Rough ISODATA Algorithm

S. Sampath, Department of Statistics, University of Madras, Chennai, Tamil Nadu, India
B. Ramya, Department of Statistics, University of Madras, Chennai, Tamil Nadu, India

ABSTRACT

Cluster analysis is a branch of data mining, which plays a vital role in bringing out hidden information in databases. Clustering algorithms help medical researchers in identifying the presence of natural subgroups in a data set. Different types of clustering algorithms are available in the literature. The most popular among them is k-means clustering. Even though k-means clustering is a popular clustering method widely used, its application requires the knowledge of the number of clusters present in the given data set. Several solutions are available in literature to overcome this limitation. The k-means clustering method creates a disjoint and exhaustive partition of the data set. However, in some situations one can come across objects that belong to more than one cluster. In this paper, a clustering algorithm capable of producing rough clusters automatically without requiring the user to give as input the number of clusters to be produced. The efficiency of the algorithm in detecting the number of clusters present in the data set has been studied with the help of some real life data sets. Further, a nonparametric statistical analysis on the results of the experimental study has been carried out in order to analyze the efficiency of the proposed algorithm in automatic detection of the number of clusters in the data set with the help of rough version of Davies-Bouldin index.

Keywords: Boundary, Clustering, Fuzzy Clustering, ISODATA, Lower Approximation, Rough Clustering, Rough Davies-Bouldin Index

INTRODUCTION

In machine learning studies, practitioners come across data sets of varying nature thereby making it necessary to develop specialized methodologies for extracting hidden information in the data set. Broadly speaking, objects in the data set may possess random, fuzzy or rough characteristics. Randomness comes into picture when uncertainty is present in the values assumed by objects in the dataset. For example, in datasets related to responses of patients, in the process of recovery from illness one can see lot of uncertainties due to various random factors, even though they receive the same type of treatment. Fuzziness comes into picture when impreciseness exists in the form of vagueness in the values assumed by the variables. For example, the laboratory measurements with respect to different diagnostic tests can hardly be treated as crisp values and in practice, they are usually of imprecise nature. Very rarely, one can get the exact status of a person with respect to the variables of these kinds because the environment in which they are recorded influences the experimental values. To deal with these kinds of fuzzy data sets machine-learning researchers use the theory of fuzzy sets developed by Zadeh (1965). Interesting discussion on this aspect can be found in Azar (2010). The third type of
data namely rough data arises in the form of coarseness. For example, two individuals may possess the same set of attributes but belong to different classes for which no proper explanation can be given. Examples of such datasets can be found in Pawlak (1982).

The idea of using rough sets in cluster analysis was initially proposed by Lingras, Yan and West (2002) in which a comparative study on conventional and rough \(k\)-means clustering had been carried out. Following this, Lingras and West (2004) considered an application of rough \(k\)-means for clustering of web users. Mitra (2004) considered an evolutionary solution for rough \(k\)-means clustering. Further studies in relation to rough \(k\)-means clustering have been carried out by Peters and Borkowski (2004), Peters (2005, 2006), Peters and Lampart (2006) and Peters, Lampart and Weber (2008). Lingras and Peters (2012) described how a core concept of rough sets, the lower and upper approximation of a set, can be used in clustering. In their work, rough clusters were shown to be useful for representing groups of highway sections, web users, and supermarket customers. A main limitation of these works is that the application of rough \(k\)-means algorithm requires the knowledge of \(k\), namely the number of clusters. This paper provides a solution for this limitation. An algorithm similar to the ISODATA algorithm of Ball and Hall (1967) is proposed for creation of rough clusters without actually assuming the value of \(k\). It is to be noted that fuzzy relative of the ISODATA algorithm was developed by Dunn (1973) and another related work based on a technique called Probabilistic relaxation labelling was also studied by Lam (1996).

The paper is organized as follows: The paper begins with brief descriptions of crisp \(k\)-means and fuzzy \(k\)-means. In order to enhance the readability of the paper it is followed by a complete detailed description on rough \(k\)-means clustering algorithm due to Lingras et al. (2002). The paper introduces a new rough clustering algorithm, namely, rough ISODATA algorithm and the concluding section discusses the outcome of the experimental studies carried out with the help of real life data sets to study the efficiency of the proposed algorithm.

### RELATED WORK

Cluster Analysis finds lot of applications in various fields of scientific research. Several applications of cluster analysis in areas like medical imaging, medical diagnostics, segmentation of patients based on their illness history as well as clinical observations are available in the literature. Different types of clustering algorithms suitable for different domains of studies are being developed for nearly five decades. Among various clustering approaches, partitioning clustering approach has attracted the attention of several researchers. A partitioning clustering algorithm creates a division of the set of data objects based on some well defined criteria. Several partitioning algorithms are available in the literature. Among these, \(k\)-means algorithm is a popular method of clustering.

#### \(k\)-Means Clustering

\(k\)-means algorithm (MacQueen, 1967) is a partitioning, prototype based algorithm. The algorithm starts by choosing \(k\) initial centroids, where \(k\) is a user-specified parameter, namely, the desired number of clusters to be formed using the data set. Each object in the data set is assigned to the nearest centroid, and the collection of objects assigned to a particular centroid forms a cluster. The centroids are revised based on the objects assigned to the clusters. The above process is repeated till either there is no change in the assignment of objects to the cluster or a well-defined termination criterion is satisfied. In order to give a formal definition of the algorithm we introduce the following notations.

Let \(N\) be the number of objects in the data set, \(k\) be the number of clusters present in the data set and \(X_n\) be the \(p\)-dimensional data vector corresponding to the \(n\)\(^{th}\) object in the data set \((n = 1, 2, \ldots, N)\). Denote by \(C_u\) the \(u\)\(^{th}\) cluster and take \(m_u\) as the centroid for the \(u\)\(^{th}\) cluster where \(u = 1, 2, \ldots, k\).

**Algorithm**

1. Choose arbitrarily \(k\) vectors in the data set as initial set of centroids.
A Neuro-Fuzzy Expert System Trained by Particle Swarm Optimization for Stock Price Prediction
www.igi-global.com/chapter/neuro-fuzzy-expert-system-trained/62711?camid=4v1a

Sign Language Recognition for Bengali Characters
www.igi-global.com/article/sign-language-recognition-for-bengali-characters/133123?camid=4v1a