A Novel Ensemble based Cluster Analysis using Similarity Matrices and Clustering Algorithm (SMCA)

Mayank Gupta
(M.Tech Scholar) Department of Computer Science & Engineering
Oriental University, Indore (M.P), India

Dhanraj Verma
(Associate. Professor) Department of Computer Science & Engineering, Oriental University, Indore (M.P), India

ABSTRACT

In today's world data analytics is gaining popularity due to user's motivation towards online data storage. This storage is not organized because of content types and data handling schemes complexity. User aims to retrieve data in lesser time with logical outcomes as desired can be achieved by applying data mining. Clustering in data mining is one of the known categorization approach used for formation of groups of similar elements having certain properties in common with other elements. This formation sometime creates noisy result in terms of formatted clusters. It depends on various factors such as distance measures, proximity values, objective functions, categorical or numerical attribute types etc. Over the last few years various schemes are suggested by different authors for improving the performance of tradition clustering algorithms. Among them, one is ensemble based clustering. Ensemble uses the mechanism for criteria selection from newly formed clusters with a defined portioning and joining methods to generate a single result instead of multiple solutions. The generation results are affected by various environmental parameters such as number of cluster, partitioning types, proximity values, objective function etc. This paper propose a novel SMCA based ensemble clustering algorithm for improvements over the existing issues defined in the paper. At the primary level of work and analytical evaluations, it shows the promising results in near future.

Keywords

Data Mining, Clustering, Ensemble, Consensus, Partitioning and Joining Criteria,, Proximity Value, Similarity Metrics and Clustering Algorithm (SMCA);

1. INTRODUCTION

In today's world the intelligence is provided to the business depending upon the extraction of data and facts from historical records. Dependencies are created on such processes which derives certain logic towards futuristic probabilities of success. To achieve this goal in forecasted manner mining is required. Data mining techniques is used for organizing the data into logical containers of classes with fixed and similar termed as clustering. Let us take a real time records using databases. Here databases are a collection of multiple records representing values in its attributes with different categories. Data clustering is the process of organizing and grouping the related data into segments called clusters. These grouping are performing on the basis of similarity between those attributable elements and properties. These are mainly used to investigate the effective analysis on data to guide futuristic operations of the organizations.

Clustering not only provides information organization in large amount but also puts them in different representations as required. It can be applied for both structured and unstructured data. In unstructured data the fixed structure of information organization is not followed but for structured data the semantic relationship holds the similarity between the objects of the same class. Some of techniques recently hold the most favorable execution of clustering includes ensembles generations, semi-supervised, multi-way and for heterogeneous data. These approaches are commonly known as cluster analysis. In marketing it is the keen interest of the managers to identify the similarity between those objects, classes or the data elements.

Clustering processes seems to provide effective categorization of data but not guarantees the clear separation and clarity of partitioned values. With existing clustering methods such as k-means, results are always depends on the initial values of the seeds elements. If the choice of cluster seed is poor and complex the categorization result are also generated as a weak partitions. Other approaches such as hierarchical methods are sensitive to outliers and distance criteria's which later on affects actual results. Thus there are some problems associated with existing normal clustering methods are given as:

The identification of distance measure: For algebraic attributes, distance procedures can be used. But recognition of measure for categorical attributes is complicated.

The number of clusters: Identifying the amount of clusters is a complicated task if the number of class labels is not known previously. A cautious analysis of number of clusters is essential to produce correct outcome.

Structure of database: Real life data may not always contain noticeably identifiable clusters. Also the order in which the tuples are prearranged may affect the results when an algorithm is executed if the distance measure used is not ideal. With a construction less data (for e.g. having lots of missing values), identification of appropriate number of clusters will not yield good results.

Types of attributes in a database: The databases may not necessarily contain distinctively numerical or categorical attributes. They may also contain other types like nominal, ordinal, binary etc. So these attributes have to be converted to categorical type to make calculations simple.

Classification of Clustering Algorithm: Clustering algorithms can be classified according to the method adopted to define the individual clusters.

Ensemble approaches for cluster formation measures multiple solutions and instead of applying any one solution a consensus combination based solution is applied. The final solution of these consensuses is different than all the elements of previously generated ensemble. Thus ensemble base solution is having diverse nature in terms of approach used of clustering (Hierarchical, k-means, neural network, fuzzifications etc) and different generation variables. It is made possible by primary nature of ensemble generation which does not looks over the initial data but examines only the individual elements of intermediate cluster generations. The consensus solution combines information from those several partitioning to find one which is most representative of them all. This paper proposes a novel ensemble based clustering method to improve the data clustering and resolves the above mentioned problems more frequently than any other method.

