Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 534 – 539

Procedia Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

Clusters of Genetic-Based Attributes Selection of Cancer Data

Vijaya Sri Kompalli[a,*] and K. Usha Rani[b]

[a] *Devineni Venkata Ramana & Dr. Hima Sekhar MIC College of Technology, Vijayawada 520 003, India*
[b] *Sri Padmavati Mahila Visvavidyalayam, Tirupati 517 502, India*

Abstract

Clustering of data simplifies the task of data analysis and results in better disease diagnosis. Well-existing K-Means clustering hard computes clusters. Due to which the data may be centered to a specific cluster having less concentration on the effect of the coupling of clusters. Soft Computing methods are widely used in medical field as it contains fuzzy natured data. A Soft Computing approach of clustering called Fuzzy C-Means (FCM) deals with coupling. FCM clustering soft computes the clusters to determine the clusters based on the probability of having memberships in each of the clusters. The probability function used, determines the extent of coupling among the clusters. In order to achieve the computational efficiency and binding of features genetic evaluation is introduced. Genetic-based features are identified having more cohesion based on the fitness function values and then the coupling of the clusters is done using K-Means clustering in one trial and FCM in another trial. Analysis of coupling and cohesion is performed on Wisconsin Breast Cancer Dataset. Nature of clusters formations are observed with respect to coupling and cohesion.

Keywords: Cluster; Coupling; Cohesion; Genetic Algorithm; Fuzzy C-Means.

1. Introduction

Cluster Techniques tend to group the data based on the similarities among calculated values. Clusters help to determine the structure and the nature of existence of the data, based on which it is easy to derive an accurate problem solution. Clusters are formed based on commonalities among data and are centered towards a point based on the measure of distance where the points are located. In an unsupervised environment it is always good to form clusters and to study the nature of data and its solution states. Clustering helps to determine the coupling and cohesion. "*Coupling*" determines the separation of clusters and "*Cohesion*" determines the closeness of values within a cluster.

Cluster Techniques[1] are mainly used where there is a need to track the continuously changing states of data values. Clusters are very much required to carry out the diagnosis of medical data. Cancer is a disease which results in an abrupt change of disease properties within no time depending on the biological conditions of the patient's clinical results. Breast cancer widely spread among women in foreign countries and in few metropolitan regions of India. The attribute values of clinical samples of Breast Cancer tend to variation to a greater extent, depending on the state of the disease. There is a shift of data points from one region to the other region due to which the structure changes and also it becomes difficult to determine the solution to diagnosis the disease.

[*] Corresponding author. Tel.: +91 9849889288.
E-mail address: vijayasri.kompalli@mictech.ac.in

doi:10.1016/j.procs.2016.06.098

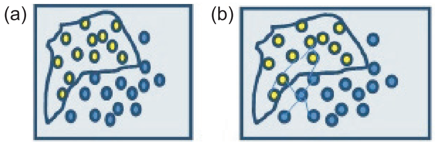

Fig. 1. (a) Cohesion; (B) Coupling.

Table 1. Wisconsin Breast Cancer Dataset.

Serial	Attribute Information	Domain Description
v1	Sample Code Number	Id Number
v2	Clump Thickness	1–10
v3	Uniformity of Cell Size	1–10
v4	Uniformity of Cell Shape	1–10
v5	Marginal Adhesion	1–10
v6	Single Epithelial Cell Size	1–10
v7	Bare Nuclei	1–10
v8	Bland Chromatin	1–10
v9	Normal Nucleoli	1–10
v10	Mitoses	1–10
v11	Class	2-Benign; 4-Malignant

Clustering identifies the location of data points, measures the distance and locates it to a region based on the centroid of the cluster. Clustering technique is implemented technically using various methods. K-Means clustering is one of the efficient methods of clustering. Another method which is widely used is Fuzzy C-Means Clustering (FCM). In this study, K-Means clusters formation, is done and also FCM clusters are formed to determine the variations of coupling and cohesion. Also, the attributes to which clustering is applied is chosen from Genetic Algorithms. Genetic Algorithms are applied in order to identify the most prominent attributes that are prone to the disease diagnosis. Using Genetic Algorithms, best fitness based attributes that contribute more to know the status of the disease are identified and only those attributes are chosen to form clusters. The reason to choose prominent attributes is to derive the computational efficiency in diagnosis and also to identify the nature of clusters in terms of most prominent attributes.

2. Literature Survey

Floating centroid method is proposed by Xiaoqian Zhang *et al.* which diminishes the sensitivity to outliers[1]; in their study a different method of centroid related changes in the clusters are specified which is somewhat similar to coupling and cohesion. In the work of Hans van Bakel, future directions specify to consider the points for redesign in any field using coupling and cohesion[2]. Genetic Algorithms are used to cluster data in various forms[3] with desirable fitness computations of the objective function used. Very few works are carried on coupling and my experiments are an attempt to show the effect of coupling and cohesion on medical data. Little effort to include theoretical approach to a practical implementation is attempted.

3. Materials and Methods

3.1 Dataset

Dataset used to perform the clustering is Wisconsin Breast Cancer Dataset (WBCD). WBCD contains a total number of 11 attributes and 699 rows. All the 16 missing values are replaced by the mean of the respective attributes. The data description of WBCD is shown in the Table 1. All the attributes are given serial representation of v1 to v11. The data distribution of clinical samples are scaled over 1–10 values.

Out of all the attributes, the first attribute is removed in the experimental data as it is not an important value required for diagnosis. The remaining attributes are evaluated based on Genetic Algorithm to derive the most prominent