

บทคัดย่อ

ปัญหาการกลุ่มข้อมูลสมดุลงานปรากฏในเซตข้อมูล ที่สัดส่วนของกลุ่มข้อมูลมีความแตกต่างกันสูง ดังนั้น การเรียนรู้ด้วยคอมพิวเตอร์มีอคติต่อกลุ่มข้อมูลที่มีสัดส่วนต่ำหรือกลุ่มน้อยมากกว่ากลุ่มข้อมูลที่มีสัดส่วนสูงหรือกลุ่มเด่น ส่งผลให้ตัวจำแนกที่ได้มีประสิทธิภาพไม่เป็นที่น่าพอใจในการทำนายกลุ่มน้อย สำหรับปัญหาประเภทนี้ กลุ่มน้อยที่มีจำนวนข้อมูลน้อยกว่าจัดเป็นข้อมูลที่มีความน่าสนใจและความสำคัญมากกว่ากลุ่มเด่นที่มีจำนวนข้อมูลมากกว่า ดังตัวอย่างต่อไปนี วิธีการใช้รับมือปัญหาการกลุ่มข้อมูลสมดุลงานที่ใช้กันอย่างแพร่หลายได้แก่ การสุ่มเพิ่ม และการสุ่มลด วิธีการแรกเป็นการสร้าง ข้อมูลน้อย เข้าไปในเซตข้อมูล เพื่อให้ขนาดของกลุ่มน้อยเพิ่มขึ้น วิธีที่สองเป็นการกำจัด ข้อมูลเด่นจากเซตข้อมูล เพื่อให้ขนาดของกลุ่มเด่นลดลง ทั้งสองวิธีจะทำการปรับขนาดของกลุ่มเด่นและกลุ่มน้อยให้มีความใกล้เคียงกัน เพื่อให้ตัวจำแนกสามารถทำนายทั้งกลุ่มเด่นและกลุ่มน้อยได้อย่างมีประสิทธิภาพ อย่างไรก็ตาม วิธีการแรกมีข้อเสียคือทำให้ขนาดของเซตข้อมูลใหญ่ขึ้น ส่งผลให้ใช้เวลาประมวลผลยาวนานขึ้น นอกจากนี้ อาจมีการสุ่มเพิ่มในบริเวณขอบทับ ของกลุ่มเด่นและกลุ่มน้อย ส่งผลให้ตัวจำแนกมีความสับสนในการทำนายในบริเวณดังกล่าว ส่วนวิธีที่สองมีข้อเสียคือ อาจกำจัดข้อมูลที่สำคัญ ออกไปจากเซตข้อมูล สำหรับงานวิจัยที่ได้ออกแบบไว้ จะทำการสุ่มเพิ่มและสุ่มลดอย่างระมัดระวัง โดยการเลือกบริเวณที่มีผลกระทบน้อยที่สุด ดังนี้ ในการสุ่มเพิ่มข้อมูลน้อย จะสุ่มเพิ่มบริเวณแกนกลางของกลุ่มน้อย ซึ่งเป็นข้อมูลที่มีความสำคัญมากกว่า ข้อมูลที่อยู่บริเวณขอบของกลุ่มน้อย นอกจากนี้ ยังเป็นการป้องกันปัญหาการเพิ่มข้อมูลในบริเวณที่ซ้อนทับกัน และในการสุ่มลด จะสุ่มลดข้อมูลเด่น จะสุ่มลดบริเวณเด่นบริเวณขอบเท่านั้น เปรียบเสมือนทำความสะอาดพื้นที่ในส่วนนี้ ให้ข้อมูลน้อยมีความสำคัญเพิ่มมากขึ้น เพื่อให้ตัวจำแนกเรียนรู้ได้ดี และที่สำคัญ ข้อมูลบริเวณขอบนี้ จัดเป็นข้อมูลที่มีความสำคัญน้อยเมื่อเปรียบเทียบกับข้อมูลบริเวณแกนกลาง

Abstract

Class imbalance learning has recently drawn considerable attention among researchers. In this area, a rare class is the class of primary interest from the aim of classification. Unfortunately, traditional machine learning algorithms fail to detect this class because a huge majority class overwhelms a tiny minority class. In this paper, we propose a new technique called CORE to handle the class imbalance problem. The objective of CORE is to strengthen the core of a minority class and weaken the risk of misclassified minority instances nearby the borderline of a majority class. These core and borderline regions are defined by the applicability of a safe level. As a result, a minority class is more crowded and dominant. The experiment shows that CORE can significantly improve the predictive performance of a minority class when its dataset is imbalance.