International Research Journal of Management Science & Technology

ISSN 2250 – 1959(0nline)
2348 – 9367 (Print)

A REFEREED JOURNAL OF

Shri Param Hans Education & Research Foundation Trust

www.IRJMST.com
www.SPHERT.org

Published by iSaRa

Active Learning for Semi-Supervised Clustering Framework for High Dimensional Data

M. Pavithra[1], Dr.R.M.S.Parvathi [2.]

Assistant Professor, Department of C.S.E, Jansons Institute of Technology, Coimbatore, India[1].

Professor & Dean- PG Studies, Sri Ramakrishna Institute of Technology, Coimbatore, India[2].

ABSTRACT

In certain clustering tasks it is possible to obtain limited supervision in the form of pairwise constraints, i.e., pairs of instances labeled as belonging to same or different clusters. The resulting problem is known as semi-supervised clustering, an instance of semi-supervised learning stemming from a traditional unsupervised learning setting. Several algorithms exist for enhancing clustering quality by using supervision in the form of constraints [2]. These algorithms typically utilize the pairwise constraints to either modify the clustering objective function or to learn the clustering distortion measure. Semi-supervised clustering employs limited supervision in the form of labeled instances or pairwise instance constraints to aid unsupervised clustering and often significantly improves the clustering performance. Despite the vast amount of expert knowledge spent on this problem, most existing work is not designed for handling high-dimensional sparse data [4]. Semi-supervised clustering uses a small amount of supervised data to aid unsupervised learning. One typical approach specifies a limited number of must-link and cannot link constraints between pairs of examples. It presents a pairwise constrained clustering framework and a new method for actively selecting informative pairwise constraints to get improved clustering performance [6]. The clustering and active learning methods are both easily scalable to large datasets, and can handle very high dimensional data. Experimental and theoretical results confirm that this active querying of pairwise constraints significantly improves the accuracy of clustering when given a relatively small amount of supervision [5].

Key words: Data Mining Knowledge Discovery in Databases Clustering Semi Supervised Clustering High Dimensional Data

1. Introduction

In a *semi-supervised clustering* setting, the focus is on clustering large amounts of unlabeled data in the presence of a small amount of supervised data. In this setting, we consider a framework that has pairwise *must-link* and *cannot link* constraints between points in a dataset (with an associated cost of violating each constraint), in addition to having distances between the points. These constraints specify that two examples must be in the same cluster (must-link) or different clusters (cannot-link) [3]. In real-world unsupervised learning tasks, e.g., clustering for speaker identification in a conversation [1], visual correspondence in multi view image processing [7], clustering multi-spectral information from Mars images [2], etc., considering supervision in the form of constraints is generally more practical than providing class labels, since true labels may be unknown a priori, while it can be easier to specify whether pairs of points belong to the same cluster or different clusters. Semi-supervised clustering with constraints performs considerably better than unsupervised clustering for the datasets we have considered (note that unsupervised clustering

corresponds to semi-supervised clustering with 0 constraints). For both the active and non-active algorithms, the clustering evaluation measures (NMI and pairwise F-measure) and the objective function improve with increasing number of pairwise constraints provided along the learning curve. Active selection of pairwise constraints, using our two phase active learning algorithm, significantly outperforms random selection of constraints. Active learning in the classification framework is a long studied problem, where different principles of query selection have been studied, e.g., reduction of the version space size [6], reduction of uncertainty in predicted label [4], maximizing the margin on training data [1], finding high variance data points by density-weighted pool-based sampling [2], etc. However, active learning techniques in classification are not applicable in the clustering framework, since the basic underlying concept of reduction of classification error and variance over the distribution of examples [9] is not well-defined for clustering. In the unsupervised setting, Hofmann et al. [19] consider a model of active learning which is different from ours – they have incomplete pairwise similarities between points, and their active learning goal is to select new data, using expected value of information estimated from the existing data, such that the risk of making wrong estimates about the true underlying clustering from the existing incomplete data is minimized. In contrast, our model assumes that we have complete similarity information between all pairs of points, along with pairwise constraints whose violation cost is a component of the objective function, and the active learning goal is to select pairwise constraints which are most informative about the underlying clustering.

