

การเพิ่มประสิทธิภาพ Feature Selection สำหรับการจำแนก กลุ่มบทความวิจัย Optimization Feature Selection for Classification of Manuscript Grouping

ปะพาตา ณ วิเชียร¹ ภาควณิ มั่นแอ¹ ญาณพัฒน์ ชูชื่น² และสุภาวดี มากอัน³
Papada N wichian¹, Pakpum Manair¹, Yanapat Chuchuen²
and Supawadee Mak-on³

บทคัดย่อ

เทคนิคเหมืองข้อมูลในปัจจุบันมีบทบาทสำคัญกับการประมวลผลข้อมูล การสกัดข้อมูล และการสืบหาสารสนเทศ ด้วยการจำแนกข้อมูลถือเป็นเทคนิคที่ได้รับความนิยม ด้วยการสร้างต้นแบบการจำแนกมีข้อจำกัดของประสิทธิภาพ การเลือกคุณลักษณะ (Feature Selection) ด้วยหลักการทางสถิติแบบ Information Gain สำหรับตัดคุณลักษณะที่มีค่าน้ำหนักน้อย ที่เลือกมาประยุกต์กับการจำแนกข้อมูล Naïve Bayes และ Decision Tree เพื่อคัดแยกกลุ่มบทความวิจัย จากกรณีศึกษา งานประชุมวิชาการเครือข่ายด้านวิทยาศาสตร์ภาคใต้ครั้งที่ 4 จำนวน 177 บทความวิจัย แบ่ง 4 สาขา ได้แก่ ด้านคณิตศาสตร์ ด้านวิทยาศาสตร์สุขภาพ ด้านวิทยาการคอมพิวเตอร์ และด้านอื่น ๆ ที่เกี่ยวข้อง ให้ค่าความถูกต้องสูงสุดที่ 84% ด้วยการจำแนกแบบ Naïve Bayes ที่มีการเลือกคุณลักษณะค่าน้ำหนัก 0.03 ให้ค่าคุณลักษณะจำนวน 444 คำ แต่การจำแนกแบบ Decision Tree ให้ค่าความถูกต้องที่ 60% ที่จำนวนคุณลักษณะ 9 คำ ด้วยค่าน้ำหนักการเลือกคุณลักษณะ 0.20 โดยสรุปการเลือกคุณลักษณะร่วมกับการจำแนกแบบ Naïve Bayes ให้ต้นแบบการจำแนกที่มีประสิทธิภาพดีกว่าด้วยชุดข้อมูลกรณีศึกษา

คำสำคัญ: การทำเหมืองข้อมูล การเลือกคุณลักษณะ การจำแนกข้อมูล

Abstract

Currently, data mining can be very useful to data computational, retrieve important and relevant information. Classification is a popular data mining function. The limitation of the classification model must use a feature selection to investigate with the real system. This paper, we used the Information Gain approach, which is the

¹ สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสงขลา

² สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสงขลา

³ สาขาวิชาเทคโนโลยีคอมพิวเตอร์ วิทยาลัยรัตนภูมิ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย

* Corresponding author e-mail: yanapat.ch@skru.ac.th

one method of feature selection, to improve performance Naïve Bayes and Decision Tree model for a manuscript grouping. The case study used 177 manuscripts from National Science Conference 4th which has 4 groups including mathematic healthcare computer science and other related fields. The results shown the Naïve Bayes model with information gain weight 0.03 has the best accuracy at 84% and 444 features. Moreover, the Decision Tree model with information gain weight 0.20 has accuracy only 60% and 9 features. Concluding, the Naïve Bayes classification method with the information gain method can apply the best classification model with this case study.

Keywords: data mining, feature selection, classification

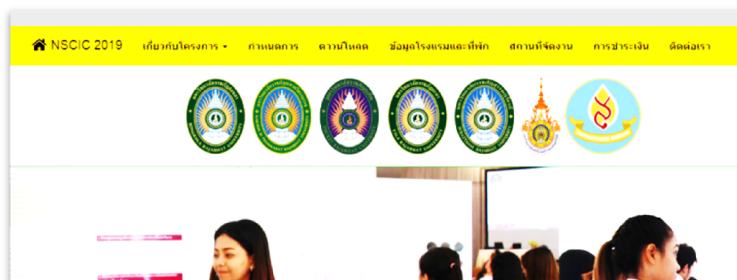
บทนำ

หลักการ Feature Selection มีความนิยมแพร่หลายขึ้นพร้อมกับการนำเทคนิคการทำเหมืองข้อมูล (Data Mining) ที่ได้รับความนิยมสำหรับการสกัดข้อมูลหรือสารสนเทศจากข้อมูลดังตัวอย่างได้แก่ การจราจรทางอากาศ การค้นหาเพลง และการแพร่ระบาดของโรค เป็นต้น ซึ่งการสร้างต้นแบบของข้อมูลในเชิงธุรกิจ หรือการนำไปใช้ประโยชน์ ด้วยกระบวนการเหมืองข้อมูลส่วนใหญ่พบว่าจะได้ชุดค่าจำนวนมากที่เรียกว่า คุณลักษณะ (Feature) หากนำมาใช้งานเพื่อการจำแนกข้อมูลจริงอาจใช้เวลามากสำหรับการเลือกคุณลักษณะ (Feature Selection) (Kan Deng, 1998) มีหลากหลายรูปแบบ ที่ได้รับความนิยม และรูปแบบพื้นฐาน คือการใช้สถิติเข้ามาช่วย และในงานวิจัยนี้ใช้หลักการที่กล่าวถึง ได้แก่ Filter approach และ Wrapper approach ซึ่งในงานวิจัยเลือกใช้กลุ่ม Filter approach ด้วยวิธีการ Information Gain ที่ใช้สัดส่วนของค่าที่ปรากฏในการพิจารณา โดยนำมาคำนวณหาค่าที่น้ำหนักเหมาะสม และนำไปเปรียบเทียบกับวิธีการจำแนกเดิมที่ไม่มีการใช้ Feature Selection ร่วมกับขั้นตอนการจำแนกพื้นฐาน Naïve Bayes และ Decision Tree บนโปรแกรม Rapidminer Studio โดยนำเสนอผลค่าความถูกต้องเปรียบเทียบในหัวข้อผลการวิจัย และอภิปรายผลการวิจัย

