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. 3, Issue 4, April 2015

An Efficient Iterative Framework for Semi-Supervised Clustering Based Batch Sequential Active Learning Approach

S.Savitha, M. Sakthi Meena

PG Scholar, Department of CSE, Velalar College of Engineering and Technology, Anna University, Chennai, India

Assistant Professor, Department of CSE, Velalar College of Engineering and Technology, Anna University, Chennai, India

ABSTRACT: Semi-supervised is the machine learning field. In the previous work, selection of pairwise constraints for semi-supervised clustering is resolved using active learning method in an iterative manner. Semi-supervised clustering derived from the pairwise constraints. The pairwise constraint depends on the two kinds of constraints such as must-link and cannot-link.In this system, enhanced iterative framework with naive batch sequential active learning approach is applied to improve the clustering performance. The iterative framework requires repeated reclustering of the data with an incrementally growing constraint set. To address incrementally growing constraint set, a batch approach is applied which selects a set of points based on query in each iterative. In the iterative algorithm, k instances select the best matches in the distribution, leading to an optimization problem that term bounded coordinated matching. Leveraging the availability of highly-effective sequential active learning method will improve performance in terms of label efficiency and accuracy with less number of iterations.

KEYWORDS: Clusters, semi-supervised, active learning, classification, sequential active learning

I. INTRODUCTION

Semi-supervised is the combination of supervised learning and unsupervised learning. Supervised learning in term of labeled data and unsupervised in term of unlabeled data. In this paper using a small amount of label data and a large amount of unlabeled data, its expensive while compare to labeled data. The goals of semi-supervised cluster are 'labeled data' in the form of must-links (two points must be in the same cluster) and cannot-links (two points cannot be in the same cluster). In data mining and machine learning tasks are provides a large unlabeled data and small number of labeled data. Many machine-learning researchers have found that unlabeled instance, when used in conjunction with a not large amount of labeled data, which improves the learning accuracy.

Each iterative determines the most important information from current clustering to according query. The general approach based on the concept of neighborhoods, contains a set of data points belong to same cluster and different neighborhoods belong to different cluster according to the queries. In this paper due to absence of point-based uncertainty its fail to measure the total amount of information and pairwise constraints captures relationship between two points, so the accuracy and performance efficiency is decreased in the active learning method.

Clustering

Clustering is the task of grouping the objects such that the objects in same group are more similar from other group. It is a main task of exploratory data mining and a common technique for statistical data analysis used in many fields including machine learning, pattern detection, image examination, information recovery and bio-informatics. Clustering is the method of unsupervised classification or unsupervised segmentation. Clustering provides a linear delineation as a starting point which improves the clustering performance. The original motivation was to develop better benchmark challenge problem for clustering. The paper focuses on problem for solving the clustering and how clustering methods are used rather than the algorithmic details of the technique, which are already many comprehensive reviews of techniques available.

Probably the best-known limitations of the partition approach is the typical (algorithmic) requirement that the number of clusters be known in advance, but there is more than that classic formalization of a natural selection process. In a nutshell, game-theoretic perspective has the following attractive features:

1. It makes no assumption on the underlying data representation: like spectral clustering, it does not require that the elements to be clustered be represented as points in a vector space.
2. It makes no assumption on the structure of the affinity matrix, being it able to work with asymmetric and even negative similarity functions alike.
3. It does not require a priori knowledge on the number of clusters (since it extracts them sequentially).
4. It leaves clutter elements unassigned (useful, e.g., in figure/ground separation or one-class clustering problems).
5. It allows extracting overlapping clusters

Active learning is a special case of semi-supervised machine learning in which a learning algorithm is able to query the user interactively and obtain the desired outputs at new records points. It is sometimes also called optimal experimental design. There are situations in which unlabeled data is abundant but labeling is expensive. In such a development, learning algorithms can actively query the user for labeled instance. This type of iterative supervised learning is called active learning. With this approach, there is a risk that the algorithm be overwhelmed by uninformative instances.

