การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลประเภทของสัตว์โดยใช้เทคนิคเหมืองข้อมูล

นิติมา ลักขณานุรักษ์[1] จารุกิตติ์ มณีสุข[2] วิภาวดี หุ่นทอง[2] ไก้รุ่ง เฮงพระพรหม[2*] และสุพจน์ เฮงพระพรหม[2]

[1]สาขาวิชาเทคโนโลยีสารสนเทศและนวัตกรรมดิจิทัล คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร
[2]สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[*]kairung2011.heng@gmail.com

บทคัดย่อ

การวิจัยในครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการจำแนกข้อมูลประเภทของสัตว์ 2 เทคนิค ได้แก่ เทคนิคการจำแนกประเภทข้อมูล JRIP และเทคนิคต้นไม้ตัดสินใจ (Decision Tree) J48 เพื่อหาประสิทธิภาพของเทคนิคที่ดีที่สุดเพื่อการจำแนกข้อมูลประเภทของสัตว์ ผลการวิจัยพบว่า เทคนิคที่ให้ประสิทธิภาพในการจำแนกข้อมูลประเภทของสัตว์ได้ดีที่สุดคือเทคนิคต้นไม้ตัดสินใจหรือ J48 โดยให้ค่าความแม่นยำในการจำแนกข้อมูลที่ 97.08 % และรองลงมา คือ JRIP ซึ่งให้ค่าความแม่นยำในการจำแนกข้อมูลที่ 87.13 % ตามลำดับ

คำสำคัญ: การจำแนกข้อมูล ต้นไม้ตัดสินใจ J48 JRIP

A Comparison of Animal Type Classification Efficiency using Data Mining Techniques

Nitima Lukkananuruk[1] Jarukit Maneesuk[1] Wipawadee Hoonthong[1] Kairung Hengpraphorm[2*] and Supojn Hengpraphorm[2]

[1]Information Technology and Digital Innovation, Faculty of Science, Silpakorn University

[2]Data Science, Faculty of Science and Technology, Nakhon Pathom Rajabhat University

[*]kairung2011.heng@gmail.com

Abstract

The aim of this research is to compare the efficiency of appropriate with the animal classification techniques. Two data classification techniques: JRIP and Decision Tree (J48) are used to find the efficiency of animal classification. The experimental results show that Decision Tree (J48) provides the best classification accuracy for animal data (the accuracy of data classification = 97.08%). The second technique is JRIP which provides the accuracy of data classification = 87.13% respectively.

Keywords: Classification, Decision Tress, J48, JRIP

1. บทนำ

สัตว์เป็นสิ่งมีชีวิตที่อาศัยอยู่บนโลกใบนี้เป็นระยะเวลายาวนาน สัตว์อาศัยกระจายอยู่ตามพื้นที่แทบทุกส่วนของโลก ทั้งในน้ำ บนพื้นดิน ใต้ดิน และในอากาศ มีการค้นพบแล้วมากกว่าล้านชนิด สัตว์บางชนิดมีความแตกต่างแบบชัดเจน แต่บางชนิดมีความคล้ายคลึงกัน จนอาจคิดได้ว่าเป็นชนิดเดียวกัน บางชนิดมีการแพร่พันธุ์ แบบกระจายครอบคลุมพื้นที่หลายประเทศ หรือข้ามทวีปก็เป็นได้

นักสัตวศาสตร์จำเป็นต้องจำแนกสัตว์ต่างๆ ออกจากกันแบบละเอียดที่สุด เพื่อสำหรับการศึกษาสัตว์ต่างๆ แบบสม่ำเสมอ การศึกษาสัตว์ทางวิทยาศาสตร์ จำเป็นต้องใช้ความละเอียดเป็นอย่างมาก มีการกำหนดหลักเกณฑ์ในการจัดหมวดหมู่ และการตั้งชื่อให้มีความรัดกุม เพื่อให้ชื่อสัตว์เหล่านั้นเป็นสากล และเป็นที่ยอมรับให้มากที่สุด และมีการหาวิธีจำแนกสัตว์ต่างๆ ออกเป็นหมวดหมู่ โดยจำแนกจากลักษณะหรือวิธีการที่จะใช้จดจำสัตว์แต่ละตัว ได้ง่าย ส่วนสัตว์ในอดีต ได้ยึดหลักทางกายวิภาคของตัวแก่เต็มวัยเป็นสำคัญ พร้อมทั้งพิจารณาถึงประวัติความเป็นมา

