมูลซึ่งเป็นแนวความคิดพื้นฐานเหล่านั้นมาใช้ในการจำแนกข้อมูลแล้วไม่ทำให้เกิดความผิดพลาดในการจำแนกข้อมูลเกินกว่าเกณฑ์ที่ผู้ใช้กำหนด (Threshold) ผลการทดลองพบว่า การนำแนวความคิดพื้นฐานมาใช้ในการจำแนกข้อมูลช่วยให้ต้นไม้ตัดสินใจที่ได้มีขนาดเล็กลงเมื่อเปรียบเทียบกับขนาดของต้นไม้ตัดสินใจที่เกิดขึ้นจากการใช้ข้อมูลเดิม นอกจากนี้ Tang & Fong (2010) ได้ทำการประยุกต์ใช้วิธีการที่เรียกว่า Attribute Value Taxonomies (AVT) ซึ่งเป็นวิธีการที่นำแนวความคิดพื้นฐานของข้อมูลซึ่งนำเสนอในรูปแบบของโครงสร้างแบบลำดับชั้นมาใช้ในการเตรียมข้อมูลสำหรับจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ โดยวิธีการนี้จะทำการแปลงค่าข้อมูลของแต่ละแอตทริบิวต์ในชุดข้อมูลด้วยข้อมูลแนวความคิดพื้นฐานที่มีความสัมพันธ์ รวมถึงยังทำการรวมแอตทริบิวต์ที่มีความสัมพันธ์กันตามโครงสร้างใน AVT เพื่อลดจำนวนแอตทริบิวต์ที่ใช้ในการสร้างต้นไม้ตัดสินใจ ซึ่งจะทำให้ได้ชุดข้อมูลใหม่จำนวน n ชุด ตามจำนวนระดับของแนวความคิดพื้นฐานที่อ้างอิง และนำชุดข้อมูลเหล่านี้ไปใช้เป็นข้อมูลนำเข้าสำหรับการสร้างต้นไม้ตัดสินใจ โดย Tang & Fong (2010) ได้ทำการทดสอบวิธีการที่นำเสนอกับข้อมูลการประมูลสินค้าจำนวน 8,000 รายการ ซึ่งผลการทดลองพบว่า ต้นไม้ตัดสินใจที่สร้างจากข้อมูลได้รับการปรับปรุงด้วยวิธีการที่นำเสนอจะมีขนาดเล็กกว่าต้นไม้ตัดสินใจที่สร้างจากชุดข้อมูลเดิม รวมทั้งช่วยให้มีค่าความถูกต้องในการจำแนกข้อมูลสูงขึ้นอีกด้วย รวมถึง Ramezankhani et al. (2014) ได้ประยุกต์ใช้เทคนิคต้นไม้ตัดสินใจในการจำแนกข้อมูลผู้ป่วยโรคเบาหวานชนิดที่ 2 (type 2 diabetes) จากข้อมูลจำนวน 6,647 รายการ โดยในการจัดเตรียมข้อมูลนั้นได้มีการประยุกต์ใช้แนวความคิดพื้นฐานของข้อมูลซึ่งนำเสนอในรูปแบบของโครงสร้างแบบลำดับชั้นเพื่อปรับปรุงคุณภาพ

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ข้อมูล เช่น ข้อมูลระดับการศึกษาของผู้ป่วยซึ่งแสดงในรูปแบบข้อมูลตัวเลขจะถูกปรับปรุงให้อยู่ใน 3 ระดับตามความสัมพันธ์ของข้อมูล ซึ่งผลการวิจัยพบว่าแบบจำลองที่พัฒนาขึ้นสามารถจำแนกข้อมูลได้อย่างมีประสิทธิภาพ โดยมีค่าความถูกต้องในการจำแนกข้อมูลเท่ากับ 90.5% นอกจากนี้ Xiahou et al. (2021) ได้นำเสนอแบบจำลองการวิเคราะห์ระดับความเสี่ยงของลูกค้าสำหรับธุรกิจประกันภัยรถยนต์ด้วยเทคนิคต้นไม้ตัดสินใจ โดยในการดำเนินงานได้ประยุกต์ใช้ความสัมพันธ์ระหว่างข้อมูลสำหรับการแปลงข้อมูลยี่ห้อรถยนต์เป็นแหล่งของเทคโนโลยีที่ผลิตรถยนต์แต่ละยี่ห้อเพื่อลดขนาดของข้อมูลที่ต้องพิจารณา ซึ่งผลการวิจัยพบว่าแบบจำลองที่พัฒนาขึ้นสามารถจำแนกข้อมูลที่มีความเสี่ยงต่ำได้อย่างมีประสิทธิภาพในขณะที่ในการจำแนกลูกค้าที่มีความเสี่ยงสูงนั้นยังมีความคลาดเคลื่อน เนื่องจากการจัดเก็บข้อมูลที่ยังไม่ครอบคลุมการดำเนินงานด้านประกันภัย

ซึ่งจากงานวิจัยข้างต้นจะพบว่าในการปรับปรุงคุณภาพของชุดข้อมูลด้วยค่าแนวความคิดพื้นฐานที่มีความสัมพันธ์กันนั้น จะเป็นการใช้ประโยชน์จากองค์ความรู้ที่เกี่ยวข้องซึ่งอยู่ในรูปแบบโครงสร้างแบบลำดับขั้น โดยในการแปลงข้อมูลจะดำเนินการแปลงข้อมูลทุกแอตทริบิวต์ที่พบแนวความคิดพื้นฐานที่เกี่ยวข้องซึ่งทำให้จำนวนข้อมูลที่ต้องพิจารณาในการสร้างต้นไม้ตัดสินใจนั้นลดลง และส่งผลให้ต้นไม้ตัดสินใจที่ได้มีขนาดเล็กลง อย่างไรก็ตามงานวิจัยข้างต้นนี้ยังมีข้อจำกัดโดยยังขาดการพิจารณาความสัมพันธ์ระหว่างข้อมูลเพื่อใช้ในการเลือกเฉพาะแอตทริบิวต์ที่มีความเกี่ยวข้องและส่งผลต่อคลาสคำตอบที่ต้องการ ซึ่งการพิจารณาความสัมพันธ์ระหว่างข้อมูลเป็นอีกวิธีหนึ่งที่ช่วยในการปรับปรุงคุณภาพของข้อมูลที่ใช้ในการสร้างแบบจำลองโดยเป็นการลดจำนวนแอตทริบิวต์ที่ไม่เกี่ยวข้องได้ ตัวอย่างเช่น Dwi Prayogo & Ikhsan (2020) ได้ทำการทดสอบวิธีการในการพิจารณาแอตทริบิวต์ที่มีความสำคัญต่อการจำแนกตัวอักษร เช่น เกนสารสนเทศ (information gain) อัตราส่วนเกน (Gain Ratio) ค่าสหสัมพันธ์ (Correlation) และ สถิติไคสแควร์ (Chi-square) เพื่อปรับปรุงคุณภาพของชุดข้อมูลโดยการลดจำนวนแอตทริบิวต์ที่ไม่มีความเกี่ยวข้อง โดยในการศึกษานี้ Dwi Prayogo & Ikhsan (2020) ได้ทำการทดสอบชุดข้อมูลที่ปรับปรุงโดยใช้อัลกอริทึมในกลุ่มของการสร้างต้นไม้ตัดสินใจ ได้แก่ อัลกอริทึม J48 อัลกอริทึม CART และอัลกอริทึม Random Forest ซึ่งผลการศึกษาพบว่าเมื่อทำการจำแนกข้อมูลด้วยแบบจำลองที่ใช้เฉพาะแอตทริบิวต์ที่มีความสำคัญนั้น จะทำให้ได้ค่าความถูกต้องในการจำแนกข้อมูลที่เพิ่มขึ้นเมื่อเปรียบเทียบกับผลลัพธ์ที่ได้จากการจำแนกข้อมูลด้วยแบบจำลองที่มีการใช้ทุกแอตทริบิวต์ในชุดข้อมูล

จากข้อจำกัดดังกล่าวในการวิจัยครั้งนี้ผู้วิจัยจึงแนวความคิดในการนำเทคนิคที่ใช้ในการพิจารณาความสัมพันธ์ระหว่างข้อมูล เช่น สถิติไคสแควร์ (Chi square) และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล (Point biserial correlation) มาใช้การพิจารณาความสัมพันธ์ระหว่างแอตทริบิวต์ข้อมูลและคลาสคำตอบที่ต้องการจำแนก เพื่อใช้ในการพิจารณาเลือกเฉพาะแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสคำตอบ และนำแอตทริบิวต์เหล่านั้นไปใช้ในการสร้างแบบจำลองการจำแนก

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ข้อมูล ร่วมกับการประยุกต์ใช้แนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีเพื่อใช้ในการแปลงข้อมูลที่ใช้สำหรับการสร้างแบบจำลองการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ ซึ่งจะการลดจำนวนแอตทริบิวต์ในชุดข้อมูล และลดจำนวนค่าของข้อมูลที่ต้องพิจารณาจะสามารถช่วยให้ต้นไม้ตัดสินใจมีขนาดลดลง หรือมีความซับซ้อนน้อยลงนั่นเอง

2. การปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจ

แนวทางหนึ่งที่นักวิจัยให้ความสนใจในการพัฒนาเพื่อปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจ คือ การประยุกต์ใช้ข้อมูลแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีในการปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจ ซึ่งการนำแนวความคิดพื้นฐานที่มีความสัมพันธ์กับข้อมูลที่ทำการศึกษามาใช้ในการสร้างต้นไม้ตัดสินใจจะช่วยลดขนาดของต้นไม้ตัดสินใจได้ โดยการลดความหลากหลายของข้อมูลที่ใช้ในการแตกกิ่งของต้นไม้ตัดสินใจ ตัวอย่างนักวิจัยที่นำเสนอวิธีการนี้ได้แก่ Zhang et al. (2002) ได้นำเสนอวิธีการปรับปรุงกระบวนการเรียนรู้ของเทคนิคต้นไม้ตัดสินใจที่มีการประยุกต์ใช้ออนโทโลยี โดยโครงสร้างแบบลำดับชั้น (Hierarchical Structure) ของแนวความคิดภายในออนโทโลยีจะถูกนำมาช่วยในการพิจารณาแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดในต้นไม้ตัดสินใจ โดยทำการเปรียบเทียบแนวความคิดภายในออนโทโลยีกับแอตทริบิวต์ในชุดข้อมูลเพื่อให้ได้โครงสร้างแบบลำดับชั้นที่มีแนวความคิดที่สอดคล้องกับข้อมูลของแอตทริบิวต์นั้น ๆ หลังจากนั้นจึงทำการคำนวณหาค่าเกนสารสนเทศของแนวความคิดในแต่ละระดับเพื่อนำมาใช้ในการพิจารณาโหนดที่เหมาะสมในการสร้างต้นไม้ตัดสินใจ เพื่อประเมินประสิทธิภาพของวิธีการที่นำเสนอ Zhang ได้นำข้อมูลการซื้อสินค้าของลูกค้าในร้านค้าแห่งหนึ่งมาทำการทดสอบ ซึ่งพบว่าการนำแนวความคิดซึ่งมีระดับของข้อมูลแตกต่างจากแนวความคิดในชุดข้อมูลเดิมมาใช้ในกระบวนการเรียนรู้ของอัลกอริทึมต้นไม้ตัดสินใจนั้น ช่วยให้สามารถจำแนกข้อมูลประเภทของลูกค้าได้อย่างมีประสิทธิภาพมากยิ่งขึ้น เช่นเดียวกันกับ Vieira & Antunes (2014) ได้นำเสนอวิธีการประยุกต์ใช้ออนโทโลยีในกระบวนการเรียนรู้ของต้นไม้ตัดสินใจ โดยข้อมูลในชุดข้อมูลที่ทำการศึกษาจะถูกนำไปใช้เป็นตัวอย่างข้อมูล (instance) ภายในออนโทโลยีเพื่อค้นหาแนวความคิดพื้นฐานที่เกี่ยวข้องโดยพิจารณาจากความสัมพันธ์แบบลำดับชั้น (superclass/subclass) ของข้อมูล ซึ่งในขั้นตอนการเรียนรู้ของต้นไม้ตัดสินใจนั้นจะพิจารณาโหนดภายในต้นไม้ตัดสินใจจากข้อมูลเดิมและแนวความคิดพื้นฐานของข้อมูลที่อ้างอิงได้จากออนโทโลยี หากข้อมูลใดที่ส่งผลให้มีค่าเกนสารสนเทศมากที่สุดข้อมูลนั้นจะถูกนำมาใช้เป็นข้อมูลสำหรับการสร้างต้นไม้ตัดสินใจ ผลการศึกษาพบว่าการนำแนวความคิดที่ได้จากออนโทโลยีมาใช้สามารถช่วยให้การจำแนกข้อมูลมีความถูกต้องมากยิ่งขึ้น

นอกจากการนำข้อมูลแนวความคิดพื้นฐานมาใช้ในกระบวนการสร้างต้นไม้ตัดสินใจแล้ว การนำระดับความสำคัญของแอตทริบิวต์ที่มีต่อคลาสที่ต้องการจำแนกมาใช้ในการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจเป็นอีกวิธีการหนึ่งที่ได้รับความสนใจจากนักวิจัย โดยค่าระดับความสำคัญของ

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

แอตทริบิวต์จะช่วยแก้ปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีข้อมูลหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจ ซึ่งแอตทริบิวต์นั้นอาจเป็นแอตทริบิวต์ที่มีระดับความสำคัญน้อยเมื่อพิจารณาจากคลาสที่ต้องการจำแนก ตัวอย่างงานวิจัยที่ทำการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจด้วยค่าระดับความสำคัญของแอตทริบิวต์ ได้แก่ งานวิจัยของ Chen et al. (2009) ที่นำเสนอวิธีการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจโดยการปรับปรุงค่าเกนสารสนเทศด้วยค่าความสำคัญของแอตทริบิวต์ โดยค่าความสำคัญของแต่ละแอตทริบิวต์จะคำนวณด้วย Correlation Function Method ซึ่งเป็นค่าที่ใช้แสดงความสัมพันธ์ระหว่างข้อมูล และนำค่าดังกล่าวไปใช้ในการปรับปรุงค่าเกนสารสนเทศซึ่งสามารถแสดงได้ดังสมการ (9)

$$Gain'(A) = ((Info(D) - Info_A(D)) \times V(A) \quad (9)$$

โดย $Gain'(A)$ คือ ค่าเกนสารสนเทศที่ทำการปรับปรุง ค่า $Info(D)$ คือ ค่าเอนโทรปีของชุดข้อมูล ค่า $Info_A(D)$ คือ ค่าเอนโทรปีของแต่ละแอตทริบิวต์ และ $V(A)$ คือ ค่าความสำคัญของแอตทริบิวต์ที่ได้ทำการปรับช่วงของข้อมูล (Normalization) ผลการทดสอบอัลกอริทึมที่นำเสนอพบว่า ต้นไม้ตัดสินใจที่ได้สามารถแสดงกฎในการจำแนกข้อมูลได้ชัดเจนมากขึ้น ส่งผลให้ผู้ใช้สามารถทำความเข้าใจกฎได้ดีขึ้น ต่อมา Soni & Pawar (2017) ได้นำเสนออัลกอริทึมต้นไม้ตัดสินใจซึ่งได้ทำการปรับปรุงการคำนวณค่าความสัมพันธ์ระหว่างข้อมูลที่เสนอโดย Chen et al. (2009) เพื่อแก้ปัญหาการซ้ำซ้อนกันของข้อมูลที่ส่งผลต่อความถูกต้องในการพิจารณาโหนดของต้นไม้ตัดสินใจ Soni & Pawar (2017) ได้ทดสอบวิธีการที่นำเสนอกับการจำแนกอารมณ์ของผู้ใช้สื่อสังคมออนไลน์ ซึ่งพบว่าวิธีการที่นำเสนอมีค่าความถูกต้องในการจำแนกข้อมูลมากกว่าค่าความถูกต้องที่ได้จากต้นไม้ตัดสินใจที่สร้างด้วยวิธีการของ Chen et al. (2009) และอัลกอริทึมต้นไม้ตัดสินใจแบบเดิม ต่อมา Zhu et al. (2018) ได้นำเสนอวิธีการในการปรับปรุงอัลกอริทึม CART โดยการประยุกต์ใช้ค่าน้ำหนักของแอตทริบิวต์ร่วมกับเทคนิค MapReduce เพื่อเพิ่มประสิทธิภาพของต้นไม้ตัดสินใจ โดยอัลกอริทึมที่นำเสนอจะทำการคำนวณค่าน้ำหนักของแต่ละแอตทริบิวต์จากความถี่ของข้อมูลที่ปรากฏในชุดข้อมูล แอตทริบิวต์ที่มีค่าน้ำหนักมากที่สุดจะหมายถึงแอตทริบิวน์นั้นมีความสำคัญกับคลาสคำตอบมากที่สุดและจะถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจ หลังจากจะทำการพิจารณาค่าที่เหมาะสมสำหรับการแตกกิ่งของต้นไม้ตัดสินใจโดยใช้ค่า Gini coefficient โดยผลการทดลองพบว่าต้นไม้ตัดสินใจที่ได้สามารถทำงานได้อย่างมีประสิทธิภาพแม้จะเป็นชุดข้อมูลที่มีขนาดใหญ่ ต้นไม้ตัดสินใจที่ได้มีความซับซ้อนลดลง รวมถึงมีความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้น เช่นเดียวกันกับ Es–Sabery & Hair (2019) ซึ่งนำแนวความคิดเกี่ยวกับความสัมพันธ์ระหว่างแอตทริบิวต์และค่าระดับความสำคัญของแต่ละแอตทริบิวต์มาใช้ในการปรับปรุงเกณฑ์การพิจารณาโหนดของต้นไม้ตัดสินใจ โดยในแต่ละครั้งที่มีการพิจารณาแอตทริบิวต์สำหรับใช้เป็นโหนดของต้นไม้ตัดสินใจ

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

อัลกอริทึมนี้จะมีการคำนวณค่าสหสัมพันธ์ระหว่างแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดตัดสินใจเดิมกับแอตทริบิวต์อื่น ๆ ที่อยู่ในชุดข้อมูล เพื่อหาค่าความสำคัญของแต่ละแอตทริบิวต์ หลังจากนั้นจะนำค่าความสำคัญที่ได้ไปปรับปรุงค่าเกนสารสนเทศและใช้ในการเลือกแอตทริบิวต์ที่เหมาะสมสำหรับการเป็นโหนดตัดสินใจต่อจากโหนดเดิม ซึ่งผลการทดลองพบว่าต้นไม้ตัดสินใจที่ได้จากอัลกอริทึมนี้มีจำนวนโหนดใบลดลง รวมทั้งมีค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้น ต่อมา Ahmed et al. (2020) ได้นำเสนออัลกอริทึมที่ชื่อว่า Decisive Decision Tree (DDT) โดยอัลกอริทึมนี้จะนำค่าความน่าจะเป็นที่แอตทริบิวต์ในชุดข้อมูลจะปรากฏร่วมกับคลาสแต่ละคลาสที่ต้องการจำแนก (Decisive value) มาใช้เป็นค่าระดับความสำคัญของแต่ละแอตทริบิวต์ และนำค่าระดับความสำคัญนี้ไปใช้ในการปรับปรุงค่าเกนสารสนเทศ โดยผลการวิจัยพบว่าอัลกอริทึม DDT สามารถช่วยลดปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีข้อมูลหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจได้ โดยต้นไม้ตัดสินใจที่สร้างด้วยอัลกอริทึม DDT จะมีการเลือกแอตทริบิวต์ที่มีค่าข้อมูลในแอตทริบิวต์น้อยเป็นโหนดรากของต้นไม้ตัดสินใจ ในขณะที่ต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึมเดิมจะเลือกแอตทริบิวต์ที่มีค่าข้อมูลจำนวนมากเป็นโหนดภายในต้นไม้ตัดสินใจ นอกจากนี้ H. Zhou et al. (2020) ได้นำเสนอแนวความคิดในการนำค่าความสำคัญของแอตทริบิวต์มาใช้เป็นเกณฑ์ในการพิจารณาโหนดภายในต้นไม้ตัดสินใจ โดยวิธีการนี้จะทำการหาค่าความสำคัญของแต่ละแอตทริบิวต์โดยใช้อัลกอริทึม ReliefF ที่ได้ทำการปรับปรุง หลังจากนั้นจะทำการเลือกเฉพาะแอตทริบิวต์ที่มีค่าความสำคัญมากกว่าค่ามัธยฐานของค่าความสำคัญที่คำนวณได้ไปใช้ในการสร้างต้นไม้ตัดสินใจ วิธีการนี้จะนำค่าระดับความสำคัญของแอตทริบิวต์เป็นเกณฑ์ในการพิจารณาโหนดภายในต้นไม้ตัดสินใจ ซึ่งผลการทดลองพบว่าวิธีการที่นำเสนอจะมีค่าความถูกต้องในการจำแนกข้อมูลสูงกว่าวิธีการสร้างต้นไม้ตัดสินใจแบบดั้งเดิม แต่หากทำการสร้างต้นไม้ตัดสินใจด้วยวิธีการที่นำเสนอโดยไม่มีการดำเนินการลดจำนวนแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสที่ต้องการจำแนกแล้วนั้น ต้นไม้ตัดสินใจที่ได้จะมีค่าความถูกต้องในการจำแนกข้อมูลใกล้เคียงกับการจำแนกข้อมูลของต้นไม้ตัดสินใจแบบเดิม รวมถึง Gang et al. (2021) ได้นำเสนออัลกอริทึมต้นไม้ตัดสินใจที่มีการประยุกต์ใช้ค่าสัมประสิทธิ์สหสัมพันธ์แบบสเปียร์แมน (Spearman rank correlation coefficient) ซึ่งแสดงขนาดความสัมพันธ์ของแต่ละแอตทริบิวต์กับคลาสคำตอบในการปรับปรุงค่าเกนสารสนเทศเพื่อแก้ปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีข้อมูลหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจ โดยได้ทำการทดสอบอัลกอริทึมที่นำเสนอกับชุดข้อมูลทางด้านการแพทย์ ซึ่งผลการทดลองพบว่าอัลกอริทึมที่นำเสนอช่วยให้กฎการจำแนกข้อมูลที่ได้จากต้นไม้ตัดสินใจมีความสอดคล้องกับความคิดเห็นของผู้เชี่ยวชาญมากกว่ากฎการจำแนกข้อมูลที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจแบบเดิม

นอกจากใช้ค่าความสัมพันธ์ระหว่างแอตทริบิวต์ในการพิจารณาระดับความสำคัญของแอตทริบิวต์ที่มีต่อการจำแนกข้อมูลแล้ว ยังมีนักวิจัยนำเสนอวิธีการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจ

491888183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ด้วยการนำค่าความสำคัญของแอตทริบิวต์ที่กำหนดโดยผู้เชี่ยวชาญมาใช้งานอีกด้วย เช่น Liu & Xie (2010) ได้นำค่าความสำคัญของแอตทริบิวต์ที่กำหนดโดยผู้เชี่ยวชาญมาใช้ในการปรับปรุงค่าเอนโทรปีของแต่ละแอตทริบิวต์ และนำค่าเอนโทรปีที่ทำการปรับปรุงมาใช้ในการคำนวณหาค่าเกนสารสนเทศเพื่อใช้ในการพิจารณาแอตทริบิวต์ที่เหมาะสมในการเป็นโหนดภายในต้นไม้ตัดสินใจ ผลจากการปรับปรุงอัลกอริทึมพบว่าต้นไม้ตัดสินใจที่ได้มีความถูกต้องในการจำแนกข้อมูลมากขึ้น รวมทั้งกฎที่ได้สามารถทำความเข้าใจได้ง่ายขึ้น นอกจากนี้ Iqbal et al. (2012) ได้พัฒนาการอัลกอริทึม Importance Aided Decision Tree (IADT) ซึ่งจะนำค่าความสำคัญของแอตทริบิวต์ที่มีต่อคลาสที่ต้องการจำแนก (Attribute Importance Score) มาใช้ในปรับปรุงค่าต่าง ๆ ที่ใช้เป็นเกณฑ์ในการพิจารณาโหนดตัดสินใจซึ่งสามารถคำนวณได้จากสมการ (10)

$$S_I(X) = (1-p) \times S(X) + p \times I_X \tag{10}$$

โดย ค่า $S_I(X)$ คือ ค่าที่ใช้ในการพิจารณาโหนดตัดสินใจที่นำค่าความสำคัญของแอตทริบิวต์ไปปรับปรุง ค่า $S(X)$ คือ ค่าที่ใช้ในการพิจารณาโหนดตัดสินใจที่คำนวณจากชุดข้อมูล เช่น เกนสารสนเทศ หรือ อัตราส่วนเกน (Gain Ratio) เป็นต้น I_x คือ ค่าความสำคัญของแต่ละแอตทริบิวต์ซึ่งจะถูกกำหนดโดยผู้เชี่ยวชาญในแต่ละสาขาหรือการคำนวณจากชุดข้อมูลที่ศึกษา และ p คือ ค่าน้ำหนักในการพิจารณาโหนดสำหรับการตัดสินใจ โดยอัลกอริทึม IADT จะทำการกำหนดค่าน้ำหนักหรือ p ในขณะที่สร้างต้นไม้ตัดสินใจในแต่ละระดับแตกต่างกัน ซึ่งจะขึ้นอยู่กับจำนวนของข้อมูลที่ใช้ในขั้นตอนของการพิจารณาโหนดตัดสินใจในแต่ละโหนด สำหรับอัลกอริทึม IADT แอตทริบิวต์ใดมีค่า $S_I(X)$ สูงที่สุดแอตทริบิวต์นั้นจะถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจ ซึ่งผลการทดสอบอัลกอริทึม IADT พบว่า ต้นไม้ตัดสินใจที่ได้มีความถูกต้องและสามารถประมวลผลได้เร็วขึ้น

จากงานวิจัยดังกล่าวจะเห็นได้ว่าการนำองค์ความรู้ซึ่งอยู่ในรูปแบบต่าง ๆ มาใช้ในปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจนั้นช่วยให้ต้นไม้ตัดสินใจที่ได้มีขนาดลดลง รวมทั้งช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลอีกด้วย อย่างไรก็ตามวิธีการที่นำเสนอยังคงมีข้อจำกัดดังนี้

1) การนำแนวความคิดพื้นฐานของข้อมูลที่มีความสัมพันธ์มาใช้ในขั้นตอนการเรียนรู้ของต้นไม้ตัดสินใจนั้นจะเพิ่มความซับซ้อนของอัลกอริทึมเนื่องจากต้องทำการพิจารณาทั้งข้อมูลซึ่งเป็นข้อมูลต้นฉบับและข้อมูลซึ่งเป็นแนวความคิดพื้นฐานของข้อมูลนั้น ๆ เพื่อให้ได้แอตทริบิวต์ที่เหมาะสมที่สุดในการใช้เป็นโหนดของต้นไม้ตัดสินใจ

2) เมื่อพิจารณาการนำค่าความสำคัญของแอตทริบิวต์ที่นำมาใช้ในการปรับปรุงเกณฑ์ในการพิจารณาโหนดของต้นไม้ตัดสินใจ พบว่าค่าความสำคัญของแอตทริบิวต์เป็นค่าที่คำนวณ

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากความสัมพันธ์ระหว่างข้อมูลซึ่งอาจเกิดความคลาดเคลื่อนได้ หากข้อมูลที่นำมาวิเคราะห์มีความผิดปกติหรือไม่สมบูรณ์

3) การพิจารณาค่าความสำคัญของแอตทริบิวต์โดยผู้เชี่ยวชาญ ค่าความสำคัญของแอตทริบิวต์ที่ได้จะขึ้นอยู่กับความรู้และประสบการณ์ของผู้เชี่ยวชาญแต่ละคน ซึ่งอาจทำให้ค่าความสำคัญของแอตทริบิวต์มีความแตกต่างกันหรืออาจละเลยการพิจารณาความสัมพันธ์ของข้อมูลอื่น ๆ ที่เกี่ยวข้องทำให้ค่าความสำคัญของข้อมูลคลาดเคลื่อนได้

จากข้อจำกัดที่เกิดขึ้นผู้วิจัยจึงมีแนวความคิดในการลดข้อจำกัดของการพิจารณาค่าระดับความสำคัญของแอตทริบิวต์ด้วยการนำองค์ความรู้ในออนโทโลยีมาใช้ในการระบุค่าความสำคัญของแอตทริบิวต์ เพื่อช่วยลดความผิดพลาดที่อาจเกิดขึ้นในขั้นตอนของการพิจารณาแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจ

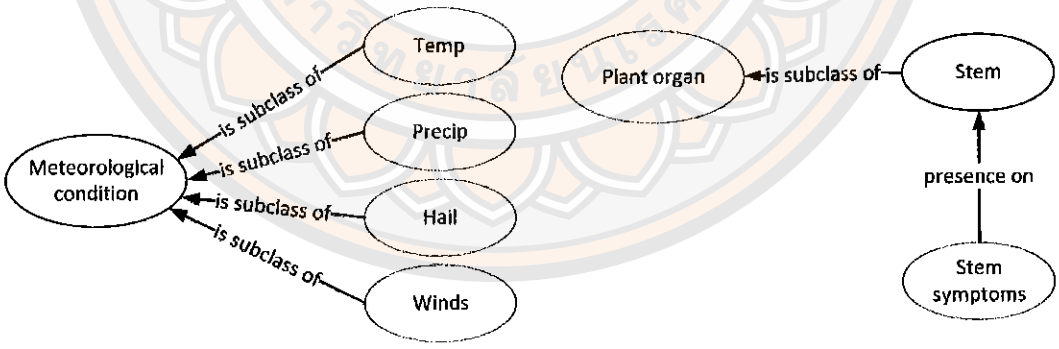
3. การประยุกต์ใช้วิธีการสรุปภาพรวมออนโทโลยีเพื่อสนับสนุนการทำเหมืองข้อมูล

ปัจจุบันออนโทโลยีมีขนาดและความซับซ้อนมากขึ้นส่งผลให้การนำองค์ความรู้ในออนโทโลยีมาช่วยสนับสนุนการวิเคราะห์ข้อมูลต้องใช้ระยะเวลาในการประมวลผลมากยิ่งขึ้น วิธีการสรุปภาพรวมออนโทโลยีเป็นวิธีการหนึ่งที่สามารถนำมาช่วยพิจารณาองค์ความรู้ที่มีความสำคัญและนำความรู้เหล่านั้นมาใช้ในการวิเคราะห์ข้อมูล ซึ่งอัลกอริทึม PageRank เป็นอัลกอริทึมหนึ่งที่นักวิจัยได้นำมาประยุกต์ใช้ เช่น Kralj et al. (2016) ได้นำอัลกอริทึม PageRank มาใช้ในการพิจารณาระดับความสำคัญของแต่ละแนวความคิดในออนโทโลยีและนำองค์ความรู้ที่มีความสำคัญเหล่านี้ไปใช้งานร่วมกับอัลกอริทึมที่มีชื่อว่า Hedwig (Vavpetič et al., 2013) ซึ่งเป็นอัลกอริทึมสำหรับการทำเหมืองข้อมูลเชิงความหมาย เพื่อสร้างแบบจำลองในการทำนายค่าข้อมูล โดยการนำเฉพาะองค์ความรู้ที่มีความสำคัญไปใช้งานจะช่วยให้อัลกอริทึมสามารถสร้างกฎในการวิเคราะห์ข้อมูลที่มีคุณภาพมากยิ่งขึ้น รวมถึงช่วยลดระยะเวลาที่ใช้ในการประมวลผลอีกด้วย นอกจากนี้ Kastrati & Imran (2019) ได้ประยุกต์ใช้อัลกอริทึม PageRank ในการหาค่าความสำคัญของแนวความคิดในออนโทโลยีและค่าระดับความสำคัญเหล่านี้จะถูกนำไปใช้ในการจำแนกข้อมูลเอกสารตามประเภทที่กำหนด โดยวิธีการของ Kastrai & Imran (2019) นี้ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีจะถูกนำไปประมวลผลร่วมกับค่าความสอดคล้องของข้อมูลในเอกสาร (concept relevance) เพื่อกำหนดค่าน้ำหนักให้กับแต่ละแนวความคิดที่ปรากฏในเอกสารนั้น ๆ และแปลงให้อยู่ในรูปแบบของเวคเตอร์เพื่อนำไปใช้ในการจำแนกข้อมูลด้วยอัลกอริทึมสำหรับการจำแนกข้อมูล ซึ่งพบว่าวิธีการนี้สามารถช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลได้ นอกจากนี้ Saranya et al. (2021) ได้นำเสนอระบบสำหรับการวินิจฉัยโรคของผู้ป่วยด้วยการประยุกต์ใช้แนวความคิดเกี่ยวกับออนโทโลยี โดยเมื่อผู้ใช้ทำการสอบถามข้อมูลที่เกี่ยวข้องกับอาการเจ็บป่วยระบบจะทำการค้นหาข้อมูลและองค์ความรู้ที่มีความสัมพันธ์กับข้อมูลสุขภาพของผู้ป่วยซึ่งอยู่ในรูปแบบของออนโทโลยีและใช้อัลทอริทึม PageRank

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ในการระบุความสำคัญของแต่ละโรคที่ปรากฏในออนโทโลยีเพื่อจัดอันดับความสำคัญของโรคที่อาจเกิดขึ้นและแสดงผลการวินิจฉัยโรค โดยผลการศึกษาพบว่าวิธีการที่นำเสนอนี้สามารถช่วยในการวินิจฉัยโรคของผู้ป่วยได้ดีขึ้น

จากงานวิจัยดังกล่าวจะพบว่าการประยุกต์ใช้ออนโทโลยีร่วมกับอัลกอริทึม PageRank สามารถช่วยสนับสนุนการวิเคราะห์ข้อมูลต่าง ๆ ได้ อย่างไรก็ตามอัลกอริทึม PageRank จะพิจารณาค่าระดับความสำคัญของแต่ละแนวความคิดจากจำนวนความถี่ของความสัมพันธ์ที่เชื่อมโยงกับแนวความคิดนั้น ๆ โดยไม่ได้นำความหมายของแต่ละความสัมพันธ์มาพิจารณา โดยแนวความคิดที่มีความถี่ในการเชื่อมโยงจากแนวความคิดอื่น ๆ สูง จะมีระดับความสำคัญสูงกว่าแนวความคิดอื่นที่มีความถี่ในการเชื่อมโยงน้อยกว่า ตัวอย่างเช่น ภาพ 9 แนวความคิด Meteorological condition ซึ่งเกี่ยวข้องกับสภาพแวดล้อมที่มีผลต่อการเกิดโรคของพืชมีการเชื่อมโยงจากแนวความคิดที่เกี่ยวข้องจำนวนสี่แนวความคิด และแนวความคิด Stem ซึ่งหมายถึง ลำต้นของต้นไม้ มีการเชื่อมโยงจากแนวความคิด Stem symptoms เพียงแนวความคิดเดียว เมื่อทำการพิจารณาระดับความสำคัญของแนวความคิดด้วยอัลกอริทึม PageRank จะทำให้แนวความคิด Meteorological condition มีระดับความสำคัญสูงกว่าแนวความคิด Stem เนื่องจากแนวความคิด Meteorological condition มีการเชื่อมโยงจากแนวความคิดอื่น ๆ มากกว่านั่นเอง ในขณะที่การวินัจฉัยโรคพืชนั้นในเบื้องต้นผู้เชี่ยวชาญจะพิจารณาอาการผิดปกติที่เกิดขึ้นในส่วนต่าง ๆ ของพืชก่อนการพิจารณาถึงสภาพแวดล้อมในขณะนั้น



ภาพ 9 ตัวอย่างแนวความคิดในออนโทโลยีโรคของถั่วเหลือง

จากข้อจำกัดที่เกิดขึ้นนี้นักวิจัยจึงได้นำเสนอวิธีการในการนำประเภทของความสัมพันธ์ในออนโทโลยีมาช่วยในการปรับปรุงการทำงานของอัลกอริทึม PageRank เช่น Jun et al. (2016) ได้นำเสนอวิธีการในการจัดลำดับความสำคัญของเว็บเพจโดยการประยุกต์ใช้อัลกอริทึม PageRank ที่มีการนำประเภทของการเชื่อมโยงข้อมูลในเอกสาร RDF มาช่วยในการทำงาน วิธีการนี้จะทำการหาค่าน้ำหนักของการเชื่อมโยงข้อมูลแต่ละประเภทและนำค่าน้ำหนักที่ได้ไปใช้ในการคำนวณหาค่าระดับ

4619881983 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ความสำคัญของแต่ละเว็บเพจ ผลศึกษาพบว่าวิธีการนี้สามารถจัดลำดับเว็บเพจได้อย่างมีประสิทธิภาพมากขึ้น โดยเว็บเพจที่มีข้อมูลที่สำคัญจะถูกจัดอันดับไว้สูงกว่าเว็บเพจที่ไม่ปรากฏข้อมูลที่สำคัญ

จากงานวิจัยที่เกี่ยวข้องผู้วิจัยจึงมีแนวความคิดที่จะนำเทคนิคการสรุปภาพรวมออนโทโลยีและความสัมพันธ์ระหว่างแต่ละแนวความคิดมาใช้ในการพิจารณาค่าระดับความสำคัญของแอตทริบิวต์ และนำค่าดังกล่าวไปใช้ในการปรับปรุงเกณฑ์ในการพิจารณาแอตทริบิวต์สำหรับใช้เป็นโหนดในต้นไม้ตัดสินใจ เพื่อช่วยลดปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีข้อมูลหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจ โดยแอตทริบิวต์ที่มีจำนวนค่าข้อมูลน้อยแต่เป็นแอตทริบิวต์ที่มีความสำคัญจะมีโอกาสถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจเพิ่มขึ้น ซึ่งสามารถช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลได้

บทสรุป

ในบทนี้ได้อธิบายถึงทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้องกับการปรับปรุงประสิทธิภาพของการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ เช่น ความรู้เกี่ยวกับเหมืองข้อมูล เหมืองข้อมูลเชิงความหมาย อัลกอริทึมต้นไม้ตัดสินใจ ออนโทโลยี วิธีการสรุปภาพรวมออนโทโลยี เป็นต้น โดยในการสร้างแบบจำลองต้นไม้ตัดสินใจนั้นอัลกอริทึมต้นไม้ตัดสินใจจะทำการสร้างแบบจำลองจากการพิจารณาชุดข้อมูลทดสอบเพื่อค้นหาแอตทริบิวต์ที่เหมาะสมสำหรับทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจ ซึ่งเป็นกระบวนการทำซ้ำเพื่อแตกกิ่งต้นไม้ตัดสินใจจนกระทั่งไม่สามารถแตกกิ่งออกไปได้อีก และนำแบบจำลองต้นไม้ตัดสินใจที่ได้ไปใช้ในการจำแนกข้อมูล

ในการเพิ่มประสิทธิภาพของการจำแนกข้อมูลของแบบจำลองต้นไม้ตัดสินใจนั้นนักวิจัยได้นำเสนอแนวทางในการดำเนินงานที่หลากหลาย เช่น การปรับปรุงข้อมูลที่ใช้สำหรับการสร้างแบบจำลองเพื่อช่วยลดจำนวนข้อมูลที่อัลกอริทึมใช้ในการสร้างแบบจำลอง ซึ่งช่วยให้ต้นไม้ตัดสินใจที่ได้มีความซับซ้อนลดลง และการปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจด้วยการปรับปรุงเกณฑ์ที่ใช้ในการพิจารณาแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจเพื่อให้อัลกอริทึมสามารถเลือกแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจได้อย่างเหมาะสมมากยิ่งขึ้น รวมถึงการนำเสนองานวิจัยที่มีการประยุกต์ใช้วิธีการสรุปภาพรวมออนโทโลยีเพื่อสนับสนุนการทำเหมืองข้อมูลซึ่งแนวทางเหล่านี้สามารถนำมาใช้ในการปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจได้ เป็นต้น

ในบทถัดไปจะอธิบายถึงขั้นตอนการดำเนินงานและกรอบแนวคิดในการดำเนินงานวิจัยเพื่อปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยี

4918891B3 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยเพื่อปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยีมีขั้นตอนการดำเนินงานดังนี้

- ข้อมูลที่ใช้ในการวิจัย
- เครื่องมือที่ใช้ในการดำเนินการวิจัย
- วิธีการดำเนินการวิจัย
- กรอบแนวคิดการวิจัย
- การวางแผนการทดลอง

ข้อมูลที่ใช้ในการวิจัย

ในการวิจัยครั้งนี้จะใช้ข้อมูลในการทดลองจำนวน 4 ชุดข้อมูลเพื่อทดสอบการทำงานของเทคนิคต้นไม้ตัดสินใจตามกรอบแนวคิดการวิจัยที่นำเสนอกับชุดข้อมูลในศาสตร์ที่หลากหลาย รวมถึงทดสอบการทำงานของวิธีการที่นำเสนอในชุดข้อมูลที่มีขนาดแตกต่างกัน โดยในการวิจัยนี้จะใช้ชุดข้อมูลมาตรฐานที่เผยแพร่โดยมหาวิทยาลัยแคลิฟอร์เนีย (University of California) รวมทั้งชุดข้อมูลที่ได้รับการเผยแพร่ในวารสารทางวิชาการจำนวนรวมทั้งสิ้น 4 ชุดข้อมูล ได้แก่ ชุดข้อมูลการเกิดโรคของถั่วเหลือง ชุดข้อมูลผู้ป่วยโรคหัวใจ (Dua & Graff, 2017) ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 (Viana dos Santos Santana et al., 2021) และชุดข้อมูลผู้ป่วยโรคไข้เลือดออก (Vianna Cardozo et al., 2018) ซึ่งมีรายละเอียดดังนี้

1. ชุดข้อมูลการเกิดโรคของถั่วเหลือง (Soybean dataset)

ชุดข้อมูลนี้ประกอบไปด้วยข้อมูลที่เกี่ยวข้องกับการปลูกถั่วเหลืองและลักษณะอาการของโรคที่เกิดขึ้นในส่วนต่าง ๆ ของถั่วเหลือง โดยชุดข้อมูลนี้มีข้อมูลทั้งสิ้น 683 เรคอร์ด ซึ่งประกอบไปด้วยแอตทริบิวต์จำนวน 36 แอตทริบิวต์ สำหรับจำแนกโรคของถั่วเหลืองออกเป็น 15 โรค โดยมีรายละเอียดของแต่ละแอตทริบิวต์ดังแสดงในตาราง 1

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 1 แอตทริบิวต์ภายในชุดข้อมูลการเกิดโรคของถั่วเหลือง

ลำดับ	ชื่อแอตทริบิวต์	ความหมาย
1	date	ช่วงเวลาที่ทำการปลูก
2	plant_stand	ลักษณะยืนต้นของพืช
3	precip	ความชื้น
4	temp	อุณหภูมิ
5	hail	การเกิดฝนหรือลูกเห็บ
6	crop-hist	ระยะเวลาที่ทำการปลูกพืชซ้ำ
7	area-damage	พื้นที่ที่เกิดความเสียหาย
8	severity	ความรุนแรงของโรค
9	seed-tmt	การเตรียมเมล็ด
10	germination	อัตราการงอกของเมล็ด
11	plant-growth	ระดับความสูงของต้น
12	leaves	ลักษณะของใบ
13	leafspots-halo	อาการใบมีจุดและมีวงรอบสีเหลือง (Halo)
14	leafspots-marg	อาการจุดบนใบมีลักษณะชุ่มน้ำ
15	leafspot-size	ขนาดของจุดบนใบ
16	leaf-shread	อาการใบร่วง
17	leaf-malf	อาการใบมีลักษณะผิดปกติ
18	leaf-mild	อาการเชื้อราบนใบ
19	stem	อาการของลำต้น
20	lodging	มีรากงอกบนลำต้น
21	stem-cankers	การเกิดแผลบนลำต้น
22	canker-lesion	สีของแผลบนลำต้น
23	fruiting bodies	มีฟรุตติ้งบอดี้ของเชื้อรา
24	external decay	ลักษณะของแผลบนลำต้น
25	mycelium	มีการเกิดไมซีเลียม
26	int-discolor	การเปลี่ยนสีของเนื้อเยื่อในลำต้น
27	sclerotia	มีการเกิดเส้นใย sclerotia
28	fruit-pods	อาการบนฝัก
29	fruit-spots	อาการเกิดจุดบนฝัก
30	seed	อาการบนเมล็ด
31	mold-growth	การเกิดเชื้อราบนเมล็ด
32	seed-discolor	การเปลี่ยนสีของเมล็ด
33	seed-size	ขนาดของเมล็ด
34	shriveling	อาการเมล็ดย่น
35	roots	อาการของราก
36	class	โรคของถั่วเหลือง

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

การวิจัยนี้ได้เลือกใช้ชุดข้อมูลการเกิดโรคของถั่วเหลืองในการทดลองเนื่องจากเป็นชุดข้อมูลที่ประกอบไปด้วยแอตทริบิวต์จำนวนมากโดยมีจำนวนแอตทริบิวต์ทั้งหมด 36 แอตทริบิวต์ และจำนวนค่าข้อมูลของแต่ละแอตทริบิวต์มีความแตกต่างกัน โดยแต่ละแอตทริบิวต์จะมีจำนวนค่าข้อมูลอยู่ระหว่าง 2 ค่า ถึง 7 ค่า ซึ่งหมายถึงชุดข้อมูลนี้มีโอกาสเกิดปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีข้อมูลหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจ

2. ชุดข้อมูลผู้ป่วยโรคหัวใจ (Heart disease dataset)

เป็นชุดข้อมูลสำหรับการจำแนกข้อมูลของผู้ป่วยโรคหัวใจและผู้ที่ไม่มีอาการโรคหัวใจ โดยมีจำนวนข้อมูลทั้งสิ้น 303 เรคอร์ด และมีแอตทริบิวต์ที่เกี่ยวข้องกับการเกิดโรคหัวใจจำนวน 14 แอตทริบิวต์ ดังแสดงในตาราง 2