The remainder works of this paper such as II.presents a related work on active learning of constraints III. Active Learning of constraints for semi-supervised clustering and the purposed method
VI. Problem statement V. Working principles of purposed system 6.experimental evaluations are presented 7.finally we conclude the paper and future enchantment.

II. RELATED WORK

Active learning has been studied extensively for supervised classification problems [10]. In contrast, the research on active learning for constraint-based clustering has been limited. As mentioned formerly, most of the offered research studied the selection of a set of initial constraints prior to performing the semi-supervised clustering. Specifically, the foremost on this topic was conducted by Basu et al. They proposed an approach called two phase approach which is same as 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 termed as the total number of clusters. The second phase (consolidate) iteratively expands the neighborhoods, each iteration it select as unsystematic point outside any neighborhood and queries it against the existing neighborhoods until a must-link is found. More recently, Mallapragada et al. [3] proposed an

improvement to Explore and Consolidate named Min-Max, which alters the consolidate phase by choosing the most uncertain point to query (as opposed to randomly).

Xu et al. proposed to select constraints by examining the spectral eigenvectors of the data similarity matrix, which is regrettably limited to two-cluster problems. In [6], constraints are selected by analyzing the co-association matrix (obtained by applying cluster ensembles to the records). A key dissimilarity of our technique from the above-mentioned work is that we iteratively select the next set of queries based on the current clustering assignment to improve the performance. This is comparable 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 [7], [9] .

Most relevant to our work is an active learning framework presented by Huang and Lam for the task of text clustering. Specifically, this structure takes a repetitive approach that is analogous to ours. In each repetition, their method performs semi-supervised clustering with the current set of constraints to produce a probabilistic clustering task. It then computes, for each pair of text, the probability belongs to the same cluster and measures the related uncertainty. To make a choice, 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 identifies the most uncertainty pair to query. If a "must-link" answer is returned, it halts and moves onto the next repetition. 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 text clustering, one could potentially apply the active learning approach to handle other types of data by assuming appropriate probabilistic models. We would like to highlight a key difference between Huang's method and our work is Huang's method makes the choice based on pairwise constraints, whereas we focus on the uncertainty of a dataset point in terms of which neighborhood it belongs to same or different cluster. This difference is subtle, but important. Pairwise uncertainty, measures only the relationship between the two points in the dataset. Depending on the query, we may need to go sequence of additional queries. Huang's method considers only the pairwise uncertainty of the first query, and fails to measure the point-based uncertainty to the ensuing queries. Our method focuses on point-based uncertainty, measures the total amount of information gained by the full sequence of queries. Furthermore, our method also takes the expected number of queries to resolve the uncertainty of a point, which has not been considered formerly. Finally, we want to mention another line of work that uses active learning to facilitate clustering [5], [8], where the goal is to cluster a set of objects by actively querying the distances between one or more pairs of dataset. This is dissimilar from the focus of this paper, where we requested only the pairwise must-link and cannot link constraints, and do not provide the specific distance values to user.

III. ACTIVE LEARNING OF CONSTRAINTS FOR SEMI-SUPERVISED CLUSTERING

Semi-Supervised clustering is the major task to improve clustering performance used by supervision. Semi-supervised clustering, which uses class labels or pairwise constraints on some instances to aid unsupervised clustering. Here using large number of unlabeled data and small number of label data, the label data is expensive while compare to unlabeled data. Most of the research on pairwise constraints, including the must-link and cannot-link constraints, specifying such that two points are belongs to same cluster (must-link) or different cluster (cannot-link) to queries. However we selected pairwise constraints which determined the relationship between two points using musk-link and cannot-link constraints, it possible to increase time and cost. In generally number of previous studies, we can lead to improved clustering performance through constraints.

The research on active learning of constraints for semi-supervised clustering is limited. Most of the previous work a set of constraints priority to performance in semi-supervised clustering; this is not a correct way to improve the clustering. The problem in active learning approach is more number of queries to achieve the good clustering performance.

