การเลือกคุณลักษณะเพื่อเพิ่มประสิทธิภาพของการจำแนกประเภท: การวิจัยเชิงสำรวจ
Features Selection to Optimization of Classification: A Survey

พัชรนิกานต์ พงษ์ธนู (Patcharanikarn Pongthanoo)[1] และวรารัตน์ สงฆ์แป้น (Wararat Songpan)[2]

[1,2]สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

[1]patcharanikarn@kkumail.com, [2]wararat@kku.ac.th

บทคัดย่อ

บทความนี้ เป็นงานวิจัยเชิงสำรวจเกี่ยวกับการเลือกคุณลักษณะเพื่อเพิ่มประสิทธิภาพของการจำแนกประเภทเหมืองข้อมูล มีวัตถุประสงค์เพื่อศึกษางานวิจัยที่เกี่ยวข้องกับการเลือกคุณลักษะของการจำแนกประเภทในการทำเหมืองข้อมูล โดยเลือกศึกษางานวิจัยไม่เกิน 10 ปีย้อนหลัง จำนวน 10 ผลงาน แล้วสรุปการนำเสนอการเลือกคุณลักษณะของแต่ละงานวิจัย ผลจากการสำรวจงานวิจัยทั้งหมด 10 ผลงาน พบว่าการเลือกคุณลักษณะซึ่งเป็นส่วนสำคัญมากในการเตรียมข้อมูลก่อนทำการสร้างตัวแบบพยากรณ์เหมืองข้อมูล จะมีการเปรียบเทียบกับวิธีการที่เป็นที่รู้จัก หรือวิธีการก่อนหน้าที่งานวิจัยอื่นๆได้นำเสนอไปแล้ว ซึ่งเมื่อเปรียบเทียบกับวิธีการก่อนหน้า งานวิจัยทั้งหมดนี้ให้เกณฑ์การวัดประสิทธิภาพสูงขึ้นทั้งหมด

คำสำคัญ: การเลือกคุณลักษณะ การทำเหมืองข้อมูล การลดขนาดของมิติข้อมูล

Abstract

This paper is a survey research of features selection to optimize the classification of data mining. The purpose is to study the research related to the selection of the classification of data mining. Selected researches of not more than 10 years, 10 papers, then summarized the presentation of the selection of characteristics of each research. Based on a total of 10 research findings, it was found that the selection of features was a very important part of the data preparation before creating the data mining model. There will be a comparison with the way it is known. Or how previous research has been presented. When comparing previous methods All these studies provide all the high performance measures.

Keyword: Feature Selection, Data Mining, Dimensionality Reduction

1. บทนำ

การทำเหมืองข้อมูล [1] [2] มี 5 ขั้นตอน คือ การทำความเข้าใจปัญหา การทำความเข้าใจข้อมูล การเตรียมข้อมูล การสร้างแบบจำลอง การประเมินและการนำไปใช้ ซึ่งขั้นตอนที่ส่งผลต่อประสิทธิภาพของตัวแบบพยากรณ์เหมืองข้อมูลมากที่สุดคือ การเลือกคุณลักษณะ (Feature Selection) ซึ่งอยู่ในขั้นตอนการเตรียมข้อมูลเพื่อช่วยเพิ่มประสิทธิภาพของอัลกอริทึมการเรียนรู้ให้เพิ่มสูงขึ้น ด้วยการเอาคุณสมบัติที่ไม่เกี่ยวข้องและซ้ำซ้อนออกนอกจากนั้น การคัดเลือกคุณลักษณะยังช่วยให้การสร้างตัวแบบพยากรณ์ทำได้รวดเร็วยิ่งขึ้น มีงานวิจัยมากมายที่ปรับปรุงการเลือกคุณลักษณะ ลดขนาดของมิติ แก้ไขปัญหาของงานวิจัยก่อนหน้า แต่ด้วยปัจจุบันมีข้อมูลมากมาย การจะนำเทคนิควิธีวิจัยก่อนหน้ามาปรับใช้กับชุดข้อมูลใหม่ๆ อาจไม่สามารถทำได้ทันที ผู้วิจัยจึงต้องการศึกษางานวิจัยที่เกี่ยวข้องกับการคัดเลือกคุณลักษณะเพื่อเพิ่มประสิทธิภาพการทำเหมืองข้อมูล โดยสำรวจงานวิจัยไม่เกิน 10 ปีย้อนหลัง จำนวน 10 ผลงาน

2. ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) [1] [2] คือ ขบวนการทำงานสกัดข้อมูล (Extract Data) ออกจากฐานข้อมูลขนาดใหญ่ เพื่อให้ได้สารสนเทศ หรือความสัมพันธ์ของข้อมูลใหม่ที่เรายังไม่รู้ โดยเป็นสารสนเทศที่มีเหตุผลและสามารถนำไปใช้ได้ ซึ่งเป็นสิ่งสำคัญที่จะช่วยตัดสินใจในการทำธุรกิจต่างๆ เหมืองข้อมูลเป็นกระบวนการสำคัญในการทำ Knowledge Discovery in database (KDD) การทำเหมืองข้อมูลมี 5 ขั้นตอน คือ การทำความเข้าใจปัญหา การทำความเข้าใจข้อมูล การเตรียมข้อมูล การสร้างแบบจำลอง การประเมิน และการนำไปใช้และตรวจสอบผลว่าบรรลุเป้าหมายที่ตั้งไว้เพียงใด

2.2 เทคนิคต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจ เป็นการเรียนรู้โดยการจำแนก (Classification) [3] ข้อมูลออกเป็นกลุ่ม (Class) ต่างๆ โดยใช้แอทริบิวท์ (Attribute) ของข้อมูลในการแยกแยะ ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่าแอตทริบิวท์ใดของข้อมูลที่เป็นตัวกำหนดการจำแนก และแอตทริบิวท์แต่ละตัวของข้อมูลมีความสำคัญมากน้อยต่างกันอย่างไร ผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจประกอบด้วย 1) โหนดภายใน (Internal Node) คือ แอทริบิวท์ (Attribute) ต่างๆ ของข้อมูล โหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้หรือหนดราก (Root Node) 2) กิ่ง (Branch Link) คือ ค่าของแอตทริบิวท์ในโหนดภายในที่แตกกิ่งนี้ออกมา 3) โหนดใบ (Leaf Node) คือ กลุ่มคลาส (Class) ต่างๆ ซึ่งเป็นผลลัพธ์ในการจำแนกข้อมูล

2.3 วิธีการเรียนรู้แบบเบย์ Naïve Bayes

วิธีการเรียนรู้แบบเบย์ [2] เป็นวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes Theorem) เข้ามาช่วยในการเรียนรู้ จุดมุ่งหมายเพื่อสร้างโมเดลที่อยู่ในรูปของความน่าจะเป็น ดังนั้นวิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดีและง่ายต่อการประมวลผล อัลกอริทึมของการเรียนรู้แบบเบย์อย่างง่ายนี้มีการทำงานที่ไม่ซับซ้อนเหมือนวิธีการอื่นๆ ซึ่งหาความน่าจะเป็นสำหรับการแยกประเภทข้อมูลที่มีความรู้ก่อนหน้า (Prior Knowledge) เขียนแทนด้วย P(ai|vj) โดยที่ ai คือแอตทริบิวท์ที่ i และ vi หมายถึงค่าของแอตทริบิวท์ aj ดังนั้น หากต้องการรู้ว่าข้อมูลชุดหนึ่งมีการแยกประเภทเป็นคลาสใด จึงมีการคำนวณหาค่าความน่าจะเป็นมากที่สุดที่มีความใกล้เคียงกับความน่าจะเป็นของทุกคลาส และผลคูณของค่าความน่าจะเป็นของคุณลักษณะของข้อมูลชุดที่มีความรู้ก่อนหน้า ดังสมการที่ 1

