IJRSET NOVEMBER 2017 Volume 4, Issue 11 Pages: 10-18

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

ENERGETIC KNOWLEDGE OF LIMITATION FOR SEMI-SUPERVISED GATHERING

[1] R. BHARATHI, [2] A. INDHUMATHI
[1] Assistant Professor, [2] Research Scholar,
[1, 2] Department of Computer Science,
[1, 2] Sengunthar College of Arts and Science,
[1, 2] Tiruchengode,Tamilnadu.

ABSTRACT: The aim of Semi-supervised clustering algorithm is to improve the clustering performance by considering the user supervision based on the pairwise constraints. In this paper, we examine the active learning challenges to choose the pairwise must-link and cannot-link constraints for semi-supervised clustering. The proposed active learning approach increases the neighborhoods based on selecting the informative points and querying their relationship among the neighborhoods. Here, the classic uncertainty-based principle is designed and novel approach is presented for calculating the uncertainty associated with each data point. Further, a selection criterion is introduced that trades off the amount of uncertainty of each data point with the probable number of queries (the cost) essential to determine this uncertainty. This permits us to select queries that have the maximum information rate. The proposed method is evaluated on the benchmark data sets and the results shows that the proposed system yields better outputs over the current state of the art.This paper describes about the methodology to effectively choose pairwise queries to produce an accurate clustering assignment. Through active learning, the number of queries is reduced to achieve a good clustering performance. We view this as an iterative process such that the decision for selecting queries should depend on what has been learned from all the previously formulated queries. In this section, we will introduce our proposed methodology

KEY WORDS: [content-based image retrieval (CBIR),Must-Link (ML), Cannot-Link (CL), Pairwise-Constrained Clustering (PCC), Multi-Dimensional Scaling (MDS), Graphical User Interface (GUI), Self-OrganizingMaps (SOM), Generative Topographic Mapping(GTM), Normalized Mutual Information (NMI), World Wide Web (WWW).]

1. INTRODUCTION

Now a day people come across a huge amount of information and store or represent it as data. One of the vital means in dealing with these data is to classify or group them into a set of segment or clusters. Clustering involves creating groups of objects which are similar, and those that are dissimilar. The clustering problem lies in finding groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function. Clustering is especially useful for organizing documents, to improve retrieval and support browsing. Clustering is often confused with classification, but there is some difference between the two. In classification, the objects are assigned to pre-

defined classes, whereas in clustering the classes are also to be defined. To be Precise, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database system the numbers of disk accesses are to be minimized. In clustering, objects having similar properties are placed in one class, and a single access to the disk makes the entire class available. Clustering algorithms can be applied in many areas, for instance, marketing, biology, libraries, insurance, city-planning, earthquakes, and www document classification.

In many data mining and machine learning tasks, there is a large supply of unlabeled data but limited labeled data, since labeled data can be expensive to generate. Consequently, semi-supervised learning, learning from a combination of both labeled and unlabeled data, has become a topic of significant recent interest. More specifically, semisupervised clustering, the use of class labels or pairwise constraints on some examples to aid unsupervised clustering, has been the focus of several recent projects. In a semi-supervised clustering setting, the focus is on clustering large amounts of unlabeled data in the presence of a small amount of supervised data. In this setting, we consider a framework that has pairwise must-link and cannot link constraints between points in a dataset (with an associated cost of violating each constraint), in addition to having distances between the points. These constraints specify that two examples must be in the same cluster (must-link) or different clusters.

In real-world unsupervised learning tasks, e.g., clustering for speaker identification in a conversation, visual correspondence in multitier image processing, clustering multi-spectral information from Mars images, etc., considering supervision in the form of constraints is generally more practical than providing class labels, since true labels may be unknown a priori, while it can be easier to specify whether pairs of points belong to the same cluster or different clusters.

Related work

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.

S. Basu, A. Banerjee, and R. Mooney, "Active Semi-Supervision for Pairwise Constrained Clustering," the Explore and Consolidate approach is used. In this approach, first selects the points using farthestfirst- traversal scheme and then iteratively selects the random point other than neighbourhoods and query that point against existing neighbourhoods to find pair-wise constraints.