วิธีดำเนินการวิจัย

1. การเตรียมข้อมูลบทความวิจัย

ขั้นตอนการเตรียมข้อมูลและเอกสารที่จะนำมาใช้ในการวิจัย บทความวิจัยงานประชุมวิชาการเครือข่ายด้านวิทยาศาสตร์ภาคใต้ ครั้งที่ 4 มีการส่งผลงานบทความแยกตามสาขาได้แก่ คณิตศาสตร์ ฟิสิกส์ เคมี ชีววิทยา จุลชีววิทยา สิ่งแวดล้อม วิทยาการคอมพิวเตอร์ เทคโนโลยีสารสนเทศ การศึกษาวิทยาศาสตร์ สาธารณสุข/สุขภาพ ซึ่งได้ทำการจัดแบ่งเป็น 4 กลุ่ม ดังนี้ 1) วิทยาการคอมพิวเตอร์ 46 บทความ 2) คณิตศาสตร์ 16 บทความ 3) สาธารณสุข 58 บทความ 4) อื่น ๆ 57 บทความ โดยใช้ส่วนจากเนื้อหา 3 ส่วนดังนี้ ชื่อบทความ บทคัดย่อ และคำสำคัญ เฉพาะในส่วนภาษาอังกฤษเท่านั้น



ภาพที่ 1 เว็บไซต์งานประชุมวิชาการเครือข่ายด้านวิทยาศาสตร์ภาคใต้ ครั้งที่ 4

```
1 class, text
2 comsci, "The Development of Scientific Argumentation Ability and Decision Making Skill
Wutungsitawas School Yala Province using the Argument-Based Inquiry Approach in th
Natural Resources The purposes of this research were (1) to develop scientific argum
decision making skills; (3) to study the good teaching practices of argument-based
participants consisted of 7 Mathayom Suksa III students in Wutungsitawas School i
obtained by purposive sampling. The research instruments were comprised of (1) the
lesson; (2) a scientific argumentation ability test; and (3) decision making skills
using content analysis. The study found that (1) the post-learning scientific argum
learned using the argument-based inquiry approach were developing increasingly (2) A
skills; (3) In exploring the best practices in scientific learning using argument-b
found out the contributions that; firstly, the argumentative issues related to stude
students to learn and get engaged to argumentative activities. Secondly, allowing st
mixed-collaboratively might boost students to enhance their social skills and promot
problem-solving skills. Students, particularly, better gained motivation, responsibi
self-contentment in engaging scientifically argumentative activities. Thirdly, tea
students to realize that evidence and warrant are matter. Lastly, student-arranged b
students to be more responsible for their work. Solve problems create confidence, va
argumentative activities. the argument-based inquiry approach, scientific argumentati
skills, Mathayom Suksa"
3 comsci, "Developing Mathematics Achievement on Frequency Distribution by Graph of Gra
Integration in the Classroom based on TPACK and SAMR Model This research aims to dev
and SAMR Model on frequency distribution by graph through technology integration in
histogram and frequency polygon through GSP program, Plickers application, Quizizz a
application then study on students' mathematics achievement and attitude in teaching
created by the researcher. The sample are 28 grade 11 students of Satree Yala School
the second semester of 2018 academic year selected by purposive technique. Pretest w
the 4 lessons created by the researcher and posttest was given after finish teaching
that students' mathematics achievement on frequency distribution by graph after lear
```

ภาพที่ 2 ตัวอย่างข้อมูลไฟล์ CSV ของบทความวิจัยเฉพาะภาษาอังกฤษ

จากภาพที่ 2 ข้อมูลทั้งหมดในไฟล์แบ่งเป็นสองส่วนหลัก คือ ส่วนป้ายชุดข้อมูล (Label) ในการระบุประเภทของบทความ และเนื้อหา จาก 3 ส่วน ได้แก่ ชื่อเรื่อง บทคัดย่อ และ คำสำคัญ เฉพาะภาษาอังกฤษ เพื่อใช้ในการจำแนกบทความวิจัย ที่มีข้อดีในการตัดคำเพราะภาษาอังกฤษจะแยกคำแล้วทำให้สะดวกในการนำเข้าไฟล์ได้ทันที แตกต่างกับภาษาไทยที่ต้องใช้โปรแกรมช่วยแยกคำในตอนเริ่มต้น

การพัฒนาโปรแกรมภาษาไพทอน กรองคำศัพท์ที่ไม่เกี่ยวข้องได้แก่ ตัวเลข เครื่องหมาย ไวยากรณ์สัญลักษณ์ทางคณิตศาสตร์ คำสรรพนาม (Pronouns) คำบุพบท (Preposition) คำสันธาน (Conjunctions) และ Verb to be ตามตัวอย่าง ได้แก่ and yet but for so nor และ neither เพื่อสร้างพจนานุกรมคำศัพท์ และใช้ชุดโปรแกรมภาษาไพทอน จากเนื้อหาภาษาอังกฤษ 3 ส่วน ในบทความวิจัย ชื่อบทความ บทคัดย่อ และคำสำคัญ

```
for v in a :
    if v in Dword :
        del Dword[v]
Dsort = sorted(Dword.items(),key=lambda x: x[1],reverse=True)
print 'keep : ',len(Dword.keys())
print Dsort
```