Many semi-supervised clustering methods have been proposed to enforce top-down structure while clustering [3, 5]. These methods allow the user to incorporate pairwise constraints, which may be either must-link (the two points/nodes belong in the same cluster) or cannot-link (the two points/nodes belong in different clusters), on the data as side information. These papers have shown that the use of pairwise constraints can significantly improve the correspondence between clusters and semantic labels when the constraints are selected well. [7] demonstrates that poorly chosen constraints can lead to worse performance than no constraints at all. Moreover, in real world problems each added constraint represents an additional real world cost, so maximizing the effectiveness of each constraint in order to minimize the total number of constraints needed is an important goal.

Currently, most work in semi-supervised clustering ignores this problem and simply selects a random constraint set (see above cited), but some work has now been done on *active* constraint selection methods [2, 4], which allow semi-supervised clustering algorithms to intelligently select constraints based on the structure of the data and/or intermediate clustering results.

2. Literature Survey

The novel scheme exploits both semi-kernel leaning and batch mode active learning for the relevance feedback in the content based image retrieval (CBIR). Particularly, learning of a kernel function from the combination of labeled and unlabeled examples is performed [7]. Then the kernel can be used for the effective identification of the informative and diverse examples for the active learning through the min-max framework [3]. Through the empirical study with the relevance feedback of the CBIR, the significant effectiveness of the proposed scheme compared to other state-of-the-art approaches is realized. Learning with the user interaction seems to be crucial in the

computer vision and pattern recognition applications. The users interact with the CBIR system, for the improvement of the retrieval quality. In such interactive procedure known as relevance feedback, the CBIR system tries to understand the information needs of the user, by learning from the feedback examples determined by the users [8]. Therefore, in order to achieve the desirable results, the traditional relevance feedback techniques have to repeat many times, due to the challenge of the semantic gap. In order for the reduction of the number of the labeled examples required by the feedback, the main key issue is the identification of the most informative unlabeled examples, for the efficient improvement in the retrieval performance [9].

3. Related Work

In the existing system, clustering is the technique of grouping a set of objects, such that the objects in the same group called as clusters are more similar to each other. More relevant to our work is an active learning framework presented by Huang and Lam 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 [1] 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 the 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. Finally, another line of work that uses active learning to facilitate clustering [8] is mentioned, 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 only pairwise must-link and cannot-link constraints are requested, and do not require the user to provide specific distance values.

4.1 Active Learning Algorithm

In the semi-supervised setting, getting labels on data point pairs may be expensive. In this section, we discuss an active learning scheme in the PCC setting in order to improve clustering performance with as few queries as possible. Formally, the scheme has access to a noiseless oracle that can assign a must-link or cannot-link label on a given pairs. In order to get pairwise constraints that are more informative than random in the PCC model, we have developed an active learning scheme for selecting pairwise constraints using the *farthest-first* traversal scheme [2]. The basic idea of farthest-first traversal of a set of points is to find points such that they are far from each other. In farthest-first traversal, we first select a starting point at random, choose the next point to be farthest from it and add it to the traversed set, then pick the following point farthest from the traversed set (using the standard notion of distance from a set [3]. Thus, the deviation of the centroid estimates falls exponentially with the number of seeds, and hence seeding results in good initial centroids. Since good initial centroids are very critical for the success of greedy algorithms such as KMeans, we follow the same principle for the pairwise case: we will try to get as many points (proportional to the actual cluster size) as

possible per cluster by asking pairwise queries, so that PCKMeans is initialized from a very good set of centroids [4].

