ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.7

Implementation & Analysis of Clustering Techniques in Bioinformatics: Cancer Research

Dr. Gurpreet Singh[1], Karan Jamla[2]
Professor[1], M.Tech Student[2]
Department of Computer Science Engineering
St. Soldier Institute of Engineering & Technology, Jalandhar, India

Abstract:
Advances in cancer medicine have traditionally come from detailed understanding of biological processes, later translated into therapeutic interventions, whose effectiveness is established by rigorous analysis of clinical trials. Computer-aided cancer prediction and risk assessment has become a very useful tool and is starting to be taken seriously by the medical community. Advanced bioinformatics and data mining techniques are used extensively to assist in predicting the chances of an individual patient's cancer occurrence as well as the population cancer rates in general. These techniques rely heavily on analyzing and comparing genetic and medical datasets, as well as environment-based and other factors. Clustering is the first line analysis methodology to mine data bases for useful research hypotheses to take onto further, more directed, study from a mathematical perspective the objects of study, be they genes, patients, mode of action drugs or disease sub-types, are points in a multidimensional space where similarity is defined, typically trough a measure of the distance between object pairs. The main classes of clustering approaches used in data mining for cancer are partitional, creating a simple partition of the data, or hierarchical, which create a tree structured hierarchy of the data. Partitional approaches form a very heterogeneous class, including techniques based on Global Optimization, Vector Quantization, Kernel functions, Fuzzy sets, and Graph theory. The HBCA proposed algorithm combines the features of BIRCH clustering algorithm whose feature of insertion and splitting is same as B-Tree algorithm and Partitioning clustering algorithm K-Means algorithm.

1. INTRODUCTION:

Cancer prediction is certainly a very complex and nondeterministic endeavor. Estimating the probability of cancer occurrences in patients requires that many factors (both genetic and non-genetic) are evaluated and properly weighted according to their significance and/or other (context sensitive) contribution factors. In recent years, rapid developments in genomics and proteomics have generated a large amount of biological data. Drawing conclusions from these data requires sophisticated computational analyses. Bioinformatics, or computational biology, is the interdisciplinary science of interpreting biological data using information technology and computer science. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data. A particular active area of research in bioinformatics is the application and development of data mining techniques to solve biological problems. Analyzing large biological data sets requires making sense of the data by inferring structure or generalizations from the data. Examples of this type of analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc. Therefore, we see a great potential to increase the interaction between data mining and bioinformatics.

2. CLUSTERING:

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data 2 concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements.

3. CLUSTERING ALGORITHMS IN DATA MINING

Based on the recently described cluster models, there are a lot of clustering that can be applied to a data set in order to partition ate the information. In this article we will briefly describe the most important ones. It is important to mention that every method has its advantages and cons. The choice of algorithm will always depende on the characteristics of the data set and what we want to do with it.

a. Centroid-based

In this type os grouping method, every cluster is referenced by a vector of values. Each object is part of the cluster whose value difference is minimal, comparing to other clusters. The number of clusters should be pre-defined, and this is the biggest problem of this kind of algorithms. This methodology is the most close to the classification subject and are vastly used for optimization problems.

b. Distributed-based

Related to pre-defined statistical models, the distributed methodology combines objects whose values belongs to the same distribution. Because of its random nature of value generation, this process needs a well-defined and complex model to interact in a better way with real data. However these processes can achieveda optimal solution and calculate correlations and dependencies.

c. Connectivity-based

On this type of algorithm, every object is related to its neighbors, depending the degree of that relationship on the distance between them. Based on this assumption, clusters are created with nearby objects, and can be described as a maximum distance limit. With this relationship between members, these clusters have hierarchal representations. The distance function varies on the focus of the analysis.

d. Density-based

These algorithms create clusters according to the high density of members of a data set, in a determined location. It aggregates some distance notion to a density standard level to group members in clusters. These kind of processes may have less performance on detecting the limit areas of the group.

