DATA CONFIDENTIALITY ON SEMI SUPERVISED CLUSTERING

Ms.D.Ranjani[1], *Santhiya.G* [2], *Swetha Narayanan* [3], *Vibisha.V* [4]
Assistant Professor, Department of Information Technology, Sri Krishna College of Technology, Kovaipudur, Coimbatore, India[1]
U.G Scholar, Department of Information Technology, Sri Krishna College of Technology, Kovaipudur, Coimbatore, India[2,3]

ABSTRACT –

Traditional cluster ensemble approaches have three limitations:(1) They have not make use of prior knowledge of the datasets given by experts.(2) Most of the conventional cluster ensemble methods cannot obtain satisfactory results while handling high dimensional data.(3) All the ensemble members are considered, even the ones without positive contributions. The random subspace technique is effective in handling high dimensional data, while the constraint propagation approach is useful for incorporating prior knowledge. In this projects to compare the gene details. First include our normal data. Then include our diabetic patient data in our dataset. After that compare our data and to show result is positive or negative. The incremental ensemble member selection process is newly designed to judiciously removed redundant ensemble members based on a newly proposed local cost function and a global cost function, Finally, a set of nonparametric tests are adopted to compares multiple semi-supervised clustering ensembled approaches over different datasets.

Index terms- handling high dimensional data, incremental ensemble member selection

I INTRODUCTION

Data mining is the process of analyzing a data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of the number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarizes the relationships identified. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Data mining can unintentionally be misused, and can then produces results which appear to be significant; but which do not actually predict future behaviour and cannot be reproduced on a new sample of data and bear little use. Often the results from investigating too many hypotheses and not performing proper statistical hypothesis testing. A simple version of this problem in machine learning is known as over fitting, but the same problem can be arise at different phases of the process and thus a train/test split - when applicable at all - may not be sufficient to prevent from happening. This section is missing information about non-classification of tasks in data mining. It only covers machine learning .

II RELATED WORK

Data Confidentiality on semi Supervised Clustering is to compare the diabetic details. First include our normal data. Then include our diabetic patient data in our dataset. After that compare our data and to show result is positive or negative. The incremental ensemble member selection process is newly designed to judiciously remove redundant ensemble members based on a newly proposed local cost function and a global cost function, Finally, a set of nonparametric tests are adopted to compare multiple semi-supervised clustering ensemble approaches over

different datasets. The four modules are doing the entire process of the diabetic patients in future.

1.Cluster ensembles: Models of consensus and weak partitions:

Clustering ensembles have emerged as a powerful methods for improving both the robustness as well as the stability of unsupervised classification solutions. Finding a consensus clustering from multiple partitions is a difficult problem that can be approached from graph based, combinatorial, or statistical perspectives. The study extends previous research on clustering ensembles in several respects. we introduce a unified representation for multiple clustering and formulate the corresponding categorical clustering problem. we propose a probabilistic model of consensus using a finite mixture of multinomial distributions in a space of clustering.

Issues:

A combined partition is found as a solution to the corresponding maximum-likelihood problem using EM algorithm. we define a new consensus function that is related to the classical infraclass variance criterion using the generalized mutual information definition. we demonstrate the efficacy of combining partitions generated by weak clustering algorithms that use data projections and random data splits. A simple explanatory model is offered for the behaviour of combinations of such weak clustering components.

2. The random subspace method for constructing decision:

Much of previous attention on decision trees focus on the splitting criteria and optimization of tree sizes. The dilemma over fitting and achieve maximum accuracy is seldom resolved.

Issues:

The classifier consists of multiple trees constructed systematically by pseudo random selection subsets of components of the feature vector.

3. Exhaustive and efficient constraint propagate:

This paper presents a novel pair wise constraint propagation approach by decomposing the challenging constraint propagation problem into a set of independent semi-supervised classification sub problems which can be solved in quadratic time using label propagation based on k-nearest neighbour graphs.

Issues:

The resulting exhaustive set of propagated pairwise constraints is further used to adjust the similarities matrix for constrained spectral clustering. Other than the traditional constraint propagation on single-source data, our approach is also extended to more challenging constraint propagation on multi-source data. This multi-source constraint propagation has an important application to cross-modal multimedia retrieval.

4. Normalized cuts and image segmentation:

We propose a novel approach for solving the perceptual grouping problem in vision. We treats image segmentation as a graph partitioning problem and propose a novel global criterion, the normalized cut, for segmenting the graph. The normalized cut criterions measure both the total dissimilarity between the different groups.

