A Survey on Low Complexity Active Learning for Semi-Supervised Clustering of Large Data Sets

P.Gandhini[#1], S.Venkatesh Babu[*2]

[#]PG Scholar, Department of Computer Science and Engineering, Christian College of Engineering and Technology, Dindigul, Tamilnadu - 624619, India.
gandhini1990@gmail.com

*Associate Professor, Department of Information Technology , Christian College of Engineering and Technology, Dindigul, Tamilnadu - 624619, India.
venkateshflower6@gmail.com

***Abstract*—The main objective of Semi-supervised clustering is improving cluster performance by user's guidance in the form of pairwise constraints named as must-link and cannot-link. The classifier uses users feedback during learning phase of classification is referred as Active learning. The active learning has the problem of satisfying the must required and not required constraints for semi supervised clustering. Active learning is achieved in an iterative manner. The concept of neighborhood of different clusters is formed according to the pair wise constraints. Under this, classic uncertainty-based principle is used and proposes a novel approach for estimating the uncertainty to each data point. This paper evaluates on eight bench mark data sets and its result achieves consistent and substantial improvements over its competitors. An active learning in iterative manner requires repeated reclustering of the data with an incrementally growing constraint set. This provides computational complexity for large data sets. To address this problem, we introduce an incremental semi-supervised clustering method that updates the existing clustering solution based on the neighborhood assignment for the new point. A naive batch active learning approach also proposed to select the top k points for active learning instead of selecting all data for training.**

***Keywords*—Active learning, clustering, semi-supervised clustering, naive batch active learning.**

I. INTRODUCTION

Data mining is extracting the information from the large group of data. Similarly data mining is mining the knowledge from the set of data. Clustering is the task that groups a set of objects in such a way, the similar objects are classed together in one cluster and dissimilar are classed in another cluster. There are many clustering methods are used to improve the cluster. In data mining many constraints are used but in semi-supervised clustering it use the pair wise constraints to improve the performance of cluster. semi-supervised clustering is done by the help of users guidance in the form of pairwise constraints such as must required and cannot required constraints specifying that two points must or must not belong to the same cluster[1],[2],[3],[4],[5]. However, if the constraints are selected improperly, they may also degrade the clustering performance [4],[5]. Moreover, obtaining pairwise constraints typically requires a user to manually inspect the data points, which can be time consuming and costly. For example, for document clustering, obtaining a must-link or cannot-link constraint requires a user to potentially scan through the documents and determine their relationship, which is feasible but costly in time. For those reasons, we would like to optimize the selection of the constraints for semi-supervised clustering, which is the topic of active learning.

Active learning is an artificially intelligence. While active learning has been extensively studied in supervised learning, the research on low complexity active learning for semi-supervised clustering of large data sets is relatively limited. Active learning has a problem of satisfying pairwise constraints for semi-supervised clustering, it can be achieved by iterative framework where in each iteration queries (i.e., constraints) are selected based on the active clustering solution and the existing pairwise constraint set. This process repeats until reaching a satisfactory solution or maximum number of queries allowed [6],[7],[8],[9]. Focus on a framework that builds on the neighborhood concept, it is introduced for different clusters in the form of pairwise constraints by choosing the most useful data points and query their kinship with the neighborhood. under this framework, builds on a classic- uncertainty based principle where uncertainty in terms of probability of the point belonging to different known neighborhoods and present a novel approach for estimating the probabilities with each data points. Introduced the selection criterion that tradeoff by normalizing the uncertainty of selected data point with the information rate.

In this paper, empirically evaluated on eight benchmark UCI data sets against a number of competing methods. The evaluation results improved the consistent and substantial performance over the current state of the art [1]. The iterative framework in active learning requires repeated reclustering of the data with an incrementally growing constraint set. This can be computationally demanding for large data sets. To solve this problem, an incremental semi-supervised clustering method that updates the existing clustering solution based on the neighborhood assignment for the new point [12]. 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 [6],[7],[8],[11]. A naive batch active learning approach would be to select the top k points that have the highest normalized uncertainty to query their neighborhoods.

