บทคัดย่อภาษาไทย

การสร้างต้นไม้ตัดสินใจเชิงอุปนัยเป็นงานที่สำคัญงานหนึ่งของการทำเหมืองข้อมูล และการเรียนรู้ของเครื่อง การทำเหมืองข้อมูลเป็นกระบวนการคัดแยกสารสนเทศที่เป็นประโยชน์และยังไม่เคยปรากฏมาก่อน เช่นรูปแบบร่วม หรือ ความสัมพันธ์ที่ซ่อนอยู่ในกลุ่มข้อมูล เทคนิคที่ใช้ในการค้นหารูปแบบร่วมที่น่าสนใจนี้มีอยู่หลากหลายเทคนิค ต้นไม้ตัดสินใจเป็นเครื่องมือชนิดหนึ่งที่นิยมใช้กันมากในงานทำเหมืองข้อมูล เทคนิคทุกชนิดที่ใช้ในงานทำเหมืองข้อมูลล้วนถูกขับเคลื่อนด้วยตัวข้อมูล แต่ข้อมูลที่ใช้ก็มักจะมีขนาดใหญ่และมีข้อผิดพลาดปะปนอยู่ ข้อผิดพลาดคือความบกพร่องที่ปรากฏแบบสุ่มที่ตำแหน่งใดก็ได้ ข้อมูลที่ผิดพลาดมีได้หลายรูปแบบ ได้แก่ การระบุค่าของคลาสผิด ค่าของแอททริบิวต์ผิด หรือข้อมูลที่มีรายละเอียดของทุกแอททริบิวต์เหมือนกันแต่มีค่าของคลาสต่างกันทำให้เกิดเป็นกรณีขัดแย้ง การมีข้อผิดพลาดทำให้ข้อมูลกลายเป็นข้อมูลรบกวน ข้อมูลรบกวนทุกประเภทล้วนมีผลต่อประสิทธิภาพการเรียนรู้ ผลกระทบที่ร้ายแรงที่สุดคือทำให้อัลกอริทึมการเรียนรู้สร้างผลลัพธ์ที่ซับซ้อนและผิดเพี้ยน ผลลัพธ์ที่มีขนาดใหญ่และซับซ้อนเกิดจากความพยายามที่จะสร้างโมเดลที่อธิบายทั้งข้อมูลที่ถูกและผิด สิ่งนี้ทำให้เกิดปัญหาที่เรียกว่า โมเดลที่จำเพาะเจาะจงมากเกินไป

อัลกอริทึมการเรียนรู้มักจะได้รับการออกแบบให้หลีกเลี่ยงปัญหาของการสร้างโมเดลที่จำเพาะเจาะจงมากเกินไป เทคนิคที่มักจะใช้จัดการไม่ให้เกิดการสร้างต้นไม้ตัดสินใจขนาดใหญ่ที่จะไปครอบคลุมข้อมูลรบกวนคือการตัดกิ่งของต้นไม้ ทั้งที่เป็นการตัดก่อนที่ต้นไม้จะมีขนาดใหญ่หรือตัดหลังจากพบว่าต้นไม้ใหญ่เกินไป เทคนิคทั้งสองประเภทมีข้อเหมือนกันคือผนวกการตัดกิ่งไว้ในขั้นตอนการสร้างต้นไม้ งานวิจัยนี้พิจารณาวิธีการจัดการกับข้อมูลรบกวนในแง่มุมที่ต่างออกไป โดยจะแยกขั้นตอนการจัดการกับข้อมูลรบกวนออกจากขั้นตอนการสร้างต้นไม้ ข้อมูลทั้งหมดที่มีทั้งข้อมูลที่ดีปะปนกับข้อมูลรบกวนจะถูกจัดกลุ่มและคัดเลือกด้วยฮิวริสติก ก่อนที่จะถูกส่งต่อไปยังขั้นตอนการสร้างต้นไม้ จากผลการทดลองพบว่าในข้อมูลปกติวิธีการใหม่ที่เสนอขึ้นนี้สามารถสร้างโมเดลที่มีความแม่นตรงสูงเทียบเท่ากับวิธี ID3 และเมื่อข้อมูลมีข้อรบกวนมากขึ้นเทคนิคที่พัฒนาขึ้นนี้ยังสามารถสร้างโมเดลที่มีความแม่นตรงสูงในขณะที่ ID3 จะให้โมเดลที่มีความแม่นตรงลดลง

บทคัดย่อภาษาอังกฤษ

Decision tree induction is a major task in data mining and machine learning. Data mining is the process of extracting useful and yet unknown information such as patterns or association hidden in stored data. Among various existing techniques applied to search for interesting patterns, decision tree is one of the most popular tools used for data mining. Most data mining techniques are data-driven, however, data repositories of interest in data mining applications can be very large and noisy. Noise is a random error in data. Noise in a data set can happen in different forms: misclassification or wrong labeled instances, erroneous or distorted attribute values, contradictory or duplicate instances having different labels. All kinds of noise can more or less affect the learning performance. The most serious effect of noise is that it can confuse the learning algorithms to produce complex and distorted results. The long and complex results are due to the attempt to fit every training data instance, including noisy ones, into the concept descriptions. This is a major cause of overfitting problem.

Most learning algorithms are designed with the awareness of overfitting problem due to noisy data. Prepruning and postprocessing are two major techniques applied to avoid growing a decision tree too deep down to cover the noisy training data. These techniques are tightly coupled to the tree induction phase. We, on the contrary, design a loosely coupled approach to deal with noisy data. Our noise-handling feature is in a separate phase from the tree induction. Both corrupted and uncorrupted data are clustered and heuristically selected prior to the application of tree induction engine. We observe from our experimental study that tree models produced from our approach are as accurate as the models generated by conventional ID3 approach. Moreover, upon highly corrupted data our approach shows a better performance than the ID3 approach.