Issues:

Segmentation-based object categorization can be viewed as a specific case of spectral clustering applied to image segmentation. We show that an efficient computational technique based on a generalized eigens problem can be used to optimize this criterion. We applied this approach to segmenting

static images and found the results to be very encouraging.

5. Constraint neighbourhood projections for semi-supervised clustering:

Semi-supervised clustering aims to incorporate the known prior knowledge into the clustering algorithm. Pair wise constraints projections are two popular techniques in semi-supervised clustering. they consider the given constraints and do not consider the neighbours around the data points constrained by the constraints.

Issues:

This paper presents a new technique by utilizing the constrained pairwise data points and their neighbours, denoted as constraint neighbourhood projections that requires fewer labelled data points (constraints) and can naturally the constraint neighbours are chosen according to the pairwise constraints and the original data points are projected into a new low-dimensional learned from the pairwise constraints and their neighbours.

III PROBLEM FORMULATION

SYSTEM ARCHITECTURE

The modules are implemented in this technique

1. Add patient's file.
2. Attribute selection.
3. Comparison.
4. Result.

1.Add patient's Details

In this module the patient medical details are added using Weka tool. The uploaded patient detail is uploaded in arff format. The uploaded patient file is then used for further process .After the patient details detected it will get started to predict the attribute for the relevant information of patient with future diseases. The file database store the details about the user at the first stage of the diseases. This data can be retrieve for the attribute selection process. The patient details are more important to analyze the further process, it is considered as the basic process for remaining process because in this module only the file created for the patient .

2.Attribute Selection

The patient's diseases are viewed in this module. The list of diseases affected to the particular patient is viewed in detail. This attribution will be help in the stage of the analysing the patient completely. The patient's attributes are related to the some diseases relevant information which gives the details to compare the stage of the patient diseases..the attribute database can store the relevant information about user depending upon the patient's file. This attribute can identify the user from the file database.

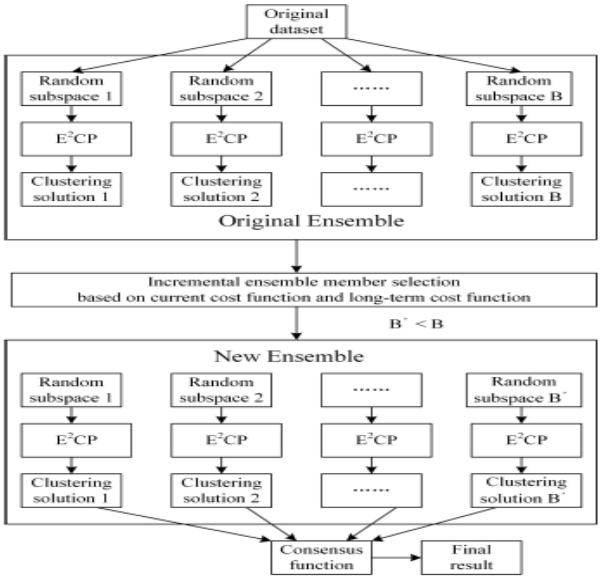
3.Comparison

The Patient report is compared with other patient's in order to predict the possibilities of occurrence other diseases. This makes the patient to be aware of other diseases.The prediction analysed by the attributes and patient's module with the help of the high dimensional data for producing the satisfactory results.the result of the every stages of the human will bemonitored and noted into file database by selecting the attribute database after this process completed the details are stored into the comparison database.it consider as the important module of the customer because each time the user

investigated by this process.this process will help to provide the basic report for the customer

4.Result

The entire history of the patient is viewed in this module. The patient report is viewed in graph format. This is final process for analyzing the patient's health stage and can be provided the awareness like what the patient should have to follow the instructions of the Experts and this illustration will be help to research the further enhancement.the result database can store the positive result of the user and also negative result of the user



DESIGN GOALS

- Software testing is required to point the defects and errors that were made during the development
- It's essential since it makes sure of the Customer's reliability and their satisfaction in the application.
- It is important to ensure the Quality of the product. Quality product delivered to the customers helps in gaining their confidence.
- Testing is necessary in order to provide the facilities to the customers like the deliver of high quality product or software application which requires lower maintenance cost and hence results into more accurate, consistent and reliable results.
- Testing is required for effective performance of the software application or product.
- It's important to ensure that the applications should not result into any fail, it can be very expensive in the future or in the later stages of the development.
- It requires to stay in the business.

CLUSTERING METHOD

Semi-Supervised Clustering

Semi-supervised learning is a class of supervised tasks and techniques that also make use of unlabeled data for training – typically a small amount of labelLed data with a large amount of unlabeled data. Many machine learning researcher have found that unlabeled data, when used in conjunction with a small amount of labelled data. The acquisition of labelled data for a problem often requires a skilled human agent or a physical experiment.

