นฤพนต์ ว่องประชานุกูล : วิธีที่เหมาะสมสำหรับการตัดกิ่งต้นไม้ตัดสินใจของการทำเหมืองข้อมูลทางด้านวิทยาศาสตร์ (A PROPER METHOD FOR DECISION TREE PRUNING IN SCIENTIFIC DATA MINING) อาจารย์ที่ปรึกษา : รศ. ดร.นิตยา เกิดประสพ, 118 หน้า. ISBN 974-533-558-4

ต้นไม้ตัดสินใจเป็นเครื่องมือหนึ่งที่นิยมนำมาใช้ในการทำเหมืองข้อมูล ที่เกี่ยวข้องกับงานการจำแนกข้อมูล โดยโมเดลสังเคราะห์ขึ้นจากกลุ่มตัวอย่างที่เรียกว่าชุดข้อมูลฝึก ในแต่ละเรคคอร์ดประกอบด้วยแอททริบิวต์จำนวนหลายแอททริบิวต์ โดยที่มีแอททริบิวต์หนึ่งแสดงกลุ่มของตัวอย่างนั้น แต่การนำเทคนิคการสร้างต้นไม้ตัดสินใจไปใช้กับข้อมูลจริง โมเดลที่สังเคราะห์ขึ้นอาจมีความซับซ้อนมากเกินไป เนื่องจากพยายามที่จะขยายโครงสร้างให้สามารถอธิบายข้อมูลรบกวนที่อาจมีอยู่ในชุดข้อมูลฝึกให้ได้ ปัญหานี้คือการเจาะจงโมเดลกับข้อมูลมากเกินไป เทคนิคที่สำคัญสำหรับแก้ไขการเจาะจงโมเดลกับข้อมูลมากเกินไปคือใช้เทคนิคการตัดกิ่งต้นไม้ตัดสินใจ เพื่อตัดกิ่งที่มีความน่าเชื่อถือน้อยออกไปจากต้นไม้ตัดสินใจ ซึ่งจะส่งผลให้ได้โครงสร้างต้นไม้ที่ใช้เวลาในการจำแนกข้อมูลได้เร็วขึ้น และปรับปรุงความสามารถของต้นไม้ให้ใช้จำแนกข้อมูลใหม่ได้อย่างถูกต้อง

งานวิจัยนี้ได้เสนอวิธีการตัดกิ่งต้นไม้ตัดสินใจ REP+ ที่พัฒนาขึ้นเพื่อเพิ่มประสิทธิภาพของต้นไม้ตัดสินใจ โดยใช้การทดสอบทางสถิติเพื่อตรวจสอบความสัมพันธ์กันอย่างมีนัยสำคัญระหว่างกลุ่มของข้อมูลที่ได้จากการทำนายของต้นไม้ตัดสินใจ และกลุ่มของข้อมูลจากชุดข้อมูลฝึก ดำเนินการทดสอบกับข้อมูลทางด้านวิทยาศาสตร์จำนวน 21 ชุดข้อมูล จากผลการวิจัยสามารถสรุปได้ว่า ต้นไม้ตัดสินใจที่สังเคราะห์ขึ้นจากงานวิจัยนี้จะมีความซับซ้อนลดลง สามารถใช้จำแนกข้อมูลได้รวดเร็วขึ้นโดยไม่ทำให้ความแม่นตรงในการจำแนกข้อมูลลดลง

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา _____

ปีการศึกษา 2548 ลายมือชื่ออาจารย์ที่ปรึกษา _____

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม _____

NARUPON WONGPRACHANUKUL : A PROPER METHOD FOR DECISION TREE PRUNING IN SCIENTIFIC DATA MINING. THESIS ADVISOR : ASSOC. PROF. NITTAYA KERDPRASOP, Ph.D., 118 PP.
ISBN 974-533-558-4

DECISION TREE/PRUNING/STATISTICAL TEST

Decision tree is one of the tools used for data mining. The main application area is classification task. The model is built from a set of records, called training set. Each record consists of a number of attribute-value pairs. One of these attributes represents class of the record. When a decision tree is built, many of the branches may be overly expanded due to noise or outliers in the training set. The built model is too complex, since it tries to classify all records in the training set including noise and outliers. This problem is called "overfitting". We use tree pruning method to remove the least reliable branches, generally resulting in faster classification and improvement in the ability of the tree to correctly classify unknown data.

This research proposed a new method for decision tree pruning, called REP+. We used the statistical test to check the significant dependency between predicted classes and actual classes in the training set. We conduct the experiments on 21 scientific data sets. The pruned trees result in reduced model complexity and faster classification while maintaining their predictive accuracy.

School of Computer Engineering

Academic Year 2005

Student's Signature Narupon

Advisor's Signature

Co-advisor's Signature