บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันในยุคของข้อมูลข่าวสาร องค์กรส่วนใหญ่มีข้อมูลที่ต้องจัดเก็บอยู่เป็นจำนวนมากมาย เช่นระบบร้านค้าปลีก จะเก็บข้อมูลพนักงานในองค์กร , ข้อมูลการซื้อขาย ข้อมูลสินค้า ข้อมูลลูกค้า เป็นต้น จะเห็นได้ว่า ยิ่งองค์กรหรือรูปแบบธุรกิจมีขนาดใหญ่เท่าไร ย่อมทำให้การเก็บสะสมข้อมูลสำหรับองค์กรต่างๆมีจำนวนมากขึ้น การเก็บข้อมูลจำนวนมากเหล่านี้ ลงในฐานข้อมูล เป็นวิธีที่นิยมใช้ในหลายองค์กร แต่ระบบการจัดการฐานข้อมูลทั่วไปไม่สามารถจัดการกับข้อมูลเหล่านี้ได้อย่างมีประสิทธิภาพ เนื่องจากใช้เวลานานในการดึงข้อมูลที่มีความสำคัญออกมาวิเคราะห์ ดังนั้นจึงได้เกิดเทคโนโลยีในการวิเคราะห์ข้อมูลที่มีความสำคัญออกมาจากแหล่งเก็บข้อมูลขนาดใหญ่ เรียกเทคโนโลยีนี้ว่า "การทำเหมืองข้อมูล" หรือ การขุดค้นข้อมูล (data mining)

การทำเหมืองข้อมูล (Data Mining) หรืออาจจะเรียกว่า การค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Databases – KDD) เป็นเทคนิคเพื่อค้นหารูปแบบ (pattern) ของจากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากสถิติ การเรียนรู้ของเครื่อง และการรู้จำแบบ หรือในอีกนิยามหนึ่ง การทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูลที่มีจำนวนมาก เพื่อค้นหารูปแบบ แนวทางและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ ซึ่งเป็นเทคโนโลยีใหม่ของการประยุกต์ใช้ข้อมูลที่เก็บอยู่ในฐานข้อมูล ให้เกิดประโยชน์สูงสุดแก่หน่วยงานที่เป็นเจ้าของข้อมูล การประยุกต์ใช้ข้อมูลที่กล่าวถึงนี้มีได้หลายแนวทาง แต่โดยทั่วไปมักจะเป็นการสรุปภาพรวมของข้อมูลในฐานข้อมูล การวิเคราะห์แนวโน้มการเปลี่ยนแปลงของข้อมูล การค้นหาความสัมพันธ์ที่ซ่อนอยู่ภายในกลุ่มของข้อมูล หรือ การค้นหาข้อมูลที่มีความสำคัญโดยที่ไม่ทราบมาก่อน แต่มีความหมายโดยนัย และคาดว่าจะมีประโยชน์จากข้อมูลในฐานข้อมูล ปัจจุบันมีการประยุกต์ใช้งาน Data Mining ในธุรกิจหลากหลายประเภท เช่น ด้านการขายปลีกและขายส่ง ด้านการเงินการธนาคาร ด้านการประดิษฐ์และการผลิต ด้านการประกันภัย ด้านความปลอดภัย ด้านการตลาด และด้านการแพทย์

หัวใจสำคัญของกระบวนการ Data mining คือส่วนของโปรแกรมที่ทำหน้าที่สังเคราะห์ความรู้ขึ้นมาจากข้อมูลจำนวนมากในฐานข้อมูล ส่วนสังเคราะห์ความรู้นี้เรียกว่า Learning algorithm ซึ่งมีผู้เสนอแนวคิดและพัฒนาอัลกอริทึมส่วนนี้ขึ้นเป็นจำนวนมาก ได้แก่ อัลกอริทึมที่ใช้หลักการของการสร้างต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึมที่ใช้หลักการทางสถิติและทฤษฎีของเบยส์ (Naïve Bayes) อัลกอริทึมที่ใช้หลักการของโครงข่ายประสาทเทียม (Neural Network) และอัลกอริทึมอื่นๆ อีกมาก ปัจจุบันได้มีนักคอมพิวเตอร์ทดสอบเปรียบเทียบความสามารถของอัลกอริทึมแต่ละประเภท เพื่อค้นหาว่าอัลกอริทึมใดมีความสามารถสูงที่สุด ผลการทดสอบส่วนใหญ่ ปรากฏว่าไม่มีอัลกอริทึมใดที่ทำงานได้ดีที่สุดในข้อมูลทุกประเภท ทั้งนี้เนื่องจากข้อมูลแต่ละประเภทมีลักษณะเฉพาะตัวที่ต่างกัน เช่น ข้อมูลทางการแพทย์ จะต่างจากข้อมูลด้านกฎหมายและต่างจากข้อมูลด้านธุรกิจ ดังนั้นจึงไม่มีอัลกอริทึมใดที่ดีที่สุดสำหรับข้อมูลทุกประเภท