ภาพที่ 3 โปรแกรมภาษาไพทอนสำหรับกรองคำศัพท์

ตามภาพที่ 3 โปรแกรมทำงานกรองคำศัพท์ตามลำดับดังนี้ ข้อความ (text) ที่ตรงกับคำศัพท์ที่เก็บอยู่อาร์เรย์ (array) จะถูกลบ ออกด้วยคำสั่ง del และ แสดงจำนวนคำศัพท์ที่เหลือ

2. การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล คือกระบวนการสกัดความสัมพันธ์หรือค้นแบบจากชุดข้อมูล เป็นกระบวนการสืบค้นความรู้ให้ปรากฏประโยชน์และข้อมูลที่สำคัญบนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very Large Database : KDD) และใช้หลักการของการพยากรณ์ ข้อมูลชุดใหม่ในอนาคต การค้นหาความรู้และความจริงที่แฝงในข้อมูลเหล่านี้ (Knowledge discovery) มีความแตกต่างจากระบบฐานข้อมูล (Database system) ที่ทำงานด้วยภาษา SQL ด้วยฐานข้อมูลนั้นสามารถแสดงได้เพียงความสัมพันธ์ในมิติพื้นฐานเท่านั้น หากต้องการข้อมูลเชิงลึกมีความต้องการใช้ทรัพยากรที่สูงขึ้น แต่การทำเหมืองข้อมูลมีวิธีการการเรียนรู้กลไก ด้วยขั้นตอนวิธีที่นิยมในการจำแนกข้อมูลอย่างเช่น ต้นไม้ตัดสินใจ และโครงข่ายประสาทเทียม เป็นต้น สำหรับเทคนิคการทำเหมืองข้อมูลที่นิยมใช้ในการทำงานวิจัยได้แก่ กฎความสัมพันธ์ (Association rule) การจำแนกประเภทข้อมูล (Data classification) การแบ่งกลุ่มข้อมูล (Data clustering) และการสร้างมโนภาพ (Visualization) โดยงานวิจัยนี้เน้นการจำแนกข้อมูลด้วยวิธีพื้นฐาน ได้แก่ นาอ์ฟเบย์ และ ต้นไม้ตัดสินใจ เพื่อสร้างต้นแบบสำหรับการจำแนกกลุ่มบทความวิจัย ซึ่งทำการปรับปรุงคุณภาพของต้นแบบด้วยวิธีการเลือกค่าคุณลักษณะ เพื่อการพัฒนาไปสู่การใช้งานจริงที่มีข้อจำกัดของทรัพยากร และเวลาในการประมวลผลผลลัพธ์ที่มีค่าความถูกต้องเหมาะสม

2.1 การเลือกค่าคุณลักษณะ (Feature Selection)

การคัดเลือกค่าสำคัญด้วยกระบวนการคัดเลือกคุณลักษณะ (feature selection) (เอกสิทธิ์ พชรวงศ์ศักดิ์, 2557) แบ่งได้เป็น 2 กลุ่ม ได้แก่ Filter approach และ Wrapper approach สำหรับงานวิจัยนี้เลือกใช้ Filter approach ด้วยหลักการคัดเลือกฟีเจอร์ โดยใช้การคำนวณค่าค่าน้ำหนักหรือค่าความสัมพันธ์ระหว่างแต่ละฟีเจอร์และคลาสต่างๆ และการคัดเลือกฟีเจอร์ด้วยการเรียงลำดับตามค่าค่าน้ำหนักที่คำนวณ และทำการเลือกฟีเจอร์ที่มีค่าน้ำหนักมากกว่าที่ต้องการมาใช้งาน

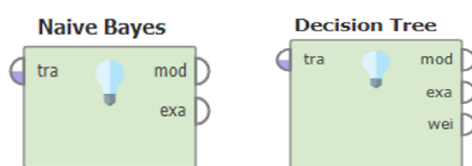
โดย Filter approach เป็นการเลือกฟีเจอร์ตามค่าค่าน้ำหนักที่คำนวณได้และเรียงลำดับตามค่าค่าน้ำหนักมากมาใช้งาน ที่คำนวณจากค่าความสัมพันธ์ระหว่างแต่ละฟีเจอร์ ซึ่งเทคนิคในการคำนวณมีดังนี้

- Information Gain การทำงานหลักจะเลือกใช้ Probability หรือความน่าจะเป็น เมื่อนำมารวมกัน Entropy เหมาะสมกับเงื่อนไขที่ไม่ได้ซับซ้อนมาก
- Chi-Square มีหลักการคำนวณน้ำหนักของแอตทริบิวต์ ด้วยหลักการทางสถิติไคสแควร์ กล่าวคือน้ำหนักของแอตทริบิวต์สูงถือว่ามีความเกี่ยวข้องสูงตามกัน

Wrapper approach เป็นการคำนวณค่าค่าน้ำหนักโดยใช้ Model แบ่งได้ 2 แบบใหญ่คือ Forward Selection คือ Model ที่ทำการเพิ่มคุณลักษณะครั้งละ 1 คุณลักษณะ ถ้าคุณลักษณะที่ใส่เพิ่มให้ประสิทธิภาพที่ดีจะเก็บไว้และเลือกคุณลักษณะอื่น ๆ มาเพิ่มต่อ และ Backward Elimination ทำงานกลับกัน คือ ตัดคุณลักษณะไม่สำคัญทิ้งไปทีละ 1 คุณลักษณะ และทำการประเมินหาประสิทธิภาพ

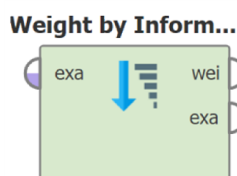
2.2 การทดสอบด้วยโปรแกรม RapidMiner Studio

