ISSN 2278-3091

Volume 6, No.2, March - April 2017

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse02622017.pdf

NOVEL APPROACH FOR SEMI SUPERVISED CLUSTERING ALGORITHM

R.Selvapriya
PG Scholar in Information Technology,
Hindusthan College of Arts and Science,Coimbatore

ABSTRACT

Semi-supervised clustering process is adopted for the improvement of the clustering performance, by considering the supervision of user in the form of pair wise constraints. This paper studies about the active learning problem of selecting pair wise cannot link and must-link constraints for semi-supervised clustering. The expansion of the neighborhoods is done by the active learning method, by selecting and querying the relationship of the informative points with the neighborhoods. Under this framework, the classic uncertainty based principle is built and a novel approach to evaluate the uncertainty related with each data point is presente. Furthermore, a selection criterion is introduced for the effective management of the amount of uncertainty of each data point with the expected number of queries required to resolve this uncertainty. So, that the selection of the queries that have the highest information rate is allowed. Evaluation of the proposed method on various benchmark data sets is performed. The experimental results demonstrate the consistent and substantial improvements of the proposed technique, with respect to the conventional state-of-the-art techniques.

KEYWORDS

Huang's Method, Min-Max Approach, Normalized Point Based Uncertainty (NPU) Clustering, Random Method Clustering, Semi-Supervised Clustering

1.INTRODUCTION

Semi-Supervised clustering aims to improve clustering performance with the help of user-provided side information. One of the most studied types of side Information is pair wise constraints, which include must-link and cannot-link constraints specifying that two points must or must not belong to the same cluster. A number of previous studies have demonstrated that, in general, such constraints can lead to improved clustering performance. However, if the constraints are select properly, they may also degrade the clustering performance. Moreover, obtaining pair wise 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, we consider active learning 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[1]. 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[2].

2.LITERATURE SURVEY

Semi-Supervised SVM Batch Mode Active Learning for Image Retrieval

Learning with user's interactions is crucial to many applications in computer vision and pattern recognition. One of them is content-based image retrieval (CBIR) where users are often engaged to interact with the CBIR system for improving the retrieval quality. Such an interactive procedure is often known as relevance feedback, where the CBIR system attempts to understand the user's information needs by learning from the feedback examples judged by users. Due to the challenge of the semantic gap, traditional relevance feedback techniques often have to repeat many runs in order to achieve desirable results. To reduce the number of labeled examples required by relevance feedback, one key issue is how to identify the most informative unlabeled examples such that the retrieval performance could be improved most efficiently. Active learning is an important technique to address this challenge.

3.CLUSTERING ENSEMBLES WITH ACTIVE CONSTRAINTS

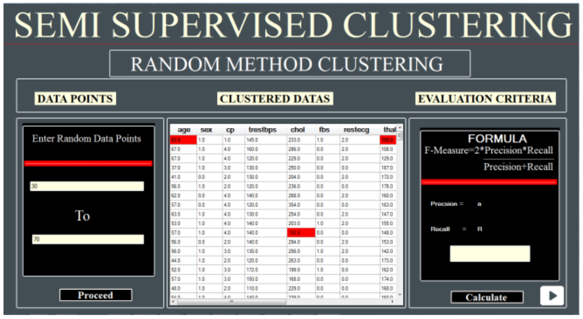
Clustering is the process of discovering homogeneous groups of data according to a given similarity measure. Clustering is well suited for data analysis.However, clustering is susceptible to several difficulties. It is well known that off-the-shelf clustering methods may discover very different structures in a given set of data[1]. This is because each clustering algorithm has its own bias resulting from the optimization of different criteria. Furthermore, there is no ground truth against which the clustering result can be validated. Thus, no crossvalidation technique can be carried out to tune input parameters involved in the clustering process.Similar to k-means, the subspace clustering algorithm LAC depends on the initial choice of centroids. Thus, we make use of the given constraints to achieve a good initialization.

Penta-training leverages the consensus achieved across such partitions to bootstrap and propagate constraintsIn particular, as expected, the largest improvements are achieved when LAC and CLAC have large standard deviations[1].

4.RESEARCH METHODOLOGY

Random Forest Algorithm

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges as. to a limit as the number of trees in the forest becomes large .The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them[3].



SEMI SUPERVISED CLUSTERING

RANDOM METHOD CLUSTERING

DATA POINTS

CLUSTERED DATAS

EVALUATION CRITERIA

Enter Random Data Points

30

To

70

Proceed

age	sex	cp	trestbps	chol	fbs	restecg	thal
63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0
67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0
67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0
37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0
41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0
56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0
62.0	0.0	4.0	140.0	268.0	0.0	2.0	160.0
57.0	0.0	4.0	120.0	354.0	0.0	0.0	163.0
63.0	1.0	4.0	130.0	254.0	0.0	2.0	147.0
53.0	1.0	4.0	140.0	203.0	1.0	2.0	155.0
57.0	1.0	4.0	140.0	192.0	0.0	0.0	148.0
56.0	0.0	2.0	140.0	294.0	0.0	2.0	153.0
56.0	1.0	3.0	130.0	256.0	1.0	2.0	142.0
44.0	1.0	2.0	120.0	263.0	0.0	0.0	173.0
52.0	1.0	3.0	172.0	199.0	1.0	0.0	162.0
57.0	1.0	3.0	150.0	168.0	0.0	0.0	174.0
48.0	1.0	2.0	110.0	229.0	0.0	0.0	168.0
54.0	1.0	4.0	140.0	239.0	0.0	0.0	160.0

FORMULA

F-Measure=2*Precision*Recall / Precision+Recall

Precsion = a

Recall = R

Calculate

Figure 1. Random Method Clusteing

These ideas are also applicable to regression. Significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. In order to grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble.

