Mining Interaction Patterns by Clustering

Madhumathi D[1] | Meghana S[2] | Rashmi P[3] | Sanjana.R[4] |
Mrs.L.Babitha[5] | R Dinesh Kumar[6]

[(1,2,3,4] UG Scholars, ,[5,6]Professor,Department of Computer Science and Engineering, KTVR Knowledge Park for Engineering and Technology, Coimbatore, India, me.dineshkumar@gmail.com[6])

***Abstract*— Human brain is complex and its difficult to understand its activities. fMRI(Functional Magnetic Resonance Imaging) provides the details to study about the brain functions fMRI process requires the effective and efficient data mining techniques to obtain the data. To collect the complex interaction patterns among brain a special type of clustering technique is required. The objects with the similar interaction pattern are to be assigned in the similar cluster. Based on this process an Interaction K-Means clustering algorithm is proposed to perform the clustering of objects in the training set of data. By the proceeding cleavage the percepts are obtained. The psychiatric disorders can be effectively recognized by means of the Artificial neural networks by which the results can be obtained effectually.*

***Keywords*— fMRI; Data Mining; Cluster; Neural Networks**

1 Introduction:

Human brain is difficult to predict and understand. Psychiatric disorders like Schizophernia and Somatoform pain disorder can be identified by the abnormalities in the brain by the fMRI (fig 1.7). fMRI relay upon the blood oxygen level dependant effect which allows the process of imaging the activities of the brain by the changes in the blood flow which is related to the energy consumption of the cells in the brain. fMRI data are the time series of the three dimensional volume images of the brain. The data is analyzed with the help of the statistical process. A statistical analysis involves in the comparing the data in groups. Data from fMRI experiments are massive in volume. Since the data represent the complex activity of the data represent the complex activity of the brain, the information is also expected to be tightly complex. To access the information more efficiently, the multivariate data mining methods are required.

The most recent findings suggest that the brain can be splitted up into different functional modules. In order to obtain the better functional complex activities of the brain it is essential to understand the interplay between the brains regions during the process of performing the tasks and the rest. So a dNovel Algorithm is proposed to obtain the healthy and diseased areas in the brain region by the process of clustering. To be more specific cluster means that the group of similar interaction pattern among the regions of the brain.

2 Interaction K-Means Clustering:

The algorithm Interaction K-Means Clustering (IKM) which minimizes the error rates.IKM is an iterative algorithm which effectually converges towards local minima of the optimization space.

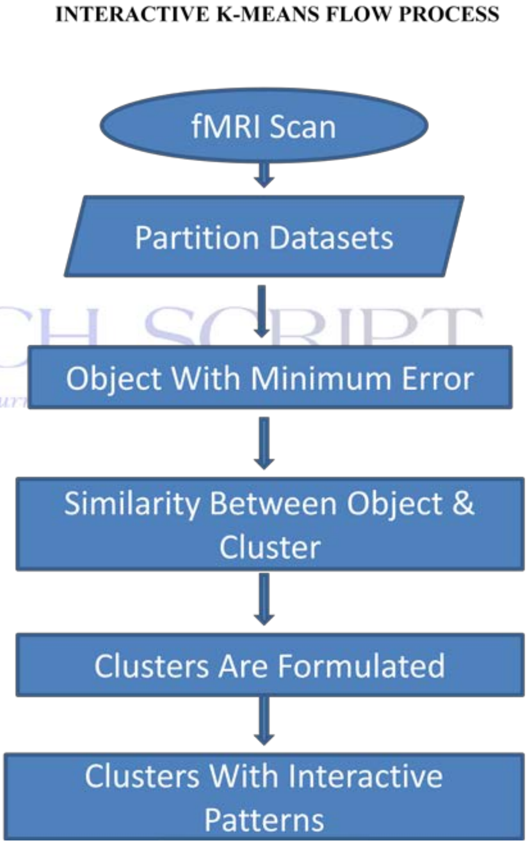
INTERACTIVE K-MEANS FLOW PROCESS
fMRI Scan
Partition Datasets
Object With Minimum Error
Similarity Between Object & Cluster
Clusters Are Formulated
Clusters With Interactive Patterns

Figure1.1-depicts the flow process of K-Means

2.1 Algorithm-IKM:

Interactive K-Means Algorithm clusters the data and obtains the interaction patterns of the brain.