The proposed active learning scheme has two phases. The first phase explores the given data to get pairwise disjoint non-null neighborhoods, each belonging to a different cluster in the underlying clustering of the data, as fast as possible [5]. Note that even if there is only one point per neighborhood, this neighborhood structure defines a correct skeleton of the underlying cluster. For this phase, we propose an algorithm Explore that essentially uses the farthest-first scheme to form appropriate queries for getting the required pairwise disjoint neighborhoods [6]. At the end of Explore, at least one point has been obtained per cluster. The remaining queries are used to consolidate this structure [7]. The cluster skeleton obtained from Explore is used to initialize pairwise disjoint non null neighborhoods

4.1.1 Explore

In the exploration phase, we use a very interesting property of the farthest-first traversal. Given a set of disjoint balls of unequal size in a metric space, we show that the farthest-first scheme is sure to get one point from each of the balls in a reasonably small number of attempts In Explore, while queries are still allowed and pairwise disjoint neighborhoods have not yet been found, the point farthest from all the existing neighborhoods is chosen as a candidate for starting a new neighborhood [8]. Queries are posed by pairing with a random point from each of the existing neighborhoods. If it cannot-linked to all the existing neighborhoods, a new neighborhood is started with. If a must-link is obtained for a particular neighborhood, is added to that neighborhood [1]. This continues till the algorithm runs out of queries or pairwise disjoint neighborhoods have been found [2].

Algorithm: Explore
Input: Set of data points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{n}$, access to an oracle that answers pairwise queries, number of clusters k, total number of queries Q.
Output: $\lambda \leq k$ disjoint neighborhoods $\{N_p\}_{p=1}^{\lambda}$ corresponding to the true clustering of $\mathcal{X}$ with at least one point per neighborhood.
Method:
1. Initialize: set all neighborhoods $\{N_p\}_{p=1}^{k}$ to null
2. Pick the first point $\mathbf{x}$ at random, add to N_1, $\lambda \leftarrow 1$
3. While queries are allowed and $\lambda < k$
   $\mathbf{x} \leftarrow$point farthest from existing neighborhoods $\{N_p\}_{p=1}^{\lambda}$
   if, by querying, $\mathbf{x}$ is cannot-linked to all existing neighborhoods
   $\lambda \leftarrow \lambda + 1$, start a new neighborhood N_λ with $\mathbf{x}$
   else
   add $\mathbf{x}$ to the neighborhood with which it is must-linked

4.1.2 Consolidation

The consolidation phase starts when at least one point has been obtained from each of the clusters [3]. The basic idea in the consolidation phase is that since we now have points from all the

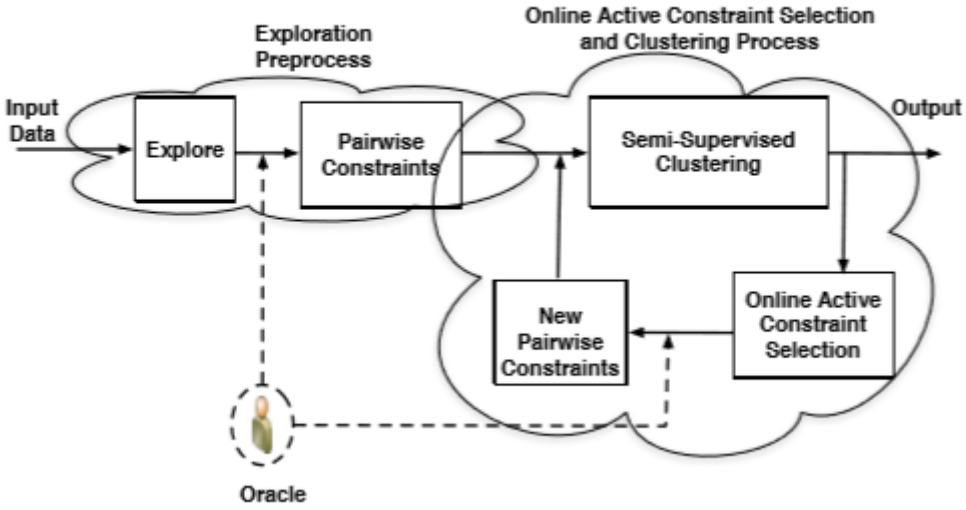
clusters, the proper neighborhood of any random point can be determined within maximum of (k-1) queries [4]. Consolidate therefore adds points to clusters at a faster rate than Explore by a factor of O (k/log k).

