ISSN:2321-1156

International Journal of Innovative Research in Technology & Science(IJIRTS)

ENHANCED SEMI SUPERVISED CLUSTERING USING BATCH ALGORITHM

R. Thilagavathy [1], C.Vimala[2], C.Sharmila[3]

Department of Information Technology , Jeppiaar Engineering College, Chennai , India.

Abstract: Semi-supervised clustering is a class of supervised clustering that aims to improve clustering performance in the form of pairwise constraints. Pairwise constraints and constraint projections are two popular techniques used in semi-supervised clustering. However, both of them only consider the given constraints and do not consider the neighbours around the data points constrained by the constraints. So, we apply a general framework that builds on the concept of neighbourhood, where it contain "labelled examples" with different clusters according to pairwise clustering. Our active learning method expands the neighbourhood by selecting data points and querying their relationship with the neighbourhood. Under this framework, we present a novel approach for computing the uncertainty associated with each data points. We further introduce to reduce the number of iterations that selects a set of points to query in each iteration.

Keyword: Active Learning, Semi-supervised Clustering, pairwise constraints

1. Introduction

Clustering is a method to segregate unclassified data into a set of meaningful sub-classes. Different methods are followed in the clustering process, one method is based on pairwise uncertainty, which includes must-link and cannot-link constraints specifying that two points must or must not belong to the same cluster. Both a must-link and cannot-link constraint define a relationship between two data instances. This approach is costly as well as time consuming method. In case of improper selection of constraints, the performance may be degraded in the clustering process.

Two methods of active learning of constraints are involved in the supervised clustering namely iterative method and non-iterative method. In the Non-iterative method, the whole set of queries is selected in a single batch. But in the iterative method repeated queries are used to improve the clustering method until reaching a satisfactory solution. Here neighbourhood based framework has been used for clustering of data, in which a set of known data points is found in the same neighbourhood.

Semi supervised clustering method is built on the uncertainty based principle. 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 neighbourhood of a selected point. We iteratively select the next set of queries based on the current clustering assignment to improve the solution. Semi- supervised clustering aims to incorporate the known prior knowledge into the clustering algorithm.

2. Previous Work:

Even though the research on active learning for constraint-based clustering has been limited, it has been studied extensively for supervised Classification. Basuet.al[8] proposed a two phase model referred as "Explore and consolidate" approach , in which incrementally selects points are taken and queried their relationship to identify disjoint neighbourhoods, further the neighbourhoods have been expanded iteratively by selecting random points outside. It gives queries against the existing neighbourhood until the satisfactory solution (must-link) is found.

Q. Xu. [11] Proposed an improvement to Explore and Consolidate (E & C approach) named Min-Max, which it is alters the consolidate phase by choosing the most uncertain point to query (as opposed to randomly).

Huang & Sam[3] have developed the document clustering based on iterative approach .This method considers pairwise uncertainty of the first query but fails to measure the benefit of queries. Even though Huang's method is applied for the purpose of document clustering, We could also go for the approach of active learning to handle other types of data. So our method focus on point-based uncertainty by measuring the total amount of information which is gained by queries of full sequence. Further our method also considers expected number of queries which resolves the uncertainty of the point, which is not taken into account previously.

T. Shi Xu.et.al [5] proposed to check the similarity matrix which is unfortunately limited to two-cluster problems for selecting constraints in dataset. In constraints are selected by analysing the co-association matrix .In order to develop the constraints we iteratively select the next set of queries based on the current clustering assignment

.This current clustering algorithm can be improved most efficiently.

In reference [4] Mallapragada et al described the important application of active learning in the NLP (Natural Language Processing), focus on how to create high quality training sample set. Ambatiet analysed word alignment model in machine translation system, which helps to reduce the data word alignment error rate by creating the half word alignment model combining unsupervised and supervised learning, and makes data concentration abnormal or makes noise sensitive.

Huang and Lam [3], [12] presented the active learning framework for text clustering. It is similar to our work because this framework [12] takes an iterative approach. Semi supervised clustering allows the current constraints set is used to make probabilistic clustering assignments in each iteration. It then 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" constraints is returned.

3. Naïve Neighbourhood Model:

As compared to the existing model, our Neighbourhood framework helps to achieve a good clustering performance. An iterative framework requires repeated re-clustering of the data. This can be computationally demanding of large data set. To solve this problem, by applying a naive batch approach that selects a set of points to query in each iteration. It gives a good result by applying clustering algorithms than traditional clustering algorithms. This Naïve batch algorithm leads to a time consuming procedure.

3.1 Neighbourhood:

A neighbourhood contains a set of data points that are known to belong to the same cluster. It is called as must-link constraints. Different neighbourhood belongs to the different clusters .It is called as cannot-link constraints. Well-formed neighbourhood can provide valuable information concerning what the underlying clusters look like. In supervised clustering, we can pose pairwise queries and it allows multiple queries for selecting a point. For eg: we explain how we form the neighbourhoods from pairwise constraints set.

