An Empirical Study on Initializing Centroid in K-Means Clustering for Feature Selection

Amit Saxena, Guru Ghasidas Vishwavidyalaya, India

John Wang, Montclair State University, USA

Wutiphol Sintunavarat, Thammasat University, Thailand

ABSTRACT

One of the main problems in K-means clustering is setting of initial centroids which can cause misclustering of patterns which affects clustering accuracy. Recently, a density and distance-based technique for determining initial centroids has claimed a faster convergence of clusters. Motivated from this key idea, the authors study the impact of initial centroids on clustering accuracy for unsupervised feature selection. Three metrics are used to rank the features of a data set. The centroids of the clusters in the data sets, to be applied in K-means clustering, are initialized randomly as well as by density and distance-based approaches. Extensive experiments are performed on 15 datasets. The main significance of the paper is that the K-means clustering yields higher accuracies in majority of these datasets using proposed density and distance-based approach. As an impact of the paper, with fewer features, a good clustering accuracy can be achieved which can be useful in data mining of data sets with thousands of features.

KEYWORDS

Centroid, Classification, Feature Selection, Information Gain, K-Means Clustering, Laplacian Score, Ranking Methods of Features in Data Sets, Variance

1. INTRODUCTION

The curse of dimensionality is a major problem in large datasets. A dimension is commonly known by names like feature or attribute or property or even column in a dataset. In order to save more and more information, many irrelevant features are also preserved in a dataset and these features can be contributing nothing while classifying the dataset for taking some inference out of it and sometimes even adding to misclassification of patterns. A dataset with large dimensionality may increase the time and space complexity wile classifying it. More specifically, the performance of a classifier depends on several factors: i) number of training instances. ii) Dimensionality, *i.e.*, number of features, and iii) complexity of the classifier (Saxena et al., 2010). Feature selection is an important component in pattern recognition (Duda et al., 2001). Feature Selection can be done in supervised or unsupervised manner. When feature selection techniques use the knowledge of class given in the data sets, it is called supervised feature selection. Feature selection without using class information is called unsupervised feature selection. For unsupervised feature selection, Mitra (Mitra et al., 2010), proposed a method that partitions original feature set into distinct subsets or clusters so that features in one cluster are highly similar while those in different clusters are dissimilar. A single feature is then selected from each cluster to form a reduced feature subset. Feature Selection for clustering is discussed in (Dash et al., 2000). Dy and Brodley (2000) presented a wrapper framework for feature

DOI: 10.4018/IJSSCI.2021010101

selection, clustering and order identification concurrently. Basu (Basu et al., 2000), discussed several methods for feature selection based on maximum entropy and maximum likelihood criteria but the proposed strategy for feature selection depends on the method used to estimate uni-variate data. Pal et al. (2000) proposed an unsupervised neuro-fuzzy feature *ranking* method. They used a criterion to measure the similarity between two patterns in the original feature space and in a transformed feature space. The transformed feature space is obtained by multiplying each feature by a coefficient *w* in interval [0,1]. This coefficient is learned through a feed-forward neural network. After training, the features are ranked according to the values of these weights. Higher values of w_i indicate higher importance and hence higher ranks. Using this rank, the required number of features is selected. A new correlation-based approach to feature selection (CFS) is presented in work from Hall (2000). CFS uses the features' predictive performances and inter-correlations to guide its search for a good subset of features. Experiments on discrete and continuous class datasets reveal that CFS can drastically reduce the dimensionality of datasets while maintaining or improving the performance of learning algorithms. The redundancy between two random variables X and Y is used to define a test of redundancy in (Heydon, 1971). This test can be used to eliminate redundant features without degrading performance of classifiers. Features that are linearly dependent on other features do not contribute towards pattern classification by linear techniques. In order to detect the linearly dependent features, a measure of linear dependence is proposed in (Das, 1971).

The present work focuses mainly on clustering i.e. unsupervised classification. For clustering, K-means (MacQueen, 1967) is among the most popular methods applied even today. In K-means clustering, the instances (patterns) are clustered on the basis of some similarity, mostly Euclidean distance. Shorter is the distance, higher will be the similarity. The patterns in a cluster have minimum Euclidean distance compared to those in other clusters. In other words, the inter-cluster distance among patterns in K-means is high, while the intra-cluster distance is low. K-means is simple, bench marked and easy to use and this is one of the reasons for its popularity. But some limitations for K-means include (a) deciding the initial centroids (b) deciding the value of K (c) even if a pattern is quite awkward but due to compulsion of putting it into one of the clusters, it is forced to fit into one cluster which reduces the clustering accuracy (CA) of a cluster. There have been various observations and proposals to modify K-means clustering. An algorithm similar to k-means, known as the Linde-Buzo-Gray (LBG) algorithm, was suggested for vector quantization (VQ) (Gersho et al., 1992) for signal compression. In this context, prototype vectors are called code words, which constitute a code book. VQ aims to represent the data with a reduced number of elements while minimizing information loss. Fuzzy c-means (FCM) is a clustering method which allows one point to belong to two or more clusters unlike *K*-means where only one cluster is assigned to each point. This method was developed by Dunn (Dunn, 1973) and improved by Bezdek (1981). The procedure of FCM (Xu et al. 2005) is similar to that of basic K-means (MacQueen, 1967). However the time complexity of K means is much less than that of FCM thus K means works faster than FCM (Ghosh et al. 2013).

Recently published Density and Distance based Centroids selection for K-means (hereafter expressed as DD based K-means) by Duan et al. (2018) reports a faster algorithm where, in a few number of iterations, the clustered are formed with the centroids which do not change in further iterations; compared to basic K-means which takes more numbers of iterations for settling to fixed centroids. Motivated by these findings and limitations of K-means, its performance under normal (initial random centroids) against DD based centroid selection is investigated in this paper. Another interest is to see the accuracies obtained by a subset of features under these two schemes. It is admitted that the usual supervised classification methods will produce higher accuracies (as target class is known during training) but the objective here is not to evaluate the proposed work on the basis of accuracies, but to observe how much variations are reported under these two schemes and under different sizes of subsets of features. The benchmark data sets with different sizes of features and patterns have been used to see the performance of DD based K-means clustering. The main contribution is to develop a new model for unsupervised feature selection. To justify it, K- means clustering for testing