Affinity Propagation with Background Knowledge using Pairwise Constraints

Saravanakumar.R[#1], Dr.C.Nandini[*2],
[#]*Associate Professor, Department of CSE, Dayananda Sagar Academy of Tech.,& Mang., Bangalore, India.*
[*]*Professor, Department of CSE, Dayananda Sagar Academy of Tech.,& Mang., Bangalore, India*

Abstract:

Data mining is the process of identifying and extracting hidden patterns and information from large databases and warehouses. Incorporating pairwise constraints into clustering algorithms is an emerging research area for machine learning and data mining communities. Already various algorithms exist to combine relative similarities between clusters from different viewpoints. But they suffer from duplicates in clusters and also lesser relevancy. The proposed Affinity propagation clustering algorithm uses semi-supervised learning to avoid data redundancy from input strings and ensures quicker retrieval. Final Clusters contain unique and relevant data. Semi-supervised learning falls between unsupervised learning (without any label training data) and supervised learning (with completely labelled training data). Thus the hybrid algorithms provides performance enhancement over its existing counterparts. Further large amount of input data can be processed precisely and even various alternative forms of similar output data can be retrieved. Hence the highest degree of accuracy can be achieved in clustering data and retrieval of the same by the improved affinity propagation algorithm.

Keywords:

Affinity propagation, clusters, pairwise constraints, semi-supervised learning.

I. INTRODUCTION

Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories. It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high performance computing.

A knowledge discovery process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation. Data mining functionalities include the discovery of concept/class descriptions (i.e., characterization and discrimination), association, classification, prediction, clustering, trend analysis, deviation analysis, and similarity analysis.

Characterization and discrimination are forms of data summarization.

Interesting data patterns can also be extracted from other kinds of information repositories, including spatial, time-related, text, multimedia, and legacy databases, and the World-Wide Web.

Clustering techniques consider data tuples as objects. They partition the objects into groups or clusters, so that objects within a cluster are similar to one another and dissimilar to objects in other clusters. Similarity is commonly defined in terms of how close the objects are in space, based on a distance function. The quality of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster. Centroid distance is an alternative measure of cluster quality, and is defined as the average distance of each cluster object from the cluster centroid.

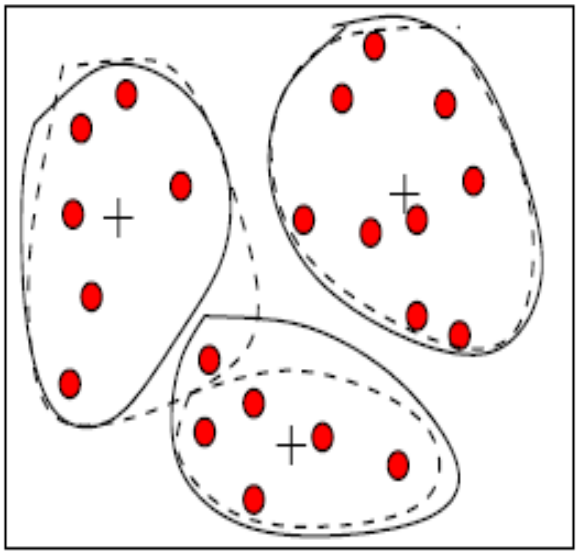

Figure 1 Clusters with centroid

Each cluster centroid is marked with a '+'. Clustering, which aims to efficiently organize the dataset, is an old problem in machine learning and data mining community. Most of the traditional clustering algorithms aim at clustering homogeneous data. However, in many real world applications, the data set to be analysed involves more than one type.

Data clustering, the unsupervised classification of samples into groups is an important research area in machine learning for several decades. A large number of algorithms have been developed for data clustering, including the k-means algorithm, mixture models, and spectral clustering. More recently, maximum margin clustering was proposed for data clustering and has shown promising performance.

The key idea of maximum margin clustering is to extend the theory of support vector machine to unsupervised learning. However, despite its success, the following three major problems with maximum margin clustering have prevented it from being applied to real-world applications. Clustering is an unsupervised learning problem, which tries to group a set of points into clusters such that points in the same cluster are more similar to each other than points in different clusters, under a particular similarity metric. In the most general formulation, the number of clusters k is also considered to be an unknown parameter. Such a clustering formulation is called a "model selection" framework, since it has to choose the best value of k under which the clustering model fits the data.

In the discriminative clustering setting, the clustering algorithm tries to cluster the data so as to maximize within-cluster similarity and minimize between-cluster similarity based on a particular similarity metric, where it is not necessary to consider an underlying parametric data generation model. In both the generative and discriminative models, clustering algorithms are generally for optimizing problems and solved by iterative methods.