4. PROPOSED ALGORITHM

The HBCA algorithm combines the features of BIRCH clustering algorithm whose feature of insertion and splitting is same as B-Tree algorithm and Partitioning clustering algorithm K-Means algorithm. The algorithm is applied on cancer dataset which is collected from a bank. The HBCA algorithm first make call to tree algorithm which is named as Kmeans algorithm that build a tree containing more than 1500 clusters on cancer dataset. The insertion and splitting of this tree algorithm is same as B Tree algorithm but in this algorithm each node of the tree stores the node or tree label, the cluster number and the number of instances in that cluster. These large numbers of clusters are difficult to predict and understand. After that the algorithm make call to K-Means clustering algorithm which clusters the leaf nodes of the clustering algorithm. In K-means we have to prior define the number of clusters. In this paper the comparison is done among proposed algorithm, K-Means and K-Medoid algorithm by changing the number of clusters.

5. RESULTS AND IMPLEMENTATION:

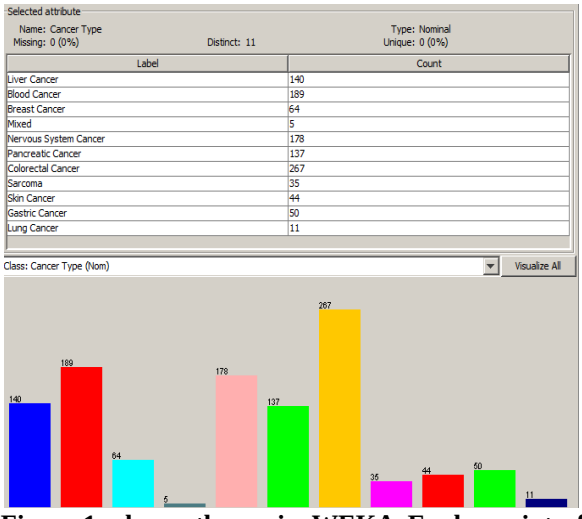

Figure.1. shows the main WEKA Explorer interface with the dataset loaded.

The last attribute cancer type is taken as a class attribute by the WEKA . This attribute contains 11 categories: The count of number of instances under each category in the dataset is shown numerically as well as graphically.

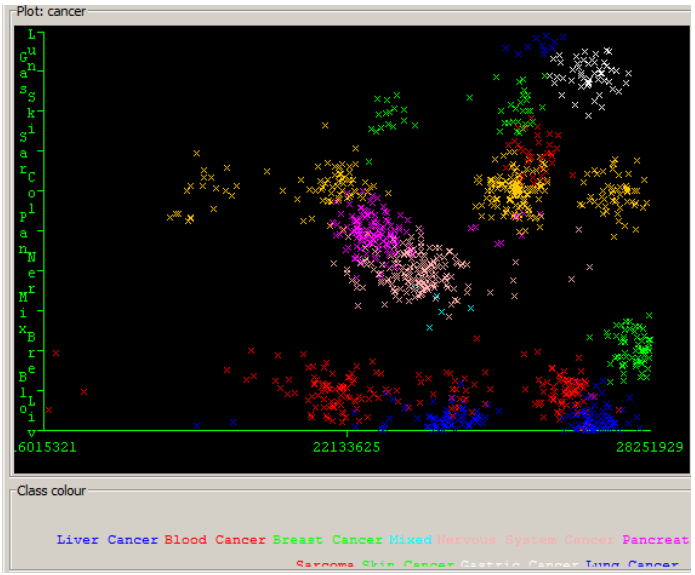

Figure.2. shows the clusters between two attributes that are PubMed Id and Cancer Type attribute. X-axis shows the PubMed Id. In figure the minimum value, middle value and maximum value ofPubMed Id attribute is depicted. Y-axis specifies the eleven Cancer Type. The eleven clusters are depicted by different eleven colors.