Clustering is an unsupervised learning problem, which tries to group of data points into clusters such that points in the same cluster are more similar to different clusters. Classification is the most limited work in active learning approach, which is allowed different types of selection queries. In the existing system in active learning approach has been mostly restricted to classification, where different principles of query selection have been studied.

In this paper, each iterative determine the most important information according to the queries. We repeat the process until we satisfactory solution or it allows the maximum number of queries. Many machine learning domain (e.g. text processing, bioinformatics), there is a large supply of unlabeled data but limited labeled data, which can be expensive to generate. In the proposed system, focus on semi-supervised clustering, which uses a small amount of label data and large amount of unlabeled data. At this end, we developed a new method for actively learning selecting point-based uncertainty and pairwise constraints for semi-supervised clustering in the iterative framework.

IV. PROBLEM METHODOLOGY

Proper selection of constraints dataset improves the clustering perform, otherwise it degrades the clustering performance. A neighborhood contains a set of data points that belongs to same cluster according to the constraints and different neighborhoods belong to similar clusters, but the neighborhood can be viewed as contains the “labeled instances” of different clusters. Pairwise uncertainty uses neighborhood based framework and it captures only relationship between two points and fails to measure the total amount of information points.

All potential pairs select the highest uncertainty regarding whether they are must-link or cannot link.

The iterative support requires repeated reclustering of data with an incrementally growing constraint set.

- ✓ Large memory
- ✓ High computational cost
- ✓ More time consumption
- ✓ Large number of iteration

Using naive batch active learning approach select the top k points that have highest normalized uncertainty to query their neighborhoods, which increases the redundant points.

V. PROPOSED SYSTEM

Semi-supervised clustering aims to improve the clustering performance in the form of pairwise constraints and point-based constraints. Leverage the availability of highly effective or uncertainty in terms of label efficiency and accuracy with the number of iterative via batch selection. The sequential approach used to select the possible points and increases the redundant points. The proposed system is to solve the optimization problem by using KMMM (Matching Mixture Model). The batch approach is the process of sequential policies selecting a batch of k samples that are “closely matches” where k>1.

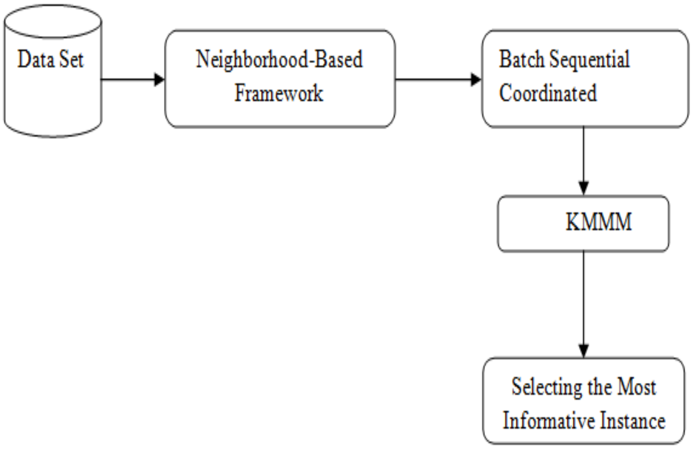
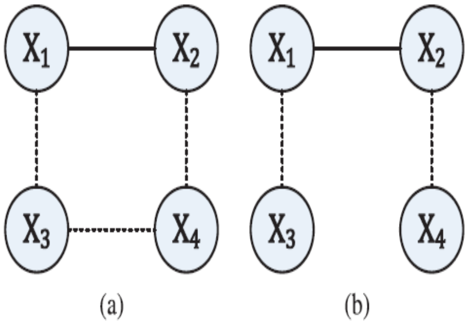

Data Set
Neighborhood-Based Framework
Batch Sequential Coordinated
KMMM
Selecting the Most Informative Instance

