หัวข้อการค้นคว้าแบบอิสระ	การจัดกลุ่มคุณลักษณะที่เหมาะสมเพื่อการสร้างต้นไม้ตัดสินใจที่มีประสิทธิภาพ
ผู้เขียน	นายประทิน กาวี
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
อาจารย์ที่ปรึกษา	อ.ดร.นริศรา เอี่ยมคณิตชาติ

บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเลือกวิธีการจัดกลุ่มคุณลักษณะที่เหมาะสมสำหรับข้อมูลแบบต่อเนื่อง ซึ่งจะทำให้ทราบวิธีการจัดกลุ่มข้อมูล จำนวนกลุ่มข้อมูล และวิธีการแบ่งคุณลักษณะที่เหมาะสม เพื่อเพิ่มความถูกต้องของต้นไม้ตัดสินใจ โดยใช้วิธีการแบ่งคุณลักษณะ 2 วิธี คือ การแบ่งคุณลักษณะแบบสองทางเลือก (Two way split) และการแบ่งคุณลักษณะแบบหลายทางเลือก (Multi way split) ซึ่งก่อนการแบ่งคุณลักษณะจะทำการจัดกลุ่มข้อมูล โดยใช้วิธีการจัดกลุ่มข้อมูล 3 วิธี ได้แก่ วิธีการจัดกลุ่มแบบอัลกอริทึม Expectation Maximization (EM) วิธีการจัดกลุ่มแบบเคมีน (K-Means) และวิธีการจัดกลุ่มแบบลำดับขั้น (Hierarchical) กำหนดจำนวนกลุ่มของการจัดกลุ่มข้อมูลเท่ากับ 2, 3, 4 และ 5 กลุ่ม ในการศึกษาครั้งนี้ใช้ชุดข้อมูลมาตรฐานจำนวน 6 ชุดได้แก่ ชุดข้อมูล Iris ชุดข้อมูล Abalone ชุดข้อมูล Breast cancer Wisconsin ชุดข้อมูล Pima Indians diabetes ชุดข้อมูล Seeds และชุดข้อมูล Ecoli ใช้ต้นไม้ตัดสินใจที่สร้างจากอัลกอริทึม J48 สำหรับการใช้เป็นเกณฑ์ในการวัดความถูกต้องของการจำแนก เมื่อวิธีการแบ่งคุณลักษณะและวิธีการจัดกลุ่มข้อมูลที่การศึกษานี้นำเสนอ

ผลการศึกษาพบว่า การสร้างต้นไม้ตัดสินใจด้วยวิธีการจัดกลุ่มแบบเคมีน จัดกลุ่มข้อมูลเป็น 3 กลุ่มกับชุดข้อมูล Iris ให้ความถูกต้องมากที่สุดถึง 97.33 % และเมื่อนำวิธีดังกล่าวไปทดสอบกับชุดข้อมูลอื่นอีก 5 ชุด พบว่าการสร้างต้นไม้ตัดสินใจโดยการแบ่งคุณลักษณะแบบสองทางเลือกมีความถูกต้องมากกว่าการแบ่งคุณลักษณะแบบหลายทางเลือก ถึง 4 ชุดข้อมูล และเมื่อทำการเปรียบเทียบผลการจำแนกโดยใช้ต้นไม้ตัดสินใจ พบว่าต้นไม้ตัดสินใจที่ใช้วิธีที่นำเสนอในการศึกษานี้ ให้ความถูกต้องมากกว่าต้นไม้ตัดสินใจแบบใช้ J48 ธรรมดาถึง 5 ชุดจากการทดสอบกับ 6 ชุดข้อมูล สรุปได้ว่าการจัดกลุ่มข้อมูลโดยใช้วิธีการจัดกลุ่มแบบเคมีน จัดกลุ่มข้อมูลเป็น 3 กลุ่ม สร้างต้นไม้ตัดสินใจโดยใช้การแบ่งคุณลักษณะแบบสองทางเลือก สามารถเพิ่มความถูกต้องให้กับต้นไม้ตัดสินใจที่ใช้อัลกอริทึม J48 ได้

Independent Study Title	An Appropriate Features Clustering to Create Efficient Decision Trees
Author	Mr.Prathin Kawee
Degree	Master of Engineering (Computer Engineering)
Advisor	Dr.Narissara Eiamkanitchat

ABSTRACT

The purpose of this study is to select a clustering method of features that is suitable for continuous data. The appropriateness of the clustering method, the number of clusters and the splitting method can increase the accuracy of decision tree. Two splitting methodologies used in this study are the two way split and the multi way split. Before the split process, 3 clustering algorithms are applied to each attribute. The clustering method in this study includes Expectation Maximization (EM), K-Means and Hierarchical. The numbers of clusters in the experimental are 2, 3, 4 and 5 clusters. Six standard dataset are used in the experimental including Iris dataset, Abalone dataset, Breast cancer Wisconsin dataset, Pima Indians diabetes dataset, Seeds dataset, and Ecoli dataset. The decision tree based on J48 algorithm creation is used for measurement the classification accuracy from the proposed splitting and clustering method in this study.

The experimental results on Iris data set shows that decision tree using K-Means of 3 clusters, has the highest accuracy, that is 97.33%. The K-Means method tested with five other data sets finds that creating a decision tree by two ways split results in the higher accuracy over multi way split in 4 data set. Based on the comparing classification result between decision tree using the method in this study and the decision tree using ordinary J48, the proposed method results higher accuracy in the 5 data set out of 6 data set. In conclusion, clustering using K-Means with 3 clusters and creating a decision tree by two way split can improve the accuracy of decision trees that use J48 algorithm.