Comparison of expectation maximization and K-means clustering algorithms with ensemble classifier model

ABSTRACT

In data mining, classification learning is broadly categorized into two categories; supervised and unsupervised. In the former category, the training example is learned and the hidden class is predicted to represent the appropriate class. The class is known, but it is hidden from the learning model. Unlike supervised, unsupervised directly build the learning model for unlabeled example. Clustering is one of the means in data mining of predicting the class based on separating the data categories from similar features. Expectation maximization (EM) is one of the representatives clustering algorithms which have broadly applied in solving classification problems by improving the density of data using the probability density function. Meanwhile, Kmeans clustering algorithm has also been reported has widely known for solving most unsupervised classification problems. Unlike EM, K-means performs the clustering by measuring the distance between the data centroid and the object within the same cluster. On top of that, random forest ensemble classifier model has reported successive perform in most classification and pattern recognition problems. The expanding of randomness layer in the traditional decision tree is able to increase the diversity of classification accuracy. However, the combination of clustering and classification algorithm might rarely be explored, particularly in the context of an ensemble classifier model. Furthermore, the classification using original attribute might not guarantee to achieve high accuracy. In such states, it could be possible some of the attributes might overlap or may redundant and also might incorrectly place in its particular cluster. Hence, this situation is believed in yielding of decreasing the classification accuracy. In this article, we present the exploration on the combination of the clustering based algorithm with an ensemble classification learning. EM and K-means clustering algorithms are used to cluster the multi-class classification attribute according to its relevance criteria andafterward, the clustered attributes are classified using an ensemble random forest classifier model. In our experimental analysis, ten widely used datasets from UCI Machine Learning Repository and additional two accelerometer human activity recognition datasets are utilized.

Keyword: Expectation maximization; K-means; Random forest; Clustering; Classification