The cost is associated with the labelling process thus may render a fully labelled training set infeasible, whereas acquisition of unlabeled data is relatively less .Semisupervised learning can be of great practical value. Semi-supervised learning is also

of theoretical interest in machine learning and as a model for human learning. Semi-supervised learning attempts to make use of this combined information to surpass the classification performance that could be obtained either by discarding the unlabeled data and doing supervised learning. The goal of learning is to infer the correct labels for the given unlabeled data . The goal of learning is to infer the correct mapping from to intuitively, we can think of the learning problem as an exam and labelled data as the few example problems that the teacher solved in class. It is unnecessary function to perform learning by way of inferring a classification rule over the entire input space; algorithms formally designed for transduction or induction are often used interchangeably.

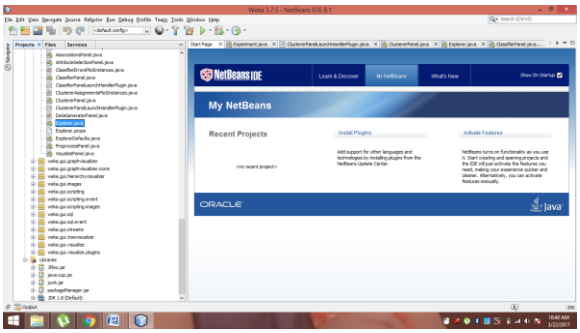


Fig 1. Weka tool

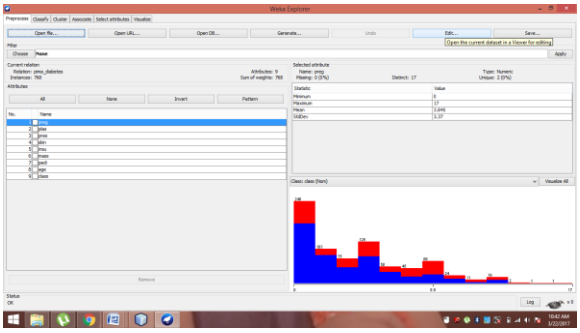


Fig 2. File Upload Page

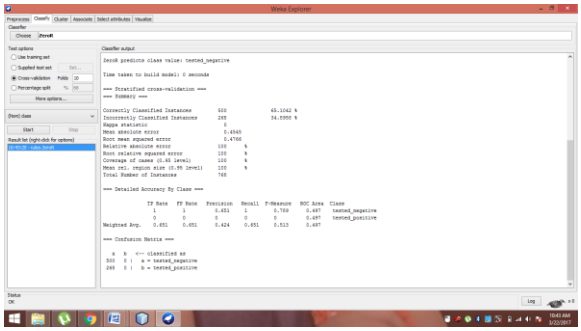


Fig 3.classification

V SYSTEM DESIGN AND IMPLEMENTATION

The executable order of the tests is called a test procedure and is also known as a test script. The test Procedure Specification is prepared then it is implemented and is called Test implementation. Test scripts is also used to describe the instructions to a test execution tool. An automation script is written in a programming language that the tool can understand. The tests that are intended to be run rather than using a test execution tool can be called as manual test script. The test procedures, or test scripts are then formed into a test execution schedule that specifies which procedures are to be run kind of superscript. Writing a test procedure is another opportunity to prioritize the tests, to ensure that the best testing is done in the time available. A good rule of thumb is 'Find the scary stuff first'.