This algorithm is not limited to fMRI data rather used in many applications. It finds the cluster of objects which are represented by means of multivariate time series which shares a common cluster interaction pattern.

Dependencies among the dimensions are considered as non redundant information. This redundant information should be preprocessed before the clustering is done.

The clusters which are obtained are composed of the objects which exhibit similar dependencies which can therefore be explicated as cluster specific interaction process. IKM is an iterative algorithm which proceeds with the following steps (fig1.1).

2.1.1 Initialization:

The IKM process is iteratively made to run several times to obtain the best overall result. This process to be proceeded by the data set is partitioned into k clusters. In this process of partitioning the clusters should be balanced in size. In order to avoid the overcome of over fitting. It repeatedly performs the following two actions iteratively.

(a)Assignment:

The object O in the cluster is assigned to the cluster is assigned to the clusters with the minimum error. In this process the total sum of errors is minimized in all the K clusters formed.

The clusters which have the longer time series tends to have tedious error rate

(b)Update:

The clusters are reformulated because of the rearrangement of the data sets. In the IKM Algorithm the similarity is measured between the objects and the clusters. In contrast to k-means and k-medoids here object representatives are not choosed.

As an iterative partitioning clustering algorithm IKM follows the same process which is being done by K-Means. The similarity measure which is used in the process of IKM is evaluated between an object and a cluster and it is not evaluated between the two objects In contrast to the K-Means and K- Medoid algorithms the data object is not choosed as a representative of the cluster. The cluster representation obtained from the IKM describes the interaction among the dimension

3 Dynamic Region Merging:

DRM (Dynamic Region Merging) algorithm is started from a set of over-segmented regions. It is because a small region can use more stable information than a single pixel and by using regions for merging can improve the computational efficiency. Here in the brain if the region gets varied it can be easily found by using the Pruning Algorithm. It is used to control the number of change points. It leads to smooth out the discrete boundaries. For simplicity and in order to validate the effectiveness the initially over-segmented regions using a more sophisticated initial segmentation lead better final segmentation results. When using segmented image, there are many regions to be merged for an meaningful segmentation. By choosing the region merging as a labeling problem, the purpose is to assign each region a label such that regions belong to the same object will have the same label.

The test of consistency or inconsistency depends on the error probabilities. The spatial relationship between the pixels has been represented by linear combination of a few training samples. If a region contains a small part of non-homogenous data, it makes a few more times of tests to verify its decision. With its equitable small error probabilities, the segmentation results will be more exact.

According to the observation, in most cases, the segmentation result is stable for a given image. By using the process of merging, the label of each region is sequentially transited from the initial stage to the final stage, which signifies as a sequence .If the region is merged with its nearest location, they will be assigned to the same name of label .To find an optimal sequence of merges which produce a union of optimal labeling for all regions, the minimization of a certain objective function is required.

4 Associate Rule Mining:

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is proposed to identify strong rules discovered in databases using different measures of interestingness.

The goal of classification is to build a model of the training data that can correctly predict the class of test set of data objects. The data given for this process is a set of objects along with their classes (supervised training set data). Once a forecast is made, it can be utilized to predict the class of the objects of test set cases for which class is not known. To measure the precision of the model, the accessible dataset is separated into training and test sets. The training set is used to build the model and test set is used to measure its accuracy. There are several problems from a wide range of domains which can be cast into classification problems.

Apriori flow process

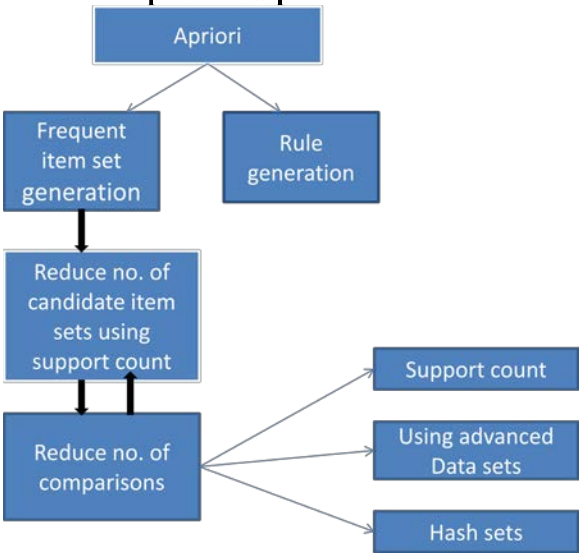
Apriori
Frequent item set generation
Rule generation
Reduce no. of candidate item sets using support count
Reduce no. of comparisons
Support count
Using advanced Data sets
Hash sets

