ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

An Improved Semi-Supervised Clustering Algorithm Based on Active Learning

S.Shalini[1], R.Raja[2]

Student/M.E (CSE), Department of CSE, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Salem, Tamilnadu, India[1]

Assistant Professor, Department of CSE, Sri Shanmugha College of Engineering and Technology, Pullipalayam, Salem, Tamilnadu, India[2]

ABSTRACT-In semi supervised clustering is one of the major tasks and aims at grouping the data objects into meaningful classes (clusters) such that the similarity of objects within clusters is maximized and the similarity of objects between clusters is minimized. The dataset sometimes may be in mixed nature that is it may consist of both numeric and categorical type of data. Naturally these two types of data may differ in their characteristics. Due to the differences in their characteristics in order to group these types of mixed data it is better to use the ensemble clustering method which uses split and merge approach to solve this problem. In this paper the original mixed dataset is splitted into numeric dataset and categorical dataset and clustered using both traditional clustering algorithms (K-Means and K-Modes) and fuzzy clustering algorithms (Fuzzy C-Means and Fuzzy C-Modes). The resultant clusters are combined using ensemble clustering methods and evaluated by both f-measure and entropy measure. It is found that splitting is more beneficial and applying fuzzy clustering algorithms yields better results than traditional clustering algorithms.

KEY WORDS-Active learning; Clustering; Semi-supervised learning.

I. INTRODUCTION

In the process of solving the practical problems by using data mining, we often encounter some cannot be labeled data. If using artificial markers, it is too costly on the one hand and on the other hand it will cause unexpected damage easily. Therefore, how to use limited prior knowledge, which comes from the data related to the small category labels and constraint condition to complete clustering analysis has become a hot issue in recent years. At present, semi-supervised clustering algorithm can be divided:

The primary category is based on the constraint of semi supervised clustering algorithm, derived from the pair wise constraints proposed by Wag staff et all must-link and cannot-link. These algorithms are determined in dependence on the above two kinds of constraints, results of the two constraint are the opposite. Among them, must-link provided that two data samples in the space belong to the must-link constraint, then the two divisions as a class; on the contrary, cannot-link provided that two data samples in the space belong to the cannot-link constraint, the two data are divided into different classes.

The second category is based on the distance of the semi-supervised clustering algorithm. and using trained adaptive distance metric to evaluate, through movement of the sample produced different distances, which constructed the restriction conditions to meet clustering. In addition, two kinds of algorithm can also be combined to implement clustering, it is the so called third category.

These three semi-supervised clustering algorithms have several common problems: first, cluster deviation, using pair wise constraints in must-link and cannot-link study, sample points around a cluster center move now and then in order to obtain the best position, the distance of sample points in the algorithm iterations is changing. Note, must-link does not

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

guarantee that all the corresponding constraint sample point is divided into one class and also cannot-link constraint cannot guarantee that it can be classified into different categories, there exist certain errors; Second, supervisory information is usually non-active ways of obtaining in semi-supervised clustering, the collection of all possible supervisory information is obviously not feasible by traversing, therefore only under limited conditions can obtain some valuable information. Because of pair wise constraints semi-supervised clustering algorithm limitations, it is often too small that information embodied in the constraint set, then influence the overall effect of clustering..In addition, the sample space is high dimensional samples, and the spacing between sample points has smaller difference, the algorithm processing ability is also poor. So how to minimize the cost reduction is a research focus.

1.1 Semi-Supervised Clustering

Based on the above viewpoints, semi-supervised clustering research can be roughly divided into three directions: Based on the constraint mechanism, based on the distance and hybrid. Related research at present basically belong to the three class, which based on the pairwise constraints algorithm include: reference[4] is based on density clustering algorithm, can deal with any shapes of clusters, and based on the constraint set to split or merge clusters; reference[5] presented an effective semi-supervised clustering algorithm and introduced fuzzy constraint thought, with minimal supervision information clustering; reference[6] puts forward a kind of distinguishing nonlinear transformation metrics in measurement and based on image retrieval to test , its effect is good.