ตาราง 2 แอตทริบิวต์ภายในชุดข้อมูลผู้ป่วยโรคหัวใจ

ลำดับ	ชื่อแอตทริบิวต์	ความหมาย
1	age	อายุผู้ป่วย
2	sex	เพศผู้ป่วย
3	Cp	รูปแบบอาการเจ็บหน้าอก
4	Thresbps	ค่าความดันโลหิต
5	Chol	ระดับคอเลสเตอรอล
6	Fbs	ระดับน้ำตาลในเลือด
7	Restecg	ผลการตรวจคลื่นไฟฟ้าหัวใจขณะพัก
8	Thalach	อัตราการเต้นของหัวใจสูงสุด
9	Exang	การเกิดอาการเจ็บหน้าอกขณะออกกำลังกาย
10	Oldpeak	ภาวะ ST-depression
11	Slope	รูปแบบของ ST segment
12	Ca	จำนวนเส้นเลือดที่พบจากการตรวจทางรังสี
13	Thal	สถานะของอัตราการเต้นของหัวใจ
14	Num	ผลการวินิจฉัยโรค

ชุดข้อมูลนี้ถูกนำมาใช้ในการทดลองเนื่องจากเป็นชุดข้อมูลขนาดเล็กโดยมีจำนวนตัวอย่างข้อมูลเพียง 303 เรคอร์ด รวมถึงจำนวนค่าข้อมูลในแต่ละแอตทริบิวต์มีความแตกต่างกัน โดยแต่ละแอตทริบิวต์มีจำนวนค่าข้อมูลอยู่ระหว่าง 2 ค่า ถึง 4 ค่า

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

3. ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 (COVID-19 dataset)

ชุดข้อมูลนี้เป็นชุดข้อมูลสำหรับการจำแนกผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และผู้ที่มีอาการปกติ จากการพิจารณาอาการบ่งชี้ต่าง ๆ ที่เกี่ยวข้องกับโรคนี้ โดยชุดข้อมูลประกอบไปด้วยข้อมูลจำนวน 3,128 เรคอร์ด และแอตทริบิวต์ที่เกี่ยวข้องจำนวน 11 แอตทริบิวต์ ดังแสดงในตาราง 3

ตาราง 3 แอตทริบิวต์ภายในชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019

ลำดับ	ชื่อแอตทริบิวต์	ความหมาย
1	Gender	เพศผู้ป่วย
2	Health professional	มีสถานะเป็นบุคลากรทางการแพทย์หรือไม่
3	Fever	อาการไข้
4	Sore throat	อาการเจ็บคอ
5	Dyspnea	อาการหายใจลำบาก
6	Olfactory disorder	ความผิดปกติทางการรับกลิ่น
7	Cough	อาการไอ
8	Coryza	อาการแบบหวัด
9	Taste disorder	ความผิดปกติทางการรับรส
10	Headache	อาการปวดหัว
11	Class	ผลการวินิจฉัยโรค

ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 ถูกนำมาใช้ในการวิจัยเนื่องจาก เป็นชุดข้อมูลที่มีจำนวนตัวอย่างข้อมูลจำนวนมากถึง 3,128 เรคอร์ด แต่ละแอตทริบิวต์จะประกอบไปด้วยจำนวน 2 ค่า คือ Positive และ Negative ซึ่งมีคุณลักษณะที่แตกต่างจากชุดข้อมูลการเกิดโรคของถั่วเหลืองและชุดข้อมูลผู้ป่วยโรคหัวใจ

4. ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก (Dengue fever dataset)

ชุดข้อมูลนี้เป็นชุดข้อมูลสำหรับการจำแนกข้อมูลผู้ป่วยออกเป็น 2 กลุ่ม คือ กลุ่มของผู้ป่วยโรคไข้เลือดออก และ กลุ่มข้อมูลผู้ป่วยโรคอื่นซึ่งมีอาการใกล้เคียงโรคไข้เลือดออก ซึ่งชุดข้อมูลนี้ประกอบไปด้วยแอตทริบิวต์ที่เกี่ยวข้องกับโรคไข้เลือดออกจำนวน 14 แอตทริบิวต์ และมีข้อมูลจำนวน 1,104 เรคอร์ด ซึ่งมีรายละเอียดดังตาราง 4

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 4 แอตทริบิวต์ภายในชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

ลำดับ	ชื่อแอตทริบิวต์	ความหมาย
1	age	อายุผู้ป่วย
2	gender	เพศผู้ป่วย
3	fever	อาการไข้
4	rash	เกิดผื่น
5	pruritus	อาการคัน
6	myalgia	อาการปวดเมื่อยกล้ามเนื้อ
7	arthralgia	อาการปวดข้อ
8	arthritis	อาการข้ออักเสบ
9	conjunctivitis	อาการเยื่อบุตาอักเสบ
10	headache	อาการปวดหัว
11	lymphadenopathy	อาการต่อมน้ำเหลืองโต
12	bleeding	ภาวะเลือดออก
13	neurological signs	อาการทางระบบประสาท
14	class	ผลการวินิจฉัยโรค

ชุดข้อมูลผู้ป่วยโรคไข้เลือดออกถูกนำมาใช้ในการวิจัยเนื่องจากเป็นชุดข้อมูลที่มีตัวอย่างข้อมูลจำนวน 1,104 เรคอร์ด ซึ่งแอตทริบิวต์ส่วนใหญ่มีจำนวนค่าข้อมูลของแอตทริบิวต์ 2 ค่า และแอตทริบิวต์ arthralgia มีจำนวนค่าข้อมูลที่แตกต่างกันจำนวน 3 ค่า

เครื่องมือที่ใช้ในการวิจัย

ในการวิจัยนี้ ได้มีการประยุกต์ใช้โปรแกรมคอมพิวเตอร์เพื่อทำการปรับปรุงประสิทธิภาพของต้นไม้ตัดสินใจดังนี้

- ภาษาคอมพิวเตอร์ไพธอน (Python) เวอร์ชั่น 3.7.6 สำหรับการพัฒนาอัลกอริทึมต่าง ๆ ที่เกี่ยวข้อง
- แพ็คเกจ Owlready2 เป็นแพ็คเกจในภาษาไพธอนเพื่อใช้ในการทำงานร่วมกับออนโทโลยี
- โปรแกรมโปรเทเจ (protégé) เวอร์ชั่น 5.2.0 เป็นเครื่องมือสำหรับการออกแบบและสร้างออนโทโลยี

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

วิธีการดำเนินการวิจัย

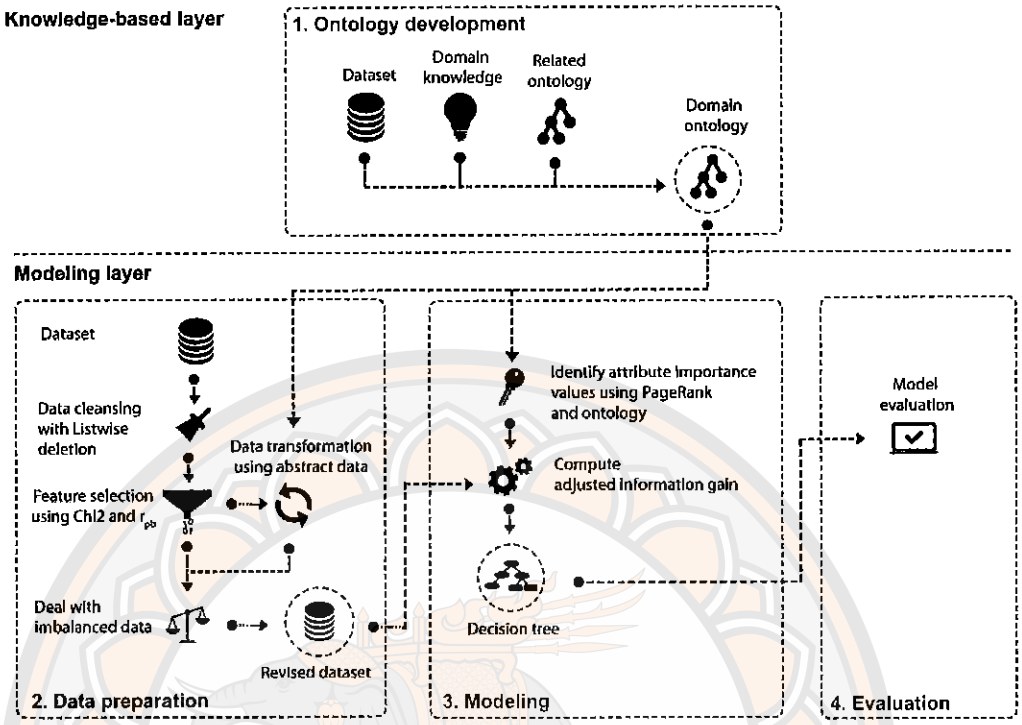
การวิจัยเพื่อปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจจะมีวิธีการดำเนินงานดังนี้

1. ศึกษาและทำความเข้าใจปัญหาที่เกิดขึ้นในการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ
2. ศึกษาทฤษฎี เอกสาร และงานวิจัยที่เกี่ยวข้อง
3. ศึกษาและรวบรวมชุดข้อมูลที่ใช้ในการทดลอง รวมทั้งออนโทโลยีที่เกี่ยวข้อง
4. ออกแบบกรอบแนวคิดในการดำเนินการวิจัย
5. ทำการออกแบบและพัฒนาออนโทโลยีเพื่อใช้ในการทดลอง
6. ทำการทดลองปรับปรุงประสิทธิภาพของต้นไม้ตัดสินใจและประเมินประสิทธิภาพของแบบจำลองที่ได้ โดยจะแบ่งการทำงานออกเป็น 2 ส่วน คือ
   - ทำการเตรียมข้อมูล (Data preparation) และการประยุกต์ใช้แนวความคิดพื้นฐานจากออนโทโลยีในการปรับปรุงข้อมูลที่ใช้สำหรับเทคนิคต้นไม้ตัดสินใจ
   - ทำการปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจ (Modeling) โดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยี
7. สรุปและวิเคราะห์ผลการวิจัย
8. รวบรวมและเขียนรายงานสรุปผลการวิจัย

การออกแบบกรอบแนวคิดการวิจัย

กรอบแนวคิดการวิจัยสำหรับการปรับปรุงประสิทธิภาพการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจโดยการประยุกต์ใช้ออนโทโลยีสามารถแสดงได้ดังภาพ 10 ซึ่งแบ่งการดำเนินงานออกเป็น 4 ส่วน คือ 1) การพัฒนาออนโทโลยี (Ontology development) 2) การจัดเตรียมข้อมูล (Data preparation) 3) การสร้างแบบจำลอง (Modeling) และ 4) การประเมินประสิทธิภาพของแบบจำลอง ซึ่งแต่ละส่วนมีรายละเอียดการดำเนินงานดังนี้

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39



ภาพ 10 กรอบแนวคิดการวิจัย

1. การพัฒนาออนโทโลยี (Ontology development)

เป็นขั้นตอนการในออกแบบและพัฒนาออนโทโลยีที่ใช้ในการสนับสนุนการจำแนกข้อมูล ในการพัฒนาและออกแบบออนโทโลยีนั้นจะดำเนินการโดยใช้เครื่องมือสำหรับการออกแบบและสร้าง ออนโทโลยี (Ontology Editor) ชื่อว่าโปรเทเจ (protégé) (Knublauch et al., 2004) ซึ่งมีขั้นตอน การดำเนินงานดังนี้

1.1 การกำหนดขอบเขตของออนโทโลยีและรวบรวมองค์ความรู้ที่เกี่ยวข้อง

การกำหนดขอบเขตของออนโทโลยีที่ใช้ในการดำเนินการวิจัยจะพิจารณาจาก ขอบเขตของชุดข้อมูลที่ใช้ในการดำเนินการวิจัยทั้ง 4 ชุดข้อมูล เพื่อใช้ในการรวบรวมออนโทโลยีที่มี การเผยแพร่ในวารสารวิชาการหรือแหล่งเรียนรู้อื่น ๆ โดยออนโทโลยีเหล่านี้ได้รับการวิเคราะห์ ออกแบบ และตรวจสอบความถูกต้องขององค์ความรู้โดยผู้เชี่ยวชาญในแต่ละศาสตร์ ออนโทโลยีที่ นำมาใช้ในการดำเนินงานวิจัยประกอบไปด้วย

- Soybean Ontology (Crop Ontology Curation, 2011) ซึ่งเป็นออนโทโลยีที่รวบรวมองค์ความรู้และแนวความคิดที่เกี่ยวข้องกับถั่วเหลือง

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

- Heart Failure Ontology (Wang, 2015) เป็นออนโทโลยีที่รวบรวมองค์ความรู้และแนวความคิดที่เกี่ยวข้องกับโรคหัวใจ

- COVID-19 Ontology (Sargsyan et al., 2020) คือ ออนโทโลยีที่รวบรวมองค์ความรู้เกี่ยวกับโรคติดเชื้อไวรัสโคโรนา 2019

- Dengue Fever Ontology (Mitraka et al., 2015) คือ ออนโทโลยีที่รวบรวมองค์ความรู้ที่เกี่ยวข้องกับโรคไข้เลือดออก

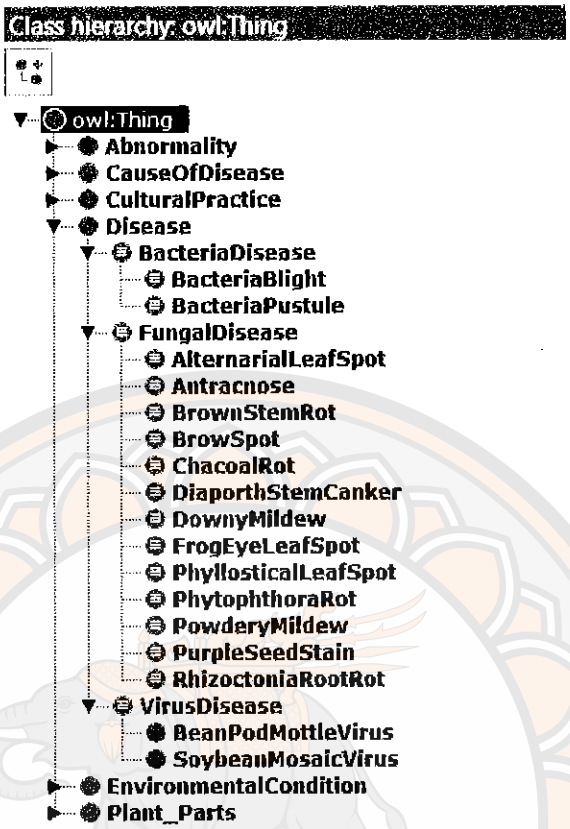
นอกจากนี้องค์ความรู้ต่าง ๆ ที่เกี่ยวข้องในการดำเนินการวิจัยจะถูกรวบรวมเพื่อใช้ในการปรับปรุงออนโทโลยีเพื่อให้ครอบคลุมสำหรับการจำแนกข้อมูล เช่น องค์ความรู้เกี่ยวกับการวินิจฉัยโรคถั่วเหลืองที่เผยแพร่โดยมหาวิทยาลัยนอร์ทดาโกต้าสเตท (North Dakota State University) (Markell & Malvick, 2018) และกฎที่ระบุโดยผู้เชี่ยวชาญในการระบุโรคของถั่วเหลืองจากงานวิจัยของ Michalski (1980) เป็นต้น ซึ่งองค์ความรู้เหล่านี้จะถูกนำมาใช้ในการปรับปรุงออนโทโลยีสำหรับการวิเคราะห์การเกิดโรคของถั่วเหลือง

1.2 การพัฒนาและปรับปรุงออนโทโลยี

ขั้นตอนนี้เป็นขั้นตอนในการปรับปรุงออนโทโลยีให้ครอบคลุมกับชุดข้อมูลที่ใช้ในการดำเนินการวิจัย โดยทำการพิจารณาแนวความคิด (Concept) ในออนโทโลยีร่วมกับแอตทริบิวต์ในชุดข้อมูลเพื่อตรวจสอบว่ามีแอตทริบิวต์ใดในชุดข้อมูลที่ไม่ได้เป็นส่วนประกอบในออนโทโลยีที่เกี่ยวข้อง และทำการกำหนดแนวความคิด ความสัมพันธ์ระหว่างแนวความคิด พร้อมทั้งทำการเพิ่มแนวความคิดนั้น ๆ ในออนโทโลยีเพื่อให้ครอบคลุมกับชุดข้อมูลที่ทำการศึกษา

สำหรับพัฒนาออนโทโลยีโรคของถั่วเหลืองนั้น จะมีการประยุกต์ใช้ Soybean Ontology (Crop Ontology Curation, 2011) ความรู้เกี่ยวกับโรคถั่วเหลือง (Markell & Malvick, 2018) และกฎการจำแนกโรคของถั่วเหลืองที่ระบุโดยผู้เชี่ยวชาญ (Michalski, 1980) มาใช้ในการดำเนินงาน โดยออนโทโลยีโรคของถั่วเหลืองที่พัฒนาขึ้นจะประกอบไปด้วย แนวความคิดที่เกี่ยวข้องกับโรคของถั่วเหลืองจำนวน 116 แนวความคิด ความสัมพันธ์ระหว่างแนวความคิดจำนวน 14 รายการ ตัวอย่างข้อมูลจำนวน 98 รายการ โดยตัวอย่างของออนโทโลยีโรคของถั่วเหลืองแสดงดังภาพ 11

4318891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39



ภาพ 11 ตัวอย่างออนโทโลยีโรคของถั่วเหลือง

สำหรับการวิเคราะห์ข้อมูลผู้ป่วยโรคหัวใจ ผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และผู้ป่วยโรคไข้เลือดออกนั้น ออนโทโลยีที่เกี่ยวข้องกับโรคนั้น ๆ จะถูกนำปรับปรุงโดยการเพิ่มแนวความคิดที่เกี่ยวข้องกับข้อมูลผู้ป่วย เช่น เพศ อายุ เพื่อให้สอดคล้องตามชุดข้อมูลที่ทำการศึกษา

ออนโทโลยีโรคหัวใจ (Wang, 2015) ประกอบไปด้วยแนวความคิดที่เกี่ยวข้องกับโรคหัวใจจำนวน 1,658 แนวความคิด ซึ่งมีความสัมพันธ์ระหว่างแนวความคิดจำนวน 2 รายการ สำหรับออนโทโลยีสำหรับโรคติดเชื้อไวรัสโคโรนา 2019 (Sargsyan et al., 2020) ประกอบไปด้วยแนวความคิดทั้งหมด 2,276 แนวความคิด มีความสัมพันธ์ระหว่างข้อมูลจำนวน 11 รายการ และตัวอย่างข้อมูลจำนวน 6 รายการ ในส่วนของออนโทโลยีสำหรับโรคไข้เลือดออก (Mitraka et al., 2015) นั้นจะประกอบไปด้วย แนวความคิดที่เกี่ยวข้องกับการเกิดโรคไข้เลือดออกจำนวน 5,030 แนวความคิดและความสัมพันธ์ระหว่างข้อมูลจำนวน 25 รายการ ซึ่งสามารถสรุปข้อมูลของออนโทโลยีที่นำไปใช้ในการปรับปรุงประสิทธิภาพของต้นไม้ตัดสินใจได้ดังตาราง 5

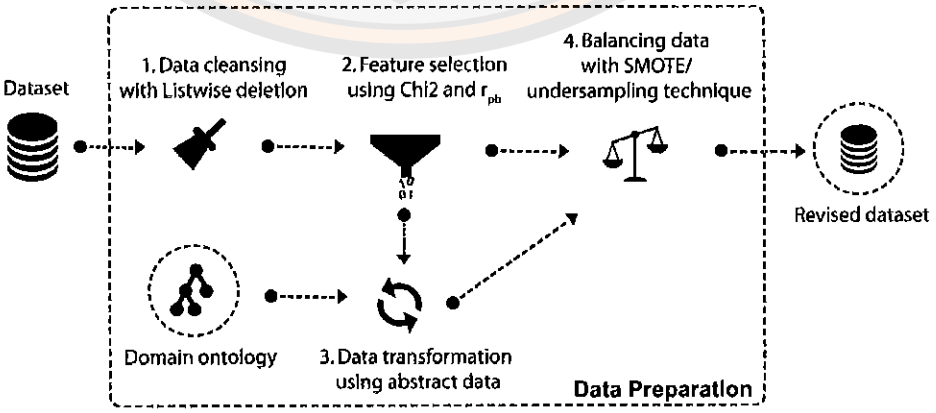
4719889183 NU iThesis 60031257 thesis / recv: 31102555 16:06:32 / seq: 39

ตาราง 5 สรุปข้อมูลออนโทโลยีที่ใช้ในการวิจัย

ออนโทโลยีที่ใช้ในการวิจัย	จำนวนแนวความคิด	จำนวนความสัมพันธ์	จำนวนตัวอย่างข้อมูล	ออนโทโลยีต้นแบบ
ออนโทโลยีโรคของถั่วเหลือง	116	14	98	Soybean Ontology(Crop Ontology Curation, 2011)
ออนโทโลยีโรคหัวใจ	1,658	2	0	Heart Failure (Wang, 2015)
ออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019	2,276	11	6	COVID-19 Ontology (Sargsyan et al., 2020)
ออนโทโลยีโรคไข้เลือดออก	5,030	25	0	Dengue Fever Ontology (Mitraka et al., 2015)

2. การจัดเตรียมข้อมูล (Data preparation)

เป็นขั้นตอนการจัดเตรียมข้อมูลเพื่อให้พร้อมสำหรับการวิเคราะห์ข้อมูลซึ่งมีการทำงานดังแสดงในภาพ 12



ภาพ 12 กระบวนการจัดเตรียมข้อมูล

4918891837 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

การจัดเตรียมข้อมูลจะประกอบไปด้วย 4 ขั้นตอน ได้แก่

2.1 การทำความสะอาดข้อมูล (Data Cleansing)

เป็นกระบวนการในการตรวจสอบความสมบูรณ์ของข้อมูล โดยการพิจารณาข้อมูลที่สูญหายเพื่อดำเนินการกับข้อมูลที่สูญหายเหล่านั้น เช่น การลบแถวของข้อมูลที่สูญหาย การเติมค่าข้อมูลที่สูญหายด้วยค่าที่เหมาะสม เป็นต้น

ในการวิจัยนี้ผู้วิจัยได้เลือกใช้เทคนิค Listwise Deletion (Baraldi & Enders, 2010) ซึ่งเป็นวิธีการที่จะนำแถวข้อมูลที่มีรายการข้อมูลสูญหายออกจากชุดข้อมูล เมื่อดำเนินการกับข้อมูลที่สูญหายแล้วจะทำให้ชุดข้อมูลการเกิดโรคถั่วเหลืองมีจำนวนข้อมูล 562 เรคอร์ด และชุดข้อมูลผู้ป่วยโรคหัวใจมีจำนวนข้อมูลทั้งหมด 297 เรคอร์ด สำหรับชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออกนั้นไม่ปรากฏข้อมูลที่สูญหาย

2.2 การเลือกคุณลักษณะของข้อมูล (Feature Selection)

เป็นกระบวนการในการพิจารณาความสัมพันธ์ระหว่างแอตทริบิวต์กับคลาสที่ต้องการจำแนกเพื่อลดจำนวนแอตทริบิวต์ที่ใช้ในการสร้างแบบจำลอง โดยแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสที่ต้องการจำแนกจะไม่ถูกนำมาใช้ในขั้นตอนการสร้างแบบจำลอง เนื่องจากหากนำแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสที่ต้องการจำแนกมาใช้งานอาจส่งผลให้อัลกอริทึมเรียนรู้ข้อมูลเหล่านั้นเพื่อทำการสร้างแบบจำลอง ทำให้แบบจำลองที่ได้มีความผิดปกติและส่งผลต่อประสิทธิภาพของการจำแนกข้อมูล รวมถึงอาจทำให้เกิดปัญหาความจำเพาะกับข้อมูลที่เรียนรู้ (Overfitting) ได้ นอกจากนี้การสร้างแบบจำลองโดยใช้แอตทริบิวต์จำนวนมากจำเป็นต้องใช้ระยะเวลาและทรัพยากรในการวิเคราะห์ข้อมูล ดังนั้นการเลือกเฉพาะแอตทริบิวต์ที่มีความสำคัญต่อคลาสที่ต้องการจำแนกจึงเป็นวิธีการหนึ่งที่ช่วยลดระยะเวลาในการวิเคราะห์ข้อมูลได้อีกด้วย (Honest, 2020) ในการวิจัยนี้จะทำการพิจารณาความสัมพันธ์ระหว่างข้อมูลโดยใช้วิธีการทางสถิติ คือ สัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล (Point biserial correlation) และค่าสถิติไคสแควร์ (Chi-square)

สำหรับสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลนั้นจะเป็นวิธีการสำหรับการพิจารณาความสัมพันธ์ระหว่างข้อมูลที่เป็นข้อมูลแบบอัตรภาคชั้น (Interval Variable) กับข้อมูลที่มีการแบ่งกลุ่มออกเป็นสองกลุ่ม ซึ่งสามารถแสดงได้ดังสมการ (11) (Verma, 2019)

$$r_{pb} = \frac{\bar{x}_p - \bar{x}_q}{s_t} \times \sqrt{pq} \qquad (11)$$

โดย r_{pb} คือ ค่าสหสัมพันธ์แบบพอยท์ไบซีเรียล

$\bar{x}_p$ คือ ค่าเฉลี่ยของตัวแปรแบบอันตรภาคชั้นของกลุ่มที่มีค่าที่หนึ่ง (p)

4918831 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

$\bar{x}_q$ คือ ค่าเฉลี่ยของตัวแปรแบบอันตรภาคชั้นของกลุ่มที่มีค่าที่สอง (q)

p คือ สัดส่วนของจำนวนข้อมูลของกลุ่มที่มีค่าที่หนึ่ง

q คือ สัดส่วนของจำนวนข้อมูลของกลุ่มที่มีค่าที่สอง

s_t คือ ค่าเบี่ยงเบนมาตรฐานของชุดข้อมูล

ค่าสถิติไคสแควร์เป็นค่าสถิติที่ใช้สำหรับการพิจารณาความเป็นอิสระต่อกันของข้อมูล หรือการพิจารณาว่าข้อมูลมีความสัมพันธ์กันหรือไม่ โดยเป็นการทดสอบสำหรับข้อมูลแบบนามบัญญัติ (Nominal Variable) ค่าสถิติไคสแควร์สามารถคำนวณได้จากสมการ (12) (McHugh, 2013)

$$\chi^2 = \frac{(O-E)^2}{E} \tag{12}$$

โดย O คือ ความถี่ที่ได้จากการสังเกต (Observed Frequency)

E คือ ค่าความถี่ที่คาดหวัง (Expected Frequency)

ในการพิจารณาความสัมพันธ์ระหว่างข้อมูลสามารถเขียนสมมติฐานหลัก (H_0) และสมมติฐานรอง (H_1) ได้ดังนี้

H_0: แอตทริบิวต์ในชุดข้อมูลไม่มีความสัมพันธ์กับคลาสคำตอบ

H_1: แอตทริบิวต์ในชุดข้อมูลมีความสัมพันธ์กับคลาสคำตอบ

หากค่า *p-value* ที่ได้จากค่าสถิติที่ทำการทดสอบมีค่าน้อยกว่าหรือเท่ากับ 0.05 จะปฏิเสธสมมติฐานหลักและยอมรับสมมติฐานรอง นั่นคือ แอตทริบิวต์ที่ทดสอบมีความสัมพันธ์กับคลาสคำตอบอย่างมีนัยสำคัญทางสถิติ ในกรณีที่ *p-value* ของสถิติที่ทดสอบมีค่ามากกว่า 0.05 จะยอมรับสมมติฐานหลักและปฏิเสธสมมติฐานรองซึ่งหมายถึง แอตทริบิวต์ที่ทำการทดสอบไม่มีความสัมพันธ์กับคลาสคำตอบอย่างมีนัยสำคัญทางสถิติ และแอตทริบิวต์นั้นจะไม่ถูกนำไปใช้ในการสร้างแบบจำลองสำหรับการจำแนกข้อมูล สถิติที่ใช้ในการทดสอบความสัมพันธ์ระหว่างข้อมูลสำหรับแต่ละชุดข้อมูลที่ศึกษาสามารถแสดงได้ดังตาราง 6

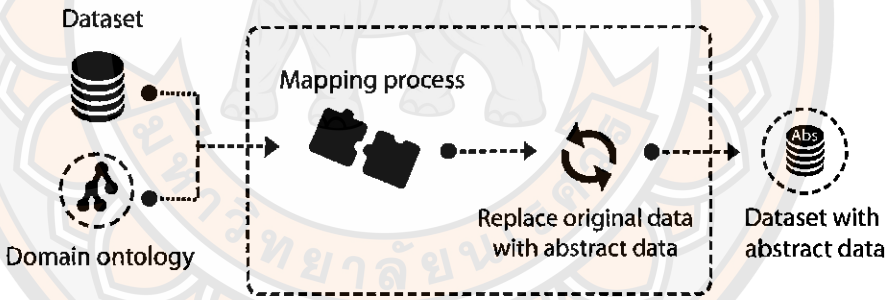
NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 6 สรุปข้อมูลสถิติที่ใช้ในการทดสอบความสัมพันธ์ระหว่างข้อมูล

ชุดข้อมูล	สถิติที่ใช้ในการทดสอบความสัมพันธ์ของข้อมูล
ชุดข้อมูลการเกิดโรคของถั่วเหลือง	ค่าสถิติไคสแควร์
ชุดข้อมูลผู้ป่วยโรคหัวใจ	สัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล และค่าสถิติไคสแควร์
ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019	ค่าสถิติไคสแควร์
ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก	สัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล และค่าสถิติไคสแควร์

2.3 การแปลงข้อมูล (Data Transformation)

เป็นขั้นตอนในการแปลงข้อมูลที่ใช้สำหรับการจำแนกข้อมูล โดยในขั้นตอนนี้จะมีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีเพื่อปรับปรุงข้อมูลดังแสดงในภาพ 13



ภาพ 13 ขั้นตอนการประยุกต์ใช้ออนโทโลยีในการปรับปรุงข้อมูล

อัลกอริทึมสำหรับกระบวนการแปลงข้อมูลโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยีจะพัฒนาด้วยภาษาไพธอน และทำการเชื่อมต่อกับออนโทโลยีซึ่งอยู่ในรูปแบบไฟล์ OWL ด้วยไลบรารี Owlready2 (Lamy, 2017) ซึ่งเป็นไลบรารีของภาษาไพธอนที่ใช้ในการเข้าถึงและจัดการออนโทโลยี กระบวนการแปลงข้อมูลนี้สามารถแบ่งขั้นตอนการทำงานออกเป็น 2 ส่วน คือ

- **การจับคู่แนวความคิดในออนโทโลยีกับข้อมูลในชุดข้อมูล** (Mapping process) ในขั้นตอนนี้เป็นกระบวนการสำหรับค้นหาแนวความคิดพื้นฐานที่มีความสัมพันธ์กับข้อมูลในชุดข้อมูล โดยแอตทริบิวต์และค่าข้อมูลของแอตทริบิวต์จะถูกนำมาตรวจสอบกับแนวความคิดภายในออนโทโลยี หากแอตทริบิวต์ในชุดข้อมูลตรงกับคลาสแม่ (Superclass) ของแนวความคิดใด ๆ

431889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ในออนโทโลยีแล้ว ค่าข้อมูลของแอตทริบิวต์จะถูกนำไปตรวจสอบกับตัวอย่างข้อมูล (Instance) ที่เป็นสมาชิกของแนวความคิดนั้น หากค่าข้อมูลของแอตทริบิวต์และตัวอย่างข้อมูลในออนโทโลยีมีค่าตรงกันแล้ว แนวความคิดซึ่งทำหน้าที่เป็นคลาสแม่ (Superclass) ของตัวอย่างข้อมูลจะถูกใช้เป็นแนวความคิดพื้นฐาน (Abstract data) ที่มีความสัมพันธ์กับค่าข้อมูลของแอตทริบิวต์ที่ทำการตรวจสอบ การจับคู่แนวความคิดในออนโทโลยีกับข้อมูลในชุดข้อมูลสามารถนำเสนอรหัสเทียม (Pseudo Code) ได้ดังภาพ 14 (Chanmee & Kesorn, 2020)

Algorithm : Abstract data identification.

Input: A list of classes in an ontology ($\{C\}$), an attribute in dataset (a_i), and a value of attribute (v_{ai})
Output: Related abstract data

```
1   FOR each class c_i where c_i ∈ C
2     IF attribute a_i match with parent class of c_i
3       IF list of instances of c_i ({I}) exist
4         FOR each instance ins_i where ins_i ∈ I
5           IF value of attribute v_ai match with instance ins_i
6             RETURN c_i
7           ENDIF
8         ENDFOR
9       ENDIF
10    ENDIF
11  ENDFOR
```

ภาพ 14 รหัสเทียม (Pseudo Code) สำหรับการระบุแนวความคิดพื้นฐานที่สัมพันธ์กับข้อมูล

- การแปลงข้อมูลในชุดข้อมูลด้วยแนวความคิดที่อ้างอิงได้จากออนโทโลยี (Replace original data with abstract data) ขั้นตอนนี้จะทำการสร้างแอตทริบิวต์ขึ้นใหม่โดยนำค่าแนวความคิดพื้นฐานที่อ้างอิงได้จากขั้นตอนก่อนหน้าไปเป็นค่าข้อมูลของแอตทริบิวต์ และแอตทริบิวต์ที่สร้างขึ้นใหม่นี้จะถูกนำไปใช้แทนแอตทริบิวต์เดิมในกระบวนการสร้างแบบจำลอง รหัสเทียมของขั้นตอนนี้สามารถแสดงได้ดังภาพ 15 (Chanmee & Kesorn, 2020)

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

Algorithm : Replacing the original data with related abstract data

Input : Dataset (D), an ontology (O)
Output : A dataset with abstract data

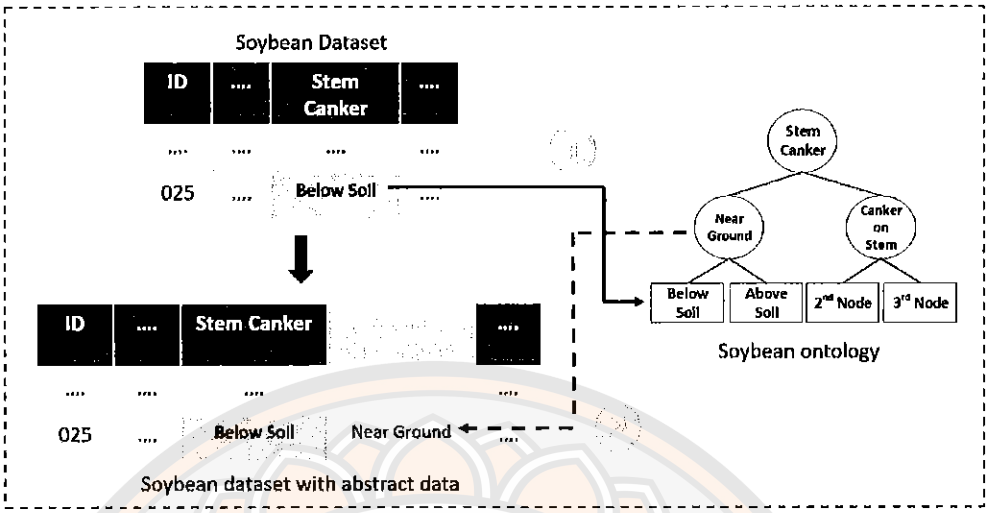
```
1   Initial the empty set {Ab} for abstract data
2   Find the list of class {C} of ontology
3   // Identifying the abstract data of each attribute's value
4   FOR each attribute a_i where a_i ∈ D
5       FOR all unique values v_ai of attribute a_i
6           InferClass = call abstract data identification algorithm({C}, a_i, v_ai)
7           update {Ab} with InferClass
8       ENDFOR
9   ENDFOR
10  //Updating the dataset with the abstract data
11  FOR each attribute a_i where a_i ∈ D
12      Count_Infer = count the number of the abstract values of attribute a_i
13      IF Count_Infer is greater than zero
14          Duplicate attribute a_i as new attribute a_infer
15      ENDIF
16      FOR each row of dataset
17          Update attribute a_infer with {Ab}
18      ENDFOR
19  ENDFOR
```

ภาพ 15 รหัสเทียม (Pseudo Code) สำหรับการแปลงข้อมูลเดิมด้วยแนวความคิดพื้นฐานที่สัมพันธ์กับข้อมูล

ตัวอย่างการอ้างอิงแนวความคิดพื้นฐานจากออนโทโลยีสามารถแสดงได้ดังภาพ 16 โดยโหนด near ground คือ คลาสแม่ (Superclass) ภายในออนโทโลยีโรคของถั่วเหลือง ซึ่งมีโหนด below soil และ โหนด above soil เป็นตัวอย่างข้อมูล (Instance) ในขั้นตอนที่ 1 ค่าข้อมูลของแอตทริบิวต์ stem canker จะถูกนำไปค้นหาองค์ความรู้ที่เกี่ยวข้องภายในออนโทโลยี ซึ่งหากพบข้อมูลที่มีความสัมพันธ์กับค่าข้อมูลดังกล่าว จะทำการสร้างแอตทริบิวต์ใหม่ชื่อ stem canker_infer และนำแนวความคิดพื้นฐานที่สัมพันธ์กับค่าข้อมูลในแอตทริบิวต์ไปแทนที่ข้อมูลเดิมดังแสดงในขั้นตอนที่ 2 ซึ่งจะนำค่า near ground ซึ่งคลาสแม่ หรืออีกนัยหนึ่งคือ เป็นแนวความคิดพื้นฐานที่สัมพันธ์กับค่า below soil ไปใช้งาน โดยจะดำเนินการเช่นนี้กับข้อมูลทุกรายการและนำข้อมูลที่อ้างอิงได้จากออนโทโลยีไปใช้เป็นข้อมูลนำเข้าสำหรับการจำแนกข้อมูล

สำหรับการวิจัยในครั้งนี้ออนโทโลยีโรคของถั่วเหลืองเป็นออนโทโลยีเดียวที่ปรากฏตัวอย่างข้อมูลภายในออนโทโลยี จึงทำให้ชุดข้อมูลการเกิดโรคของถั่วเหลืองเป็นชุดข้อมูลเดียวที่มีการดำเนินการแปลงข้อมูล

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39



ภาพ 16 ตัวอย่างการอ้างอิงแนวความคิดพื้นฐานจากออนโทโลยี

2.4 การจัดการกับข้อมูลที่ไม่สมดุล (Deal with imbalanced data)

ปัญหาข้อมูลที่ไม่สมดุล (imbalanced data) (Kaur et al., 2019) เป็นปัญหาที่เกิดขึ้นในกรณีที่ชุดข้อมูลมีการแบ่งข้อมูลออกเป็นกลุ่มต่าง ๆ ในอัตราส่วนที่แตกต่างกัน โดยกลุ่มหลัก (Majority class) จะมีข้อมูลในกลุ่มเป็นจำนวนมาก ในขณะที่กลุ่มอื่น ๆ (Minority class) จะมีจำนวนข้อมูลในกลุ่มน้อยกว่า ซึ่งการนำชุดข้อมูลที่มีปัญหาข้อมูลไม่สมดุลไปใช้ในการสร้างแบบจำลองจะส่งผลต่อความถูกต้องของแบบจำลองการจำแนกข้อมูลที่ได้ โดยค่าความถูกต้องในการจำแนกข้อมูลที่ได้อาจเป็นค่าความถูกต้องสำหรับข้อมูลกลุ่มหลักเพียงกลุ่มเดียว ในขณะที่การจำแนกข้อมูลในกลุ่มอื่น ๆ ยังคงมีความคลาดเคลื่อน

ในการวิจัยครั้งนี้ชุดข้อมูลการเกิดโรคของถั่วเหลือง และชุดข้อมูลผู้ป่วยโรคไข้เลือดออก เป็นชุดข้อมูลที่พบปัญหาความไม่สมดุลของข้อมูล เพื่อแก้ปัญหาดังกล่าวการวิจัยนี้จึงได้เลือกเทคนิค Synthetic Minority Oversampling Technique (SMOTE) และวิธีการสุ่มลดข้อมูล (Undersampling technique) โดยมีรายละเอียดดังนี้

- ชุดข้อมูลการเกิดโรคของถั่วเหลือง จะทำการแก้ปัญหาข้อมูลที่ไม่สมดุลด้วยเทคนิค SMOTE ซึ่งเป็นวิธีการสุ่มตัวอย่างข้อมูลในกลุ่มรองให้มีจำนวนใกล้เคียงกับจำนวนข้อมูลกลุ่มหลัก (Chawla et al., 2002) เนื่องจากเป็นชุดข้อมูลที่มีข้อมูลเพียง 562 เรคคอร์ด

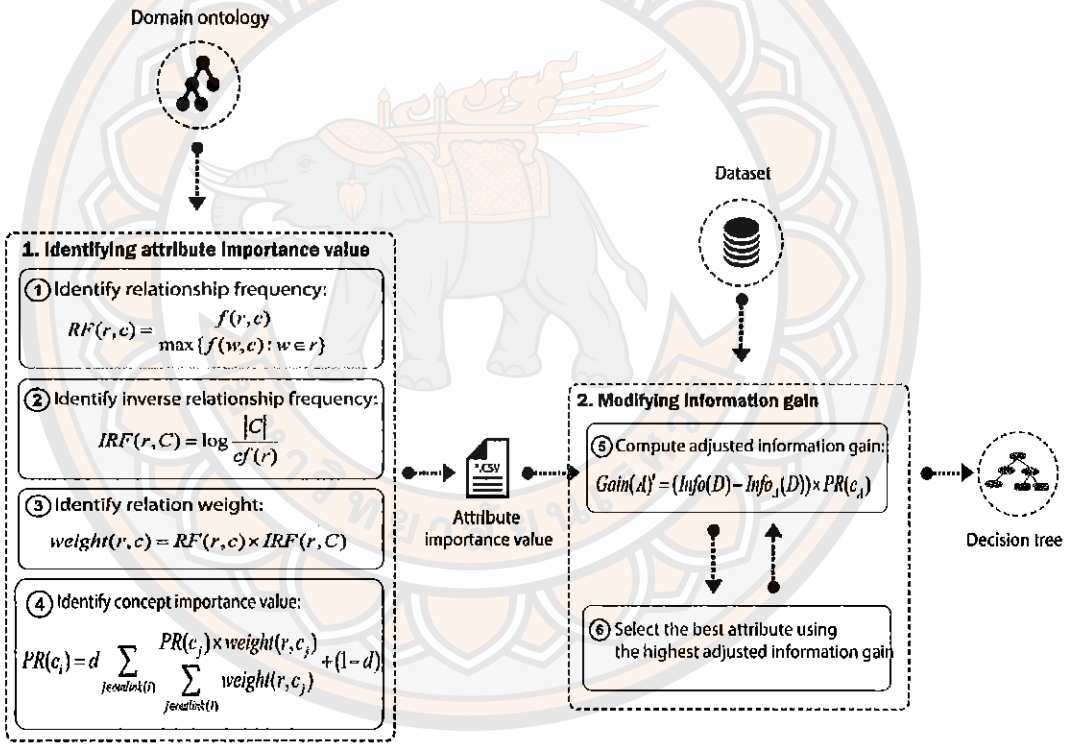
- ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก จะถูกปรับปรุงข้อมูลด้วยวิธีการสุ่มลดข้อมูล (Undersampling technique) ซึ่งเป็นวิธีการในการสุ่มลดจำนวนข้อมูลของกลุ่มหลักให้มี

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จำนวนลดลง และมีจำนวนใกล้เคียงกับจำนวนข้อมูลในกลุ่มรอง (Kaur et al., 2019) เนื่องจากเป็นชุดข้อมูลที่มีจำนวนข้อมูล 1,286 เรคคอร์ดซึ่งเพียงพอต่อการสร้างแบบจำลองการจำแนกข้อมูล

3. การสร้างแบบจำลอง (Modeling)

เป็นขั้นตอนในการสร้างแบบจำลองการจำแนกข้อมูลโดยใช้อัลกอริทึมต้นไม้ตัดสินใจและองค์ความรู้ภายในออนโทโลยีเพื่อช่วยแก้ปัญหาความลำเอียงในการพิจารณาแอตทริบิวต์ที่มีค่าหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจ ซึ่งจะแบ่งการทำงานออกเป็น 2 ขั้นตอน คือ การระบุค่าความสำคัญของแต่ละแอตทริบิวต์ (Identifying attribute Importance value) และการปรับปรุงค่าเกนสารสนเทศ (Modifying information gain) สามารถแสดงขั้นตอนการทำงานดังภาพ 17



ภาพ 17 กระบวนการสร้างแบบจำลองโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยี

3.1 การระบุค่าระดับความสำคัญของแต่ละแอตทริบิวต์ (Identifying attribute Importance value)

เป็นขั้นตอนการคำนวณหาค่าระดับความสำคัญของแต่ละแอตทริบิวต์โดยการประยุกต์ใช้อัลกอริทึม Weighted Semantic PageRank (Jun et al., 2016) ซึ่งเป็นอัลกอริทึมสำหรับการวิเคราะห์ระดับความสำคัญของเว็บเพจที่มีการนำคำอธิบายของข้อมูลอาร์ดีเอฟ (RDF

4918891B3
NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

metadata) มาใช้ในการคำนวณ สำหรับการวิจัยนี้โครงสร้างและความสัมพันธ์ของข้อมูลภายในออนโทโลยีจะถูกนำมาใช้เพื่อระบุระดับความสำคัญของแต่ละแนวความคิดในออนโทโลยี และนำค่าระดับความสำคัญของแนวความคิดที่ได้ไปใช้เป็นค่าระดับความสำคัญของแอตทริบิวต์ในขั้นตอนการปรับปรุงค่าเกนสารสนเทศ

- **ขั้นตอนการหาค่าระดับความสำคัญของแนวความคิด** การหาค่าความสำคัญของแนวความคิดในออนโทโลยีจะดำเนินการโดยการหาค่าความถี่ของความสัมพันธ์ในออนโทโลยี (Relationship frequency) ค่าส่วนกลับของความถี่ของความสัมพันธ์ (Inverse relationship frequency) ค่าน้ำหนักของแต่ละความสัมพันธ์ (Relation weight) เพื่อนำไปใช้ในการหาค่าความสำคัญของแต่ละแนวความคิดตามลำดับ

