Journal of Engineering Sciences

Vol 11, Issue 7,July/2020

ISSN NO:0377-9254

MULTISTAGE APPROACH OF MEDICAL DATA USING CLUSTERING

K.Sasirekha, Dr.V.Kathiresan

Ms.K.Sasirekha,Research *Scholor,Department of CS,Dr.SNS Rajalakshmi College of Arts & Science*

Dr.V.Kathiresan, *Director,Department of MCA,Dr.SNS Rajalakshmi College of Arts & Science*

***Abstract*—** Data mining refers to the process of retrieving knowledge by discovering novel and relative patterns from large datasets. Clustering and Classification are two distinct phases in data mining that work to provide an established, proven structure from a voluminous collection of facts. Accurate and fast approaches for automatic medical data classification are vital for clinical diagnosis of heart disease.

The analysis of medical data is currently a key topic in biostatistics and machine learning applications to medical research. The effective data acquisition is subject to many obstacles. In medical care several restrictions arise from ethical and experimental issues. The high costs associated with medical analysis lead to less frequent tests, which in turn results in longitudinal data that is usually sparse and incomplete, with unequal and long sampling periods, which may severely hamper the analysis and the correct identification of significant covariates. In fact, most of this data translates into time series, since the corresponding patients are medically followed for a period of time, which leads to the relevancy of longitudinal data analysis methods as a way to better understand and interpret medical data

Keywords: Clustering , Classification, Supervised and Unsupervised Learning

1. INTRODUCTION

Medical area has huge amount of data that require processing and analysis in order to extract useful information that sometimes might save a human life. Medical data include patient records, test results, or some type of images such as X-rays, MRI, ECG, EEG and CT scans. In order to analyze these data, supervised and unsupervised learning techniques are necessary to facilitate data handling and decision making. This research concerns only about supervised and unsupervised learning algorithms. These techniques were heavily employed in medical field. In the previous work, fuzzy clustering is employed to generate the weights of every instance in the dataset to which class it belongs to introduce additional significant features added to the data [9]. The data is then fed to modify SVM for classification process. The results are not in the expected level of the researcher and further it can be improved using the alternative approach of the proposed one. It overcomes the other works presented in the literature.

2. This research shows that the proposed novel approach can be employed as a powerful tool to facilitate final decision of medical diagnosis and can be successfully applied for various medical data classification. Although the result of this research is promising, number of general directions remains open to extend this work. This research can be extended to investigate other real-world problems of different domains. Also, testing the scalability of the enhanced multistage system is an interesting subject. Furthermore, the dataset used in this research is a well-known two-class problem; and in our work can be evaluating the performance of the proposed system on other multiclass problems [3]

3. In our proposed multistage approach including various procedures for dimensionality reduction, consensus clustering of randomized samples, followed by the use of a fast supervised classification algorithm. The performance of the classification algorithm at the final stage is crucial for the effectiveness of this technique. It can be regarded as an indication of the reliability, quality and stability of the combined consensus clustering.

3.PROBLEM SPECIFICATION

Medical data collection from different sources, classification of information according to their property, traits and arranging the information according to the sub group is very important. Due to this reason, to implement the data mining and

machine leaning techniques are useful to get the user decisions. Machine learning techniques can be classified into two categories, called supervised learning and unsupervised learning. This research work, focus on to analyze clusters of medical data records obtained via unsupervised clustering techniques and compare the performance through classification algorithm on the medical data. New feature selection is accomplished in the class of the dimensionality aid supervised approach referred to as igPCA that attempts to choose a subset of the predictor features in line with the guidelines achieve. After that the easy k-means clustering set of rules randomly chooses K segments as centroids of clusters at the initialization stage. This proposed manner introduce the consensus serve as with the combo of CSPA, HGPA, and MCLA, which might be used to combine unbiased clustering into one final consensus clustering. The result of consensus clustering is used to train the classification set of rules known as SVM. The performance of the classification algorithm at the final stage is crucial for determining the effectiveness of the classification algorithm. Performance evaluation can be regarded as an indication of the reliability, quality and stability of the combined consensus clustering.

4. OBJECTIVE OF THE RESEARCH

- To improve an accurate and computationally efficient means of classifying medical data sets in the hope that it may provide some valuable information worth looking at for various physiological and biomedical analyses.
- To segment, novel multistage approach combines dimensionality reduction algorithms, multiple unsupervised clustering algorithms and supervised classification algorithms in such a way that efficient and accurate profiling of very large and highly dimensional medical data sets can be achieved.
- To enhance prototypes reduce the influence of the differences in distribution and density of the data on the clustering result, while the similarity-driven merging helps determine a suitable number of clusters, starting from an overestimated number of clusters.
- To improve the consensus serve as with the combo of CSPA, HGPA, and MCLA, this might be used to combine unbiased clustering into one final consensus clustering. The result of consensus clustering is used to train the classification set of rules known as SVM are proposed.

