ABSTRACT

Diagnosis using data mining has a number of obstacles that may affects classification performance, including use of irrelevant features and presence of outliers in dataset. To avoid irrelevant features, it is necessary to conduct feature selection in fertility dataset. In previous research, features selection used Decision tree cannot increase accuracy. In other research that used Clustering-Based Decision Forest (CBDF) method can produced five best features, but does not explain its effect on enhancement of algorithms performance.

This study will analyze the effect of outliers elimination in fertility dataset on feature selection performance methods. The feature selection algorithm used in this study were Wrapper, Principal Component Analysis (PCA), and Gain Ratio, applied to classification method that are Multilayer Perceptron (MLP), Decision Tree, and Support Vector Machines (SVM). While the algorithm used to identification of outliers is K-Nearest Neighbor (KNN).

Outliers removal in dataset fertility can improve performance of classification majority significantly. It is also able to improve majority performance of feature selection. The best method of study produced by MLP - Gain ratio and MLP - Wrapper using dataset without outliers.

Keywords: Data mining, outlier, K-NN, Decision tree, MLP, SVM, feature selection, Gain Ratio, PCA, Wrapper.

INTISARI

Diagnosis menggunakan *data mining* memiliki sejumlah kendala yang dapat mempengaruhi performa klasifikasi, di antaranya adalah penggunaan fitur yang tidak relevan dan adanya *outlier* pada *dataset*. Cara menghindari fitur yang tidak relevan dapat dilakukan dengan metode seleksi fitur, sebagaimana telah dilakukan pada *dataset fertility* sebelumnya menggunakan *Decision tree* namun tidak memberikan peningkatan akurasi. Penelitian selanjutnya menggunakan metode *Clustering-Based Decision Forest* (CBDF) menghasilkan lima fitur terpilih, namun tidak menjelaskan pengaruhnya terhadap peningkatan performa klasifikasi.

Penelitian ini akan menganalisis pengaruh penghapusan *outlier* pada *dataset fertility* terhadap kinerja metode seleksi fitur. Algoritme seleksi fitur yang digunakan dalam penelitian ini adalah *Wrapper, Principal Component Analisys* (PCA), dan *Gain Ratio* yang diterapkan pada metode klasifikasi *Multilayer Perceptron* (MLP), *Decision Tree,* dan *Support Vector Machines* (SVM). Sedangkan untuk melakukan identifikasi *outlier* menggunakan algoritme K-*Nearest Neighbor* (KNN).

Penghapusan *outlier* pada *dataset fertility* dapat meningkatkan mayoritas performa klasifikasi secara signifikan. Selain itu juga mampu meningkatkan mayoritas performa seleksi fitur. Metode terbaik dari penelitian ini dihasilkan oleh MLP - *Gain ratio* dan MLP - *Wrapper* menggunakan *dataset* tanpa *outlier.*

Kata kunci: *Data mining*, *outlier,* k-NN, *Decision tree*, MLP, SVM, seleksi fitur, *Gain Ratio*, PCA, *Wrapper*.