ค่าความถี่ของความสัมพันธ์จะคำนวณโดยการนำจำนวนของความสัมพันธ์ที่สนใจซึ่งปรากฏในแนวความคิดหารด้วยจำนวนความสัมพันธ์ใด ๆ ที่มีค่ามากที่สุดที่ปรากฏในแนวความคิดเดียวกัน ดังแสดงในสมการ (13)

$$RF(r,c) = \frac{f(r,c)}{\max\{f(w,c) : w \in r\}} \tag{13}$$

โดย $RF(r,c)$ คือ ค่าความถี่ของความสัมพันธ์ r ที่ปรากฏในแนวความคิด c

$f(r,c)$ คือ จำนวนความสัมพันธ์ r ที่พบในแนวความคิด c

$f(w,c)$ คือ จำนวนความสัมพันธ์ w ที่พบในแนวความคิด c

ค่าส่วนกลับของความถี่ความสัมพันธ์ คือ ค่าที่แสดงระดับความสำคัญของแต่ละความสัมพันธ์ในออนโทโลยี ซึ่งสามารถคำนวณได้จากสมการ (14)

$$IRF(r,C) = \log \frac{|C|}{cf(r)} \tag{14}$$

โดย $IRF(r,C)$ คือ ค่าส่วนกลับของความถี่ความสัมพันธ์ r ที่ปรากฏในออนโทโลยี

$|C|$ คือ จำนวนแนวความคิดทั้งหมดในออนโทโลยี

$cf(r)$ คือ จำนวนแนวความคิดทั้งหมดที่มีความสัมพันธ์ r ปรากฏเพื่อเชื่อมโยงไปยังแนวความคิดอื่น

สำหรับค่าน้ำหนักของแต่ละความสัมพันธ์ในออนโทโลยีนั้นสามารถคำนวณได้จากสมการ (15)

$$weight(r,c) = RF(r,c) \times IRF(r,C) \tag{15}$$

4916889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

โดย $weight(r,c)$ คือ ค่าน้ำหนักของความสัมพันธ์ r ที่มีการเชื่อมโยงแนวความคิด c ไปยังแนวความคิดอื่น

$RF(r,c)$ คือ ค่าความถี่ของความสัมพันธ์ r ที่ปรากฏในแนวความคิด c

$IRF(r,C)$ คือ ค่าส่วนกลับของความถี่ความสัมพันธ์ r ที่ปรากฏในออนโทโลยี

รหัสเทียมสำหรับการคำนวณค่าส่วนกลับของความถี่ความสัมพันธ์ และค่าน้ำหนักของแต่ละความสัมพันธ์ในออนโทโลยีสามารถแสดงได้ดังภาพ 18 และ ภาพ 19 ตามลำดับ

Algorithm : Identifying Inverse relationship frequency

Input : Ontology graph (G), Relationships (R)
Output : IRF values
1 Initial the empty set of IRF values $\{IRF\}$
2 *total_concept* = count all concepts in ontology
3 **FOR** each relationships r_i where $r_i \in R$ and $r_i \in G$
4 *start_node* = count concepts that use r_i as the outgoing link
5 $IRF[r_i]$ = log(*total_concept* / *start_node*)
6 **ENDFOR**
7 **RETURN** $\{IRF\}$

ภาพ 18 รหัสเทียม (Pseudo Code) สำหรับการคำนวณค่าส่วนกลับของความถี่ของความสัมพันธ์ในออนโทโลยี

จากภาพ 18 ในการคำนวณหาค่าส่วนกลับของความถี่ความสัมพันธ์นั้น ออนโทโลยีที่นำมาใช้จะถูกพิจารณาในลักษณะกราฟของออนโทโลยี (Ontology graph, G) ซึ่งประกอบไปด้วย แนวความคิดที่ทำหน้าที่เป็นโหนดต้นทาง แนวความคิดที่เป็นโหนดปลายทาง และความสัมพันธ์ระหว่างแนวความคิด โดยในกรณีที่แนวความคิดในออนโทโลยีมีความสัมพันธ์แบบ Subclass-of แนวความคิดที่ทำหน้าที่เป็นคลาสลูกจะถูกกำหนดเป็นโหนดต้นทาง และแนวความคิดที่เป็นคลาสแม่จะถูกกำหนดเป็นโหนดปลายทางที่มีความสัมพันธ์ Subclass-of เชื่อมโยงแนวความคิดทั้งสอง ในกรณีแนวความคิดมีความสัมพันธ์ลักษณะอื่น ๆ จะพิจารณาแนวความคิดที่เป็นโหนดต้นทางและโหนดปลายทางตามความหมายของการเชื่อมโยงแนวความคิด เช่น ในออนโทโลยีโรคของถั่วเหลือง ปรากฏความสัมพันธ์ของแนวความคิดดังนี้ Bacteria Blight has-symptom leaf-shred ดังนั้น แนวความคิด Bacteria Blight จะทำหน้าที่เป็นโหนดต้นทาง และแนวความคิด leaf-shred จะทำหน้าที่เป็นโหนดปลายทางซึ่งทำการเชื่อมโยงกันด้วยความสัมพันธ์ has-symptom เป็นต้น

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ขั้นตอนในการคำนวณหาค่าส่วนกลับของความถี่ความสัมพันธ์จะมีการทำงานดังนี้

1) หาค่าจำนวนแนวความคิดทั้งหมดที่ปรากฏในออนโทโลยี
2) ทำการวนรอบการทำงานตามลิสต์รายการความสัมพันธ์ทั้งหมดที่ปรากฏอยู่ในออนโทโลยี (Relationships, *R*) เพื่อนับจำนวนแนวความคิดที่มีความสัมพันธ์ที่สนใจ (r_i) เชื่อมโยงไปยังแนวความคิดอื่น ๆ และนำค่าจำนวนแนวความคิดที่ได้ไปคำนวณค่าส่วนกลับของความถี่ของความสัมพันธ์ตามสมการ (14)
3) เมื่อวนรอบการทำงานครบทุกรายการความสัมพันธ์ อัลกอริทึมจะคืนค่าส่วนกลับของความถี่ของความสัมพันธ์ทั้งหมดที่คำนวณได้เพื่อนำไปใช้งานต่อไป

Algorithm : Identifying weight of relationship

```
    Input: Ontology graph (G), Relationships (R)
    Output: weigh of relationships
1   Initial the empty set of relationship's weight {relation_weight}
2   IRF = call the algorithm Identifying Inverse relationship frequency (G,R)
3   FOR each node ci where ci ∈ G
4       relation_of_ci = set of the relationships that start with node ci
5       IF count(relation_of_ci) != 0
6           max_relationship = max(number of each relationship in relation_of_ci)
7           FOR each relationship ri of node ci where ri ∈ R and ri ∈ G
8               // Identifying the relationship frequency values of relationship ri
9               RF = frequency of ri / max_relationship
10              // Identify the weight of relationship ri which outgoing link of node ci
11              relation_weight[ci,ri] = RF×IRF[ri]
12          ENDFOR
13      ENDIF
14  ENDFOR
15  RETURN {relation_weight}
```

ภาพ 19 รหัสเทียม (Pseudo Code) การคำนวณค่าน้ำหนักของแต่ละความสัมพันธ์ในออนโทโลยี

จากภาพ 19 การคำนวณหาค่าน้ำหนักของความสัมพันธ์ในออนโทโลยีนั้นจะนำเข้าข้อมูลจากออนโทโลยีในรูปแบบกราฟของออนโทโลยี โดยมีขั้นตอนการทำงานดังนี้

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

1) เรียกใช้อัลกอริทึม Identifying Inverse Relationship frequency ในภาพ 18 เพื่อหาค่าส่วนกลับของความถี่ของความสัมพันธ์ทั้งหมดในออนโทโลยี

2) วนรอบการทำงานตามแนวความคิดทั้งหมดที่ปรากฏในออนโทโลยี โดย

- ทำการรวบรวมรายการความสัมพันธ์ที่มีการเชื่อมโยงออกจากแนวความคิดที่พิจารณา
- ตรวจสอบว่าแนวความคิดนั้นเป็นมีความสัมพันธ์ใด ๆ เชื่อมโยงออกไปยังแนวความคิดอื่นหรือไม่ ถ้ามีให้ทำการหาจำนวนของความสัมพันธ์ที่มีความถี่สูงที่สุด เพื่อนำไปใช้ในการคำนวณหาค่าความถี่ของความสัมพันธ์ดังสมการ (13)
- ทำการวนรอบการทำงานตามลิสต์รายการความสัมพันธ์ทั้งหมดเพื่อคำนวณค่าน้ำหนักของความสัมพันธ์ตามสมการ (15)

3) เมื่อวนรอบการทำงานจนครบทุกแนวความคิดทั้งหมดในออนโทโลยี อัลกอริทึมจะคืนค่าน้ำหนักของความสัมพันธ์ทั้งหมดเพื่อนำไปใช้ในการทำงานขั้นต่อไป

หลังจากการคำนวณค่าน้ำหนักของความสัมพันธ์ในออนโทโลยีแล้ว ค่าน้ำหนักของความสัมพันธ์ที่ได้จะถูกนำไปใช้ในการหาค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโดยใช้อัลกอริทึม Weighted Semantic PageRank ซึ่งสามารถแสดงได้ดังสมการ (16)

$$PR(c_i) = d \sum_{j \in outlink(i)} \frac{PR(c_j) \times weight(r, c_j)}{\sum_{k \in outlink(j)} weight(r_k, c_j)} + (1-d) \qquad (16)$$

โดย $PR(c_i)$ คือ ค่าระดับความสำคัญของแนวความคิด i

$PR(c_j)$ คือ ค่าระดับความสำคัญของแนวความคิด j ซึ่งมีการเชื่อมโยงมายังแนวความคิด i

$weight(r, c_j)$ คือ ค่าน้ำหนักของความสัมพันธ์ r ที่เชื่อมโยงแนวความคิด j ไปยังแนวความคิด i

$weight(r_k, c_j)$ คือ ค่าน้ำหนักของความสัมพันธ์ k ที่เชื่อมโยงจากแนวความคิด j ไปยังแนวความคิดใด ๆ

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

d คือ damping factor ซึ่งเป็นตัวแปรสำหรับถ่วงน้ำหนักเพื่อป้องกันปัญหาค่าระดับความสำคัญคลาดเคลื่อนจากกรณีที่แนวคิดใด ๆ ไม่มีการเชื่อมโยงไปยังแนวความคิดอื่น ๆ (rank leak) หรือ กรณีแนวคิดมีการเชื่อมโยงกันแบบวนรอบ (rank sink) โดย d จะมีค่าเท่ากับ 0.85

การคำนวณหาค่าระดับความสำคัญของแนวความคิดในออนโทโลยีนั้นจะเป็นการทำงานในสมการ (16) แบบวนซ้ำ เมื่อค่าระดับความสำคัญของแนวความคิดไม่มีการเปลี่ยนแปลงจึงจะหยุดการคำนวณ รหัสเทียมสำหรับการคำนวณค่าระดับความสำคัญของแนวความคิดในออนโทโลยีแสดงดังภาพ 20

Algorithm : Weighted Semantic PageRank

Input: Ontology graph (G), Relationships (R)
Output: Concept importance value

Initial empty set for concept importance value {concept_weight}
d = 0.85
all_weight = call the algorithm *Identifying weight of relationship (G,R)*
N = number of concepts in G
// Initial default importance value of each concept
FOR each node c_i where $c_i \in G$
    $PR(c_i) = 1/N$
ENDFOR
// Identify concept importance value
REPEAT
    FOR each node c_i that has links from node c_j
    // Summation of the weight of relationship r_i that associated with c_j, $r_i \in R$ and $r_i \in G$
        $sum\_weight[c_j] = \sum(all\_weight[c_j, r_i])$
        $PR(c_i) = d \times \left(\frac{PR(c_j) \times all\_weight[c_j, r_i]}{sum\_weight[c_j]} \right) + (1-d)$
        update {*concept_weight*} with $PR(c_i)$
    ENDFOR
UNTIL the importance value $PR(c_i)$ of all concepts does not change
RETURN {*concept_weight*}

ภาพ 20 รหัสเทียม (Pseudo Code) สำหรับการคำนวณค่าระดับความสำคัญของแนวความคิดในออนโทโลยีด้วย Weighted Semantic PageRank

จากภาพ 20 แนวความคิดและความสัมพันธ์ระหว่างข้อมูลในออนโทโลยีจะถูกนำไปใช้ในรูปแบบกราฟของออนโทโลยีเพื่อคำนวณหาค่าระดับความสำคัญของแต่ละแนวความคิด โดยมีขั้นตอนการทำงานดังนี้

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

1) กำหนดค่าตัวแปรสำหรับถ่วงน้ำหนัก หรือ damping factor ให้มีค่าเท่ากับ 0.85

2) คำนวณหาค่าน้ำหนักของทุกความสัมพันธ์ที่ปรากฏในออนโทโลยีโดยการเรียกใช้อัลกอริทึม Identifying weight of relationship ซึ่งแสดงในภาพ 19

3) นับจำนวนแนวความคิดทั้งหมดในออนโทโลยีเพื่อใช้ในการกำหนดค่าระดับความสำคัญเริ่มต้นให้แต่ละแนวความคิด โดยค่าระดับความสำคัญเริ่มต้นของแต่ละแนวความคิดจะมีค่าเท่ากับหนึ่งหารด้วยจำนวนแนวความคิดทั้งหมด

4) ทำการวนรอบตามแนวความคิดที่มีการเชื่อมโยงกันในออนโทโลยีโดย

- ทำการหาผลรวมของค่าน้ำหนักของความสัมพันธ์ของแนวความคิดอื่น ที่มีการเชื่อมโยงมายังแนวคิดที่สนใจเพื่อใช้ในการคำนวณค่าระดับความสำคัญของแนวความคิดนั้น ๆ
- คำนวณหาค่าระดับความสำคัญของแนวความคิดตามสมการ (16)

5) ทำซ้ำกระบวนการในข้อที่ 4 จนกระทั่งค่าระดับความสำคัญของทุกแนวความคิดไม่มีการเปลี่ยนแปลงจึงหยุดการคำนวณ และคืนค่าระดับความสำคัญของแนวความคิดที่คำนวณได้

โดยค่าระดับความสำคัญของแนวความคิดในออนโทโลยีที่คำนวณได้จะถูกจัดเก็บในรูปแบบไฟล์ CSV เพื่อนำไปใช้ในการปรับปรุงค่าเกนสารสนเทศในขั้นตอนต่อไป

3.2 การปรับปรุงค่าเกนสารสนเทศ (Modifying information gain)

ในขั้นตอนนี้จะนำค่าระดับความสำคัญของแนวความคิดในออนโทโลยีที่คำนวณได้จากขั้นตอนก่อนหน้ามาทำหน้าที่เป็นค่าระดับความสำคัญของแอตทริบิวต์เพื่อใช้ในการปรับปรุงค่าเกนสารสนเทศ ซึ่งเกนสารสนเทศที่ทำการปรับปรุงสามารถคำนวณได้จากสมการ (17)

$$Gain(A)' = (Info(D) - Info_A(D)) \times PR(c_A) \quad (17)$$

โดย $Gain(A)'$ คือ ค่าเกนสารสนเทศที่ทำการปรับปรุง

$Info(D)$ คือ ค่าเอนโทรปีของชุดข้อมูล

$Info_A(D)$ คือ ค่าเอนโทรปีของแอตทริบิวต์

$PR(c_A)$ คือ ค่าระดับความสำคัญของแอตทริบิวต์

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ค่าเกนสารสนเทศที่ทำการปรับปรุงนี้จะถูกนำไปใช้เป็นเกณฑ์ในการพิจารณาแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดภายในต้นไม้ตัดสินใจ โดยแอตทริบิวต์ที่มีค่าเกนสารสนเทศที่ทำการปรับปรุงสูงที่สุดจะถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจ ซึ่งการพิจารณาโหนดภายในต้นไม้ตัดสินใจด้วยค่าเกนสารสนเทศที่ทำการปรับปรุงนี้จะช่วยลดปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีข้อมูลหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจได้ โดยแอตทริบิวต์ที่มีค่าเกนสารสนเทศน้อยแต่มีค่าระดับความสำคัญสูงจะมีโอกาสถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจมากขึ้น รหัสเทียมสำหรับอัลกอริทึมการสร้างต้นไม้ตัดสินใจที่มีการประยุกต์ใช้องค์ความรู้จากออนโทโลยีสามารถแสดงดังภาพ 21

Algorithm : Semantic Decision Tree

Input: Dataset (D), Target attribute (a_{target}), Attribute importance values (PR)
Output: Decision tree
Initial empty set for decision tree $Tree = \{\}$
IF samples in D are all the same class or other stopping criteria is invoked
    Create leaf node that correspond to the most frequency class of D
ENDIF
FOR each attribute a_i where $a_i \in D$ and $a_i \neq a_{target}$
    //Compute adjusted information gain of attribute a_i
    $Gain(a_i)' = (Info(D) - Info_{a_i}(D)) \times PR(a_i)$
ENDFOR
a_{best} = Attribute that obtains the highest $Gain(a_i)'$
$Tree$ = Create a node of the decision tree based on attribute a_{best}
D_j = Create sub-dataset from D based on attribute a_{best}
FOR each attribute a_j where $a_j \in D_j$ and $a_j \neq a_{target}$
    //call recursive algorithm: Semantic Decision Tree
    $Tree_j$ = call the algorithm *Semantic Decision Tree*(D_j, a_{target}, PR)
    Attach $Tree_j$ to the corresponding branch of tree $Tree$
ENDFOR
RETURN $Tree$

ภาพ 21 รหัสเทียม (Pseudo Code) สำหรับการสร้างต้นไม้ตัดสินใจที่มีการประยุกต์ใช้องค์ความรู้จากออนโทโลยี หรือ Semantic Decision Tree

จากภาพ 21 ชุดข้อมูล (Dataset) แอตทริบิวต์ที่ทำหน้าที่เป็นคลาสคำตอบ (Target attribute) และค่าระดับความสำคัญของแต่ละแอตทริบิวต์ที่คำนวณได้จากอัลกอริทึม Weighted Semantic PageRank จะถูกนำมาใช้ในการสร้างต้นไม้ตัดสินใจเชิงความหมาย ซึ่งมีขั้นตอนการทำงานดังนี้

1) ตรวจสอบข้อมูลในชุดข้อมูลว่าทุกแถวในชุดข้อมูลนั้นอยู่ในคลาสคำตอบเดียวกัน หรือ ตรวจสอบว่าข้อมูลตรงกับเงื่อนไขสำหรับหยุดการสร้างต้นไม้ตัดสินใจ

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

หรือไม่ เช่น สร้างต้นไม้ตัดสินใจจนถึงระดับความสูงที่กำหนด ซึ่งหากเงื่อนไขเป็นจริง จะทำการสร้างโหนดใบของต้นไม้ตัดสินใจโดยให้คลาสคำตอบที่มีจำนวนแถวข้อมูลสูงที่สุดเป็นค่าคำตอบสำหรับโหนดใบนั้น หลังจากนั้นจะไปทำงานในขั้นตอนถัดไป

2) ทำการวนรอบการทำงานสำหรับทุกแอตทริบิวต์ที่ไม่ได้ทำหน้าเป็นคลาสคำตอบในชุดข้อมูล เพื่อทำการคำนวณหาค่าเกนสามารถสนเทศที่ถูกปรับปรุงด้วยค่าน้ำหนักของแอตทริบิวต์นั้น ๆ และทำการเลือกแอตทริบิวต์ที่มีค่าเกนสารสนเทศที่ทำการปรับปรุงสูงที่สุดเพื่อทำหน้าที่เป็นโหนดภายในต้นไม้ตัดสินใจ

3) แบ่งข้อมูลออกเป็นชุดย่อย ๆ ตามค่าของแอตทริบิวต์ที่ถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจจากขั้นตอนก่อนหน้า

4) ทำซ้ำกระบวนการทำงานทั้งหมดจนกระทั่งข้อมูลตรงตามเงื่อนไขสำหรับการหยุดสร้างต้นไม้ตัดสินใจ จึงหยุดการทำงานและคืนค่าต้นไม้ตัดสินใจเชิงความหมาย

4. การประเมินประสิทธิภาพของแบบจำลอง

ขั้นตอนนี้เป็นขั้นตอนในการประเมินประสิทธิภาพของต้นไม้ตัดสินใจที่สร้างขึ้น โดยการพิจารณาค่าความถูกต้องในการจำแนกข้อมูล (Accuracy)

ค่าความถูกต้องในการจำแนกข้อมูล (Accuracy) สามารถคำนวณได้จากสมการ (18) (Han et al., 2011)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

โดย TP (True Positive) คือ จำนวนข้อมูลที่มีการจำแนกคลาสที่สนใจได้ถูกต้อง

TN (True Negative) คือ จำนวนข้อมูลที่มีการจำแนกคลาสที่ไม่สนใจได้ถูกต้อง

FP (False Positive) คือ จำนวนข้อมูลที่มีการจำแนกข้อมูลผิด โดยทำการจำแนกคลาสอื่น ๆ เป็นคลาสที่กำลังสนใจ

FN (False Negative) คือ จำนวนข้อมูลที่มีการจำแนกข้อมูลผิด โดยทำการจำแนกคลาสที่สนใจเป็นคลาสอื่น ๆ

การวางแผนการทดลอง

ในการวิจัยครั้งนี้ได้แบ่งการทดลองออกเป็น 2 ส่วน ได้แก่ การทดลองสำหรับขั้นตอนการเตรียมข้อมูล และ การทดลองสำหรับขั้นตอนการสร้างแบบจำลองการจำแนกข้อมูล ซึ่งมีการทดลองแสดงดังภาพ 22

NU ithesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ภาพ 22 แผนการทดลองสำหรับการปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมาย

1. การทดลองสำหรับขั้นตอนการเตรียมข้อมูล

สำหรับขั้นตอนการเตรียมข้อมูลโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยีนั้นจะแบ่งการทดลองออกเป็น 4 การทดลอง ดังนี้

1.1 การพิจารณาความสัมพันธ์ระหว่างข้อมูล

การทดลองนี้มีวัตถุประสงค์เพื่อพิจารณาความสัมพันธ์ระหว่างแอตทริบิวต์กับคลาสที่ต้องการจำแนกเพื่อลดจำนวนของแอตทริบิวต์ที่ใช้ในการสร้างแบบจำลองการจำแนกข้อมูล

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39
4918891B3

ซึ่งการพิจารณาความสัมพันธ์ระหว่างข้อมูลนี้จะดำเนินการโดยใช้สัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล (Point biserial correlation) และค่าสถิติไคสแควร์ (Chi-square)

1.2 การอ้างอิงแนวความคิดพื้นฐานจากออนโทโลยี

การทดลองนี้มีวัตถุประสงค์เพื่อประยุกต์ใช้องค์ความรู้ในออนโทโลยีในการปรับปรุงข้อมูลที่ใช้สำหรับการจำแนกข้อมูล ในการทดลองนี้ค่าข้อมูลของแต่ละแอตทริบิวต์ภายในชุดข้อมูลจะถูกนำไปค้นหาตัวอย่างข้อมูลในออนโทโลยีซึ่งมีความสัมพันธ์กับค่าข้อมูลนั้น และนำค่าแนวความคิดในออนโทโลยีที่มีความสัมพันธ์กับตัวอย่างข้อมูลไปใช้เป็นแนวความคิดพื้นฐาน (Abstract data) เพื่อปรับปรุงชุดข้อมูล

1.3 การปรับปรุงข้อมูลด้วยแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยี

การทดลองนี้มีวัตถุประสงค์เพื่อการสำรวจการเปลี่ยนแปลงที่เกิดขึ้นในชุดข้อมูลเมื่อมีการนำค่าข้อมูลพื้นฐานที่อ้างอิงได้จากออนโทโลยีมาใช้ในการปรับปรุงข้อมูลที่มีความสัมพันธ์

1.4 การจำแนกข้อมูลที่มีการนำแนวความคิดพื้นฐานมาใช้งาน

การทดลองนี้มีวัตถุประสงค์เพื่อตรวจสอบผลของการปรับปรุงข้อมูลด้วยองค์ความรู้ในออนโทโลยีที่มีต่อประสิทธิภาพการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจ การทดลองนี้จะใช้อัลกอริทึม ID3 ในการจำแนกข้อมูลเพื่อทำการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลที่ได้จากชุดข้อมูลเดิมและชุดข้อมูลที่ได้ทำการปรับปรุงด้วยองค์ความรู้ในออนโทโลยี ซึ่งจะทำการทดลองจำแนกข้อมูลจำนวน 30 ครั้ง โดยทำการสุ่มข้อมูลเพื่อแบ่งเป็นข้อมูลสำหรับการเรียนรู้ (training data) จำนวนร้อยละ 70 จากข้อมูลทั้งหมด และเป็นข้อมูลสำหรับการทดสอบ (test data) จำนวนร้อยละ 30 ในการทดลองนี้จะทำการทดสอบกับชุดข้อมูลการเกิดโรคของถั่วเหลืองเนื่องจากเป็นชุดข้อมูลเดียวที่สามารถอ้างอิงแนวความคิดพื้นฐานในออนโทโลยีได้

การพิจารณาประสิทธิภาพของการจำแนกข้อมูลในการทดลองนี้จะพิจารณาจากค่าความถูกต้องในการจำแนกข้อมูล ความสูงของต้นไม้ตัดสินใจ และ จำนวนโหนดที่ใช้ในการสร้างต้นไม้ตัดสินใจ โดยจะพิจารณาประสิทธิภาพการจำแนกข้อมูลใน 2 กรณี คือ ในกรณีที่ต้นไม้ตัดสินใจมีความสูงที่มากที่สุด และในกรณีที่มีการใช้ค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสม

2. การทดลองสำหรับขั้นตอนการสร้างแบบจำลองการจำแนกข้อมูล

สำหรับการทดลองสร้างแบบจำลองการจำแนกข้อมูลโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยีจะประกอบไปด้วยการทดลอง 8 การทดลอง เพื่อพิจารณาประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจที่พัฒนาขึ้นในประเด็นต่าง ๆ เช่น ความถูกต้องในการจำแนกข้อมูล การเกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting) ความทนทานกับชุดข้อมูลที่ผิดปกติ รวมถึงโครงสร้างของต้นไม้ตัดสินใจที่เปลี่ยนแปลงไปเมื่อมีการประยุกต์ใช้องค์ความรู้ภายในออนโทโลยี

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

การประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อไม่ปรากฏองค์ความรู้ที่เกี่ยวข้องในออนโทโลยี โดยมีรายละเอียดของการทดลองดังนี้

2.1 การคำนวณค่าระดับความสำคัญจากออนโทโลยี

การทดลองนี้มีวัตถุประสงค์เพื่อทำการหาค่าระดับความสำคัญของแต่ละแนวความคิดที่อยู่ในออนโทโลยี และนำค่าระดับความสำคัญของแนวความคิดที่มีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูลมาใช้เป็นค่าระดับความสำคัญของแอตทริบิวต์เพื่อใช้ในการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจ ซึ่งการหาค่าระดับความสำคัญของแนวความคิดในออนโทโลยีจะประยุกต์ใช้แนวความคิดของการสรุปภาพรวมออนโทโลยี (Ontology Summarization) ในการดำเนินการ

2.2 การทดสอบประสิทธิภาพการจำแนกข้อมูลของต้นไม้ตัดสินใจเชิงความหมาย

การทดลองนี้มีวัตถุประสงค์เพื่อทดสอบประสิทธิภาพในการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจที่มีการประยุกต์ใช้องค์ความรู้ในออนโทโลยี โดยค่าระดับความสำคัญของแนวความคิดในออนโทโลยีที่คำนวณได้จากขั้นตอนก่อนหน้าจะถูกนำมาใช้เป็นค่าระดับความสำคัญของแอตทริบิวต์เพื่อปรับปรุงค่าเกนสารสนเทศที่ใช้ในการพิจารณาแอตทริบิวต์สำหรับเป็นโหนดของต้นไม้ตัดสินใจ

ในการทดลองนี้จะทำการเปรียบเทียบค่าความถูกต้องในการจำแนกข้อมูลของอัลกอริทึมที่ได้รับการปรับปรุงด้วยองค์ความรู้ในออนโทโลยีกับอัลกอริทึม ID3 ที่มีการใช้ค่าเกนสารสนเทศในการพิจารณาแอตทริบิวต์สำหรับใช้เป็นโหนดของต้นไม้ตัดสินใจ โดยทำการทดลองจำแนกข้อมูลทั้ง 4 ชุดข้อมูล และทำการทดลองจำแนกข้อมูลจำนวน 30 ครั้ง โดยแต่ละครั้งจะทำการสุ่มข้อมูลเพื่อแบ่งข้อมูลออกเป็นข้อมูลสำหรับการเรียนรู้จำนวนร้อยละ 70 จากข้อมูลทั้งหมด และข้อมูลสำหรับการทดสอบจำนวนร้อยละ 30

2.3 การทดสอบความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting)

การทดลองนี้มีวัตถุประสงค์เพื่อสำรวจโอกาสในการเกิดความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้เมื่อมีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีในการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจ โดยทำการเปรียบเทียบกับอัลกอริทึม ID3

ในการพิจารณาโอกาสในการเกิดความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้นั้น จะทำการพิจารณาจากค่าความแตกต่างระหว่างค่าความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้กับค่าความถูกต้องในการจำแนกข้อมูลสำหรับการทดสอบ หากอัลกอริทึมใดมีค่าความแตกต่างของค่าความถูกต้องในการจำแนกข้อมูลทั้งสองชุดน้อยจะหมายถึง อัลกอริทึมนั้นมีโอกาสในการเกิดความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้น้อยนั่นเอง ซึ่งการพิจารณาค่าความแตกต่างของค่าความถูกต้องของการจำแนกข้อมูลนี้จะพิจารณาผลของการจำแนกข้อมูลในทุกระดับความสูงของต้นไม้ตัดสินใจเพื่อระบุระดับความสูงที่มีโอกาสเกิดความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ของแต่ละอัลกอริทึม

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

นอกจากนี้ในการทดลองนี้จะมีการประยุกต์ใช้วิธีการทางสถิติ เช่น การทดสอบที (T-Test) หรือการทดสอบวิลคอกซัน (Wilcoxon Signed Rank Test) เพื่อพิจารณาความแตกต่างระหว่างผลลัพธ์ที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจที่ปรับปรุงกับผลลัพธ์ที่ได้จากอัลกอริทึม ID3 โดยมีการกำหนดสมมติฐานทางสถิติดังนี้

H_0: ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และค่าความถูกต้องในการจำแนกข้อมูลทดสอบที่ได้จากอัลกอริทึม ID3 และค่าดังกล่าวที่ได้จากอัลกอริทึมที่ปรับปรุงไม่แตกต่างกัน

H_1: ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และค่าความถูกต้องในการจำแนกข้อมูลทดสอบที่ได้จากอัลกอริทึม ID3 และค่าดังกล่าวที่ได้จากอัลกอริทึมที่ปรับปรุงแตกต่างกัน

โดยหากค่า *p-value* ที่ได้จากการทดสอบสถิติมีค่าน้อยกว่าหรือเท่ากับ 0.05 จะปฏิเสธสมมติฐานหลัก (H_0) และยอมรับสมมติฐานรอง (H_1) นั่นคือค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และค่าความถูกต้องในการจำแนกข้อมูลทดสอบที่ได้จากทั้งสองอัลกอริทึมมีค่าแตกต่างกันอย่างมีนัยสำคัญทางสถิติ

2.4 การทดสอบผลของข้อมูลที่ผิดปกติต่อต้นไม้ตัดสินใจเชิงความหมาย

การทดลองนี้มีวัตถุประสงค์เพื่อตรวจสอบผลกระทบของข้อมูลที่ผิดปกติที่มีต่ออัลกอริทึมต้นไม้ตัดสินใจที่ได้รับการปรับปรุงด้วยองค์ความรู้ในออนโทโลยี โดยทำการเปรียบเทียบผลของการจำแนกข้อมูลด้วยอัลกอริทึมที่ได้รับการปรับปรุงกับผลของการจำแนกข้อมูลด้วยอัลกอริทึม ID3 ซึ่งจะทำการทดสอบในชุดข้อมูลที่มีจำนวนข้อมูลที่ผิดปกติที่แตกต่างกัน ซึ่งประกอบไปด้วยชุดข้อมูลที่ไม่มีข้อมูลผิดปกติ ชุดข้อมูลที่มีข้อมูลผิดปกติร้อยละ 10 ชุดข้อมูลที่มีข้อมูลผิดปกติร้อยละ 20 และชุดข้อมูลที่มีข้อมูลผิดปกติร้อยละ 30

ในการทดลองนี้ข้อมูลที่มีคลาสคำตอบที่ผิดปกติ (Class noise) (Gupta & Gupta, 2019) ซึ่งหมายถึง ในแถวข้อมูลใด ๆ มีการระบุคลาสคำตอบของแถวข้อมูลนั้นไม่ถูกต้องจะถูกนำมาใช้ในการทดสอบผลกระทบที่มีต่ออัลกอริทึม สำหรับการสร้างชุดข้อมูลที่ประกอบไปด้วยข้อมูลที่ผิดปกติจะมีการดำเนินการดังนี้

- สำหรับชุดข้อมูลที่มีคลาสจำนวนสองคลาส ซึ่งได้แก่ ชุดข้อมูลผู้ป่วยโรคหัวใจ ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออก คลาสคำตอบของแถวข้อมูลจะถูกเปลี่ยนเป็นคลาสที่ตรงข้าม เช่น ในชุดข้อมูลผู้ป่วยโรคหัวใจหากแถวข้อมูลใดระบุคลาสคำตอบเป็น “ผู้ป่วยโรคหัวใจ” ก็จะถูกเปลี่ยนเป็น “ไม่ใช่ผู้ป่วยโรคหัวใจ” เป็นต้น

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

- สำหรับชุดข้อมูลที่มีคลาสคำตอบมากกว่าสองคลาส ซึ่งได้แก่ ชุดข้อมูลการเกิดโรคของถั่วเหลือง คลาสคำตอบของแถวข้อมูลจะถูกเปลี่ยนเป็นคลาสคำตอบอื่นที่มีจำนวนแถวข้อมูลในคลาสใกล้เคียงกัน

2.5 การทดสอบปรับพารามิเตอร์ที่เหมาะสมสำหรับเพิ่มประสิทธิภาพการจำแนกข้อมูล

การทดลองนี้มีวัตถุประสงค์เพื่อตรวจสอบประสิทธิภาพที่ดีที่สุดจากการวิเคราะห์ข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจที่มีการประยุกต์ใช้องค์ความรู้ในออนโทโลยี ด้วยการปรับปรุงค่าความสูงของต้นไม้ตัดสินใจด้วยค่าที่เหมาะสม ซึ่งการค้นหาค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสมจะใช้เทคนิคการค้นหาแบบกริด (Grid search) โดยกำหนดความสูงของต้นไม้ตัดสินใจให้มีค่าระหว่าง 1 ถึง ค่าความสูงของต้นไม้ตัดสินใจที่สูงที่สุดสำหรับแต่ละชุดข้อมูล

2.6 การพิจารณาโครงสร้างของต้นไม้ตัดสินใจเชิงความหมาย

สำหรับการทดลองนี้มีวัตถุประสงค์เพื่อสำรวจโครงสร้างของต้นไม้ตัดสินใจที่เปลี่ยนแปลงไปเมื่อมีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีในการปรับปรุงค่าเกนสารสนเทศ โดยจะทำการเปรียบเทียบกับโครงสร้างของต้นไม้ตัดสินใจที่สร้างอัลกอริทึม ID3 ซึ่งมีการใช้ค่าเกนสารสนเทศเป็นเกณฑ์ในการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับสร้างต้นไม้ตัดสินใจ

2.7 การประมาณค่าระดับความสำคัญเมื่อไม่ปรากฏองค์ความรู้ในออนโทโลยี

การทดลองนี้มีวัตถุประสงค์เพื่อทดสอบการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อแอตทริบิวต์นั้นไม่ปรากฏเป็นองค์ความรู้ในออนโทโลยี ในการทดลองนี้จะทำการประมาณค่าระดับความสำคัญของแอตทริบิวต์ด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด (k-NN) โดยทำการสุ่มแอตทริบิวต์ในชุดข้อมูลให้เป็นแอตทริบิวต์ที่ไม่ปรากฏแนวความคิดในออนโทโลยีจำนวน 3 ชุด คือ

- ชุดข้อมูลที่มีจำนวนแอตทริบิวต์ซึ่งไม่ปรากฏเป็นแนวความคิดในออนโทโลยีจำนวนร้อยละ 10 ของแอตทริบิวต์ในชุดข้อมูล
- ชุดข้อมูลที่มีจำนวนแอตทริบิวต์ซึ่งไม่ปรากฏเป็นแนวความคิดในออนโทโลยีจำนวนร้อยละ 20 ของแอตทริบิวต์ในชุดข้อมูล
- ชุดข้อมูลที่มีจำนวนแอตทริบิวต์ซึ่งไม่ปรากฏเป็นแนวความคิดในออนโทโลยีจำนวนร้อยละ 30 ของแอตทริบิวต์ในชุดข้อมูล

ทำการประมาณค่าระดับความสำคัญของแอตทริบิวต์นั้น ๆ และนำค่าระดับความสำคัญที่ได้ไปใช้ในการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเพื่อทดสอบประสิทธิภาพในการจำแนกข้อมูลเมื่อมีการประมาณค่าระดับความสำคัญของแอตทริบิวต์

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

2.8 การเปรียบเทียบประสิทธิภาพของต้นไม้ตัดสินใจเชิงความหมายกับอัลกอริทึมอื่น ๆ

การทดลองนี้วัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจที่มีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีกับประสิทธิภาพในการจำแนกข้อมูลด้วยอัลกอริทึมในกลุ่มของอัลกอริทึมต้นไม้ตัดสินใจอื่น ๆ จำนวน 3 อัลกอริทึม ได้แก่ อัลกอริทึม C4.5 อัลกอริทึม Classification and Regression Tree (CART) และ อัลกอริทึม Mutual Information Decision Tree (MIDT) (Fang et al., 2017) ในการจำแนกข้อมูลจะทำการกำหนดความสูงของต้นไม้ตัดสินใจที่เหมาะสมสำหรับแต่ละอัลกอริทึมเพื่อให้ได้ค่าความถูกต้องในการจำแนกข้อมูลที่มีประสิทธิภาพมากที่สุด ซึ่งในการทดลองนี้จะทำการจำแนกข้อมูลจำนวน 30 ครั้ง โดยทำการสุ่มข้อมูลเพื่อแบ่งข้อมูลออกเป็นข้อมูลสำหรับการเรียนรู้จำนวนร้อยละ 70 จากข้อมูลทั้งหมด และข้อมูลสำหรับการทดสอบจำนวนร้อยละ 30 เช่นเดียวกับการทดลองอื่น ๆ

บทสรุป

ในบทนี้ได้กล่าวถึงข้อมูล เครื่องมือ และขั้นตอนต่าง ๆ ที่เกี่ยวข้องในการดำเนินการวิจัยเพื่อปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยี โดยทำการทดลองในชุดข้อมูลจำนวน 4 ชุด ซึ่งประกอบไปด้วย ชุดข้อมูลการเกิดโรคของถั่วเหลือง ชุดข้อมูลผู้ป่วยโรคหัวใจ ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และ ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

การทดลองเพื่อปรับปรุงประสิทธิภาพในการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจนั้นแบ่งออกเป็น 2 ส่วน ได้แก่ การเตรียมข้อมูล และ การสร้างแบบจำลอง โดยทั้งสองขั้นตอนมีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีเพื่อช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูล สำหรับขั้นตอนการเตรียมข้อมูลนั้นออนโทโลยีจะถูกนำไปใช้ในกระบวนการแปลงข้อมูลเพื่อลดจำนวนค่าข้อมูลที่อัลกอริทึมใช้ในการสร้างต้นไม้ตัดสินใจและส่งผลให้ต้นไม้ตัดสินใจมีความซับซ้อนลดลง ในส่วนของการสร้างแบบจำลองนั้นองค์ความรู้ในออนโทโลยีจะถูกนำมาประยุกต์ใช้ในรูปแบบของค่าระดับความสำคัญของแอตทริบิวต์ในการปรับปรุงเกนสารสนเทศซึ่งเป็นเกณฑ์ในการพิจารณาแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจเพื่อลดปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีข้อมูลหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจ โดยแอตทริบิวต์ที่มีค่าเกนสารสนเทศน้อยแต่เป็นแอตทริบิวต์ที่มีความสำคัญในศาสตร์นั้น ๆ มีโอกาสถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากขึ้น

ในบทถัดไปจะแสดงถึงผลการทดลองของการเตรียมข้อมูลและการประยุกต์ใช้ออนโทโลยีในการแปลงข้อมูลสำหรับการสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจ

4913881B3 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

บทที่ 4
การประยุกต์ใช้ออนโทโลยีในการเตรียมข้อมูลสำหรับเทคนิคต้นไม้ตัดสินใจ

บทนี้จะนำเสนอผลการทดลองในขั้นตอนการเตรียมข้อมูลด้วยการประยุกต์ใช้ออนโทโลยีเพื่อปรับปรุงข้อมูลและนำข้อมูลที่ได้ไปใช้ในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจ โดยมีรายละเอียดผลการดำเนินงานดังนี้

- ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูล
- ผลการอ้างอิงแนวความคิดพื้นฐาน (Abstract data) จากออนโทโลยี
- ผลการปรับปรุงข้อมูลด้วยแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยี
- ผลการจำแนกข้อมูลที่มีการนำแนวความคิดพื้นฐานมาใช้งาน
- สรุปผลการวิจัย

ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูล

การทดลองนี้มีวัตถุประสงค์เพื่อพิจารณาความสัมพันธ์ระหว่างแอตทริบิวต์ในชุดข้อมูลและคลาสคำตอบที่ต้องการจำแนก และคัดเลือกเฉพาะแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสคำตอบเท่านั้นไปใช้ในขั้นตอนต่อไป โดยการนำแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสคำตอบไปใช้ในการสร้างแบบจำลองจะทำให้อัลกอริทึมเรียนรู้ข้อมูลเหล่านั้นและสร้างต้นไม้ตัดสินใจที่ผิดปกติ ซึ่งส่งผลต่อความความถูกต้องในการจำแนกข้อมูล รวมถึงต้นไม้ตัดสินใจที่ได้มีความซับซ้อนอีกด้วย

ในการทดลองนี้จะใช้สถิติไคสแควร์ (χ^2) และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียล (r_{pb}) ในการพิจารณาความสัมพันธ์ระหว่างข้อมูลในแต่ละชุดข้อมูล หาก *p-value* ของค่าสถิติที่ทำการทดสอบในแอตทริบิวต์ใดมีค่าน้อยกว่าหรือเท่ากับ 0.05 แอตทริบิวต์นั้นจะเป็นแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสคำตอบในชุดข้อมูล ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลสามารถแสดงรายละเอียดได้ดังนี้

- การพิจารณาความสัมพันธ์ระหว่างข้อมูลของชุดข้อมูลการเกิดโรคของถั่วเหลือง

ชุดข้อมูลนี้จะใช้สถิติไคสแควร์ในพิจารณาความสัมพันธ์ระหว่างข้อมูล มีผลการทดลองดังตาราง 7

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 7 ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์สำหรับชุดข้อมูลการเกิดโรคของถั่วเหลือง

แอตทริบิวต์	χ^2	*p-value*	แอตทริบิวต์	χ^2	*p-value*
date	188.18**	< 0.001	canker-lesion	374.16**	< 0.001
plant-stand	70.76**	< 0.001	fruiting bodies	148.06**	< 0.001
precip	149.78**	< 0.001	external decay	89.03**	< 0.001
temp	176.80**	< 0.001	mycelium	79.15**	< 0.001
area-damage	121.63**	< 0.001	int-discolor	274.74**	< 0.001
severity	198.04**	< 0.001	sclerotia	118.01**	< 0.001
seed-tmt	12.66**	0.049	fruit-pods	595.80**	< 0.001
plant-growth	256.60**	< 0.001	fruit-spots	568.06**	< 0.001
leaves	149.96**	< 0.001	seed	41.65**	< 0.001
leafspots-halo	416.37**	< 0.001	mold-growth	39.98**	< 0.001
leafspots-marg	415.72**	< 0.001	seed-discolor	49.24**	< 0.001
leafspot-size	431.18**	< 0.001	seed-size	85.72**	< 0.001
leaf-shread	58.19**	< 0.001	shriveling	122.89**	< 0.001
leaf-malf	10.97**	0.012	*hail*	*3.02*	*0.389*
leaf-mild	21.65**	< 0.001	*crop-hist*	*4.39*	*0.884*
stem	227.02**	< 0.001	*germination*	*3.85*	*0.697*
lodging	94.14**	< 0.001	*roots*	*4.02*	*0.260*
stem-cankers	594.05**	< 0.001			

จากค่าสถิติไคสแควร์ที่แสดงความสัมพันธ์ระหว่างแอตทริบิวต์ที่เกี่ยวข้องกับโรคของถั่วเหลืองและโรคที่ต้องการจำแนกในตาราง 7 พบว่ามีแอตทริบิวต์ที่มีค่า *p-value* มากกว่า 0.05 จำนวน 4 แอตทริบิวต์ ได้แก่ hail crop-hist germination และ roots ซึ่งหมายถึงแอตทริบิวต์เหล่านี้ไม่มีความสัมพันธ์กับโรคของถั่วเหลืองที่ต้องการจำแนกอย่างมีนัยสำคัญทางสถิติ ส่งผลให้ในชุดข้อมูลการเกิดโรคของถั่วเหลืองมีแอตทริบิวต์ที่นำไปใช้ในขั้นตอนต่อไปจำนวนรวมทั้งสิ้น 31 แอตทริบิวต์