Semi-supervised clustering uses a small amount of supervised data to aid unsupervised learning. One typical approach specifies a limited number of must-link and cannot link constraints between pairs of examples. This paper presents a pairwise constrained clustering framework and a new method for actively selecting informative pairwise constraints to get improved clustering performance. The clustering and active learning methods are both easily scalable to large datasets, and can handle very high dimensional data. Experimental and theoretical results confirm that this active querying of pairwise constraints significantly improves the accuracy of clustering when given a relatively small amount of supervision.

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.

Semi-supervised clustering allows a user to specify available prior knowledge about the data to improve the clustering performance. A common way to express this information is in the form of pair-wise constraints. A number of studies have shown that, in general, these constraints improve the resulting data partition. How- ever, the choice of constraints is critical since improperly chosen constraints might actually degrade the clustering performance. We focus on constraint (also known as query) selection for improving the performance of semi-supervised clustering algorithms. We present an active query selection mechanism, where the queries are selected using a min-max criterion. Experimental results on a variety of datasets, using MPCK-means as the underlying semi-clustering algorithm, demonstrate the superior performance of the proposed query selection procedure.

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.
Semi-supervised document clustering, which takes into account limited supervised data to group unlabeled documents into clusters, has received significant interest recently. Because of getting supervised data may be expensive, it is important to get most informative knowledge to improve the clustering performance. This paper presents a semi-supervised document clustering algorithm and a new method for actively selecting informative instance-level constraints to get improved clustering performance. The semi-supervised document clustering algorithm is a Constrained DBSCAN (Cons-DBSCAN) algorithm, which incorporates instance-level constraints to guide the clustering process in DBSCAN. An active learning approach is proposed to select informative document pairs for obtaining user feedbacks. Experimental results show that Cons-DBSCAN with our proposed active learning approach can improve the clustering performance significantly when given a relatively small amount of constraints.

L. Breiman, "Random Forests," They present a method i.e random forest that 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.

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 a.s. 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. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost, but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

B. Gowthami, and Mr. K.Selvaraj, "Active learning of constraints for semi-supervised Clustering" The novel scheme exploits both semi-kernel learning and batch mode active learning for relevance feedback in CBIR. In particular, a kernel function is first learned from a mixture of labeled and unlabeled examples. The kernel will then be

used to effectively identify the informative and diverse examples for active learning via a min-max framework. An empirical study with relevance feedback of CBIR showed that the proposed scheme is significantly more effective than other state-of-the-art approaches. 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. In particular, we presented a unified learning framework for incorporating both labeled and unlabeled data to improve the retrieval accuracy, and developed a new batch mode active learning algorithm based on the min-max framework. The empirical results with relevance feedback of CBIR showed the advantages of the proposed solution compared to the other state of-the-art methods.

Ms.A.Savithamani ,Mr.M.Mohanraj et al, the active learning along with incremental clustering problems, which is pointing at the problem of category detection accuracy in the traditional active learning based detection algorithms. Those algorithms does not produce high precision and performs only low forecasting accuracy under the situation of small sample training, and puts forward the algorithm of Support Vector Machine. The proposed system has implemented to deal the above problem and Aimed at the important influence of ACO_BSVM with ant primary direction on classification performance. The proposed system adopts the improved SVM along with ant colony and top K methods of selection appropriate labels and characteristics parameters. This algorithm is significantly will produce higher results than the other algorithm in training and the detection speed, and have a high enhance of the detection rates of attacking sample. This paper introduces a new machine learning based data classification algorithm that is applied to disease detection.

Statistical approaches assume that the data follows some standard or predetermined distributions, and this type of approach aims to find the outliers which deviates from such distributions. For distance-based methods, the distances between eachdata point of interest and its neighbors are calculated. If the result is above some predetermined threshold, the target instance will be considered as an outlier.

Mikhail Bilenko et al, Semi-supervised clustering employs a small amount of labeled data to aid unsupervised learning. Previous work in the area has utilized supervised data in one of two approaches: 1) constraint-based methods that guide the clustering algorithm towards a better grouping of the data, and 2) distance-function learning methods that adapt the underlying similarity metric used by the clustering algorithm. This paper provides new methods for the two approaches as well as presents a new semi-supervised clustering algorithm that integrates both of these techniques in a uniform, principled framework. Experimental results demonstrate that the unified approach produces better clusters than both individual approaches as well as previously proposed semisupervised clustering algorithms.

