กิตติพงศ์ ชมบุญ : เทคนิคการจำแนกประเภทข้อมูลส่วนน้อยบนข้อมูลไม่สมดุลด้วยวิธีการแบ่งข้อมูล (CLASSIFICATION TECHNIQUE FOR MINORITY CLASS ON IMBALANCED DATASET WITH DATA PARTITIONING METHOD)
อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ, 98 หน้า

การจำแนกประเภทข้อมูลโดยใช้ข้อมูลที่ไม่สมดุลนั้นเป็นปัญหาสำคัญในการทำเหมืองข้อมูลที่น่าสนใจเนื่องจากการทำเหมืองข้อมูลด้วยข้อมูลที่ไม่สมดุลซึ่งมีข้อมูลคลาสส่วนใหญ่ และคลาสส่วนน้อยอยู่ปะปนกันนั้น ข้อมูลส่วนใหญ่จะมีคุณสมบัติบางประการที่บดบังคุณสมบัติของข้อมูลส่วนน้อย ทำให้การจำแนกประเภทข้อมูลส่วนน้อยนั้นไม่สามารถจำแนกได้อย่างมีประสิทธิภาพ ยกตัวอย่างเช่นข้อมูลไม่สมดุลของผลวินิจฉัยผู้ป่วยที่เป็นโรคมะเร็ง โดยมีข้อมูลส่วนใหญ่เป็นข้อมูลผู้ป่วยปกติ และมีข้อมูลส่วนน้อยที่เป็นโรคมะเร็ง ซึ่งเมื่อทำการจำแนกประเภทข้อมูลด้วยวิธีการจำแนกประเภทข้อมูลแบบปกติที่ให้ความสำคัญกับข้อมูลทุกคลาสเท่าเทียมกันนั้นจะทำให้ประสิทธิภาพในการจำแนกประเภทข้อมูลผู้ป่วยที่เป็นโรคมะเร็งซึ่งเป็นข้อมูลส่วนน้อยนั้นมีประสิทธิภาพไม่ดีเท่าที่ควร ดังนั้นในงานวิจัยนี้จึงได้นำเสนอเทคนิคการจำแนกประเภทข้อมูลที่มีขนาดของคลาสไม่สมดุลด้วยวิธีการแบ่งข้อมูล

ในงานวิจัยนี้เป็นการเสนอแนวคิดเพื่อแก้ปัญหาที่คุณสมบัติบางประการของข้อมูลส่วนใหญ่บดบังคุณสมบัติของข้อมูลส่วนน้อยจึงทำให้ประสิทธิภาพในการจำแนกข้อมูลส่วนน้อยนั้นไม่มีประสิทธิภาพดีเท่าที่ควร ด้วยวิธีการแบ่งข้อมูล โดยในงานวิจัยนี้จะแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ส่วนที่มีการซ้อนทับกัน และส่วนที่ไม่มีการซ้อนทับกัน โดยในแต่ละข้อมูลนั้นจะมีโมเดลในการจำแนกประเภทข้อมูล 2 โมเดล การทำนายคลาสของข้อมูลใหม่จะใช้ทั้งสองโมเดลประกอบกันเพื่อเพิ่มประสิทธิภาพในการจำแนกข้อมูลส่วนน้อยให้มีความถูกต้องสูงขึ้น ผลที่ได้จากการวิจัยคือการใช้วิธีการแบ่งข้อมูลด้วยการวัดระยะแบบ Euclidean และการใช้อัลกอริทึม SVM Linear kernel ให้ประสิทธิภาพในการจำแนกข้อมูลส่วนน้อยที่ดีที่สุด

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2558 ลายมือชื่ออาจารย์ที่ปรึกษา_____

KITTIPONG CHOMBOON : CLASSIFICATION TECHNIQUE FOR MINORITY CLASS ON IMBALANCED DATASET WITH DATA PARTITIONING METHOD. THESIS ADVISOR ASSOC. PROF. KITTISAK KERDPRASOP, Ph.D., 98 PP.

DATA PARTITIONING / IMBALANCED DATA/ DATA MINING

Classification using imbalanced dataset is a challenging problem in the data mining research area. The difficulty is due to the fact that the number of data instances in the minority class is much less than the number of instances in the majority class. The majority data can over-shadow the minority data and make the classification performance of the minority class unacceptable. For instance, the imbalance between non-cancer patients and patients with breast cancer. The minority of breast cancer records in the majority group of non-cancer patients can absent when classifying with traditional techniques. This thesis, therefore, proposes a partitioning technique to handle the imbalanced dataset problem.

This research solves the imbalanced dataset problem by partitioning data into two groups: overlap and non-overlap data. Each partition has its own classification model. To predict the future event, both classifiers are used in order to improve the minority class prediction. The experimental results show that partitioning technique based on Euclidean distance measure when applied to the SVM with linear kernel yields the best performance in classifying minority data.

School of Computer Engineering Student's Signature_____

Academic Year 2015 Advisor's Signature_____