- การพิจารณาความสัมพันธ์ระหว่างข้อมูลของชุดข้อมูลผู้ป่วยโรคหัวใจ

ชุดข้อมูลนี้จะใช้สถิติไคสแควร์และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลในการพิจารณาความสัมพันธ์ระหว่างข้อมูล ซึ่งสามารถแสดงผลลัพธ์ได้ดังตาราง 8

4918896183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 8 ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลสำหรับชุดข้อมูลผู้ป่วยโรคหัวใจ

แอตทริบิวต์	χ^2	*p*-value	แอตทริบิวต์	r_{pb}	*p*-value
Sex	23.03**	< 0.001	Age	0.23**	< 0.001
Cp	77.28**	< 0.001	Threstbps	0.15**	0.008
Fbs	***0.003***	***0.956***	***Chol***	***0.80***	***0.168***
Restecg	9.58**	0.008	Thalach	-0.42**	< 0.001
Exang	52.73**	< 0.001	Oldpeak	0.42**	< 0.001
Slope	43.47**	< 0.001	Ca	0.46**	< 0.001
Thal	82.46**	< 0.001			

จากตาราง 8 พบว่าชุดข้อมูลผู้ป่วยโรคหัวใจมีแอตทริบิวต์จำนวน 2 แอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสคำตอบอย่างมีนัยสำคัญทางสถิติ นั่นคือมีค่า *p-value* มากกว่า 0.05 ซึ่งประกอบไปด้วยแอตทริบิวต์ Fbs และ Chol ทำให้ชุดข้อมูลนี้มีแอตทริบิวต์ที่สามารถนำไปใช้ในขั้นตอนต่อไปจำนวน 11 แอตทริบิวต์

- การพิจารณาความสัมพันธ์ระหว่างข้อมูลของชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019

ชุดข้อมูลนี้จะใช้สถิติไคสแควร์ในการพิจารณาความสัมพันธ์ระหว่างข้อมูล ซึ่งมีผลลัพธ์ดังตาราง 9

ตาราง 9 ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์สำหรับชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019

แอตทริบิวต์	χ^2	*p*-value	แอตทริบิวต์	χ^2	*p*-value
gender	316.60**	< 0.001	olfactory disorders	136.90**	< 0.001
Health professional	127.91**	< 0.001	cough	11.01**	0.001
fever	549.51**	< 0.001	coryza	42.75**	< 0.001
sore throat	10.88**	0.001	taste disorders	60.30**	< 0.001
dyspnea	80.34**	< 0.001	headache	5.10**	0.024

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากตาราง 9 จะพบว่าทุกแอตทริบิวต์ในชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 มีค่า *p-value* น้อยกว่า 0.05 ซึ่งหมายถึงทุกแอตทริบิวต์ในชุดข้อมูลมีความสัมพันธ์กับคลาสคำตอบอย่างมีนัยสำคัญทางสถิติ ดังนั้นสำหรับชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 จะมีแอตทริบิวต์ที่สามารถนำไปใช้ในขั้นตอนต่อไปได้ทั้งหมด 10 แอตทริบิวต์

- การพิจารณาความสัมพันธ์ระหว่างข้อมูลของชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

ชุดข้อมูลนี้จะใช้สถิติไคสแควร์และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลในการพิจารณาความสัมพันธ์ระหว่างข้อมูล โดยมีผลลัพธ์ดังแสดงในตาราง 10

ตาราง 10 ผลการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลสำหรับชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

แอตทริบิวต์	χ^2	*p*-value	แอตทริบิวต์	r_{pb}	*p*-value
gender	20.51**	< 0.001	age	-0.097**	0.001
fever	153.43**	< 0.001			
rash	589.62**	< 0.001			
pruritus	463.07**	< 0.001			
myalgia	62.61**	< 0.001			
arthralgia	77.42**	< 0.001			
arthritis	9.12**	0.003			
conjunctivitis	74.20**	< 0.001			
headache	195.39**	< 0.001			
lymphadenopathy	13.40**	< 0.001			
bleeding	*1.13*	*0.288*			
neurological signs	*0.30*	*0.584*			

จากตาราง 10 พบว่าในการพิจารณาความสัมพันธ์ระหว่างข้อมูลในชุดข้อมูลผู้ป่วยโรคไข้เลือดออกมีแอตทริบิวต์จำนวน 2 แอตทริบิวต์ที่ค่า *p-value* ของสถิติที่ทำการทดสอบมีค่ามากกว่า 0.05 ได้แก่ แอตทริบิวต์ bleeding และ แอตทริบิวต์ neurological signs ซึ่งหมายถึงแอตทริบิวต์ดังกล่าวไม่มีความสัมพันธ์กับคลาสคำตอบอย่างมีนัยสำคัญทางสถิติ ดังนั้นชุดข้อมูลผู้ป่วยโรค

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ไข้เลือดออกจึงมีแอตทริบิวต์ที่สามารถนำไปใช้ในการดำเนินงานในขั้นตอนต่อไปได้รวมทั้งสิ้น 10 แอตทริบิวต์

จากการทดลองนี้จะพบว่าการพิจารณาความสัมพันธ์ระหว่างแอตทริบิวต์และคลาสคำตอบของชุดข้อมูลสามารถช่วยลดจำนวนแอตทริบิวต์ที่นำไปใช้ในขั้นตอนการสร้างแบบจำลองการจำแนกข้อมูลได้ อย่างไรก็ตามการพิจารณาความสัมพันธ์ระหว่างแอตทริบิวต์ที่เป็นปัจจัยที่ใช้ในการจำแนกข้อมูลด้วยวิธีการทางสถิติ หรือการพิจารณาความสัมพันธ์ของแต่ละแอตทริบิวต์จากองค์ความรู้ในออนโทโลยี เป็นอีกแนวทางหนึ่งที่สามารถช่วยในการค้นหาแอตทริบิวต์ที่อาจมีความซ้ำซ้อนกัน ส่งผลให้สามารถลดจำนวนแอตทริบิวต์ที่ใช้สำหรับการสร้างแบบจำลองได้

ผลการอ้างอิงแนวความคิดพื้นฐานจากออนโทโลยี

การทดลองนี้มีวัตถุประสงค์เพื่อประยุกต์ใช้องค์ความรู้ในออนโทโลยีในการปรับปรุงชุดข้อมูลที่ทำการศึกษาโดยการนำค่าแนวความคิดพื้นฐาน (Abstract data) ที่มีความสัมพันธ์กับข้อมูลในแต่ละแอตทริบิวต์มาใช้แทนข้อมูลเดิม การค้นหาแนวความคิดพื้นฐานในออนโทโลยีที่มีความสัมพันธ์กับข้อมูลในชุดข้อมูลนั้นจะดำเนินการโดยใช้อัลกอริทึมที่พัฒนาขึ้นด้วยภาษาไพธอนและแพ็คเกจ Owlready2 เพื่อทำการจับคู่ข้อมูลในชุดข้อมูลที่ตรงกับตัวอย่างข้อมูล (Instance) ในออนโทโลยี ดังแสดงในภาพ 14 และนำค่าแนวความคิดพื้นฐานที่ได้ไปแทนที่ค่าข้อมูลเดิมที่มีความสัมพันธ์ด้วยอัลกอริทึมในภาพ 15

ผลการทดลองพบว่ามีเพียงออนโทโลยีโรคของถั่วเหลืองที่สามารถอ้างอิงแนวความคิดพื้นฐานที่มีความสัมพันธ์กับข้อมูลในชุดข้อมูล ผลของการอ้างอิงแนวความคิดพื้นฐานแสดงดังตาราง 11

ตาราง 11 แนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีโรคของถั่วเหลือง

แอตทริบิวต์	ค่าข้อมูลเดิม	ค่าแนวความคิดพื้นฐาน
stem-canker	below-soil, above-soil	near ground
	above-sec-nde	canker on stem
fruit-pods	diseased, few-present	presented symptom on fruit pod
fruit spots	colored, brown-w/blk-specks	colored fruit spots

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากตาราง 11 พบว่ามีแอตทริบิวต์จำนวน 3 แอตทริบิวต์ ได้แก่ stem-canker fruit-pods และ fruit spots ที่ค่าข้อมูลของแอตทริบิวต์สามารถจับคู่กับข้อมูลในออนโทโลยีและสามารถอ้างอิงไปยังแนวความคิดพื้นฐานที่เกี่ยวข้องได้ เช่น แอตทริบิวต์ stem-canker ซึ่งมีค่าข้อมูล below-soil และ above-soil ที่สามารถอ้างอิงไปยังแนวความคิดพื้นฐาน near ground ในขณะที่ค่าข้อมูล above-sec-nde สามารถอ้างอิงไปยังแนวความคิด canker on stem เป็นต้น

ผลการปรับปรุงข้อมูลด้วยแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยี

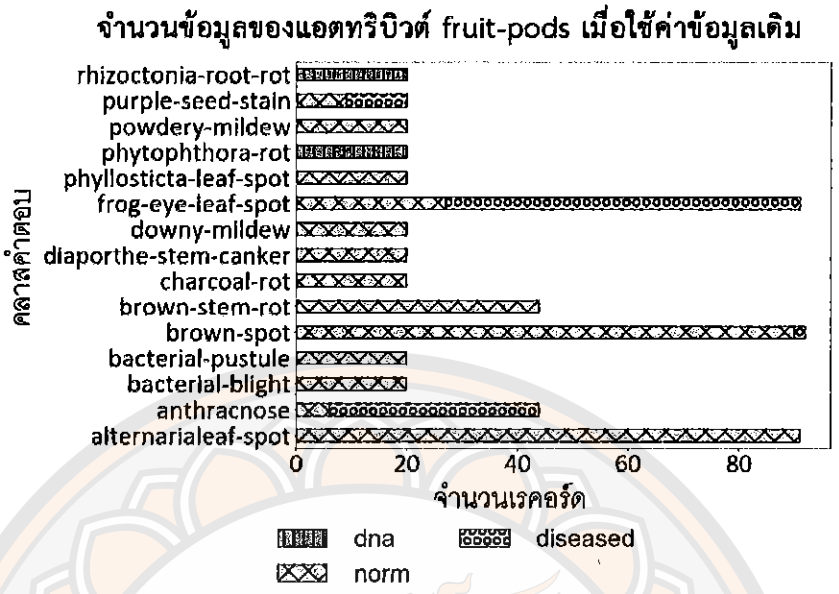
การทดลองนี้มีวัตถุประสงค์เพื่อสำรวจการเปลี่ยนแปลงที่เกิดขึ้นกับชุดข้อมูลเมื่อมีการนำค่าข้อมูลพื้นฐานที่อ้างอิงได้จากออนโทโลยีมาใช้ในการปรับปรุงชุดข้อมูล โดยสามารถแสดงการเปลี่ยนแปลงของชุดข้อมูลได้ดังภาพ 23 ถึง ภาพ 25

ภาพ 23 พบว่าสำหรับแอตทริบิวต์ stem-canker เมื่อทำการปรับปรุงค่าข้อมูลในแอตทริบิวต์ด้วยค่าแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยี จะทำให้จำนวนค่าข้อมูลของแอตทริบิวต์ stem-canker เดิมซึ่งมีจำนวน 4 ค่า ได้แก่ absent above-sec-nde above-soil และ below-soil ลดลงเหลือจำนวน 3 ค่า ได้แก่ absent, canker on stem และ near ground ซึ่งส่งผลให้เกิดการเปลี่ยนแปลงกับคลาสที่ต้องจำแนกเพียง 1 คลาส คือ คลาส phytophthora-rot ที่มีค่าข้อมูลที่เกี่ยวข้องจากจำนวน 2 ค่า คือ above-soil และ below-soil เหลือเพียง 1ค่า คือ near ground ซึ่งเป็นค่าแนวความคิดพื้นฐานที่มีความสัมพันธ์กับข้อมูลเดิม

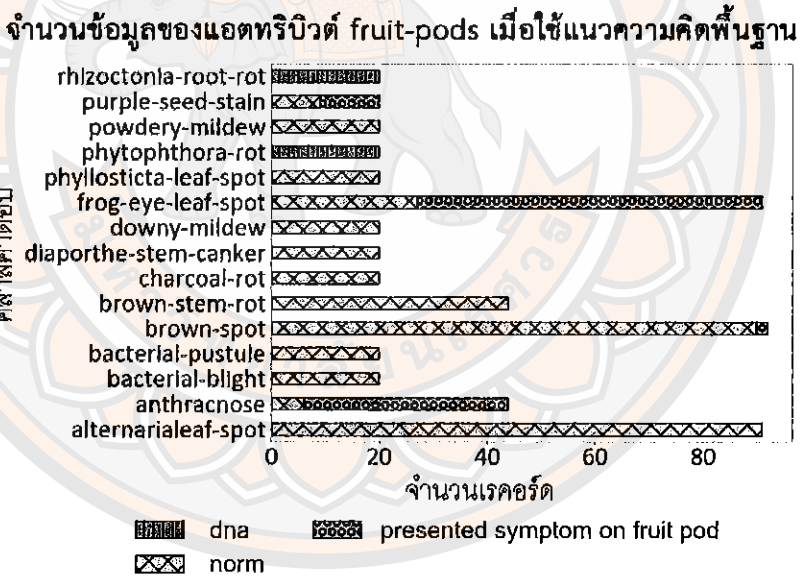
4918891B3 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ภาพ 23 จำนวนข้อมูลในแอตทริบิวต์ stem-canker โดย (ก) เมื่อใช้ข้อมูลเดิม และ(ข) เมื่อใช้ค่าแนวความคิดพื้นฐาน

4918881831 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39



(ก)



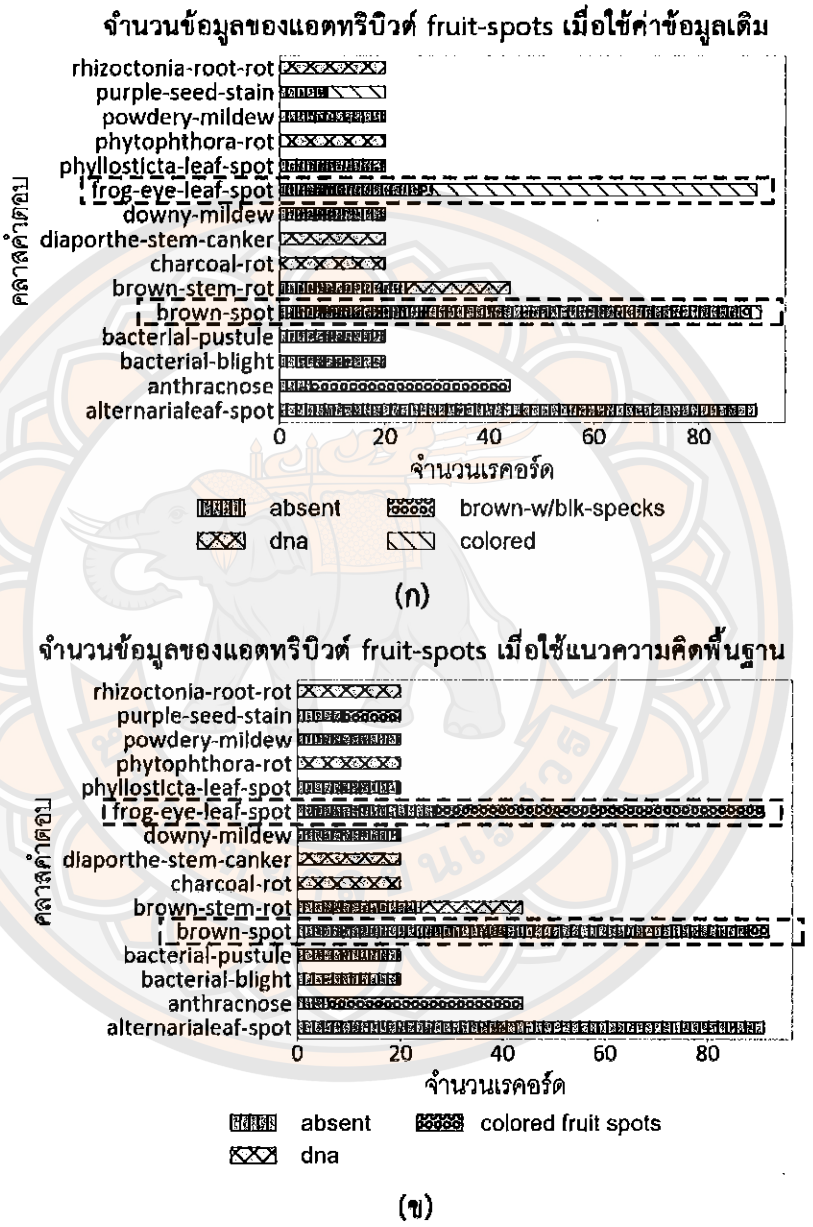
(ข)

ภาพ 24 จำนวนข้อมูลในแอตทริบิวต์ fruit-pods โดย (ก) เมื่อใช้ข้อมูลเดิม และ(ข) เมื่อใช้ค่าแนวความคิดพื้นฐาน

จากภาพ 24 พบว่าแอตทริบิวต์ fruit-pods นั้น เมื่อนำค่าแนวความคิดพื้นฐานที่สามารถอ้างอิงได้จากออนโทโลยีไปแทนที่ค่าข้อมูลเดิมจะไม่ทำให้เกิดการเปลี่ยนแปลงจำนวนค่าข้อมูลที่เกี่ยวข้องในแต่ละคลาส เนื่องจากในชุดข้อมูลเดิมมีเพียงค่า disease เพียงค่าเดียวที่สามารถอ้างอิงไปยังแนวความคิดพื้นฐานที่เกี่ยวข้องในออนโทโลยีได้ นั่นคือ ค่า presented symptom on fruit pod

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

สำหรับการเปลี่ยนแปลงของชุดข้อมูลเมื่อทำการแทนที่ค่าข้อมูลเดิมของแอตทริบิวต์ fruit-spots ด้วยค่าแนวคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีนั้นสามารถแสดงดังภาพ 25



ภาพ 25 จำนวนข้อมูลในแอตทริบิวต์ fruit-spots โดย (ก) เมื่อใช้ข้อมูลเดิม และ(ข) เมื่อใช้ค่าแนวความคิดพื้นฐาน

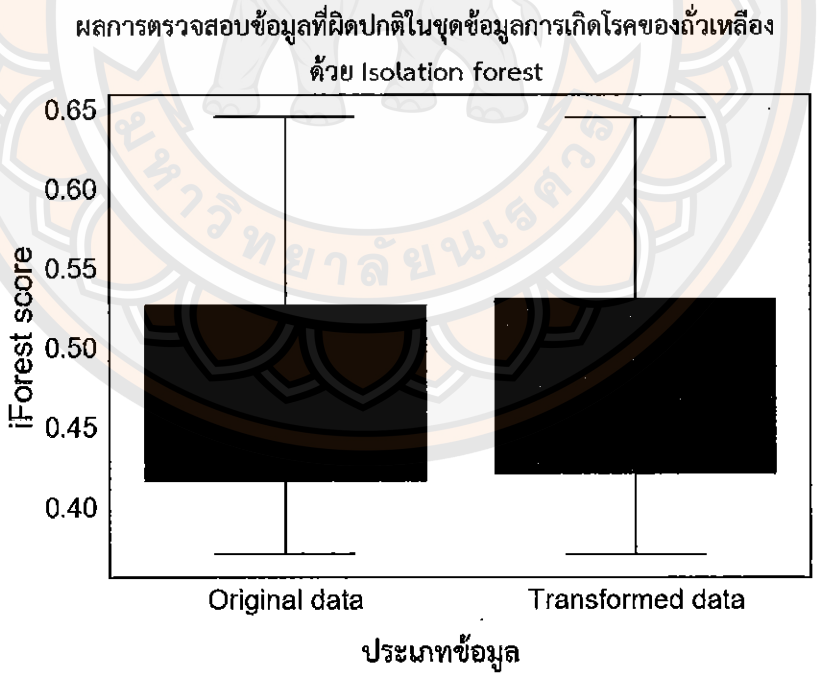
จากภาพ 25 จะพบว่าเมื่อนำค่าแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีไปแทนที่ค่าข้อมูลเดิมของแอตทริบิวต์ fruit-spots จะทำให้แอตทริบิวต์นี้ซึ่งมีจำนวนค่าข้อมูลที่เกี่ยวข้องจำนวน

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

4 ค่า ประกอบไปด้วย absent dna brown-w/blk-specks และ colored ลดลงเหลือ 3 ค่า คือ absent dna และ colored fruit spots โดยค่า colored fruit spots คือค่าแนวความคิดพื้นฐานที่มีความสัมพันธ์กับค่า brown-w/blk-specks และ colored สำหรับการเปลี่ยนแปลงของจำนวนค่าข้อมูลในแต่ละคลาสเมื่อทำการแทนที่ค่าแนวความคิดพื้นฐานของแอตทริบิวต์ fruit-spots นั้น พบว่าเกิดการเปลี่ยนแปลงกับคลาสจำนวน 2 คลาส คือ fog-eye-leaf-spot และ brown-spot โดยจำนวนค่าข้อมูลที่เกี่ยวข้องกับคลาสเหล่านี้จะมีจำนวนลดลงจากทั้งหมด 3 ค่า เป็น 2 ค่า

นอกจากการพิจารณาการเปลี่ยนแปลงจำนวนค่าข้อมูลของแอตทริบิวต์ในแต่ละคลาสเมื่อมีการนำแนวความคิดพื้นฐานที่มีความสัมพันธ์มาปรับปรุงข้อมูลแล้ว การทดลองนี้ยังทำการสำรวจข้อมูลที่ผิดปกติ (outlier) ที่อาจเกิดขึ้นจากการแปลงข้อมูลด้วยค่าแนวความคิดพื้นฐานโดยใช้เทคนิค Isolation forest โดยหากค่าคะแนนที่ได้จากเทคนิค Isolation forest มีค่าเข้าใกล้ 1 จะหมายถึงข้อมูลนั้นเป็นข้อมูลที่ผิดปกติ (outlier) (Wang et al., 2021) ซึ่งในการทดลองนี้หาค่าคะแนนของเทคนิค Isolation forest โดยใช้ไลบรารี Scikit Learn (Pedregosa et al., 2011)

ผลการตรวจสอบข้อมูลที่ผิดปกติของชุดข้อมูลการเกิดโรคของถั่วเหลืองสามารถแสดงดังภาพ 26



ภาพ 26 ผลการตรวจสอบข้อมูลที่ผิดปกติด้วย Isolation forest

จากภาพ 26 จะพบว่าทั้ง 2 ชุดข้อมูล คือ ชุดข้อมูลการเกิดโรคของถั่วเหลือง (original data) และ ชุดข้อมูลการเกิดโรคของถั่วเหลืองที่ปรับปรุงด้วยแนวความคิดพื้นฐานที่อ้างอิงจากออนโทโลยี (transformed data) มีค่าคะแนนที่ได้จากเทคนิค Isolation forest ใกล้เคียงกัน โดยค่า

4918881183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

คะแนนที่สูงที่สุดของชุดข้อมูลเดิมมีค่าเท่ากับ 0.645 จึงสามารถสรุปได้ว่าชุดข้อมูลเดิมนี้ไม่ปรากฏข้อมูลที่ผิดปกติ ในขณะเดียวกันชุดข้อมูลที่ปรับปรุงด้วยค่าแนวความคิดพื้นฐานมีค่าคะแนนที่ได้จากเทคนิค Isolation forest ที่มากที่สุดเท่ากับ 0.645 เช่นกัน ซึ่งหมายถึงไม่ปรากฏข้อมูลที่ผิดปกติในชุดข้อมูลนี้

ผลจากการสำรวจการเปลี่ยนแปลงของข้อมูลที่เกิดขึ้นนั้นจะพบว่า การแทนที่ข้อมูลด้วยแนวความคิดพื้นฐานที่มีความสัมพันธ์ในออนโทโลยีช่วยลดขนาดของชุดข้อมูลลงด้วยการลดจำนวนของค่าข้อมูลในแอตทริบิวต์ที่เกี่ยวข้องโดยไม่ทำให้เกิดข้อมูลที่ผิดปกติในชุดข้อมูล อย่างไรก็ตามเนื่องจากแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีโรคของถั่วเหลืองนี้มีความแตกต่างจากข้อมูลเดิมเพียง 1 ระดับ จึงทำให้เกิดการเปลี่ยนแปลงในชุดข้อมูลเพียงเล็กน้อย ซึ่งหากข้อมูลแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีมีความแตกต่างจากข้อมูลอื่นมากกว่า 1 ระดับ จำเป็นต้องพิจารณาถึงระดับของแนวความคิดพื้นฐานที่เหมาะสมที่จะนำมาใช้ในการปรับปรุงข้อมูลเพื่อให้การจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจมีประสิทธิภาพ

ผลการจำแนกข้อมูลที่มีการนำแนวความคิดพื้นฐานมาใช้งาน

การทดลองนี้มีวัตถุประสงค์เพื่อตรวจสอบผลการปรับปรุงข้อมูลด้วยองค์ความรู้ในออนโทโลยีที่มีต่อประสิทธิภาพของการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ ในการทดลองนี้จะใช้ชุดข้อมูลการเกิดโรคของถั่วเหลืองในการทดลองเนื่องจากเป็นชุดข้อมูลเดียวที่สามารถอ้างอิงแนวความคิดพื้นฐานจากออนโทโลยี โดยทำการแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลสำหรับการเรียนรู้ (training data) จำนวนร้อยละ 70 ของจำนวนข้อมูลทั้งหมด และข้อมูลสำหรับการทดสอบ (test data) จำนวนร้อยละ 30 ของจำนวนข้อมูลทั้งหมด พร้อมทั้งทำการแก้ปัญหาข้อมูลที่ไม่สมดุล (imbalanced data) ด้วยเทคนิค Synthetic Minority Oversampling Technique (SMOTE) หลังจากนั้นจะทำการพิจารณาผลของการนำองค์ความรู้ในออนโทโลยีในการปรับปรุงข้อมูลจากค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึม ID3 ซึ่งผลการทดลองสามารถแสดงดังตาราง 12

ตาราง 12 ผลการจำแนกข้อมูลเมื่อมีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีในการปรับปรุงข้อมูล

เกณฑ์การประเมิน	ชุดข้อมูลเดิม	ชุดข้อมูลที่ปรับปรุงด้วยออนโทโลยี	ค่าที่เปลี่ยนแปลง
ค่าความถูกต้อง (Accuracy)	87.53%	88.17%	+0.64%
ความสูงของต้นไม้ตัดสินใจ	8	8	-
จำนวนโหนด	105	105	-

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากตาราง 12 จะพบว่าชุดข้อมูลที่ทำการปรับปรุงข้อมูลด้วยแนวความคิดพื้นฐานในออนโทโลยีทำให้ค่าความถูกต้องในการจำแนกข้อมูลเทคนิคต้นไม้ตัดสินใจเพิ่มขึ้น 0.64% โดยเมื่อทำการจำแนกข้อมูลในชุดข้อมูลดั้งเดิมค่าความถูกต้องในการจำแนกข้อมูลมีค่าเป็น 87.53% และเมื่อทำการปรับปรุงชุดข้อมูลด้วยองค์ความรู้ในออนโทโลยีค่าความถูกต้องในการจำแนกข้อมูลจะมีค่าเป็น 88.17% เมื่อพิจารณาความสูงของต้นไม้ตัดสินใจและจำนวนโหนดภายในต้นไม้ตัดสินใจซึ่งเป็นเกณฑ์ที่ใช้ในการพิจารณาความซับซ้อนของต้นไม้ตัดสินใจพบว่าชุดข้อมูลทั้งสองแบบนี้ให้ค่าความสูงและจำนวนโหนดของต้นไม้ตัดสินใจไม่แตกต่างกัน

ซึ่งการที่ค่าความสูงและจำนวนโหนดของต้นไม้ตัดสินใจของชุดข้อมูลที่ปรับปรุงด้วยแนวความคิดพื้นฐานที่ได้จากออนโทโลยีมีค่าไม่แตกต่างจากค่าความสูงและจำนวนโหนดเมื่อทำการสร้างต้นไม้ตัดสินใจด้วยชุดข้อมูลเดิมนั้น เนื่องจากชุดข้อมูลที่ทำการปรับปรุงเกิดการเปลี่ยนแปลงจำนวนค่าของข้อมูลเพียงเล็กน้อย โดยเกิดการเปลี่ยนแปลงค่าของข้อมูลในแอตทริบิวต์ stem-canker และแอตทริบิวต์ fruit-spots ดังแสดงในภาพ 23 และภาพ 25 ซึ่งการลดลงของจำนวนค่าข้อมูลของแต่ละแอตทริบิวต์อาจไม่เพียงพอที่จะส่งผลต่อการเปลี่ยนแปลงความสูงและจำนวนโหนดที่ใช้ในการสร้างต้นไม้ตัดสินใจ แต่อย่างไรก็ตามการเปลี่ยนแปลงจำนวนค่าข้อมูลในแต่ละแอตทริบิวต์ส่งผลให้เกิดการปะปนกันของค่าข้อมูลในแต่ละคลาสลดลง เช่น การแทนค่าข้อมูลเดิมในแอตทริบิวต์ fruit-spots จะทำให้คลาส fog-eye-leaf-spot และ brown-spot มีการปะปนกันของข้อมูลภายในคลาสจาก 3 ค่า เหลือเพียง 2 ค่า ซึ่งส่งผลให้เทคนิคต้นไม้ตัดสินใจสามารถจำแนกข้อมูลได้ดีขึ้น

ทำการปรับปรุงประสิทธิภาพของการจำแนกข้อมูลโดยการค้นหาความสูงของต้นไม้ตัดสินใจที่เหมาะสม (optimal depth) ที่จะทำให้มีค่าความถูกต้องในการจำแนกข้อมูลสูงที่สุด เนื่องจากค่าความสูงของต้นไม้ตัดสินใจเป็นพารามิเตอร์หนึ่งที่ส่งผลต่อประสิทธิภาพของเทคนิคต้นไม้ตัดสินใจ (Wen & Xu, 2021) โดยการค้นหาแบบกริด (Grid search) จะถูกนำมาใช้เพื่อค้นหาความสูงของต้นไม้ที่เหมาะสมที่สุด ซึ่งผลการจำแนกข้อมูลโดยการปรับปรุงค่าความสูงของต้นไม้ตัดสินใจสามารถแสดงดังตาราง 13

ตาราง 13 ผลการจำแนกข้อมูลเมื่อมีการใช้ค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสม

เกณฑ์การประเมิน	ชุดข้อมูลเดิม	ชุดข้อมูลที่ปรับปรุงด้วยออนโทโลยี	ค่าที่เปลี่ยนแปลง
ค่าความถูกต้อง (Accuracy)	88.19%	88.80%	+0.61%
ความสูงของต้นไม้ตัดสินใจที่เหมาะสม	5	5	-
จำนวนโหนด	90	88	2

49188918 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากตาราง 13 พบว่าค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสมสำหรับชุดข้อมูลเดิมและชุดข้อมูลที่ปรับปรุงด้วยค่าแนวความคิดพื้นฐานในออนโทโลยีมีค่าเท่ากัน โดยมีค่าเท่ากับ 5 เมื่อพิจารณาค่าความถูกต้องในการจำแนกข้อมูลพบว่าเมื่อทำการแนกชุดข้อมูลเดิม ค่าความถูกต้องในการจำแนกข้อมูลจะมีค่าเท่ากับ 88.19% และค่าความถูกต้องในการจำแนกชุดข้อมูลที่มีการปรับปรุงด้วยแนวความคิดพื้นฐานในออนโทโลยีมีค่าเท่ากับ 88.80% โดยมีค่าความถูกต้องในการจำแนกข้อมูลมากกว่าชุดข้อมูลเดิม 0.61% สำหรับจำนวนโหนดของต้นไม้ตัดสินใจนั้น เมื่อทำการจำแนกชุดข้อมูลเดิม ณ ความสูงของต้นไม้ตัดสินใจที่เหมาะสมจะทำให้มีจำนวนโหนดเท่ากับ 90 โหนด ในขณะที่เมื่อทำการจำแนกข้อมูลที่ทำการปรับปรุงด้วยแนวความคิดพื้นฐานในออนโทโลยี ณ ความสูงของต้นไม้ตัดสินใจที่เหมาะสมจะมีจำนวนโหนดเท่ากับ 88 โหนด ซึ่งการลดลงของจำนวนโหนดภายในต้นไม้ตัดสินใจนั้นเกิดจากการนำค่าแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีมาแทนที่ข้อมูลที่มีความสัมพันธ์ ทำให้การปะปนกันของข้อมูลในคลาสคำตอบลดลงและส่งผลอัลกอริทึมหยุดการแตกกิ่งของต้นไม้ตัดสินใจได้เร็วขึ้น

การปรับปรุงประสิทธิภาพการจำแนกข้อมูลด้วยค่าความสูงที่เหมาะสมนั้นจำนวนโหนดของชุดข้อมูลที่ปรับปรุงด้วยแนวความคิดพื้นฐานที่ได้จากออนโทโลยีมีค่าใกล้เคียงกับจำนวนโหนดของต้นไม้ตัดสินใจที่สร้างจากข้อมูลเดิมเช่นเดียวกันกับผลลัพธ์ที่ได้จากการทดลองสร้างต้นไม้ตัดสินใจที่ไม่มีการกำหนดความสูงของต้นไม้ตัดสินใจที่เหมาะสม อย่างไรก็ตามจำนวนโหนดของต้นไม้ตัดสินใจเมื่อมีการใช้แนวความคิดพื้นฐานในชุดข้อมูลจะมีจำนวนโหนดลดลง ซึ่งแสดงให้เห็นว่าการใช้แนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีเพื่อปรับปรุงข้อมูลที่มีความสัมพันธ์นั้นสามารถช่วยลดจำนวนโหนดภายในต้นไม้ตัดสินใจได้ ซึ่งเป็นวิธีการหนึ่งที่ช่วยลดความซับซ้อนของแบบจำลองต้นไม้ตัดสินใจ

สรุปผลการวิจัย

การปรับปรุงข้อมูลเพื่อเพิ่มประสิทธิภาพการจำแนกข้อมูลของต้นไม้ตัดสินใจนั้น ผู้วิจัยได้ประยุกต์ใช้องค์ความรู้ที่อยู่ในออนโทโลยีเพื่อสนับสนุนกระบวนการจัดเตรียมข้อมูล โดยทำการพิจารณาความสัมพันธ์ระหว่างข้อมูลด้วยสถิติไคสแควร์และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลเพื่อนำแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสคำตอบออกจากชุดข้อมูล ซึ่งการใช้เฉพาะแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสคำตอบจะช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ เนื่องจากจำนวนแอตทริบิวต์ที่ใช้ในการสร้างต้นไม้ตัดสินใจนั้นเป็นเกณฑ์หนึ่งที่ใช้ในการพิจารณาความซับซ้อนของเทคนิคต้นไม้ตัดสินใจ (Su & Zhang, 2006)

หลังจากขั้นตอนการพิจารณาความสัมพันธ์ระหว่างข้อมูล องค์ความรู้ในออนโทโลยีจะถูกนำมาใช้ในการพิจารณาค่าแนวความคิดพื้นฐานที่มีความสัมพันธ์กับค่าข้อมูลของแต่ละแอตทริบิวต์ในชุดข้อมูลที่ทำการศึกษา และนำค่าแนวความคิดพื้นฐานที่อ้างอิงได้มาแทนที่ค่าที่มีความสัมพันธ์ใน

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ชุดข้อมูลเดิม ซึ่งการนำค่าแนวความคิดพื้นฐานมาใช้ในการปรับปรุงข้อมูลจะส่งผลให้ชุดข้อมูลมีจำนวนค่าข้อมูลของแอตทริบิวต์ลดลง ส่งผลให้การปะปนกันของค่าข้อมูลในแต่ละคลาสคำตอบลดลง ซึ่งช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลได้ นอกจากนี้การที่ค่าข้อมูลของแอตทริบิวต์มีความสัมพันธ์กับคลาสคำตอบเพียงคลาสเดียวเป็นเงื่อนไขหนึ่งที่ทำให้หยุดการแตกกิ่งของต้นไม้ตัดสินใจ ซึ่งหากมีการปะปนกันของค่าข้อมูลของแอตทริบิวต์ในคลาสคำตอบก็จะส่งผลให้ต้นไม้ตัดสินใจมีความซับซ้อนมาก โดยต้นไม้ตัดสินใจอาจมีความลึกหรือมีจำนวนโหนดภายในต้นไม้ตัดสินใจจำนวนมาก ดังนั้นการนำค่าแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีมาใช้ในการปรับปรุงข้อมูลสามารถช่วยลดจำนวนโหนดหรือความสูงของต้นไม้ตัดสินใจได้หากค่าแนวความคิดพื้นฐานที่นำมาใช้นั้นทำให้เกิดการเปลี่ยนแปลงมากพอในชุดข้อมูล

ถึงแม้ว่าการนำค่าแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีมาใช้ในการปรับปรุงข้อมูลจะสามารถเพิ่มประสิทธิภาพการจำแนกข้อมูลของต้นไม้ตัดสินใจได้ ระดับของค่าแนวความคิดพื้นฐานที่นำมาใช้ในการปรับปรุงข้อมูลนั้นเป็นประเด็นหนึ่งที่ควรให้ความสำคัญหากมีการนำวิธีการนี้ไปใช้งาน เนื่องจากหากนำค่าแนวความคิดพื้นฐานที่มีความแตกต่างจากข้อมูลเดิมมากเกินไป หรือเป็นข้อมูลเป็นแนวความคิดแบบทั่วไป (General concept) ไม่เฉพาะเจาะจงมากเกินไป อาจส่งผลให้สูญเสียข้อมูลที่สำคัญได้

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

บทที่ 5
การปรับปรุงกระบวนการสร้างต้นไม้ตัดสินใจโดยการประยุกต์ใช้ออนโทโลยี

บทนี้จะนำเสนอผลการทดลองในขั้นตอนการสร้างแบบจำลอง (Modeling) โดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยีเพื่อปรับปรุงกระบวนการพิจารณาแอตทริบิวต์สำหรับทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจ ซึ่งช่วยลดปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลาย (Multi-valued bias) โดยจะเรียกอัลกอริทึมต้นไม้ตัดสินใจที่ได้รับการปรับปรุงว่า อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย (Semantic decision tree, SDT) ซึ่งมีรายละเอียดผลการดำเนินงานดังนี้

- ผลการคำนวณค่าระดับความสำคัญจากออนโทโลยี
- ผลการทดสอบประสิทธิภาพการจำแนกข้อมูลของต้นไม้ตัดสินใจเชิงความหมาย
- ผลการทดสอบความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting)
- ผลการทดสอบผลของข้อมูลที่ผิดปกติต่อต้นไม้ตัดสินใจเชิงความหมาย
- ผลการทดสอบปรับพารามิเตอร์ที่เหมาะสม (Optimization) สำหรับเพิ่มประสิทธิภาพการจำแนกข้อมูล
- ผลการพิจารณาโครงสร้างของต้นไม้ตัดสินใจเชิงความหมาย
- ผลการประมาณค่าระดับความสำคัญเมื่อไม่ปรากฏองค์ความรู้ในออนโทโลยี
- ผลการเปรียบเทียบประสิทธิภาพของต้นไม้ตัดสินใจเชิงความหมายกับอัลกอริทึมอื่น ๆ
- สรุปผลการวิจัย

ผลการคำนวณค่าระดับความสำคัญจากออนโทโลยี

การทดลองนี้มีวัตถุประสงค์เพื่อคำนวณค่าระดับความสำคัญของแนวความคิดในออนโทโลยี เพื่อนำไปใช้เป็นค่าระดับความสำคัญของแอตทริบิวต์ที่มีความสัมพันธ์กับแนวความคิดนั้น ๆ เพื่อใช้สำหรับการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจในขั้นตอนต่อไป การหาค่าระดับความสำคัญของแนวความคิดในออนโทโลยีจะดำเนินการประยุกต์ใช้แนวความคิดการสรุปภาพรวมออนโทโลยี (Ontology Summarization) ซึ่งในการทดลองนี้ใช้อัลกอริทึม Weighted Semantic PageRank (Jun et al., 2016) ที่พิจารณาความสัมพันธ์ระหว่างแนวความคิดเพื่อระบุความสำคัญของแต่ละแนวความคิด โดยค่าระดับความสำคัญของแนวความคิดที่มีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูลที่ทำการศึกษาสามารถแสดงดังตาราง 14 ถึง ตาราง 17

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 14 ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคของถั่วเหลืองซึ่งมีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูล

แนวความคิด / แอตทริบิวต์	ค่าระดับความสำคัญ	แนวความคิด / แอตทริบิวต์	ค่าระดับความสำคัญ
leaves	2.21	canker_lesion	0.26
stem	1.87	leaf_shread	0.23
seed	1.09	area_damaged	0.22
precip	0.62	fruiting_bodies	0.21
temp	0.60	seed_discolor	0.20
leafspot_halo	0.55	int_discolor	0.19
leafspot_size	0.52	plant_stand	0.18
leafspots_marg	0.52	severity	0.17
seed_size	0.48	lodging	0.17
stem_cankers	0.48	sclerotia	0.17
crop_hist	0.46	plant_growth	0.17
leaf_mild	0.46	leaf_malf	0.15
fruit_spots	0.46	shriveling	0.15
external_decay	0.42	mycelium	0.15
date	0.33	seed_tmt	0.15
fruit_pods	0.30		

ตาราง 14 พบว่าค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคของถั่วเหลืองที่มีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูลนั้น แอตทริบิวต์ที่มีค่าระดับความสำคัญสูงที่สุด คือ แอตทริบิวต์ leaves ซึ่งหมายถึงลักษณะของใบต้นถั่วเหลือง โดยมีค่าระดับความสำคัญเท่ากับ 2.21 รองลงมาได้แก่ แอตทริบิวต์ stem ซึ่งหมายถึง ลักษณะของลำต้นถั่วเหลืองที่มีค่าระดับความสำคัญเท่ากับ 1.87 และแอตทริบิวต์ seed คือ ลักษณะเมล็ดถั่วเหลืองซึ่งมีค่าระดับความสำคัญเท่ากับ 1.09 ตามลำดับ

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 15 ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคหัวใจซึ่งมีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูล

แนวความคิด / แอตทริบิวต์	ค่าระดับความสำคัญ	แนวความคิด / แอตทริบิวต์	ค่าระดับความสำคัญ
Cp	0.38	Thalach	0.15
Thal	0.28	Exang	0.15
Sex	0.28	Oldpeak	0.15
Trestbps	0.21	Slope	0.15
Age	0.15	Ca	0.15
Restecg	0.15		

ตาราง 15 พบว่าค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคหัวใจที่มีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูลมีค่าระดับความสำคัญสูงสุดเท่ากับ 0.38 คือ แอตทริบิวต์ Cp ซึ่งหมายถึงรูปแบบอาการเจ็บหน้าอกของผู้ป่วยโรคหัวใจ รองลงมาคือแอตทริบิวต์ Thal ซึ่งหมายถึงสถานะของอัตราการเต้นของหัวใจ และแอตทริบิวต์ sex ซึ่งแสดงเพศของผู้ป่วยที่มีค่าระดับความสำคัญเท่ากับ 0.28 ตามลำดับ

ตาราง 16 ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019 ซึ่งมีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูล

แนวความคิด / แอตทริบิวต์	ค่าระดับความสำคัญ	แนวความคิด / แอตทริบิวต์	ค่าระดับความสำคัญ
Olfactory disorder	0.41	Fever	0.15
Gender	0.29	Headache	0.15
Dyspnea	0.28	Taste disorder	0.15
Cough	0.28	Coryza	0.15
Sore throat	0.15	Health professional	0.15

ตาราง 16 แสดงค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019 ที่มีความสัมพันธ์กับแอตทริบิวต์ภายในชุดข้อมูล โดยแอตทริบิวต์ที่มีค่าระดับความสำคัญสูงที่สุด คือ Olfactory disorder ซึ่งหมายถึง ความผิดปกติทางการรับกลิ่น โดยมีค่าระดับความสำคัญ

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

เท่ากับ 0.41 รองลงมาคือ แอตทริบิวต์ Gender หรือ เพศ โดยมีค่าระดับความสำคัญเท่ากับ 0.29 สำหรับแอตทริบิวต์ที่มีความสำคัญอันดับที่สามคือ แอตทริบิวต์ Dyspnea ซึ่งหมายถึงอาการหายใจลำบาก และแอตทริบิวต์ Cough ซึ่งหมายถึงอาการไอ โดยมีค่าระดับความสำคัญเท่ากับ 0.28 ตามลำดับ

ตาราง 17 ค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโรคไข้เลือดออกซึ่งมีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูล

แนวความคิด / แอตทริบิวต์	ค่าระดับความสำคัญ	แนวความคิด / แอตทริบิวต์	ค่าระดับความสำคัญ
fever	0.28	myalgia	0.15
age	0.21	arthralgia	0.15
gender	0.21	arthritis	0.15
headache	0.15	conjunctivitis	0.15
rash	0.15	lymphadenopathy	0.15
pruritus	0.15		

จากตาราง 17 พบว่าค่าระดับความสำคัญของแอตทริบิวต์ที่มีค่าสูงที่สุดคือ แอตทริบิวต์ fever ซึ่งหมายถึงอาการไข้ในผู้ป่วยไข้เลือดออก โดยมีค่าระดับความสำคัญเท่ากับ 0.28 รองลงมาคือ แอตทริบิวต์ age ซึ่งหมายถึงอายุ และแอตทริบิวต์ gender ซึ่งหมายถึงเพศ โดยมีค่าระดับความสำคัญเท่ากับ 0.21

