An Active Learning for Weakly Supervised Clustering

Ms.A.Savithamani , Mr.M.Mohanraj

Abstract-**This paper researches 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.**

Index Terms- semi-supervised clustering, active learning, Batch Support Vector Machine

1. INTRODUCTION

1.1 OVERVIEW

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, marketting, biology, libraries, insurance, city-planning, earthquakes, and www document classification.

OVERVIEW OF ACTIVE LEARNING

Active learning is a special case of semi-supervised machine learningin which a learning algorithm is able to interactively query the user or some other information source to obtain the desired outputs at new data points.

There are situations in which unlabeled data is abundant but manually labeling is expensive. In such a scenario, learning algorithms can actively query the user for labels. This type of iterative supervised learning is called active learning. Since the learner chooses the examples, the number of examples to learn a concept can often be much lower than the number required in normal supervised learning. With this approach, there is a risk that the algorithm be overwhelmed by uninformative examples. Recent developments are dedicated to hybrid active learning and active learning in a single-pass (on-line) context, combining concepts from the field of Machine Learning with adaptive, incremental learning policies in the field ofOnline machine learning.

OVERVIEW OF SEMI-SUPERVISED LEARNING

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls betweenunsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-

learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning.

Active learning and semi-supervised learning both traffic in making the most out of unlabeled data. As a result, there are a few conceptual overlaps between the two areas that are worth considering. For example, a very basic semi-supervised technique is self-training in which the learner is first trained with a small amount of labeled data, and then used to classify the unlabeled data. Typically the most confident unlabeled instances, together with their predicted labels, are added to the training set, and the process repeats. A complementary technique in active learning is uncertainty sampling where the instances about which the model is leastconfident are selected for querying.

Similarly, co-training and multi-view learning use ensemble methods for semisupervised learning. Initially, separate models are trained with the labeled data which then classify the unlabeled data, and "teach" the other models with a few unlabeled examples (using predicted labels) about which they are most confident. This helps to reduce the size of the version space, i.e., the models must agree on the unlabeled data as well as the labeled data. Query-by-committee is an active learning compliment here, as the committee represents different parts of the version space, and is used to query the unlabeled instances about which they do not agree.

1.2 OBJECTIVE OF THE RESEARCH

- The work aims to develop an algorithm that combines the logic of both methods to produce a high-performance of semi supervised clustering with active learning system.
- The proposed system aims at increasing the clustering performance through the combination of B-SVM (Batch-Support vector machine) classification and incremental Ant clustering.
- The goal of the proposed system is applying the active learning of constrains to identify the best label of objects and clustering them accordingly.
- This aims at producing least false alarm rate and improving clustering performance.
- Reducing training data by applying the historical data as an input. So this aims at reducing the training overhead.
- Active clustering aims to identify a small group of instances which deviate remarkably from the existing data

1.3 SCOPE OF THE RESEARCH

The pair-wise implementation helps to improve the clustering performance. The proposed system overcomes the re-clustering problem by applying the incremental semi supervised method, which utilizes ant clustering and oversampling method. The proposed oversampling method, which is a semi supervised technique, utilizes the previous top K labels as training data for data learning. This performs top –k algorithm for finding best labels for fast clustering. Using the above the proposed system reduces the training phase and improves the clustering speed.

2. LITERATURE REVIEW

2.1 PROBLEM DEFINITION

The proposed system deals the active learning problem of selecting pair wise must-link and cannot-link constraints for semi supervised clustering. In common the research on active learning for constraint-based clustering has been limited in the research. Most of the existing research studied the selection of a set of initial constraints prior to performing semi-supervised clustering. Several studies

do not deal the active learning process, which incurs more training overhead. The problem addressed in this thesis is how to effectively choose pair wise queries to produce an accurate clustering assignment.

The proposed system defines the problem as follows.

Given a set of data instances D = {x1; . . . ; xn}, this assumes that there exists an underlying class structure that assigns each data instance to one of the c classes. Every data instances may not have proper label for clustering. Training process for semi supervised clustering is very tedious, so the system uses neighbor labels for the given data set.

2.2 EXISTING SYSTEM

Obtaining pair wise constraints typically requires a user to manually inspect the data points in question, which can be time consuming and costly. While active learning has been extensively studied in supervised learning the research on active learning of constraints for semi-supervised clustering is relatively limited. Most of the existing work on this topic has focused on selecting an initial set of constraints prior to performing semi-supervised clustering.

The existing approaches can be divided into three categories:

1. Distribution (statistical)
2. Distance
3. Density based methods.

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 each data 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.