$$V_{NB} = argmax\, P(v_j) * \prod_{i=1}^{n} P(a_i|v_j) \quad (1)$$

2.4 อัลกอริทึม k-Nearest Neighbor

อัลกอริทึม k-Nearest Neighbor [3] เป็นขั้นตอนวิธีเพื่อนบ้านที่ใกล้ที่สุด โดยใช้วิธีการหาระยะห่างระหว่างคุณสมบัติของแต่ละข้อมูล ซึ่งเหมาะสำหรับข้อมูลที่เป็นตัวเลข โดยกำหนดจำนวนเพื่อนบ้าน k แล้วคำนวนระห่าง (Distance) ของข้อมูลที่ต้องการพิจารณากับชุดข้อมูลสอนจากสมการระยะทางยูคริเดียน (Euclidean Distance) คำตอบของข้อมูลที่ต้องการทำนายคือ กลุ่มที่มีจำนวนมากที่สุดในกลุ่มของชุดข้อมูล k ตัวแรก

2.5 การเลือกคุณลักษณะ (Feature Selection)

การเลือกคุณลักษณะเป็นขั้นตอนการสำคัญ [3] ในการเตรียมข้อมูลก่อนการทำเหมืองข้อมูล เป็นเทคนิคการลดขนาดของมิติ ลดจำนวนคุณลักษณะ และคัดเลือกคุณลักษณะที่มีความสำคัญต่อการพยากรณ์เหมืองข้อมูล และอาจกล่าวได้ว่าประสิทธิภาพของการจำแนกประเภทเหมืองข้อมูลขึ้นอยู่กับคุณลักษณะ (Feature) ที่นำมาใช้ ซึ่งนอกจากจะช่วยลดระยะเวลาในการสร้างตัวแบบพยากรณ์ให้เร็วขึ้นแล้ว ยังช่วยตัดคุณลักษณะที่ไม่จำเป็นต่อการสร้างตัวแบบพยากรณ์ได้อีกด้วย

2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยของ Ghaemi, M., & Feizi-Derakhshi, M. R. (2016). [4] ได้เสนอวิธีการเลือกคุณลักษณะเป็นปัญหาพื้นที่ว่างในการค้นหาแบบไม่ต่อเนื่อง โดยใช้อัลกอริทึมการเพิ่มประสิทธิภาพการป่า (Forest Optimization Algorithm - FSFOA) เปรียบเทียบกับวิธีการอื่น ๆ คือ HGAFS, PSO และ SVM-FuzCoc ผลการทดลองแสดงให้เห็นว่า FSFOA สามารถปรับปรุงความถูกต้องในการจำแนกประเภทของตัวจำแนกประเภทได้ 8 ชุดข้อมูลจากทั้งหมด 11 ชุดข้อมูล

Gao, W., Hu, L., Zhang, P., & Wang, F. (2018). [5] เสนอการเลือกคุณลักษณะแบบไฮบริดคือ Minimal Redundancy-Maximal New Classification Information (MR-MNCI) ที่รวม เกณฑ์การเลือกคุณลักษณะทั้งสองกลุ่มเข้าด้วยกัน โดยตัดคุณลักษณะที่ซ้ำซ้อนของของทั้งสองกลุ่ม และเลือกคุณลักษณะที่เป็นอิสระจากคลาสทั้งสองกลุ่มออก แล้วเปรียบเทียบวิธีการคัดเลือกคุณลักษณะกับอีก 7 วิธี (IWFS, MRI, JMIM, DRGS, MIFS, MIM and mRMR) ทดสอบกับข้อมูล 12 ชุด ผลการทดลอง วิธี MR-MNCI มีประสิทธิภาพในการจัดหมวดหมู่สูงที่สุด

Zhang, P., Gao, W., & Liu, G. (2018). [6] เสนอวิธี Feature Selection based on Weighted Relevancy (WRFS) ซึ่งจะวัดค่าสัมประสิทธิ์ที่ใช้ข้อมูลร่วมกัน เพื่อหาความสมดุลระหว่างสองคุณลักษณะ และเปรียบเทียบกับวิธีการคัดเลือกคุณลักษณะ 3 วิธีในชุดข้อมูลมาตรฐาน 18 ชุด ผลการทดลองวิธี WRFS มีประสิทธิภาพดีกว่าวิธีอื่นๆ โดยวัดค่าความถูกต้อง AUC และคะแนน F1