จากค่าระดับความสำคัญของแนวความคิดในออนโทโลยีที่มีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูลที่ทำการศึกษาจะพบว่า สำหรับชุดข้อมูลการเกิดโรคของถั่วเหลืองนั้น ค่าระดับความสำคัญของแต่ละแอตทริบิวต์มีค่าที่แตกต่างกันเนื่องจากในออนโทโลยีโรคของถั่วเหลืองมีจำนวนการเชื่อมโยงกันระหว่างแนวความคิดสูง ซึ่งหมายถึงแต่ละแนวความคิดมีความสัมพันธ์กับแนวความคิดอื่น ๆ เป็นจำนวนมาก ในขณะที่ออนโทโลยีโรคหัวใจ ออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019 และออนโทโลยีโรคไข้เลือดออกนั้น แนวความคิดที่มีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูลที่ศึกษาเป็นแนวความคิดที่มีความสัมพันธ์กับแนวความคิดอื่น ๆ ในออนโทโลยีน้อย โดยส่วนใหญ่เป็นแนวความคิดที่ทำหน้าที่เป็นคลาสลูก (subclass) และไม่มีความสัมพันธ์ลักษณะอื่นกับแนวความคิดใด ๆ ในออนโทโลยี ดังนั้นจึงทำให้แนวความคิดเหล่านี้มีค่าระดับความสำคัญเท่ากับ 0.15 ซึ่งเป็นค่าระดับความสำคัญของแนวความคิดที่ไม่มีการเชื่อมโยงจากแนวความคิดอื่น ๆ

4918893 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

เมื่อพิจารณาค่าระดับความสำคัญของแนวความคิดเพศ ในออนโทโลยีโรคหัวใจ ออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019 และออนโทโลยีโรคไข้เลือดออก มีค่าระดับความสำคัญสูงเป็นอันดับต้น เมื่อเปรียบเทียบกับแนวความคิดอื่น ๆ เนื่องจากการพิจารณาค่าระดับความสำคัญของแนวความคิดด้วยอัลกอริทึม Weighted Semantic PageRank นี้มีแนวความคิดจากเทคนิค Term Frequency-Inverse Document Frequency (TF-IDF) (Qaiser & Ali, 2018) ซึ่งเป็นวิธีการพิจารณาความสำคัญของคำในเอกสาร โดยคำที่ปรากฏบ่อยครั้งในหลายเอกสารจะเป็นคำที่มีค่าน้ำหนักน้อยกว่าคำอื่น ๆ ที่พบในเอกสารจำนวนน้อย ดังนั้นแนวความคิดเพศมีการเชื่อมโยงกับแนวความคิดอื่น ๆ ด้วยความสัมพันธ์ต่าง ๆ เช่น ในออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019 แนวความคิดเพศเชื่อมโยงกับแนวความคิดผู้ป่วยด้วยความสัมพันธ์ has-gender ซึ่งเป็นความสัมพันธ์ที่ปรากฏน้อยครั้งจึงมีค่าน้ำหนักที่สูงและส่งผลให้แนวความคิดเพศมีค่าระดับความสำคัญสูง ในขณะที่แนวความคิดอื่น ๆ ในออนโทโลยีนี้มีการเชื่อมโยงระหว่างแนวความคิดด้วยความสัมพันธ์ Subclass-of ซึ่งเป็นความสัมพันธ์ที่ปรากฏบ่อยครั้งจึงมีค่าน้ำหนักน้อย และทำให้ค่าระดับความสำคัญของแนวความคิดมีค่าน้อยได้

ผลการทดสอบประสิทธิภาพการจำแนกข้อมูลของต้นไม้ตัดสินใจเชิงความหมาย

การทดลองนี้มีวัตถุประสงค์เพื่อตรวจสอบประสิทธิภาพของการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจด้วยการประยุกต์ใช้องค์ความรู้ในออนโทโลยีซึ่งอยู่ในรูปแบบของค่าระดับความสำคัญของแอตทริบิวต์ ค่าระดับความสำคัญของแอตทริบิวต์จะถูกนำไปใช้ในการปรับปรุงค่าเกนสารสนเทศดังแสดงในสมการ (17)

การทดลองนี้จะใช้ชุดข้อมูลทั้ง 4 ชุด คือ ชุดข้อมูลการเกิดโรคของถั่วเหลือง ชุดข้อมูลผู้ป่วยโรคหัวใจ ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออก โดยชุดข้อมูลการเกิดโรคถั่วเหลืองนั้นจะใช้ชุดข้อมูลที่ได้ทำการปรับปรุงข้อมูลด้วยแนวความคิดพื้นฐาน (Abstract data) ในออนโทโลยีจากการทดลองในบทที่ 4 และมีการดำเนินการดังนี้

- ข้อมูลที่ใช้ในการศึกษาจะถูกแบ่งออกเป็น 2 ส่วน คือ ข้อมูลสำหรับการเรียนรู้ (training data) จำนวนร้อยละ 70 ของจำนวนข้อมูลทั้งหมด และข้อมูลสำหรับการทดสอบ (test data) จำนวนร้อยละ 30 ของจำนวนข้อมูลทั้งหมด

- ทำการแก้ปัญหาข้อมูลที่ไม่สมดุล (imbalanced data) ในชุดข้อมูลผู้ป่วยโรคไข้เลือดออกด้วยวิธีการสุ่มลดข้อมูล (Undersampling technique)

- ทำการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย หรือ SDT กับ อัลกอริทึม ID3 ซึ่งเป็นอัลกอริทึมต้นไม้ตัดสินใจที่ใช้ค่าเกนสารสนเทศเป็นเกณฑ์ในการพิจารณาแอตทริบิวต์สำหรับเป็นโหนดของต้นไม้ตัดสินใจ

4918891813 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ผลการทดสอบประสิทธิภาพของการจำแนกข้อมูลด้วยอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายสามารถแสดงดังตาราง 18

ตาราง 18 ความถูกต้องในการจำแนกข้อมูลด้วยอัลกริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย

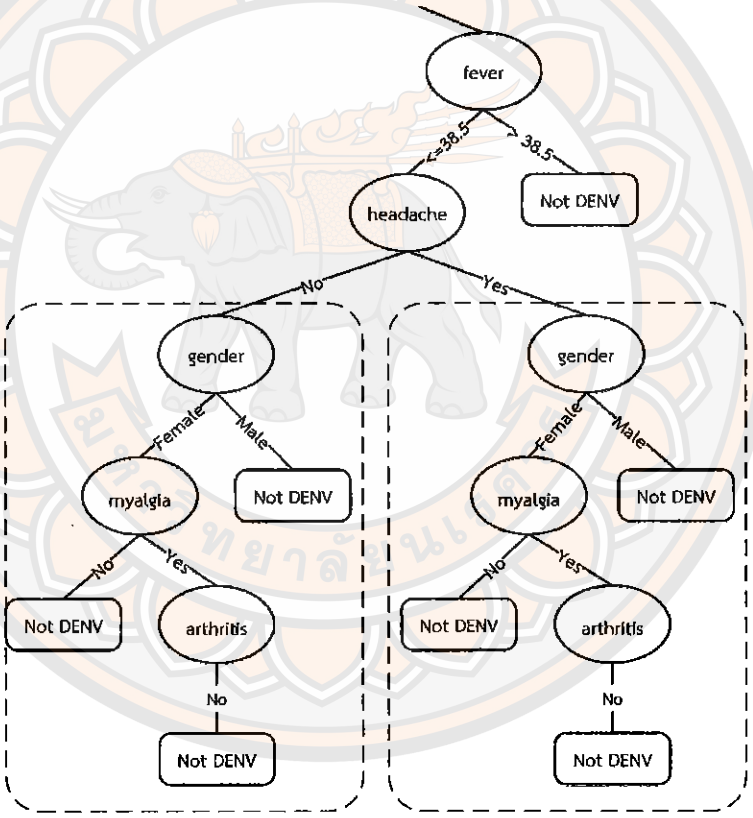
ชุดข้อมูล	ค่าเฉลี่ยความถูกต้องในการจำแนกข้อมูล		ค่าที่เปลี่ยนแปลง
	ID3	SDT	
ชุดข้อมูลการเกิดโรคของถั่วเหลือง	88.17%	92.13%	+3.96%
ชุดข้อมูลผู้ป่วยโรคหัวใจ	73.78%	77.37%	+3.59%
ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019	87.68%	88.66%	+0.98%
ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก	88.93%	89.58%	+0.65%

จากตาราง 18 จะพบว่าอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีค่าความถูกต้องในการจำแนกข้อมูลสูงกว่าอัลกอริทึม ID3 ในทุกชุดข้อมูลที่ทำการศึกษา สำหรับชุดข้อมูลการเกิดโรคของถั่วเหลืองค่าเฉลี่ยความถูกต้องในการจำแนกข้อมูลเมื่อใช้อัลกอริทึม ID3 เท่ากับ 88.17% และค่าเฉลี่ยของความถูกต้องในการจำแนกข้อมูลมีค่าเพิ่มเป็น 92.13% เมื่อทำการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย สำหรับการจำแนกข้อมูลชุดข้อมูลผู้ป่วยโรคหัวใจ ค่าเฉลี่ยของความถูกต้องในการจำแนกข้อมูลเมื่อใช้อัลกอริทึม ID3 มีค่าเท่ากับ 73.78% และเมื่อจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะทำให้ค่าเฉลี่ยของความถูกต้องในการจำแนกข้อมูลมีค่าเพิ่มขึ้นเป็น 77.37% สำหรับการจำแนกข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 นั้น เมื่อทำการจำแนกข้อมูลด้วยอัลกอริทึม ID3 จะมีค่าเฉลี่ยความถูกต้องในการจำแนกข้อมูลเท่ากับ 87.68% และค่าเฉลี่ยความถูกต้องในการจำแนกข้อมูลจะมีค่าเท่ากับ 88.66% เมื่อทำการจำแนกข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 ด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย ในส่วนของการจำแนกข้อมูลผู้ป่วยโรคไข้เลือดออกนั้นค่าเฉลี่ยความถูกต้องในการจำแนกข้อมูลจะมีค่าเท่ากับ 88.93% เมื่อทำการจำแนกข้อมูลด้วยอัลกอริทึม ID3 และค่าเฉลี่ยความถูกต้องในการจำแนกข้อมูลจะเพิ่มเป็น 89.58% เมื่อจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย

จากตาราง 18 จะพบว่าค่าความถูกต้องในการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจเชิงความหมายสำหรับชุดข้อมูลข้อมูลผู้ป่วยโรคไข้เลือดออกมีค่าความถูกต้องเพิ่มขึ้นเพียง 0.65%

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

เมื่อเปรียบเทียบกับค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึม ID3 เนื่องจากการสร้างแบบจำลองการจำแนกข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจนั้น อัลกอริทึมจะพยายามทำการเลือกแอตทริบิวต์ที่เหมาะสมที่สามารถแบ่งแยกคลาสที่ต้องการจำแนกได้ โดยดำเนินการเช่นนี้จนกระทั่งสามารถจำแนกข้อมูลออกเป็นคลาสคำตอบ กล่าวอีกนัยหนึ่งคือ ดำเนินการจนกระทั่งไม่สามารถแตกกิ่งของต้นไม้ตัดสินใจได้อีก จากวิธีการดังกล่าวเมื่อดำเนินการสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม ID3 จากชุดข้อมูลผู้ป่วยโรคไข้เลือดออก อัลกอริทึมจะใช้แอตทริบิวต์ทั้งหมดในการสร้างต้นไม้ตัดสินใจ และทำให้เกิดความซ้ำซ้อนของต้นไม้ย่อย (subtree) ภายในต้นไม้ตัดสินใจที่สร้างขึ้น ดังแสดงในภาพ 27



ภาพ 27 ตัวอย่างต้นไม้ย่อย (subtree) ที่มีความซ้ำซ้อนเมื่อสร้างต้นไม้ตัดสินใจด้วยอัลกอริทึม ID3

จากภาพ 27 จะพบว่าเมื่ออัลกอริม ID3 เมื่อทำการแตกกิ่งของต้นไม้ตัดสินใจจากค่าของแอตทริบิวต์ headache ซึ่งหมายถึง อาการปวดศีรษะนั้น อัลกอริทึมจะพยายามแตกกิ่งจากชุดข้อมูลที่ได้เรียนรู้ และเกิดต้นไม้ย่อยที่มีความซ้ำซ้อนหลังจากการเลือกแอตทริบิวต์ headache โดยการที่ต้นไม้ตัดสินใจเกิดความซ้ำซ้อนกันของต้นไม้ย่อยนั้นจะหมายถึงต้นไม้ตัดสินใจเกิดปัญหา

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

แฟรกเมนเตชัน (fragmentation) คือ ปัญหาที่ต้นไม้ตัดสินใจพยายามแตกกิ่งเพื่อจำแนกข้อมูลออกเป็นคลาสใด ๆ โดยพิจารณาจากข้อมูลจำนวนน้อย ซึ่งทำให้เกิดปัญหาความจำเพาะกับข้อมูลที่เรียนรู้ (Overfitting) (Rokach, 2016) และส่งผลต่อความถูกต้องในการจำแนกข้อมูลที่ไม่เคยเรียนรู้มาก่อน สำหรับการสร้างแบบจำลองการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเพื่อจำแนกข้อมูลผู้ป่วยโรคไข้เลือดออกนั้น ค่าระดับความสำคัญของแอตทริบิวต์จะช่วยให้แอตทริบิวต์ที่มีความสำคัญกับคลาสที่ต้องการจำแนกมีโอกาสถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากขึ้นและเกิดการเปลี่ยนแปลงโครงสร้างของต้นไม้ตัดสินใจ อย่างไรก็ตามต้นไม้ตัดสินใจเชิงความหมายที่ได้ยังคงปรากฏความซ้ำซ้อนกันของต้นไม้ย่อยในบางส่วน เมื่อนำไปจำแนกชุดข้อมูลทดสอบจึงทำให้ค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้นเพียง 0.65%

จากผลการทดลองพบว่าอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีประสิทธิภาพในการจำแนกข้อมูลที่ดีกว่าอัลกอริทึม ID3 ซึ่งใช้ค่าเกนสารสนเทศในการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับสร้างต้นไม้ตัดสินใจจากการพิจารณาค่าของข้อมูลที่ใช้ในการศึกษา ในกรณีที่ชุดข้อมูลที่มีขนาดเล็ก เช่น ชุดข้อมูลผู้ป่วยโรคหัวใจซึ่งมีจำนวนข้อมูลเพียง 297 แถว เมื่อนำมาใช้ในการสร้างแบบจำลองการจำแนกข้อมูลด้วยอัลกอริทึม ID3 อาจส่งผลต่อประสิทธิภาพของแบบจำลองเนื่องจากอาจขาดข้อมูลที่มีความสำคัญต่อการวิเคราะห์ข้อมูล เช่น ค่าความดันโลหิตเป็นปัจจัยหนึ่งที่แสดงให้เห็นถึงความเสี่ยงของการเกิดโรคหัวใจ โดยผู้ที่มีค่าความดันโลหิตสูงซึ่งหมายถึงผู้มีค่าความดันโลหิตมากกว่า 120 มิลลิเมตรปรอท มีโอกาสที่จะเกิดโรคหัวใจมากกว่าผู้ที่มีระดับความดันโลหิตปกติ (Gray et al., 2011) โดยเป็นปัจจัยเสี่ยงที่มีนัยสำคัญมากกว่าอายุของผู้ป่วย (Onat, 2001) แต่เมื่อพิจารณาความสามารถในการจำแนกข้อมูลออกเป็นคลาสต่าง ๆ ของแต่ละแอตทริบิวต์ในชุดข้อมูลผู้ป่วยโรคหัวใจจากค่าเกนสารสนเทศของชุดข้อมูลในตาราง 19 จะพบว่าแอตทริบิวต์ Trestbps ซึ่งแสดงค่าระดับความดันโลหิตของผู้ป่วยเป็นแอตทริบิวต์ที่สามารถใช้ในการแบ่งแยกข้อมูลออกเป็นคลาสต่าง ๆ ได้น้อยกว่าแอตทริบิวต์ Age โดยแอตทริบิวต์ Trestbps เป็นแอตทริบิวต์ที่มีค่าเกนสารสนเทศน้อยที่สุด ซึ่งเกิดจากจำนวนข้อมูลในชุดข้อมูลไม่เพียงพอ ทำให้แอตทริบิวต์ Trestbps ซึ่งมีนัยสำคัญต่อการเกิดโรคหัวใจมีโอกาสถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจน้อยกว่าแอตทริบิวต์อื่น ๆ ที่มีนัยสำคัญต่อการเกิดโรคน้อยกว่า

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 19 ค่าเกนสารสนเทศของแอตทริบิวต์ในชุดข้อมูลผู้ป่วยโรคหัวใจ

ลำดับ	แอตทริบิวต์	ค่าเกนสารสนเทศ
1	Cp	0.180
2	Thal	0.171
3	Ca	0.157
4	Slope	0.141
5	Oldpeak	0.140
6	Exang	0.139
7	Thalach	0.119
8	**Age**	**0.066**
9	Sex	0.052
10	Restecg	0.023
11	**Trestbps**	**0.021**

ในขณะที่อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะนำองค์ความรู้ในออนโทโลยีซึ่งอยู่ในรูปแบบค่าระดับความสำคัญของแอตทริบิวต์มาช่วยในการพิจารณาแอตทริบิวต์ที่เหมาะสมร่วมกับค่าเกนสารสนเทศที่คำนวณได้จากชุดข้อมูล ทำให้ในการพิจาณาแอตทริบิวต์สำหรับต้นไม้ตัดสินใจเชิงความหมายนั้นแอตทริบิวต์ที่มีค่าเกนสารสนเทศน้อยแต่มีความสำคัญในโดเมนที่ศึกษามีโอกาสถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจเชิงความหมายมากขึ้น และส่งผลต่อความถูกต้องในการจำแนกข้อมูล

จากการทดลองประสิทธิภาพในการจำแนกข้อมูลของต้นไม้ตัดสินใจเชิงความหมายกับชุดข้อมูลทั้ง 4 ชุดข้อมูลที่มีลักษณะแตกต่างกันพบว่า อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะสามารถทำงานได้ดีกับชุดข้อมูลที่แต่ละแอตทริบิวต์มีจำนวนค่าข้อมูลที่แตกต่างกันซึ่งหมายถึงชุดข้อมูลนั้นมีโอกาสเกิดปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลาย เช่น ชุดข้อมูลการเกิดโรคของถั่วเหลือง และชุดข้อมูลผู้ป่วยโรคหัวใจ ซึ่งมีค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้น 3.96% และ 3.59% เมื่อใช้อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายในการจำแนกข้อมูล ในขณะที่ชุดข้อมูลที่แต่ละแอตทริบิวต์มีจำนวนค่าข้อมูลในแอตทริบิวต์ใกล้เคียงกัน หรือในแอตทริบิวต์มีค่าข้อมูลเพียงสองค่า เช่น ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 ซึ่งแต่ละแอตทริบิวต์มีค่าข้อมูลเป็น Positive และ Negative นั้น เมื่อทำการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะมีค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้นเพียง 0.98%

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ในการทดลองนี้ได้ทำการพิจารณาความซับซ้อนของออนโทโลยีที่มีต่อความถูกต้องในการจำแนกข้อมูล โดยการพิจารณาความซับซ้อนของออนโทโลยีจะพิจารณาจากเกณฑ์ ComplexOnto (Kumar & Baliyan, 2018) ซึ่งมีองค์ประกอบที่เกี่ยวข้อง 4 องค์ประกอบ ได้แก่ ความหนาแน่นของความสัมพันธ์ (Link density) จำนวนความสัมพันธ์ต่อแนวความคิด (Link per concept) จำนวนความสัมพันธ์ในออนโทโลยี (Link richness) และ ค่าความซับซ้อนไซโคลมาติก (Cyclomatic complexity) ผลลัพธ์ของการพิจารณาความซับซ้อนของออนโทโลยีที่ใช้ในการวิจัยสามารถแสดงดังตาราง 20

ตาราง 20 ค่าความซับซ้อนของออนโทโลยี

ออนโทโลยีที่ใช้ในการวิจัย	Link density	Link per concept	Link richness	Cyclomatic complexity	ComplexOnto score
ออนโทโลยีโรคของถั่วเหลือง	305×10^{-6}	1604×10^{-4}	1349×10^{-4}	22.00	5.57
ออนโทโลยีโรคหัวใจ	0.7×10^{-6}	6×10^{-4}	5×10^{-4}	413.00	103.25
ออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019	4.6×10^{-6}	53×10^{-4}	46×10^{-4}	355.00	88.75
ออนโทโลยีโรคไข้เลือดออก	2×10^{-6}	50×10^{-4}	42×10^{-4}	888.00	222.00

จากตาราง 20 พบว่าออนโทโลยีโรคไข้เลือดออกเป็นออนโทโลยีที่มีค่าความซับซ้อน (ComplexOnto score) สูงที่สุด คือ 222.00 ในขณะที่ออนโทโลยีโรคของถั่วเหลืองมีค่าความซับซ้อนน้อยที่สุด โดยมีค่าเท่ากับ 5.57 และเมื่อพิจารณาความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายในตาราง18 พบว่าค่าความถูกต้องในการจำแนกข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 มีค่าใกล้เคียงกับค่าความถูกต้องในการจำแนกข้อมูลผู้ป่วยโรคไข้เลือดออก ในขณะที่ค่าซับซ้อนที่ได้จากออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019 มีความแตกต่างกับค่าความซับซ้อนของออนโทโลยีโรคไข้เลือดออกเท่ากับ 133.25 และเมื่อพิจารณาค่าความถูกต้อง

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ในการจำแนกข้อมูลผู้ป่วยโรคหัวใจซึ่งมีค่าเท่ากับ 77.37% ซึ่งมีค่าน้อยกว่าค่าความถูกต้องในการจำแนกข้อมูลของผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 ที่มีค่าเท่ากับ 89.58% ในขณะที่ค่าความซับซ้อนของออนโทโลยีที่เกี่ยวข้องกับทั้งสองชุดข้อมูลมีค่าใกล้เคียงกัน ดังนั้นจึงสามารถสรุปได้ว่าค่าความซับซ้อนของออนโทโลยีไม่มีความสัมพันธ์โดยตรงกับค่าความถูกต้องในการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย

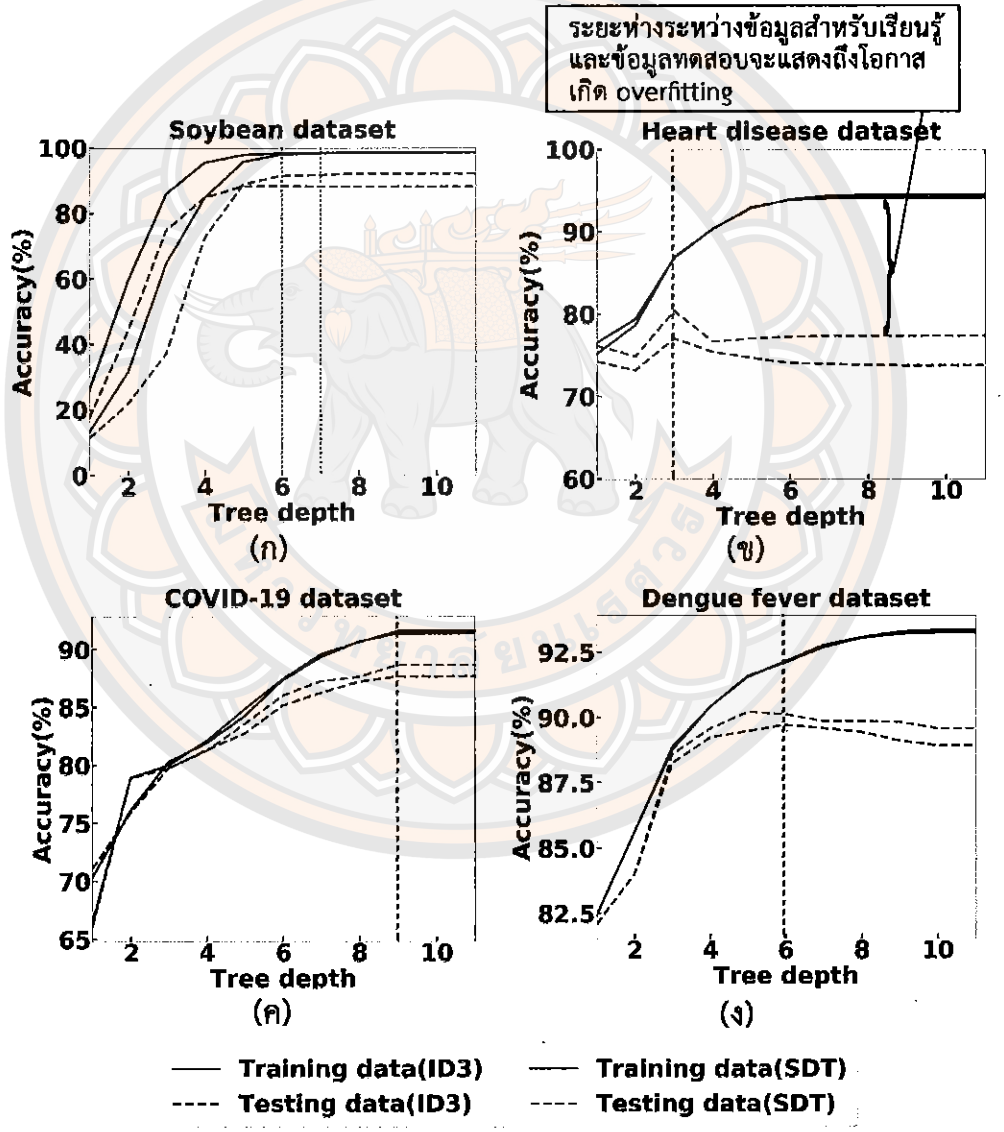
ปัจจัยที่ส่งผลต่อประสิทธิภาพการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย คือ ค่าระดับความสำคัญของแนวความคิดในออนโทโลยี ซึ่งค่าระดับความสำคัญของแนวความคิดจะมีความสัมพันธ์กับค่าจำนวนความสัมพันธ์ต่อแนวความคิด (Link per concept) และจำนวนความสัมพันธ์ในออนโทโลยี (Link richness) โดยเกี่ยวข้องกับความถี่ในการเชื่อมโยงความสัมพันธ์ระหว่างแนวความคิดและจำนวนประเภทความสัมพันธ์ที่ใช้ในอัลกอริทึม Weighted Semantic PageRank ซึ่งหากค่าจำนวนความสัมพันธ์ต่อแนวความคิดและจำนวนความสัมพันธ์ในออนโทโลยีมีค่าสูงจะหมายถึง ในออนโทโลยีมีแนวความคิดที่มีการเชื่อมโยงกับแนวความคิดอื่น ๆ ด้วยความสัมพันธ์แตกต่างกัน ส่งผลให้ค่าระดับความสำคัญของแต่ละแนวความคิดที่คำนวณได้มีค่าแตกต่างกัน การที่ค่าระดับความสำคัญของแนวความคิดมีค่าแตกต่างกันนั้นจะทำให้เมื่อนำค่าระดับความสำคัญเหล่านี้ไปปรับปรุงค่าเกนสารสนเทศจะทำให้เกิดการเปลี่ยนแปลงลำดับของแอตทริบิวต์ในขั้นตอนการพิจารณาโหนดของต้นไม้ตัดสินใจ ซึ่งทำให้แอตทริบิวต์ที่มีความสำคัญกับคลาสที่ต้องการจำแนกมีโอกาสถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากขึ้น และส่งผลให้ความถูกต้องในการจำแนกข้อมูลมีค่าเพิ่มขึ้น

ผลการทดสอบความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting)

การทดลองนี้มีวัตถุประสงค์เพื่อสำรวจโอกาสในการเกิดความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting) ของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย ซึ่งปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้นี้เป็นประเด็นที่นักวิทยาการข้อมูลให้ความสำคัญเพื่อป้องกันไม่ให้เกิดการแปลผลที่ผิดพลาด โดยในการทดลองนี้จะทำการเปรียบเทียบความถูกต้องในการจำแนกข้อมูลที่ใช้สำหรับการเรียนรู้ (training data) และ ความถูกต้องในการจำแนกข้อมูลสำหรับการทดสอบ (test data) ของอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย โดยหากค่าความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้มีค่ามากกว่าค่าความถูกต้องในการจำแนกข้อมูลสำหรับการทดสอบมาก จะหมายถึงแบบจำลองนั้นมีโอกาสในการเกิดความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้สูง ผลลัพธ์ของการทดลองนี้สามารถแสดงดังภาพ 28

4918819183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ภาพ 28 เส้นทึบจะแสดงค่าความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้ เส้นประจะหมายถึงค่าความถูกต้องในการจำแนกข้อมูลทดสอบ และเส้นประในแนวตั้งจะหมายถึงตำแหน่งที่เกิดการจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ หรือ overfitting ของแต่ละอัลกอริทึม จากผลการจำแนกข้อมูลของทั้งสองอัลกอริทึมจะพบว่าค่าความถูกต้องในการจำแนกข้อมูลจะมีค่ามากขึ้นเมื่อความสูงของต้นไม้ตัดสินใจเพิ่มขึ้น อย่างไรก็ตามสำหรับการจำแนกข้อมูลสำหรับการทดสอบนั้นค่าความถูกต้องในการจำแนกข้อมูลจะเพิ่มขึ้นและค่าความถูกต้องในการจำแนกข้อมูลจะเริ่มลดลงเนื่องจากเกิดการจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้



ภาพ 28 ผลการทดสอบความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting) ของชุดข้อมูลต่าง ๆ โดย (ก) ชุดข้อมูลการเกิดโรคถั่วเหลือง (ข) ชุดข้อมูลผู้ป่วยโรคหัวใจ (ค) ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และ (ง) ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

สำหรับภาพ 28 (ก) จะแสดงผลการจำแนกข้อมูลของชุดข้อมูลการเกิดโรคของถั่วเหลือง โดยค่าความถูกต้องในการจำแนกข้อมูลสำหรับทดสอบจะเริ่มลดลงที่ความสูงของต้นไม้ตัดสินใจมีค่าเท่ากับ 6 เมื่อใช้อัลกอริทึม ID3 และเมื่อใช้อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายค่าความถูกต้องในการจำแนกข้อมูลทดสอบจะเริ่มลดลงเมื่อต้นไม้ตัดสินใจมีความสูงเท่ากับ 7

สำหรับผลการจำแนกชุดข้อมูลผู้ป่วยโรคหัวใจจะแสดงดังภาพ 28 (ข) โดยค่าความถูกต้องในการจำแนกข้อมูลสำหรับการทดสอบของทั้งสองอัลกอริทึมจะเริ่มลดลงเมื่อต้นไม้ตัดสินใจมีความสูงเท่ากับ 3 ซึ่งหมายถึงเกิดการจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ขึ้น โดยสังเกตได้จากค่าความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และค่าความถูกต้องในการจำแนกข้อมูลทดสอบมีความแตกต่างกันมาก ซึ่งปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ที่เกิดขึ้นในชุดข้อมูลนี้มีสาเหตุจากชุดข้อมูลผู้ป่วยโรคหัวใจเป็นชุดข้อมูลที่มีขนาดเล็ก โดยเมื่อทำการวิเคราะห์ชุดข้อมูลที่มีขนาดเล็กด้วยอัลกอริทึมต้นไม้ตัดสินใจ ปริมาณข้อมูลจะไม่เพียงพอให้อัลกอริทึมใช้ในการเรียนรู้เพื่อสร้างแบบจำลองและส่งผลให้เกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ได้ซึ่งสอดคล้องตามงานวิจัยของ Song & Lu (2015)

ผลการจำแนกข้อมูลสำหรับชุดข้อมูผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 แสดงดังภาพ 28 (ค) โดยค่าความถูกต้องในการจำแนกข้อมูลสำหรับการทดสอบของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีค่าสูงกว่าค่าความถูกต้องที่ได้จากอัลกอริทึม ID3 ซึ่งทำให้ความแตกต่างของค่าความถูกต้องสำหรับการจำแนกข้อมูลสำหรับการเรียนรู้และการจำแนกข้อมูลทดสอบของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีค่าน้อยกว่าอัลกอริทึม ID3 ซึ่งหมายถึงอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีโอกาสในการเกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้น้อยกว่าอัลกอริทึม ID3

สำหรับผลการจำแนกข้อมูลของชุดข้อมูลผู้ป่วยโรคไข้เลือดออกจะแสดงดังภาพ 28 (ง) โดยพบว่าค่าความถูกต้องในการจำแนกข้อมูลทดสอบของทั้งสองอัลกอริทึมจะมีค่าลดลงเมื่อต้นไม้ตัดสินใจมีความสูงเท่ากับ 6 อย่างไรก็ตามความแตกต่างระหว่างค่าความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และค่าความถูกต้องในการจำแนกข้อมูลทดสอบของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีค่าน้อยกว่าอัลกอริทึม ID3 เช่นเดียวกับผลที่ได้จากข้อมูลชุดอื่น ๆ

ในการทดลองนี้ได้ทำการเปรียบเทียบค่าความแตกต่างระหว่างค่าความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และค่าความถูกต้องในการจำแนกข้อมูลสำหรับการทดสอบของทั้งสองอัลกอริทึม ซึ่งการเปรียบเทียบค่าความแตกต่างที่ได้นี้จะเป็นการพิจารณาโอกาสในการเกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ โดยใช้วิธีการทดสอบวิลคอกซัน (Wilcoxon Signed Rank Test) ซึ่งเป็นวิธีการทางสถิติที่ใช้ในการทดสอบความแตกต่างของข้อมูลสองกลุ่ม เนื่องจากใน

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ข้อมูลค่าความแตกต่างของความถูกต้องในการจำแนกข้อมูลบางส่วนมีการแจกแจงไม่ปกติ โดยมีสมมติฐานทางสถิติดังนี้

H_0: ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบของอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายไม่แตกต่างกัน

H_1: ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบของอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายแตกต่างกัน

ผลของการทดสอบสำหรับแต่ละชุดข้อมูลสามารถแสดงได้ดังตาราง 21 ถึง ตาราง 24

ตาราง 21 ผลลัพธ์การทดสอบวิลคอกซันของชุดข้อมูลการเกิดโรคของถั่วเหลือง

ความสูงของต้นไม้ตัดสินใจ	ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบ (±S.D.)		*Z*	*p-value*
	ID3	SDT		
1	8.29±6.78	1.89±5.10	-3.692**	< 0.001
2	15.26±6.87	9.68±6.19	-3.980**	< 0.001
3	11.03±5.42	27.46±4.51	-4.762**	< 0.001
4	**10.69±2.74**	**12.21±4.26**	**-1.594**	**0.111**
5	9.58±2.56	6.90±2.04	-3.918**	< 0.001
6	10.02±2.29	6.43±1.68	-4.679**	< 0.001
7	10.33±2.36	6.69±1.90	-4.659**	< 0.001
8	10.38±2.31	6.43±1.73	-4.700**	< 0.001
9	10.38±2.31	6.46±1.70	-4.700**	< 0.001
10	10.38±2.31	6.46±1.70	-4.700**	< 0.001
11	10.38±2.31	6.46±1.70	-4.700**	< 0.001
12	10.38±2.31	6.46±1.70	-4.700**	< 0.001

จากตาราง 21 จะพบว่าในสำหรับการจำแนกข้อมูลการเกิดโรคของถั่วเหลืองนั้นค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบของอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายที่ความสูงของต้นไม้ตัดสินใจ

4918883183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

เท่ากับสี่มีค่าไม่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติโดยมีค่า *p-value* เท่ากับ 0.111 ในขณะที่เมื่อพิจารณาข้อมูล ณ ความสูงอื่น ๆ ของต้นไม้ตัดสินใจพบว่า ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบของทั้งสองอัลกอริทึมมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติ โดยค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบจากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายส่วนใหญ่จะมีค่าน้อยกว่าค่าที่ได้จากอัลกอริทึม ID3

ตาราง 22 ผลลัพธ์การทดสอบวิลคอกซันของชุดข้อมูลผู้ป่วยโรคหัวใจ

ความสูงของต้นไม้ตัดสินใจ	ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบ (±S.D.)		*Z*	*p-value*
	ID3	SDT		
1	2.29±4.15	−1.00±5.23	3.992-**	< 0.001
2	6.18±5.98	3.79±4.82	2.606**	0.009
3	9.77±4.26	6.55±3.84	3.733-**	< 0.001
4	**15.00±4.64**	**13.69±4.65**	**−1.243**	**0.214**
5	18.18±4.90	15.92±3.54	−2.386**	0.017
6	19.86±5.49	16.59±3.82	−3.157**	0.002
7	20.38±5.51	16.70±3.62	−3.744**	< 0.001
8	20.61±5.51	16.74±3.61	−3.816**	< 0.001
9	20.70±5.45	16.74±3.61	−3.857**	< 0.001
10	20.70±5.45	16.74±3.61	−3.857**	< 0.001
11	20.70±5.45	16.74±361.	−3.857**	< 0.001
12	20.70±5.45	16.74±3.61	−3.857**	< 0.001

จากตาราง 22 พบว่าค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบของอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายส่วนใหญ่มีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติโดยมีค่า *p-value* ของการทดสอบวิลคอกซันน้อยกว่า 0.05 ยกเว้นเมื่อต้นไม้ตัดสินใจมีความสูงเท่ากับ 4 ค่า *p-value* มีค่าเท่ากับ 0.214 ซึ่งหมายถึง เมื่อต้นไม้ตัดสินใจมีค่าความสูงเท่ากับ 4 ค่าเฉลี่ยของค่าความแตกต่าง

ระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบของทั้งสองอัลกอริทึมไม่แตกต่างกัน

ตาราง 23 ผลลัพธ์การทดสอบวิลคอกซันของชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019

ความสูงของต้นไม้ตัดสินใจ	ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบ(±S.D.)		*Z*	*p-value*
	ID3	SDT		
1	-0.75±2.16	0.38±1.88	-2.258**	0.024
2	0.22±1.79	0.06±1.57	-1.184	0.236
3	0.54±1.68	0.25±1.81	-2.596**	0.009
4	0.62±1.64	0.86±1.34	-1.178	0.239
5	1.50±1.57	1.27±1.35	-0.998	0.318
6	2.19±1.24	1.42±1.33	-3.075**	0.002
7	3.10±1.03	2.33±1.23	-3.260**	0.001
8	3.47±1.13	2.98±1.20	-3.054**	0.002
9	3.72±1.14	2.94±1.09	-4.412**	< 0.001
10	3.72±1.14	2.94±1.09	-4.412**	< 0.001
11	3.72±1.14	2.94±1.09	-4.412**	< 0.001
12	3.72±1.14	2.94±1.09	-4.412**	< 0.001

ตาราง 23 แสดงผลของการทดสอบวิลคอกซันของชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 ซึ่งพบว่าเมื่อต้นไม้ตัดสินใจมีความสูงเท่ากับ 2, 4 และ 5 ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบที่ได้จากอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ โดยมีค่า *p-value* มากกว่า 0.05 ส่วนในระดับความสูงอื่น ๆ ค่าเฉลี่ยของความแตกต่างระหว่างค่าความถูกต้องในการจำแนกข้อมูลสำหรับการเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบจะแตกต่างกัน โดยค่าที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะมีค่าน้อยกว่าค่าที่ได้จากอัลกอริทึม ID3

4918689183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 24 ผลลัพธ์การทดสอบวิลคอกซันของชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

ความสูงของต้นไม้ตัดสินใจ	ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบ (±S.D.)		*Z*	*p-value*
	ID3	SDT		
1	0.34±2.32	0.34±2.32	0.000	1.000
2	1.61±2.13	1.62±2.14	−1.000	0.317
3	0.50±2.13	0.29±2.05	−0.507	0.612
4	1.18±1.89	0.82±2.06	−2.679**	0.007
5	2.07±1.94	1.38±1.94	-2.718**	0.007
6	2.43±2.14	1.97±1.79	-2.126**	0.033
7	3.18±1.84	2.83±1.86	-2.283**	0.022
8	3.63±1.97	3.18±2.16	-2.950**	0.003
9	4.18±1.89	3.38±2.07	-4.185**	< 0.001
10	4.42±1.88	3.67±2.02	-4.054**	< 0.001
11	4.42±1.88	3.67±2.02	-4.054**	< 0.001
12	4.42±1.88	3.67±2.02	-4.054**	< 0.001

ตาราง 24 พบว่าสำหรับชุดข้อมูลผู้ป่วยไข้เลือดออกเมื่อต้นไม้ตัดสินใจมีค่าความสูงน้อย ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบของทั้งสองอัลกอริทึมไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติ ในขณะที่เมื่อความสูงของต้นไม้ตัดสินใจมีมากกว่า 4 ค่าเฉลี่ยของค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบของทั้งสองอัลกอริทึมจะแตกต่างกัน โดยค่าที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะมีค่าน้อยกว่าค่าที่ได้จากอัลกอริทึม ID3

จากผลการทดลองในแต่ละชุดข้อมูลจะเห็นได้ว่า เมื่อต้นไม้ตัดสินใจมีค่าความสูงน้อย อัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะมีโอกาสในการเกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ หรือ overfitting ไม่แตกต่างกัน แต่เมื่อความสูงของต้นไม้ตัดสินใจเพิ่มขึ้นค่าความแตกต่างระหว่างความถูกต้องในการจำแนกข้อมูลเรียนรู้และความถูกต้องในการจำแนกข้อมูลทดสอบที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะมีค่าน้อยกว่าค่าดังกล่าวที่ได้จากอัลกอริทึม ID3 ดังแสดงในภาพ 28 ดังนั้นจึงสามารถสรุปได้ว่าอัลกอริทึมต้นไม้

4918891813 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตัดสินใจเชิงความหมายจะมีโอกาสในการเกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้หรือ overfitting น้อยกว่าอัลกอริทึม ID3 เนื่องจากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีการนำค่าความสำคัญของแอตทริบิวต์ที่ได้จากออนโทโลยีมาช่วยในกระบวนการพิจารณาแอตทริบิวต์สำหรับเป็นโหนดภายในต้นไม้ตัดสินใจ ซึ่งส่งผลให้ได้แอตทริบิวต์ที่ไม่ได้ขึ้นต่อกันโดยตรงกับค่าข้อมูลที่ใช้ในการเรียนรู้เพื่อสร้างแบบจำลอง จึงสามารถช่วยลดโอกาสในการเกิดความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ได้

ผลการทดสอบการทนทานต่อข้อมูลที่ผิดปกติต่อต้นไม้ตัดสินใจเชิงความหมาย

การทดลองนี้มีวัตถุประสงค์เพื่อตรวจสอบผลกระทบของข้อมูลที่ผิดปกติ (Noise) ที่มีต่ออัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย โดยทำการเปรียบเทียบผลการจำแนกข้อมูลที่ได้จากอัลกอริทึม ID3 กับผลการจำแนกข้อมูลที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย ในการทดลองนี้จะทำการวิเคราะห์ชุดข้อมูลที่มีปริมาณข้อมูลที่ผิดปกติที่แตกต่างกัน ซึ่งประกอบไปด้วยชุดข้อมูลที่ไม่มีข้อมูลผิดปกติ ชุดข้อมูลที่มีข้อมูลผิดปกติร้อยละ 10 ชุดข้อมูลที่มีข้อมูลผิดปกติร้อยละ 20 และชุดข้อมูลที่มีข้อมูลผิดปกติร้อยละ 30

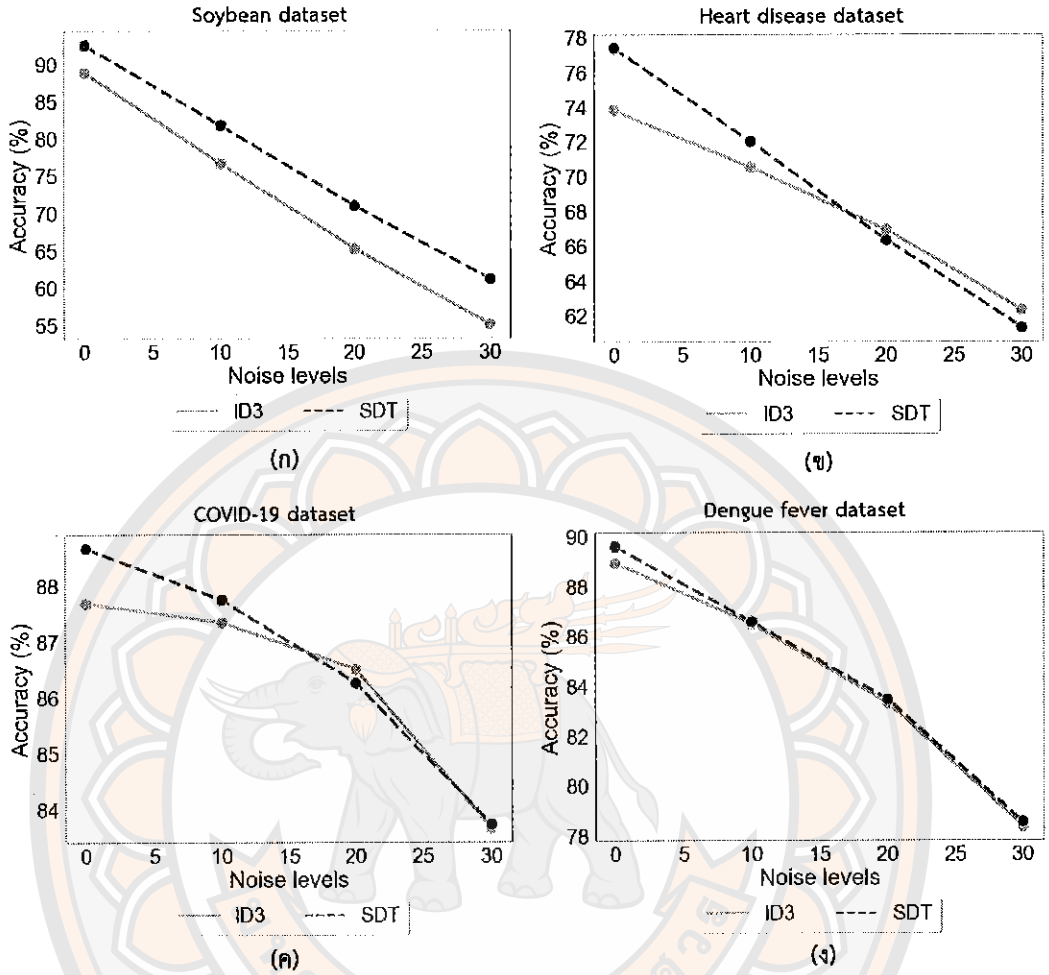
ในการทดลองนี้ข้อมูลที่ผิดปกติจะหมายถึงข้อมูลที่มีคลาสคำตอบที่ผิดปกติ (Class noise) ซึ่งหมายถึง แถวข้อมูลมีการระบุคลาสคำตอบของแถวข้อมูลนั้นไม่ถูกต้อง การสร้างข้อมูลที่ผิดปกติจะดำเนินการโดยการสุ่มจำนวนแถวข้อมูลในชุดข้อมูลเพื่อกำหนดเป็นข้อมูลผิดปกติตามจำนวนที่กำหนด พร้อมทั้งดำเนินการดังนี้