เครื่องมือในการวิเคราะห์ข้อมูลด้วย Data Mining เป็นที่นิยมไปทั่วโลก การมีซอฟต์แวร์ที่ช่วยให้การวิเคราะห์ได้ง่ายขึ้น เหมาะกับการวิเคราะห์ข้อมูลที่มีขนาดใหญ่และจำนวนมากในงานวิจัยนี้ได้สร้าง Model จำแนกข้อมูล Decision Tree และ Naïve Bayes โดยขั้นตอนในการดำเนินการจำแนกบทความวิจัยตามสาขาด้านวิทยาศาสตร์ด้วยเทคนิคการทำเหมืองข้อมูล ได้แก่ Naïve Bayes และ Decision Tree ที่นิยมในการทำเหมืองข้อมูลในปัจจุบัน



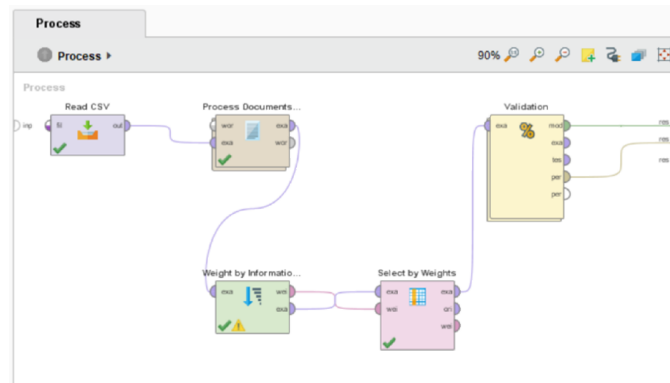
ภาพที่ 4 ชุดเครื่องมือ Naïve Bayes และ Decision Tree ในการจำแนกข้อมูลบนโปรแกรม RapidMiner Studio

เครื่องมือที่สำคัญของการเลือกคุณลักษณะด้วย Information Gain ซึ่งในโปรแกรม RapidMiner Studio มีการเลือกใช้งานร่วมกันกับชุดจำแนกข้อมูล Naïve Bayes และ Decision Tree ในภาพที่ 5 และ ภาพที่ 6



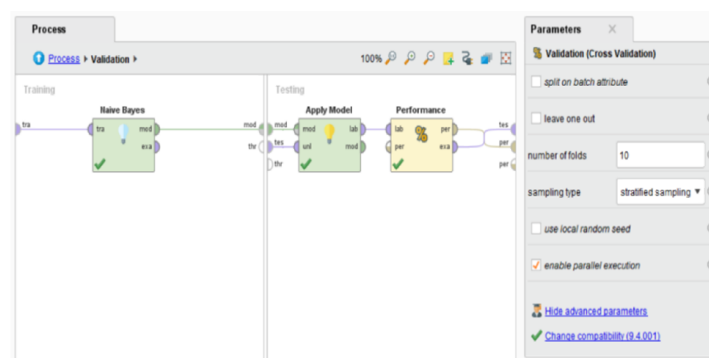
ภาพที่ 5 ชุดเครื่องมือเลือกคุณลักษณะด้วย Information Gain บนโปรแกรม RapidMiner Studio

การทดสอบเมื่อเตรียมข้อมูลของบทความวิจัย และทำการแยกคำที่ไม่ต้องการ ได้แก่ ตัวเลข เครื่องหมายไวยากรณ์สัญลักษณ์ทางคณิตศาสตร์ คำสรรพนาม (Pronouns) คำบุพบท (Preposition) คำสันธาน (Conjunctions) และ Verb to be ตามตัวอย่าง ได้แก่ and yet but for so nor และ neither เพื่อเข้ามาใช้การประมวลผลเอกสาร (Process Documents) ที่มีขั้นตอน TFIDF พร้อมกับการให้ค่าน้ำหนักที่มีการทดสอบค่าน้ำหนักตั้งแต่ 0.01 ถึง 0.25 เพื่อนำผลของค่าความถูกต้อง และ จำนวนคุณลักษณะ มาพิจารณาค่าน้ำหนักที่ดีที่สุด สำหรับการใช้งานในต้นแบบจริง เพื่อเปรียบเทียบประสิทธิภาพกับ การจำแนกด้วยต้นแบบที่มีคุณลักษณะตามชุดข้อมูล



ภาพที่ 6 การประมวลผลและประเมินประสิทธิภาพการวิจัยบนโปรแกรม RapidMiner Studio

ตามภาพที่ 6 มีส่วนของ Validation ซึ่งเป็นการประเมินประสิทธิภาพ ที่ภายในประกอบด้วย Naïve Bayes และ Decision Tree ในส่วนนี้มีชุดเครื่องมือ Apply Model และ Performance ทำหน้าที่นำต้นแบบที่สร้างขึ้น มาทำการทดสอบในรูปแบบ 10 folds คือการสลับชุดข้อมูลย่อย เพื่อสร้างต้นแบบ และทดสอบ ที่นิยมในการทำเหมือนข้อมูล ตามภาพที่ 7



ภาพที่ 7 ตัวอย่างการจำแนกข้อมูลด้วยเทคนิค Naïve Bayes บนโปรแกรม RapidMiner Studio

