The 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)

Labeled Clustering

A Unique Method to Label Unsupervised Classes

Muhammad Shaheen[1], Saeed Iqbal[2], Fazl –e – Basit[3]
Department of Computer Science
National University of Computer & Emerging Sciences
Peshawar, Pakistan
[1]Muhammad.shaheen@nu.edu.pk, [2]saeediqbalkhattak@gmail.com, [3]fazl.basit@nu.edu.pk

***Abstract*— This paper proposes a method to label unsupervised classes. Clustering is an unsupervised classification technique that is used to group data on the basis of similarity measures. K-Means clustering is one of the methods which is used to classify given dataset in number of groups on the basis of Euclidean distance of data points in Cartesian system. On the basis of similarity and dissimilarity the data points are divided into multiple clusters which do not have identification labels. K means clustering can given a broadened insight into the data if the resulting groups bear some identification. A unique method for labeling unsupervised classes by using correlation analysis and frequent membership function is proposed. The method is applied to custom world energy dataset and divided world nations into five labeled clusters which increased the opportunities for energy sector to derive valuable patterns for guided decisions. The results showed minor deviations from the real energy scenario because of the factors discussed in the paper.**

Keywords- K Means Clustering, Unsupervised classification, Labels, Correlation analysis, Sustainability indicators

I. INTRODUCTION

Clustering is to find the likely grouping among different objects. In broader sense, clustering is categorized as: (1) Hierarchical clustering; and (2) Partitioned clustering. Hierarchical clustering is based on hierarchical breakdown of the set of items using some criterion. Partitioned clustering construct various partitions and assesses them by some criterion. K-means and Gaussian Mixture Models are considered as Partitioned clustering techniques.

Methods of classification in the field of Data Mining are categorized as: (1) Supervised Classification and (2) Unsupervised Classification. User is known to inputs with class labels in supervised classification while class labels are not applied to input data in Unsupervised Classification. Clustering techniques are used to group unlabeled pattern of data and is known as unsupervised classification technique. A generic clustering process would involve pattern presentation, definition of pattern proximity measure and data abstraction and grouping.

For clustering huge datasets, number of methods has been explored. Gao and Hitchcock presented nomenclature of these methods [7]. In this paper, we shall propose a method based upon squared error partitioned K-mean clustering.

According to Al-Sultan and Khan [8], in K-mean clustering first of all cluster centers are randomly selected. Euclidean distance (ED) is found from each cluster center for every data point and dataset is allotted to the closest cluster center. Following these new cluster centers the criterion function is computed [8]. Finally, cluster centers are computed repeatedly for calculating new assignments until no new allocation remains possible. K-mean algorithm converges to a local minimum by dividing the datasets into required number of clusters [7, 15]. Since unsupervised classification does not produce class labels and group the data on the basis of intra cluster similarity and inter cluster dissimilarity, K-Means clustering also follows the convention. Hence it does not specify the cluster labels. In certain applications cluster labels are necessary to make results fruitful in decision making [11]. Comprehensive information regarding k-mean algorithm is available in the existing literature [5, 16].

In this paper we propose a unique technique for labelling clusters obtained after applying K-means clustering to a dataset. The labels for classified datasets are derived by a statistical method. This labelling facilitated user to see supervised results of unsupervised classes. The supervision is proposed on the basis of rank of each cluster member. The algorithm is applied on energy dataset of different countries to rank their status with respect to sustainable energy development in the country. Sustainability is defined as, ‘‘The development that meets the needs of today without compromising the ability of future generations to meet their needs’’ [1]. Energy sources are the backbone of of a country’s economy and to till date they are mostly obtained from hydrocarbons. Energy distributors are seriously working on extraction of hydrocarbons to meet the energy needs of the consumers as good percentage of people are using other alternative energy sources such as animals and other low cost non-commercial fuels. The two major concerns are: (1) Developing nations are still lacking in resolving the energy dependent socio-eco-economic aspects; and (2) Large scale economic recession is observed despite the fact that remarkable energy planning for energy extraction and dissemination is made.

The proposed algorithm can classify the world nations with respect to their value of sustainable energy development. The