2. BACKGROUND

Data Mining is a process to extract the logical results from the previously measured data with intelligence for extracting the records in real time. As of now, mining shows a drastic change over the technologies working to resolves its issues. Many new things and concepts are coming frequently in the market to provide more stability and performance over the mined results based on applications. Clustering is one of those approaches used to categorize the data into separate clusters having characteristics and feature as a partition criteria. Among the different types of clustering mechanism, ensembles based techniques are promising the result stability, performance and will resolve the classical clustering issues. Ensemble based clustering is provided the solution of situation where number of clustering solutions can be formed and to identify which one is correct again a unified solution of combining all the results is required. This cumulative clustering is known as ensemble clustering with a common solution of all the clustered inputs called as consensus which is better than existing clustering methodologies.

Most of the existing ensemble based clustering methods are used to partition the data with a standard solution of consensus. As of now variable methods for cluster analysis is used with ensembles such as hierarchical clustering.

It requires components for clear boundary based partitions such as connectivity metrics, vectors or indicators and typically represented using dendograms. But as making the particulars for clustering methods designing of objective functions generates the complexity in normal programs and taken as a critical and challenging task. It analyses the similarity between the several data element and place them to some common group with common attributes defined by the above objective functions. These similarities are accessed using the similarity functions by its distance measures. Larger is the distance more different characteristics will be in between the data element. Such approaches are used for statistical data analysis, machine learning's, bioinformatics and image recognitions.

2.1 Understanding Consensus

It has evolved as an important solution for critical clustering problems through a unified cumulative solution. Here the common characteristics are used for generations of solution as a aggregation process of previously performed clustering (or Partitions). Consensus is defined is the situation where there are number of different (input) clustering have been obtained for a particular dataset and is desired to find a single solution known as which is far improved than existing clustering outputs. It is taken as the solution of the problem of adaptive clustering information about the similar data set parsed from heterogeneous sources or from different runs of the same algorithmic approaches. When cast as an optimization problem, consensus clustering is known as median partition, and has been shown to be NP-complete. Consensus clustering for unsupervised learning is analogous to ensemble learning in supervised learning. This ensemble based clustering is further divided into two categories: Hard Ensemble Clustering and Soft Ensemble Clustering.

Hard Ensemble Clustering: In this, multiple partitioned data clusters are grouped into a single comprehensive cluster without accessing the basic features of the clustering algorithms used for determination of these clusters with a common objective function. The aim is to identify the high quality based objective functions with low computational cost. It is applied by designing various consensus functions defined particularly for hard generation of ensembles. These are Cluster-based similarity partitioning algorithm (CSPA), Hyper-graph partitioning algorithm (HGPA) and Meta-clustering algorithm (MCLA).

Soft Ensemble Clustering: In this method, soft ensembles are represented using concatenation prior probability operator calculated by probability distribution of the traditional clustering algorithm. It is used to identify the distance measure between the two instances of partitioned data which measures the similarity and distance. It uses KL (Kullback-Liebler) divergence for distance analysis using probability distribution model. As similar to hard ensemble clustering, the soft approaches are applied by designing a separate consensus functions given as: sCSPA, sMCLA and sHBGF. These functions are the extended versions of the functions defined with hard clustering.

Thus by using the effective combinations and selection of some of the existing objective function in collaboration with new operators the cluster formation and analysis is improved. This paper focuses on such aspects and suggested some improvements over the traditional ensemble generation mechanism for better clustering partitioned results.

3. LITERATURE SURVEY

Over the last few years various approaches related to the multidimensional clustering had been proposed by the researchers. Most of them focus on specific objective function generation and a cumulative solution generation using ensemble based hierarchical clustering. The approaches are motivated to generate the single set of clusters from the numbers of clusters based on similarity metrics. The suggested methods provide multiple run results on single instances of combining all the results with different datasets. Some of them is given below as surveyed study.

Hierarchical clustering based on ensemble generation by combining the partitioned results with a single solution is proposes in the paper [4]. The work designs a new objective functioned named as ultra metric distance for providing the data based learning's with clearly separated results in terms of clusters. The paper uses dendograms as a representation criterion for ensemble and consensus formation. The approach uses various definition criteria such cophentic difference, maximum edge distance, partition membership divergence, cluster membership divergence and sub-dendograms divergence. At the root level of analysis the approach is proving its effectiveness.

