บทคัดย่อ

การสูญหายของค่าข้อมูลในขั้นตอนการเตรียมข้อมูลนั้นถือว่าเป็นปัญหาที่สำคัญอย่างยิ่งในกระบวนการสืบค้นสารสนเทศ หากไม่สามารถหาวิธีที่มีประสิทธิภาพมาทำการประมาณค่าที่สูญหายของข้อมูลได้อาจจะทำให้ผลที่ได้จากการวิเคราะห์เกิดความผิดพลาด ในงานวิจัยนี้จึงได้นำเสนอและวิเคราะห์ประสิทธิภาพของวิธีการประมาณค่าที่สูญหายของข้อมูลโดยอาศัยหลักการของ k - Nearest Neighbor ที่มีการปรับวิธีการให้น้ำหนักกับกลุ่มข้อมูลที่จะนำมาทำการประมาณค่าที่สูญหายของค่าข้อมูล โดยทำการศึกษาเพื่อเปรียบเทียบประสิทธิภาพของวิธีการให้น้ำหนักแบบเดิมที่เรียกว่า Distance-weight และวิธีการให้น้ำหนักที่ได้นำเสนอในงานวิจัยนี้หรือที่เรียกว่า Fuzzy-weight ในการวัดประสิทธิภาพของแต่ละวิธีจะใช้การคำนวณเพื่อหาค่าความแตกต่างระหว่างค่าของข้อมูลจริงกับค่าที่คำนวณได้จากอัลกอริทึม ซึ่งเรียกวิธีนี้ว่า Mean Square Error และเพื่อให้เข้าใจผลการทดลองมากยิ่งขึ้นจึงได้ใช้ค่า Performance Index เข้ามาช่วยในการนำเสนอผลการทดลองด้วย ข้อมูลที่ใช้ในการวิจัยประกอบด้วย ข้อมูลปริมาณน้ำฝน ข้อมูลอุณหภูมิที่สูงที่สุดในแต่ละวัน ข้อมูลปริมาณน้ำไหลเข้าอ่างเก็บน้ำ และข้อมูลไมโครอาร์เรย์ของยีสต์ Saccharomyces Cerevisiae จากการศึกษาพบว่าโดยเฉลี่ยแล้วการประมาณค่าที่สูญหายของข้อมูลด้วยวิธีการให้น้ำหนักกับกลุ่มข้อมูลแบบ Distance-weight เหมาะสำหรับชุดข้อมูลที่มีการสูญหายของค่าข้อมูลในระดับที่ไม่สูงมากนัก (10% - 20%) แต่ในระดับการสูญหายของค่าข้อมูลที่ 30 เปอร์เซ็นต์เป็นต้นไป พบว่าการประมาณค่าที่สูญหายของข้อมูลด้วยวิธีการให้น้ำหนักแบบ Fuzzy-weight มีประสิทธิภาพสูงกว่าวิธีการ Distance-weight สำหรับค่าฟัซซี่พารามิเตอร์ที่มากกว่า 3.0 และจำนวนเวกเตอร์ที่เหมาะสมในการนำไปใช้ประมาณค่าที่สูญหายของข้อมูลมีค่าตั้งแต่ 5 ถึง 20 $(5 \le k \le 20)$ สำหรับทุก ๆ ชุดข้อมูล

ABSTRACT

Missing value is an important topic in Knowledge Discovery from Databases (KDD) system. Some techniques used in KDD process give incorrect results when some values of information are missed. This thesis proposes and evaluates a "Fuzzy-weight" method to estimate missing values based on k - Nearest Neighbor algorithm. The performance of this method is evaluated by using Mean Square Error (MSE) and Performance Index (PI). We evaluate this method by experimenting on four datasets – RainFall, Maximum Temperature, Gaug high and Saccharomyces Cerevisiae Microarray. The number of neighborhoods (k) varies from 5 – 20 for the Fuzzy-weight method, the fuzzy parameter (f) is set to greater than 3.0. The experimental results are compared to those provided by an original method called "Distance-weight". The results indicate that the Distance-weight method is a suitable estimator for missing value at the missing level of 10-20%. However, our proposed method gives better estimation performance at the missing level of 30%-50%.