4.2 Active Constraint Selection for Semi-Supervised Clustering

In this section, we first present the framework for our active semi supervised clustering, then describe details of each framework component and two novel node uncertainty models/definitions for use inactive constraint selection [5]. Throughout the rest of the paper, let X be a data set with n nodes, X = {x1,x2,⋯ ,xn}and xi ∈ Rd with k total clusters.

4.2.1 Active Semi-Supervised Clustering Framework

We now present our framework for active clustering—recall the basic flow of the algorithm depicted earlier. We divide the whole framework into two parts: the exploration preprocess and the online active selection/clustering process [6]. We begin by running the explorations preprocess once, before computing any clustering results. The goal is to obtain a set of exemplar nodes Q0, with (hopefully but not necessarily) at least one representing each true cluster [7].



4.3 SVM Batch Mode Active Learning

The traditional SVM active learning method employs the notion of version space for measuring the risk in the active learning task [5]. Given the training data L and a Mercer kernel K, the version space is defined as a set of hyper planes that can separate the training data in the feature space HK induced by the Mercer kernel. More formally, the version space can be expressed as,

$$\mathcal{V} = \{f \in \mathcal{H}_K | \forall i \in \{1, \ldots, l\}, \quad y_i f(\mathrm{x}_i) > 0\}.$$

The idea of SVM active learning is to find an optimal unlabeled example that will result in the maximal reduction of the version space. More details can be found in [8]. Although the above idea works well for selecting a single unlabeled [4]. The above optimization is a standard quadratic

programming (QP) problem that can be solved effectively by existing convex optimization software packages [1]. Finally, given the estimated qi, we will select the unlabeled examples with the largest probabilities qi.

4.4 Graph Sampling based Active Semi-Supervised Learning

We now relate the sampling theory developed for graph signals to active semi-supervised learning and propose our solution to the problem. As noted earlier, if the edges of the graph represent similarity between the nodes, then a graph signal defined using the membership functions of a particular class tends to be smooth [2]. We showed how to estimate the sampling cut-off frequency for a set of vertices. In practice, class membership signals are not strictly band limited (see Figure 3). Thus we will be approximating a non-band limited signal with one that is band limited to the cut-off frequency of the chosen vertex set [4]. The key observation in our work is that, even though we cannot recover the "true" membership signal exactly from its samples, an active learning approach should aim at selecting the sampling set with maximum cut-off frequency [5]. This is obviously true since $PW\omega(G) \subset PW\omega 0(G)$ if $\omega \leq \omega 0$ and thus, for any signal, its best approximation with a signal from $PW\omega 0(G)$ can be no worse (in terms of l2 error) than its best approximation with a signal from $PW\omega(G)$. In this setting, predicting the labels of the unknown data points using the labeled data amounts to reconstructing a band limited graph signal from its values on the sampling set [6]. Thus, based on the above reasoning the active learning strategy, given a target number of data points to be labeled, should be to find a set S, with that size, so that the cut-off frequency of S is maximized [7].

4.5 Batch Mode Active Learning

The key of batch mode active learning (BMAL) is to ensure the selected instances of both in formativeness and diversity. BMAL method [3] based on farthest-first traversal (we call it BMAL_FFT) is based on the intuition that for two examples, the larger the distance between them, the smaller redundancy the information they provide. BMAL_FFT works as follows. First, it selects an instance x from CP randomly or according to its uncertainty for the learning model, and adds x to query set Q [5]. Then it selects the next instance xi according to equation (2) and adds xi to Q. BMAL_FFT repeats the above selection procedure until the needed number of instances has been selected. BMAL_FFT is a global search method which may be not efficient for very large-scale text classification problem. In this paper, BMAL_FFT selects instances from CP set, which has a much smaller search space and whose instances are more informative than the whole unlabeled data set U [7].