Kamyab, S., & Eftekhari, M. (2016). [7] เสนอเทคนิค Multimodal Optimization (MO) ในการแก้ปัญหาการเลือกคุณลักษณะเพื่อเพิ่มประสิทธิภาพ ในการค้นหาคุณลักษณะที่เหมาะสม ผลลัพธ์ที่ได้และการวิเคราะห์ทางสถิติพบว่าประสิทธิภาพของวิธี MO ให้ค่าถูกต้องดีกว่าเมื่อเทียบกับวิธีการอื่น

Jalilvand, A., & Salim, N. (2017). [8] เสนอวิธี Feature Unionization (FU) หรือการรวมคุณลักษณะ เพื่อลดขนาดการจัดหมวดหมู่คุณลักษณะ ทดสอบชุดข้อมูลจำนวน 9 ชุด ใช้การจำแนกความประเภท 3 แบบ ผลการทดลองวิธี FU มีประสิทธิภาพดีกว่าวิธีการคัดเลือกคุณลักษณะแบบอื่น โดยให้ความถูกต้องสูงสุดทั้ง 9 ชุดข้อมูลในวิธี SVM

Gu, N., Wang, D., Fan, M., & Meng, D. (2014). [9] เสนอแนวทางการจำแนกแบบกึ่งกำกับดูแล เรียกอัลกอริทึมนี้ว่า Kernel-based Sparse Regularized Least-Squared Classification (KSR-LSC) สำหรับการจัดหมวดหมู่แบบกึ่งกำกับดูแลให้มีประสิทธิภาพแตกต่างกันในแต่ละชุดข้อมูล โดยเปรียบเทียบวิธีการ 9 วิธี (KSR-LSC, S-RLSC, S-RLSC-un, SRC, SRC-un, 1-NN, GRF, LapRLSC, LNP) ผลการทดลองประสิทธิภาพของอัลกอริทึม KSR-LSC สูงสุดในทุกชุดทดสอบ

Akhtar, M. S., Gupta, D., Ekbal, A., & Bhattacharyya, P. (2017). [10] เสนอการเลือกคุณลักษณะการจำแนกประเภทโดยใช้ particle swarm optimization (PSO) เพื่อวิเคราะห์ความเชื่อมั่นด้านรูปแบบ การวิเคราะห์ความเชื่อมั่นเชิงคุณภาพจะดำเนินการสองขั้นตอน คือ การแยกและการจำแนกความเชื่อมั่นในชุดข้อมูลที่มีการตัดบางคุณลักษณะออกแล้ว ใช้ขั้นตอนวิธีการเรียนรู้ 3 ประเภท การเรียนรู้ คือ Maximum Entropy (ME), Randomly Field แบบมีเงื่อนไข (CRF) และ Support Vector Machine (SVM) การทดลองการแยกแยะคำกว้างยาวและการวิเคราะห์ความเชื่อมั่นในสองประเภทโดเมน ผลการทดลองงานวิจัยนี้มีประสิทธิภาพดีขึ้นกว่าวิธีการอื่น

Uysal, A. K. (2016). [11] เสนอการปรับปรุงการเลือกคุณลักษณะหรือ An improved global feature selection scheme (IGFSS) ซึ่งเป็นขั้นตอนสุดท้ายในการเลือกคุณลักษณะร่วมกัน เพื่อแก้ไขชุดข้อมูลคุณลักษณะที่เป็นตัวแทนมากขึ้น วิธีการคัดเลือกคุณลักษณะจะใช้ใน IGFSS เพื่อกำหนดลักษณะตามความสามารถในการจำแนก

แยกแยะในชั้นเรียน และป้ายกำกับ) label featuresจะ (ใช้ในขณะที่สร้างชุดคุณลักษณะ ผลการทดสอบ วิธี IGFSS ช่วยปรับปรุงประสิทธิภาพของการจัดหมวดหมู่ดีขึ้น โดยวัดจากค่าประสิทธิภาพ 2 ค่าคือ Micro-F1 และ Macro-F1

