Abstract

Outlier detection can be viewed as a classification problem if a training data set with class labels is available. Generally, in a typical medical dataset such as a cancer data set, if there are samples available with class information, then it is possible to apply a classification based outlier detection method. The general idea of classification-based outlier detection method is to train a classification model that can distinguish normal data from outliers [7]. Previous work had implemented and evaluated using the three classifications based outlier detection algorithms and found that the k-neighborhood algorithm was capable of identifying and classifying the outliers better than the other two compared algorithm in terms of accuracy, f-score, Sensitivity/Recall, error rate. Further, the cpu time of the k-neighborhood algorithm also minimum [23]. In this work, the performance of outlier detection using feature selection algorithms are evaluated but the results clearly shows that the impact of feature selection algorithm on the cancer dataset is very low and does not improve the overall classification performance.
References


16. Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining Concepts and Techniques (Third Edition)", Morgan Kaufmann Publishers is an imprint of Elsevier, © 2002 by Elsevier Inc.


Index Terms

Computer Science Information Sciences

Keywords

Outlier, Gini Index, Information Gain, Chi-Square