บทคัดย่อ 187471

การจัดกลุ่มข้อมูลโดยทั่วไป จะทำการแบ่งข้อมูลออกเป็นกลุ่มๆ โดยข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีความคล้ายคลึงกัน ซึ่งการจัดกลุ่มข้อมูลแบบถือว่าเป็น การจัดกลุ่มข้อมูลแบบไม่มีผู้สอน นั่นคือไม่มีการระบุชนิดของข้อมูลก่อนทำการจัดกลุ่ม แต่ในทางปฏิบัติการจำแนกประเภทของข้อมูล ข้อมูลที่เข้ามาอาจมีการระบุชนิดของข้อมูลก่อนทำการจัดกลุ่ม ในงานวิจัยนี้นำเสนอการจัดกลุ่มข้อมูลอีกวิธีหนึ่งที่ใช้ระยะห่างของข้อมูล และการระบุชนิดของข้อมูล ซึ่งเรียกการจัดกลุ่มข้อมูลแบบนี้ว่า การจัดกลุ่มข้อมูลแบบมีผู้สอน โดยข้อมูลทั้งหมดจะถูกแบ่งลงกริด ใน feature space เช่น ในกรณีที่ feature space เป็น 2 มิติ ลักษณะของกริดจะเป็นสี่เหลี่ยม ถ้าเป็น 3 มิติลักษณะของกริดจะเป็นลูกบาศก์ หรืออาจจะเป็น n มิติ ดังนั้นข้อมูลที่อยู่ในกริดเดียวกันจะมีความคล้ายคลึงกันมาก และข้อมูลที่อยู่ในกริดที่ติดกัน ถ้ามีการระบุชนิดของข้อมูลเป็นชนิดเดียวกัน จะสามารถจัดให้อยู่ในกลุ่มเดียวกันได้ การรวมกริดดังกล่าว จะทำจนกระทั่งไม่สามารถรวมกริดที่อยู่ติดกันและมีการระบุชนิดของข้อมูลเป็นชนิดเดียวกัน วิธีการนี้ใช้เวลาในการประมวลผลที่รวดเร็วและให้จำนวนกลุ่มที่เหมาะสม ซึ่งไม่ใช่แค่ข้อมูลที่อยู่ภายในกลุ่มเดียวกันจะอยู่ใกล้กันเท่านั้น แต่ยังรวมถึงชนิดของข้อมูลที่เหมือนกัน หรือเป็นการผสมของข้อมูลชนิดเดียวกัน (มีอัตราส่วนคลาสที่เหมือนกัน)

ABSTRACT 187471

In normal clustering, similar data are clustered together. Hence, data that located near each other in the feature space are clustered in to the same cluster. This clustering is unsupervised; that is, data are un-labeled. But in many applications such as data classification, train data are labeled and trained. This paper introduces another kind of clustering where both the distance in feature space and the data label are used in clustering process (supervised clustering). The data are separated into grids in feature space. For example, if the feature space has 2 dimensions, the grid shape is square. If the feature space has 3 dimensions, the grid shape is cube, the feature space will be n dimensions. The grids with a similar mixture of the class label are grouped in the same cluster. This process is repeated until no neighborhood grid that has similar grids can be merged. The clustering is fast and produces a good cluster where not only data in the same cluster are located near each other in the feature space but they must have a similar label or a similar mixture of multiple labels (same class ratio).