In the paper [5], the author addresses and outlined the hierarchical clustering combinations and introduces a novel algorithmic framework for combining the results into single entity. According to the suggested approach the similarity based data descriptive metrics of inputs are aggregated into transitive consensus matrix to for the final hierarchies. The aggregation is performed on the basis of some rules for combination with identified distance metrics. The evaluation of the approach is also provided with certain better results with cluster formations.

In the paper [6], application specific cluster formation is analysed on some prior datasets with improved ensemble generation. The approach uses a R package clue for extensible computational environments and ensemble generations and works as a data structure for hierarchical generations. Such ensembles can be obtained, for example, by varying the hyper parameters of a base clustering algorithm, by re-sampling or reweighting the set of objects, or by employing several different base clusterers. The representation is provided on dendograms with proximity calculation for minimization of difference.

Among the several hierarchical clustering approaches diversification of the results are also considered as a important factor for results improvements as given by [7]. It improves the quality of cluster by applying ensembles. It combines the diverse results of clusters obtained from the traditional clustering methods and merges the data. Evaluation of the suggested approach is also termed as two layer ensemble generation and provides more feasible diversification of results than any tradition LDA or k-means based approach.

Some other representation based approaches are suggested such as hyper graph partitioning [8], mutual model clustering [9], and co-association based functions clustering [10]. The approach provides a different ways of ensemble generation using incremental learning’s. The results are tracked using micro array datasets for individual algorithms. The above methods are also uses various internal cluster validation indices used for ensemble formation with hard generation consensus functions. [11].

These ensemble methods to define the dissimilarity measure through combining assignments of observations from a sequence of data partitions produced by multiple clustering [12]. Some of the authors had worked with dissimilarity measure which is applied as subjective and data dependent values. Experimental evaluation on gene expression data are used to illustrate the application of the ensemble method to discovering sample classes. Continuing the above method, paper [13] shows novel mechanism for improved analysis and cluster separations. The approach uses ensemble generation process with existing clustering methods such as k-means analysis for using hard functions such as CSPA and HGPA. The approaches are compared using modularity functions and data metrics using similarity associations.

In the paper [14], an application based ensembles generation for spam filtering is applied for removal of bulk spam emails. The approach uses a filtering method for previously analysed data for spam characteristic detection and proximity identification with actual and desired results. This automatic process is based on text retrieval schemes and tools for handling text documents is abstract vector formats for applying machine learning’s. The approach is also capable of defining various classifiers and detecting the most suitable one. Furthermore the purpose of the ensemble idea is defined and evaluated more precisely.

Continuing the above study it is identified that some of the author’s put their efforts towards parametric arrangements of data as suggested by [15]. A partitioned cluster generation based on dendograms generation is improved for reducing the error occurring due to cutting the edges of dendograms trees. Here the partitioned boundaries are clearly defined with fixed outlier values and giving the improved results. For the paper uses the classical document clustering problems which is not giving an optimal solutions. Here a comparison is made with the all document clustering approaches using ensemble based clustering’s.

4. PROBLEM STATEMENT

Cluster analysis is the identification of similarities between the elements of the cluster. The process requires selection of cluster head by which similarities are calculated. Similarity metrics gives the values in terms of its distance measures, so if distance measure is more dissimilarity exists and if the distance measure is less similarity is identified. Clustering can be applied by several methods such as hierarchical, k-Means etc. For generating the ensemble this partition criteria and the merging decision needs to be defined accurately so as to get clearly separated boundaries. The traditional clustering methods are unable to resolve such issues of effective ensemble generation and taken as major work area for this paper. The problems associated with them which remain unsolved are:

Problem 1: The identification of distance measure for categorical attributes is not properly defined which leads to incorrect placement of elements in different clusters.

Problem 2: Identification of accurate number of cluster is not stated with its proximity values and hence labelling of class based clusters are computationally complicated.

Problem 3: Merging decision using consensus function is not explicitly defined which creates incorrect merging and cause higher distance measure of similar elements.

Problem 4: Global objective function is not taken as a base parameters which leads cluster head selection more complex for less size categorical attributes of dataset.

Thus from the above identified issues in existing work the major problem area of work suggested in this paper is related to efficient ensemble generation using hierarchical clustering with some similarity association matrices for non parametric datasets. The boundaries are overlooked to measure clear separation and partition and merging are implicitly defined.

