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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 4, April 2017

A Survey on Semi-supervised Clustering using Neighbourhood Framework

Kolekar Sangita [1], Kamble Priyanka [2], Kadam Sonali[3]

B.E. Student, Department of Computer Engineering, S.S.P.M.'s College of Engineering, Kankavli, India

ABSTRACT: Semi-supervised clustering is combination of supervised clustering and unsupervised clustering. It has an important impact on clustering. This paper introduces a method of clustering based on pair-wise constraints. This method uses neighbourhood framework and select most informative point. By performing the query against all data points, data points are clustered.

KEYWORDS: Clustering, Pair-wise constraints, Neighbourhood framework

I. INTRODUCTION

Clustering is method of creating number of clusters (groups) of large amount of data. The data points that lies in same cluster are similar to each other in terms of their features and the data points that lies in different clusters are dissimilar to each other in terms of their features. So, it is advantageous to have similar data together.

Three types of clustering methods are there: Unsupervised, Supervised, Semi-supervised. In unsupervised clustering, the features of all the data points are already known, according to those features clustering is performed. So it is one of simple method of clustering. In case of supervised clustering, the features of data points are not known, the features need to be extracted before performing clustering. Semi-supervised clustering is the combination of unsupervised clustering and supervised clustering. It employs both labelled and unlabeled data. The features of some data points are known but not of all data points.

In this paper, the neighbourhood concept is used. A neighbourhood is set data points that are belonging to same cluster and different neighbourhoods are belonging to different clusters. The most informative point is selected from the remaining data points. The most informative point is then query against the existing neighbourhoods. For performing query, the pair-wise constraints are used. The pair-wise constraints can be must-link constraint and cannot-link constraint. If the result of the query gives must-link constraint then that point is belonging to that neighbourhood and if it gives cannot-link constraint then that point is query against other neighbourhood. If no must-link found with any of existing neighbourhoods, the data point is considered as a new neighbourhood. But the maximum number of possible neighbourhoods is maximum number of possible clusters to form. The data point which is not belonging to any cluster is called as outlier. This process repeats until reach to solution or reach to maximum number of queries allowed. So, different neighbourhoods shows the different clusters formed of the given large amount data.

In supervised learning each point only requires one query to obtain its label, in semi-supervised clustering, we can only use pair-wise constraints and it typically takes multiple queries to determine the neighbourhood of a selected data point.

II. RELATED WORK

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

In this paper, the Explore and Consolidate approach is used. In this approach, first selects the points using farthest-first-traversal scheme and then iteratively selects the random point other than neighbourhoods and query that point against existing neighbourhoods to find pair-wise constraints.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 4, April 2017

2. P. Mallapragada, R. Jin, and A. Jain, "Active Query Selection for Semi-Supervised Clustering,"
Using random point for query may degrade performance of clustering. So here, the min-max method is used which chooses the most uncertain point to query against the neighbourhoods. So, it improves the performance of clustering.

3 R. Huang and W. Lam, "Semi-Supervised Document Clustering via Active Learning with Pairwise Constraints"
The active learning framework is used for document clustering. This framework uses an iterative approach. Here, for each pair of documents, the probability of them belonging to the same cluster is computed and measures the associated uncertainty. By checking the pair-wise constraints it performs clustering.

III. EXISTING METHOD

Two existing methods of clustering i.e. Supervised clustering and unsupervised clustering.

1] Supervised clustering:

In Supervised clustering, we are given data set and already know what our correct output should look like, having the idea that there is a relationship between the input and the output. In case of supervised clustering data points are known & learning with a teacher's presence as the desired output is known. It is limited to small and simple model.
Eg: given several images of faces and not faces(i.e. labelled data), a supervised clustering will eventually learn and be able to predict whether or not an unseen image is a face.

2]Unsupervised clustering:

In Unsupervised clustering, no idea what our results should look like. We can derive this structure by clustering the data based on relationship among the variables in the data. Unsupervised clustering is known as learning without a teacher's presence as desired output is unknown. With unsupervised clustering it is possible to learn larger and more complex models than with supervised clustering.

IV. PROPOSED METHOD

Proposed method includes two parts i.e. Neighbourhood based framework & Selecting Most Informative Point.

– Neighbourhood based framework:
The data instances that are connected by must-link constraints are belongs to the same class and those which are connected by cannot-link constraints are belongs to the different classes.Given a set of constraints denoted by C, we can identify a set of l neighbourhood 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 neighbourhoods 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 neighbourhood 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 neighbourhood.

Algorithm 1
Input: A set of data points D; the total number of classes c; the maximum number of
Pair-wise queries T.
Output: a clustering of D into c clusters.
1. Select a random point and create an initial neighbourhood.
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 neighbourhood.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 4, April 2017

6. Query most informative point against any point of given neighbourhood.
7. Update the constraint on returned answer.
8. If must link found, add data point to given neighbourhood otherwise create a neighbourhood and add the data points in that neighbourhood.
9. Stop.

– Selecting Most Informative Point:

Given a set of existing neighbourhoods, we would like to select an instance such that knowing its neighbourhood 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 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. The most informative points are selected so that to improve the performance of clustering.

Estimating neighbourhood 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 neighbourhood 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 neighbourhood Ni as:

$$p(\mathbf{x} \in N_i) = \frac{\frac{1}{|N_i|}\sum_{\mathbf{x}_j \in N_i} M(\mathbf{x}, \mathbf{x}_j)}{\sum_{p=1}^{l} \frac{1}{|N_p|}\sum_{\mathbf{x}_j \in N_p} M(\mathbf{x}, \mathbf{x}_j)}, \qquad (1)$$

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

Normalizing Uncertainty with Expected Cost-

We query selected instance against the existing neighbourhoods to determine to which neighbourhood it belongs. Given a selected data instance, it may take multiple pair-wise queries to decide its neighbourhood. 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:

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 4, April 2017

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

Where l is the total number of neighbourhood. 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
Input: set of data points, cluster assignment and a set of neighbourhood.
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.
3. Find probability of x for given neighbourhood
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

We have used the method of semi-supervised clustering using pair-wise constraints and concept of most informative point.

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"