Pinheiro, R. H., Cavalcanti, G. D., Correa, R. F., & Ren, T. I. (2012). [12] เสนอวิธีการกรองสำหรับการเลือกคุณลักษณะหรือ A filtering method for feature selection called)ALOFT(วิธีการนี้มุ่งเน้นไปที่ลักษณะเฉพาะของโดเมนแบ่งประเภทของข้อความ เปรียบเทียบประสิทธิภาพของวิธีการ ALOFT กับวิธีการอื่น แล้วทดสอบโดยใช้ข้อมูล 3 ชุด (Reuters-21578, 20 Newsgroup และWebKB) ใช้การจำแนกประเภท 2 วิธีคือ k-Nearest Neighbor และ Naïve Bayes ผลทดลอง ALOFT ได้ผลลัพธ์ที่เท่ากันหรือดีกว่าวิธีการอื่น

3. วิธีดำเนินงาน

บทความนี้เป็นงานวิจัยเชิงสำรวจ โดยศึกษาเกี่ยวกับการเลือกคุณลักษณะที่ดีที่สุด และมีความสำคัญมากที่สุดในการสร้างตัวแบบพยากรณ์ โดยสำรวจและศึกษางานวิจัยจำนวน 10 ผลงาน

4. ผลการดำเนินงาน

ตารางที่ 1 เปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการเลือกคุณลักษณะในขั้นตอนของการเตรียมข้อมูล

Authors	Pre-processing	Classifiers	Output
Ghaemi, M., & Feizi-Derakhshi, M. R. (2016). [4]	- 8 Method comparisons. (SFSSBS-SFFS, NSM, SVM-FuzCoc, HGAFS, FS-NEIR, UFSACO, PSO, Feature Selection using Forest Optimization Algorithm (FSFOA))	Three classifiers, KNN (K=1, K=3, K=5) C4.5, SVM	FSFOA can improve the classification accuracy of classifiers is the best.
Gao, W., Hu, L., Zhang, P., & Wang, F. (2018). [5]	- MR-MNCI, our method is compared with four state-of-the-art feature selection methods (IWFS, MRI, JMIM and DRGS) and three competing methods (MIFS, MIM and mRMR).	Two classifiers, SVM and Naïve Bayes	The MR-MNCI method outperforms the seven other methods in terms of three metrics: average classification accuracy, highest average accuracy.
Zhang, P., Gao, W., & Liu, G. (2018). [6]	WRFS is compared to the six other representative methods , JMI, MIFS, mRMR, IWFS, MRI and JMIM	Two classifiers k-Nearest-Neighbors (3NN) and SVM	The experimental results show that WRFS outperforms (13/18 datasets) the other methods. WRFS can effectively identify the relevant features while eliminating the redundant and irrelevant features.

ตารางที่ 1 เปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการเลือกคุณลักษณะในขั้นตอนของการเตรียมข้อมูล (ต่อ)