A. Semi-supervised Learning

In many practical learning domains (for example: text processing, bioinformatics), it has a large number of unlabeled data but small amount of labeled data and in most cases it can be expensive to generate the labeled data. Subsequently, semi-supervised learning is learning from a conjunction of both labeled and unlabeled data during training [1].

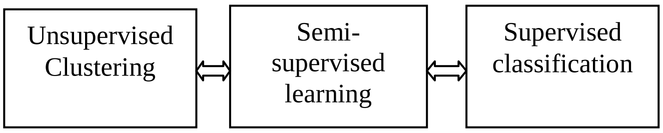

Fig. 1. Semi-Supervised Learning

1) Semi-supervised Clafficiation:

It has a fixed known set of groups and class-labeled training data is used to induce in this function. In this setting, training can also have additional unlabeled data, similarly resulting in a more accurate classification function.

2) Semi-supervised Clustering:

It uses limited labeled data, which is available to aid the unsupervised clustering process. Semi-supervised clustering consists of two approaches such as Constraint-based and Distance-based methods. In constraint-based methods, the clustering algorithm itself modified so that the user provided class labels or pairwise constraints are used to direct the algorithm towards a suitable data partitioning. This is done by the modified clustering objective function and that also included the self satisfaction of constraints, enforcing the constraints during initializing and constraining based on labeled examples. In distance-based or metric-based methods, an existing clustering algorithm has distance metric and that is employed; whereas, the metric is first guided to satisfy the class labels or constraints in the semi-supervised data. More number of distance measures that have been used for metric-based clustering including Euclidean distance guided by shortest-path algorithm, string-edit distance learned using EM, KL divergence adapted using gradient descent and Mahalanobis distances trained using convex optimization[1],[3].

B. Active learning Algorithm

Active learning is a subfield of machine learning and, more generally artificial intelligence. Active learning algorithm is a branch of classification algorithm, because of relatively wide research direction and application, domestic and foreign scholars have put forward many topics. References use source domain data to study the target domain with active learning algorithm, trying to simplify the sample point label complexity. Tomanek et al described the important application of active learning in the Natural Language Processing (NLP), 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.

C. Neighborhood-based Framework

A neighborhood contains a set of data instances that are known to belong to the same class (that is connected by must link constraints). Furthermore, different neighborhoods are connected by cannot-link constraints and, thus, are known to belong to different classes. The nodes denote data instances, and the solid lines denote must link constraints while the dashed lines denote cannot link constraints. Each neighborhood is required to have a cannot link constraint with all other neighborhoods [1]. Therefore, Fig. 2a contains three neighborhoods: {x1,x2},{x3], and {x4}, whereas Fig. 2b contains only two known neighborhoods, which can be either {x1,x2},{x3} or {x1,x2},{x4}.

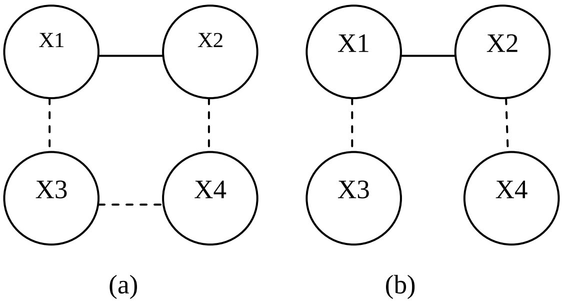

Fig.2. Two examples to show how to identify neighborhoods from a set of pairwise constraints.