ดังนั้นงานวิจัยนี้ จึงมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการจำแนกข้อมูลประเภทสัตว์ โดยทำการศึกษากับ 2 เทคนิค ได้แก่ เทคนิคการจำแนกประเภทข้อมูล JRip และเทคนิคต้นไม้ตัดสินใจ (Decision Tree) เพื่อหาเทคนิคที่ดีที่สุดสำหรับการจำแนกข้อมูลประเภทของสัตว์

2. วัตถุประสงค์

เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการจำแนกประเภทข้อมูลสัตว์ด้วยเทคนิค JRip และเทคนิคต้นไม้ตัดสินใจ (Decision Tree)

3. งานวิจัยที่เกี่ยวข้อง

ณัฐวดี หงษบุญมี และพงศนรินทร ศรรุง (2561 : บทคัดย่อ) ศึกษาเรื่อง การประยุกตใชเทคนิคจำแนกข้อมูลแบบต้นไม้ตัดสินใจเพื่อการวินิจฉัยโรคในโคเบื้องตนบนโทรศัพทมือถือ ผลการวิจัยพบว่า การรวบรวมขอมูลปจจัยที่เกี่ยวของกับการวินิจฉัยโรคโนโคจากกลุมเกษตกรกรผู้เลี้ยงโคและผู้เชี่ยวชาญในเขตจังหวัดพิษณุโลก สร้างโมเดลวินิจฉัยโรคใชเทคนิคตนไม้ตัดสินใจ เปรียบเทียบอัลกอริทึมจำนวน 3 อัลกอริทึม ได้แก J48 RandomTree และ REPTree แลวทำการทดสอบประสิทธิภาพโมเดลด้วยวิธีการตรวจสอบแบบไขว เพื่อที่จะหาโมเดลการวินิจฉัยโรคที่มีประสิทธิภาพดีที่สุด จากผลการวิจัยพบว่าตนไม้ตัดสินใจแบบอัลกอริทึม RandomTree มีประสิทธิภาพดีที่สุด ค่าความถูกตองเทากับ 99.47%

ฐิติมา ช่วงชัย (2559 : บทคัดย่อ) ศึกษาเรื่อง การวิเคราะห์หารูปแบบการเรียนรู้โดยใช้เหมืองข้อมูลของนักศึกษาต่อการจัดทำปริญญานิพนธ์ ผลการวิจัยพบว่า การจัดการเรียนรู้ให้เหมาะกับรูปแบบการเรียนรู้ของผู้เรียนโดยศึกษากับนักศึกษาระดับปริญญาตรี สาขาการจัดการเทคโนโลยีการผลิตและสารสนเทศ การทำเหมืองข้อมูลโดยใช้โปรแกรม WEKA เป็นเครื่องมือในการวิเคราะห์ เลือกวิธีวิเคราะห์ผลด้วยรูปแบบ Rule Based Classification ด้วยวิธี Decision Table, JRip และ PART และรูปแบบ Decision Tree Classification ด้วยวิธี LMT, J48 และ Random Tree จากการวิเคราะห์ผลพบว่ารูปแบบของ Rule Based classification ด้วยวิธีการ PART ให้ค่าความถูกต้องสูงสุด (84.12%) วิธีการ JRip ให้ค่าความถูกต้อง (53.82%) และวิธีการ Decision Table ให้ค่าความถูกต้อง (21.18%) ส่วนรูปแบบของ Decision Tree ด้วยวิธีการ Random Tree ให้ค่าความถูกต้องสูงสุด (100%) วิธีการ LMT ให้ค่าความถูกต้อง (90.58%) และวิธีการ J48 ให้ค่าความถูกต้อง (78.24%)