2.3 RELATED WORK

Semi-supervised clustering [1] 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 pair wise constrained clustering framework and a new method for actively selecting informative pair wise 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 pair wise constraints significantly improves the accuracy of clustering when given a relatively small amount of supervision. In this paper, they have presented a pair wise constrained clustering framework and a new theoretically well motivated method for actively selecting good pair wise constraints for semi-supervised clustering.

Clustering with constraints is an active area of machine learning [3] and data mining research. Previous empirical work has convincingly shown that adding constraints to clustering improves performance, with respect to the true data labels. However, in most of these experiments, results are averaged over different randomly chosen constraint sets, thereby masking interesting properties of individual sets. They demonstrate that constraint sets vary significantly in how useful they are for constrained clustering; some constraint sets can actually decrease algorithm performance. They create two quantitative measures, in formativeness and coherence that can be used to identify useful constraint sets.

3.RESEARCH METHODOLOGY

3.1 PROPOSED SYSTEM

The existing iterative framework requires repeated re-clustering of the data with an incrementally growing constraint set. This can be computationally demanding for large data sets. To address this problem, it would be interesting to consider an incremental semi-supervised clustering method that updates the existing clustering solution based on the neighborhood assignment for the new point.

An alternative way to lower the computational cost is to reduce the number of iterations by applying a batch approach that selects a set of points to query in each iteration.

3.2 CONTRIBUTION OF THE PROPOSED WORK

The followings are the contributions of the proposed system.

- The existing iterative framework requires repeated re-clustering of the data with an incrementally growing constraint set. This can be computationally demanding for large data sets. To address this problem, the system introduces an incremental semi-supervised clustering method that updates the existing clustering solution based on the neighborhood assignment for the new point.
- An alternative way to lower the computational cost is to reduce the number of iterations by applying a batch approach that selects a set of points to query in each iteration of the dataset.
- A naive batch active learning approach would be to select the top k points that have the highest normalized uncertainty to query their neighborhoods.
- SVM based batch active learning approach has been applied to select the top k points that have the highest normalized uncertainty to query their neighborhoods.
- Ant Colony Optimization algorithm has been used for neighbor label selection. This reduces the need of training dataset. This also handles the data inconsistency.
- The proposed system also updates the previous dataset and perform the Top-k result.

3.3 METHODOLOGY

Clustering means the act of partitioning an unlabeled dataset into groups of similar objects. The goal of clustering is to group sets of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate clusters. The proposed system performs semi supervised clustering.

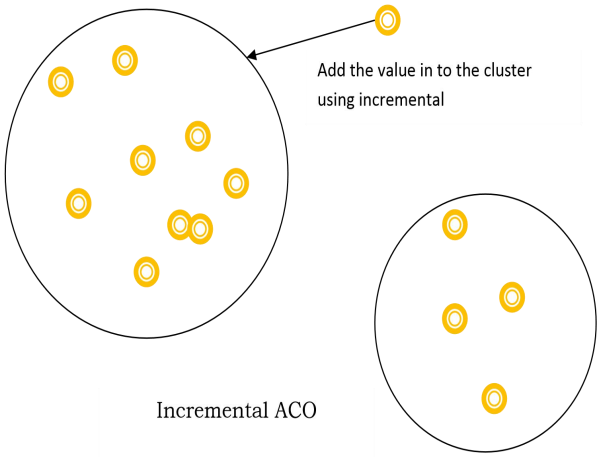
1. Incremental Ant Clustering
2. B-SVM (Batch-Support vector machine) classification.
3. Oversampling and update sampling methods.
4. Top-K algorithm
5. Normalized Mutual information

Incremental Ant Clustering:

The system effectively utilizes the incremental ant colony. In general ant colony in the natural world has an intellective character—ants can release a chemical substance called pheromone, they can carry food back to their nest in the shortest route without any visual aid. In literature several authors proposed the “Ant system” method based on such character of ants, and used in several methods, which received great best results.

Further on, M. Dorigo named all ant colony algorithms as Ant Colony Optimization (ACO) in general, which proposed an unique framework model. This algorithm has not only great robustness, positive feedback characteristic and also with parallel and distributed computing feature.

Suppose ant colony scale as N, randomly distribute the ant colonies in solution space, then according to the initialized position of the ants’ distribution, follow the difference of optimization problem to confirm the initialized pheromone size of ant f:



After one searching round, ants will make the next searching round by the movement experience accordingly. This proposed algorithm movement regulation contains two steps.