- สำหรับชุดข้อมูลที่มีคลาสคำตอบจำนวนสองคลาส ซึ่งประกอบไปด้วย ชุดข้อมูลผู้ป่วยโรคหัวใจ ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออก คลาสคำตอบของแถวข้อมูลที่ถูกสุ่มเลือกจะถูกเปลี่ยนเป็นคลาสคำตอบที่ตรงข้าม เช่น ในชุดข้อมูลผู้ป่วยโรคหัวใจหากแถวข้อมูลใดระบุคลาสคำตอบเป็น “ผู้ป่วยโรคหัวใจ” ก็จะถูกเปลี่ยนเป็น “ไม่ใช่ผู้ป่วยโรคหัวใจ” เป็นต้น

- สำหรับชุดข้อมูลที่มีคลาสคำตอบมากกว่าสองคลาส ซึ่งได้แก่ ชุดข้อมูลการเกิดโรคของถั่วเหลือง คลาสคำตอบของแถวข้อมูลจะถูกเปลี่ยนเป็นคลาสคำตอบอื่นที่มีจำนวนแถวข้อมูลในคลาสใกล้เคียงกัน เช่น หากชุดข้อมูลมีคลาสคำตอบเป็น “โรคราแป้ง (powdery mildew)” จะถูกเปลี่ยนเป็น “โรคราน้ำค้าง (downy mildew)” ซึ่งเป็นคลาสคำตอบที่มีจำนวนแถวข้อมูลใกล้เคียงกันเป็นต้น

ผลของการวิเคราะห์ชุดข้อมูลที่มีปริมาณข้อมูลรบกวนแตกต่างกันสามารถแสดงดังภาพ 29

4918819183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39



ภาพ 29 ผลการจำแนกข้อมูลที่มีปริมาณข้อมูลรบกวนแตกต่างกัน โดย (ก) ชุดข้อมูลการเกิดโรคของถั่วเหลือง (ข) ชุดข้อมูลผู้ป่วยโรคหัวใจ (ค) ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และ (ง) ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

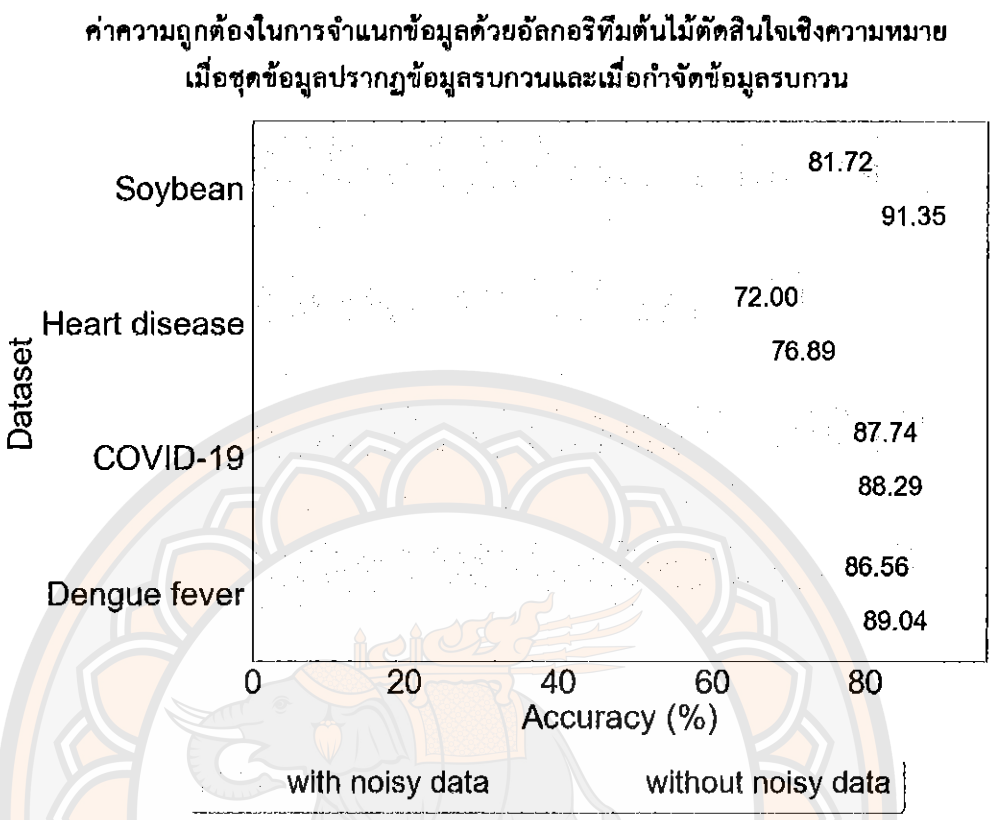
ภาพ 29 แสดงผลการจำแนกข้อมูลเมื่อชุดข้อมูลปรากฏข้อมูลที่ผิดปกติในจำนวนที่แตกต่างกัน โดยเส้นสีเขียวจะแสดงผลการจำแนกข้อมูลที่ได้จากอัลกอริทึม ID3 และเส้นประสีน้ำเงินแสดงผลการจำแนกข้อมูลที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย จากภาพ 29 จะพบว่าจำนวนข้อมูลที่ผิดปกติในชุดข้อมูลจะส่งผลต่อค่าความถูกต้องในการจำแนกทั้งสองอัลกอริทึม โดยเมื่อชุดข้อมูลมีจำนวนข้อมูลที่ผิดปกติเพิ่มขึ้นจะส่งผลให้ค่าความถูกต้องในการจำแนกข้อมูลลดลง เนื่องจากการนำชุดข้อมูลที่มีข้อมูลผิดปกติมาใช้ในการวิเคราะห์ข้อมูล อัลกอริทึมจะทำการสร้างแบบจำลองโดยการเรียนรู้ข้อมูลที่ผิดปกติและเมื่อนำแบบจำลองนั้นไปใช้ในการจำแนกข้อมูลที่เข้ามาใหม่ (unseen data) ส่งผลให้ค่าความถูกต้องที่ได้มีค่าน้อยลงอย่างมีนัยสำคัญ

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากภาพ 29 (ก) และภาพ 29 (ง) จะพบว่าสำหรับการจำแนกชุดข้อมูลการเกิดโรคของถั่วเหลืองและชุดข้อมูลผู้ป่วยไข้เลือดออกนั้น ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะมีค่าสูงกว่าค่าความถูกต้องในการจำแนกข้อมูลที่ได้จากอัลกอริทึม ID3 ในทุกชุดข้อมูลที่มีปริมาณข้อมูลที่ผิดปกติแตกต่างกัน ในขณะที่เมื่อทำการจำแนกชุดข้อมูลผู้ป่วยโรคหัวใจและชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะลดลงอย่างรวดเร็วเมื่อในชุดข้อมูลปรากฏข้อมูลรบกวนเพิ่มขึ้นร้อยละ 10 และ ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะเริ่มมีค่าน้อยกว่าค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึม ID3 เมื่อทำการจำแนกชุดข้อมูลที่ปรากฏข้อมูลที่ผิดปกติร้อยละ 20 ในชุดข้อมูลผู้ป่วยโรคหัวใจและชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 ดังแสดงในภาพ 29 (ข) และ ภาพ 29 (ค)

เมื่อพิจารณาการเปลี่ยนแปลงของค่าความถูกต้องในการจำแนกข้อมูลเมื่อภายในชุดข้อมูลมีปริมาณข้อมูลที่ผิดปกติแตกต่างกัน ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะลดลงมากกว่าค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึม ID3 ซึ่งสามารถอธิบายได้ว่าอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายนั้นจะได้รับผลกระทบจากข้อมูลที่ผิดปกติมากกว่าอัลกอริทึม ID3 เนื่องจากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายนี้ยังคงใช้ค่าเกนสารสนเทศเป็นพื้นฐานในการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับเป็นโหนดของต้นไม้ตัดสินใจ เมื่อปรากฏข้อมูลที่ผิดปกติในชุดข้อมูลจึงทำให้ค่าเกนสารสนเทศมีความคลาดเคลื่อนได้ และการนำค่าระดับความสำคัญของแอตทริบิวต์มาช่วยในการปรับปรุงค่าเกนสารสนเทศนั้นอาจไม่เพียงพอที่จะส่งผลให้สามารถเลือกแอตทริบิวต์ที่เหมาะสมสำหรับกระบวนการสร้างต้นไม้ตัดสินใจส่งผลให้ค่าความถูกต้องในการจำแนกข้อมูลลดลงเมื่อปรากฏข้อมูลที่ผิดปกติในชุดข้อมูล แต่อย่างไรก็ตามเมื่อนำอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายไปใช้ในการวิเคราะห์ชุดข้อมูลที่ไม่ปรากฏข้อมูลที่ผิดปกติจะทำให้ได้ค่าความถูกต้องในการจำแนกข้อมูลที่สูงกว่าอัลกอริทึมต้นไม้ตัดสินใจแบบเดิม ดังนั้นเพื่อให้การจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีประสิทธิภาพสูงที่สุด กระบวนการในการจัดเตรียมข้อมูลเพื่อให้ข้อมูลมีข้อมูลรบกวนน้อยที่สุดจึงเป็นประเด็นที่ต้องให้ความสำคัญสำหรับผู้ใช้งาน ดังตัวอย่างในภาพ 30 ซึ่งแสดงค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเมื่อชุดข้อมูลปรากฏข้อมูลที่ผิดปกติร้อยละ 10 และเมื่อมีการจัดเตรียมข้อมูลโดยการลบแถวข้อมูลที่ผิดปกติ จะพบว่าค่าความถูกต้องในการจำแนกข้อมูลมีค่าเพิ่มขึ้นเมื่อมีการกำจัดข้อมูลรบกวนออกจากชุดข้อมูล

4618866183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39



ภาพ 30 ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเมื่อชุดข้อมูลปรากฏข้อมูลรบกวนและเมื่อกำจัดข้อมูลรบกวน

ผลการทดสอบปรับพารามิเตอร์ที่เหมาะสมสำหรับเพิ่มประสิทธิภาพการจำแนกข้อมูล

การทดลองนี้มีวัตถุประสงค์เพื่อตรวจสอบพารามิเตอร์ที่ดีที่สุดเมื่อใช้อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายในการจำแนกชุดข้อมูลที่ทำการศึกษาทั้ง 4 ชุดข้อมูลโดยการปรับปรุงค่าความสูงของต้นไม้ตัดสินใจ ซึ่งเทคนิคการค้นหาแบบกริด (Grid search) จะถูกนำมาใช้เพื่อค้นหาความสูงของต้นไม้ที่เหมาะสม ผลของการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเมื่อมีการใช้ค่าระดับความสูงของต้นไม้ที่เหมาะสมสามารถแสดงดังตาราง 25

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

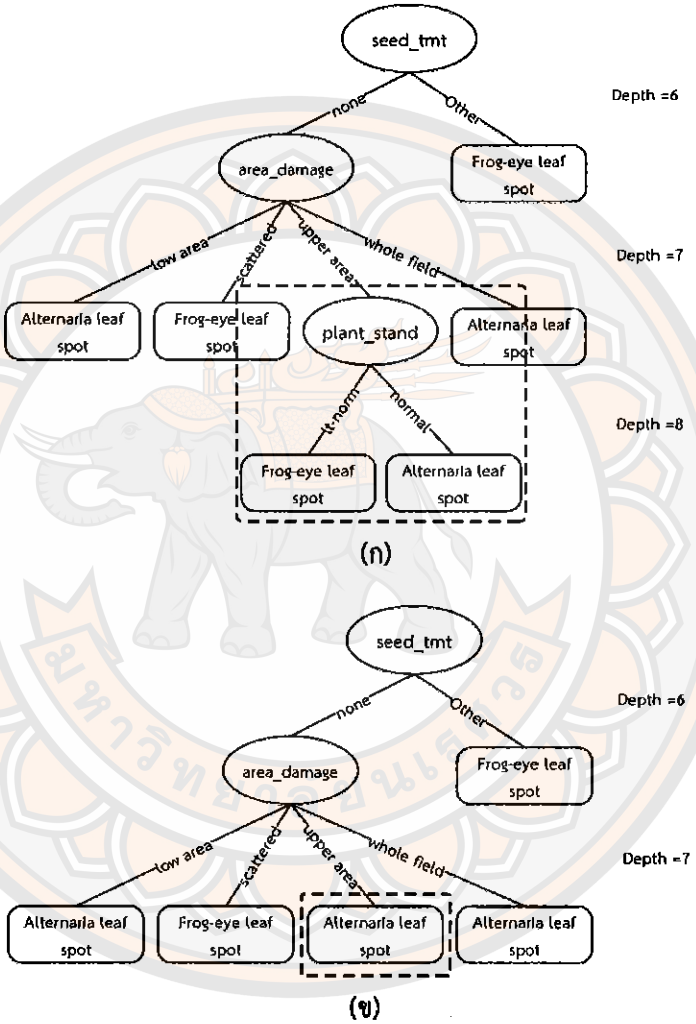
ตาราง 25 ผลการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเมื่อมีการใช้ค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสม

ชุดข้อมูล	ก่อนการปรับพารามิเตอร์			หลังปรับพารามิเตอร์			ค่าความถูกต้องที่เปลี่ยนแปลง
	ค่าความถูกต้อง	ความสูงของต้นไม้ตัดสินใจ	จำนวนโหนด	ค่าความถูกต้อง	ความสูงของต้นไม้ตัดสินใจ	จำนวนโหนด	
ชุดข้อมูลการเกิดโรคของถั่วเหลือง	92.13%	8	103	92.47%	7	94	+0.34%
ชุดข้อมูลผู้ป่วยโรคหัวใจ	77.37%	7	60	80.93%	3	34	+3.56%
ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019	88.66%	9	326	88.66%	9	326	-
ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก	89.58%	10	169	91.05%	6	72	+1.47%

จากตาราง 25 พบว่าสำหรับชุดข้อมูลการเกิดโรคของถั่วเหลืองค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสมจะมีค่าเท่ากับ 7 ซึ่งทำให้ค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้น 0.34% หรือมีค่าเท่ากับ 92.47% และจำนวนโหนดที่ใช้ในการสร้างต้นไม้ตัดสินใจลดลงจาก 103 โหนด เป็น 94 โหนด ซึ่งในการทำงานของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายนั้นอัลกอริทึมจะพยายามทำการพิจารณาแอตทริบิวต์ที่สามารถจำแนกข้อมูลออกเป็นคลาสต่าง ๆ และมีความสำคัญในแต่ละโดเมนเพื่อทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจ โดยจะดำเนินการเช่นนี้ไปเรื่อย ๆ จนกระทั่งข้อมูลที่ศึกษานั้นอยู่ในคลาสเดียวกันหรือไม่สามารถแบ่งข้อมูลได้อีก แล้วจึงสร้างต้นไม้ตัดสินใจที่มีความสูงที่มากที่สุดที่เป็นไปได้จากข้อมูลที่ได้ทำการเรียนรู้ ซึ่งการสร้างต้นไม้ตัดสินใจจนถึงความสูงที่มากที่สุดจะทำให้ต้นไม้ตัดสินใจที่ได้มีความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ (Overfitting) เมื่อนำไปจำแนกข้อมูลที่ไม่เคยเรียนรู้มาก่อนจะส่งผลให้ค่าความถูกต้องในการจำแนกข้อมูลลดลง ซึ่งการปรับพารามิเตอร์ความสูงของต้นไม้ตัดสินใจนั้นจะเป็นการควบคุมความสูงของต้นไม้ตัดสินใจเพื่อไม่ให้อัลกอริทึมทำการสร้างต้นไม้ตัดสินใจจนกระทั่งถึงความสูงที่มากที่สุด ซึ่งจะช่วยลดความจำเพาะกับข้อมูลที่ได้ทำการเรียนรู้และเมื่อนำไปใช้จำแนกข้อมูลที่ไม่เคยเรียนรู้มาก่อนจะส่งผลให้

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39
491889183

ค่าความถูกต้องในการจำแนกข้อมูลมีค่าเพิ่มขึ้น โดยแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจซึ่งอยู่ในระดับความสูงที่มากกว่าความสูงที่กำหนดจะถูกแปลงให้เป็นโหนดใบโดยมีคลาสคำตอบที่มีจำนวนตัวอย่างมากที่สุดเป็นคลาสคำตอบของโหนดใบนั้น ภาพ 31 แสดงตัวอย่างของต้นไม้ย่อยในต้นไม้ตัดสินใจเชิงความหมายเมื่อทำการปรับความสูงของต้นไม้ตัดสินใจ



ภาพ 31 ตัวอย่างต้นไม้ย่อยในต้นไม้ตัดสินใจเชิงความหมาย โดย (ก) ต้นไม้ย่อยที่มีความสูงมากที่สุด และ (ข) ต้นไม้ย่อยเมื่อมีการปรับความสูงของต้นไม้ตัดสินใจเท่ากับ 7

ภาพ 31 (ก) แสดงต้นไม้ย่อยในต้นไม้ตัดสินใจเชิงความหมายซึ่งมีความสูงเท่ากับ 8 โดยเมื่อแอตทริบิวต์ plant_stand ซึ่งหมายถึงลักษณะการยืนต้นของต้นถั่วเหลืองมีค่าน้อยกว่าปกติ (lt-norm) จะทำการจำแนกข้อมูลว่าเกิดโรคใบจุดตากบ (Frog-eye leaf spot) ในขณะที่หากลักษณะการยืนต้นของต้นถั่วเหลืองเป็นปกติ (normal) จะทำการจำแนกข้อมูลเป็นโรคใบจุดที่เกิด

4918891831 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากเชื้อ Alternaria (Alternaia leaf spot) และเมื่อทำการปรับพารามิเตอร์ความสูงของต้นไม้ตัดสินใจให้มีค่าเท่ากับ 7 นั้น แอตทริบิวต์ plant_stand (ลักษณะการยืนต้นของถั่วเหลือง) จะถูกแปลงให้เป็นโหนดใบของต้นไม้ตัดสินใจโดยมีโรคโรคใบจุดที่เกิดจากเชื้อ Alternaria ทำหน้าที่เป็นคลาสคำตอบ เนื่องจากเป็นคลาสที่มีจำนวนตัวอย่างมากที่สุด ดังแสดงในภาพ 31 (ข)

ในส่วนของการจำแนกข้อมูลผู้ป่วยโรคหัวใจนั้น ค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสมจะมีค่าเท่ากับ 3 โดยทำให้ค่าความถูกต้องในการจำแนกข้อมูลมีค่าเท่ากับ 80.93% ซึ่งมีค่าเพิ่มขึ้นจากเดิม 3.56% และจำนวนโหนดลดลงเหลือ 34 โหนด เช่นเดียวกันกับการจำแนกข้อมูลผู้ป่วยโรคไข้เลือดออกที่มีค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้น 1.47% โดยมีค่าเท่ากับ 91.05% เมื่อทำการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจที่มีความสูงเท่ากับ 6 รวมทั้งมีจำนวนโหนดที่ใช้ในการสร้างต้นไม้ตัดสินใจลดลงจาก 169 โหนด เป็น 72 โหนด

สำหรับชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 นั้น ค่าความสูงที่เหมาะสมจะมีค่าเท่ากับค่าความสูงที่สูงที่สุดของต้นไม้ตัดสินใจ ซึ่งมีค่าความสูงเท่ากับ 9 ดังนั้น ค่าความถูกต้องในการจำแนกข้อมูลที่มากที่สุดของชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 จึงมีค่าเท่ากับ 88.66% การที่ความสูงที่เหมาะสมของต้นไม้ตัดสินใจของชุดข้อมูลนี้มีค่าเท่ากับความสูงที่มากที่สุดที่เป็นไปได้จากการวิเคราะห์ข้อมูล เนื่องจากชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 นี้เป็นชุดข้อมูลที่มีแอตทริบิวต์ที่เกี่ยวข้องจำนวน 11 แอตทริบิวต์ และแต่ละแอตทริบิวต์มีค่าข้อมูลที่เกี่ยวข้องจำนวน 2 ค่า ซึ่งในการสร้างแบบจำลองเพื่อจำแนกข้อมูลออกเป็นสองคลาสตามที่กำหนดในชุดข้อมูล อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะนำแอตทริบิวต์ที่เหมาะสมและค่าข้อมูลของแอตทริบิวต์นั้น ๆ มาทำการสร้างต้นไม้ตัดสินใจ ซึ่งเมื่อทำการปรับปรุงประสิทธิภาพของแบบจำลองโดยการลดความสูงของต้นไม้ตัดสินใจ อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะทำการสร้างรูปแบบสำหรับการตัดสินใจ (decision pattern) ที่มีจำนวนแอตทริบิวต์ลดลงซึ่งแตกต่างจากรูปแบบการตัดสินใจแบบเดิมที่ได้จากต้นไม้ตัดสินใจที่มีความสูงที่มากที่สุด ซึ่งรูปแบบการตัดสินใจที่สร้างขึ้นใหม่นี้อาจไม่สามารถจำแนกข้อมูลออกเป็นคลาสที่กำหนดได้อย่างชัดเจนจึงส่งผลให้ค่าความถูกต้องในการจำแนกข้อมูลลดลง ด้วยเหตุนี้จึงทำให้ค่าความสูงที่เหมาะสมของต้นไม้ตัดสินใจมีค่าเท่ากับค่าความสูงที่มากที่สุด

จากการทดลองนี้จะพบว่าค่าความสูงของต้นไม้ตัดสินใจที่เหมาะสมสามารถเพิ่มประสิทธิภาพของต้นไม้ตัดสินใจได้ โดยทำให้ค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้น ในขณะที่จำนวนโหนดที่ใช้ในการสร้างต้นไม้ตัดสินใจลดลง ซึ่งการลดลงของความสูงของต้นไม้ตัดสินใจหรือการลดลงของจำนวนโหนดที่ใช้ในการสร้างต้นไม้ตัดสินใจจะหมายถึงความซับซ้อน (complexity) ของต้นไม้ตัดสินใจลดลงเช่นกัน เนื่องจากความสูงของต้นไม้ตัดสินใจและจำนวนโหนดภายในต้นไม้ตัดสินใจเป็น

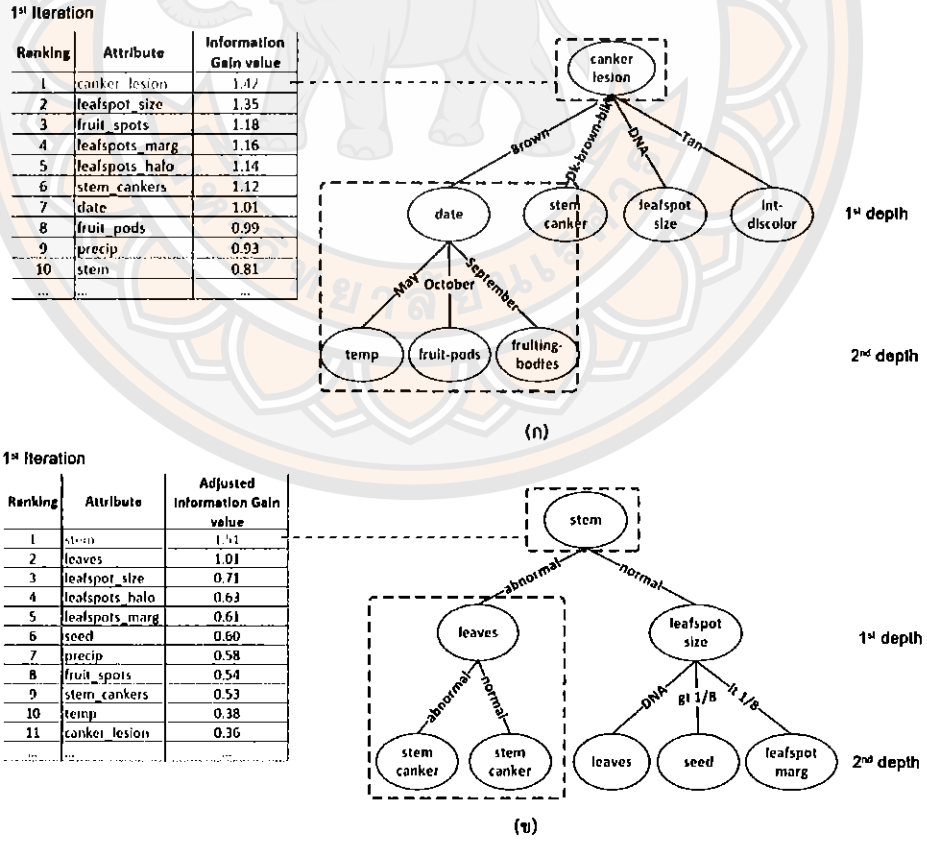
4918889183 NU ithesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

เกณฑ์หนึ่งที่ใช้ในการพิจารณาของความซับซ้อนของต้นไม้ตัดสินใจ (Kotsiantis, 2013) เมื่อต้นไม้ตัดสินใจมีความซับซ้อนลดลงจะส่งผลให้ผู้ใช้งานสามารถแปลความหมายของผลลัพธ์ได้ดียิ่งขึ้น

ผลการพิจารณาโครงสร้างของต้นไม้ตัดสินใจเชิงความหมาย

การทดลองนี้มีวัตถุประสงค์เพื่อทำการสำรวจโครงสร้างของต้นไม้ตัดสินใจเมื่อมีการประยุกต์ใช้องค์ความรู้ในออนโทโลยี โดยทำการเปรียบเทียบโครงสร้างของต้นไม้ตัดสินใจที่สร้างโดยอัลกอริทึม ID3 กับต้นไม้ตัดสินใจที่สร้างด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย

เนื่องจากต้นไม้ตัดสินใจที่สร้างขึ้นจากอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีโครงสร้างที่ซับซ้อนโดยมีโหนดภายในต้นไม้ตัดสินใจจำนวนมาก ดังนั้นในการทดลองนี้จึงนำเสนอตัวอย่างบางส่วนของต้นไม้ตัดสินใจที่ได้จากทั้งสองอัลกอริทึมเพื่อให้ทราบถึงความแตกต่างของโครงสร้างของต้นไม้ตัดสินใจเมื่อมีการนำองค์ความรู้ในออนโทโลยีมาช่วยในกระบวนการพิจารณาโหนดของต้นไม้ตัดสินใจ ตัวอย่างโครงสร้างต้นไม้ตัดสินใจจากอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายแสดงดังภาพ 32 และ ภาพ 34



ภาพ 32 ตัวอย่างโครงสร้างต้นไม้ตัดสินใจสำหรับชุดข้อมูลการเกิดโรคของถั่วเหลือง โดย (ก) อัลกอริทึม ID3 และ (ข) อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย

NU Thesis 6003257 thesis / rec: 31102565 16:06:32 / seq: 39
4918891B3

ภาพ 32 แสดงตัวอย่างของต้นไม้ตัดสินใจสำหรับชุดข้อมูลการเกิดโรคของถั่วเหลือง โดยภาพ 32 (ก) คือ ส่วนหนึ่งของต้นไม้ตัดสินใจที่สร้างโดยใช้อัลกอริทึม ID3 และ ภาพ 32 (ข) คือ ส่วนหนึ่งของต้นไม้ตัดสินใจที่สร้างโดยใช้อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย ซึ่งจากภาพจะพบว่าต้นไม้ตัดสินใจที่ได้มีโหนดรากที่แตกต่างกัน โดยต้นไม้ตัดสินใจที่สร้างด้วยอัลกอริทึม ID3 แอตทริบิวต์ canker lesion (สีของแผลบนลำต้น) ทำหน้าที่เป็นโหนดราก เนื่องจากเป็นแอตทริบิวต์ที่มีค่าเกนสารสนเทศสูงที่สุด ในขณะที่เมื่อทำการสร้างแบบจำลองการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย แอตทริบิวต์ stem (อาการของลำต้น) ซึ่งเป็นแอตทริบิวต์ที่มีค่าเกนสารสนเทศที่ได้รับการปรับปรุงสูงที่สุดทำหน้าที่เป็นโหนดรากของต้นไม้ตัดสินใจ จากภาพ 32 (ก) จะพบว่าแอตทริบิวต์ stem นี้มีค่าเกนสารสนเทศอยู่ในอันดับที่ 10 และหากใช้ค่าเกนสารสนเทศในการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับใช้เป็นโหนดของต้นไม้ตัดสินใจแอตทริบิวต์ stem จะไม่ถูกเลือก ถึงแม้ว่าจะเป็นแอตทริบิวต์ที่มีความสำคัญในอับดับที่ 2 (ตาราง 14) และมีความสำคัญมากกว่าแอตทริบิวต์ canker-lesion ก็ตาม นอกจากนี้เมื่อพิจารณาค่าข้อมูลของแอตทริบิวต์จะพบว่าแอตทริบิวต์ canker-lesion เป็นแอตทริบิวต์ที่มีค่าข้อมูลที่แตกต่างกันจำนวน 4 ค่า ในขณะที่แอตทริบิวต์ stem มีจำนวนค่าของแอตทริบิวต์เพียง 2 ค่า ซึ่งแสดงให้เห็นว่าอาจเกิดปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายเมื่อใช้ค่าเกนสารสนเทศในการพิจารณาโหนดสำหรับต้นไม้ตัดสินใจ เนื่องจากในการสร้างต้นไม้ตัดสินใจนั้นจะมีการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับเป็นโหนดของต้นไม้ตัดสินใจโดยการใช้ค่าเอนโทรปี ที่จะมีการพิจารณาว่ามีการปะปนกันของข้อมูลในแต่คลาสมากน้อยเพียงใด โดยพิจารณาจากสัดส่วนของข้อมูลที่ปรากฏในแต่ละคลาส หากแอตทริบิวต์มีจำนวนค่าข้อมูลที่แตกต่างหลายค่า จะส่งผลให้ชุดข้อมูลถูกแบ่งออกเป็นหลาย ๆ กลุ่มตามจำนวนค่าของแอตทริบิวต์นั้น ๆ ซึ่งอาจทำให้มีการปะปนกันของข้อมูลในแต่ละคลาสลดลงเมื่อพิจารณาจากข้อมูลในแต่ละกลุ่ม จึงทำให้ค่าเอนโทรปีที่ได้มีค่าน้อยและส่งผลให้ค่าเกนสารสนเทศมีค่ามากกว่าแอตทริบิวต์ที่มีจำนวนค่าข้อมูลน้อยกว่า ดังตัวอย่างในภาพ 33

ID	Age1	Age2	Class
1	10 - 19	10 - 14	Y
2	10 - 19	15 - 19	N
3	10 - 19	15 - 19	N
4	10 - 19	15 - 19	Y
5	20 - 29	20 - 24	Y
6	20 - 29	20 - 24	Y

Entropy ของ Age1

$$= \frac{4}{6} \times (-(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4})) + \frac{2}{6} \times (-(\frac{2}{2}\log_2\frac{2}{2}))$$
$$= 0.67$$

Entropy ของ Age2

$$= \frac{1}{6} \times (-(\frac{1}{1}\log_2\frac{1}{1})) + \frac{3}{6} \times (-(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3})) + \frac{2}{6} \times (-(\frac{2}{2}\log_2\frac{2}{2}))$$
$$= 0.46$$

ภาพ 33 ค่าเอนโทรปีของแอตทริบิวต์เมื่อมีจำนวนค่าข้อมูลในแอตทริบิวต์แตกต่างกัน

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากภาพ 33 จะเห็นได้ว่าแอตทริบิวต์ Age1 ซึ่งเกี่ยวข้องกับข้อมูลอายุของคน โดยจะมีการจัดกลุ่มข้อมูลออกเป็น 2 กลุ่มที่แตกต่างกันจะมีค่าเอนโทรปีเท่ากับ 0.67 ในขณะที่แอตทริบิวต์ Age2 ซึ่งเป็นข้อมูลอายุของคนคนเดียวกันแต่มีการจัดกลุ่มอายุออกเป็น 3 กลุ่ม จะมีค่าเอนโทรปีลดลงโดยมีค่าเอนโทรปีเท่ากับ 0.46 ดังนั้นแอตทริบิวต์ Age2 ซึ่งมีค่าข้อมูลในแอตทริบิวต์จำนวน 3 ค่าจึงมีค่าเกนสารสนเทศสูงกว่าแอตทริบิวต์ Age1 ซึ่งมีค่าข้อมูลที่แตกต่างกันจำนวน 2 ค่า และส่งผลให้แอตทริบิวต์ Age2 มีโอกาสในการถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากกว่าแอตทริบิวต์ Age1

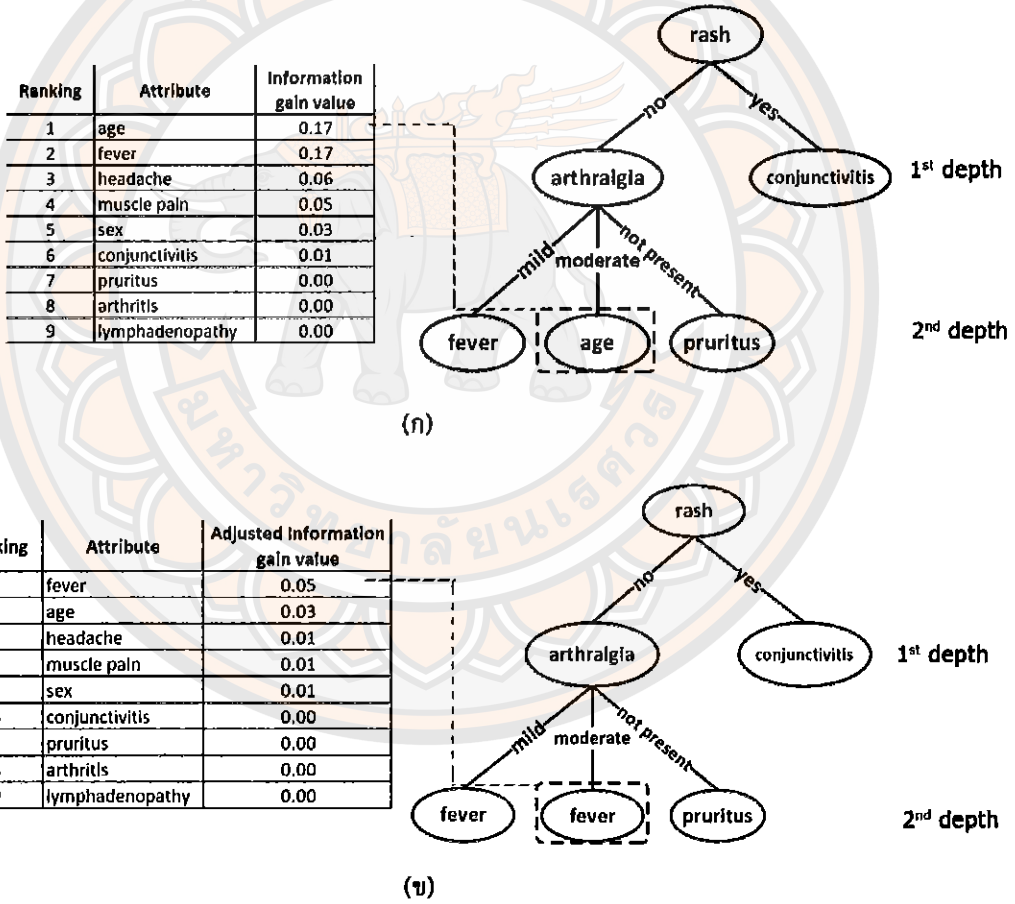
ดังนั้นการนำค่าระดับความสำคัญของแอตทริบิวต์ที่อ้างอิงได้จากออนโทโลยีมาช่วยปรับปรุงค่าเกนสารสนเทศนั้นจะช่วยในการลดปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายได้ โดยแอตทริบิวต์ที่มีจำนวนค่าของแอตทริบิวต์น้อยและมีความสำคัญในโดเมนมีโอกาสที่จะถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากขึ้น

ในความเป็นจริงการพิจารณาการเกิดโรคในพืชนั้นผู้เชี่ยวชาญจะพิจารณาจากอาการที่ปรากฏบนส่วนต่าง ๆ ของพืชเพื่อระบุโรคที่เกิดขึ้น และเมื่อพืชปรากฏอาการที่มีลักษณะใกล้เคียงกัน รูปแบบของอาการที่เกิดขึ้นร่วมกันรวมถึงสภาพแวดล้อมจะถูกนำมาพิจารณาเพื่อระบุโรคที่ชัดเจน (Grogan, 1981) เช่น โรคขอบใบแห้ง (Bacterial Blight) และ โรคใบจุดสีน้ำตาล (Brown spot) เป็นโรคพืชที่มีอาการใกล้เคียงกัน โดยเป็นโรคที่ทำให้ใบของต้นถั่วเหลืองเกิดจุดสีน้ำตาลเช่นเดียวกัน และมักเกิดความสับสนในการจำแนกโรค (Yang, 2000) อย่างไรก็ตามทั้งสองโรคจะมีอาการที่เกิดร่วมกับโรคนั้น ๆ แตกต่างซึ่งสามารถใช้ในการระบุโรคที่เกิดขึ้นได้ เช่น โรคขอบใบแห้งจะไม่ปรากฏอาการใบร่วง ในขณะที่โรคใบจุดสีน้ำตาลจะปรากฏอาการใบร่วง นอกจากนี้ข้อมูลเกี่ยวกับวิธีการปลูกถั่วเหลืองยังสามารถนำมาใช้ประกอบการระบุโรคที่เกิดขึ้นได้อีกด้วย โดยการปลูกต้นถั่วเหลืองติดต่อกันมักจะทำให้เกิดโรคใบจุดสีน้ำตาล เป็นต้น เมื่อพิจารณากฎการตัดสินใจที่ได้จากต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะมีลำดับการพิจารณาปัจจัยต่าง ๆ ที่ส่งผลกับการเกิดโรคใกล้เคียงกับการพิจารณาของผู้เชี่ยวชาญมากกว่ากฎการตัดสินใจที่ได้จากอัลกอริทึม ID3 ดังจะเห็นได้จากภาพ 32 (ก) แอตทริบิวต์ date ซึ่งเกี่ยวข้องกับช่วงเวลาที่เกิดโรคในถั่วเหลืองซึ่งเป็นปัจจัยภายนอกที่เกี่ยวข้องปรากฏอยู่ในต้นไม้ตัดสินใจในความสูงระดับที่ 1 แล้วจึงมีการพิจารณาอาการที่เกิดขึ้นกับส่วนอื่น ๆ ของถั่วเหลือง เช่น fruit-pods (อาการที่เกิดขึ้นบนฝักถั่วเหลือง) และ fruiting bodies (การปรากฏฟรุตติ้งบอดี้ของเชื้อราบนต้นถั่วเหลือง) รวมถึงปัจจัยภายนอกอื่น ๆ เช่น แอตทริบิวต์ temp หรือ อุณหภูมิ เมื่อต้นไม้ตัดสินใจมีความสูงในระดับที่ 2 ในขณะที่ต้นไม้ตัดสินใจที่สร้างโดยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย (ภาพ 32 (ข)) เมื่อต้นไม้ตัดสินใจมีความสูงระดับที่ 1 แอตทริบิวต์ leaves ซึ่งหมายถึง ส่วนใบของถั่วเหลืองที่มักปรากฏอาการของโรคพืช และตามด้วยแอตทริบิวต์ stem canker ซึ่งหมายถึง การเกิดแผลบนลำต้นของพืชซึ่งเป็นอาการหนึ่งที่แสดงให้เห็นว่าเกิดความผิดปกติกับต้นถั่วเหลืองได้ถูกนำมาใช้เมื่อต้นไม้ตัดสินใจมี

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ความสูงในระดับที่ 2 โดยแอตทริบิวต์ที่เป็นปัจจัยภายนอกอื่น ๆ จะถูกนำมาใช้ในการพิจารณาการเกิดโรคของถั่วเหลืองเมื่อต้นไม้ตัดสินใจมีความลึกมากขึ้น ดังนั้นจึงอาจกล่าวได้ว่า รูปแบบของกฎการตัดสินใจที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายนี้มีความคล้ายคลึงกับการพิจารณาการเกิดโรคของถั่วเหลืองโดยผู้เชี่ยวชาญมากขึ้นซึ่งช่วยให้ผู้ใช้งานสามารถทำความเข้าใจผลลัพธ์ได้ง่ายขึ้น

อีกหนึ่งตัวอย่างที่แสดงความแตกต่างของโครงสร้างต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 และอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายของชุดข้อมูลผู้ป่วยโรคไข้เลือดออกสามารถแสดงดังภาพ 34



ภาพ 34 ตัวอย่างโครงสร้างต้นไม้ตัดสินใจสำหรับชุดข้อมูลผู้ป่วยโรคไข้เลือดออก โดย (ก) อัลกอริทึม ID3 และ(ข) อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย

ภาพ 34 แสดงบางส่วนของต้นไม้ตัดสินใจของชุดข้อมูลผู้ป่วยโรคไข้เลือดออก โดยภาพ 34 (ก) คือ ต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 และ ภาพ 34 (ข) คือ ต้นไม้ตัดสินใจที่ได้

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

จากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย จากภาพ 34 จะพบว่าโครงสร้างของต้นไม้ตัดสินใจที่ได้จากอัลกอริทึม ID3 และ อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายแตกต่างกันในบางจุด เนื่องจากค่าระดับความสำคัญของแอตทริบิวต์ส่วนใหญ่ในชุดข้อมูลมีค่าใกล้เคียงกัน (ตาราง 17) เมื่อนำค่าระดับความสำคัญเหล่านี้ไปใช้ในการปรับปรุงค่าเกนสารสนเทศจึงทำให้อันดับของแอตทริบิวต์ที่เหมาะสมสำหรับนำไปใช้เป็นโหนดของต้นไม้ตัดสินใจใกล้เคียงกับอันดับของแอตทริบิวต์ที่ได้จากวิธีการเดิม อย่างไรก็ตามค่าระดับความสำคัญของแอตทริบิวต์ที่ได้จากออนโทโลยีสามารถช่วยในการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับใช้เป็นโหนดของต้นไม้ตัดสินใจได้ โดยจากภาพ 34 (ก) เมื่อต้นไม้ตัดสินใจมีความสูงระดับ 2 เมื่อพิจารณาค่าเกนสารสนเทศที่ได้จากชุดข้อมูลจะพบว่า แอตทริบิวต์ age ซึ่งแสดงอายุของผู้ป่วย และ แอตทริบิวต์ fever ที่แสดงอาการมีไข้ มีค่าเกนสารสนเทศเท่ากัน คือ 0.17 สำหรับอัลกอริทึม ID3 นั้น เมื่อปรากฏแอตทริบิวต์ที่มีค่าเกนสารสนเทศเท่ากัน อัลกอริทึมจะทำการสุ่มเลือกแอตทริบิวต์ใดแอตทริบิวต์หนึ่งมาทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจ ซึ่งในตัวอย่างนี้อัลกอริทึมทำการสุ่มเลือกแอตทริบวิต์ age มาเป็นโหนดของต้นไม้ตัดสินใจ ดังนั้นวิธีการสุ่มเลือกอาจจะเป็นวิธีที่ไม่มีประสิทธิภาพมากนัก ซึ่งสำหรับอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายนั้นเมื่อค่าเกนสารสนเทศของแอตทริบิวต์มีค่าเท่ากัน ค่าระดับความสำคัญของแอตทริบิวต์ที่นำมาปรับปรุงค่าเกนสารสนเทศจะช่วยเลือกแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจได้ดีขึ้น ดังนั้น แอตทริบิวต์ fever ซึ่งมีค่าระดับความสำคัญสูงกว่าแอตทริบิวต์ age จึงถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจดังแสดงในภาพ 34 (ข) ซึ่งวิธีการดังกล่าวส่งผลให้กำการตัดสินใจที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีความใกล้เคียงกับการวินิจฉัยโรคโดยผู้เชี่ยวชาญที่มีการพิจารณาอาการต่าง ๆ ที่เกิดขึ้นเพื่อระบุโรคที่เกิดขึ้นกับผู้ป่วย (Tang & Ooi, 2012) ในขณะที่ข้อมูลอื่น ๆ ของผู้ป่วย เช่น อายุ และ เพศ จะเป็นปัจจัยเสี่ยงที่ส่งผลต่อความรุนแรงของโรค (Vicente et al., 2017)

จากภาพ 32 และ ภาพ 34 จะพบว่าแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดใด ๆ ภายในต้นไม้ตัดสินใจสามารถมีแอตบิวต์เดียวกันทำหน้าที่เป็นโหนดตัดสินใจต่อจากโหนดนั้นได้ เช่น ภาพ 32 (ข) แอตทริบิวต์ leaves (ลักษณะของใบ) จะมีแอตทริบิวต์ stem canker (การเกิดแผลบนลำต้น) ทำหน้าที่เป็นโหนดตัดสินใจต่อจากแอตทริบิวต์ leaves จำนวน 2 โหนด รวมถึงภาพ 34 (ข) แอตทริบิวต์ arthralgia (อาการปวดข้อ) มีแอตทริบิวต์ pruritus (อาการคัน) จำนวน 1 โหนด และแอตทริบิวต์ fever (อาการไข้) จำนวน 2 โหนด ทำหน้าที่เป็นโหนดตัดสินใจต่อจากแอตทริบิวต์ arthralgia เป็นต้น เนื่องจากการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับเป็นโหนดของต้นไม้ตัดสินใจนั้น จะพิจารณาจากแอตทริบิวต์ที่มีค่าของเกณฑ์การพิจารณาโหนดที่ดีที่สุด ณ ชุดข้อมูลย่อยของแต่ละรอบการทำงาน ดังนั้นแอตทริบิวต์ที่มีค่าเกนสารสนเทศที่ปรับปรุงที่ดีที่สุดในการทำงานแต่ละ

