หัวข้อดุษฎีนิพนธ์ การลดข้อมูลในกระบวนการฝึกแบบมีผู้สอนโดยใช้การจัดกลุ่มอินพุต-เอาต์พุต

ผู้เขียน นายอนุสรณ์ ยอดใจเพ็ชร

ปริญญา วิศวกรรมศาสตรดุษฎีบัณฑิต (วิศวกรรมไฟฟ้า)

คณะกรรมการที่ปรึกษา

รศ. ดร. นิพนธ์ ธีรอำพน	อาจารย์ที่ปรึกษาหลัก
รศ. ดร. เสริมศักดิ์ เอื้อตรงจิตต์	อาจารย์ที่ปรึกษาร่วม
ผศ. ดร. ศันสนีย์ เอื้อพันธ์วิริยะกุล	อาจารย์ที่ปรึกษาร่วม

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอวิธีการลดข้อมูลที่ใช้ในกระบวนการฝึกแบบมีผู้สอน การลดข้อมูลแบ่งออกได้เป็น 2 แบบ คือ การลดข้อมูลสำหรับฝึกเพื่อการทำนายและการลดข้อมูลสำหรับฝึกเพื่อการจำแนก ประโยชน์ของการลดข้อมูลที่ใช้ในการฝึก คือ พื้นที่สำหรับการจัดเก็บข้อมูลลดลง ใช้หน่วยความจำสำหรับการประมวลผลน้อยลง และลดเวลาในการประมวลผล อย่างไรก็ตาม การลดจำนวนข้อมูลลงก็มีผลเสียที่ตามมาคือในบางกระบวนการฝึกแบบมีผู้สอนต้องการข้อมูลที่ใช้สำหรับฝึกจำนวนมาก ข้อมูลทางอินพุตและเอาต์พุตต้องมีรายละเอียดมากที่สุดเท่าที่เป็นไปได้ หรือข้อมูลนั้นครอบคลุมทั้งปริภูมิของอินพุตและเอาต์พุต การลดข้อมูลที่ใช้ฝึกสอนลงก็ส่งผลให้ความถูกต้องของการทำนายหรือการจำแนกลดลงตามไปด้วยตามปริมาณของข้อมูลที่ลดลงได้

ผู้วิจัยได้ทราบถึงปัญหาของความถูกต้องที่ลดลง จึงได้ออกแบบกระบวนการลดข้อมูลให้เหมาะกับชนิดของการฝึกสอนคือ 1.การลดข้อมูลสำหรับฝึกเพื่อการทำนาย จะใช้หลักการแบ่งข้อมูลออกเป็นส่วนๆตามข้อมูลเอาต์พุตโดยที่ข้อมูลแต่ละส่วนยังคงมีคุณสมบัติหรือข้อมูลที่สำคัญเหมือนเช่นเดิม จากนั้นลดจำนวนอินพุตโดยจัดกลุ่มข้อมูลทางฝั่งอินพุตเลือกเอาข้อมูลที่เป็นตัวแทนของกลุ่มนั้นออกมาโดยจะต้องมีความสัมพันธ์กับข้อมูลทางเอาต์พุต จากนั้นนำข้อมูลที่ลดลงแล้วในแต่ละส่วนมารวมเข้าหากันก็สามารถลดจำนวนข้อมูลสำหรับฝึกสอนลงได้ 2.การลดข้อมูลสำหรับฝึกเพื่อการจำแนก จะใช้หลักการคงอยู่และเก็บข้อมูลที่มีความสำคัญต่อการจำแนกบริเวณขอบของแต่ละกลุ่มข้อมูล แล้วกำจัดส่วนที่มีผลต่อการจัดกลุ่มบริเวณตรงกลางของกลุ่มข้อมูลออกไป