จากปัญหาดังกล่าวผู้วิจัยจึงมีแนวคิด ที่จะศึกษา ค้นหาลักษณะ และเปรียบเทียบประสิทธิภาพ อัลกอริทึมสังเคราะห์ความรู้ ที่เหมาะสมกับข้อมูลด้านการแพทย์ เพื่อศึกษาเปรียบเทียบอัลกอริทึมกลุ่มใดหรือประเภทใด ที่มีประสิทธิภาพการสังเคราะห์ หรือ การเรียนรู้ที่ดีที่สุดสำหรับการจำแนกประเภทของโรคทางการแพทย์ และศึกษาวิธีการปรับปรุงประสิทธิภาพอัลกอริทึมนั้นด้วยเทคนิคต่างๆ เพื่อจะเพิ่มขีดความสามารถการวิเคราะห์โรคให้แม่นยำมากขึ้น รวมถึงศึกษาเปรียบเทียบวิธีการลดคุณลักษณะที่เหมาะสมกับข้อมูลทางการแพทย์

1.2 วัตถุประสงค์

1.2.1 เพื่อศึกษาและวิเคราะห์อัลกอริทึมที่มีประสิทธิภาพในการจำแนกประเภทของโรคที่เหมาะสมกับโมเดลการวิเคราะห์โรคอัตโนมัติ

1.2.2 เพื่อศึกษาวิธีการเรียนรู้แบบ Single learning กับ Multiple learning ที่ส่งผลต่อประสิทธิภาพในการจำแนกประเภทของโรค ที่เหมาะสมกับโมเดลการวิเคราะห์โรคอัตโนมัติ

1.2.3 เพื่อศึกษาวิธีการลดคุณลักษณะ ที่ส่งผลต่อประสิทธิภาพในการจำแนกประเภทของโรค ที่เหมาะสมกับโมเดลการวิเคราะห์โรคอัตโนมัติ

1.3 ขอบเขตของการวิจัย

งานวิจัยนี้มุ่งเน้นการค้นหาเทคนิคด้านเหมืองข้อมูล เพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ เพื่อค้นหาอัลกอริทึมที่เหมาะสมที่สุดสำหรับฐานข้อมูลทางการแพทย์ โดยใช้อัลกอริทึมพื้นฐาน 7 อัลกอริทึม ซึ่งประกอบด้วย Naïve Bayes, Multilayer Perceptron, Radial Basis

Function Network, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Ripper รวมถึงการศึกษาเปรียบเทียบการลดคุณลักษณะที่เหมาะสมด้วย วิธี Correlation-based Feature Subset Selection (CFS) และวิธี Feature selection method based on correlation measure and relevance & redundancy analysis (FCBF) รวมถึงทดสอบกับอัลกอริทึมประเภท Single learning กับ Multiple learning โดยเพิ่มประสิทธิภาพด้วยวิธี Bagging และ Boosting โดยทดสอบกับข้อมูลทางการแพทย์ทั้ง 13 ชุด ซึ่งประกอบด้วย breast-cancer,breast-w,diabetes,heart-c,heart-statlog,hepatitis,hypothyroid,leukemia,liver-disorders,lung-cancer,lymphography,postoperative-patient,primary-tumor เท่านั้น

1.4 เครื่องมือที่ใช้ในงานวิจัย

1.4.1 เครื่องคอมพิวเตอร์ PC PentiumCore2 Duo 2.4 GHz หน่วยความจำหลัก 4 GB

1.4.2 ซอฟต์แวร์ Java JDK 1.6 และ Weka-3-6-2

1.4.3 กลุ่มข้อมูลทางการแพทย์ของมหาวิทยาลัยแห่งรัฐแคลิฟอร์เนียเมืองเออร์ไวน์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัตินั้น ผลที่ได้จากงานวิจัยนี้ สามารถใช้เป็นแนวทางในการพัฒนาซอฟต์แวร์เฉพาะทาง ด้านการแพทย์ที่เกี่ยวข้องกับการวินิจฉัยและตรวจรักษาโรค เพื่อทดแทนบุคลากรทางการแพทย์ที่ขาดแคลนได้เป็นอย่างดีต่อไปในอนาคต