ISSN:0975-9646

Active Learning of constraints using incremental approach in semi-supervised clustering

Jamil Ahmed Sk [#1] , Manjunath Prasad [*2] ,Abdullah Gubbi[*3], Hasibur Rahman[#4]

[#1,*2] *Department of computer science*[*3,#4] *Department of EC*
[1,3,4]*P.A. College of Engineering, Mangalore, Karnataka, India.*
[2] *NMAMIT,NITTE Mangalore, Karnataka, India.*

***Abstract*— Semi-supervised clustering aims to improve clustering performance by considering user-provided side information in the form of pairwise constraints. We study the active learning problem of selecting must-link and cannot-link pairwise constraints for semi-supervised clustering. We consider active learning in an iterative framework; each iteration queries are selected based on the current clustering outcome and constraints available. We use the neighborhood framework where the pairwise points having the must-link belong to the same neigborhood and cannot-link pairwise points belong to the different neighborhood. If two points belong to the same neighborhood then they belong to the same cluster and viceversa. We will use the Glass Identification Data Set from the UCI machine learning repositories and investigate the improvement in clustering time using the Incremental Clustering.**

***Keywords*— Active learning, Semi-supervised clustering, Incremental Approach, Pairwise constraints.**

I. INTRODUCTION

Semi-supervised clustering is a technique that make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-Supervised learning falls between unsupervised learning and supervised learning. Active learning systems attempt to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an expert. In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data. Active learning is well motivated in many modern machine learning problems where data may be abundant but labels are scarce or expensive to obtain.

Semi-supervised clustering aims to improve the clustering by providing a user provided side information. Pairwise constraints is one of the types of side information, which include must-link and cannot-link constraints that specifies whether the two points must or must not belong to the same cluster.

In this paper, we consider iterative framework in active learning. Each iteration finds the most important information toward improving the current clustering model and form queries accordingly. The response given in the queries are used to update the clustering. This process continues until we reach a satisfactory outcome or we reach a maximum number of queries.

Such a iterative framework is used active learning. We focus on the neighborhood framework. A points belonging to the same cluster as the constraints is said to belong to the same neighborhood and points belonging to different clusters as per the constraints is said to have different neighborhood. Valuable information can be furnished from a well formed neighborhood about the demography of the cluster.

We propose an incremental clustering approach where the most informative point is selected based on the uncertainty of its belonging to a cluster. The user provides side information based on which these most informative point is assigned the cluster. In this approach, there is an improvement of time from the method proposed in [1].

The remainder of the paper is organized as follows: Section 2 presents a brief review of the related work. Section 3 presents framework of the problem and the proposed solution. The Experimental results are discussed in Section 4. Finally, we end our discussion in Section 5 with Conclusion.

II. LITERATURE SURVEY

In paper [6],[7],[8][9][10][11], Active learning has been extensively used in the field of supervised classification. The research based on the active learning for the constraint based clustering as been limited. The first work on this field is that of Basu et al[2].They used two phase approach i.e. Explore and Consolidate (E & C). In the first phase(Explore), it incrementally selects points using the farthest first query scheme(FFQS) and queries their relationship to identify n disjoint neighborhoods, where n is the number of clusters. In the second phase(Consolidate), it iteratively expands the neighborhoods, where in each iteration it selects a random point outside any neighborhood and queries it until a must-link is found in the existing neighborhoods. This method performs better than the unconstrained k means. The disadvantage of this method is that it is sensitive to outlier and in the second phase the points were chosen randomly which could give an incorrect clustering result.

In [4],P. Mallapragada et al, proposed a method named Min-Max which was an improvement over Explore and Consolidate. In Min-Max method, modifies the consolidate phase by choosing the most uncertain point as opposed to randomly selected point.

In [5], Q. Xu et al, proposed to select constraints by examining the spectral eigen vectors of the similarity matrix, which is limited to two-cluster problems.

In [3], R. Huang et al, proposed a framework that takes an iterative approach. In each iteration, with current set of constraints performs semi-supervised clustering to produce a probabilistic clustering assignment. It computes the probability of them belonging to the same cluster and measures the associated uncertainty. To select, it considers all unconstrained pairs that has exactly on document already assigned to one of the existing neighborhoods and identify the most uncertain pair to query among them. If a must link is returned as answer, it terminates and moves onto the next iteration. Otherwise, it will query the unassigned point against the existing neighborhoods until a must-link is found. This method focuses on the pairwise uncertainty for the first query and then fails to measure the ensuing queries.

In [1], Xiong et al, they proposed a neighborhood based approach and incrementally expands the neighborhoods by posing paiwise queries. They devised an instance based selection criterion that identifies in each iteration the best instance to include into the existing neighborhoods. The selection criterion trades off two factors, the information content of the instance, which is measured by the uncertainty about which neighborhood the instance belongs to; and the cost of acquiring this information, which is measured by the expected number of queries required to determine its neighborhood.

