ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

Enhanced Semi-Supervised Clustering

Vardam Dinesh Dattatray[1], Patale Bhagyashri[2]

P.G. Student, Department of Computer, S.K.N.S.I.T.E.'s Lonawala, Maharashtra, India

Associate Professor, Department of Computer, S.K.N.S.I.T.E.'s Lonawala, Maharashtra, India

ABSTRACT:Semi-supervised clustering uses user supervision in the form of pairwise constraint .in this paper, uses neighbourhood framework. Where uses "labelledexamples" of different clusters according to the pairwise constraints. Select most informative point and query against data point present within neighbourhood. In this way data points are clustered using semi supervised clustering.

KEYWORDS:clustering, semi supervised learning

I. INTRODUCTION

Clustering is nothing but Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups. Three types of clustering techniques are used. Unsupervised, semi supervised, supervised clustering. In case of unsupervised clustering all the data points are known. So very to cluster. In case of supervised clustering data points are not known but total supervision is required for clustering of data points .In semi supervised clustering, combine the features of supervised and unsupervised clustering. It uses user-provided side information to improve the clustering performance. In many learning tasks, unlabeled data is plentiful but labeled data is limited and expensive to generate. Consequently, *semi-supervised learning*, which employs both labeled and unlabeled data.Obtaining pairwise constraints typically requires a user to manually inspect the data points in question, 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 in question 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.

In this paper, consider active learning [4] of constraints in an iterative framework. Specifically, in each iteration we determine what is the most important information toward improving the current clustering model and form queries accordingly. The responses to the queries (i.e., constraints) are then used to update (and improve) the clustering. This process repeats until we reach a satisfactory solution or we reach the maximum number of queries allowed. Such an iterative framework is widely used in active learning for supervised classification and has been generally observed to outperform non iterative methods, where the whole set of queries is selected in a single batch.

In this paper, we are going to use the concept of neighborhood. A neighborhood contains a set of data points that are known to belong to the same cluster according to the constraints and different neighborhoods are known to belong to different clusters. Simply put, neighborhoods can be viewed as containing the "labeled examples" of different clusters. Well-formed neighborhoods can provide valuable information regarding what the underlying clusters look like. Analogous to supervised active learning, an active learner of constraints will then seek to select the most informative data point to include in the neighborhoods. Once a point is selected, we query the selected point against the existing neighborhoods to determine to which neighborhood it belongs.

Here, define the uncertainty in terms of the probability of the point belonging to different known neighborhoods and propose a novel nonparametric approach using random forest for estimating the probabilities. Different from supervised learning where each point only requires one query to obtain its label, in semi-supervised clustering, we can only pose pairwise queries and it typically takes multiple queries to determine the neighborhood of a selected point. In general, points with higher uncertainty will require larger number of queries. This suggests that there is a tradeoff between the amount of information we acquire by querying about a point, and the expected number of queries (cost) for acquiring this information. I propose to balance this tradeoff by normalizing the amount of uncertainty of each data point by the expected number of queries required to resolve this uncertainty, and as such, select queries that have the highest rate of information.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

II. RELATED WORK

1. S. Basu, A. Banerjee, and R. Mooney, "Active Semi-Supervision for Pairwise Constrained Clustering,"

They proposed a two-phase approach [1], 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 neighbourhoods, where c is the total number of clusters. The second phase (consolidate) iteratively expands the neighbourhoods, where in each iteration it selects a random point outside any neighbourhood and queries it against the existing neighbourhoods until a must-link is found.

2. P. Mallapragada, R. Jin, and A. Jain, "Active Query Selection for Semi-Supervised Clustering,"

In previous method, random point is select for query, so it may degrade clustering performance. This author [2] make improvement to previous method. That is called as min-max method which modifies the consolidate phase by choosing the most uncertain point to query. In previous method random point is select to query.