1.2 Active Learning Algorithm

Active learning algorithm is a branch of classification algorithm, because of the relatively wide research direction and application, domestic and foreign scholars have put forward many topics. Reference use source domain data to study the target domain with active learning algorithm, trying to simplify the sample point label complexity. In reference[8] Tomanek et al described the important application of active learning in the NLP (Natural Language Processing), focus on how to create high quality training sample set. Ambati et al analyzed word alignment model in machine translation system, which helps to reduce the data word alignment error rate by creating the half word alignment model combining unsupervised and supervised learning, and makes data concentration abnormal or makes noise sensitive.

II. RELATED WORK

Active learning has been studied extensively for supervised classification problems. In contrast, the research on active learning for constraint based clustering has been limited. As mentioned previously, most of the existing research studied the selection of a set of initial constraints prior to performing semi supervised clustering. .They proposed a two-phase approach, which we refer to as the Explore and Consolidate (E & C) approach. The first phase (Explore) incrementally selects points using the farthest-first traversal scheme and queries their relationship to identify c disjoint neighborhoods, where c is the total number of clusters. The second phase (Consolidate) iteratively expands the neighborhoods, where in each iteration it selects a random point outside any neighborhood and queries it against the existing neighborhoods until a must-link is found an improvement to Explore and Consolidate named Min-Max, which modifies the Consolidate phase by choosing the most uncertain point to query (as opposed to randomly). To select constraints by examining the spectral eigenvectors of the similarity matrix, which is unfortunately limited to two-cluster problems? In constraints are selected by analyzing the co-association matrix (obtained by applying cluster ensembles to the data). A key distinction of our method from the above mentioned work is that we iteratively select the next set of queries based on the current clustering assignment in order to improve the solution. This is analogous to supervised active learning where data points are selected iteratively based on the current classification model such that the model can be improved most efficiently.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

More relevant to our work is an active learning framework presented for the task of document clustering. Specifically, this framework takes an iterative approach that is similar to ours. In each iteration, their method performs semi-supervised clustering with the current set of constraints to produce a probabilistic clustering assignment. It then computes, for each pair of documents, the probability of them belonging to the same cluster and measures the associated uncertainty. To make a selection, it focuses on all unconstrained pairs that has exactly one document already "assigned to" one of the existing neighborhoods by the current constraint set, and among them identifies the most uncertain pair to query. If a "must-link" answer is returned, it stops and moves onto the next iteration. Otherwise, it will query the unassigned point against the existing neighborhoods until a "must-link" is returned. While Huang's method is developed specifically for document clustering, one could potentially apply the underlying active learning approach to handle other types of data by assuming appropriate probabilistic models. We would like to highlight a key distinction between Huang's method and our work, that is Huang's method makes the selection choice based on pair wise uncertainty whereas we focus on the uncertainty of a point in terms of which neighborhood it belongs to. This difference is subtle, but important.

Pairwise uncertainty captures only the relationship between the two points in the pair. Depending on the outcome of the query, we may need to go through a sequence of additional queries. Huang's method only considers the pair wise uncertainty of the first query, fails to measure the benefit of the ensuing queries. This is why our method instead focuses on point-based uncertainty, which measures the total amount of information gained by the full sequence of queries as a whole. Furthermore, our method also takes into account the expect number of queries to resolve the uncertainty of a point, which has not been considered previously.

Finally, we want to mention another line of work that uses active learning to facilitate clustering where the goal is to cluster a set of objects by actively querying the distances between one or more pairs of points. This is different from the focus of this paper, where we only request pairwise must-link and cannot-link constraints, and do not require the user to provide specific distance values.