III. METHODOLOGY

The problem addressed in this paper is how to effectively choose pairwise queries to assign an accurate clustering . Through active learning, we aim to achieve query efficiency, i.e., we would like to reduce the number of queries/questions asked 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. We start with the formulation of the active learning problem.

Problem Statement

Formally, we consider that they are c distinct classes that assigns each instances to one of the classes, a set of Data instances $D=\{x_1,x_2,....,x_n\}$ and y is the unknown label of each instances .Then each $x_i \in y_i$, where $y_i \in \{1,2,.....,c\}$, for all $i \in \{1,2,....n\}$. We need to query a pair of instances for find whether they belong to the same neighborhood or different neighborhood i.e. whether the pair of instances are "must link"(ML) or "cannot-link"(CL) with the given constraints and the current clustering scenario. If the pair of point belongs to the same neighborhood then they reply that it is must- link(ML) otherwise cannot-link(CL).

TABLE 1:
RULES SHOWING THE MUST-LINK(ML) AND CANNOT-LINK(CL)

Rules	**(x_i,x_j) (1)**	**(x_j,x_k) (2)**	**(1)^(2)=>(x_i,x_k) (3)**
Rule 1	ML	ML	ML
Rule 2	ML	CL	CL

Proposed Solution

In [1], they proposed the method where a most informative point is selected and queried to find the clustering assignment given the side information which is the constraints provided by the user. In this method, every time the clustering is done with the most informative point and the constraint from the scratch and hence it is time consuming.

We propose an incremental clustering approach wherein the most informative point is selected and the side information is provided. Then the clustering is done only by re-assigning these most information points to the clusters. After the clustering more informative point will be generated , if the user is satisfied with the result of clustering and give no side information then clustering stops and we find and compare the time with the method proposed in [1].

In Fig. 1, the schematic diagram of the system design below shows the proposed methodology.

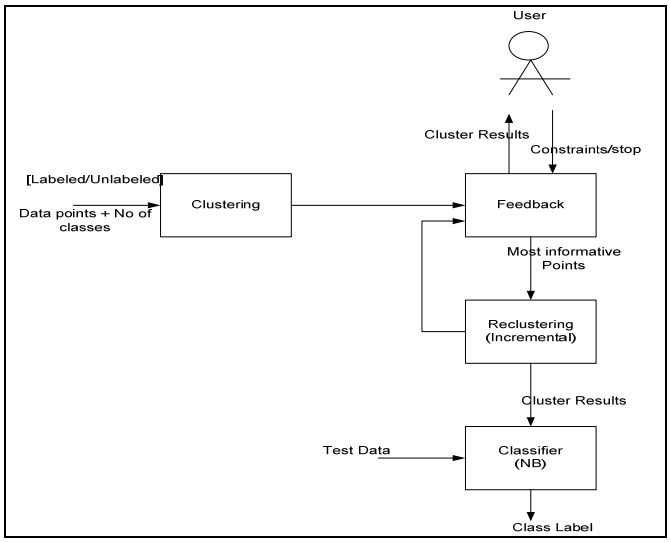

Fig. 1 A schematic diagram showing the proposed method..

IV. EXPERIMENT

In this section, we discuss about the data set used for the experiment, experimental settings and the comparison on the proposed method with current state- of- the-art method.

Experimental Data Sets

We have used the Glass Identification Data Set from UCI Machine Learning Repository. There are 9 Attributes used in the Data Set are Refractive Index(RI), Sodium(Na), Magnesium(Mg), Aluminium(Al), Silicon(Si), Potassium(K), Calcium(Ca), Barium(Ba) and Iron(Fe). The number of classes are 7 , which refer to the type of glass as building windows float processed, building windows non float processed, vehicle windows float processed, vehicle windows non float processed, containers, tableware and headlamps. The details of the data set is given in the table below:

Table 2 CHARACTERISTIC FEATURES OF THE DATASET

Data Set Characteristics	Multivariate	No. Of Instances	105
Attribute Characteristics	Real	No. Of Attributes	9

Experimental Result

This section presents the experiment results on running on the Normal Clustering and the Incremental Clustering. The system is loaded with unlabelled and labelled data. The experiment has two different clustering approach- Normal clustering and the Incremental clustering. In Normal Clustering(NC), the data is presented to the system and the most informative point is generated in the feedback form. The expert will decide the data point if there is ambiguity and he needs to fill the feedback form and the procedure is iteratively repeated. This procedure is repeated till all the data points are correctly assigned with the correct labels. The convergence time is measured and plotted using JFreeChart.

Next, we choose Incremental Clustering wherein the data points are assigned with the cluster if there is no confusion(ambiguity).In case any ambiguity only that data point is assigned with the label by the expert. The labels for the rest of the data points are unaltered. Finally the time for the incremental clustering is noted. It has been experimentally found in the investigation that the time for clustering using incremental approach is better than the normal clustering as the number of iteration are less.