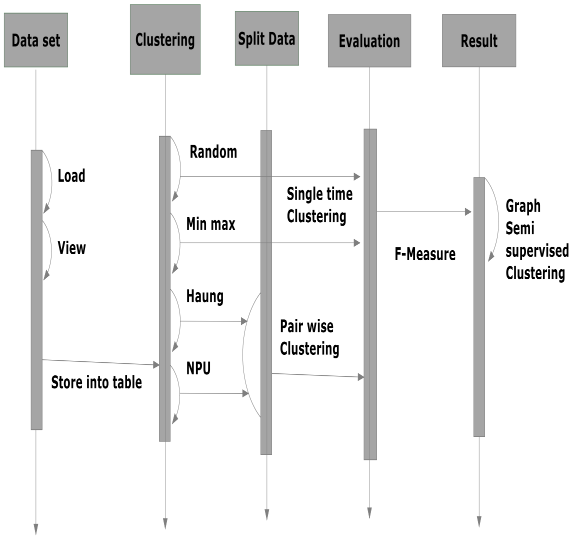
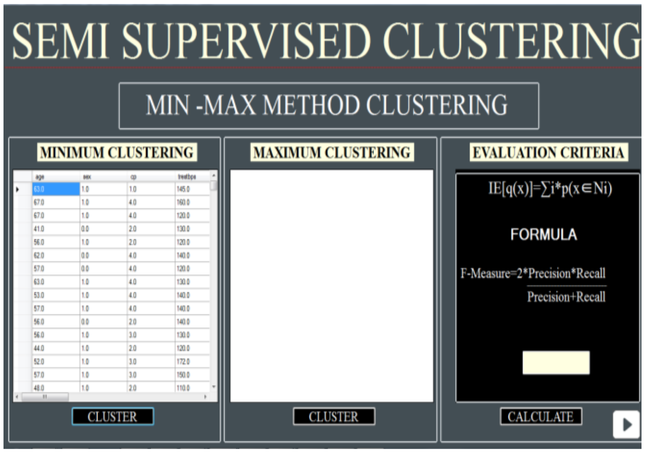


Figure 2.Sequence Diagram

Min max method clustering:

This techniques are very important in data mining and knowledge discovery area as it can be used as basis for most complex and powerful methods[2]. One of these techniques is the Maximum Likelihood and its main goal is to adjust a statistic model with a specific data set, estimating its unknown parameters so the function that can describe all the parameters in the dataset.

In other words, the method will adjust some variables of a statistical model from a dataset or a known distribution, so the model can "describe" each data sample and estimate others. It was realized that clustering can be based on probability models to cover the missing values. This provides insights into when the data should conform to the model and has led to the development of new clustering methods such as Expectation Maximization (EM) that is based on the principle of Maximum Likelihood of unobserved variables in finite mixture models[4].



SEMI SUPERVISED CLUSTERING

MIN -MAX METHOD CLUSTERING

MINIMUM CLUSTERING

age	sex	cp	trestbps
63.0	1.0	1.0	145.0
67.0	1.0	4.0	160.0
67.0	1.0	4.0	120.0
41.0	0.0	2.0	130.0
56.0	1.0	2.0	120.0
62.0	0.0	4.0	140.0
57.0	0.0	4.0	120.0
63.0	1.0	4.0	130.0
53.0	1.0	4.0	140.0
57.0	1.0	4.0	140.0
56.0	0.0	2.0	140.0
56.0	1.0	3.0	130.0
44.0	1.0	2.0	120.0
52.0	1.0	3.0	172.0
57.0	1.0	3.0	150.0
48.0	1.0	2.0	110.0

MAXIMUM CLUSTERING

EVALUATION CRITERIA

IE[q(x)]=∑i*p(x∈Ni)