By ablating the metric-based and constraint-based components of our unified method, we present experimental results comparing and combining the two approaches on multiple datasets. The two methods for

semi-supervision individually improve clustering accuracy, and our unified approach integrates their strengths. Finally, we demonstrate that the semi-supervised metric learning in our approach outperforms previously proposed methods that learn metrics prior to clustering, and that learning multiple cluster specific metrics can lead to better results.

Ian Davidson et al, Clustering with constraints is an active area of machine learning and data mining research. Previous empirical work has convincingly shown that adding constraints to clustering improves the performance of a variety of algorithms. However, in most of these experiments, results are averaged over different randomly chosen constraint sets from a given set of labels, thereby masking interesting properties of individual sets. We demonstrate that constraint sets vary significantly in how useful they are for constrained clustering; some constraint sets can actually decrease algorithm performance. We create two quantitative measures, informativeness and coherence, that can be used to identify useful constraint sets. We show that these measures can also help explain differences in performance for four particular constrained clustering algorithms.
The operating assumption behind all constrained clustering methods is that the constraints provide information about the true (desired) partition, and that more information will increase the agreement between the output partition and the true partition. Therefore, if the constraints originate from the true partition labels, and they are noise-free, then it should not be possible for them to decrease clustering accuracy. However, as we show in this section, this assumption does not always hold. The experimental methodology adopted by most previous work in constrained clustering involves generating constraints by repeatedly drawing pairs of data points at random from the labeled subset (which may be the entire data set). If the labels of the points in a pair agree, then an ML constraint is generated; otherwise, a CL constraint is generated. Once the set of constraints has been generated, the constrained clustering algorithm is run several times and the average clustering accuracy is reported. Learning curves are produced by repeating this process for different constraint set sizes, and the typical result is that, on average, when more constraints are provided, clustering accuracy increases. However, the focus on characterizing average behavior has obscured some interesting and exceptional behavior that results from specific constraint sets. In this work, we will empirically demonstrate such cases and provide insight into the reasons for this behavior.

PROBLEM DESCRIPTION
EXISTING SYSTEM:

Existing system presented an evolutionary algorithm to induce fuzzy rules that exploits labeled and unlabeled training data. Existing system compared it to existing fuzzy semi-supervised algorithms. Our MDL-based approach outperformed the other semi-supervised algorithms on the artificial example datasets, where certain flexibility was required to model the distribution and where the given labeled examples were less representative. Additionally Existing system have shown the applicability of our semi-supervised rule learner on a real-world problem.

PROPOSED SYSTEM:

A number of approaches have been proposed for models like, for example, neural networks or support vector machines, that are generally hardly human understandable. Little has been done on the semi supervised extraction of (interpretable) fuzzy rules. The methods described in the following sections are able to induce fuzzy models in a partially supervised manner. It is probably easier to support an unsupervised algorithm with additional labels than vice versa. Thus, it is not surprising that there are a number of semi-supervised extensions of fuzzy clustering.

MODULES

SYSTEM MODEL

The methodology to effectively choose pairwise queries to produce an accurate clustering assignment. Through active learning, the number of queries is reduced to achieve a good clustering performance. We view this as an iterative process such that the decision for selecting queries should depend on what has been learned from all the previously formulated queries. Will introduce our proposed methodology.

MEASURING UNCERTAINTY

In uncertainty-based sampling for supervised learning, an active learner queries the instance about which the label uncertainty is maximized. Numerous studies have investigated different approaches for measuring uncertainty given probabilistic predictions of the class labels. In our context, one can take a similar approach and measure the uncertainty of each data instance belonging to different clusters. Instead, our approach estimates the probability of each instance belonging to each neighborhood using a similarity based approach, where the similarity measure is learned under the supervision of the current clustering solution. This learning-based approach allows us to transfer the knowledge that we have learned from the constraints to the similarity measures.

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. 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. Note that random forest has previously been successfully applied to estimating similarities between unsupervised objects. In that work, a random forest classifier is built to distinguish the observed data from synthetically generated data, whereas our work builds the random forest classifier to distinguish the different clusters. Because the clusters are identified by applying constraint-based clustering to the data using the constraint set K, thus the resulting proximities can be also viewed as a supervised similarity measure learned indirectly using the constraint set K.