The composition and cardinality of the sets M and C can significantly impact upon the improvements achieved by semi-supervised algorithms. In addition, as the number of data objects n increases, the number of possible constraints also significantly increases. If constraints are selected at random, a large number may be required before any noticeable improvement in clustering accuracy is achieved. To illustrate this, Figure 1 shows the effect of adding constraints for randomly chosen pairs on the normalized mutual information (NMI) [4] scores produced when the PCKM algorithm is applied to the 3-news-similar text dataset. Even after the addition of 1000 constraints, little significant increase in accuracy is evident. For many semi-supervised tasks, it will be the case that the oracle is a human expert. Since it is unrealistic to expect a human to respond to so many queries, an intelligent strategy for choosing constraints is desirable.

III. METHODOLOGY

The problem addressed in this paper is how to effectively choose pair wise queries in order to produce an accurate clustering assignment. Through active learning, we aim to achieve query efficiency, i.e., we would like to reduce the number of queries/questions asked in order to achieve a good clustering performance. We view this as an iterative process such that the decision for selecting queries should depend on what has been learned from all the previously formulated queries. In this section, we will introduce our proposed method. Below we will begin by providing a precise formulation of our active learning problem.

3.1 Problem Solution

To evaluate the proposed method on the eight benchmark datasets against a number of competing methods. The evaluation results indicate that our method achieves consistent and substantial improvements over its competitors. There are a number of interesting directions to extend our work. The iterative framework requires repeated re clustering of the data with an incrementally growing constraint set. This can be computationally demanding for large datasets. To address this problem, it would be interesting to consider an incremental semi-supervised clustering method that updates the existing

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

clustering solution based on the neighborhood assignment for the new point. An alternative way to lower the computational cost is to reduce the number of iterations by applying a batch approach that selects a set of points to query in each iteration. A naive batch active learning approach would be to select the top k points that have the highest normalized uncertainty to query their neighborhoods. However, such a strategy will typically select highly redundant points. Designing a successful batch method requires carefully trading-off the value (normalized uncertainty) of the selected points.

Also in this paper the original mixed dataset is splitted into numeric dataset and categorical dataset and clustered using both traditional clustering algorithms (K-Means and K-Modes based as existing) and to implement fuzzy clustering algorithms (Fuzzy C-Means and Fuzzy C-Modes). The resultant clusters are combined using ensemble clustering methods and evaluated by both f-measure and entropy measure. It is found that splitting is more beneficial and applying fuzzy clustering algorithms yields better results than traditional clustering algorithms.

3.2 Implementation

The link based algorithm is used to implement for the data clustering. In this algorithm is used to clustering the data from the data base. Already the data will be stored in the data base. The two way to clustering from the data base one is categorical dataset another one is numerical dataset.

The general framework for pairwise constrains of cluster ensembles. Essentially, solutions achieved from different base clustering are aggregated to form a final the basic process of cluster ensembles. It first applies multiple bases clustering to a data set to search with semantic based to obtain diverse clustering decisions. Then these solutions are combined to establish the final clustering result using a consensus function. This multilevel methodology involves two major tasks generating a cluster ensemble and producing the final partition, normally referred to as a consensus function .it matches the fuzzy cluster approach in following base.

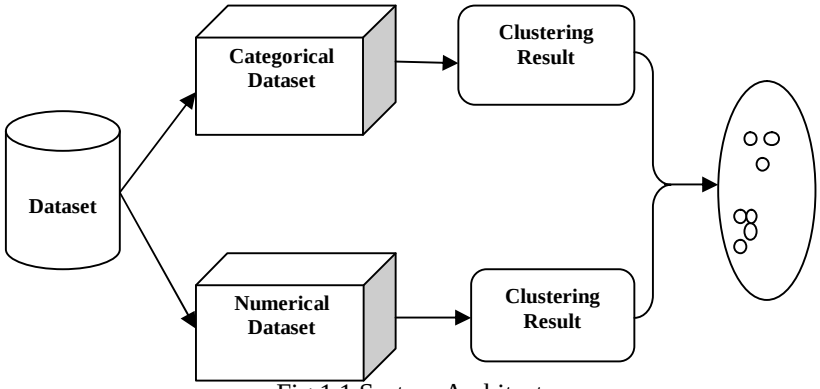

Fig 1.1 System Architecture

3.2.1 Categorical Dataset

The categorical dataset is already data will be stored in the database. The manual data clustering from the data base. The data clustering is used to link based algorithm to cluster the data from the data base. The related data will be cluster from the data set in the categorical data set. Only the related data to be clustering from the data set.