2.3 งานวิจัยที่เกี่ยวข้อง

การจำแนกกลุ่มข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลได้รับการประยุกต์ใช้งานต่อเนื่องทั้งในการวิจัย และการทำงานจริง โพสต์เกมด้วยเทคนิคเหมืองข้อมูล (นุรุลีมาน เบ็งบุงอและคณะ, 2561) เปรียบเทียบโมเดลสำหรับสังคมออนไลน์ ด้วยการนำเทคนิคเหมืองข้อมูลเพื่อมาพัฒนาต้นแบบโมเดลตัวจำแนกที่มีประสิทธิภาพที่ดีด้วยขั้นตอน Naïve Bayes และ Decision Tree พบค่านิยมมีค่าความถี่ที่สูง ได้แก่ ดี แข็ง เทพ ป้อม และแรงค์ เป็นต้น มาเป็นแนวคิดในการออกแบบการวิจัยสำหรับจำแนกกลุ่มบทความวิจัย และ มีการประยุกต์ใช้ในการวางแผนเรียน เพื่อแนะนำรายวิชา ร่วมกับการจัดแผนการศึกษา ตัวกระบวนการค้นหาเฉพาะที่ รายการทาบู (ธาดา หวังธรรมมั่งและคณะ, 2560) แต่งานวิจัยนี้นำขั้นตอนวิธีการจำแนกเหล่านี้มาใช้กับชุดข้อมูลบทความของบทความวิจัย เพื่อจัดกลุ่มของบทความวิจัยในเบื้องต้นสำหรับการจัดหาผู้ทรงคุณวุฒิประจำสาขาในการพิจารณาบทความ ของระบบบริหารจัดการงานประชุมวิชาการ และเพิ่มประสิทธิภาพ

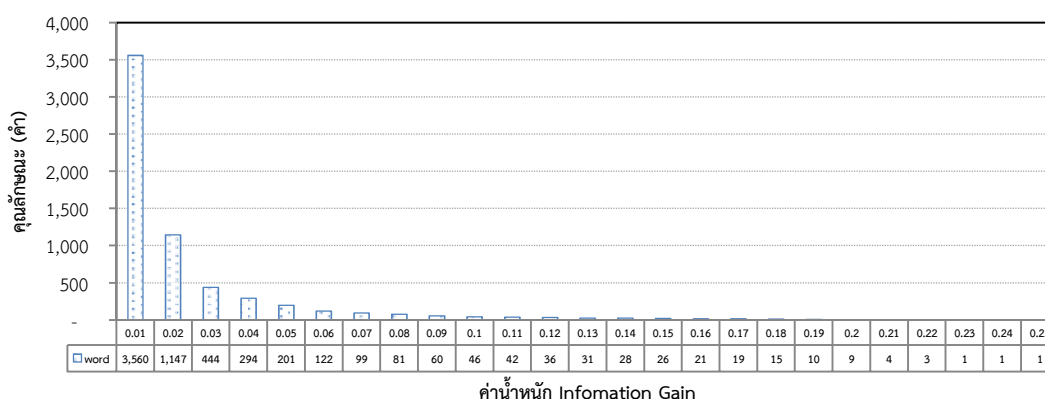
ด้วยการคัดแยกคำ ซึ่งแตกต่างไปจากการใช้เพียงการทดสอบสมมติฐานทางสถิติ (ก้องนาคินทร์ จรรย์วุฒิศและคณะ, 2561)

ผลการวิจัย

ตามขั้นตอนที่กล่าวไว้สำหรับการทดสอบ ให้ผลลัพธ์ที่น่าเสนอไว้ ในส่วนแรกคือค่าน้ำหนักของการเลือกคุณลักษณะด้วย Information Gain เพื่อให้ได้ค่าน้ำหนักที่เหมาะสมกับต้นแบบตามภาพที่ 8 และภาพที่ 9 สำหรับการเปรียบเทียบค่าความถูกต้องระหว่าง Naïve Bayes และ Decision Tree ที่นำเสนอพร้อมอธิบายในส่วนของการวิจัยที่มีการปรับปรุงน้ำหนักในการเลือกคุณลักษณะด้วย Information Gain

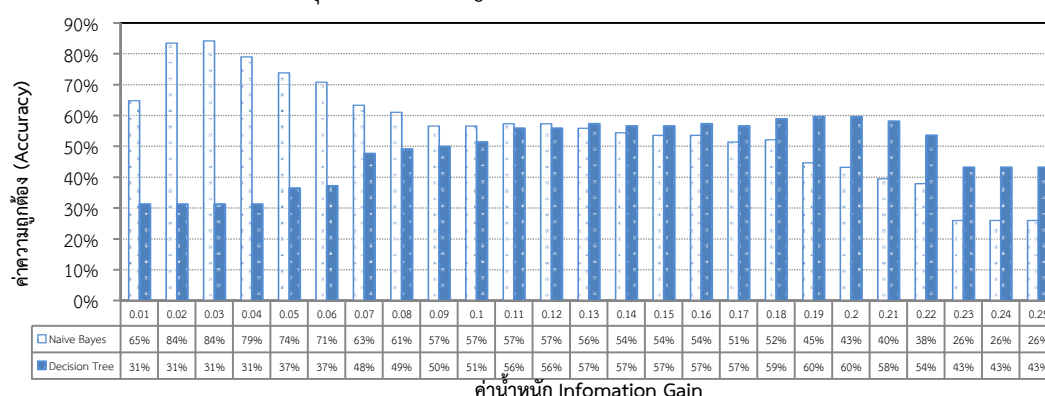
1. ประสิทธิภาพการเลือกคุณลักษณะด้วย Information Gain

ภาพที่ 8 แสดงค่าน้ำหนักของการเลือกคุณลักษณะด้วย Information Gain ที่มีจำนวนของคุณลักษณะ หรือคำ แสดงควบคู่กัน โดยประสิทธิภาพของค่าความถูกต้องนำเสนอในภาพที่ 9



ภาพที่ 8 จำนวนคุณลักษณะตามค่าน้ำหนักของ Information Gain

ข้อมูลในกราฟพบว่า เมื่อเพิ่มค่าน้ำหนักไปที่มากกว่า 0.06 จำนวนคุณลักษณะลดลงที่ค่า 122 ซึ่งมีค่ามากกว่า 10% ของคุณลักษณะเดิมจากชุดข้อมูล และเหลือเพียง 1 คุณลักษณะที่ค่าน้ำหนัก 0.23 ขึ้นไป ด้วยสาเหตุของค่าส่วนใหญ่กระจายตัวที่ระดับน้อยกว่า 0.20 เป็นต้นไป