5. PROPOSED SMCA APPROACH

This paper proposes a novel SMCA (Similarity Metrics and Clustering Algorithms) method to resolve the existing issues in ensemble based cluster generation with improved cluster analysis and formation. The approach is used to combine the various objects into a single solution o ensemble using consensus designing without any access to the features of identifying these cluster partitions. Initially the data is analyzed for resultant knowledge through multiple solutions using ensembles. The cluster formation can be taken as optimization problems having NP hard computational complexity. The paper also proposes a novel consensus creating rules by which boundaries of the cluster are clearly separated from the other cluster. The distance measures are used for defining the similarity and dissimilarity between the objects of the same cluster and distant clusters. According to these defined metrics the partitioning and the cluster formation is reconstructed. The approach uses hierarchical clustering approach whose main objective is to from multiple clusters arranged from top to bottom according to their similarity of points. Multiple variants are used to visualize these clustered values and one of them is dendograms.

It is diagrammatical type for hierarchies' representation using tree based structures which describes the order of points where tow data elements or groups are merging together with some common characteristics. Here the proposed approach uses hierarchical clustering because of its some feature that it does not assume initially some fixed number of clusters. In process to that the hierarchies using dendograms are separated at any points for further increase in new clusters counts at proper levels for better quality results as a cluster.

5.1 Description

According to the suggested approach initially the data sets are selected from the sources to mine the results in the form of clusters. These data sets could be taken from the same data source or some heterogeneous records. For clustering the aim is to measure the similarities between the various elements of the common group. The data is passed to the mining evaluator in which the separate or same mining algorithms are executed on the data with different parameters. The algorithm can also be made in common with random initializations executed at multiple times. Aim is to get the clustered results. Here the approach is giving the selection of two basic clustering algorithms: K-means and Hierarchical. The outputs generated by these mechanisms are unstructured and are not labelled. This cluster data is not having clearly separated boundaries and even the objective function is not properly measured in this. Here the difference between the elements located at the boundaries of the two clusters is not identified.

Now the output of clustering algorithms is partitioned into multiple parts whose rules are explicitly defined by the clustering type. Here the numbers of clusters are not fixed as there where dependencies in existing approach. These generated partitions are termed as ensembles but their generation is not complete without any objective functions used for consolidated output. Thus each ensemble formation must require some combining algorithms which combine the multiple solutions into a single set.

But these combinations of partitions are as formal, and hence require some combination criteria. This work uses similarity metrics generated from median partition approach which is used to define similarity between of the element between the two clusters. The similarity association of two objects is a measure of how closely they are associated in the different hierarchical clustering. The paper also assumes that it is a measure of proximity of the two objects. Same process is also used for dissimilarity identification. From both these concepts the distance between the elements and their density near boundary line is identified.

Proximity (cluster 1, cluster 2) =∑proximity (p1, p2)/size (cluster 1)* size (cluster 2)

After identification of similarity metrics the similar clusters are grouped together with combining consensus criteria (Hard Function). Here the proximity is also calculated using proximity analysis function. It gives the errors percentage values or deviation of results in percentage. It could be also of inter cluster or intra cluster proximity. Now the generated output in the form of analysed clusters is labelled using predicted labelling method. Labelling improves the futuristic identification of elements is multiple clusters.