FORMULA

F-Measure=2*Precision*Recall / Precision+Recall

Figure 3. Min-Max Clust

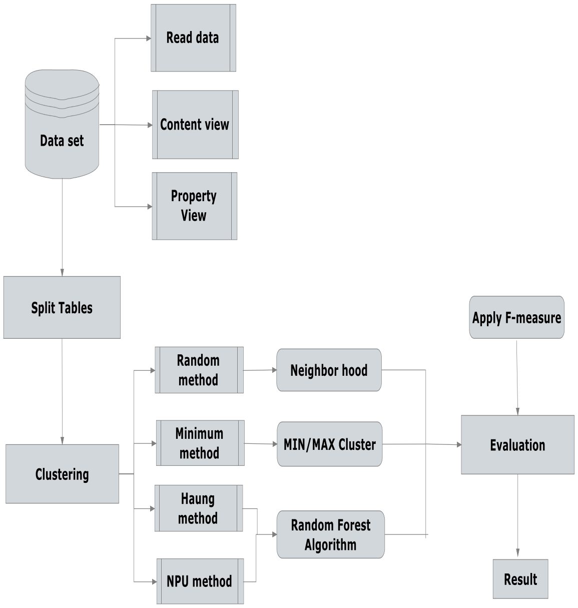


Figure 4.Flow Diagram for Semi Clustering

5.IMPLEMENTATION

Active learning in an iterative manner where in each iteration queries are selected based on the current clustering solution and the existing constraint set. We apply a general framework that builds on the concept of neighborhood, where neighborhoods contain "labeled examples" of different clusters according to the pair wise constraints. Our active learning method expands the neighborhoods by selecting informative points and querying their relationship with the neighborhoods.

Under this framework, we build on the classic uncertainty-based principle and present a novel approach for computing the uncertainty associated with each data point. We further introduce a selection criterion that trades off the amount of uncertainty of each data point with the expected number of queries (the cost) required to resolve this uncertainty. This allows us to select queries that have the highest information rate[5].

One way to interpret the neighborhoods is to view them as the “labeled examples” of the underlying classes because instances belonging to different neighborhoods are guaranteed to have different class labels, and instances of the same neighborhood must belong to the same class. A key advantage of using the neighborhood concepts is that by leveraging the knowledge of the neighborhoods, we can acquire a large number of constraints via a small number of queries.

6 RESULTS AND CONCLUSION

Our method takes a neighborhood-based approach, andincrementally expands the neighborhoods by posing pairwisequeries. We devise an instance-based selection criterionthat identifies in each iteration the best instance toinclude into the existing neighborhoods. The selectioncriterion trades off two factors, the information content ofthe instance, which is measured by the uncertainty aboutwhich neighborhood the instance belongs to; and the cost ofacquiring this information, which is measured by the expected number of queries required to determine its neighborhood.We empirically evaluate the proposed method on theeight benchmark data sets against a number of competingmethods.

The evaluation results indicate that our methodachieves consistent and substantial improvements over itscompetitors.There are a number of interesting directions to extendour work. The iterative framework requires repeatedreclustering of the data with an incrementally growingconstraint set. This can be computationally demanding forlarge data sets. To address this problem, it would beinteresting to consider an incremental semi-supervisedclustering method that updates the existing clusteringsolution based on the neighborhood assignment for thenew point. An alternative way to lower the computationalcost is to reduce the number of iterations by applying abatch approach that selects a set of points to query in eachiteration[6].

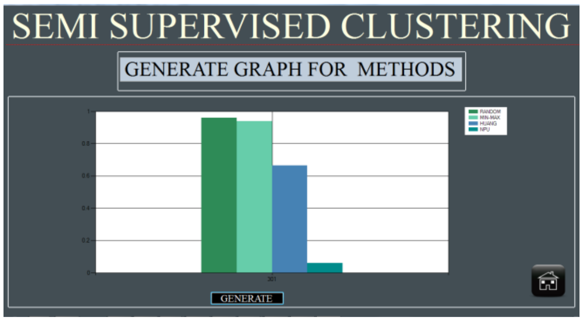


Figure 3. General Graph of all Methods

7 . FUTURE ENHANCEMENT

The iterative framework requires repeated re clustering of the data with an incrementally growing constraint set. This can be computationally demanding for large data sets. To address this problem, it would be interesting to consider an incremental semi-supervised clustering method that updates the existing clustering solution based on the neighborhood assignment for the new poin[4]t. 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. A naive batch active learning approach would be to select the top k points that have the highest normalized uncertainty to query their neighborhoods. However, such a strategy will typically select highly redundant points. Designing a successful batch method requires carefully trading off the value (normalized uncertainty) of the selected points and the diversity among them—a direction that we plan to pursue for future work.

REFERENCES

[1] 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.

[2] S. Basu, I. Davidson, and K. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman & Hall, 2008.

[3] M. Bilenko, S. Basu, and R. Mooney, “Integrating Constraints and Metric Learning in Semi-Supervised Clustering,” Proc. Int’l Conf. Machine Learning, pp. 11-18, 2004.

[4] I. Davidson, K. Wagstaff, and S. Basu, “Measuring Constraint-Set Utility for Partitional Clustering Algorithms,” Proc. 10th European Conf. Principle and Practice of Knowledge Discovery in Databases, pp. 115-126, 2006.

[5] D. Greene and P. Cunningham, “Constraint Selection by Committee: An Ensemble Approach to Identifying Informative Constraints for Semi-Supervised Clustering,” Proc. 18th European Conf. Machine Learning, pp. 140-151, 2007.

[6] D. Cohn, Z. Ghahramani, and M. Jordan, “Active Learning with Statistical Models,” J. Artificial Intelligence Research, vol. 4, pp. 129- 145, 1996.