International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 12 4184 - 4188

An Efficient Learning of Constraints For Semi-Supervised Clustering using Neighbour Clustering Algorithm

T.Saranya
Research Scholar
Snr sons college
Coimbatore, Tamilnadu
saran2585@gmail.com

Dr. K.Maheswari
Associate Professor
Snr sons college
Coimbatore, Tamilnadu
maheswarisnr@gmail.com

***Abstract*—** Data mining is the process of finding the previously unknown and potentially interesting patterns and relation in database. Data mining is the step in the knowledge discovery in database process (KDD) .The structures that are the outcome of the data mining process must meet certain condition so that these can be considered as knowledge. These conditions are validity, understandability, utility, novelty, interestingness. Researcher identifies two fundamental goals of data mining: prediction and description.

The proposed research work suggests the semi-supervised clustering problem where to know (with varying degree of certainty) that some sample pairs are (or are not) in the same class. A probabilistic model for semi-supervised clustering based on Shared Semi-supervised Neighbor clustering (SSNC) that provides a principled framework for incorporating supervision into prototype-based clustering. Semi-supervised clustering that combines the constraint-based and fitness-based approaches in a unified model. The proposed method first divides the Constraint-sensitive assignment of instances to clusters, where points are assigned to clusters so that the overall distortion of the points from the cluster centroids is minimized, while a minimum number of must-link and cannot-link constraints are violated. Experimental results across UCL Machine learning semi-supervised dataset results show that the proposed method has higher F-Measures than many existing Semi-Supervised Clustering methods.

Keywords-*k-means clustering;neighbourhood; dataset;centroids*

*****

I. INTRODUCTION

Data Mining, "The Extraction of hidden predictive information from large databases", is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve [15]. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

The main goal of data mining is to provide a comprehensive review of different clustering techniques. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept.

1.1.1 The Scope of Data Mining

Data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviours [16]**. Data Mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data Mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- **Automated discovery of previously unknown patterns [16]**. Data Mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

1.1.2 Architecture for Data Mining

Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on.

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse [17]. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure.

The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure.

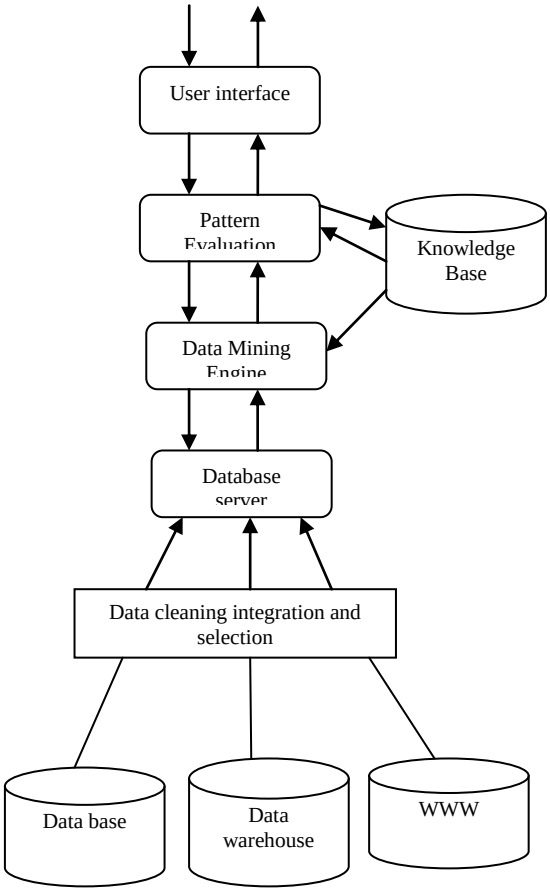


Figure 1. Integrated data mining architecture

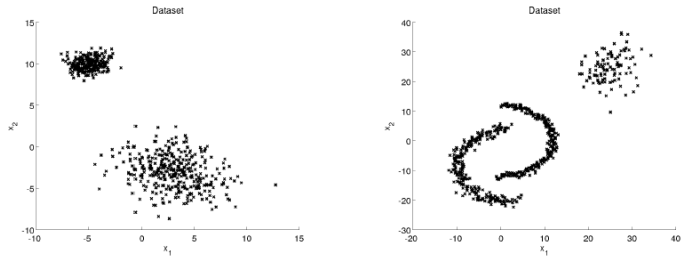
1.3 Clustering

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups.

1.4 K – means Clustering

A well known general clustering procedure is the k-means algorithm [22]. This is most often implemented with the Euclidean distance. Given a set of cluster representatives, in the first step each point is assigned to its closest representative. The second step updates the cluster representatives by setting them equal to the mean of the data vectors assigned to them in the previous step.

The k-means algorithm is optimal for clustering dense, spherically shaped and linearly separable clusters [16]. Fig. 2 (a) shows an example of such a situation, and the k-means algorithm can be expected to give a good clustering results.



(a) Dataset 1. (b) Dataset 2.

Figure 2. Two datsets.