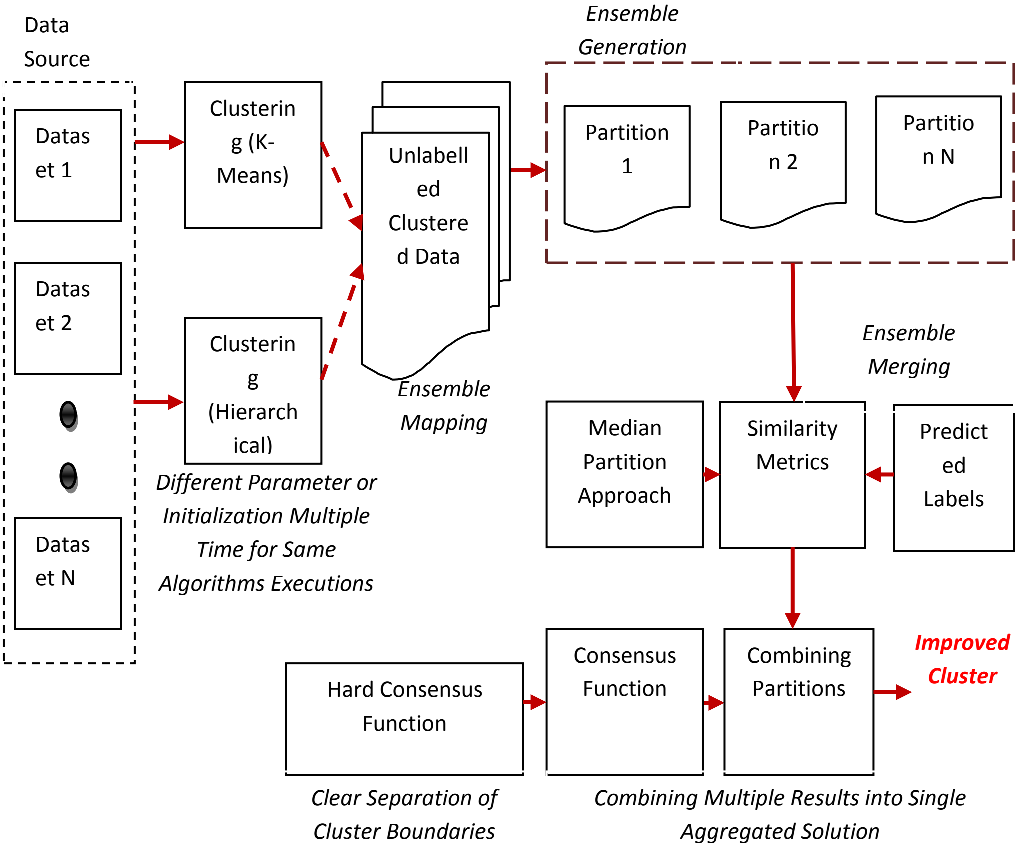
Data Source
Dataset 1
Dataset 2
Dataset N
Clustering (K-Means)
Clustering (Hierarchical)
Different Parameter or Initialization Multiple Time for Same Algorithms Executions
Unlabelled Clustered Data
Ensemble Mapping
Ensemble Generation
Partition 1
Partition 2
Partition N
Ensemble Merging
Median Partition Approach
Similarity Metrics
Predicted Labels
Hard Consensus Function
Consensus Function
Combining Partitions
Improved Cluster
Clear Separation of Cluster Boundaries
Combining Multiple Results into Single Aggregated Solution

Figure 1: A Novel Ensemble Based Cluster Analysis Using Similarity Matrices & Clustering Algorithm (SMCA)

5.2 Representation of Multiple Partitions

For representing the partitions using ensemble based approach dendograms are used having correct visualizations mapping and combinations characteristics of results. It gives a range of representations for the same data sets but the views are same & the partition logic is individually mentioned. It needs to classify the limits properly & place the element in accurate cluster. To categorize the correct element of all cluster strength of bond or dependencies is calculated which later shown in weak or strong manner. Assign those associations a value term as weights of vertex or dendograms.

5.3 Benefits of the Approach

Over the last few years various clustering techniques are evolve using different objective functions. As similar to that, here in ensemble generation the objective is provided by designing the consensus function which works as joining the partitions. For this specific reason the suggested work uses cluster ensemble generation using similarity metrics and clustering algorithms (SMCA). The aim is to increase predictive accuracy of cluster analysis. It seems that the work is capable of achieving its goal with above novel mechanisms. There are some identified areas where the clustering is benefitted from the proposed work given as:

Quality cluster formation with clearly separated boundaries and outliers

Effective and well formed objective function is defined using consensus

Proximity and size based merging a decision are evaluated for improved results

Robust clustering with implicit and explicit similarity association designing

Scalable and variable clustering choices for generating the ensemble partitions

Knowledge reuse and multi-view clustering can be applied

Distributed computation wit low cost operations is provided

5.4 Performance Metrics

For evaluating the suggested approach results over some of the existing mechanism some comparative parameters are selected based on their coverage values of internal structure working. The parameters and factors affecting the result are regularly maintained by assessing the actual values of these parameters. For proving the efficiency of suggested SMCA identified parameters are:

Number of Clusters: This factor is used for identification of actual clusters generated after applying the proposed approach. By this comparison can be easily made of the measures of quality of a cluster algorithm using external criterion.