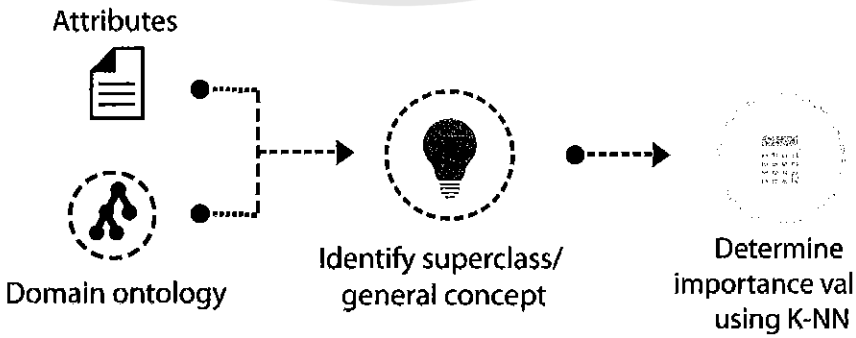
4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ครั้งจึงถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจ จึงทำให้บางครั้งปรากฏแอตทริบิวต์เดียวกันทำหน้าที่เป็นโหนดตัดสินใจต่อจากแอตทริบิวต์ใด ๆ ได้

จากผลการพิจารณาโครงสร้างของต้นไม้ตัดสินใจสามารถสรุปได้ว่า การนำองค์ความรู้ในออนโทโลยีซึ่งอยู่ในรูปแบบของค่าระดับความสำคัญของแอตทริบิวต์มาใช้ในการปรับปรุงค่าเกนสารสนเทศสามารถช่วยปรับปรุงประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจได้ โดยช่วยให้กฎการตัดสินใจที่ได้มีความใกล้เคียงกับการพิจารณาของผู้เชี่ยวชาญซึ่งช่วยให้ผู้ใช้งานสามารถทำความเข้าใจผลลัพธ์ได้ดีขึ้น นอกจากนี้ค่าระดับความสำคัญของแอตทริบิวต์ที่ได้จากโครงสร้างและความสัมพันธ์ของแนวความคิดในออนโทโลยีสามารถช่วยลดปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายซึ่งส่งผลต่อประสิทธิภาพในการจำแนกข้อมูลเมื่อแอตทริบิวต์ที่ไม่มีความสำคัญถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจ โดยแอตทริบิวต์ที่จำนวนค่าของแอตทริบิวต์น้อยแต่มีความสำคัญในโดเมนมีโอกาสได้รับเลือกเป็นโหนดของต้นไม้ตัดสินใจมากขึ้น

ผลการประมาณค่าระดับความสำคัญเมื่อไม่ปรากฏองค์ความรู้ในออนโทโลยี

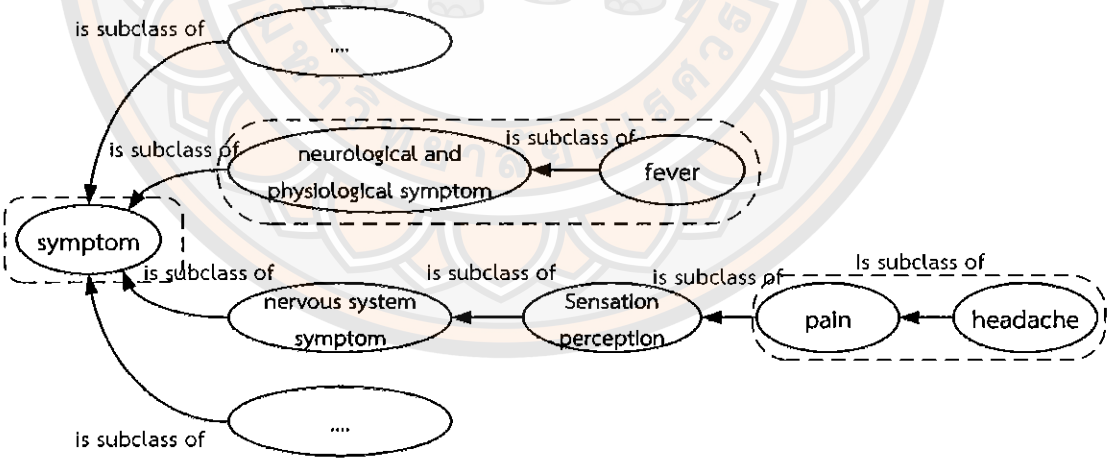
การทดลองนี้มีวัตถุประสงค์เพื่อทดสอบการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อแอตทริบิวต์นั้นไม่ปรากฏเป็นองค์ความรู้ในออนโทโลยีและผลลัพธ์ที่ได้จากการจำแนกข้อมูลเมื่อมีการนำค่าระดับความสำคัญนั้นไปใช้ในการจำแนกข้อมูล โดยในการประมาณค่าระดับความสำคัญของแอตทริบิวต์ที่ไม่ปรากฏในออนโทโลยีจะประยุกต์ใช้แนวความคิดในการเติมค่าข้อมูลที่สูญหาย (Missing value imputation) ด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด (k-NN) ในการดำเนินการ โดยค่าระดับความสำคัญของแอตทริบิวต์ที่ไม่ปรากฏในออนโทโลยีนั้นจะถูกพิจารณาเป็นข้อมูลสูญหายและถูกประมาณค่าด้วยค่าระดับความสำคัญของแอตทริบิวต์อื่น ๆ ที่มีความคล้ายคลึงกันจำนวน k ค่า ซึ่งสามารถแสดงขั้นตอนการทำงานได้ดังภาพ 35



ภาพ 35 ขั้นตอนการประมาณค่าระดับความสำคัญสำหรับแอตทริบิวต์ที่ไม่ปรากฏในออนโทโลยี

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

การประมาณค่าระดับความสำคัญของแอตทริบิวต์ที่ไม่ปรากฏในออนโทโลยีนั้นจะดำเนินการโดยการพิจารณาความสัมพันธ์ของแนวความคิดในออนโทโลยี ซึ่งแนวความคิดที่ทำหน้าที่เป็นคลาสแม่ (Superclass) ของแต่ละแอตทริบิวต์ และแนวความความคิดที่เป็นแนวความคิดทั่วไปของแอตทริบิวต์นั้น ๆ จะถูกนำไปใช้ในการพิจารณาความคล้ายคลึงกันของแต่ละแอตทริบิวต์ ซึ่งการพิจารณาแนวความคิดทั่วไปของแต่ละแอตทริบิวต์จะขึ้นอยู่กับความต้องการของผู้ใช้แต่ละคน เช่น ในภาพ 36 เป็นการพิจารณาความสัมพันธ์ระหว่างข้อมูลของแอตทริบิวต์ fever ซึ่งหมายถึงอาการไข้ และแอตทริบิวต์ headache ที่แสดงอาการปวดศีรษะในออนโทโลยีโรคไข้เลือดออก เมื่อพิจารณาคลาสแม่ (superclass) ที่เกี่ยวข้องกับแอตทริบิวต์ fever หรืออาการไข้ จะพบว่าแนวความคิด neurological and physiological symptom ซึ่งหมายถึงอาการที่เกิดขึ้นกับระบบประสาทและร่างกายทำหน้าที่เป็นคลาสแม่ และเมื่อพิจารณาคลาสแม่ของแอตทริบิวต์ headache หรืออาการปวดศีรษะนั้น จะพบแนวความคิด pain หรืออาการเจ็บปวดทำหน้าที่เป็นคลาสแม่ สำหรับแนวความคิดที่เป็นแนวความคิดทั่วไปของทั้งสองแอตทริบิวต์นี้จะพิจารณาเลือกแนวความคิด symptom ที่หมายถึงอาการป่วย ซึ่งเป็นแนวความคิดทั่วไปที่มีความสัมพันธ์กับทั้งสองแอตทริบิวต์นี้ รวมถึงมีความสัมพันธ์กับแอตทริบิวต์อื่น ๆ ในชุดข้อมูล



ภาพ 36 ตัวอย่างการพิจารณาแนวความคิดในออนโทโลยีโรคไข้เลือดออกสำหรับการประมาณค่าระดับความสำคัญของแอตทริบิวต์

หลังจากการพิจารณาแนวความคิดที่ทำหน้าที่เป็นคลาสแม่และแนวความคิดทั่วไปของแต่ละแอตทริบิวต์แล้ว ข้อมูลเหล่านี้จะถูกนำมาใช้ในการประมาณค่าระดับความสำคัญของแอตทริบิวต์

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ที่ไม่ปรากฏในออนโทโลยีด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด ซึ่งจะทำการคำนวณค่าความคล้ายคลึงของแอตทริบิวต์ด้วยวิธีการวัดระยะห่างแบบยุคลิด (Euclidean distance) ดังสมการ (19)

$$dis(x_a, x_b) = \sqrt{\sum_{i=1}^{n} (x_{a,i} - x_{b,i})^2} \tag{19}$$

โดย $dis(x_a, x_b)$ คือ ระยะห่างระหว่างข้อมูล x_a และ x_b

n คือ จำนวนค่าคุณสมบัติทั้งหมดที่เกี่ยวข้องกับข้อมูล

$x_{a,i}$ คือ ค่าของคุณสมบัติ i ของข้อมูล x_a

$x_{b,i}$ คือ ค่าของคุณสมบัติ i ของข้อมูล x_b

ตัวอย่างเช่น ในกรณีที่แอตทริบิวต์ lymphadenopathy ซึ่งหมายถึงอาการต่อมน้ำเหลืองโตในชุดข้อมูลผู้ป่วยโรคไข้เลือดออกไม่ปรากฏเป็นแนวความคิดในออนโทโลยีโรคไข้เลือดออก โดยผู้ใช้ทราบเพียงว่าแอตทริบิวต์นี้เป็นอาการหนึ่งที่อาจปรากฏในผู้ป่วยโรคไข้เลือดออก และมีตัวอย่างการพิจารณาคลาสแม่และแนวความคิดทั่วไปที่เกี่ยวข้องกับแอตทริบิวต์ที่เกี่ยวข้องดังตาราง 26 จะสามารถคำนวณค่าระยะห่างระหว่างแอตทริบิวต์ fever และ lymphadenopathy ได้ดังต่อไปนี้

ตาราง 26 ตัวอย่างผลการพิจารณาคลาสแม่และแนวความคิดทั่วไปของแอตทริบิวต์ต่าง ๆ ในชุดข้อมูลผู้ป่วยโรคไข้เลือดออก

แอตทริบิวต์	symptom	neurological and physiological symptom	Person attribute	pain	skin and integumentary tissue symptom	Other symptom
fever	1	1	0	0	0	0
lymphadenopathy	1	0	0	0	0	0

$$dis(lymphadenopathy, fever) = \sqrt{(1-1)^2 + (0-1)^2 + 0 + 0 + 0 + 0}$$
$$= 1$$

จากผลการคำนวณจะพบว่าแอตทริบิวต์ lymphadenopathy และแอตทริบิวต์ fever มีค่าระยะห่างระหว่างข้อมูลเท่ากับ 1 ซึ่งจะดำเนินการเช่นนี้กับทุกแอตทริบิวต์เพื่อระบุค่าระยะห่างระหว่างแอตทริบิวต์ lymphadenopathy กับแอตทริบิวต์อื่น ๆ จนครบทุกแอตทริบิวต์ หากแอตทริบิวต์ lymphadenopathy มีค่าระยะห่างกับแอตทริบิวต์ใด ๆ น้อยที่สุดจะหมายถึง

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

แอตทริบิวต์ lymphadenopathy มีความคล้ายคลึงกับแอตทริบิวต์นั้นมากที่สุด ซึ่งเทคนิคเพื่อนบ้านใกล้ที่สุดจะทำการพิจารณาแอตทริบิวต์ที่มีระยะห่างกับแอตทริบิวต์ lymphadenopathy น้อยที่สุดจำนวน k แอตทริบิวต์ ดังนั้นจะสามารถทำการประมาณค่าระดับความสำคัญของแอตทริบิวต์ lymphadenopathy ได้จากค่าระดับความสำคัญของแอตทริบิวต์ที่มีความคล้ายคลึงเหล่านั้น

ในการวิจัยนี้ได้ทำการทดลองเพื่อหาค่า k ที่เหมาะสมสำหรับการประมาณค่าระดับความสำคัญของแอตทริบิวต์ในแต่ละชุดข้อมูล โดยกำหนดให้ k มีค่าเท่ากับ 3, 5 และ 7 ตามลำดับ ซึ่งผลการทดลองในการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อแต่ละแอตทริบิวต์เป็นแอตทริบิวต์ที่ไม่ปรากฏเป็นองค์ความรู้ในออนโทโลยีด้วยค่า k ที่แตกต่างกันมีค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (Mean Absolute Error: MAE) ดังตาราง 27

ตาราง 27 ค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ยจากการประมาณค่าระดับความสำคัญของแอตทริบิวต์

ชุดข้อมูล	ค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ย (MAE)		
	k=3	k=5	k=7
ชุดข้อมูลการเกิดโรคของถั่วเหลือง	0.1610	**0.1377**	0.1471
ชุดข้อมูลผู้ป่วยโรคหัวใจ	0.0718	0.0700	**0.0664**
ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019	0.0770	**0.0710**	0.0790
ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก	0.0155	0.0145	**0.0136**

จากตาราง 27 พบว่าสำหรับค่า k ที่เหมาะสมสำหรับการประมาณค่าระดับความสำคัญของแอตทริบิวต์ด้วยเทคนิคเพื่อนบ้านใกล้ที่สุดสำหรับชุดข้อมูลโรคของถั่วเหลืองและชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 คือ 5 เนื่องจากทำให้ค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ยน้อยที่สุด สำหรับชุดข้อมูลผู้ป่วยโรคหัวใจและชุดข้อมูลผู้ป่วยโรคไข้เลือดออกนั้นจะมีค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ยน้อยที่สุดเมื่อ k มีค่าเท่ากับ 7

ทำการทดสอบประมาณค่าระดับความสำคัญของแอตทริบิวต์โดยทำการสุ่มเลือกแอตทริบิวต์จำนวนร้อยละ 10 ร้อยละ 20 และ ร้อยละ 30 ของแอตทริบิวต์ในชุดข้อมูลให้เป็นแอตทริบิวต์ที่ไม่ปรากฏแนวความคิดในออนโทโลยีตามลำดับ สำหรับการสุ่มแอตทริบิวต์ที่มีค่าระดับความสำคัญสูญหายจะมีจำนวนแอตทริบิวต์ในแต่ละชุดข้อมูลดังตาราง 28

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ตาราง 28 จำนวนแอตทริบิวต์ที่ใช้สำหรับทดสอบการประมาณค่าระดับความสำคัญที่สูญหาย

ชุดข้อมูล	จำนวนแอตทริบิวต์ที่ค่าระดับความสำคัญสูญหาย (แอตทริบิวต์)		
	ร้อยละ 10	ร้อยละ 20	ร้อยละ 30
ชุดข้อมูลการเกิดโรคของถั่วเหลือง	3	6	9
ชุดข้อมูลผู้ป่วยโรคหัวใจ	1	2	3
ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019	1	2	3
ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก	1	2	3

ผลการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 10 ของจำนวนแอตทริบิวต์ในชุดข้อมูลสามารถแสดงดังตาราง 29

ตาราง 29 ผลการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 10

ชุดข้อมูล	แอตทริบิวต์	ค่าระดับความสำคัญ		MAE
		ค่าจริง	ค่าที่ประมาณด้วย k-NN	
ชุดข้อมูลการเกิดโรคของถั่วเหลือง	fruit_spots	0.46	0.46	0.04
	fruiting_bodies	0.21	0.21	
	seed_tmt	0.15	0.28	
ชุดข้อมูลผู้ป่วยโรคหัวใจ	Oldpeak	0.15	0.17	0.02
ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019	Fever	0.15	0.18	0.03
ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก	rash	0.15	0.15	0.00

จากตาราง 29 จะพบว่าเมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 10 ของจำนวนแอตทริบิวต์ในชุดข้อมูล เมื่อใช้เทคนิคเพื่อนบ้านใกล้ที่สุดประมาณค่าระดับความสำคัญของแอตทริบิวต์สำหรับชุดข้อมูลผู้ป่วยโรคหัวใจ ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออกมีค่าใกล้เคียงกับค่าจริงของแต่ละแอตทริบิวต์ โดยมีค่าความ

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

คลาดเคลื่อนสมบูรณ์เฉลี่ย (MAE) อยู่ระหว่าง 0.00 - 0.03 เนื่องจากแนวความคิดในออนโทโลยีที่เกี่ยวข้องกับชุดข้อมูลเหล่านี้ส่วนใหญ่มีความสัมพันธ์แบบลำดับชั้น เช่น Subclass-of ซึ่งเมื่อมีการนำข้อมูลแนวความคิดที่เป็นคลาสแม่และแนวความคิดทั่วไปที่เกี่ยวข้องกับแต่ละแอตทริบิวต์มาใช้เป็นส่วนประกอบในการประมาณค่าระดับความสำคัญของแอตทริบิวต์จึงทำให้ค่าระดับความสำคัญที่ประมาณค่าได้มีค่าใกล้เคียงกับค่าจริง สำหรับชุดข้อมูลการเกิดโรคของถั่วเหลืองนั้นจะมีค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ยเท่ากับ 0.04 โดยเมื่อพิจารณาแอตทริบิวต์ seed_tmt ซึ่งหมายถึงขั้นตอนการเตรียมเมล็ด จะพบว่าค่าระดับความสำคัญของแอตทริบิวต์ที่ได้จากเทคนิคการประมาณค่านี้แตกต่างจากค่าจริง ซึ่งเกิดจากการความสัมพันธ์ของแนวความคิดในออนโทโลยีโรคของถั่วเหลืองนั้นประกอบไปด้วยความสัมพันธ์แบบลำดับชั้นและความสัมพันธ์อื่น ๆ เช่น has-symptom ซึ่งเป็นความสัมพันธ์ที่แสดงความสัมพันธ์ระหว่างโรคและอาการผิดปกติต่าง ๆ ที่เกิดขึ้นกับต้นถั่วเหลือง เมื่อนำเพียงข้อมูลคลาสแม่และแนวความคิดทั่วไปที่เกี่ยวข้องมาใช้ในการประมาณค่าระดับความสำคัญจึงอาจยังไม่เพียงพอในการระบุความคล้ายคลึงระหว่างแอตทริบิวต์และส่งผลให้การประมาณค่าระดับความสำคัญของแอตทริบิวต์มีความคลาดเคลื่อนมากกว่าในชุดข้อมูลอื่น ๆ

สำหรับการประมาณการค่าระดับความสำคัญของแอตทริบิวต์เมื่อค่าระดับความสำคัญสูญหายร้อยละ 20 และ ร้อยละ 30 นั้น สามารถแสดงดังตาราง 30 และตาราง 31 ตามลำดับ

ตาราง 30 ผลการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 20

ชุดข้อมูล	แอตทริบิวต์	ค่าระดับความสำคัญ		MAE
		ค่าจริง	ค่าที่ประมาณด้วย k-NN	
ชุดข้อมูลการเกิดโรคของถั่วเหลือง	fruit_spots	0.46	0.30	0.09
	fruiting_bodies	0.21	0.23	
	seed_tmt	0.15	0.28	
	temp	0.6	0.62	
	canker_lesion	0.26	0.23	
	External_decay	0.42	0.23	
ชุดข้อมูลผู้ป่วยโรคหัวใจ	Oldpeak	0.15	0.17	0.03
	Restecg	0.15	0.19	
ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019	Fever	0.15	0.19	0.05
	Taste disorder	0.15	0.20	
ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก	rash	0.15	0.15	0.01
	lymphadenopathy	0.15	0.17	

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39
4918891B3

ตาราง 31 ผลการประมาณค่าระดับความสำคัญของแอตทริบิวต์เมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 30

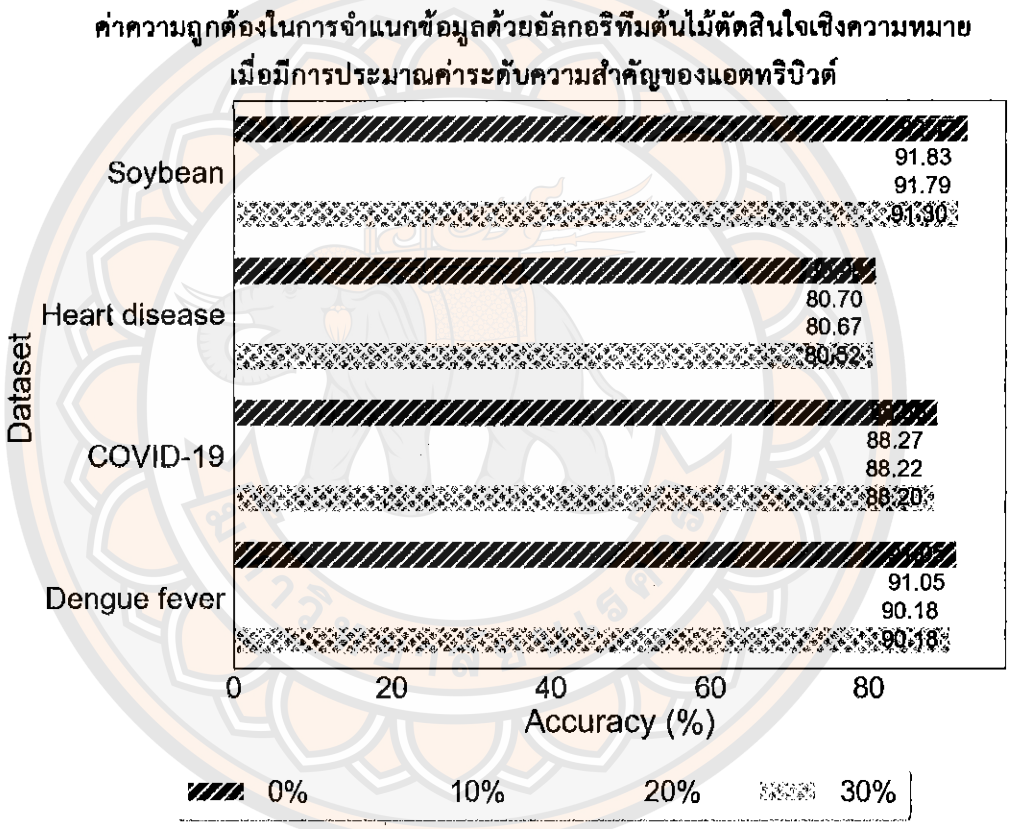
ชุดข้อมูล	แอตทริบิวต์	ค่าระดับความสำคัญ		MAE
		ค่าจริง	ค่าที่ประมาณด้วย k-NN	
ชุดข้อมูลการเกิดโรคของถั่วเหลือง	fruit_spots	0.46	0.30	0.09
	fruiting_bodies	0.21	0.25	
	seed_tmt	0.15	0.28	
	temp	0.6	0.62	
	canker_lesion	0.26	0.25	
	external_decay	0.42	0.25	
	leafspot_size	0.52	0.38	
	severity	0.17	0.22	
	lodging	0.17	0.25	
ชุดข้อมูลผู้ป่วยโรคหัวใจ	Oldpeak	0.15	0.17	0.04
	Restecg	0.15	0.18	
	Age	0.15	0.22	
ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019	Fever	0.15	0.20	0.05
	Taste disorder	0.15	0.20	
	Coryza	0.15	0.20	
ชุดข้อมูลผู้ป่วยโรคไข้เลือดออก	lymphadenopathy	0.15	0.17	0.01
	rash	0.15	0.15	
	gender	0.21	0.21	

จากตาราง 30 และ ตาราง 31 จะพบว่าเมื่อค่าระดับความสำคัญที่สูญหายมีจำนวนมากขึ้นจะส่งผลให้ค่าระดับความสำคัญที่ได้จากเทคนิคเพื่อนบ้านใกล้ที่สุดมีความคลาดเคลื่อนมากขึ้น โดยค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ยของแต่ละชุดข้อมูลจะมีค่าเพิ่มขึ้น เช่น ในชุดข้อมูลผู้ป่วยโรคหัวใจมีค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ยจะเพิ่มขึ้นจาก 0.03 เป็น 0.04 เมื่อมีจำนวนค่าระดับความสำคัญสูญหายร้อยละ 20 และเมื่อมีจำนวนค่าระดับความสำคัญสูญหายร้อยละ 30 จะทำให้ค่าความคลาดเคลื่อนสมบูรณ์เฉลี่ยมีค่าเพิ่มขึ้นเป็น 0.05 เป็นต้น เนื่องจากเทคนิคเพื่อนบ้านใกล้ที่สุดจะใช้

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

วิธีการพิจารณาค่าสูญหายจากข้อมูลที่คล้ายคลึงกันภายในชุดข้มูล ดังนั้นหากภายในชุดข้อมูลมีจำนวนข้อมูลที่มีลักษณะคล้ายคลึงกันน้อยอาจส่งผลให้การประมาณค่าข้อมูลที่สูญหายมีความคลาดเคลื่อนได้

การทดสอบผลของการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายโดยนำค่าระดับความสำคัญของแอตทริบิวต์ที่ได้จากเทคนิคเพื่อนบ้านใกล้ที่สุดในตาราง 29 ถึง ตาราง 31 มาประยุกต์ใช้สามารถแสดงได้ดังภาพ 37



ภาพ 37 ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเมื่อมีการประมาณค่าระดับความสำคัญของแอตทริบิวต์

จากภาพ 37 จะพบว่าเมื่อนำค่าระดับความสำคัญของแอตทริบิวต์ที่ทำการประมาณค่าด้วยเทคนิคเพื่อนบ้านใกล้ที่สุดเมื่อแอตทริบิวต์ในชุดข้อมูลไม่ปรากฏองค์ความรู้ที่เกี่ยวข้องในออนโทโลยีไปใช้ในกระบวนการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะทำให้ค่าความถูกต้องในการจำแนกข้อมูลมีค่าลดลงเมื่อเปรียบเทียบกับค่าความถูกต้องในการจำแนกข้อมูลเมื่อทุกแอตทริบิวต์สามารถอ้างอิงค่าระดับความสำคัญได้จากองค์ความรู้ในออนโทโลยี เมื่อจำนวน

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

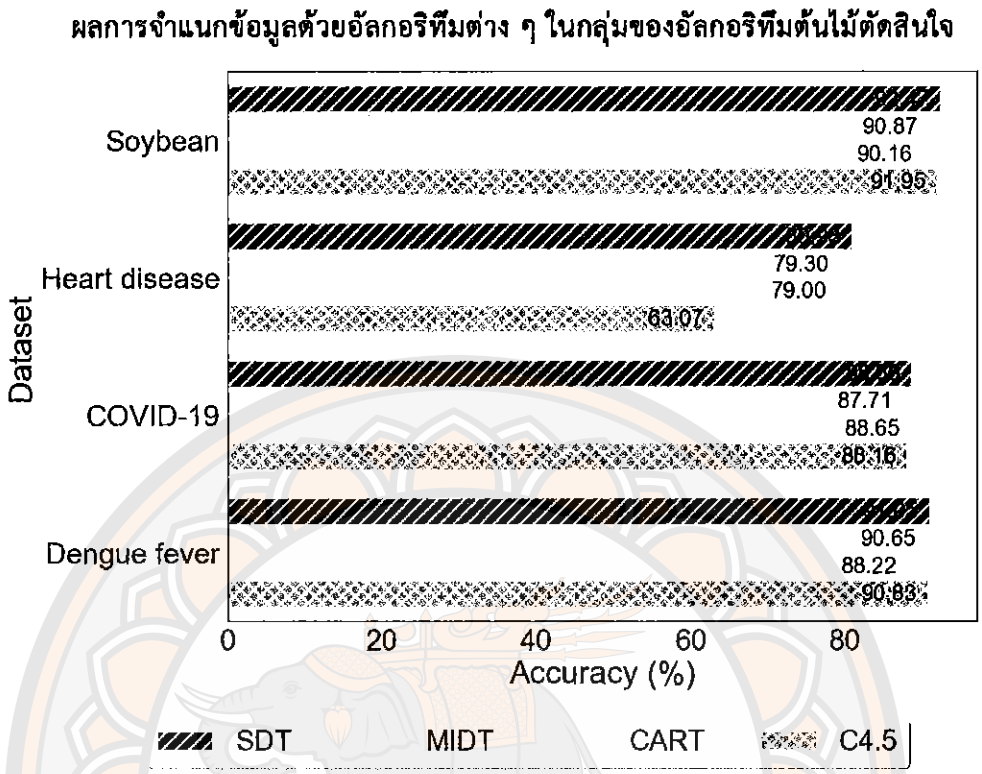
แอตทริบิวต์ที่ไม่สามารถอ้างอิงค่าระดับความสำคัญจากองค์ความรู้ในออนโทโลยีมีจำนวนมากขึ้นจะส่งผลให้ค่าความถูกต้องในการจำแนกข้อมูลมีค่าลดลงมากขึ้น โดยค่าความถูกต้องในการจำแนกข้อมูลของชุดข้อมูลการเกิดโรคของถั่วเหลืองจะมีค่าลดลง 0.64% เมื่อมีค่าระดับความสำคัญสูญหายร้อยละ 10 และเมื่อมีจำนวนค่าระดับความสำคัญสูญหายร้อยละ 30 จะทำให้ค่าความถูกต้องในการจำแนกข้อมูลลดลง 1.17% ในขณะที่ชุดข้อมูลผู้ป่วยโรคหัวใจ ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออกจะมีค่าความถูกต้องในการจำแนกข้อมูลลดลง 0.41% 0.46% และ 0.87% ตามลำดับ เมื่อมีค่าระดับความสำคัญของแอตทริบิวต์สูญหายร้อยละ 30

จากการทดลองนี้สามารถสรุปได้ว่าการประมาณค่าระดับความสำคัญของแอตทริบิวต์ด้วยเทคนิคเพื่อนบ้านใกล้ที่สุดสามารถช่วยแก้ปัญหาในกรณีที่แอตทริบิวต์ไม่ปรากฏเป็นองค์ความรู้ในออนโทโลยีได้ โดยการพิจารณาความคล้ายคลึงระหว่างแอตทริบิวต์ซึ่งช่วยให้สามารถประมาณค่าระดับความสำคัญของแอตทริบิวต์ได้ใกล้เคียงกับค่าจริง รวมถึงมีผลการจำแนกข้อมูลที่ใกล้เคียงกับผลการจำแนกข้อมูลเมื่อทุกแอตทริบิวต์ในชุดข้อมูลปรากฏในออนโทโลยีที่เกี่ยวข้อง อย่างไรก็ตามเมื่อแอตทริบิวต์ไม่ปรากฏเป็นองค์ความรู้ในออนโทโลยีมีจำนวนมากจะส่งผลให้การประมาณค่าระดับความสำคัญมีความคลาดเคลื่อนมากขึ้นและอาจส่งผลให้ความถูกต้องในการจำแนกข้อมูลลดลงมากขึ้น

ผลการเปรียบเทียบประสิทธิภาพของต้นไม้ตัดสินใจเชิงความหมายกับอัลกอริทึมอื่น ๆ

การทดลองนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายกับอัลกอริทึมอื่น ๆ ในกลุ่มของอัลกอริทึมต้นไม้ตัดสินใจเหมือนกัน ซึ่งประกอบไปด้วย อัลกอริทึม C4.5 อัลกอริทึม Classification and Regression Tree (CART) และอัลกอริทึม Mutual Information Decision Tree (MIDT) (Fang et al., 2017) ซึ่งเป็นอัลกอริทึมต้นไม้ตัดสินใจที่ใช้ค่าสารสนเทศร่วม (Mutual information) เป็นเกณฑ์ในการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับเป็นโหนดของต้นไม้ตัดสินใจ ในการทดลองได้ดำเนินปรับพารามิเตอร์ให้กับแต่ละอัลกอริทึมเพื่อให้ได้ประสิทธิภาพที่ดีที่สุดสำหรับการจำแนกข้อมูล ซึ่งผลลัพธ์ของการจำแนกข้อมูลสามารถแสดงดังภาพ 38

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39



ภาพ 38 ผลการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายและอัลกอริทึมอื่น ๆ ในกลุ่มของอัลกอริทึมต้นไม้ตัดสินใจ

จากภาพ 38 พบว่าค่าความถูกต้องในการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย หรือ SDT มีค่ามากกว่าค่าความถูกต้องในการจำแนกข้อมูลที่ได้จากอัลกอริทึมอื่น ๆ ที่ได้ทำการทดสอบทุกอัลกอริทึม โดยเมื่อพิจารณาค่าความถูกต้องในการจำแนกชุดข้อมูลผู้ป่วยโรคหัวใจจะพบว่า ค่าความถูกต้องในการจำแนกข้อมูลที่ได้จากอัลกอริทึม C4.5 จะมีค่าเท่ากับ 63.07% เนื่องจากอัลกอริทึม C4.5 เป็นอัลกอริทึมที่สามารถทำได้ดีเมื่อชุดข้อมูลมีข้อมูลจำนวนมาก (Roy & Garg, 2017) เมื่อนำอัลกอริทึม C4.5 มาใช้ในการจำแนกข้อมูลผู้ป่วยโรคหัวใจซึ่งมีปริมาณข้อมูล 297 แถว ซึ่งจำนวนข้อมูลอาจไม่เพียงพอสำหรับการวิเคราะห์ข้อมูลจึงทำให้ค่าความถูกต้องในการจำแนกข้อมูลไม่สูงมากนัก ในขณะที่เมื่อทำการจำแนกข้อมูลผู้ป่วยโรคหัวใจด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย ค่าความถูกต้องในการจำแนกข้อมูลจะมีค่าเท่ากับ 80.93% ด้วยเหตุนี้จึงสามารถกล่าวได้ว่า ขนาดของชุดข้อมูลสำหรับการจำแนกข้อมูลส่งผลกระทบต่อความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายเพียงเล็กน้อย

4918891B3 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

สำหรับชุดข้อมูลการเกิดโรคของถั่วเหลือง ชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 และชุดข้อมูลผู้ป่วยโรคไข้เลือดออกค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะมีค่ามากกว่าค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึม C4.5 โดยเมื่อทำการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้น 0.52% สำหรับชุดข้อมูลการเกิดโรคของถั่วเหลือง ค่าความถูกต้องในการจำแนกข้อมูลสำหรับชุดข้อมูลผู้ป่วยโรคติดเชื้อไวรัสโคโรนา 2019 จะเพิ่มขึ้น 0.50% และค่าความถูกต้องในการจำแนกข้อมูลผู้ป่วยโรคไข้เลือดออกเพิ่มขึ้น 0.22% ถึงแม้ว่าอัลกอริทึม C4.5 จะมีประสิทธิภาพในการจำแนกข้อมูลที่ดีกว่าอัลกอริทึม ID3 (Hssina et al., 2014) แต่จากการทดลองนี้จะพบว่าอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมาย หรือ SDT ซึ่งมีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีเพื่อปรับปรุงอัลกอริทึม ID3 นั้น มีค่าความถูกต้องในการจำแนกข้อมูลมากกว่าอัลกอริทึม C4.5

จากผลการทดลองจึงสามารถสรุปได้ว่าอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายมีประสิทธิภาพในการจำแนกข้อมูลมากกว่าอัลกอริทึม MIDT อัลกอริทึม CART และอัลกอริทึม C4.5 โดยการประยุกต์ใช้ค่าน้ำหนักของแอตทริบิวต์ที่อ้างอิงได้จากออนโทโลยีในกระบวนการพิจารณาแอตทริบิวต์สำหรับเป็นโหนดของต้นไม้ตัดสินใจนั้นสามารถทำงานได้ดีกับทั้งชุดข้อมูลที่มีขนาดเล็กดังตัวอย่างเช่น ชุดข้อมูลผู้ป่วยโรคหัวใจ หรือแม้กระทั่งชุดข้อมูลที่มีแอตทริบิวต์จำนวนมากดังตัวอย่างเช่น ชุดข้อมูลการเกิดโรคของถั่วเหลืองซึ่งมีแอตทริบิวต์ที่เกี่ยวข้องถึง 31 แอตทริบิวต์

สรุปผลการวิจัย

การปรับปรุงประสิทธิภาพในการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจ ผู้วิจัยได้ประยุกต์ใช้องค์ความรู้ในออนโทโลยีในรูปแบบของค่าระดับความสำคัญของข้อมูลในการปรับปรุงกระบวนการพิจารณาแอตทริบิวต์ที่เหมาะสมสำหรับใช้เป็นโหนดภายในต้นไม้ตัดสินใจ โครงสร้างและความสัมพันธ์ระหว่างแนวความคิดภายในออนโทโลยีจะถูกนำมาใช้ในการกำหนดค่าระดับความสำคัญของแต่ละแนวความคิดด้วยอัลกอริทึม Weighted Semantic PageRank และนำค่าระดับความสำคัญของแนวความคิดที่คำนวณได้ไปใช้เป็นค่าระดับความสำคัญของแอตทริบิวต์ที่มีความสัมพันธ์ในกระบวนการสร้างต้นไม้ตัดสินใจ

ค่าระดับความสำคัญของแต่ละแอตทริบิวต์จะถูกนำไปใช้ในการปรับปรุงค่าเกนสารสนเทศซึ่งเป็นเกณฑ์ที่ใช้ในการพิจารณาค่าแอตทริบิวต์ที่เหมาะสมสำหรับใช้เป็นโหนดของต้นไม้ตัดสินใจเพื่อลดปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลาย ซึ่งหมายถึงการที่แอตทริบิวต์ที่มีค่าข้อมูลหลายค่ามีโอกาสที่จะถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากกว่าแอตทริบิวต์ที่มีจำนวนค่าข้อมูลน้อยกว่านั่นเอง การนำค่าระดับความสำคัญของแอตทริบิวต์มาช่วยในการปรับปรุงค่าเกนสารสนเทศนี้จะช่วยให้แอตทริบิวต์ที่มีจำนวนค่าข้อมูลน้อยแต่มีความสำคัญในโดเมนที่ศึกษามีโอกาสถูกเลือกเป็น

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

โหนดภายในต้นไม้ตัดสินใจได้มากขึ้น ส่งผลให้ค่าความถูกต้องในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายนี้สูงกว่าค่าความถูกต้องที่ได้จากอัลกอริทึมต้นไม้ตัดสินใจที่ใช้ค่าเกนสารสนเทศในกระบวนการพิจารณาโหนดของต้นไม้ตัดสินใจ หรืออัลกอริทึม ID3 รวมทั้งอัลกอริทึม CART และ อัลกอริทึม MIDT อีกด้วย

อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายสามารถทำงานได้ดีกับชุดข้อมูลที่แต่ละแอตทริบิวต์มีจำนวนค่าข้อมูลที่แตกต่างกัน รวมถึงชุดข้อมูลที่มีขนาดเล็กซึ่งอาจขาดข้อมูลที่มีความสำคัญต่อการวิเคราะห์ข้อมูล อย่างไรก็ตามอัลกอริทึมนี้จะมีประสิทธิภาพลดลงเมื่อทำการจำแนกข้อมูลในชุดข้อมูลที่แต่ละแอตทริบิวต์มีจำนวนค่าข้อมูลเพียง 2 ค่า โดยทำให้ค่าความถูกต้องในการจำแนกข้อมูลเพิ่มขึ้นเพียงเล็กน้อย

นอกจากนี้การนำองค์ความรู้ภายนอกมาช่วยสนับสนุนกระบวนการสร้างต้นไม้ตัดสินใจยังช่วยลดโอกาสในการเกิดความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ หรือ Overfitting รวมถึงยังสามารถทำงานได้อย่างมีประสิทธิภาพแม้ในชุดข้อมูลจะปรากฏข้อมูลที่ผิดปกติ เนื่องจากในกระบวนการสร้างต้นไม้ตัดสินใจเชิงความหมายนี้จะใช้ทั้งข้อมูลที่ปรากฏในชุดข้อมูลและองค์ความรู้ภายในออนโทโลยีในการพิจารณาแอตทริบิวต์ที่ใช้เป็นโหนดของต้นไม้ตัดสินใจ องค์ความรู้ที่อยู่ในออนโทโลยีจะช่วยให้แอตทริบิวต์ที่มีนัยสำคัญในแต่ละโดเมนถูกนำไปใช้ในการสร้างต้นไม้ตัดสินใจมากขึ้น ซึ่งทำให้สามารถจำแนกข้อมูลได้อย่างถูกต้องกับชุดข้อมูลที่ไม่เคยพบมาก่อน หรือชุดข้อมูลที่ปรากฏข้อมูลที่ผิดปกติ

อย่างไรก็ตามประเด็นที่จำเป็นต้องให้ความสำคัญเมื่อมีการนำอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายไปใช้ในการจำแนกข้อมูล คือ องค์ความรู้ที่อยู่ในภายในออนโทโลยี เนื่องจากค่าระดับความสำคัญของแอตทริบิวต์ที่นำมาใช้ในกระบวนการพิจารณาโหนดของต้นไม้ตัดสินใจจะได้มาจากการพิจารณาแนวความคิดและความสัมพันธ์ระหว่างแนวความคิดในออนโทโลยี หากมีแอตทริบิวต์บางส่วนไม่ปรากฏเป็นแนวความคิดภายในออนโทโลยีอาจทำให้ค่าระดับความสำคัญของแอตทริบิวต์มีความคลาดเคลื่อน ทำให้ต้นไม้ตัดสินใจที่ได้มีโครงสร้างที่ผิดปกติซึ่งส่งผลต่อประสิทธิภาพในการจำแนกข้อมูล นอกจากนี้การเชื่อมโยงระหว่างแนวความคิดภายในออนโทโลยีเป็นอีกประเด็นที่ควรให้ความสำคัญ เนื่องจากการคำนวณค่าระดับความสำคัญของแอตทริบิวต์จะใช้อัลกอริทึม PageRank เป็นพื้นฐาน ซึ่งอัลกอริทึมในกลุ่มนี้จะพิจารณาค่าความถี่ของการเชื่อมโยงระหว่างแนวความคิดในการระบุค่าระดับความสำคัญ ซึ่งหากแนวความคิดในออนโทโลยีมีการเชื่อมโยงระหว่างแนวความคิดที่ไม่ครบถ้วนหรือมีความผิดพลาดแล้ว จะส่งผลต่อค่าระดับความสำคัญของแอตทริบิวต์และประสิทธิภาพของการจำแนกข้อมูล

4918881183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

บทที่ 6
สรุปผลการวิจัย

บทนี้จะนำเสนอสรุปผลการดำเนินการวิจัยและประเด็นต่าง ๆ ที่เกี่ยวข้อง ซึ่งประกอบไปด้วยหัวข้อดังต่อไปนี้

- สรุปผลการวิจัย
- ข้อค้นพบที่ได้และการบรรลุวัตถุประสงค์การวิจัย
- ข้อกำจัดและแนวทางในการวิจัย
- การนำไปใช้ประโยชน์
- บทความทางวิชาการจากงานวิจัย

สรุปผลการวิจัย

การวิจัยเรื่องการปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมาย ผู้วิจัยได้ทำการประยุกต์ใช้องค์ความรู้ซึ่งอยู่ในรูปแบบของออนโทโลยีในการปรับปรุงประสิทธิภาพการจำแนกข้อมูลของอัลกอริทึมต้นไม้ตัดสินใจซึ่งมีการดำเนินการใน 2 ขั้นตอน คือ ขั้นการจัดเตรียมข้อมูล และขั้นตอนการสร้างแบบจำลองการจำแนกข้อมูล โดยสามารถสรุปผลการดำเนินได้ดังนี้

1. ขั้นตอนการจัดเตรียมข้อมูล

สำหรับขั้นตอนนี้ผู้วิจัยได้ใช้วิธีการทางสถิติและองค์ความรู้ในออนโทโลยีในปรับปรุงข้อมูลเพื่อนำไปใช้ในการจำแนกข้อมูลด้วยอัลกอริทึมต้นไม้ตัดสินใจ โดยสถิติไคสแควร์และสัมประสิทธิ์สหสัมพันธ์แบบพอยท์ไบซีเรียลจะถูกนำมาใช้ในการพิจารณาความสัมพันธ์ระหว่างแอตทริบิวต์และคลาสคำตอบของแต่ละชุดข้อมูลเพื่อลดจำนวนแอตทริบิวต์ที่ไม่มีความสัมพันธ์กับคลาสคำตอบ หลังจากนั้นองค์ความรู้ในออนโทโลยีจะถูกนำมาใช้ในกระบวนการแปลงข้อมูลเพื่อลดจำนวนค่าข้อมูลของแอตทริบิวต์ซึ่งเป็นปัจจัยที่มีผลต่อขนาดและความซับซ้อนของต้นไม้ตัดสินใจ โดยค่าข้อมูลของแต่ละแอตทริบิวต์จะถูกจับคู่กับตัวอย่างข้อมูล (Instance) ในออนโทโลยีเพื่อค้นหาแนวความคิดที่มีความสัมพันธ์กับตัวอย่างข้อมูลนั้น และนำค่าแนวความคิดที่ได้ใช้เป็นแนวความคิดพื้นฐาน (Abstract data) สำหรับแทนที่ค่าข้อมูลของแอตทริบิวต์ที่มีความสัมพันธ์ การแทนที่ข้อมูลในชุดข้อมูลด้วยแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีนอกจากจะช่วยลดจำนวนค่าข้อมูลของแอตทริบิวต์ซึ่งส่งผลให้ได้ต้นไม้ตัดสินใจที่มีความลึกหรือความสูงลดลงแล้ว ยังช่วยลดการปะปนกันของข้อมูลในแต่ละคลาสคำตอบซึ่งช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลได้