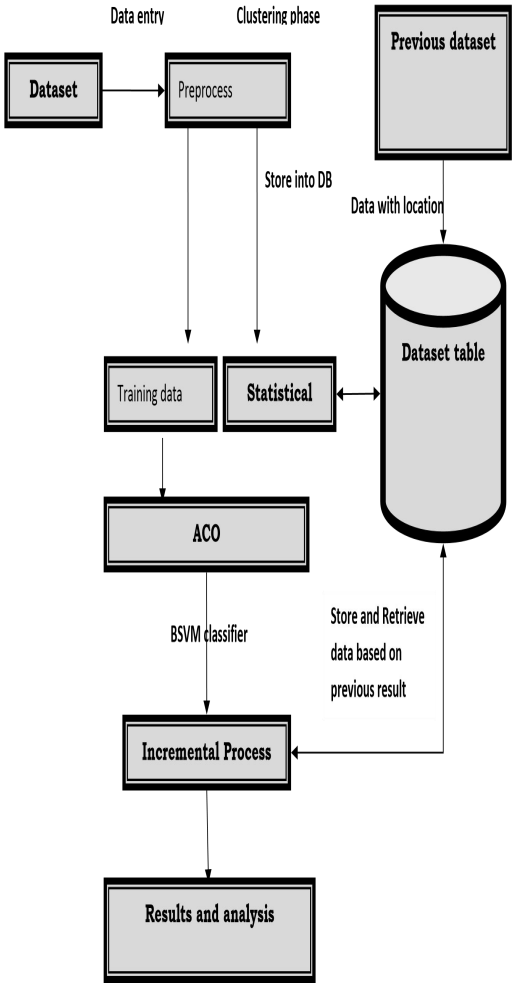
Step 1: one is to select individual target through dynamic random, move the other ants to individual target except the optimal ants from the last iteration, this call it as overall long step search;

Step 2: the second step refers from the detective theory which follows the above step1 in pattern search. Let the optimal ants have short step partial elaborate search in the neighborhood, in order to find the optimal label.

After finishing overall search and partial search in the neighborhood watch method, update label in the cluster.

BSVM Process

The proposed system uses SVM based semi supervised clustering algorithm, which is the new classification method proposed with batch processing. It develops on the basis of statistical model. The basic thought of BSVM is first input the sample and through the kernel function map to the higher dimensional eigenspace, then looking for the optimum boundary in the eigenspace through the maximizing classification interval and the classification interval is maximized and can be transformed into quadratic programming problem.



Overall diagram

3.4 MODULES

1.Data set

- The first module is the process of uploading datasets.
- The proposed system uses statlog Heart dataset.
- The dataset has a list of descriptions in the table. The dataset may contain 270 records. This modules collects those data's and stores into the database for further process.

2.Preprocessing and training

Preprocessing is the process of elimination, which eliminates duplicate and incomplete data's from the dataset before processing. It does not include redundant records in the train set as well, so the classifiers

will not be biased towards more frequent records. Here are no duplicate records in the proposed test sets; therefore, the performance of the learners is not biased by the methods which have better detection rates on the frequent records.

3.Ant clustering

The ant colony in the natural world has an intellective character. The system implements the ant colony techniques for intrusion detection. This module describes the "Ant system" method based on such character of ants, which received great lab results. This modules implements the ant clustering phase.

After the third module clustering, the neighbors of those marked objects are stored in the SVM training data file, which is used by the component SVM.

All ant colony algorithms as Ant Colony Optimization (ACO) in general, which proposed an unique framework model. This algorithm has not only great robustness, positive feedback characteristic and also with parallel and distributed computing feature.

4.Classification

It establishes the enhanced model of ACO_BSVM-based classifier, which is a hybrid of the SVM classifier and the Ant Colony classifier. By repeating the processes of SVM training and AC clustering, the detection classifier is established and stored in the common storage Disk, which is used in the testing phase. This will finally used to display the results.

5.Reports and results

The final module provides the classification results and test bed approach toshow the accuracy and effectiveness of the proposed system. The results will be generated as a graphical form.

4. IMPLEMENTATION AND RESULTS

4.1 DATASETS

In the experiments, the system use benchmark data set from UCI repository. Which not have been used in previous studies on constraint based clustering. the data sets include statlog-heart.

4.2 EXPERIMENT STEPS

Dataset Collection and Upload Process

The first module is the process of uploading datasets. The first module creates dataset for ACO_BSVM implementation from UCI repository.

Preprocessing

The dataset will be preprocessed before starting the clustering implementation. This step eliminates the duplicate and missing items in the uploaded dataset.

ACO_BSVM Process

The investigating data have 1000 observing sample, there exists missing value in these sample. After eliminating the missing element, the system performs the ACO_BSVM for every attribute. The ACO_BSVM implementation process identifies the frequency of every value from the dataset.

BSVM has been implemented to identify the class of the given test data property and its label. This is based on the SVM based approach which performs the best segmented ie clustered label identification process and class analysis.