Figure1.2 Apriori process

Association rule mining disclose the inherent relationships among the data attributes. The classical model of association rule mining employs the support measure, which treats each transaction identically. In contrast, diverse transactions have different weights in Real-life data sets. Class based rule mining is a special kind of ARM in which we are interested in finding class based association rules. In class based association rule the outcome of the rule is always a class label.

4.1 Rule strength measures:

It is calculated by using the formulas with the X and Y occurrence given below:

$$\text{Support} = \frac{\text{Count (XUY)}}{n}$$

$$\text{Confidence} = \frac{\text{Count (XUY)}}{\text{X. Count}}$$

The support and confidence outcome are used to find association rules (fig 1.2). The association rules that satisfy minimum support and confidence threshold should be specified by the user. General association rules mining approach can predict any attribute not just the class attribute and can predict the values of more than one attributes. It is the proportion of the dataset that is (correctly) covered by a set of rules.

There are different approaches developed for associative classification. An algorithm has three main steps. Rule discovery process extracts all rules from training set of data. These rules are called class association rules. Rule selection process picks the subset of rules from all obtained rules on the basis of their predictive precision to make a classifier. Confidence measure is used for selecting rules. Higher confidence rules habitually give higher predictive accuracy conclusively classification process classifies the unperceived data samples. An unperceived data sample is assigned the class of the rule that has highest confidence value and which also resembles with the data sample.

5 Classification:

ANN (Artificial Neural Network) is used as a classifier. Supervised learning classifier identifies the set of possible patterns in advance. Support vector machine (SVM) focus only on the difficult patterns that is to be clustered.SVM searches for the nearest pattern to be clustered. It uses input as images can give an accurate result compared to artificial neural network with hand designed features.

The Classifier maps the input data to a category. It generally presented as systems of interconnected neurons which can compute values from inputs and are capable of pattern recognition to its acceptability nature. The classifier produces the values which will be accurate when unerring to the training set of data. Based on the characteristics feature it detects whether the region is sway with the disease or not. An Artificial Neural Network (ANN) is an information processing pattern that is transitive by the way biological nervous systems, as the brain, process information. The key element of this methodology is the novel structure of the information processing system. It is produced of a large number of highly interconnected processing elements (neurons) working in unison to unravel specific problems as pattern recognition or data classification, through a learning process. Each neuron is bind to its neighbors with varying coefficients of connectivity that represent the strengths of these connections. All are composed of units (neurons) and connections between them, which together induce the behavior of the network. The choice of the network type depends on the problem to be solved.

6 Interpretation result of Clustering:

The major advantage is that the interaction patterns can be obtained. By the process of differentiation among the clusters, the objects are found out (fig 1.4). Then the errors in the cluster objects are identified and the total number of errors are summed up and calculated. The cluster ranking should be choosed and provided. If the cluster is with the positive sign then it indicates that the errors are minimal. If it is found with the negative sign then it indicates that the clusters are with the maximum number of errors.

Figure 1.3- overall system flow

7 Interaction among brain regions:

7.1 Functional Magnetic Resonance Imaging:

The goal of fMRI data analysis is to detect correlations between brain activation and task the subject performs during the scan. It also aims to detect interdependence with the specific intellective states, such as memory and recognition, induced in the subject. The BOLD signature of activation is proportionally weak; however, so further references of noise in the acquired data must be carefully restrained. This means that a sequence of proceeding steps must be performed on the acquired images before the actual statistical search for task-related activation can begin.fig (1.3) shows the project flow

The data sets obtained from the MRI experiments generate the series of 3-D volume images of the brain. Each image consists of about certain number of voxels and the interval points. Initially the standard preprocessing technique is applied including the process of realignment, normalization and smoothing. Here the approach is based on the set of time series. The brain atlas with the predefined mask of regions is used here. Physiologically relevance components and rejects the components with noise.