49188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

2. ขั้นตอนการสร้างแบบจำลองการจำแนกข้อมูล

ในขั้นตอนนี้แนวความคิดและความสัมพันธ์ระหว่างแนวความคิดภายในออนโทโลยีจะถูกนำไปใช้ในการระบุค่าระดับความสำคัญของแนวความคิดโดยการประยุกต์ใช้เทคนิคการสรุปภาพรวมออนโทโลยี (Ontology Summarization) และนำค่าระดับความสำคัญของแนวความคิดที่ได้ไปใช้เป็นค่าระดับความสำคัญของแอตทริบิวต์เพื่อปรับปรุงค่าเกนสารสนเทศที่ใช้ในขั้นตอนการสร้างต้นไม้ตัดสินใจ

ค่าเกนสารสนเทศที่ถูกปรับปรุงด้วยค่าระดับความสำคัญของแอตทริบิวต์นี้จะถูกใช้เป็นเกณฑ์ในการพิจารณาแอตทริบิวต์ที่ทำหน้าที่เป็นโหนดของต้นไม้ตัดสินใจเพื่อช่วยลดปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีข้อมูลหลากหลายเป็นโหนดภายในต้นไม้ตัดสินใจที่เกิดขึ้นเมื่อมีการใช้ค่าเกนสารสนเทศเป็นเกณฑ์ในการพิจารณาโหนดของต้นไม้ตัดสินใจ ซึ่งองค์ความรู้ที่ได้จากออนโทโลยีซึ่งอยู่ในรูปแบบของค่าระดับความสำคัญของแอตทริบิวต์นี้จะช่วยให้แอตทริบิวต์ที่มีความสำคัญแต่มีจำนวนค่าข้อมูลในแอตทริบิวต์น้อยมีโอกาสถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจมากขึ้น และส่งผลให้แบบจำลองที่ได้สามารถจำแนกข้อมูลได้ถูกต้องมากขึ้น รวมถึงมีกฎการตัดสินใจที่สามารถทำความเข้าใจได้ง่ายและมีความใกล้เคียงกับการพิจารณาของผู้เชี่ยวชาญมากขึ้น

นอกจากนี้การนำองค์ความรู้จากออนโทโลยีมาช่วยในการสร้างต้นไม้ตัดสินใจยังช่วยลดโอกาสในการเกิดปัญหาความจำเพาะกับข้อมูลที่ใช้ในการเรียนรู้ หรือ Overfitting รวมถึงสามารถจำแนกข้อมูลได้อย่างมีประสิทธิภาพแม้จะเป็นชุดข้อมูลที่ปรากฏข้อมูลที่ผิดปกติ (Noise) อีกด้วย

ผลการวิจัยในการปรับปรุงประสิทธิภาพการจำแนกข้อมูลต้นไม้ตัดสินใจในทั้ง 2 ขั้นตอนสามารถช่วยลดข้อจำกัดที่เกิดขึ้นกับเทคนิคต้นไม้ตัดสินใจได้ ดังนี้

- การพิจารณาความสัมพันธ์ระหว่างข้อมูลช่วยลดจำนวนแอตทริบิวต์ที่ไม่มีส่วนเกี่ยวข้องกับคลาสคำตอบสามารถช่วยแก้ปัญหาคุณภาพของข้อมูลที่มีผลต่อประสิทธิภาพการจำแนกข้อมูลของต้นไม้ตัดสินใจ

- การนำแนวความคิดพื้นฐานที่อ้างอิงได้จากออนโทโลยีมาใช้ปรับปรุงข้อมูลที่มีความสัมพันธ์ในชุดข้อมูลทำให้แอตทริบิวต์มีจำนวนค่าข้อมูลในแอตทริบิวต์ลดลง ซึ่งช่วยแก้ปัญหาประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจเมื่อทำงานกับแอตทริบิวต์ที่มีค่าข้อมูลจำนวนมากได้

- การปรับปรุงค่าเกนสารสนเทศด้วยค่าระดับความสำคัญของแอตทริบิวต์ที่พิจารณาจากองค์ความรู้ในออนโทโลยีทำให้แอตทริบิวต์ที่มีความสำคัญมีโอกาสถูกเลือกเป็นโหนดของต้นไม้ตัดสินใจมากขึ้น ซึ่งสามารถช่วยลดปัญหาการลำเอียงไปยังแอตทริบิวต์ที่มีค่าข้อมูลหลากหลายที่มักเกิดขึ้นเมื่อใช้เกนสารสนเทศเป็นเกณฑ์ในการเลือกโหนดของต้นไม้ตัดสินใจได้

4818881183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ข้อค้นพบที่ได้และการบรรลุวัตถุประสงค์การวิจัย

การวิจัยเรื่องการปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมายสามารถสรุปข้อค้นพบในงานวิจัยได้ดังนี้

1. การประยุกต์ใช้ออนโทโลยีในการสนับสนุนการเตรียมข้อมูลสำหรับการวิเคราะห์ข้อมูล

การพิจารณาความสัมพันธ์ระหว่างข้อมูลร่วมกับการแปลงข้อมูลในชุดข้อมูลด้วยค่าแนวความคิดพื้นฐาน (Abstract data) ที่มีความสัมพันธ์ซึ่งอ้างอิงได้จากออนโทโลยี สามารถช่วยลดความหลากหลายของค่าข้อมูลในแต่ละแอตทริบิวต์ ส่งผลให้ต้นไม้ตัดสินใจมีความซับซ้อนลดลงและมีประสิทธิภาพในการจำแนกข้อมูลเพิ่มขึ้น ซึ่งสอดคล้องตามวัตถุประสงค์การวิจัยที่จะนำเสนอวิธีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีในการสนับสนุนกระบวนการจัดเตรียมข้อมูล ซึ่งวิธีการที่ได้จากการวิจัยนี้เป็นแนวทางที่ช่วยปรับปรุงข้อมูลเพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูลได้

2. การปรับปรุงเกณฑ์การพิจารณาโหนดของต้นไม้ตัดสินใจด้วยออนโทโลยี

- การนำค่าระดับความสำคัญของแอตทริบิวต์ที่ได้จากการพิจารณาโครงสร้างและความสัมพันธ์ของแนวความคิดในออนโทโลยีมาใช้การปรับปรุงค่าเกนสารสนเทศนั้น ทำให้ได้วิธีการหาเกนสารสนเทศแบบใหม่ ที่ไม่มีนักวิจัยท่านใดนำเสนอมาก่อน และใช้เป็นเกณฑ์ใหม่สำหรับการพิจารณาโหนดของต้นไม้ตัดสินใจ

- การนำค่าระดับความสำคัญของแอตทริบิวต์ที่ได้จากออนโทโลยีมาช่วยในการปรับปรุงค่าเกนสารสนเทศสามารถช่วยลดปัญหาความลำเอียงในการเลือกแอตทริบิวต์ที่มีค่าหลากหลายเป็นโหนดของต้นไม้ตัดสินใจได้ โดยแอตทริบิวต์ที่มีจำนวนค่าของข้อมูลในแอตทริบิวต์น้อยจะมีโอกาสถูกเลือกเป็นโหนดภายในต้นไม้ตัดสินใจมากขึ้นหากแอตทริบิวต์นั้นเป็นแอตทริบิวต์ที่มีความสำคัญในโดเมนที่ศึกษา

- การนำค่าระดับความสำคัญของแอตทริบิวต์ที่ได้จากออนโทโลยีมาใช้ในการปรับปรุงค่าเกนสารสนเทศสามารถช่วยลดโอกาสในการเกิดปัญหาความจำเพาะกับข้อมูลที่เรียนรู้ หรือ Overfitting ได้

จากข้อค้นพบข้างต้นสามารถสรุปได้ว่าการประยุกต์ใช้องค์ความรู้ในออนโทโลยีสามารถช่วยเพิ่มประสิทธิภาพในการคำนวณค่าเกนสารสนเทศที่เป็นเกณฑ์ในการพิจารณาแอตทริบิวต์สำหรับเป็นโหนดภายในต้นไม้ตัดสินใจ และส่งผลให้สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจได้ ซึ่งสอดคล้องตามวัตถุประสงค์ในการวิจัยที่ต้องการพัฒนาวิธีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีร่วมกับเทคนิคต้นไม้ตัดสินใจในการปรับปรุงประสิทธิภาพการจำแนกข้อมูลของเทคนิคต้นไม้ตัดสินใจโดยการประยุกต์ใช้องค์ความรู้ในออนโทโลยีช่วยในการคำนวณค่าเกนสารสนเทศ

4918818183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ข้อค้นพบของงานวิจัยซึ่งมีความแตกต่างจากงานวิจัยอื่น ๆ ที่ทำการปรับปรุงประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจสามารถสรุปได้ดังนี้

- งานวิจัยนี้ทำการปรับปรุงค่าเกนสารสนเทศด้วยค่าระดับความสำคัญของแอตทริบิวต์ที่พิจารณาจากองค์ความรู้และความสัมพันธ์ของข้อมูลในออนโทโลยี ในขณะที่งานวิจัยอื่น ๆ จะพิจารณาค่าระดับความสำคัญจากค่าสหสัมพันธ์ (correlation) ของข้อมูล หรือกำหนดโดยผู้เชี่ยวชาญในแต่ละศาสตร์

- งานวิจัยนี้ทำการพิจารณาค่าระดับความสำคัญของแนวความคิดในออนโทโลยีโดยการประยุกต์ใช้อัลกอริทึม Weighted Semantic PageRank ที่ใช้ในการจัดอันดับความสำคัญของเว็บไซต์

- วิธีการที่นำเสนอในงานวิจัยนี้สามารถประยุกต์ใช้กับข้อมูลอย่างหลากหลายเนื่องจากออนโทโลยีและชุดข้อมูลเป็นอิสระต่อกัน โดยออนโทโลยีที่นำมาใช้ในการปรับปรุงค่าเกนสารสนเทศในงานวิจัยนี้เป็นออนโทโลยีที่ได้รับการเผยแพร่ในแหล่งเรียนรู้ต่าง ๆ ซึ่งสามารถนำมาใช้ได้ทันทีโดยที่ผู้วิเคราะห์ข้อมูลไม่จำเป็นต้องพัฒนาออนโทโลยีขึ้นเอง ผู้วิเคราะห์ข้อมูลสามารถนำออนโทโลยีที่ได้รับการเผยแพร่ไปใช้เชื่อมโยงกับชุดข้อมูลที่ต้องการได้ ในขณะที่งานวิจัยอื่น ๆ มักดำเนินการออกแบบและพัฒนาออนโทโลยีขึ้นเองเพื่อให้สอดคล้องตามข้อมูลและวัตถุประสงค์ของงานจึงทำให้มีความจำเพาะระหว่างออนโทโลยีและชุดข้อมูลที่ใช้งาน

ข้อจำกัดและแนวทางในการวิจัย

การวิจัยเรื่องการปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมายมีข้อจำกัดและสามารถนำไปใช้เป็นแนวทางในการวิจัยได้ดังนี้

ในส่วนของการแปลงข้อมูลโดยใช้องค์ความรู้ในออนโทโลยี ในการวิจัยนี้จะใช้แนวความคิดพื้นฐานซึ่งมีค่าแตกต่างจากค่าข้อมูลเดิมเพียงหนึ่งระดับเนื่องจากโครงสร้างของออนโทโลยีที่ใช้ในการศึกษา ซึ่งหากออนโทโลยีที่ใช้มีการเชื่อมโยงของแนวความคิดที่สามารถอ้างอิงไปยังแนวความคิดที่เกี่ยวข้องได้มากกว่าหนึ่งระดับ อัลกอริทึมที่นำเสนอจะไม่สามารถอ้างอิงไปยังแนวความคิดนั้น ๆ ได้ ดังนั้นจึงควรปรับปรุงอัลกอริทึมให้สามารถอ้างอิงไปยังแนวความคิดที่มีความสัมพันธ์มากกว่าหนึ่งระดับพร้อมทั้งทำการตรวจสอบความเหมาะสมในการนำค่าแนวความคิดเหล่านั้นมาใช้ในการแปลงข้อมูลเพื่อลดปัญหาที่อาจสูญเสียข้อมูลที่สำคัญได้

ในส่วนของการพิจารณาค่าความสำคัญของแอตทริบิวต์จากออนโทโลยีนั้น เนื่องจากการวิจัยในครั้งนี้ทำการทดลองโดยมีเงื่อนไขที่ทุกแอตทริบิวต์ของชุดข้อมูลจะมีแนวความคิดที่มีความสัมพันธ์ปรากฏอยู่ภายในออนโทโลยี ดังนั้นจึงทำให้สามารถทำการปรับปรุงค่าเกนสารสนเทศด้วยค่าระดับความสำคัญของแต่ละแอตทริบิวต์ได้ แต่หากในออนโทโลยีไม่ปรากฏแนวความคิดที่สอดคล้องกับ

4188183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

แอตทริบิวต์ภายในชุดข้อมูลแล้วอาจทำให้ประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายลดลง ซึ่งในการวิจัยครั้งนี้ได้นำเสนอวิธีการประมาณค่าระดับความคัญของแอตทริบิวต์ที่ไม่ปรากฏแนวความคิดที่สอดคล้องในออนโทโลยีด้วยเทคนิคเพื่อนบ้านใกล้ที่สุด (k-NN) อย่างไรก็ตามค่าระดับความสำคัญที่ประมาณการจากวิธีการที่นำเสนอยังคงมีความคลาดเคลื่อน ดังนั้นจึงอาจมีการพัฒนาเพื่อแก้ปัญหาในกรณีที่ไม่ปรากฏแนวความคิดในออนโทโลยีที่มีความสัมพันธ์กับแอตทริบิวต์ในชุดข้อมูลให้ค่าระดับความสำคัญมีความถูกต้องมากยิ่งขึ้น เช่น การนำออนโทโลยีอื่น ๆ ที่เกี่ยวข้องมาร่วมในการดำเนินการ เป็นต้น

4918889183
NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

การนำไปใช้ประโยชน์

การวิจัยเรื่องการปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมายนี้สามารถนำไปประยุกต์ใช้เพื่อสนับสนุนในงานด้านต่าง ๆ ที่จำเป็นต้องอาศัยผู้เชี่ยวชาญในการวิเคราะห์ข้อมูล ซึ่งอัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายนี้สามารถนำไปใช้เพื่อแบ่งเบาภาระของผู้เชี่ยวชาญในแต่ละงานได้ เช่น

- ด้านการเกษตรหากมีการนำระบบการวินิจฉัยการเกิดโรคของพืชที่มีการประยุกต์ใช้อัลกอริทึมต้นไม้ตัดสินใจเชิงความหมายจะสามารถช่วยให้เกษตรกรสามารถทราบถึงโรคที่เกิดขึ้นในพืช รวมทั้งทราบถึงปัจจัยต่าง ๆ ที่เกี่ยวข้องกับโรคพืชนั้น ๆ และสามารถหาแนวทางในการดำเนินการหรือป้องกันโรคพืชรวดเร็วขึ้น
- ด้านการแพทย์หากมีการนำอัลกอริทึมต้นไม้ตัดสินใจไปใช้ในการพัฒนาระบบวินิจฉัยโรคเบื้องต้นจะมีส่วนช่วยให้ผู้ป่วยหรือผู้ที่เกี่ยวข้องทราบถึงความเสี่ยงในการเกิดโรค และทำให้สามารถเข้าสู่ระบบการรักษาที่เหมาะสมได้เร็วขึ้น

นอกจากนี้ผลการวิจัยในครั้งนี้ยังสามารถนำไปใช้ประโยชน์ในด้านการวิจัยได้อีกด้วย โดยนักวิจัยสามารถนำแนวความคิดที่มีการประยุกต์ใช้องค์ความรู้ในออนโทโลยีในการปรับปรุงอัลกอริทึมต้นไม้ตัดสินใจไปใช้ในการพัฒนาเพื่อปรับปรุงอัลกอริทึมอื่น ๆ ที่ใช้ในการจำแนกข้อมูลได้

บทความทางวิชาการจากงานวิจัย

การวิจัยเรื่องการปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยใช้เทคนิคต้นไม้ตัดสินใจเชิงความหมายนี้ได้มีการเผยแพร่ผลการวิจัยในงานประชุมวิชาการ วารสารวิชาการ ๆ รวมถึงรางวัลที่เกี่ยวข้อง ดังนี้

1. การประชุมวิชาการ

- ศิริจรรยา จันทร์มี, ไกรศักดิ์ เกษร. (2562). การจำแนกโรคของถั่วเหลืองโดยการประยุกต์ใช้เทคนิคต้นไม้การตัดสินใจเชิงความหมาย. ใน มหาวิทยาลัยศรีนครินทรวิโรฒ, คณะวิทยาศาสตร์, *งานประชุมวิชาการวิทยาศาสตร์วิจัย ครั้งที่ 11* (น.863-872). กรุงเทพฯ.

2. วารสารวิชาการ

- Chanmee, S., & Kesorn, K. (2020). Data quality enhancement for decision tree algorithm using knowledge-based model. *Current Applied Science and Technology*, 259-277.

- Chanmee, S., & Kesorn, K. (2021). Semantic data mining in the information age: A systematic review. *International Journal of Intelligent Systems*, 36(8), 3880-3916.

- Chanmee, S., & Kesorn, K. (2022). Exploiting a knowledge base for intelligent decision tree construction to enhance classification power. *Engineering and Applied Science Research*, 49(4), 545-561.

3. รางวัลที่ได้รับ

- Best session presentation of the 3rd Asia Joint Conference on Computing (AJCC2022)

4918389183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

บรรณานุกรม

Ahmed, S. T., Al-Hamdani, R., & Croock, M. S. (2020). Developed third iterative dichotomizer based on feature decisive values for educational data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 209-217.

Ali, M. M., & Rajamani, L. (2012). Decision tree induction: Priority classification. *Proceedings of IEEE-International Conference On Advances In Engineering, Science And Management*, (pp. 668-673).

Amjad, M., Ali, Z., Rafiq, A., Akhtar, N., Israr Ur, R., & Abbas, A. (2019). Empirical Performance Analysis of Decision Tree and Support Vector Machine based Classifiers on Biological Databases. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *10*(9), 309-318.

Anand, S. S., Bell, D. A., & Hughes, J. G. (1995). The role of domain knowledge in data mining. *Proceedings of the 4th international conference on Information and knowledge management*, (pp. 37-43).

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5-37.

Blake, R., & Mangiameli, P. (2011). The Effects and Interactions of Data Quality and Problem Complexity on Classification. *Journal of Data and Information Quality*, *2*(2), 1-28.

Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, *56*(18), 3825-3833.

Chanmee, S., & Kesorn, K. (2020). Data Quality Enhancement for Decision Tree Algorithm using Knowledge-Based Model. *Current Applied Science and Technology*, 259-277.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Chen, J., Luo, D.-l., & Mu, F.-x. (2009). An improved ID3 decision tree algorithm.

4918893 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

Proceedings of the 4th International Conference on Computer Science Education, (pp. 127-130).

Crop Ontology Curation, T. (2011). *Soybean Ontology.* Retrieved 24 August 2019, from http://www.cropontology.org/ontology/CO_336/Soybean.

Damak, W., Rebai, I., & Kallel, I. K. (2014). Semantic object recognition by merging decision tree with object ontology. *Proceedings of the 1st International Conference on Advanced Technologies for Signal and Image Processing*, (pp. 65-70).

Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning. *Archives of Computational Methods in Engineering, 27*(4), 1071-1092.

Dietrich, D., Heller, B., & Yang, B. (2015). *Data science and big data analytics: discovering, analyzing, visualizing and presenting data.* Indianapolis,USA: Wiley.

Dou, D., Wang, H., & Liu, H. (2015). Semantic data mining: a survey of ontology-based approaches. *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing*, (pp. 244-251).

Dua, D., & Graff, C. (2017). *UCI machine learning repository.* University of California, Irvine, School of Information and Computer Sciences.

Dwi Prayogo, R., & Ikhsan, N. (2020). Attribute Selection Effect on Tree-Based Classifiers for Letter Recognition. *Proceedings of the 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, (pp. 13-18).

Es-Sabery, F., & Hair, A. (2019). An Improved ID3 Classification Algorithm Based On Correlation Function and Weighted Attribute*. *Proceedings of the 2019 International Conference on Intelligent Systems and Advanced Computing Sciences*, (pp. 1-8).

Fang, L., Jiang, H., & Cui, S. (2017). An improved decision tree algorithm based on mutual information. *Proceedings of the 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, (pp. 1615-1620).

4918816183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

Gahar, R. M., Arfaoui, O., Hidri, M. S., & Hadj-Alouane, N. B. (2018). An Ontology-driven MapReduce Framework for Association Rules Mining in Massive Data. *Procedia Computer Science, 126*, 224-233.

Gang, M., Liumei, Z., Yimin, C., & Quancheng, Z. (2021). Application Research of ID3 Attribute Optimization Algorithm Based on Correlation Coefficient. *Proceedings of the 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, (pp. 279-283).

Gray, L., Lee, I. M., Sesso, H. D., & Batty, G. D. (2011). Blood Pressure in Early Adulthood, Hypertension in Middle Age, and Future Cardiovascular Disease Mortality. *Journal of the American College of Cardiology, 58*(23), 2396-2403.

Grogan, R. G. (1981). The Science and Art of Plant-Disease Diagnosis. *Annual Review of Phytopathology, 19*(1), 333-351.

Gupta, S., & Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science, 161*, 466-474.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques* (3 ed.). USA: Morgan Kaufinann.

Hand, D. J. (2007). Principles of data mining. *Drug Safety, 30*(7), 621-622.

Honest, N. (2020). A survey on Feature Selection Techniques. *GIS SCIENCE JOURNAL, 7*, 353-358.

Iqbal, M. D. R. A., Rahman, S., Nabil, S. I., & Chowdhury, I. U. A. (2012). Knowledge based decision tree construction with feature importance domain knowledge. *Proceedings of the 7th International Conference on Electrical and Computer Engineering*, (pp. 659-662).

Jun, H.-G., Im, D.-H., & Kim, H.-J. (2016). An RDF metadata-based weighted semantic pageRank algorithm. *International Journal of Web & Semantic Technology (IJWesT), 7*, 11-24.

Kastrati, Z., & Imran, A. S. (2019). Performance analysis of machine learning classifiers on improved concept vector space models. *Future Generation Computer Systems, 96*, 552-562.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. *ACM Computing*

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

Surveys, *52*(4), 79:71-79:36.

Knublauch, H., Fergerson, R. W., Noy, N. F., & Musen, M. A. (2004). The Protégé OWL plugin: An open development environment for semantic web applications. *Proceedings of The Semantic Web – ISWC 2004, (pp. 229-243).*

Kononenko, I. (1984). Experiments in automatic learning of medical diagnostic rules. *Technical report, Jozef Stefan Institute*.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., & Lawley, M. (2016). Information retrieval as semantic inference: a Graph Inference model applied to medical search. *Information Retrieval Journal*, *19*(1), 6-37.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261-283.

Kralj, J., Vavpetič, A., Dumontier, M., & Lavrač, N. (2016). Network Ranking Assisted Semantic Data Mining. *Proceedings of Bioinformatics and Biomedical Engineering*, (pp. 752-764).

Kudoh, Y., Haraguchi, M., & Okubo, Y. (2003). Data abstractions for decision tree induction. *Theoretical Computer Science*, *292*(2), 387-416.

Kumar, S., & Baliyan, N. (2018). Quality evaluation of ontologies. In *Semantic Web-Based Systems* (pp. 19-50). Springer.

Kuo, Y.-T., Lonie, A., Sonenberg, L., & Paizis, K. (2007). Domain ontology driven data mining: A medical case study. *Proceedings of the 2007 International Workshop on Domain Driven Data Mining*, (pp. 11-17).

Lamy, J.-B. (2017). Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine*, *80*, 11-28.

Ławrynowicz, A. (2017). *Semantic Data Mining: An Ontology-based Approach*. Germany: IOS Press.

Liu, Y., & Xie, N. (2010). Improved ID3 algorithm. *Proceedings of the 3rd International Conference on Computer Science and Information Technology*, (pp. 465-468).

Maimon, O. Z., & Rokach, L. (2014). *Data Mining With Decision Trees: Theory And Applications (2nd Edition)*. World Scientific Publishing Company.

Markell, S., & Malvick, D. (2018). Soybean disease diagnostic series — publications.

4918893183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

NDSU North Dakota state University. Retrieved 11 November 2019, from https://www.ag.ndsu.edu/publications/crops/soybean-disease-diagnostic-series

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica, 23*(2), 143-149.

Michalski, R. S. (1980). Learning by being told and learning from examples : An experimental comparison of the two methods of knowledge acquisition in the context of development an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems, 4*(2), 125-161.

Miller, G. A. (1995). WordNet: a lexical database for English. *Commun. ACM, 38*(11), 39–41.

Mitraka, E., Topalis, P., Dritsou, V., Dialynas, E., & Louis, C. (2015). Describing the Breakbone Fever: IDODEN, an Ontology for Dengue Fever. *PLOS Neglected Tropical Diseases, 9*(2), e0003479.

Onat, A. (2001). Risk factors and cardiovascular disease in Turkey. *Atherosclerosis, 156*(1), 1-10.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Pouriyeh, S., Allahyari, M., Liu, Q., Cheng, G., Arabnia, H. R., Atzori, M., & Kochut, K. (2018). Graph-Based Methods for Ontology Summarization: A Survey. *Proceedings of the 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering*, (pp. 85-92),

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications, 181*(1), 25-29.

Ramezankhani, A., Pournik, O., Shahrabi, J., Khalili, D., Azizi, F., & Hadaegh, F. (2014). Applying decision tree for identification of a low risk population for type 2 diabetes. Tehran Lipid and Glucose Study. *Diabetes Research and Clinical Practice, 105*(3), 391-398.

Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion, 27*, 111-125.

Roy, S., & Garg, A. (2017). Analyzing performance of students by using data mining

4918389183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

techniques a literature survey. *Proceedings of the 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, (pp. 130-133),

Saranya, D., Gomathi, D., & Chinnasamy, P. (2021). Automatic Ontology Framework for Personal Health Service Using Semantic Web. *Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI)*, (pp. 1-7).

Sargsyan, A., Kodamullil, A. T., Baksi, S., Darms, J., Madan, S., Gebel, S., Keminer, O., Jose, G. M., Balabin, H., DeLong, L. N., Kohler, M., Jacobs, M., & Hofmann-Apitius, M. (2020). The COVID-19 Ontology. *Bioinformatics*, *36*(24), 5703-5705.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526-534.

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: an application in indirect lending. *Decision Support Systems*, *46*(1), 287-299.

Song, Y.-y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130-135.

Soni, V. K., & Pawar, S. (2017). Emotion based social media text classification using optimized improved ID3 classifier. *Proceedings of the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing*, (pp. 1500-1505).

Staab, S., & Studer, R. (2009). *Handbook on ontologies* (2 ed.). Heidelberg, Germany: Springer Science & Business Media

Su, J., & Zhang, H. (2006). A fast decision tree learning algorithm. *Proceedings of the 21st national conference on Artificial intelligence*, (pp. 500-505)

Tang, A., & Fong, S. (2010). A taxonomy-based classification model by using abstraction and aggregation. *Proceedings of the 6th International Conference on Advanced Information Management and Service*, (pp. 448-454).

Tang, K. F., & Ooi, E. E. (2012). Diagnosis of dengue: an update. *Expert Review of Anti-infective Therapy*, *10*(8), 895-907.

Vavpetič, A., Novak, P. K., Grčar, M., Mozetič, I., & Lavrač, N. (2013). Semantic Data

491889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

Mining of Financial News Articles. *Proceeding of the 16th International Conference on Discovery Science,* (pp. 294-307).

Verma, J. P. (2019). Non-parametric Correlations. In J. P. Verma (Ed.), *Statistics and Research Methods in Psychology with Excel* (pp. 523-565). Singapore: Springer.

Viana dos Santos Santana, İ., Silveira, C. M. d., Sobrinho, A., Chaves e Silva, L., Dias da Silva, L., Freire de Souza Santos, D., Candeia, E., & Perkusich, A. (2021). A Brazilian dataset of symptomatic patients for screening the risk of COVID-19. *Mendeley Data, 5*.

Vianna Cardozo, S., Maniero, V., Rangel, P., Camargo, T., Souza, M., Forte, J., & Lamas, C. (2018). Databases of a clinico-ecological study of a triple epidemic. *Mendeley Data, 1*.

Vicente, C. R., Junior, C. C., Fröschl, G., Romano, C. M., Cabidelle, A. S. A., & Herbinger, K. H. (2017). Influence of demographics on clinical outcome of dengue: a cross-sectional study of 6703 confirmed cases in Vitória, Espírito Santo State, Brazil. *Epidemiology & Infection*, *145*(1), 46-53.

Vieira, J., & Antunes, C. (2014). Decision tree learner in the presence of domain knowledge. *Proceeding of Chinese Semantic Web and Web Science Conference,* (pp. 42-55).

Wang, H., Jiang, W., Deng, X., & Geng, J. (2021). A new method for fault detection of aero-engine based on isolation forest. *Measurement*, *185*, 110064.

Wang, L. (2015). *Heart Failure Ontology*. Retrieved 5 August 2021, from https://bioportal.bioontology.org/ontologies/HFO

Wen, Y., & Xu, W. (2021). Research on Influencing Factors of Fatigue Driving Based on Decision Tree. *Proceedings of the 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA),* (pp. 520-524).

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques* (4th ed.). San Francisco, CA, USA: Morgan Kaufmann.

Xiahou, J., Xiao, M., He, X., & Cui, X. (2021). Research on the Applications of Data Mining in the Analysis of Vehicle Insurance Industry. *Proceedings of the 4th Advanced*

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), (pp. 1095-1099).

Yang. (2000). *Soybean Bacterial Blight and Brown Spot*. Iowa State University. Retrieved 7 April 2022, from https://store.extension.iastate.edu/product/Soybean-Bacterial-Blight-and-Brown-Spot.

Zhang, J., Silvescu, A., & Honavar, V. (2002). Ontology-driven induction of decision trees at multiple levels of abstraction. *Proceedings of International Symposium on Abstraction, Reformulation, and Approximation,* (pp. 316-323).

Zhang, X., Cheng, G., & Qu, Y. (2007). Ontology summarization based on rdf sentence graph. *Proceedings of the 16th international conference on World Wide Web,* (pp. 707-716).

Zhou, B., Svetashova, Y., Pychynski, T., Baimuratov, I., Soylu, A., & Kharlamov, E. (2020). SemFE: Facilitating ML Pipeline Development with Semantics. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management,* (pp. 3489-3492).

Zhou, H., Zhang, J., Zhou, Y., Guo, X., & Ma, Y. (2020). A feature selection algorithm of decision tree based on feature weight. *Expert Systems with Applications*, *164*, 113842.

Zhu, F., Tang, M., Xie, L., Zhu, & Haodong. (2018). A Classification Algorithm of CART Decision Tree based on MapReduce Attribute Weights. *International Journal of Performability Engineering*, *14*(1), 17.

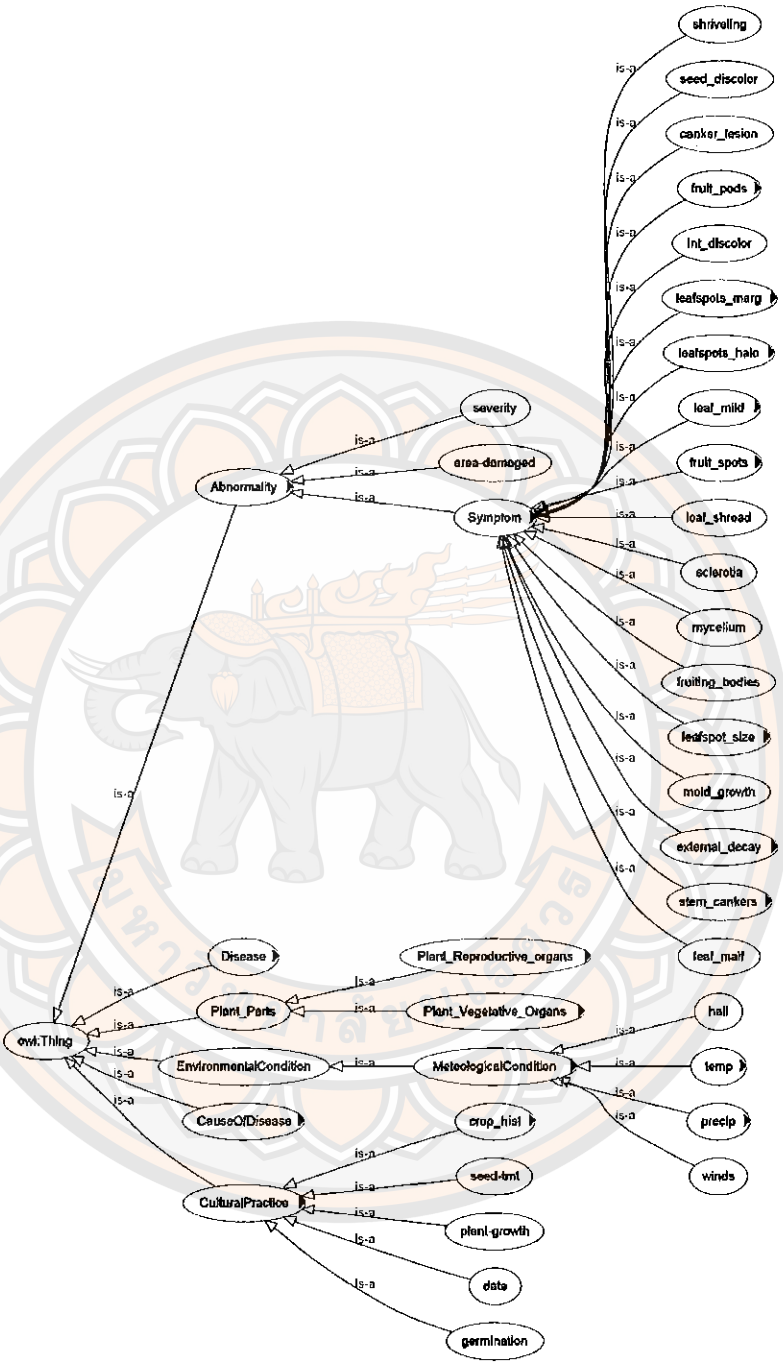
มาลี กาบมาลา, ลำปาง แม่นมาตย์ และ ครรชิต มาลัยวงศ์. (2549). ออนโทโลยี: แนวคิดการพัฒนา Ontologies: Approach for Building. Journal of Information Science, 24(1-3), 24-49.

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39
4918896183

ภาคผนวก

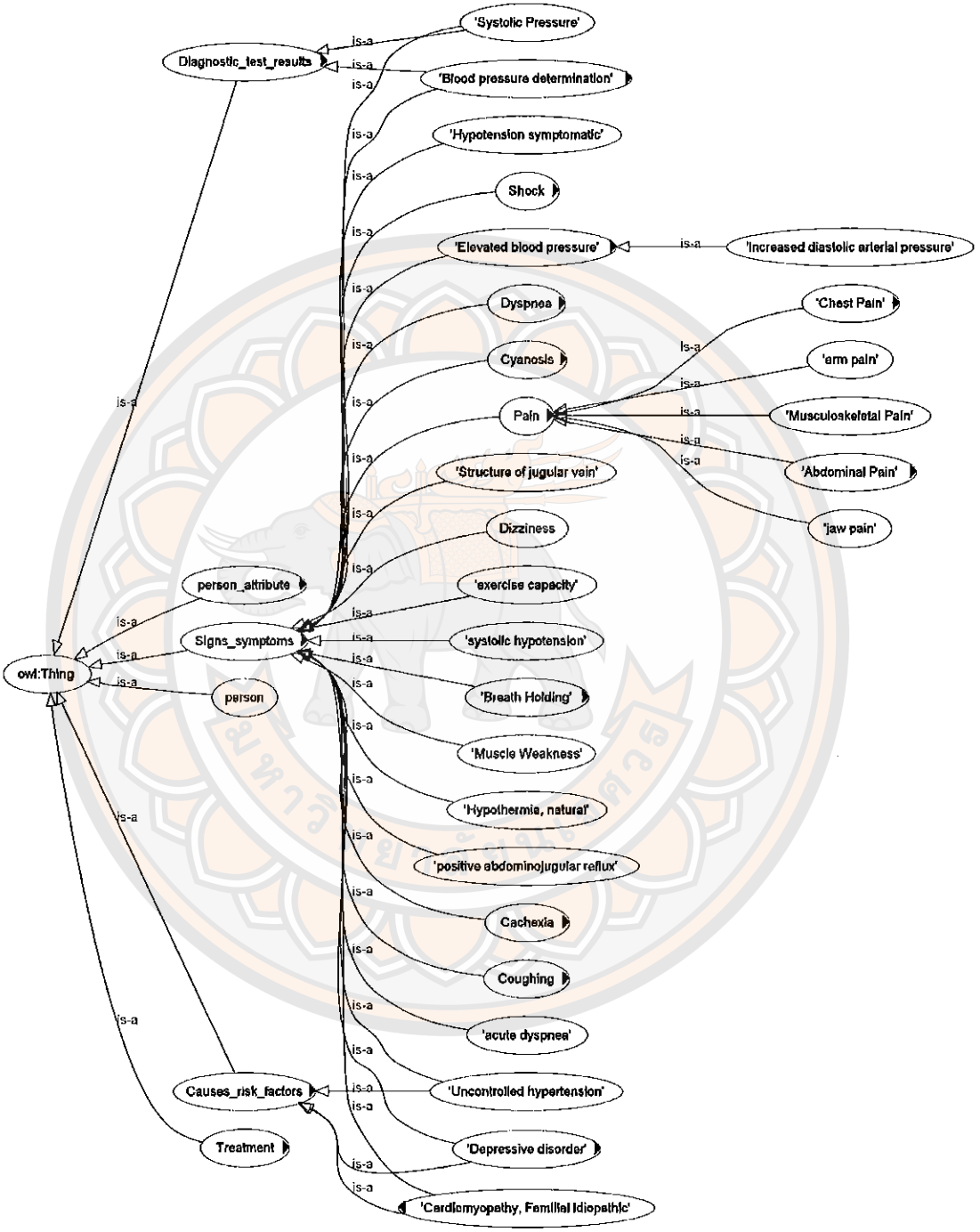
ภาคผนวก ก ตัวอย่างแนวความคิดในออนโทโลยีโรคของถั่วเหลือง



ภาพ 39 ตัวอย่างแนวความคิดในออนโทโลยีโรคของถั่วเหลือง

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

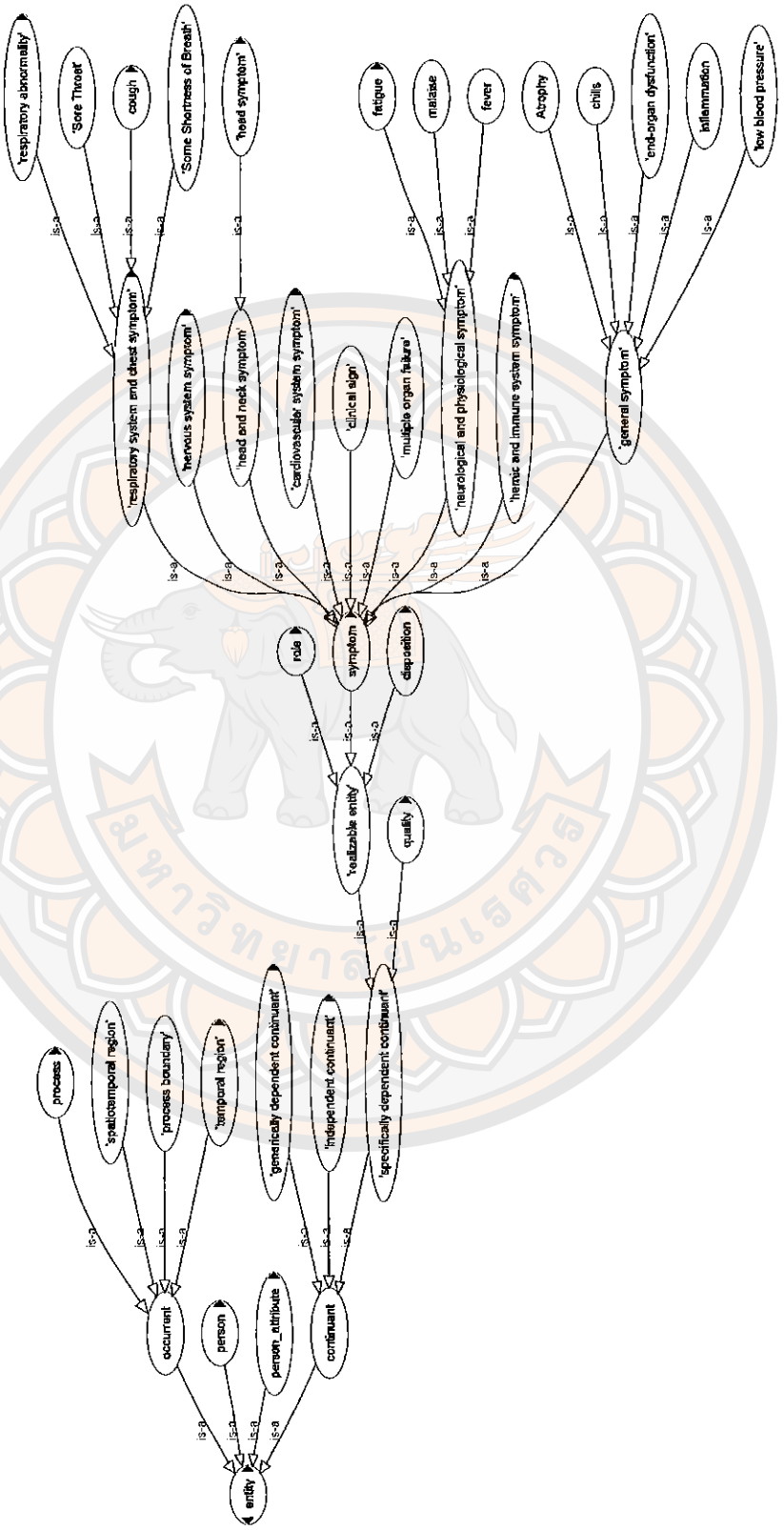
ภาคผนวก ข ตัวอย่างแนวความคิดในออนโทโลยีโรคหัวใจ



ภาพ 40 ตัวอย่างแนวความคิดในออนโทโลยีโรคหัวใจ

4918891183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

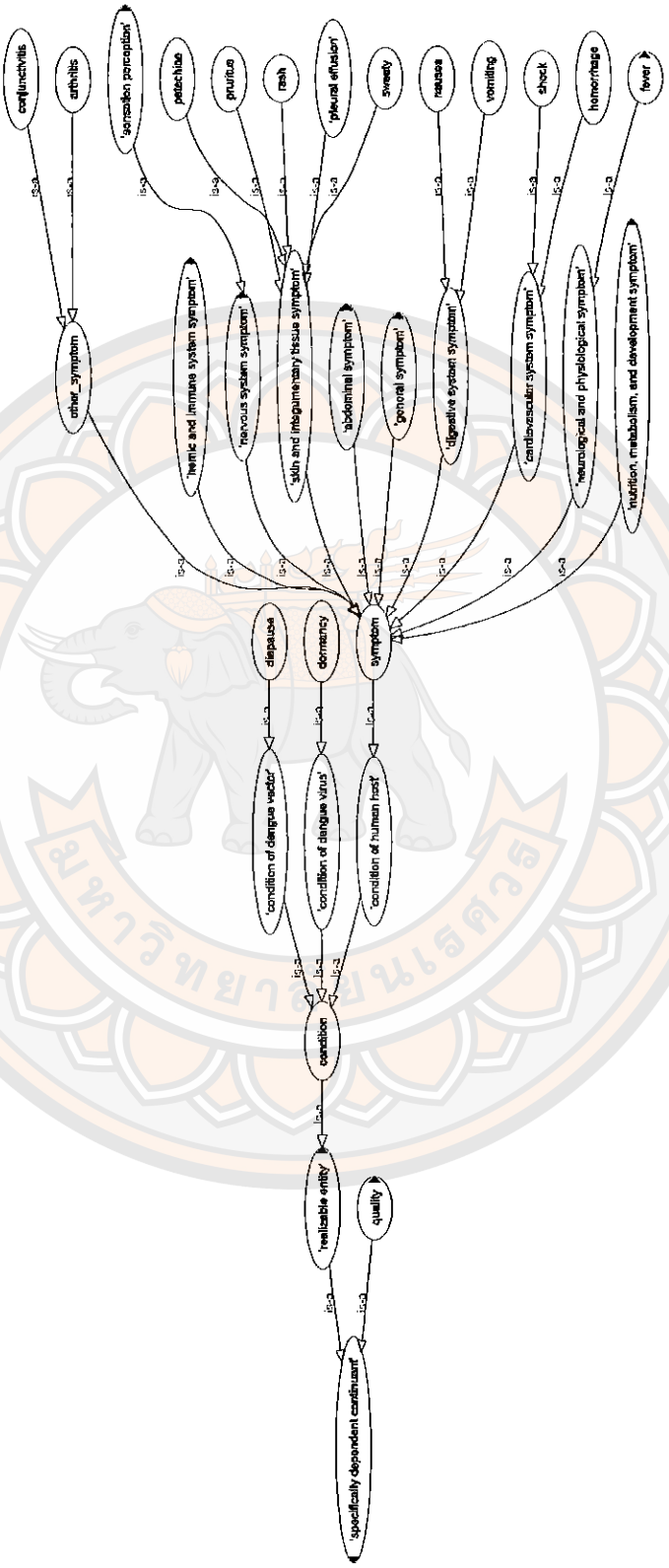
ภาคผนวก ค ตัวอย่างแนวความคิดในออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019



ภาพ 41 ตัวอย่างแนวความคิดในออนโทโลยีโรคติดเชื้อไวรัสโคโรนา 2019

4918889183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39

ภาคผนวก ง ตัวอย่างแนวความคิดในออนโทโลยีโรคไข้เลือดออก



ภาพ 42 ตัวอย่างแนวความคิดในออนโทโลยีโรคไข้เลือดออก

4210661183 NU iThesis 60031257 thesis / recv: 31102565 16:06:32 / seq: 39