5. Conclusions

In this paper, we have presented a pairwise constrained clustering framework and a new theoretically well-motivated method for actively selecting good pairwise constraints for semi-supervised clustering. The active learning and pairwise constrained clustering algorithms are both linear and hence suitable for real world clustering applications, as they can be easily scaled to large datasets and can handle very high-dimensional data. The Explore stage of the active learning scheme is currently sensitive to outliers in the data. Note however that it is as sensitive to outliers as the baseline algorithm KMeans. Outlier sensitivity can be handled by density weighted point selection in Explore, where we could have a modified farthest-first traversal that selects distant

points from dense regions of the data space [7]. Such a formulation of active learning would be more robust to outliers, and can be used with more outlier-robust clustering algorithms, e.g., KMedian [8]. .Our paper makes two primary contributions: first, we describe a powerful general framework for online active semi-supervised clustering based on node uncertainty; second, we propose two methods for actively sampling constraints by transforming the pair-uncertainty problem in to an uncertainty problem. We test our online active framework and selection criteria with two different semi-supervised clustering algorithms, against a number of existing active selection methods (including active clustering and active learning techniques), and find our method to be the most effective and robust of those surveyed.

6. Future Work

We test our online active framework and selection criteria with two different semi-supervised clustering algorithms, against a number of existing active selection methods (including active clustering and active learning techniques), and find our method to be the most effective and robust of those surveyed [4]. In the future we hope to explore new node selection criteria. In particular, we wish to examine the possibility of a nonparametric global uncertainty measure, and of a compound uncertainty measure that considers both local and global structure [6].

REFERENCES

[1]. S. Basu, A. Banerjee, and R.J. Mooney, 'Active semi-supervision for pairwise constrained clustering', in *ICDM*, pp. 333–344, (2014).

[2]. S.J. Huang, R. Jin, and Z.H. Zhou, 'Active learning by querying informative and representative examples'. NIPS, (2010)

[3]. S. Basu, M. Bilenko, and R.J. Mooney, 'A probabilistic framework for semi-supervised clustering', in *SIGKDD*, pp. 59–68. ACM, (2013).

[4]. P. Jain and A. Kapoor, 'Active learning for large multi-class problems', in *CVPR*, pp. 762–769. IEEE, (2009).

[5]. P.K. Mallapragada, R. Jin, and A.K. Jain, 'Active query selection for semi-supervised clustering', in *ICPR*, pp. 1–4. IEEE, (2008).

[6]. Q. Xu, M. Desjardins, and K. Wagstaff, 'Active constrained clustering by examining spectral eigenvectors', in *Discovery Science*, pp. 294– 307. Springer, (2005).

[7]. S.Rajan, J. Ghosh, and M.M. Crawford, "An active learning approach to hyper spectral data classification," IEEE Trans. Geosci. Remote Sens., vol. 46, no. 4, pp. 1231–1242, Apr. 2008.

[8]. D. Cohn, Z. Ghahramani, and M. Jordan, "Active Learning with Statistical Models," J. Artificial Intelligence Research, vol. 4, pp. 129- 145, 2016.

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

[10]. 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.

Explore Innovate Educate

Shri Param Hans Education & Research Foundation Trust

www.SPHERT.org

भारतीय भाषा, शिक्षा, साहित्य एवं शोध

ISSN 2321 – 9726

WWW.BHARTIYASHODH.COM

INTERNATIONAL RESEARCH JOURNAL OF MANAGEMENT SCIENCE & TECHNOLOGY

ISSN – 2250 – 1959 (0) 2348 – 9367 (P)

WWW.IRJMST.COM

INTERNATIONAL RESEARCH JOURNAL OF COMMERCE, ARTS AND SCIENCE

ISSN 2319 – 9202

WWW.CASIRJ.COM

INTERNATIONAL RESEARCH JOURNAL OF MANAGEMENT SOCIOLOGY & HUMANITIES

ISSN 2277 – 9809 (0) 2348 - 9359 (P)

WWW.IRJMSH.COM

INTERNATIONAL RESEARCH JOURNAL OF SCIENCE ENGINEERING AND TECHNOLOGY

ISSN 2454-3195 (online)

WWW.RJSET.COM

INTERNATIONAL RESEARCH JOURNAL OF MANAGEMENT SCIENCE AND INNOVATION

WWW.IRJMSI.COM