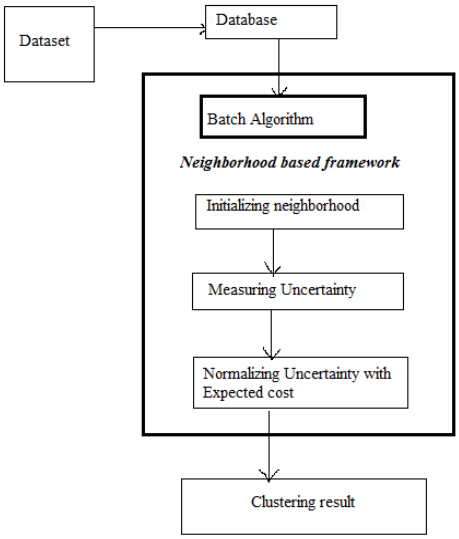
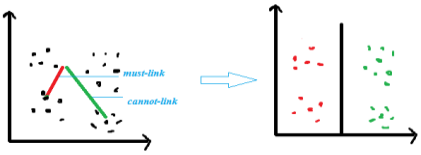


Figure (1) System Architecture



Figure(2) Must-link and cannot link constraints

Figure 2 explains this semi supervised clustering graph diagram allows the users to know about which documents are related to must-link and cannot-link constraints.

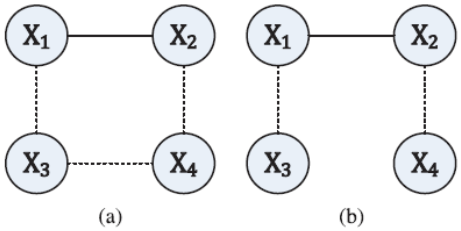


Figure (3) two examples to show how to identify neighbourhood from a set of pairwise constraints

In this Figure 3, the nodes represents data instances, and the solid lines represents must-link constraints while the dashed lines represents the cannot link constraints. Therefore, fig3 (a) denotes three neighbourhoods, {x1; x2}, {x3}, and {x4}, whereas fig3 (b) denotes only two known neighbourhoods, which {x1; x2}; {x3} or {x1; x2}; {x4}.

3.2 Selecting the most informative distance:

In a set of Existing neighbourhood, we would like to select an instance such that knowing its neighbourhood. It will gain more information about the data. If we predict with high certainty to which neighbourhood an instance belongs based on the current understanding of the clustering structure. It will not lead to any gain of information. To apply uncertain based sampling for selecting most informative instance. We present our approach for dealing with two issues.

3.3 Measuring Uncertainty:

To measure uncertainty of each data points belongs to different clusters. For example, we take a model based clustering approach. In this model, User allows to assign a query for selecting a data point. This is sensitive to the current clustering solution. If the clustering cannot be properly by assumption mode, it will not provide correct reliable uncertainty.

3.4 Normalizing uncertainty with expected cost:

If we want to select a particular data point, we can consider the number of queries to reach must-link constraint as accost associated with search instances. If must-link is returned we can stop with only one query. Otherwise we will continue our queries until we will reach our target.

4. Experimental Analysis:

In this section, we outline the detail of our experiment.

4.1 Clustering Evaluation:

We have used two metrics in our experiments. First one is Normalized Mutual Information (NMI). Second one is F-measure.

A.NMI

NMI is used to evaluate the clustering assignment against the class labels. It determines the amount of statistical information by random variables representing the clustering assignments and underlying class labels. It is used to measure the information between the two random variables. If C is the random variables it considering the clustering assignments of the instances and K is considering the class labels on the instances. The NMI is defined as

$$NMI = \frac{2I(C; K)}{H(C) + H(K)}$$

Where,

I(X: Y) = H(X)-H(X/Y), H(X) denotes entropy and H(X/Y) denotes the conditional entropy X given Y.

B. F-measure:

It is defined as the harmonic mean of precision and recall. It is used to evaluate how well we can predict the pairwise relationship between each pair of instances ,for every pair of instances do not explicit constraints between them ,the decision to cluster this pair into same or different cluster is considered to be correct if it matches with the underlying class label available for the instances. Pairwise F-measure is followed by equation

$$F\text{-measure} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where,

Precision is

$$\frac{\#\text{Pair Correctly Predicted In Same Cluster}}{\#\text{Total Pair Predicted In same Cluster}},$$

Recall is

$$\frac{\#\text{Pair Correctly Predicted In Same Cluster}}{\#\text{Total Pair Actually In Same Cluster}}$$

4.2 Dataset:

In our experiments, we use UCI data sets that have been used on constraint based clustering in previous

studies. Our data sets consists breast, pen-based recognition of handwritten digits ecoli, glass identification, statlog-heart, Parkinsons and wine. We removed the smallest three classes for ecoli data set only contain 2, 2, and, 5 instances, respectively. The characteristics of the eight data sets are shown in Figure 4.