7.2 Schizophrenia:

Schizophrenia is identified by the impaired interaction among the distributed brain regions exactly the striatum. The treatment for schizophrenia and antidopaminergic is increased dopamin activity in the striatum. The Spontaneous influence between intrinsic brain networks and including striatum is aberrant in patients. Intrinsic brain networks are identified by Synchronous brain activity at rest. Intrinsic brain networks represented by Independent component analysis of fMRI data resulted in 9 ICs by spatial maps of brain activity and its related time series. The time series shows 26 multidimensional time series objects of health controls in patients. Non linear model of clustering reflects Granger casualty which separates patients from control with high cluster purity of 84.6% for the striatum model. Every cluster consists of only 13 persons in which only two persons have been incorrectly clustered. The changed Influence on the striatum was identified for many intrinsic brain networks, reflecting aberrant regulation for striatum activity.IKM achieves good results for synthetic data and real world data for various domains. It achieves excellent results for electroencephalogram (EEG) and fMRI. The Algorithm is scalable and accurate for noisy data. The interaction pattern identified by IKM are easy to implement and visualization. Different regions of time series should consider different model. It intend to work on methods for exact initialization of IKM. The existing strategies of K-mean cannot be straightly transferred to IKM due to special cluster notion. It also investigates feature selection for interaction based clustering.

7.3 Somatoform Pain Disorder:

Somatoform Pain Disorder has no medical explanation which has severe impact on quality of living of the affected person where the main symptom is severe and prolongment pain. The cause for the disorder cannot be fully understood but it is determined by altered mechanisms of observing and processing pain. The aim of the experiment is to cluster persons based on interaction patterns of ROI (Regions of Interest) inside the brain during experiment. Every person is identified by multivariate time series with 90 dimensions and 325 time points.IKM (Interactive K-Means) technique does not require multivariate time series with 90
length. The Naïve and ICACLUS use only 216 time points for clustering, which shows an information loss. The IKM algorithm is efficient to the result of all comparison methods. The first cluster consists of nine subjects including Somatoform Pain disorder and four healthy controls. The next cluster consist of nine healthy controls and four subjects including Somatoform Pain disorder. The accidental model are represented by color coding.

7.4 Independent component analysis:

Time series clustering is a dynamic field with applications in medicine, astronomy and economics. In spite of the prevalence of multivariate time series only a few algorithms have been especially designed for their clustering.

Independent component analysis (ICA) is a linear transformation method in which the goal is to find a linear representation of non-Gaussian data so that the resulting components are statistically independent, or at least as independent as possible. This technique originally gained popularity for applications in blind source separation (BSS).

8 Experimental Results:

The below figures are the results obtained from the overall system process

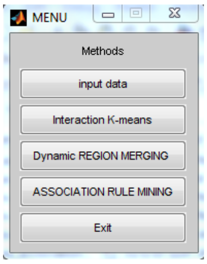

Figure-1.6Mining

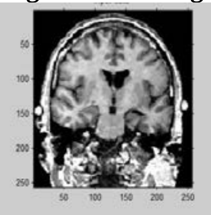

Figure- 1.7 scanned brain image (fMRI)

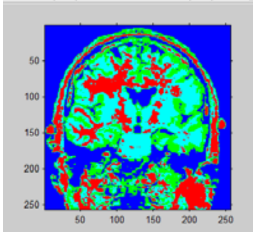

Figure 1.4- the k-means clustering

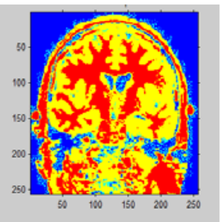

Figure1.5-the Dynamic region merging

9 Conclusion:

In this system the Interaction K-means was implemented to produce better results. The multivariate time series were introduced to detect the region of the brain in an efficient way. The interactions between the clusters were determined and studied to get the exact region that is defected. The clustering is done by splitting the brain region into various regions. These regions are ranked in a way with the defected region in the highest priority. Same way, the clustering is followed for the normal brain image too. On fMRI data the studies on Somatoform Pain Disorder and Schizophrenia, detects very interesting and meaningful interaction patterns. This technique originally gained popularity for applications in blind source separation (BSS), that is the process of extracting one or more obscure signals from noise.