3 R. Huang and W. Lam, "Semi-Supervised Document Clustering via Active Learning with Pairwise Constraints"

They present [3] active learning framework for document clustering. This framework takes an iterative approach. 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 neighbourhoods 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 neighbourhoods until a "must-link" is returned.

3. L. Breiman, "Random Forests,"

They present a method i.e random forest that compute the similarity between a pair of instances by sending them down the decision trees in the random forest and count the number of times they reach the same leaf, normalized by the total number of trees. This will result in a value between 0 and 1, with 0 for no similarity and 1 for maximum similarity.

III. EXISTING METHOD

1.Random

This policy selects random pairwise queries that are not included in or implied by the current set of constraints C. It is not a neighborhood-based approach, and is a commonly used baseline for active learning studies.

2.Min-Max

Min-Max is a neighborhood-based approach that works in two phases. The first phase, Explore, builds c disjoint neighborhoods using farthest-first traversal, where c is the total number of clusters. The second phase, Min-Max, incrementally expands the neighborhoods by selecting a point to query using a distance-based Min-Max criterion.

3.Huang's Method

The language model- based method by Huang and Lam is only applicable to document clustering because it assumes a specific language model for each cluster. This model is used toestimate the probability of each document belonging to different clusters. In our experiments, to apply the underlying active learning method to general-type data, we replace the language with a discriminative approach for estimating the probabilities. In particular, we train a random forest classifier using the cluster labels as classes. We then estimate for each data point x its probability of belonging to different clusters using the out-of-bag probabilistic prediction for x. That is, we consider all decision trees in the random forest that were trained without using the data point x, and use them to estimate the probability of x belonging to different clusters.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

4. This method uses neighborhood approach [6],A neighborhood contains a set of data instances that are known to belong to same class i.e. connected by must –link constraints. Furthermore, different neighborhoods are connected by can-link constraints and thus are known to different classes.

Given a set of existing neighborhoods, we would like to select an instance such that knowing its neighborhood will allow us to gain maximal information about the underlying clustering structure of the data. 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.

IV.PROPOSED SYSTEM

In proposed method, we present two parts .In first part present a neighborhood based framework.

A neighborhood Framework

Contains a set of data instances that are known to belong to the same class (i.e., connected by must-link constraints). Furthermore, different neighborhoods are connected by cannot-link constraints and, thus, are known to belong to different classes. Given a set of constraints denoted by C, we can identify a set of l neighborhood and c is the total number of classes. Consider a graph representation of the data where vertices represent data instances, and edges represent must-link constraints. The neighborhoods are simply the connected components of the graph that have cannot-link constraints between one another. Note that if there exists no cannot-link constraints, we can only identify a single known neighborhood even though we may have multiple connected components because some connected components may belong to the same class. In such cases, will treat the largest connected component as the known neighborhood.

Algorithm 1

Input: A set of data points D; the total number of classesc;
the maximum number of pairwise queries T.
Output: a clustering of D into c clusters.

1. Select a random point and create an initial neighborhood.
2. Repeat steps 3 through 8.
3. Apply semi supervised clustering with given data points and constraints.
4. Select most informative point.
5. Find the probability of most informative point against each given neighborhood.
6. Query most informative point against any point of given neighborhood
7. Update the constraint on returned answer.
8. If must link found, add data point to given neighborhood otherwise create a neighborhood and add the data points in that neighborhood.
9. Stop

Selecting the Most Informative Instance

Given a set of existing neighborhoods, we would like to select an instance such that knowing its neighborhood will allow us to gain maximal information about the underlying clustering structure of the data.