พรพิรุณ โอ่งอินทร์ และคณะ (2558 : บทคัดย่อ) ศึกษาเรื่อง การวิเคราะห์ปัจจัยที่มีความสัมพันธ์ต่อการเกิดอาชญากรรมด้วยวิธีการทำเหมืองข้อมูล ผลการวิจัยพบว่า การวิเคราะห์ปัจจัยที่มีความสัมพันธ์ต่อการเกิดอาชญากรรมในพื้นที่เขตเมืองและนอกเขตเมืองด้วยการทำเหมืองข้อมูลโดยนำขั้นตอนวิธีริปเปอร์ (Ripper Algorithm) ได้มีการวิเคราะห์ข้อมูลจากการหาสถิติทั่วไป และใช้การทำเหมืองข้อมูลเพื่อวิเคราะห์หาความสัมพันธ์ระหว่างประเภทคดีเวลา สถานที่เกิดเหตุ ลักษณะผู้กระทำความผิด โดยนำหลักการขั้นตอนวิธี Ripper มาใช้วิเคราะห์ข้อมูล ซึ่งในโปรแกรม Weka นั้นเป็นการเลือกใช้วิธี JRip จากนั้นนำกฎที่ได้จากการศึกษามาสอบถามผู้เชี่ยวชาญเพื่อค้นหากฎที่น่าสนใจและสามารถนำไปใช้วางแผนแก้ปัญหาอาชญากรรมได้จริง

อรนุช พันโท และมนต์ชัย เทียนทอง (2557 : บทคัดย่อ) ศึกษาเรื่อง การเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการเรียนรู้ VARK ด้วยเทคนิคเหมืองข้อมูล ผลการวิจัยพบว่า การเปรียบเทียบประสิทธิภาพของการจำแนกรูปแบบการเรียนรู้ VARK ด้วยเทคนิคเหมืองข้อมูล 3 วิธีคือ วิธีแบบเบย์ วิธีต้นไม้ตัดสินใจ และวิธีฐานกฎ โดยการศึกษาครั้งนี้ได้รวบรวมข้อมูลด้วยแบบสอบถามจากผู้เรียนในระดับปริญญาตรี มหาวิทยาลัยราชภัฏเชียงใหม่ จำนวน 900 คน โดยใช้โปรแกรม WEKA พบว่า ประสิทธิภาพการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ มีประสิทธิภาพสูงสุด

ปราลี มณีรัตน์ (2554 : บทคัดย่อ) ศึกษาเรื่อง การสร้างโมเดลการจัดการระบบนักศึกษาสัมพันธ์โดยใช้เทคนิคเหมืองข้อมูล ผลการวิจัยพบว่า การศึกษาและวิเคราะห์หาปัจจัยที่ส่งผลต่อความพึงพอใจของนักศึกษา ในการสร้างระบบบริหารความสัมพันธ์นักศึกษา (Student Relationship Management) โดยใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) ซึ่งใช้การแบ่งกลุ่มข้อมูล (Clustering) โดยอัลกอริทึม K Mean และเทคนิค JRip การหาความสัมพันธ์ของระหว่างปัจจัยร่วมกัน ผลลัพธ์ที่ได้จากการใช้การแบ่งกลุ่มข้อมูล (Clustering) โดยใช้เทคนิค K mean แบ่งกลุ่มนักศึกษาตามชั้นปีที่ 1,2,3 และ 4 ออกเป็นกลุ่มต่างๆ เป็นวิธีที่ช่วยให้สามารถวิเคราะห์คุณลักษณะและความพึงพอใจในด้านต่างๆ ของนักศึกษาแต่ละกลุ่ม ได้อย่างถูกต้องและชัดเจน และใช้เทคนิค JRip ช่วยในการค้นหากฎความสัมพันธ์ในแต่ละกลุ่มที่สนใจศึกษา เพื่อวิเคราะห์ในรายละเอียดว่ากลุ่มนั้นมีความสัมพันธ์ระหว่างปัจจัยใดร่วมกัน