Sum of Squared Error (SSE): For each point, the error is the distance to the nearest cluster and can be measured by identifying the cumulative error between them using SSE. Here x is a data point in cluster Ci and mi is the representative point for cluster Ci can show that mi corresponds to the center (mean) of the cluster. One easy way to reduce SSE is to increase K, the number of clusters a good clustering with smaller K can have a lower SSE than a poor clustering with higher K

SSE=∑i=1 to K ∑dist2 (mi, x)

Rand Measure: This index computes how similar the clusters (returned by the clustering algorithm) are to the benchmark classifications. One can also view the Rand index as a measure of the percentage of correct decisions made by the algorithm.

F-Measure (FM): It is used to compare cluster qualities. It quantifies how the clustering fits the actual classification of data where the most desirable value is 1 and the least desirable value is 0.It shows the result in the form of curve graphs syntactically processed values of clustering. In this each curve represents the F-measure for one document subsets.

6. CONCLUSION

Ensemble based clustering have evolved in multiple ways over the last few years and shows effective results over the existing approaches. These ensemble generations is the identification of similarity between the elements of the clusters and place them at near or far from the cluster head as according to their calculated distance measure. But sometimes these decisions are giving the accurate cluster formation due to noisy data and categorical attributes. It needs to be improved and proximity ranges have to be defined explicitly. Ensemble operation includes partitioning and merging decision at various points and given by the consensus function. Here the work identified that few parameters had made visible affects over the cluster formation and is not taken previously. Thus, this work proposes a novel MCA based ensemble cluster generation approach. Here the criteria of cluster formation are defined by which improvements over the analysed data are measured. In the near future more promising results are expected after the practical evaluations on various suggested parameters

7. REFERENCES

[1] Mahmood Hossain1, Susan M. Bridges2, Yong Wang2, and Julia E. Hodges2, "An Effective Ensemble Method for Hierarchical Clustering" in ACM at ISBN 978-1-4503-1084-0/12/06, 2012.

[2] Ayan Acharya, Eduardo R. Hruschka, Joydeep Ghosh & Sreangsu Acharyya, "Transfer Learning with Cluster Ensembles" in JMLR: Workshop and Conference Proceedings 27:123{133, 2012.

[3] Hanan G. Ayad and Mohamed S. Kamel, "Cluster-Based Cumulative Ensembles" in Pattern Analysis and Machine Intelligence Lab, 2009.

[4] Li Zheng & Tao Li, "Hierarchical Ensemble Clustering" in School of Computing and Information Sciences, 2010.

[5] Abdolreza Mirzaei & Mohammad Rahmati, "A Novel Hierarchical-Clustering-Combination Scheme Based on Fuzzy-Similarity Relations" in IEEE Transaction of fuzzy system, Vol 18, No 1, Feb 2010.

[6] Kurt Hornik , "A CLUE for CLUster Ensembles" in Journal of Statistical Software, Volume 14, Issue 12. September 2005.

[7] Dong Nguyen & Djoerd Hiemstra, "Ensemble Clustering for Result Diversification in Human Media Interaction" at http://mirex.sourceforge.net.

[8] Marcin Pelka, “Ensemble method for clustering of interval valued symbolic data” in Statistics in Transition Vol. 13, No. 2, pp. 335—342, June 2012.

[9] Prachi Joshi1 and Dr. Parag Kulkarni, “Incremental Learning: Areas and Methods –A Survey” in IJDKP Vol.2, No.5, September 2012.

[10] Harun Pirim, Dilip Gautam, Tanmay Bhowmik, Andy D. Perkins & Burak Ekşioglu, “ Performance of an ensemble clustering algorithm on biological data set” in Mathematical and Computational Applications, Vol. 16, No. 1, pp. 87-96, 2011.

[11] L.I. Kuncheva, S.T. Hadjitodorov & L.P. Todorova, “Experimental Comparison of Cluster Ensemble Methods” in CLBME.

[12] Proling Dechang Chen, Zhe Zhang, Zhenqiu Liu and Xiuzhen Cheng, “An Ensemble Method of Discovering Sample Classes Using Gene Expression” in Uniformed Services University of the Health Sciences, 2010.

[13] Bryan Orme & Rich Johnson, “Improving K-Means Cluster Analysis: Ensemble Analysis Instead of Highest Reproducibility Replicates” in Sawtooth Software, Inc Research Series, 2008.

[14] Robert Neumayer, “Clustering Based Ensemble Classification for Spam Filtering” in Vienna University of Technology, 2011.

[15] Edgar Gonz`alez & Jordi Turmo, “Non-Parametric Document Clustering by Ensemble Methods” in TALP Research Center ISSN 1135-5948, 2008.