ภาพที่ 9 การเปรียบเทียบค่าความถูกต้องระหว่าง Naïve Bayes และ Decision Tree ในแต่ละน้ำหนักของ Information Gain

ผลของกราฟในภาพที่ 9 สำหรับการจำแนกแบบ Naïve Bayes มีค่าความถูกต้องสูงที่สุดที่ 84% ที่ค่าน้ำหนัก 0.03 และต่ำกว่า 20% ที่ค่าน้ำหนักมากกว่า 0.18 แตกต่างจากการจำแนกข้อมูลแบบ Decision Tree ให้ค่าความถูกต้องที่ค่าน้ำหนัก 0.20 ที่ 60% ด้วยพื้นฐานของการจำแนกแตกต่างกัน และมีค่าสูงตั้งแต่ค่าน้ำหนัก 0.08 ถึง 0.23

2. ประสิทธิภาพการเลือกคุณลักษณะด้วย Information Gain ร่วมกับเทคนิคการจำแนกข้อมูล Naïve Bayes

โดยภาพรวมสำหรับค่าความถูกต้องในส่วนที่กล่าวมา ในรายละเอียด ที่นำเสนอในเชิงลึกของค่าความแม่นยำ (precision) แยกตามกลุ่มของบทความวิจัย ที่น้ำหนัก 0.01 และ 0.03 (กรณีดีที่สุดของ Naïve Bayes) และ ค่าน้ำหนัก 0.20 (กรณีดีที่สุดของ Decision Tree) พร้อมกับค่าระลึก (Recall)

2.1 กรณีการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.01

accuracy: 64.89% +/- 16.02% (micro average: 64.93%)

	true comsci	true health	true math	true other	class precision
pred. comsci	27	5	3	4	69.23%
pred. health	7	33	8	10	56.90%
pred. math	0	0	1	0	100.00%
pred. other	2	6	2	26	72.22%
class recall	75.00%	75.00%	7.14%	65.00%	

ภาพที่ 10 ค่าความแม่นยำและค่าความถูกต้องของการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.01 ร่วมกับเทคนิคการจำแนกข้อมูล Naïve Bayes

ตามภาพที่ 10 ค่าความแม่นยำสูงสุดด้านคณิตศาสตร์ (math) กลุ่มอื่นๆ (other) ด้านคอมพิวเตอร์ (comsci) และ ด้านสุขภาพ (health) ตามลำดับ ค่าความระลึกมีค่าสูงสุดเท่ากันในด้าน คอมพิวเตอร์ และ ด้านสุขภาพ

2.2 กรณีการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.03

ที่ค่าน้ำหนัก 0.03 การจำแนกแบบ Naïve Bayes ปรากฏค่าความถูกต้องสูงสุด ตามภาพที่ 9 โดยมีรายละเอียด ค่าความแม่นยำ ด้านคณิตศาสตร์ ที่ 100% กลุ่มอื่นๆ (other) และด้านคอมพิวเตอร์ เท่ากันที่ 89.47% และ ด้านสุขภาพ (health) 75.93%

accuracy: 84.29% +/- 9.07% (micro average: 84.33%)

	true comsci	true health	true math	true other	class precision
pred. comsci	34	1	2	1	89.47%
pred. health	2	41	6	5	75.93%
pred. math	0	0	4	0	100.00%
pred. other	0	2	2	34	89.47%
class recall	94.44%	93.18%	28.57%	85.00%	

ภาพที่ 11 ค่าความแม่นยำและค่าความถูกต้องของการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.03 ร่วมกับเทคนิคการจำแนกข้อมูล Naïve Bayes

ภาพที่ 11 สำหรับผลของค่าระลึกมีค่าสูงสุด ด้านคอมพิวเตอร์ที่ 94.44% ด้านสุขภาพที่ 93.18% กลุ่มอื่น ๆ ที่ 85.00% แต่น้อยกว่าครึ่งสำหรับด้านคณิตศาสตร์ที่ 28.57%

2.3 กรณีการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.2

accuracy: 43.24% +/- 10.82% (micro average: 43.28%)

	true comsci	true health	true math	true other	class precision
pred. comsci	26	9	0	4	66.67%
pred. health	1	0	0	2	0.00%
pred. math	9	33	14	16	19.44%
pred. other	0	2	0	18	90.00%
class recall	72.22%	0.00%	100.00%	45.00%	

ภาพที่ 12 ค่าความแม่นยำและค่าความถูกต้องของการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.2 ร่วมกับเทคนิคการจำแนกข้อมูล Naïve Bayes

ภาพที่ 12 การจำแนกแบบ Decision Tree ปรากฏค่าความถูกต้องสูงสุด ตามภาพที่ 9 โดยมีรายละเอียด ค่าความแม่นยำ สำหรับผลของค่าระลึกราคาสูงสุด กลุ่มอื่น ๆ ที่ 90.00% ด้านคอมพิวเตอร์ที่ 66.67% แต่มีค่า 19.44% สำหรับด้านคณิตศาสตร์ และ 0% สำหรับด้านสุขภาพ

3. ประสิทธิภาพการเลือกคุณลักษณะด้วย Information Gain ร่วมกับเทคนิคการจำแนกข้อมูล Decision Tree

เช่นเดียวกับภาพรวมสำหรับค่าความถูกต้อง ของเทคนิคการจำแนกข้อมูล Decision Tree ที่นำเสนอในเชิงลึกของค่าความแม่นยำ (precision) แยกตามกลุ่มของบทความวิจัย ที่ น้ำหนัก 0.01 และ 0.03 (กรณีที่ดีที่สุดของ Naïve Bayes) 0.20 (กรณีที่ดีที่สุดของ Decision Tree) พร้อมกับค่าระลึกราคา (Recall)