First learn a similarity measure using a random forest-based approach. In particular, leverage the current clustering assignment by creating a labeled training set. Using this training set, build a random forest classifier (containing 50 decision trees) that predicts the cluster label. Random forest is an ensemble learning algorithm that learns a collection of decision trees. Each decision tree is trained using a randomly bootstrapped sample of the training set and the test for

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

each node of the tree is selected from a random subset of the features. While one could also use other supervised classifiers, we choose random forest because it is not prone to over fitting, and as described below it provides a natural definition of similarities (proximities) among instances once a random forest is built.
Given the learned random forest classifier, we compute the similarity between a pair of instances by sending them down the decision trees in the random forest and count the number of times they reach the same leaf, normalized by the total number of trees. This will result in a value between 0 and 1, with 0 for no similarity and 1 for maximum similarity.

Estimating neighborhood probability
Given the similarity matrix M generated by previous steps, let M(xi,xj)denote the similarity between instance xi and instance xj. For any unconstrained data point x, we assume that its probability of belonging to a neighborhood Ni to be proportional to the average similarity between x and the instances in Ni. More formally, we estimate the probability of an instance x belonging to neighborhood N_i as

$$p(x \in N_i) = \frac{\sum_{j=1}^{k=l} \frac{|N_i^k|}{l} \sum^{x_j \in N_i^k} M(x, x_j)}{\frac{|N_i^t|}{l} \sum^{x_j \in N_i^t} M(x, x_j)}, \qquad (1)$$

Finally, we measure the uncertainty of an instance by the entropy of its neighborhood membership.

$$H(\mathcal{N} \mid \mathbf{x}) = -\sum_{i=1}^{l} p(\mathbf{x} \in N_i) \log_2 p(\mathbf{x} \in N_i), \qquad (2)$$

Where l is number of neighborhood.

Normalizing Uncertainty with Expected Cost

We query selected instance against the existing neighborhoods to determine to which neighborhood it belongs. Given a selected data instance, it may take multiple pairwise queries to decide its neighborhood. In particular, we can consider the number of queries required to reach a must-link as the cost associated with each data instance.

The expectation computed by following equation

$$\mathbb{E}[q(\mathbf{x})] = \sum_{i=1}^{l} i * p(\mathbf{x} \in N_i), \qquad (3)$$

Where l is the total number of neighborhood.

However, these two objectives are at odds and we trade off them by selecting the instance that maximizes the ratio between them

$$\mathbf{x}^* = \underset{\mathbf{x} \in \mathcal{U}}{\operatorname{argmax}} \frac{H(\mathcal{N} \mid \mathbf{x})}{\mathbb{E}[q(\mathbf{x})]}, \qquad (4)$$

Algorithm 2 for selection of most informative point

Input: set of data points, cluster assignment and a set of neighborhood.
Output: The most informative data point.
1. Learn the random forest classifier and compute similarity matrix M.
2. Every x belonging to data point D do the following steps.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly Peer Reviewed Journal)

Vol. 5, Issue 1, Januray 2016

3. Find probability of x for given neighborhood
4. Compute entropy using equation num.2
5.Compute expectation using equation num.3
6.Compute most informative point x* using equation num.4
7. Stop

V.CONCLUSION

In this paper, we present an iterative active learning framework to select pairwise constraints for semi-supervised clustering and propose a novel method for selecting the most informative queries. A neighborhood-based approach, and incrementally expands the neighborhoods by posing pairwise queries.

REFERENCES

[1] Sicheng Xiong, Javad Azimi, and Xiaoli Z. Fern," Active Learning of Constraints for Semi-Supervised Clustering".
[2] S. Basu, A. Banerjee, and R. Mooney, "Active Semi-Supervision for Pairwise Constrained Clustering," Proc. SIAM Int'l Conf. Data Mining,pp. 333-344, 2004.
[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] D. Cohn, Z. Ghahramani, and M. Jordan, "Active Learning with Statistical Models," J. Artificial Intelligence Research, vol. 4, pp. 129-145, 1996.
[5] L. Breiman, "Random Forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001
[6]Active Learning of Constraints for Semi-Supervised Clustering,"Sicheng Xiong, Javad Azimi, and Xiaoli Z. Fern"