Our experimental evaluation demonstrates that the interaction based cluster notion is a valuable complement to existing methods for clustering multivariate time series. IKM achieves better results on synthetic data and on real world data from various domains, but especially better results on EEG and fMRI data. Our algorithm is hardy against noise. The interaction patterns detected by IKM are easy to be visualized. Nonlinear models show their superiority in the corresponding real world data.

10 Future Work

In ongoing and future work, we plan to extend our ideas to differential equations. We want to deliberate different models for different regions of the time series. We intend to work on methods for suitable initialization of IKM, since existing strategies for K-means cannot be straightforwardly transferred to IKM because of the special cluster notion. We are also criticizing in feature selection for interaction-based clustering. The proposed system can be extended by adding a efficient way to calculate the interaction among the neurons of the brain by concentrating on time consumption in an effective way by using the segmentation among the regions and also by classifying the regions for analyzing the disease by concentrating on time consumption.

References:

[1] M. D. Fox and M. E. Raichle, “Spontaneous fluctuations in brain activity observed with functional magnetic resonance imagining,” Nat. Rev. Neurosci., vol. 8, no. 9. pp. 700-711, 2007.

[2] C.Sorg et al.,”Selective changes of resting-state networks in individuals at risk for alzheimer’s disease,”PNAS, vol. 104, no. 47, pp. 18760-18765, 2007.

[3] C. Li, L. Khan, and B.Prabahakaran,”Feature selection for classification of variable length multiattribute motions,” in Multimedia Data Mining and Knowledge Discovery, V.A Petrushin and L.Khan,EDS. London, U.K.: Springer, 2007.

[4] X. Ge and P. Smyth, “Deformable Markov model templates for time-series pattern matching,” in Proc. KDD, New York, NY, USA, 2000, pp.81-90.

[5] X. Wang, A. Wirth, and L. Wang, “Structure-based statistical features and multivariate time series clustering,” in Proc. ICDM, Omaha, NE, USA, 2007, pp.351-360.

[6] E.H.C. Wu and P.L.H.Yu,”Independent component analysis for clustering multivariate time series data,” in Proc. ADMA, Wuhan, china,2005, pp. 474-482.

[7] L. Owsley, L. Atlas, and G. Bernard, “Automatic clustering of vector time-series for manufacturing machine monitoring ,” in Proc. IEEE ICASSP, vol. 4. Munich, Germany,1997,pp. 3393-3396.

[8] X. Z. Wang and C. McGreevy, “Automatic classification for mining process operational data,” Ind. Eng. Chem. Res., vol. 37, no. 6, pp.2215-2222, 1998 [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ie970620h.

[9] F. Morchen, “Time series feature extraction for data mining using DWT and DFT,” Dept. Math. Computer. Sci., University of Marburg, Germany, Tech. Rep. 33, 2003.

[10] A. Sudjianto and G. S. Wasserman, “A nonlinear extension of principal component analysis for clustering and spatial differentiation,” IIE, vol.28, no.12, pp. 1023-1026,1996.

[11] A.Trouve and Y.Yu, “Unsupervised clustering trees by nonlinear principal component analysis,” Pattern Recognit. Image Anal., vol. 2, pp. 108-112, 2001.

[12] D. T. Larose, Data Mining Methods and Models. Hoboken, NJ, USA: Wiley, 2006.

[13] D. Arthur, B. Manthey, and H. Roglin, “Smoothed analysis of the k-means method,” J.ACM, vol. 58, no. 5,pp. 19:1-19:31, Oct. 2011[Online]. Available: http://doi.acm.org/10.1145/2027217.

[14] C. Sorg et al., “Increased intrinsic brain activity in the striatum reflects symptom dimensions in schizophrenia,” Schizophr Bill., vol.39, no. 2, pp. 413-421, NOV.2007.

[15] M. Halkidi, Y.Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” J. Intell. Inf. Syst., vol. 17, no. 2-3, pp. 107-145, 2001.

[16] B. Scholkopf, A. J. Smola, and K.-R. Muller, “Nonlinear component analysis as a kernel eigenvalue problem,” Neural Comput., vol. 10, no.5, pp. 1299-1319, 1998.