Characteristics of the Data Sets

Datasets	# of Classes	# of Features	# 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

Figure (4) Characteristics of data

4.3. Further Result Analysis:

Huang's method is our closest competitor. The difference between our experiment and Huang's method is that our method considers point-base uncertainty whereas Huang's method is based on pairwise uncertainty.

4.3.1. Comparing Huang's method with unnormalized point based uncertainty (UPU):

We consider our method without the normalization step and compare it with Huang's method in Fig 5. The x-axis shows the performance of Huang's method, and the y-axis shows the performance of our method without normalization, i.e., directly using (2) as the selection criterion (referred to as unnormalized point-based uncertainty), both measured in NMI (Normalized mutual information). Each point in the figure corresponds to a particular data set with a particular query size.

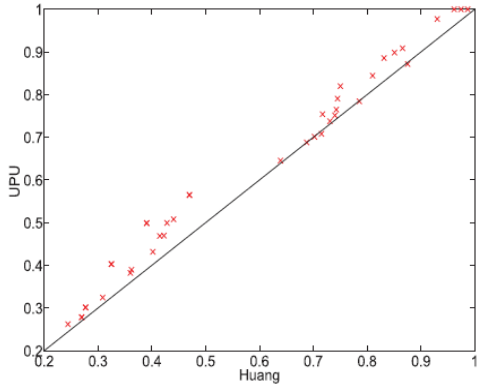


Figure (5) Comparing Haung's method with UPU

5. Conclusion

In this paper, we have presented an active learning framework which is used to select pairwise constraints and introduce a novel method for selecting queries .We take neighborhood approach which discover clusters based on the neighborhood characteristics of data. The results indicates that our method reduce time consumption .This is the main advantage of this model. Our proposed system, a Batch Active Clustering Approach was used to calculate the point wise constraints based on the regions in a semi supervised clustering. Our proposed system balance the 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. Our proposed system focuses on the point-based uncertainty, allowing us to select the queries according to the total amount of information gained by the full sequence of queries as a whole. 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. Computational time should be reduced Query efficiency is very high.

6. Future work:

In the process of solving the problems by using data mining we often encountered cannot be labeled data. To overcome these problems, we introduce a Naïve batch algorithm where in select a set of points to query in each iteration. It is more beneficial and it is mainly used to reduce the computational cost and time consumption.

References

[1]. Y. Guo & d schuurmans, "discrimination batch mode active learning ,"process advances in neural information process system , pp 593-600 , 2008.

[2]. Hoi, R-Jin, J.zhu & M.lyu,"semi Supervised svm batch mode active learn for image retrieval," Proc.Ieee conf. Compute vision and pattern recognition ,"pp 1-7,2008.

[3]. Mallapragada,R.Jin,A.Jain," active query and selection for semi-supervised clustering," proc.Int'l conference and Pattern recognition, pp. 1-4 2008.

[4]. T.Shi, S.Harvath," Supervised Learning With random Forest Predictors, "Computational and Graphical statistics,"pp118-138, 2006.

[5]. O.Shamir and N.Fishy,"Spectral Clustering on a Budget, "Machine Learning Research Proc.Track, vol.15, pp.661-669,2011.

[6]. D, Greene P. Cunnigham,"Constraint Selection by Committee: AN Ensemble Approach to Identifying Informative Constraints for semi-Supervised clustering,"Proc.18th European conf. Machine Learning, pp.140.140-151, 2007.

[7] D. Cohn, Z. Schuurmans, and M.Jordan "Active Learning with Statistical Models," J. Artificial Intelligence Research, Vol.4, pp. 129-145, 1996.

[8] M. –F. Balcan, A. Blum, and A. Gupta. Approximate clustering without the Approximation. In SODA, pages 1065- 1077, 2009.

[9] M. Little, P. McSharry, S. Roberts D. Costello, and I. Moroz, "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection," BioMedical Eng. OnLine, vol. 6, no. 1, p. 23, 2007.

[10] Q. Xu, M. Desjardins, and K. Wagstaff, "Active Constrained Clustering by Examining Spectral Eigenvectors," Proc. Eighth Int'l Conf. Discovery Science, pp. 294-307, 2005.

[11] L. Breiman, "Random Forests,"Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[12] P. Sen, G. M. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-ad.Collective classi_cation in network data. AI Magazine, 29(3):93-106, 2008.

[13] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time. In STOC, pages 296-305, 2001.

[14] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling inprivate data analysis. In STOC, pages 75-84, 2007

8. Bibliographies:

First author – C.Vimala is currently doing her Final year **B. TECH Information Technology** at **Jeppiaar Engineering College** in Chennai, Tamil Nadu .She has presented a paper based on cloud computing in national level technical symposium and has attended many workshops.

Second author – C.Sharmila is currently doing her Final year **B. TECH Information Technology** at **Jeppiaar Engineering College** in Chennai, Tamil Nadu .She has presented a paper based on cloud computing in national level technical symposium and has attended many workshops.