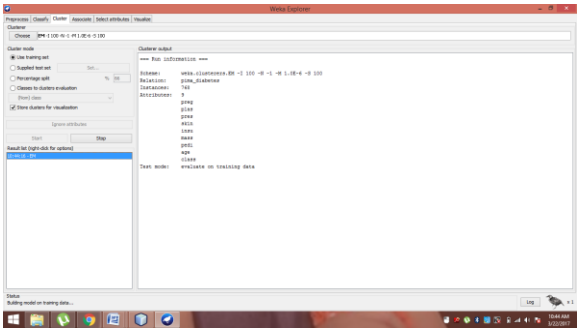


Fig 3. Cluster page

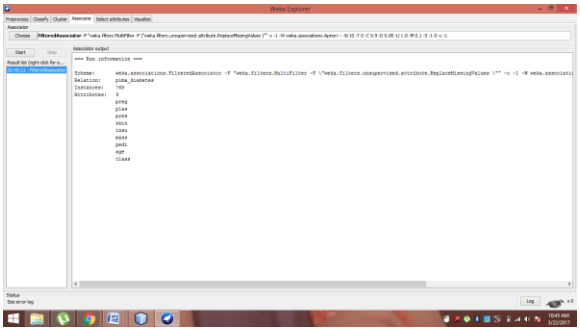


Fig 4. Association page

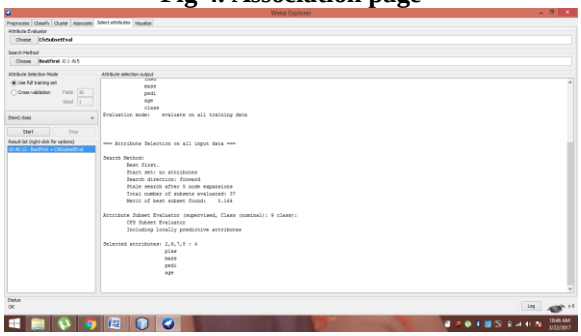


Fig 5.attribute selection

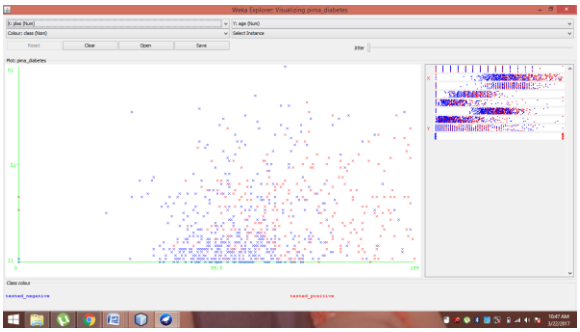


Fig 6. Visualizing

VI CONCLUSION

In this system our research focus is on semi-supervised clustering, which uses a small amount of supervised data in the form of class labels or pairwise constraints on some examples to aid unsupervised clustering. Semi-supervised clustering can be either search-based. Our main goal in the proposed thesis is to study search-based semi-supervised clustering algorithms and apply them to different domains.

We have shown how supervision can be provided to clustering in the form of labelled data points or pairwise constraints. We have also developed an active learning frame work for selecting informative constraints in the pairwise constrained semi-supervised clustering model, and proposed a method for unifying search- based and similarity-based techniques in semi-supervised clustering. Some of the issues we want to investigate include: effect of noisy , probabilistic or incomplete supervision in clustering; model selection techniques for automatic selection, ensemble semi-supervised clustering

VII REFERENCES

[1] A. P. Topchy, A. K. Jain, W. F. Punch, "Cluster ensembles: Models of consensus and weak partitions", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1866-1881, Dec. 2005.

[2] T. K. Ho, "The random subspace method for constructing decision ", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832-844, Aug. 1998.

[3] Z. Lu, Y. Peng, "Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications", *Int. J. Comput. Vis.*, vol. 103, no. 3, pp. 306-325, 2013.

[4] J. Shi, J. Malik, "Normalized cuts and image segmentation", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888-905, Aug. 2000.

[5] H. Wang, T. Li, T. Li, Y. Yang, "Constraint neighborhood projections for semi-supervised clustering", *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 636-643, May 2014.

[6] N.X. Vinh, J. Epps, J. Bailey, "Information theoretic measures for clusterings comparison: Variants properties normalization and correction for chance", *J. Mach. Learn. Res.*, vol. 11, pp. 2837-2854, 2010.

[7] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, A. Schliep, "Clustering cancer gene expression data: A comparative study", *BMC Bioinformatics*, vol. 9, no. 497, 2008.

[8] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, F.-S. Gou, "Semi-supervised linear discriminant clustering", *IEEE Trans. Cybern.*, vol. 44,

no. 7, pp. 989-1000, Jul. 2014.

[9] "UCI machine learning repository [http://archive.ics.uci.edu/ml]".

[10] D. Greene, P. Cunningham, "Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering", *Proc. Conf. Mach. Learn.*, pp. 140-151, 2007.

[11] "An extension on ”Statistical comparisons of classifiers over multiple data sets” for all pairwise comparisons", *J. Mach. Learn. Res.*.

[12] E. Akbari, H.M. Dahlan, R. Ibrahim, H. Alizadeh, "Hierarchical cluster ensemble selection", *Eng. Appl. Artif. Intell.*, vol. 39, pp. 146-156, 2015.

[13] N. Iam-On, T. Boongoen, S. Garrett, C. Price, "A link-based approach to the cluster ensemble problem", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2396-2409, Dec. 2011.

[14] H. G. Ayad, M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 16-173, Jan. 2008.

[15] A. L. N. Fred, A. K. Jain, "Combine multiple clusterings using evidence accumulation", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835-850, Jun. 2005.