4. ทฤษฎีที่เกี่ยวข้อง

การทำเหมืองข้อมูล (Data Mining) คือ การแสวงหาความรู้ (knowledge) จากฐานข้อมูลหรือคลังข้อมูล ด้วยการวิเคราะห์ข้อมูล ซึ่งรูปแบบของการทำเหมืองข้อมูลได้รวมความรู้จากหลากหลายแขนงเข้าไว้ด้วยกัน ประกอบด้วยระบบฐานข้อมูล (Database) ระบบการเรียนรู้ของเครื่องจักร (Machine Learning) วิทยาศาสตร์สารสนเทศ (Information Science) และสถิติ(Statistics)

เทคนิคการจำแนกกลุ่มข้อมูล (Classification)

เป็นเทคนิคที่สามารถจำแนกกลุ่มของข้อมูลด้วยคุณลักษณะต่างๆ ที่ได้มีการกำหนดไว้ก่อนแล้ว เทคนิคประเภทนี้มีความเหมาะสมในการใช้พยากรณ์ข้อมูล ซึ่งการเรียนรู้ในลักษณะเรียกว่าการเรียนรู้แบบมีผู้สอน (Supervised Learning)

เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

เป็นเทคนิคที่ใช้ในการแยกหมวดหมู่ข้อมูลหรือแบ่งประเภท ซึ่งเป็นเทคนิคหนึ่งของการจำแนกกลุ่มข้อมูล (Classification)

เทคนิคการจำแนกประเภทข้อมูล (JRip)

เทคนิคที่พัฒนามาจากเทคนิค IREP ซึ่งเป็นเทคนิคที่ใช้ในการจำแนกกลุ่มข้อมูล (Classification) ประกอบด้วย 3 ขั้นตอนหลัก คือ 1) ขั้นตอนการสร้างกฎเพื่อใช้ในการจำแนกประเภทข้อมูล 2) ขั้นตอนการปรับแต่งกฎ 3) ขั้นตอนการลบกฎ

5.วิธีดำเนินการวิจัย

1 ขั้นตอนการดำเนินการวิจัยประกอบด้วย 3 ขั้นตอน ดังนี้

1.1 ศึกษาและวิเคราะห์ปัญหาที่เกี่ยวข้องกับงานวิจัย

1.2 วิเคราะห์ข้อมูลที่ได้จากการศึกษาในขั้นที่ 1 โดยวิเคราะห์จากข้อมูลที่เลือกมาด้วยโปรแกรม Weka 3.8.4

1.3 วิเคราะห์ผลการจำแนกประเภทขอมูล เปนขั้นตอนการทดสอบการจำแนกประเภทขอมูล หลังจากทำการทดสอบการจำแนกขอมูลประเภทของสัตว์ที่มีความเหมาะสมมากที่สุด โดยทำการเปรียบเทียบคาเฉลี่ยความถูกตองการจำแนกประเภทขอมูลของแตละกลุมการทดสอบ ดวยวิธี JRip และต้นไม้ตัดสินใจ

2 เครื่องมือการวิจัย

2.1 Weka 3.8.4

3. กลุ่มเป้าหมาย/ ประชากรและกลุ่มตัวอย่าง

3.1 กลุ่มตัวอย่าง คือ ข้อมูลประเภทของสัตว์จำนวน 100 ตัวอย่าง 15 คุณลักษณะ ซึ่งมีจำนวน 7 ประเภท