Authors	Pre-processing	Classifiers /Models	Output
Kamyab, S., & Eftekhari, M. (2016). [7]	- compared Multimodal Optimization (MO) methods with Binary GA (BGA) and binary PSO (BPSO)	- 6 additional classifiers including SVM, PART, jRip, BFTree, Bayes Net and J48	MO methods, their results can significantly be improved.
Jalilvand, A., & Salim, N. (2017). [8]	- two filter feature selection algorithms (i.e., IG and CHI2) - Feature unionization (FU)	3 classification algorithms SVM, Naive Bayes and KNN(3NN)	The experimental results confirmed the efficiency of the Feature unionization (FU) approach.
Gu, N., Wang, D., Fan, M., & Meng, D. (2014). [9]	Kernel-based Sparse Regularized Least-Squared Classification (KSR-LSC) algorithm for semi-supervised classification is effective and compared with 9 methods (KSR-LSC, S-RLSC, S-RLSC-un, SRC, SRC-un, 1-NN, GRF, LapRLSC, LNP)	KSR-LSC method for semi-supervised classification	Classification accuracy of KSR-LSC algorithm is the highest on all test sets
Akhtar, M. S., Gupta, D., Ekbal, A., & Bhattacharyya, P. (2017). [10]	Aspect based sentiment analysis is performed in two steps, viz. aspect term extraction and sentiment classification.	Three classifiers, namely Maximum Entropy (ME), Conditional Random Field (CRF) and SVM	A PSO based ensemble technique can improve the classification accuracy.
Uysal, A. K. (2016). [11]	Compared an improved global feature selection scheme (IGFSS), IG, IG+IGFSS, GI, GI+IGFSS, DFS, DFS+IGFSS)	Two different classifiers Naïve bayes and SVM	DFS+IGFSS can improve the classification accuracy of classifiers is the best.
Pinheiro, R. H., Cavalcanti, G. D., Correa, R. F., & Ren, T. I. (2012). [12]	a filtering method for feature selection called ALOFT (At Least One FeaTure)	Two different classifiers k-Nearest Neighbor (KNN) and Naïve Bayes	For all the datasets, the best or near-best results of ALOFT are generated using CHI as FEF

5. สรุป

จากการสำรวจงานวิจัยที่เกี่ยวข้องกับการเลือกคุณลักษณะซึ่งเป็นกระบวนการสำคัญ ในการเตรียมข้อมูลก่อนทำการสร้างตัวแบบพยากรณ์เหมืองข้อมูล ทำให้ช่วยลดขนาดมิติของข้อมูล ส่งผลต่อการคัดเลือกคุณลักษณะที่มีความสำคัญต่อการพยากรณ์ ลดระยะเวลาในการพยากรณ์ และทำให้ได้ตัวแบบพยากรณ์มีประสิทธิภาพสูงยิ่งขึ้น งานวิจัยครั้งต่อไป ผู้วิจัยจะได้นำขั้นตอนการเลือกคุณลักษณะที่ให้ค่าประสิทธิภาพสูงในการสำรวจงานวิจัยครั้งนี้ มาวิเคราะห์กับข้อมูลจริง เพื่อวิเคราะห์ประสิทธิภาพและเปรียบเทียบกับขั้นตอนวิธีอื่นต่อไป

เอกสารอ้างอิง

[1] P.N. Tan, M. Steinbach, V. Kumar, “Introduction to Data Mining”, 1st ed. USA: Pearson, 2014.

[2] Daniel T. Larose, “Discovering Knowledge in Data”, 1st ed., John Wiley, 2005.

[3] Jiawei Han, Micheline Kameber and Jain Pei, *“Data Mining:Concepts and Techniques”*, Morgan Kaufmann publishers, San Francisco, CA, 2006.

[4] Ghaemi M, Feizi-Derakhshi MR. Feature selection using Forest Optimization Algorithm. Pattern Recognition. 2016;60:121-9.

[5] Gao W, Hu L, Zhang P, Wang F. Feature selection by integrating two groups of feature evaluation criteria. Expert Systems with Applications. 2018;110:11-9.

[6] Zhang P, Gao WF, Liu GX. Feature selection considering weighted relevancy. Applied Intelligence. 2018;48(12):4615-25.

[7] Kamyab S, Eftekhari M. Feature selection using multimodal optimization techniques. Neurocomputing. 2016;171:586-97.

[8] Jalilvand A, Salim N. Feature unionization: A novel approach for dimension reduction. Applied Soft Computing. 2017;52:1253-61.

[9] Gu NN, Wang D, Fan MY, Meng DY. A kernel-based sparsity preserving method for semi-supervised classification. Neurocomputing. 2014;139:345-56.

[10] Akhtar MS, Gupta D, Ekbal A, Bhattacharyya P. Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. Knowledge-Based Systems. 2017;125:116-35.

[11] Uysal AK. An improved global feature selection scheme for text classification. Expert Systems with Applications. 2016;43:82-92.

[12] Pinheiro RHW, Cavalcanti GDC, Correa RF, Ren TI. A global-ranking local feature selection method for text categorization. Expert Systems with Applications. 2012;39(17):12851-7.