II. RELATED WORK

Clustering can be done with several techniques and algorithms. After the clustering process, the obtained clusters are represented with examples, which can include all or part of the features that appear in the cluster members. During cluster-based query processing, only those clusters that contain examples similar to the query are considered for further comparisons with cluster members, e.g., documents that are stored already and not the newly made documents.

Existing cluster analysis methods are either targeting attribute data or relationship data. The systems are interested in incorporating the pairwise constraints into the recently proposed maximum margin clustering. MMC utilizes the maximum margin principle adopted in the supervised learning and tries to find the hyper planes that partition the data into different clusters with the largest margins between them over all the possible labelling.

These systems with any of the fast variants will always provide sensible clustering solutions. Therefore, one can evaluate the complex system and adjust critical system parameters without having to worry for dependence of system performance on the clustering method employed but in case similarity constraints they lead to complex problems and thus results in redundancy.

III. SYSTEM ANALYSIS

In the system implementation, semi-supervised learning has captured a great deal of attentions. Semi-supervised learning is a machine learning paradigm in which the model is constructed using both labelled and unlabelled data for training typically a small amount of labelled data and a large amount of unlabelled data. In this proposed system it retrieve the data from training data or labelled data and extract the feature of the data and compare with labelled data and unlabelled data.

It reduce the human work that need not train all data in the label data it occupy less memory this method user to make an accurate clustering. In clustering process, semi-supervised learning is a class of machine learning techniques that make use of both labelled and unlabelled data for training - typically a small amount of labelled data with a large amount of unlabelled data.

Semi-supervised learning falls between unsupervised learning (without any labelled training data) and supervised learning (with completely labelled training data). Many machine-learning researchers have found that unlabelled data, when used in conjunction with a small amount of labelled data, can produce considerable improvement in learning accuracy.

Pairwise constraints clustering contain the word comparison for more effective clustering. The semi supervised algorithm overcomes the untrained data process. Another drawback of the existing clustering is word pair process. Same meaning words come in different words example computer, computers, computing all belongs to the computer word. In existing system using K-mean the document will make cluster in over lapping or in different cluster that not relevant to any cluster groups. In this proposed system we overcome over lapping problems and avoid miss similar cluster problem too.

In particular, the cluster assignments that are inconsistent with the constraints are excluded from the partition function when computing the posterior probability for the cluster memberships. One problem with treating the side information as hard constraints is that we may not be able to find feasible solutions that are consistent with all the constraints.

To overcome this problem, a number of studies view the side information as soft constraints. The key idea is to penalize, not to exclude, the cluster assignments that are inconsistent with the given pairwise constraints. Many present probabilistic models for semi-supervised clustering where the pairwise constraints are incorporated into the clustering algorithms through the Bayesian priors.

Another approach to semi-supervised clustering is to first learn a distance metric from the given pairwise constraints. The pairwise similarity between any two examples is then computed based on the affinity, and a maximum margin clustering algorithm is applied to the computed similarity matrix. The key to this approach is to effectively learn a distance metric from the side information. Finally, a few studies cluster data points by a similarity matrix that is directly modified according to the pair wise constraints.

IV. ARCHITECTURE

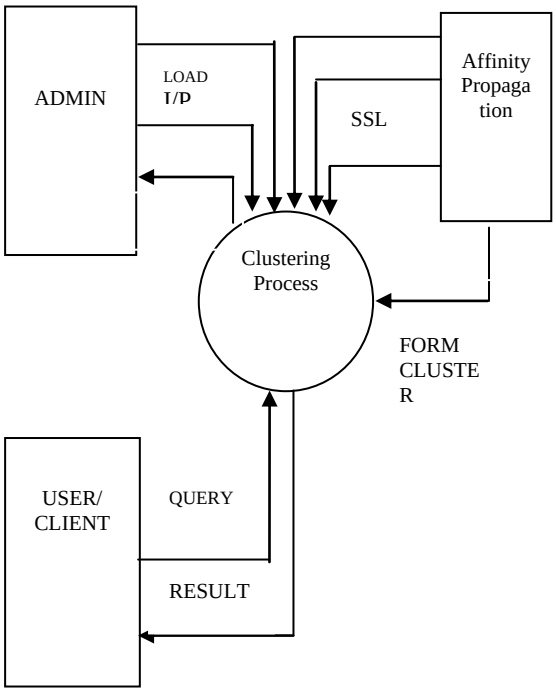

Figure 2 Architecture

The architecture diagram gives a brief overview of the system at a very high level from an end-user's point of view. It is intended for target audience who may not be interested in knowing the underlying system details. It is defined as the process of a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system. The Architectural Design for the system is shown in the Figure 2.