3.2.2 Numerical Dataset

The numerical dataset is data to be stored from the data base already stored from the data from the data base. The link based algorithm is used to clustering the data from the data base. They will be not only related data clustering data all

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

data to be clustering from the data base. It can be is used to fuzzy clustering algorithm clustering the data from the data set. The final clustering the data from the dataset is used to categorical dataset and numerical dataset to be clustered from the data base and then to display the finial data clustering dataset from the data base.

IV. EXPERIMENTAL RESULTS

In this section, we present the experimental results on two datasets. The first one, called client IP addresses (numerical dataset). The second dataset, called book details (categorical dataset).

Fig 1.2 Login Form

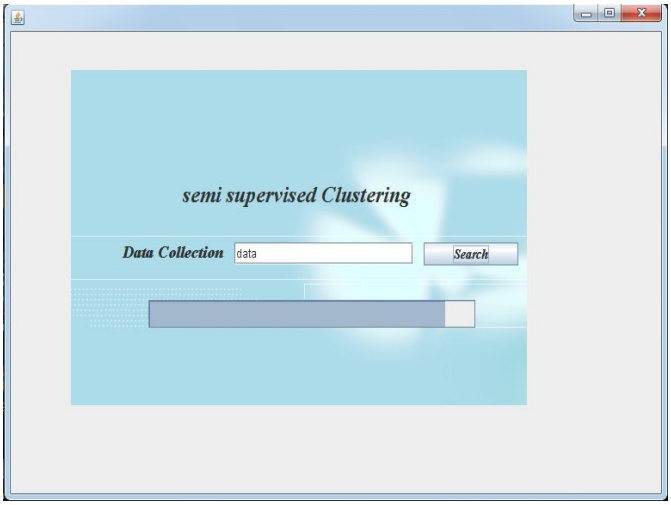

Fig 1.3 Semi-Supervised Clustering

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

- Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
- Select methods for presenting information.
- Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

- Confirm an action.