ESTIMATION OF NEIGHBORHOOD PROBABILITY

Let S denotes the similarity matrix generated by previous steps, let S(yi,yj) denotes the similarity between instance yi and instance yj. For any unconstrained data point y, we assume that its Probability of belonging to a neighborhood Hi to be proportional to the average2 similarity between y and the instances in Hi. More formally, we estimate the probability of an instance y belonging to neighborhood Hi,

$$p(y \in H_i) = \frac{\frac{1}{|H_i|}\sum_{y_j \in H_i} S(y,y_j)}{\sum_{p=1}^{m} \frac{1}{|H_p|}\sum_{y_j \in H_p} S(y,y_j)}$$

where H_i indicates the number of instances in neighborhood H_i, and m is the total number of existing neighborhoods. Note that in the early stages of our algorithm, when all neighborhoods are small, it is possible for an unconstrained data point y to have zero average similarity with every neighborhood. In such cases, we assign equal probabilities to all neighborhoods for y . This will essentially treat instance x as highly uncertain, making it a good candidate to be selected by our algorithm. This behavior is reasonable because it will encourage the discovery of more neighborhoods in early stages. Finally, we measure the uncertainty of an instance by

the entropy of its neighborhood membership, which we denote as (H/y).

$$E(H|y) = -\sum_{i=1}^{m} p(y \in H_i) \log_2 p(y \in H_i)$$

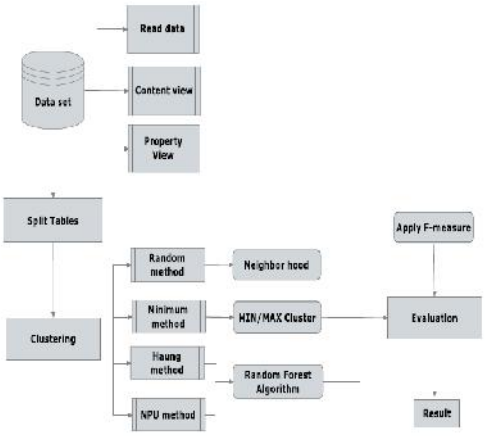
Where m is the total number of existing neighborhoods.

BASELINE METHODS

To demonstrate the effectiveness of the proposed method, we first compare its performance to a set of competing methods, including a random policy, the Min-Max approach introduced to make it applicable to non-document data types. Below we briefly explain these baseline methods.

Random: This policy selects random pairwise queries that are not included in or implied by the current set of constraints K. It is not a neighborhoodbased approach, and is a commonly used baseline for active learning studies.

Min-Max:. Min-Max is a neighborhood-based approach that works in two phases. The first phase builds disjoint neighborhoods using farthest-first traversal. The second phase incrementally expands the neighborhoods by selecting a point to query using a distance-based Min-Max criterion.



BASELINE METHOD 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

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.
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.

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:

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

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

$$H(\mathcal{N} \mid \mathbf{x}) = -\sum_{i=1}^{l} p(\mathbf{x} \subset N_i) \log_2 p(\mathbf{x} \subset N_i),$$

Where l is number of neighbourhood.

CLUSTERING WITH K-MEANS

K-Means is a clustering algorithm based on iterative relocation that partitions a dataset into K clusters, locally minimizing the total squared Euclidean distance between the data points and the cluster centroids. Let X = fxigNi=1; xi 2 m be a set of data points, xidbe the d-th component of xi, f¹hgK h=1 represent the K cluster centroids, and li be the cluster assignment of a point xi, where li 2 f1; : : : ;Kg. The Euclidean K-Means algorithm creates a K-partitioning fXhgK h=1 of X so that the objective functionPxi2X kxi¡ ¹lik2 is locally minimized. It can be shown that the K-Means algorithm is essentially an EM algorithm on a mixture of K Gaussians under assumptions of identity covariance of the Gaussians, uniform mixture component priors and expectation under a particular type of conditional distribution. In the Euclidean K-Means formulation, the squared L2-norm kxi¡ ¹lik2 = (xi ¡ ¹li)T(xi ¡ ¹li) between a point xi and its corresponding cluster centroid ¹li is used as the distance measure, which is a direct consequence of the identity covariance assumption of the underlying Gaussians.