การทดลองวัดประสิทธิภาพของวิธีการลดข้อมูลจะทดสอบด้วยชุดข้อมูลที่สังเคราะห์ขึ้นมาเพื่อใช้พิสูจน์สมมุติฐาน และชุดข้อมูลมาตรฐานสำหรับการเปรียบเทียบที่ได้รับมาจากฐานข้อมูลต่างๆ เพื่อเปรียบเทียบความสามารถในการลดข้อมูลกับวิธีการอื่นๆ ที่มีมาก่อนหน้างานวิจัยนี้ ค่าที่ใช้ในการเปรียบเทียบคือค่าเฉลี่ยความถูกต้อง ค่าเฉลี่ยอัตราส่วนการลดลงของข้อมูล และค่าร้อยละของการปรับปรุง

จากผลของการทดลองการลดข้อมูลสำหรับการทำนายสามารถสรุปได้ว่าปริมาณของข้อมูลที่ได้หลังจากผ่านกระบวนการลดข้อมูลนอกจากจะขึ้นอยู่กันตัวของข้อมูลที่ใช้เป็นอินพุตแล้วยังจะขึ้นอยู่กับตัวแปรที่กำหนดในช่วงเริ่มต้นคือจำนวนของการแบ่งข้อมูลออกเป็นส่วนย่อยที่เท่ากันและจำนวนของการแบ่งระดับข้อมูล โดยที่ความสามารถในการลดข้อมูลจะแปรผกผันกับปริมาณของตัวแปรทั้งสองนี้ แต่สิ่งดีที่ได้รับกลับมาคือค่าความถูกต้องที่มากขึ้นจนเกือบเท่ากับความถูกต้องของข้อมูลต้นฉบับ เมื่อพิจารณาผลการทดลองลดข้อมูลสำหรับการคัดแยกเฉพาะผลที่ได้จากพื้นฐานเดียวกันด้วยค่าเฉลี่ยอันดับจากความถูกต้อง ความสามารถในการลดข้อมูล และความสามารถที่จะนำไปใช้กับข้อมูลที่มีขนาดใหญ่นั้นพบว่าวิธีการที่นำเสนอเป็นวิธีการที่ดีที่สุด

Dissertation Title	Data Reduction in Supervised Training Process Using Input-Output Clustering	
Author	Mr. Anusorn Yodjaiphet	
Degree	Doctor of Philosophy (Electrical Engineering)	
Advisory Committee	Assoc. Prof. Dr. Nipon Theera-Umpon	Advisor
	Assoc. Prof. Dr. Sermsak Uatrongjit	Co-advisor
	Asst. Prof. Dr. Sansanee Auephanwiriyakul	Co-advisor

ABSTRACT

This thesis presents data reduction techniques in supervised learning. We divide reduction methods into two types, i.e., instance reduction for regression and instance reduction for classification. An advantage of data reduction for training data is that storage spaces, memory, and computation time required for training the full set of data are reduced. However, some supervised learning algorithms need a large number of training data. Input-output data must be abundant or at least sufficiently provided to cover the input and output data space. The reduction of the training data also possibly leads to degraded prediction accuracy for larger reduction ratios.

Due to this problem, we design two data reduction processes suitable for each type of training data sets. That is, for regression data, the instance reduction reduces the training data set by separating all the training data into small parts according to the output values. This is performed such that each part still preserves the same property and important information. After that, the input samples from each part are clustered and selectively chosen by using the relation between input and output data. Finally, all selected samples are combined together. For the instance reduction for classification data, the important data are retained whenever they are near the boundary of each class, or otherwise removed.

We prove the hypothesis by experimental results with synthesized data sets and also evaluate the performance of the proposed method on standard benchmark data sets acquired from various databases. We compare the experimental results with many state-of-the-art algorithms and assess the performance of the proposed algorithms by comparing the average accuracy, reduction ratios, and percent improvements.

From the experimental result of data reduction for regression problems, the reduction rate of a training data set after the reduction process depends on various properties of the data set and the setting parameters such as the number of split data and the number of quantization levels. The reduction rate goes down when these setting parameter values increase and it is in contrast to the regression accuracy which increases nearly close to the regression resulted from the original data. In addition, if we consider the classification results from all methods based on the same proposed method by the average accuracy, reduction rates, and the ability to apply to a large data set. The proposed method also gives the best result.