4. การทดลองใช้โปรแกรม Weka version 3.8.4

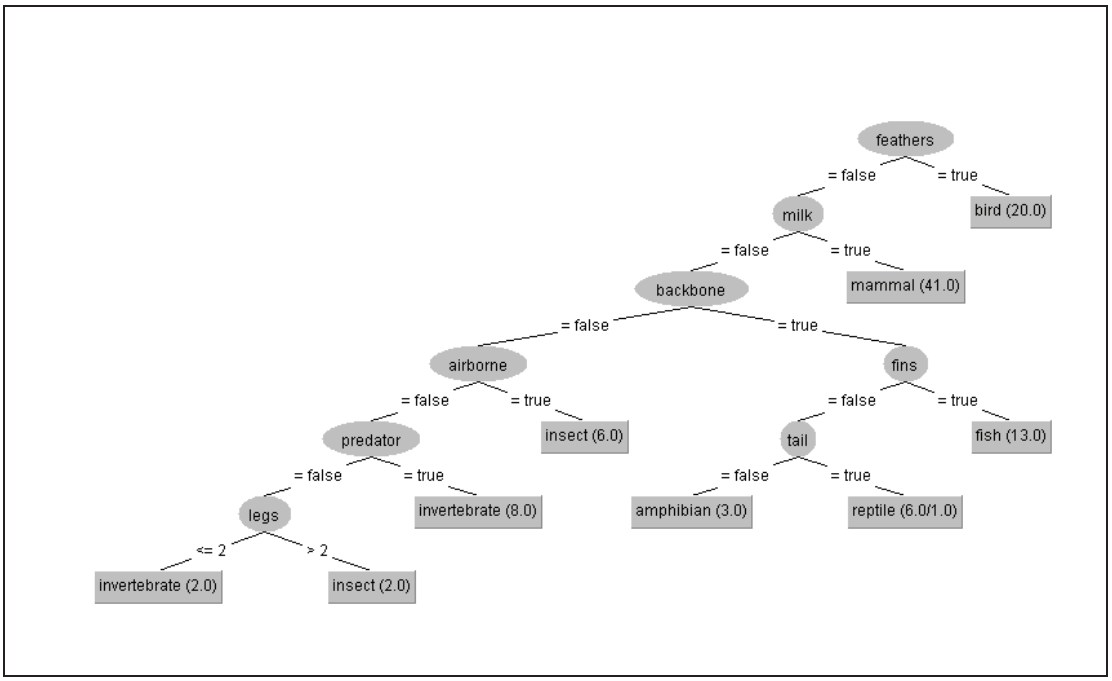
การทดลองใช้ โปรแกรม Weka version 3.8.4 ที่พัฒนาขึ้นเพื่อจำแนกข้อมูลประเภทของสัตว์ด้วยเทคนิคเหมืองข้อมูล จำนวน 101 ตัวอย่าง โดยใช้ 2 เทคนิค คือ JRip และ ต้นไม้ตัดสินใจ (J48) จะแสดงในภาพที่ 1 และ ภาพที่ 2

```
JRIP rules:
===========

(animal = frog) => type=amphibian (2.0/0.0)
(legs >= 6) and (predator = false) => type=insect (7.0/0.0)
(backbone = false) => type=invertebrate (11.0/1.0)
(breathes = false) => type=fish (14.0/1.0)
(feathers = true) => type=bird (20.0/0.0)
 => type=mammal (47.0/6.0)

Number of Rules : 6
```

ภาพที่ 1 ภาพแสดงการจำแนกข้อมูลด้วยเทคนิค JRip



ภาพที่ 2 ภาพแสดงการจำแนกข้อมูลด้วยเทคนิค J48

5. ผลการทดลอง

ผลการเปรียบเทียบประสิทธิภาพเทคนิคการจำแนกข้อมูลประเภทของสัตว์ด้วย 2 เทคนิค ได้แก่ JRIP และต้นไม้ตัดสินใจ จะนำเสนอในรูปแบบตาราง ดังแสดงในตารางที่ 1

