ในปัจจุบันมนุษย์มีความสามารถในการเก็บบันทึกข้อมูลไว้เป็นจำนวนมหาศาล แต่ปริมาณที่มากเกินไปนี้กลับเป็นอุปสรรคต่อการวิเคราะห์ ตีความ และนำความรู้ที่ได้มาใช้ประโยชน์ต่อการตัดสินใจ การทำเหมืองข้อมูลจึงเกิดขึ้นในฐานะของศาสตร์แห่งการวิเคราะห์ข้อมูลอัตโนมัติ งานวิเคราะห์ข้อมูลนี้เป็นได้หลายรูปแบบ เช่น การวิเคราะห์เพื่อจำแนกข้อมูลได้อัตโนมัติ เพื่อค้นหาความสัมพันธ์ภายในข้อมูล ไปจนถึงเพื่อการตรวจจับรูปแบบที่เบี่ยงเบนไปจากข้อมูลปกติ

งานวิจัยนี้เน้นที่การทำเหมืองข้อมูลประเภทการจำแนกข้อมูลอัตโนมัติ โดยเจาะจงที่กลุ่มข้อมูลการวินิจฉัยโรค โดยมุ่งหวังเพื่อเอื้อประโยชน์กับงานทางการแพทย์ จุดมุ่งหมายหลักของการวิจัย คือ ทดสอบอัลกอริทึมต่างๆในการทำเหมืองข้อมูล เพื่อค้นหาอัลกอริทึมที่เหมาะสมกับข้อมูลการวินิจฉัยโรค งานวิจัยนี้ยังได้ตรวจสอบเทคนิคการเรียนรู้หลายครั้ง เพื่อเพิ่มความแม่นตรงในการทำนายและจำแนกประเภทข้อมูล โดยเน้นการศึกษาที่สองเทคนิค คือ bagging และ boosting

ผลการทดสอบอัลกอริทึมพื้นฐาน 4 อัลกอริทึมกับข้อมูล 12 ชุด พบว่าอัลกอริทึมที่ใช้หลักการต้นไม้ตัดสินใจ ทำงานได้ดีกับข้อมูลประเภทข้อความและสัญลักษณ์ที่มีจำนวนคลาสเพียงสองคลาส เมื่อจำนวนแอททริบิวต์เพิ่มมากขึ้นอัลกอริทึมประเภทนี้จะมีประสิทธิภาพลดลงอย่างชัดเจน ในขณะที่อัลกอริทึมที่ใช้หลักการเบย์ส์ไม่ได้รับผลกระทบจากจำนวนแอททริบิวต์ หรือจากจำนวนคลาสแต่อย่างใด

เทคนิคการเรียนรู้หลายครั้งสามารถเพิ่มประสิทธิภาพการจำแนกข้อมูลได้ แต่มีข้อยกเว้นในกรณีที่ข้อมูลมีการกระจุกตัวในบางคลาสมากเกินไป ประสิทธิภาพการจำแนกจะไม่เพิ่มขึ้นไปจนถึงลดลงในบางครั้ง บทสรุปของงานวิจัยนี้ได้เสนอแนะโมเดลในการคัดเลือกอัลกอริทึมและเทคนิคที่เหมาะสมโดยจะต้องพิจารณาร่วมกับลักษณะของข้อมูล

We are flooded with a huge volume of data and information. The tremendous amount of data, collected and stored in large databases, has far exceeded the human ability to analyze and extract valuable information for the purpose of decision-making support. Data mining has thus emerged as a new technology that can intelligently transform the vast amount of data into useful information and knowledge. Data mining tasks can vary from classification, association, to deviation detection.

This research focuses on the classification data mining. We have investigated the performance of four basic classification algorithms on twelve data sets, all are taken from a specific domain of medical diagnosis. Our main objective is to discover the appropriate technique for classifier induction on the medical data sets. Mulitple learning techniques such as bagging and boosting have also been employed to study the improvement on inducing a more accurate and sensitive model.

On the single-learning approach, we have found that the decision-tree induction algorithm performs well on the binary-class nominal data sets. However, the performance of the tree-based classifiers significantly degraded on high-dimensional data sets. The naive Bayes algorithm, on the contrary, is not affected by neither the dimension nor the number of classes.

The multiple-learning approach can, in general, improve the accuracy of the classification model. Nevertheless, we have discovered that if the distribution of data in each class is highly non-uniform, the multiple-learning techniques cannot improve,or even lower, the classifier's accuracy. We conclude our experimentations with the proposed decision model to suggest users how to choose the algorithm most appropriate for their specific medical data set.