Chapter 4
Genetically-Modified K-Medoid Clustering Algorithm for Heterogeneous Data Set

Dhayanithi Jaganathan
Sona College of Technology, India

Akilandeswari Jeyapal
Sona College of Technology, India

ABSTRACT

In recent days, researchers are doing research studies for clustering of data which are heterogeneous in nature. The data generated in many real-world applications like data form IoT environments and big data domains are heterogeneous in nature. Most of the available clustering algorithms deal with data in homogeneous nature, and there are few algorithms discussed in the literature to deal the data with numeric and categorical nature. Applying the clustering algorithm used by homogenous data to the heterogeneous data leads to information loss. This chapter proposes a new genetically-modified k-medoid clustering algorithm (GMODKMD) which takes fused distance matrix as input that adopts from applying individual distance measures for each attribute based on its characteristics. The GMODKMD is a modified algorithm where Davies Boudlin index is applied in the iteration phase. The proposed algorithm is compared with existing techniques based on accuracy. The experimental result shows that the modified algorithm with fused distance matrix outperforms the existing clustering technique.

INTRODUCTION

The nature of data with high fluctuation and different characteristics are called heterogeneous data. In general, integrating the heterogeneous data is very difficult to meet the business information requirements. In recent days the data generated from IoT are often heterogeneous nature. The heterogeneous data are further classified into four different characteristics namely, numeric, binary, nominal and ordinal. The data are in measurable form or numeric forms are called numerical data. The data that falls on two states 0 or 1 are called binary data. The data which simply names or label something without any

DOI: 10.4018/978-1-5225-9902-9.ch004

ordered is called Nominal data. Ordinal data are extension of nominal data is follows an order. Apart from the characteristics of the data, it also important to know much about those data is in the form of metadata management. For the better interpretation of heterogeneous data detailed metadata information are required. In many cases it is difficult to collect those metadata.

Grouping objects into similar clusters is the prime motive of any clustering techniques. Similarity or Dissimilarity is measured by how far the objects are close enough together. Majority of similarity measure have been studied and tested in the literature. The similarity measure are falls in two categorized first the data with numerical value and second the data with conceptually categorical. The similarity measures available for one type of data are not suitable for other type of data. The challenges of clustering heterogeneous data concentrate on designing in tackling the difficulties raised by complex and dynamic characteristics, volume of data, and defining the good similarity measure to know the similarity between the objects in order to group them together. More focused research on similarity or dissimilarity measure for heterogeneous dataset was already carried out by many researchers. Study is needed for defining perfect similarity or dissimilarity measures of heterogeneous types.

Machine learning is the design of algorithms that permit machines to develop behaviors based on empirical data. Most of research work carried out in machine learning is that make the computer to automatically learn by themselves. Machine Learning is defined as any algorithm can learn by themselves based on the Experience (E) obtained from certain Task (T) and the Performance measure (P) of that Task T is keep on improving by their Experience (E). Based on the outcome of the algorithm machine learning can be classified in to two types namely, supervised and unsupervised. In supervised learning, function generates to map the input to desired outputs and in unsupervised learning, a set of inputs were modeled like clustering. Machine Learning is performed by various strategies and techniques namely, Inductive Logic Programming, Simulated Annealing, Neural Nets and Evolutionary Strategies. The first three techniques are beyond the scope of this chapter and an only evolutionary strategy is currently focused.

Genetic Algorithm is a heuristic search which is widely used in search optimization and finding the optimal solution based on natural evolution. Genetic Algorithm is a subset of evolutionary algorithm in which the offspring of the next generation is incurred by fittest individuals of current generation. Genetic Algorithm comprises of five phases namely, Population Initialization, fitness function, selection, cross over and mutation. It is necessary to incorporate Genetic Algorithm with clustering techniques because clustering is the key task in the process of acquiring knowledge. The cluster analysis is usually observed by measuring the natural association of members in the clusters i.e., the natural association of members within the group is high compared to the members in different group. Even for a small set of elements (25) to be clustered in small set of groups (5) arise a very large number of possibilities (2,436,684,974,110,751). The clustering task is incorporated with Genetic Algorithm leads to minimize the within cluster variance.

This paper focuses on two aspects, 1) Formulate the fused distance matrix for the heterogeneous data types and 2) Genetically modified K-Mediod clustering algorithm with modified Davies Bouldin Index as the fitness function.