3.1 กรณีการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.01

accuracy: 31.43% +/- 6.27% (micro average: 31.34%)

	true comsci	true health	true math	true other	class precision
pred. comsci	0	0	0	0	0.00%
pred. health	34	41	13	38	32.54%
pred. math	2	3	1	2	12.50%
pred. other	0	0	0	0	0.00%
class recall	0.00%	93.18%	7.14%	0.00%	

ภาพที่ 13 ค่าความแม่นยำและค่าความถูกต้องที่การเลือกคุณลักษณะ Information Gain น้ำหนัก 0.01 ร่วมกับเทคนิคการจำแนกข้อมูล Decision Tree

ค่าน้ำหนักพื้นฐานของการเลือกคุณลักษณะ ให้ผลค่าความแม่นยำ ด้านสุขภาพที่ 32.54% และ ด้านคณิตศาสตร์ที่ 12.50% แต่ปรากฏด้านอื่น ๆ ที่ 0% สำหรับค่าความระลึกราคา ด้านสุขภาพสูงสุดที่ 93.18% แต่ด้านอื่น ๆ มีค่าใกล้เคียง 0%

3.2 กรณีการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.03

accuracy: 31.43% +/- 6.27% (micro average: 31.34%)

	true comsci	true health	true math	true other	class precision
pred. comsci	1	1	0	0	50.00%
pred. health	33	40	13	38	32.26%
pred. math	2	3	1	2	12.50%
pred. other	0	0	0	0	0.00%
class recall	2.78%	90.91%	7.14%	0.00%	

ภาพที่ 14 ค่าความแม่นยำและค่าความถูกต้องของการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.03 ร่วมกับเทคนิคการจำแนกข้อมูล Decision Tree

ภาพที่ 14 สำหรับการจำแนกแบบ Decision Tree มีค่าความแม่นยำด้านคอมพิวเตอร์ที่ 50% และให้ผลลัพธ์ใกล้เคียงกับที่น้ำหนัก 0.01 เช่นเดียวกับค่าความระลึก

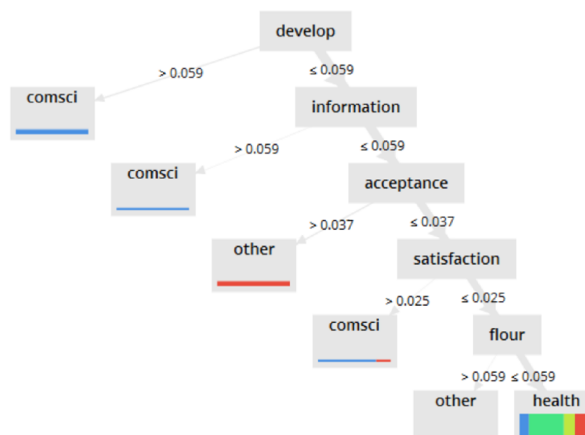
3.3 กรณีการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.2

accuracy: 59.73% +/- 12.30% (micro average: 59.70%)

	true comsci	true health	true math	true other	class precision
pred. comsci	24	1	0	1	92.31%
pred. health	12	41	14	24	45.05%
pred. math	0	0	0	0	0.00%
pred. other	0	2	0	15	88.24%
class recall	66.67%	93.18%	0.00%	37.50%	

ภาพที่ 15 ค่าความแม่นยำและค่าความถูกต้องของการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก 0.2 ร่วมกับเทคนิคการจำแนกข้อมูล Decision Tree

ค่าน้ำหนักที่ดีที่สุดที่ 0.20 ซึ่งการจำแนกแบบ Decision Tree ให้ผลลัพธ์ค่าความถูกต้องสูงที่สุดที่ 59.73% มีค่าความแม่นยำด้านคอมพิวเตอร์ 92.31% กลุ่มอื่น ๆ 88.24% แต่น้อยกว่า 50% สำหรับด้านสุขภาพ และด้านคณิตศาสตร์ที่ 0% ซึ่งต้นไม้มัดสินใจของต้นแบบที่ค่าน้ำหนัก 0.20 ตามภาพที่ 16



ภาพที่ 16 ค่าความแม่นยำและค่าความถูกต้องของการเลือกคุณลักษณะด้วย Information Gain น้ำหนัก

อภิปรายผลการวิจัย

การทดสอบเพื่อให้ได้ค่าน้ำหนักที่ดีที่สุด และค่าความถูกต้องสูงสุดสำหรับการเลือกคุณลักษณะด้วย Information Gain ตามผลลัพธ์ในภาพที่ 8 และภาพที่ 9 สำหรับการเปรียบเทียบค่าความถูกต้องระหว่าง Naïve Bayes และ Decision Tree การเลือกคุณลักษณะด้วย Information Gain กรณีจำแนกข้อมูลด้วย Naïve Bayes ได้ผลลัพธ์ค่าน้ำหนัก 0.03 ให้ค่าความถูกต้องที่ 84% ที่มีค่าสูงกว่าค่าความถูกต้องในการจำแนกแบบ Decision Tree ที่ค่าน้ำหนัก 0.20 ให้ค่าความถูกต้องเพียง 60% ด้วยค่าน้ำหนักที่เลือกคุณลักษณะ Information Gain ผู้วิจัยจึงเลือกที่ 0.03 และการจำแนกแบบ Naïve Bayes และ Decision Tree มาเปรียบเทียบกับชุดทดสอบที่ไม่มีการเลือกคุณลักษณะที่มีจำนวนคุณลักษณะครบถ้วน เพื่อศึกษาความสามารถของต้นแบบเมื่อได้ลดจำนวนของคุณลักษณะในตารางที่ 1