This is not the case if we assume that in addition to the points in the upper right corner, each of the two half circles in fig.2 should be different clusters. Clearly, while it is easy to separate the corner cluster from the other two, it is not possible to define a straight line that separates the two half circles.

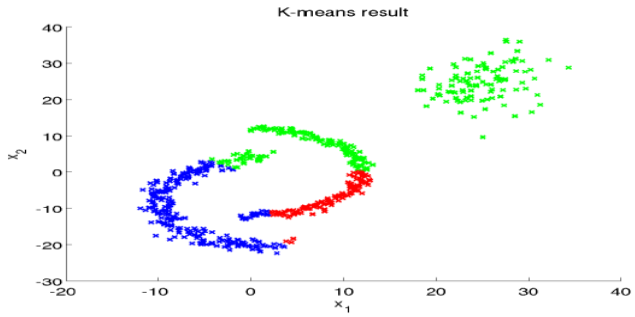


Figure 3. K-means solution

Figure 3 shows the clustering result of the k-means algorithm for three clusters. This is not a good result, but is it possible solve this without defining an ad hoc decision curve between the two half circles? This is where spectral clustering methods excel. By performing a nonlinear data transformation, the data is transformed to some space where it is easier to separate between the clusters.

II. RELATED WORK

S. Basu, et.al (2004**)** [1] proposed a semi-supervised clustering uses a small amount of supervised data to aid unsupervised learning. This paper presents a pairwise constrained clustering framework and a new method for actively selecting informative pairwise constraints to get improved clustering performance.

M. Bilenko, et.al (2004) [2] proposed a semi-supervised clustering employs a small amount of labeled data to aid unsupervised learning. Experimental results demonstrate that the unified approach produces better clusters than both individual approaches as well as previously proposed semi-supervised clustering algorithms.

I. Davidson et.al. **(2006)** [3] Proposed the Clustering with constraints is an active area of machine learning and data mining research.

D. Greene and P. Cunningham, (2007) [4] proposed the number of clustering algorithms have been proposed for use in tasks where a limited degree of supervision is available.D. Cohn, et.al (1996) [5] discussed for many types of machine learning algorithms, one can compute the statistically "optimal" way to select training data.

Y. Guo and D. Schuurmans, (2008) [6] proposed active learning sequentially selects unlabeled instances to label with

the goal of reducing the effort needed to learn a good classifier.

III. METHODOLOGY

A. A Neighbourhood based Framework

A neighborhood 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* neighbourhoods N = {*N1*, …,*Nl*}, such that $l \leq c$ 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, which are denoted by *Ni D.* i ϵ {*1*, …, *l*}, 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, we will treat the largest connected component as the known neighborhood.

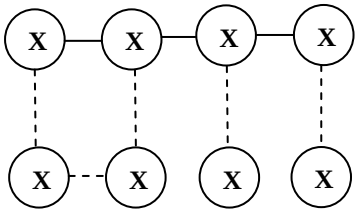


Figure 4. Two examples to show how to identify neighborhoods from a set of pairwaise constraints

B. K Means Clustering Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume *k* clusters) fixed a priori. The main idea is to define *k* centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point re-calculate the k new centroids as centers of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop k centroids change their location step by step until no more changes are done. The K means is one of the simplest clustering technique and it is commonly used in medical imaging, biometrics and related fields.

The algorithm composed of the following steps:

1. Place *K* points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroids.
3. When all the objects have been assigned, recalculate the position of the K centroids.
4. Repeat step 2 and 3 until the centroid no longer move. This produce a separation of the object into groups from which metric o be minimized can be calculated.

In this research work optimized Shared-Neighbour cluster algorithm is used for detecting outlier data points. When employing the outlier approach with help of kernel mapping are extracted from an data collection and clustered typically using Neighbour cluster.

C. Data sets

The Active learning of constraints based Shared Neighbor clustering and Fitness Mapping (kernel mapping) method uses eight benchmark UCI data sets that have been used in previous studies on constraint based clustering [1], [4].

Out data sets include breast [22], pen-based recognition of handwritten digits (3, 8, 9), ecoli, glass identification, statlog-heart, parkinsons [23], statlog image segmentation, and wine. For the ecoli data set, we removed the smallest three classes, which only contain 2, 2, and, 5 instances, respectively. The characteristics of the eight data sets are shown in Table 1.

TABLE I. CHARACTERISTICS OF THE DATASETS

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

IV. EXPERIMENT RESULT

There were two different experimental setups. In the first setup, a single data set was clustered for many different K-s (number of clusters), to see if there is any difference when the number of clusters is varied. In the second setup, 8 different data sets were all clustered by the number of classes in the data (the number of different labels).

Two evaluation criteria are used in our experiments. First, we use normalized mutual information (NMI) to evaluate the clustering assignments against the ground-truth class labels [21].