Table.1.Analysis of No. of Iterations, Error Rate, No. of Clusters

	Kmeans	**HBCA**	**KMedoid**
No. of Iterations	39	31	44
Error Rate	7185.19	7185.19	7184.89
No. of Clusters	6	6	6

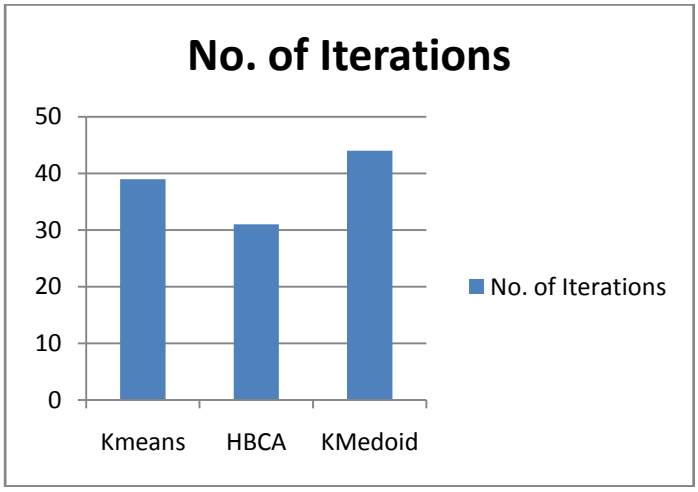

Figure.3. Analysis between No. of Iterations

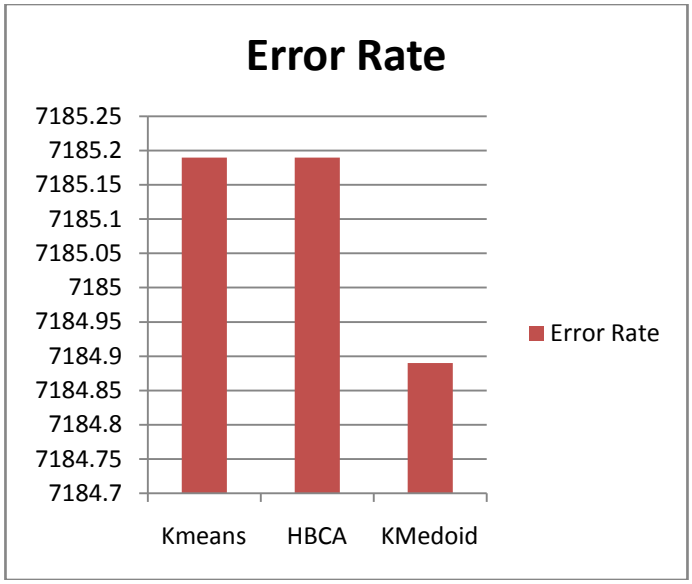

Figure.4. Analysis Between Error Rate

6. CONCLUSION:

In this research , study is being done on clustering algorithms. The features of traditional KMeans & KMedoid algorithms are combined and a new algorithm HBCA proposed. The comparison of proposed algorithm is done with the existing algorithm KMeans & KMedoidon Cancer dataset using WEKA data mining tool. The results by changing the No. of iterations, Error Rates value specifies that the proposed method gives better performance than KMeans&KMedoidby reducing the sum of square error which signifies that HBCA have high intra classification similarity and is more accurate. Also the proposed algorithm can handle large datasets more effectively.

7. REFERENCES

[1]. T. Zhang, R. Ramakrishnan, M. Linvy, “BIRCH: an efficient data Classification method for very large databases” (1996) ACM SIGMOD International Conference on Management of Data.

[2]. L.Kovács, L. Bednarik, “Parameter Optimization for BIRCH Pre-Classification Algorithm”, 2011 IEEE International Symposium on Computational Intelligence and Informatics, Budapest, Hungary.

[3]. W.-L. C. T.-H. Lin, "A Classification-Based Approach for Automatic Social Network Construction",2010 IEEE Second International Conference on Social Computing (SocialCom).

[4]. Wei-Lun Chang, Tzu-Hsiang Lin, “A Classification-Based Approach for Automatic Social Network Construction”, 2010 IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust

[5]. Ji Dan, QiuJianlin, Gu Xiang, Chen Li, He Peng “A Synthesized Data Mining Algorithm Based on Classification and Decision Tree” 2010 IEEE International Conference on Computer and Information Technology.

[6]. R.Xu, “Survey of Classification Algorithms” 2005 IEEE Trans.Neural Networks.