Fig 5. Working Principle

A. **Neighborhood-Based Framework**

A set of constraints denoted by C, set of l neighborhoods N= {N_1,…,N_l}, such that l ≤ c and c is the total number of classes. In Graphical representation of the data, the vertices represent data instances and edges represent must-link constraints. The neighborhood denoted by N_i belongs to D, i € {1… l} are connected components of the graph that have cannot-link constraints. A neighborhood contains a set of data points belong to the same cluster according to the constraints and neighborhoods belong to different clusters. Neighborhoods contain the "labeled instance" belongs to different class labels or same class labels. Neighborhood concept is leveraging the knowledge of neighborhood, to obtain a large number of constraints via a small number of queries. Active learning expands the neighborhoods by selecting the most informative data point and querying against the known neighborhoods. Nodes are denoted data instances, solid lines represented by must-link constraints and dash lines represented by cannot-link constraints.



B. **Batch Sequential Coordinated Matching with KMMM**

Given a dataset Dl of labeled instances, consider how to select the next batch of k instances to be labeled. For instances, a common measure of informativeness or class uncertainty of an instance with respect to the currently learned classifier. By picking the top k most informative instances measure selects the clusters of nearby instances that are quite redundant. The main idea behind of this batch approach is to leverage such sequential policies by selecting a batch of k > 1 samples that are closely matched. To select a batch B of k unlabeled instance with best matches from the sequential policy conditioned on the currently labeled instances. The batch B is considered to be a good match if it has high probability as compared to other set under the dataset distribution P_k, if the P_k distribution is not a closer set it is used to directly optimize the probability of B in trivial sequential policies.

KMMM

Given a KMMM model and a set of k points S ={x_1. . . .x_k}, there are k possible of S can be generated, each corresponding to one possible matching of the k points to the k components. Given such a matching m, let m denotes the index of the model component that is matched to point xi and let M denotes the set of all possible matching.

C. **Selecting the Most Informative Instance**

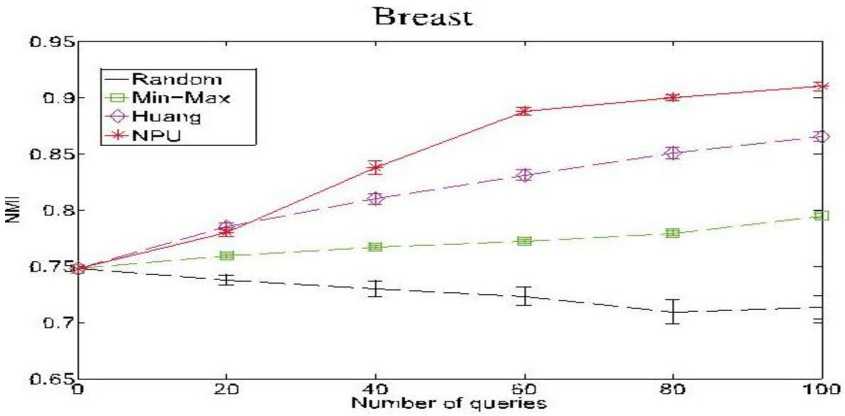
Given a set of existing neighborhoods, select an instance such that knowing its neighborhood allow to maximum information about clustering structure of the data. If predicted with high uncertainty to neighborhood an instance belongs to current clustering structure and low certainty instance will not lead to any gain of information to query. Similar observations have been used to motivate the uncertainty-based sampling principle for active learning of classifiers.

VI. EXPERIMENTAL RESULTS

In our experiments we use UCI data sets that have been used in the previous work on clustering. Data sets include breast, heart, ecoli, cancer, etc., In this section we evaluate the performance of our proposed method in comparison with exiting method. Previous work on semi-supervised clustering has similar result on random selected on data sets.