Fig. 2 shows that on clustering the data set with two iteration using both the approach- the time taken by Normal clustering is 7797ms and the time taken by Incremental clustering is 1687ms. The Speedup Achieved using incremental approach is 4.62 i.e. the clustering assignment of the experiment data has improved the clustering time by 4.62 times.

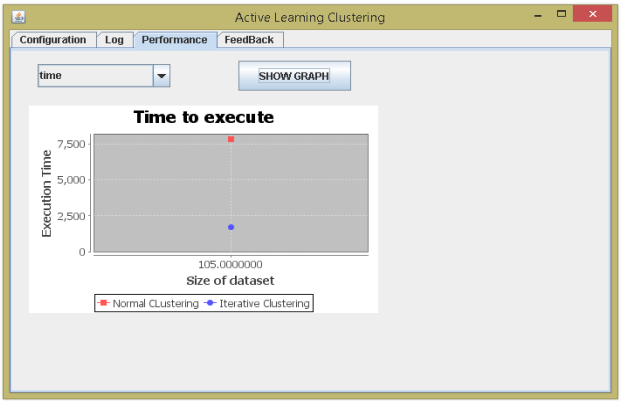

Fig. 2 Iterative clustering showing better convergence time than Normal Clustering with expert side information

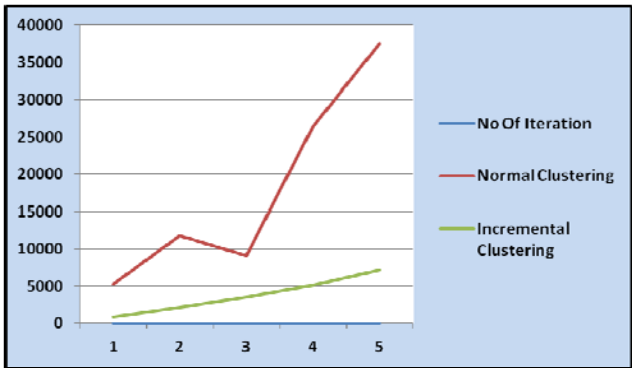

Fig. 3 Plotted graph indicates than Incremental clustering converges with less time than Normal clustering. X axis indicates the No. of iteration and the Y axis indicates the time taken for clustering in milliseconds(ms)

The graph in Fig. 3 shows the time taken by both the approach on different iteration numbers i.e. 1,3,5,7,10. In this case the line which will be lower is the better approach as it indicates a better result with less clustering time. The graph indicates that the Incremental Cluster shows better result than Normal Clustering with the increase in the iterations.

V. Conclusions

In this paper, we study an incremental active learning framework to select pairwise constraints for semi-supervised clustering. Our method takes a neighborhood-based approach, and incrementally expands the neighborhoods by posing pairwise queries. The uncertain points are selected and user side information is used to decide in the assignment of that point in the cluster.

We can expand the work by combining the concept of soft computing with the Active learning of constraints for the Semi-Supervised clustering. After labelling the dataset, we can compare different classifier techniques to find the better among the classifiers. We can also use semi-supervised clustering on the application domain where unsupervised clustering has been used and investigate the better technique for that domain.

References

[1] Sicheng Xiong, Javad Azimi, and Xiaoli Z. Fern, "Active Learning of Constraints for Semi-Supervised Clustering",IEEE Transactions on Knowledge and Data Engineering, Vol 26, No 1,January 2014.

[2] S.Basu, I.Davidson,andK.Wagstaff,Constrained Clustering:Advances in Algorithms, Theory, and Applications. Chapman & Hall, 2008.

[3] R. Huang and W. Lam, "Semi-Supervised Document Clustering via Active Learning with Pairwise Constraints," Proc. Int'l Conf.Date Mining, pp. 517-522, 2007.

[4] P. Mallapragada, R. Jin, and A. Jain, "Active Query Selection for Semi-Supervised Clustering," Proc. Int'l Conf. Pattern Recognition, pp. 1-4, 2008.

[5] Q.Xu, M. Desjardins, and K. Wagstaff, "Active Constrained Clustering by Examining Spectral Eigenvectors," Proc. Eighth Int'l Conf. Discovery Science, pp. 294-307, 2005.

[6] D. Cohn, Z. Ghahramani, and M. Jordan, "Active Learning with Statistical Models," J. Artificial Intelligence Research, vol. 4, pp. 129-145, 1996.

[7] Y. Guo and D. Schuurmans, "Discriminative Batch Mode Active Learning," Proc. Advances in Neural Information Processing Systems, pp. 593-600, 2008.

[8] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Batch Mode Active Learning and Its Application to Medical Image Classification," Proc. 23rd Int'l Conf. 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," Proc. IEEE Conf. 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," Proc. Advances in Neural Information Processing Systems, pp. 892-900, 2010.

[11] B. Settles, "Active Learning Literature Survey," technical report, 2010.

[12] O. Shamir and N. Tishby, "Spectral Clustering on a Budget ", J. Machine Learning Research-Proc Track, vol 15, pp. 661-669,2011.