D. UCI Data Sets

TABLE1
Characteristics of the Data Sets

Datasets	# of Classes	# of Features	# of Examples
Breast	2	9	683
Digits-389	3	16	3165
Ecoli	5	7	327
Glass	6	9	214
Heart	2	13	270
Parkinsons	2	22	195
Segment	7	19	2310
Wine	3	13	178

In this paper, used eight benchmark UCI data sets that have been used in constraint based clustering[1], [3], includes breast, pen-based recognition of handwritten digits(3,8,9), ecoli, glass identification, statlog-heart, parkinsons, statlog-image segmentation and wine[1]. The characteristics of the eight data sets are shown in above Table 1.

II. PROPOSED WORK

A. Incremental Semi-supervised Clustering Method

In this method, many other techniques are used, to avoid reclustering the entire data sets and to reduce the computational cost. Efficient incremental constrained clustering algorithm used basic sufficient conditions and nontrival sufficient conditions to provide instructions to a user, based on the properties that are fulfilled by constraints so that the given partition could be effectively updated the satisfied new constraints. In this approach the greedy incremental algorithm was introduced. It has two types one is center-greedy and another is distance-greedy. The center greedy algorithm combines a center for each cluster and merges the two clusters whose centers are closest. The center of the old cluster with large radius becomes the new center. The distance-greedy algorithm integrates those two clusters which minimize the diameter of the resulting integrated cluster.

B. Naive Batch Active Learning Approach

A naive approach to batch active learning is to simply apply an existing sequential selection rule k times to generate a batch, example: selecting the k minimum margin examples. This approach however will often perform poorly since it will tend to ignore redundancy among selected examples. To overcome this issue, there has been a small amount of work on batch active learning, which provides different heuristic approaches for incorporating batch diversity into the selection method.

This paper considers a general approach for batch active learning, which refer as simulation matching. We are motivated by the observation that sequential methods are generally more example-efficient than their batch counterparts, since each example is selected with more information. Indeed in theory the best sequential strategy will never be worse than the best batch strategy, since one could simulate the batch approach and then select the examples sequentially. Leveraging the availability of highly-effective sequential active learning methods, we view a given sequential method as a gold standard whose performance, in terms of label efficiency, we would like to approach using batch selection.

We use Monte-Carlo to estimate the posterior distribution over examples selected by the sequential method, and then select a batch of k examples that best matches these distributions. A key contribution of this work is to instantiate the notion of best match by developing a novel matching objective called bounded coordinated matching. This problem is called as NP-hard, so that we propose an efficient greedy algorithm that optimizes the objective with an approximation bound.

We propose the criteria such as minimum redundancy, maximum uncertainty and maximum impact to express the informativeness of a set of instances, and formulize the batch mode active learning problem as choosing a set of instances by maximizing an objective function that combines both link as well as content information.

III. RELATED WORK

Many researches where done in active learning for semi-supervised clustering using pairwise constraints[1],[2],[3],[5],[10],[11], and also improves the clustering accuracy. Sometimes if the constraints are selected improperly then they may also degrade the clustering performance[4],[5]. Basu et al. [2] proposed random ordering in consolidate phase but Mallapragada et al. [10] poses an improvement in Explore and Consolidate named as Min-Max, that modifies the Consolidate phase by selecting the most uncertain point to query(opposed randomly). [3] They used Metric Pariwise Constrained K-Means (MPCK-Means) to have different shapes and used unified approach to get better cluster and it is difficult to have high dimensional data sets, but this can be achieved in this survey paper. Davidson et al. [4] introduced two quantitative measures that used to identify useful constraint set for increasing the performance and it has high computational effort whereas this may also reduced in the proposed work of the survey paper. In supervised active learning the data pints are selected iteratively based on current clustering classification model and that can be improved efficiently[6],[7],[8],[9]. [11] Has high quality and efficient sequential active learning policies applied for k steps. Having optimization problem in terms of bounded coordinated matching that is NP-hard and they uses efficient greedy algorithm to tackle this problem and also used eight bench mark data sets that is highly effective. This is what same in proposed work of this paper. Limited behavior and batch Bayesian optimization where used in previous work, that is not suitable for active learning [11]. Davidson et al. [12] proposed sufficient conditions for efficiently adding a clustering to assuage the new and old constraints rather than rerunning the whole data sets, this is done by the help of users feedback. This work is more relevant to proposed method of this paper. [12] The maximum number of additional constraints allowed before complete reclustering is not known and also there is a problem in incrementally modifying a distance function, step by step these are all addressed in the proposed task. Xiong [1] introduced some methods and it also used eight UCI data sets to achieve the improvement in clustering. Here the computational demand occurs for large data sets due to large number of iterations and not considered the trading off values, these can be conclude in this paper by using efficient incremental semi-supervised clustering method and also uses the naive batch active learning approach to lower the computational cost by selecting the top k points that have highest normalized uncertainty.