Dimensionality Reduction using igPCA

Reduction of large datasets can be performed by reducing the number of analyzed parameters (dimensions) or by decreasing the number of analyzed cases. The dimensionality reduction can be carried out through statistical methods, primarily Principal Component Analysis (PCA) [11] or by using feature selection techniques [14, 15]. Dataset cardinality reduction can be achieved by sampling, grouping or instance selection methods [16].

In this research, we propose a modification to the application of PCA method called igPCA (in-group Principal Component Analysis). It introduces the pre-processing phase that arranges the related features into groups of similar distribution. Further, applied our method to reduce data derived from ECG signals to improve storage and inference process in solving arrhythmia classification problem.

The proposed in-group feature extraction method (igPCA) is based on principal component analysis incorporating diversity in distribution of various parameters.

5.2. Clustering with Consensus Function

The fundamental objective of this algorithm is to determine the optimal and minimum number of clusters in combination with K-SVD as a dictionary learning method and CS theory as a random sampling approach.

The K-SVD improves the K-Means clustering process for adapting dictionaries in order to achieve sparse ECG signal representation. Therefore, the main objective of KSVD is to train the suitable dictionary to generate compressed ECG signals. The K-Means algorithm based on K-SVD method states that the sparse pattern of a dataset can be recovered from a set of low-dimensional random linear measurements. In the K-SVD learning method in combination with CS theory, rather than measuring each sample and then computing a compressed representation, we can measure and collect a compressed representation of the dataset directly. Clustering ensembles have emerged as a powerful method for improving both the robustness as well as the stability of unsupervised classification solutions [10]. However, finding a consensus clustering from multiple partitions is a difficult problem that can be

approached from graph-based, combinatorial or statistical perspectives. This study extends previous research on clustering ensembles in several respects [7].

This proposed multistage approach, introduce a novel consensus function is proposed based on a dual similarity measure - the similarity between initial clusters and the similarity between candidate clusters and uncertain objects. Our consensus function does not require a given cluster number as a parameter. The algorithm is hence called Dual-Similarity Clustering Function, DSCF. In this section, we address this problem and describe the consensus functions that are used to combine independent clustering into one final consensus clustering. Here, further investigated the performance of three consensus clustering algorithms incorporated in the proposed multistage scheme. These methods are cluster-based similarity partitioning algorithm (CSPA), hypergraph partitioning algorithm (HGPA), and the metaclustering algorithm (MCLA).

PERFORMANCE EVALUATION

In this work, the performance of the proposed dimensionality reduction, cluster ensemble and classification is evaluated and the performance results are compared with existing fuzzy clustering approach.

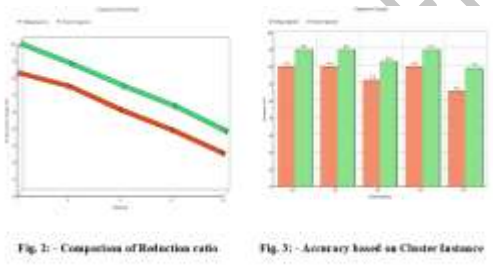

Fig. 2: - Comparison of Reduction ratio

Fig. 3: - Accuracy based on Cluster Instance

CONCULSION AND FUTURE WORK

This research introduces the essential needs of Supervised and Unsupervised learning for medical data classification. The proposed a multistage approach based on consensus function clustering for medical data classification. The system involves three main stages. In the first phase implies dimensionality reduction algorithm, and the second phase implemented the multiple unsupervised clustering algorithms and finally, these fast classification algorithms classified the whole data set in such a way that efficient and accurate profiling of very large and highly dimensional medical data sets can be achieved.

REFERENCES:

1. Consensus Function Based on Clusters Clustering and Iterative Fusion of Base Clusters,Musa Mojarad,Hamid Parvin,Samad Nejatian, and Vahideh Rezaie

2. Using Closed Patterns to Solve the consensus, Clustering Problem,Atheer Al-NajdiNicolas Pasquier and Frédéric Precioso

3. emporal Data Clustering via a Weighted Clustering Ensemble With Different Representations YunYang