CONSTRAINTS CAN DECREASE PERFORMANCE

The operating assumption behind all constrained clustering methods is that the constraints provide information about the true (desired) partition, and that more information will increase the agreement between the output partition and the true partition. Therefore, if the constraints originate from the true partition labels, and they are noise-free, then it should not be possible for them to decrease clustering accuracy. However, as we show in this section, this assumption does not always hold. The experimental methodology adopted by most previous work in constrained clustering involves generating constraints by repeatedly drawing pairs of data points at random from the labeled subset (which may be the entire data set). If the labels of the points in a pair agree, then an ML constraint is generated; otherwise, a CL constraint is generated. Once the set of constraints has been generated, the constrained clustering algorithm is run several times and the average clustering accuracy is reported. Learning curves are produced by repeating this process for different constraint set sizes, and the typical result is that, on average, when more constraints are provided, clustering accuracy increases. However, the focus on characterizing average behavior has obscured some interesting and exceptional behavior that results from specific constraint sets. In this work, we will empirically demonstrate such cases and provide insight into the reasons for this behavior.

We begin by examining the behavior of four different constrained clustering algorithms on several standard clustering problems. The two major types of constrained clustering techniques are (a) direct constraint satisfaction and (b) metric learning. The techniques of the first category try to satisfy the constraints during the clustering algorithm; the latter techniques treat an ML (or CL) constraint as specifying that the two points in the constraint and their surrounding points should be nearby (or well separated) and tries to learn a distance metric to achieve this purpose.

Data Set	Algorithm							
	CKM		PKM		MKM		MPKM	
	Unconst.	Const.	Unconst.	Const.	Unconst.	Const.	Unconst.	Const.
Glass	69.0	69.4	43.4	68.8	39.5	56.6	39.5	67.8
Ionosphere	58.6	58.7	58.8	58.9	58.9	58.9	58.9	58.9
Iris	84.7	87.8	84.3	88.3	88.0	93.6	88.0	91.8
Wine	70.2	70.9	71.7	72.0	93.3	91.3	93.3	90.6

Table 1- Average Performance (Rand Index) Of Four Constrained Clustering Algorithms

DEALING WITH OVERLAPPING CLUSTERING

Constraints have been used to improve clustering performance by incorporating some background knowledge in a clustering problem. In a study on constraint based cluster using constraints can sometimes decrease this performance. They introduce the notion of coherence between constraints, and show that the more incoherent a constraint set is, the more chance it has to decrease clustering performance. Two constraints are called incoherent if they carry information that is a priori contradictory. For instance, in figure the must-link constraint (in blue) implies that the left area must be clustered with the right area, while the cannot-link constraint (in red) says the opposite.

This toy example illustrates the variety of clustering algorithms: differentalgorithms will produce different partitionings. Moreover, in a real clustering problem, we cannot say one of these partitionings is better as we do not knowthe true labels. Even on the same dataset, two users might be interested in adifferent partitioning of the data. Only if some constraints are specified can webuild a system that selects the algorithm best fitting a user requirements.

EMPIRICAL RISK MINIMIZATION FOR ACTIVE LEARNING

In supervised learning, the target of learning is to find the optimal classifier which is expected to generalize well on the unseen data. The empirical risk minimization (ERM)is a successful guideline for designing machine learning and data mining methods. It minimizes an upper bound of the true risk under the unknown data distribution. This upper bound is approximated by the summation of empirical risk on the available data and a properly designed regularization term, which constrains the complexity of the candidate classifiers. Assume we are given a data source D, with unknown distribution p(z) = p(x, y) for sample z = {x, y}, and a finite data set S with n points, which are i.i.d. sampled from the same distribution, p(z). Using the Rademacher complexity to describe the complexity of the function class, we obtain the uniform convergence property between the true risk and the empirical risk

$$E_D(l(\mathbf{z})) \leq \hat{E}_S(l(\mathbf{z})) + 2R_n(\mathcal{L}) + \sqrt{\frac{\ln 1/\delta}{n}},$$

HUMAN ACTIVE LEARNING IN CONSTRAINED CLUSTERING

This paper approach these problems by developing an interactive tool that helps users efficiently select effective constraints during the clustering process. The main objectives to build the interactive tool can be sum up as follows. 1) To provide an interactive environment in which users can visually recognize the proximity of data, and give constraints easily by mouse manipulation. 2) To provide hints for the better selection strategies through the interaction process between the interactive system and users. In addition to the 2-D visual arrangement of a dataset and the constraint assignment function, our prototype tool has distance metric learning and k-means clustering that can

be quickly executed as the background process. Using these functions, the users can compare the results of the clustering before and after the constraints addition easily. We consider such interactions helpful for providing hints for better selection strategies. Although there are many data mining tools that have clustering function, we have not found any other tool that realizes interactive constraint assignment through the interactive clustering process.