In the client side the user/client gives a query to search and obtains the result for it. The administrator side after getting the query creates a training set and loads the input data. In order to get the clustered data i.e. related data is obtained through clustering process that takes place in the server side.

Clustering digital objects (e.g., text documents) by identifying a subset of representative examples plays an important role in recent text mining and information retrieval research. During cluster-based query processing, only those clusters that contain examples similar to the query are considered for further comparisons with cluster members, e.g., documents. This strategy, sometimes called Cluster-Based Retrieval.

The Affinity Propagation Algorithm is used to cluster the data. If a user gives a keyword to search only through this algorithm the cluster is made and provided to the user. Semi-supervised learning is a machine learning paradigm in which the model is constructed using both labelled and unlabelled data for training typically a small amount of labelled data and a large amount of unlabelled data.

Semi-supervised clustering is to first learn a distance metric from the given pairwise constraints. The semi-supervised is applied on the data to be cluster so that it matches the similarities of the existing data and avoid pairwise constraints problem.

Final Clusters contain unique and relevant data. Large amount of input data can be processed precisely and even various alternative forms of similar output data can be retrieved. Based on the client request or the phrase entered and the administrator loaded data the data is made to the cluster that it exactly matches on the similar constraints by applying affinity propagation clustering. The system overcomes over lapping problems and avoids miss similar cluster problem too.

V. CONCLUSION

Existing clustering and data retrieval methods lack accuracy to a large extent. Further they suffer from duplicates and take longer processing time. The proposed Pairwise constraints affinity propagation algorithm which uses semi-supervised learning will overcome the entire drawback and achieve maximum accuracy in clustering and data retrieval. Further loss can provide more robust penalization for the pairwise constraints. The proposed constrained algorithm will effectively improve the baseline MMC. It will also outperform typical semi-supervised clustering counterparts in terms of accuracy and efficiency. Thus the clustering of data and retrieval can be predicted using the proposed algorithm.

REFERENCES

[1] Antonio Augusto Chaves, Luiz Antonio Nogueira Lorena, 'Clustering Search Algorithm for the Capacitated Centered Clustering Problem'.

[2] Basu.S, Bilenko.M, and Mooney.RJ, (2004) 'A Probabilistic Framework for Semi-Supervised Clustering,' Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 59-68.

[3] Basu.S, Banerjee.A, and Mooney.R.J, (2004) 'Acive Semi-Supervision for Pairwise Constraints,' ICDM 04.

[4] Bilenko.M, Basu.S, and MooneyR.J, (2004) 'Integrating Constraints and Metric Learning in Semi-Supervised Clustering,' Proc. Int'l Conf. Machine Learning, pp. 81-88.

[5] Dhillon.I.S, Mallela.S, and ModhaD.S, (2003) 'Information-Theoretic Co-Clustering,' Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98.

[6] Ding.C, He.X, Zha.H, Gu.M, and Simon.H, (2001) 'A Min-Max Cut Algorithm for Graph Partitioning and Data

Clustering,' Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 107-114.

[7] Ester.M, Ge.R, Gao.B.J, Hu.Z, and Ben-Moshe.B, (2006) 'Joint Cluster Analysis of Attribute Data and Relationship Data: The Connected K-Center Problem,' Proc. SIAM Int'l Conf. Data Mining, pp. 25-46.

[8] Hoi.S.C.H, Liu.W, Lyu.M.R, and Ma.W.Y, (2006) 'Learning Distance Metrics with Contextual Constraints for Image Retrieval,' Proc. IEEE CS Conf. Computer Vision and Pattern Recognition, pp. 2072-2078.

[9] Kulis.B, Basu.S, Dhillon.I, and Mooney.R, (2005) "Semi-Supervised Graph Clustering: A Kernel Approach," Proc. Int'l Conf. Machine Learning, pp. 457-464.

[10] Law.M, Topchy.A, and Jain.A.K, (2005) 'Model-Based Clustering with Probabilistic Constraints,' Proc. SIAM Int'l Conf. Data Mining, pp. 641-645.

[11] Wagstaff.k, Cardie.C, and Schroedl.S, (2001) 'Constrained K-Means Clustering with Background Knowledge,' Proc. Int'l Conf. Machine Learning, pp. 577-584.

[12] Xing.E.P, Ng. A.Y,Jordan. M.I, and Russell.S, (2003) 'Distance Metric Learning with Application to Clustering with Side-Information,' Advances in Neural Information Processing Systems, vol. 15, pp. 521-528.

[13] Zha.H, He.X, Ding.C, Simon.H, and Gu.M, (2001) 'Spectral Relaxation for K-Means Clustering,' Proc. Neural Info. Processing Systems (NIPS), pp. 1057-1064.