TABLE II. COMPARISON ON F-MEASURE ON ECOLI DATSETS

K	**20**	**40**	**60**	**80**	**100**
Random	0.63	0.74	0.70	0.65	0.65
MinMax	0.64	0.77	0.83	0.85	0.85
Huang	0.68	0.76	0.80	0.833	0.82
NPU	0.67	0.76	0.80	0.833	0.829
SSNC	0.69	0.81	0.86	0.88	0.86

The above table shows the quality measurement of clustering on Ecoli dataset . It shows the performance measures for different methods using varied range of K values.

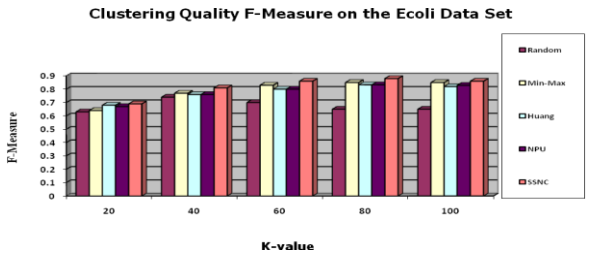


Figure 5. Clustering Quality F-Measure on Ecoli Datasets.

TABLE III. COMPARISON ON F-MEASURE ON GLASS IDENTIFICATION DATASET

K	20	40	60	80	100
Random	0.44	0.403	0.41	0.41	0.413
MinMax	0.432	0.418	0.463	0.484	0.493
Huang	0.480	0.481	0.476	0.474	0.473
NPU	0.493	0.492	0.481	0.496	0.495
SSNC	0.51	0.52	0.499	0.59	0.61

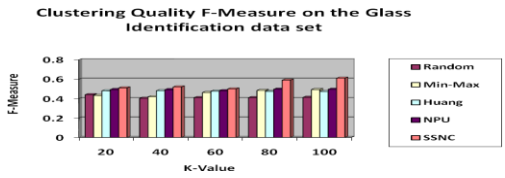


Figure 6. Clustering Quality F-Measure on Glass Identification Dataset.

TABLE IV. COMPARISON ON F-MEASURE ON IRIS DATSET

K	20	40	60	80	100
Random	0.59	0.60	0.63	0.65	0.68
MinMax	0.593	0.615	0.666	0.70	0.74
Huang	0.59	0.60	0.65	0.69	0.73
NPU	0.68	0.72	0.76	0.75	0.79
SSNC	0.69	0.75	0.77	0.76	0.81

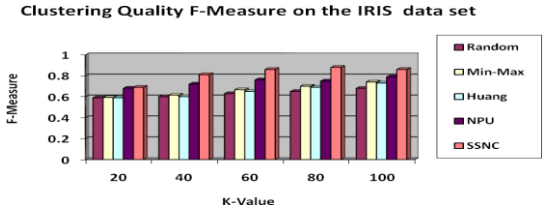


Figure 7. Clustering Quality F-Measure on IRIS Dataset.

V. CONCLUSION

An iterative active learning framework with shared neighbour cluster is to select pairwise constraints for semi-supervised kernels clustering and propose a novel method for selecting the most informative queries. The proposed Shared Semi-supervised Neighbor clustering (SSNC) method had proven to be more robust than the Active learning of iterative frame work baseline on both synthetic and real-world data, as well as in the presence of high levels of artificially introduced noise.

The proposed method takes a shared neighborhood-based approach, and incrementally expands the neighborhoods by posing pairwise queries. We devise an instance-based kernel selection criterion that identifies in each iteration the best instance to include into the existing neighborhoods. 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.

The kernel mapping with shared neighbor clustering can easily be extended to incorporate additional pair-wise constrains such as requiring points with the same label to come into view in the same cluster with just an extra layer of function hubs. The model is flexible enough for information other than explicit constraints such as two points being in different clusters or even higher-order constraints (e.g., two of three points must be in the same cluster).

REFERENCES

[1] S. Basu, A. Banerjee, and R. Mooney, "Active Semi-Supervision for Pair-wise Constrained Clustering," Proc. SIAM Int'l Conf. Data Mining, pp. 333-344, 2004.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Semi-Supervised SVM Batch Mode Active Learning for Image Retrieval," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-7, 2008.

[9] S. Huang, R. Jin, and Z. Zhou, "Active Learning by Querying Informative and Representative Examples," Proc. Advances in Neural Information Processing Systems, pp. 892-900, 2010.

[10] 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.

[11] P. Mallapragada, R. Jin, and A. Jain, "Active Query Selection for Semi-Supervised Clustering," Proc. Int'l Conf. Pattern Recognition, pp. 1-4, 2008.

[12] 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.

[13] M. Al-Razgan and C. Domeniconi, "Clustering Ensembles with Active Constraints," Applications of Supervised and Unsupervised Ensemble Methods, pp. 175-189, Springer, 2009.

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

[15] An Introduction to Data Mining (Discovering hidden value in your data warehouse) see http://www.thearling.com/text/dmwhite/dmwhite.htm

[16] Doug Alexander, "Data Mining" see http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

[17] Hyperion, "The Role of the OLAP Server in a Data Warehousing Solution".