IV. CONCLUSION

This paper evaluate on the eight benchmark data sets against a number of competing methods. The evaluation results indicate that our method achieves consistent and substantial improvements over its competitors. Proposed method on updating the new data in existing clustering with the help of neighborhood concept using incremental semi supervised clustering method and to lower the computational cost by reducing the number of iteration using batch approach. A naive batch active learning approach would be to select the top k points that have the highest normalized uncertainty to query their neighborhoods. A naive approach to batch active learning is to simply apply an existing sequential selection rule k times to generate a batch. Thus the result produced selects the highly redundant points when compare with the existing system.

REFERENCES

[1] S. Xiong, J. Azimi, & Z. Fern, "Active Learning of Constraints for Semi-Supervised Clustering", IEEE Trans. Knowledge and Data Engineering, 2014.

[2] S. Basu, A. Banerjee, & R. Mooney, "Active Semi-Supervision for Pairwise Constrained Clustering", International Conference on Data Mining, 2004.

[3] M. Bilenko, S. Basu, & R. Mooney, "Integrating Constraints and Metric Learning in Semi-Supervised Clustering", International Conference on Machine Learning, 2004.

[4] I. Davidson, K. Wagstaff, & S. Basu, "Measuring Constraint-Set Utility for Partitional Clustering Algorithms", European Conference, 2006.

[5] D. Greene & P. Cunningham, "Constraint Selection by Committee: An Ensemble Approach to Identifying Informative Constraints for Semi-Supervised Clustering", European Conference on Machine Learning, 2007.

[6] Y. Guo & D. Schuurmans, "Discriminative Batch Mode Active Learning", Advances in Neural Information Processing Systems, 2008.

[7] S. Hoi, R. Jin, J. Zhu, & M. Lyu, "Batch Mode Active Learning and Its Application to Medical Image Classification", International Conference on Machine learning, 2006.

[8] S. Hoi, R. Jin, J. Zhu, & M. Lyu, "Semi-Supervised SVM Batch Mode Active Learning for Image Retrieval", IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[9] S. Huang, R. Jin, & Z. Zhou, "Active Learning by Querying Informative and Representative Examples", Advances in Neural Information Processing Systems, 2010.

[10] P. Mallapragada, R. Jin, & A. Jain, "Active Query Selection for Semi-Supervised clustering", International Conference on Pattern Recognition, 2001.

[11] J. Azimi, A. Fern, X. Fern, & G. Borradaile, "Batch Active Learning Via Coordinated Matching," Proc. 29th Int'l Conf. Machine Learning, 2012.

[12] I. Davidson, M. Ester, & S. Ravi, "Efficient Incremental Constrained Clustering," ACM KDD, 2007.

[13] R. Huang & W, Lam, "Semi-Supervised Document Clustering Via Active Learning with Pairwise Constraints", International Conference on Data Mining, 2007.

[14] D. Cohn, Z. Gharamani, & M. Jordan, "Active Learning with Satistical Modes', Journal on Artificial Intelligence Research, 1990.

[15] Q. Xu, M. Desjardins, & K. Wagstaff, "Active Constrained Clustering by Examining spectral Eigenvectors", International Conference on Discovery Science, 2005.