TABLE 1:Characteristics of the Data Sets

Datasets	# of Classes	# of Feature	# 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



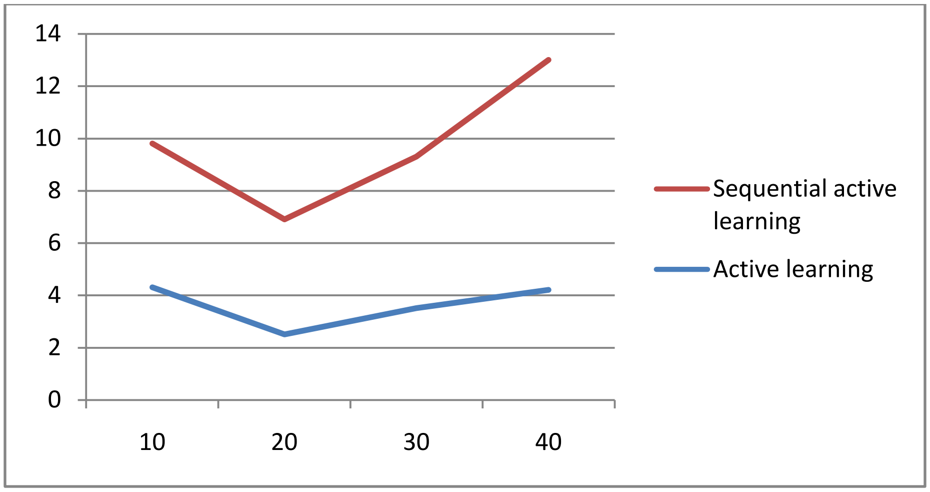
Evaluation Criteria

Two evaluation criteria are used in our experiments. A First criterion is normalized mutual information (NMI) to evaluate the clustering performance and its measures the mutual information between the two random variables. Another criterion is F-measure to evaluate the pairwise relationship between each pair of instances. It defined as the harmonic mean of precision and recall.

$$\mathrm{Precision} = \frac{\#\mathrm{PairsCorrectlyPredictedInSameCluster}}{\#\mathrm{TotalPairsPredictedInSameCluster}}$$

$$Recall = \frac{\#\mathrm{PairsCorrectlyPredictedInSameCluster}}{\#\mathrm{TotalPairsActuallyInSameCluster}}$$

$$F - measure = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{2\ \mathrm{Precision} + \mathrm{Recall}}$$



VII. CONCLUSION AND FUTURE WORK

In this paper we studied about neighborhood-based approach and selecting the most informative queries and proposed a novel method for batch sequential coordinated matching with KMMM.

The system enhances the iterative framework with batch sequential active learning approach. Each iterative selects a point-based and pairwise uncertainty which improves the clustering performance. The batch sequential approach select the k points that have highest uncertainty to query using neighborhood and also information is measured by the expected number of query. This system improves accuracy in less number of iterations and clustering performance.

Our method takes a point-based uncertainty and pairwise uncertainty which improve the performance. Extend our work to develop and apply an incremental semi-supervised clustering method that updates the previous clustering performance.

REFERENCES

[1] B. Settles, “Active Learning Literature Survey,” technical report, 2010.
[2] S. Basu, I. Davidson and K. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall, 2008.
[3] P. Mallapragada, R. Jin and A. Jain, “Active Query Selection for Semi-Supervised Clustering,” Proc. Int’l Conf. Pattern Recognition, pp. 1-4, 2008.
[4] SichengXiong, JavadAzimi, 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
[5] O. Shamir and N. Tishby, “Spectral Clustering on a Budget,” J. Machine Learning Research - Proc. Track, vol. 15, pp. 661-669, 2011.
[6] M. Al-Razgan and C. Domeniconi, “Clustering Ensembles with Active Constraints,” Applications of Supervised and Unsupervised Ensemble Methods, pp. 175-189, Springer, 2009.
[7] Y. Guo and D. Schuurmans, “Discriminative Batch Mode Active Learning,” Proc. Advances in Neural Information Processing Systems, pp. 593-600, 2008.
[8] K. Voevodski, M. Balcan, H. Ro¨ glin, S. Teng and Y. Xia, “Active Clustering of Biological Sequences,” J. Machine Learning Research, vol. 13, pp. 203-225, 2012.
[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.