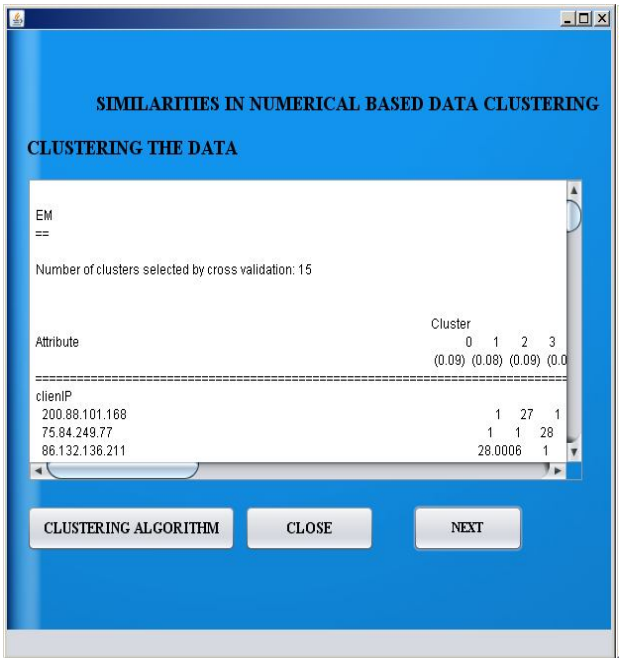

Fig 1.4 Numerical Dataset

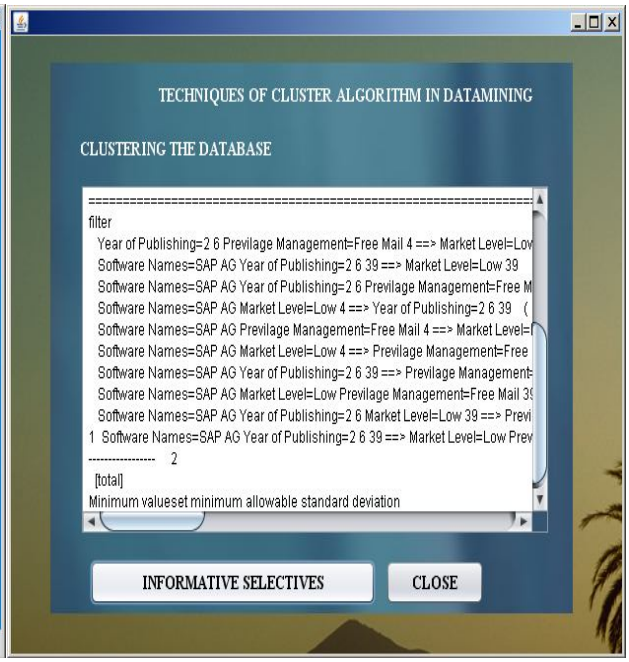

Fig 1.5 Categorical Dataset

V. CONCLUSION AND FUTURE WORK

The link-based cluster ensemble approach to categorical data clustering. It transforms the original uncompromising data matrix to an information preserve statistical difference to which an effective graph partition technique can be directly applied. The problem of construct the is resourcefully resolved by the similarity among uncompromising labels or clusters using the Weighted Triple excellence comparison algorithm. The observed study with different ensemble type's validity procedures and data sets suggests that the proposed link based method usually achieves higher clustering results compared to those of the traditional uncompromising data algorithms and standard cluster ensemble techniques. The prominent future work includes an extensive study regarding the behavior of other link based similarity measures within this problem context. Also the new method will be applied to specific domains including tourism and medical data sets.

REFERENCES

[1] S. Basu, A. Banerjee, and R. Mooney, "Active semi-supervision for pairwise constrained clustering," in SIAM International Conference on Data Mining, pp. 333–344, 2004.
[2] S. Basu, I. Davidson, and K. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall, 2008.
[3] M. Bilenko, S. Basu, and R. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in International Conference on Machine learning, pp. 11–18, 2004.
[4] I. Davidson, K. Wagstaff, and S. Basu, "Measuring constraint-set utility for partitional clustering algorithms," Knowledge Discovery in Databases, pp. 115–126, 2006.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

[5] D. Greene and P. Cunningham, "Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering," European Conference on Machine Learning, pp. 140–151, 2007.

[6] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," Journal of Artificial Intelligence Research, vol. 4, pp. 129–145, 1996.

[7] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," Advances in Neural Information Processing Systems, pp. 593– 600, 2007.

[8] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Batch mode active learning and its application to medical image classification," in International Conference on Machine learning, pp. 417–424, 2006.

[9] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," in Computer Vision and Pattern Recognition, pp. 1–7, 2008.

[10] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and representative examples," Advances in Neural Information Processing Systems, pp. 892–900, 2010.

[11] B. Settles, "Active learning literature survey," tech. rep., 2010.

[12] R. Huang and W. Lam, "Semi-supervised Document Clustering . via Active Learning with Pairwise Constraints," in International Conference on Date Mining, pp. 517–522, 2007.

[13] P. Mallapragada, R. Jin, and A. Jain, "Active query selection for semi-supervised clustering," in International Conference on Pattern Recognition, pp. 1–4, 2008.

[14] Q. Xu, M. Desjardins, and K. Wagstaff, "Active constrained clustering by examining spectral eigenvectors," in Discovery Science, pp. 294–307, 2005.

[15] L. Breiman "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.