PERFORMANCE ANALYSIS

In our experiments with high-dimensional text documents, we used datasets created from the 20 Newsgroups collection.3 It has messages collected from 20 different Usenet newsgroups, 1000 messages from each newsgroup. From the original dataset, a reduced dataset News-all20 was created by taking a random subsample of 100 documents from each of the 20 newsgroups – this subsample is a more difficult dataset to cluster than the original 20 Newsgroups, since each cluster has fewer documents. News-all20 has 2000 points in 16089 dimensions. By selecting 3 categories from the reduced dataset News-all20, two other datasets were created: News-sim3 that consists of 3 newsgroups on similar topics (comp.graphics, comp.os.ms-windows, comp.windows.x) with significant cluster overlap, and News-diff3 that consists of 3 newsgroups on different topics (alt.atheism, rec.sport.baseball, sci.space) with well-separated clusters. News-sim3 has 300 points in 3225 dimensions, while News-diff3 had 300 points in 3251 dimensions. Another dataset we used in our experiments is a subset of Classic3 containing 400 documents – 100 Cranfield documents from aeronautical system papers, 100 Medline documents from medical journals, and 200 Cisi documents from information retrieval papers. This Classic3-subset dataset was specifically designed to create clusters of unequal size, and has 400 points in 2897 dimensions.

Similarities between data points in the text datasets were computed using cosine similarity, following SPKMeans [11]. SPKMeans maximizes the average cosine similarity between points and cluster centroids, so that the objective function monotonically increases with every iteration till convergence. All the text datasets were preprocessed following the methodology. For experiments on low-dimensional data, we selected the UCI dataset Iris, which has 150 points in 4 dimensions. The Euclidean metric was used for computing distances between points in this dataset, following KMeans. In this case, the objective function, which is the average squared Euclidean distance between points and cluster centroids, decreases at every iteration till convergence. The Iris dataset was not pre-processed in any way.

CONCLUSION

Here this paper shown that using individual metrics for different clusters, as well as performing feature generation via a full weight matrix in contrast to feature weighting with a diagonal weight matrix, can lead to improvements over our basic algorithm. Extending our approach to high-dimensional datasets, where Euclidean distance performs poorly, is the primary avenue for future research. Other interesting topics for future work include selection of most informative pairwise constraints that would facilitate accurate metric learning and obtaining good initial centroids, as well as methodology for handling noisy constraints and cluster initialization sensitive to constraint costs

.

FUTURE WORK

The methods for the extraction of fuzzy classification rules from data that we mentioned in the last sections are either supervised or unsupervised. Both learning paradigms have their drawbacks. The main drawback ofsupervised learning is clearly its need for supervision, i.e. the need to present labels together with the objects.The result ofunsupervised learning, however, strongly

depends on a number ofprior assumptions (explicitly or implicitly). Thus, it depends on an appropriate choice of, e.g. attributes scaling, distance measure, distribution function and expected number of classes or clusters, whether the clusters found in the data space correspond to any "meaningful" classes of objects. Hence, unsupervised learning does in many cases not yield satisfactory results, and supervised learning is much more common in practice.

BIBLOGRAPHY

[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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] B. Sugato, "Wekaut, A Modified Version of Weka." http://www. cs.utexas.edu/users/ml/risc/code/, 2011.

[13] L. Kuncheva and S. Hadjitodorov, "Using Diversity in Cluster Ensembles," Proc. Int'l Conf. Systems, Man and Cybernetics, vol. 2, pp. 1214-1219, 2004.

[14] I. Davidson, S. Ravi, and L. Shamis, "A SAT-Based Framework for Efficient Constrained Clustering," Proc. SIAM Int'l Conf. Data Mining, 2010.