ตารางที่ 1 ผลการเปรียบเทียบประสิทธิภาพเทคนิคการจำแนกข้อมูลประเภทของสัตว์ด้วยเทคนิค JRIP และต้นไม้ตัดสินใจ

Techniques	Accuracy
Decision Tree (J48)	97.08%**
JRIP	87.13 %

จากตารางสามารถสรุปได้ว่าเทคนิคที่ให้ประสิทธิภาพในการจำแนกข้อมูลประเภทของสัตว์ได้ดีที่สุดคือเทคนิคต้นไม้ตัดสินใจหรือ J48 โดยให้ค่าความแม่นยำในการจำแนกข้อมูลที่ 97.08 % และรองลงมา คือ JRIP ซึ่งให้ค่าความแม่นยำในการจำแนกข้อมูลที่ 87.13 % ตามลำดับ

6.สรุปผลและข้อเสนอแนะ

6.1 สรุปผล

ผลการวิจัย ครั้งนี้โดยพัฒนาและเปรียบเทียบตัวแบบการจำแนกข้อมูลทั้ง 2 เทคนิค ได้แก่ JRIP และต้นไม้ตัดสินใจซึ่งผลการประเมินประสิทธิภาพตัวแบบที่มีค่าความแม่นยำมากที่สุด คือ ต้นไม้ตัดสินใจซึ่งได้ค่าที่สูงที่สุดโดยวัดจากค่าความถูกต้องในการจำแนกข้อมูล (Accuracy)

6.2 ข้อเสนอแนะ

การวิจัยครั้งนี้เป็นการวิจัยที่ใช้โปรแกรม Weka 3.8.4 ซึ่งเป็นการเริ่มต้นศึกษาและทำการวิจัยเท่านั้น ดังนั้นสำหรับการทดสอบประสิทธิภาพในการจำแนกข้อมูลของเทคนิคทางเหมืองข้อมูลนั้น ยังสามารถทดสอบด้วยวิธีการต่างๆ ได้อีกมากมาย หากต้องการทดสอบแม่นยำของประสิทธิภาพของการจำแนกข้อมูลชนิดนี้จึงควรทำการทดสอบในเทคนิคอื่นๆ นอกเหนือจากนี้ด้วย

7. เอกสารอ้างอิง

ณัฐวดี หงษบุญมี และพงศนรินทร ศรรุง. (2561). การประยุกตใชเทคนิคจำแนกข้อมูลแบบต้นไม้ตัดสินใจเพื่อการวินิจฉัยโรคในโคเบื้องตนบนโทรศัพทมือถือ. **วารสารวิทยาศาสตรและเทคโนโลยี,** 20(1), 44-58.

ฐิติมา ช่วงชัย. (2559). การวิเคราะห์หารูปแบบการเรียนรู้โดยใช้เหมืองข้อมูลของนักศึกษาต่อการจัดทำปริญญานิพนธ์. **วารสารบัณฑิตศึกษา,** 10 (2), 53-62.

อรนุช พันโท และมนต์ชัย เทียนทอง. (2558). การเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการเรียนรู้ VARK ด้วยเทคนิคเหมืองข้อมูล. **วรสารเทคโนโลยีอุตสาหกรรม,** 4(1), 1-11.

พรพิรุณ โอ่งอินทร์, วรรณา ศิริแสงตระกูล, และพุธษดี ศิริแสงตระกูล. (2557). **การวิเคราะห์ปัจจัยที่มีความสัมพันธ์ต่อการเกิดอาชญากรรมด้วยวิธีการทำเหมืองข้อมูล.** การประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาแห่งชาติ, 34, 172-180.

ปราลี มณีรัตน์. (2554). **การสร้างโมเดลการจัดการระบบนักศึกษาสัมพันธ์โดยใช้เทคนิคเหมืองข้อมูล**. รายงานการวิจัย. มหาวิทยาลัยศรีปทุม. กรุงเทพฯ.