ตารางที่ 1 การเปรียบเทียบค่าความถูกต้อง และจำนวนคุณลักษณะของ การเลือกคุณลักษณะ Information Gain

Operators	Naïve Bayes	Decision Tree	weight	word
ไม่ใช้การเลือกคุณลักษณะ	64.93%	31.34%		3562
การเลือกคุณลักษณะ Information Gain	84.33%	31.34%	0.03	444

ตามตารางที่ 1 ปรากฏผลลัพธ์สำหรับการจำแนกแบบ Naïve Bayes ด้วยค่าความถูกต้อง 84.33% ที่สูงกว่าการไม่เลือกคุณลักษณะ แต่การจำแนกแบบ Decision Tree ให้ค่าความถูกต้องเท่ากับการไม่เลือกคุณลักษณะ โดยอธิบายผลลัพธ์ที่ปรากฏด้วยค่าคุณลักษณะที่ลดลงสนับสนุนการใช้ความน่าจะเป็นทำให้เพิ่มประสิทธิภาพการจำแนกแบบ Naïve Bayes ด้วยคุณลักษณะ 444 คำ ปรากฏข้อสนับสนุนตามภาพที่ 9 เนื่องจากเมื่อคุณลักษณะลดลงน้อยกว่า 10% ของจำนวนคุณลักษณะค่าความถูกต้องมีการปรับลงอย่างชัดเจน ในกรณีการจำแนกแบบ Decision Tree มีค่าความถูกต้องสูงที่จำนวนคุณลักษณะ 9 คำ ด้วยเนื่องจากต้นแบบมีขนาดลดลงจากจำนวนคุณลักษณะ

สรุปผลการวิจัย

ด้วยผลการวิจัย และการเปรียบเทียบประสิทธิภาพด้วยค่าความถูกต้องของต้นแบบดั้งเดิมที่ไม่ใช้การเลือกคุณลักษณะ ให้ค่าความถูกต้องของการจำแนกแบบ Naïve Bayes ด้วยหลักการของความน่าจะเป็น กับจำนวนคุณลักษณะ 444 คำ ที่ 84.33% สูงกว่าอย่างชัดเจน และตอบข้อสมมติฐานของจำนวนคุณลักษณะที่ลดลงมีผลกระทบกับค่าความถูกต้องของการจำแนกกลุ่มทั้งสองแบบ พบว่า Naïve Bayes ให้ประสิทธิภาพดีที่ค่าน้ำหนัก 0.03 แต่ Decision Tree จำแนกกลุ่มได้ด้อยกว่าและปรากฏประสิทธิภาพดีที่คุณลักษณะที่ระดับน้อยกว่า 10 คำ ซึ่งในรายละเอียดไม่ปรากฏความแม่นยำแต่ละกลุ่มแตกต่างกัน หรือกล่าวได้ว่ามีความแม่นยำเฉพาะเพียงบางกลุ่มเท่านั้น แต่การศึกษาวิจัยเพียงบางเทคนิคเท่านั้น เพราะเทคนิคการทำเหมือนข้อมูลมีขั้นตอนวิธีหลากหลายที่น่าสนใจได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (Support vector machine) และโครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) ที่อาจพัฒนาต่อยอดในอนาคตได้

กิตติกรรมประกาศ

ขอขอบคุณคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสงขลา และหลักสูตร
วิทยาการคอมพิวเตอร์ที่สนับสนุนข้อมูล และเครื่องมือสำหรับการทำวิจัยให้บรรลุผลสำเร็จ

เอกสารอ้างอิง

Kan Deng. (1998). OMEGA: ON-LINE MEMORY-BASED GENERAL PURPOSE SYSTEM
CLASSIFIER. The Robotics Institute School of Computer Science Carnegie
Mellon University Pittsburgh, PA 15213.

ก้องนาคินทร์ จรรย์วุฒิวงศ์ และดวงรัตน์ โกยกิจเจริญ (2561). “การรับรู้คุณภาพบริการที่ส่งผลต่อ
การตัดสินใจเลือกใช้บริการสายการบินต้นทุนต่ำ ท่าอากาศยานนานาชาติภูเก็ต.” ใน
การประชุมวิชาการระดับชาติ มหาวิทยาลัยราชภัฏภูเก็ต ครั้งที่ 11. ภูเก็ต:
มหาวิทยาลัยราชภัฏภูเก็ต, 739 – 746.

ธาดา หวังธรรมมั่ง เสกสรรสุวรรณมณี ธัชชัย เองฉ้วน และแสงสุรีย์ วสุพงศ์อัยยะ(2560). “การ
พัฒนาโปรแกรมช่วยจัดแผนการเลือกวิชาลงทะเบียนโดยใช้เทคนิคการ ค้นหาเฉพาะที่
และรายการทาบู.” ใน วารสารมหาวิทยาลัยทักษิณ ปีที่ 22, ฉบับที่ 2. สงขลา:
มหาวิทยาลัยทักษิณ, 275 – 283.

นุรุลอี้มาน เบ็งบูงอ ฮัสมีะ เร๊ะคุมหลี กฤษณ์วรา รัตนโอภาส และยุพดี อินทส (2561). “การ
จำแนกโพสต์เกมโดยใช้เทคนิคเหมืองข้อมูล : กรณีศึกษาการโพสต์บนเฟซบุ๊ก.” ใน การ
ประชุมวิชาการระดับชาติ มหาวิทยาลัยราชภัฏภูเก็ต ครั้งที่ 11. ภูเก็ต: มหาวิทยาลัย
ราชภัฏภูเก็ต, 547 – 555.

เอกสิทธิ์ พัทธวงศ์ศักดิ์ (2557). การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมนิง เบื้องต้น. พิมพ์ครั้งที่
ที่ 1. กรุงเทพมหานคร : บริษัท เอเชีย ดิจิตอลการพิมพ์ จำกัด.