BSVM based labeling has been created in this module. The user can give the partition threshold. Aset of data instances in the original data set is taken as predefined input. This data may be contaminated by noise and incorrect data labelling etc., this data might be error free, because this is going to be used as training data. So the cleaning is done using before updating the data.

Type	Neighborhood based method	ACO_BSVM
Precision (%)	90.7	99.5
Training Time(ms)	5.6	2.3
Testing Time (ms)	3.4	2.1
Efficiency	Ordinary	Better

Detecting Labels (Results)

This is for detecting the cluster label from the user test data. When the user gives the input to the system, the system calculates the threshold value for every attribute value for the new input. And then compare that new S_t value with the threshold value which is calculated in earlier. Final results will be identified individually and updated in the database using oversampling method.

Evaluation Criteria

Two evaluation criteria are used in our experiments. First, we use normalized mutual information (NMI) to evaluate the clustering assignments against the groundtruth class labels. NMI considers both the class label and clustering assignment as random variables, and measures the mutual information between the two random variables, and normalizes it to a zero-to-one range.

Second, we consider F-measure as another criterion to evaluate how well we can predict the pairwise relationship between each pair of instances in comparison to the relationship defined by the ground-truth class labels [1]. F-measure is defined as the harmonic mean of precision and recall

4.3 COMPARISION

Our method takes a neighborhood-based approach, and incrementally expands the neighborhoods by posing pairwise queries. We devise an instance-based 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.

5. CONCLSUION AND FUTURE ENHANCEMENT

Active learning is a growing area of research in machine learning, no doubt fueled by the reality that data is increasingly easy or inexpensive to obtain but difficult or costly to label for training. Over the past two decades, there has been much work in formulating and understanding the various ways in which queries are selected from the learner's perspective. This has generated a lot of evidence that the number of labeled examples necessary to train accurate models can be effectively reduced in a variety of applications.

An alternative way to lower the computational cost is to reduce the number of iterations by applying a batch approach that selects a set of points to query in each iteration. A naive batch active learning approach would be to select the top k points that have the highest normalized uncertainty to query their neighborhoods. However, such a strategy will typically select highly redundant points.

The optimized ACO_BSVM algorithm has been expanded with the new optimal classification algorithms, which can handle large category dataset more rapidly, accurately and effectively, and keep the good scalability at the same time. The algorithm mainly created to perform active learning process in the given data, but this should disperse the value data in the dealing process. So, this should do further improvement to the algorithm to adapt the mixed data directly.

REFERENCES

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

[2] M. Bilenko, S. Basu, and R. Mooney, "Integrating Constraints and Metric Learning in Semi-Supervised Clustering," Proc. Int'l Conf. Machine Learning, pp. 1118, 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] Y. Guo and D. Schuurmans, "Discriminative Batch Mode Active Learning," Proc. Advances in Neural Information Processing Systems, pp. 593-600, 2008.

[5] 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. [6] P. Mallapragada, R. Jin, and A. Jain, "Active Query Selection for Semi-Supervised Clustering," Proc. Int'l Conf. Pattern Recognition, pp. 1-4, 2008.

[7] 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.

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

175-189, Springer, 2009.

[9] 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.

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

[12] K. Voevodski, M. Balcan, H. Ro¨ glin, S. Teng, and Y. Xia, "Active Clustering of Biological Sequences," J. Machine Learning Research, vol. 13, pp. 203-225, 2012.

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

[14] 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.

[15] 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.

[16] 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.

[17] Zhu, Xiaojin; Goldberg, Andrew B. (2009). Introduction to semi-supervised learning. Morgan & Claypool. ISBN 9781598295481.

[18] Balcan, M.-F., & Blum, A. (2006). An augmented pac model for semi-supervised learning. In O. Chapelle, B. Scholkopf¨ and A. Zien (Eds.), Semi-supervised learning. MIT Press.

[19] Brefeld, U., & Scheffer, T. (2006). Semi-supervised learning for structured output variables. *ICML06,*

23rd International Conference on Machine Learning. Pittsburgh, USA.

[20] Chapelle, O., Sindhwani, V., & Keerthi, S. S. (2006b). Branch and bound for semisupervised support vector machines. Advances in Neural Information Processing Systems (NIPS).

[21] Chapelle, O., Zien, A., & Scholkopf,¨ B. (Eds.). (2006c). Semi-supervised learning. MIT Press.

[22] Culp, M., & Michailidis, G. (2007). An iterative algorithm for extending learners to a semi supervised setting. The 2007 Joint Statistical Meetings (JSM).

[23] Johnson, R., & Zhang, T. (2007). Two-view feature generation model for semi supervised learning. The 24th International Conference on Machine Learning.

[24] Mann, G. S., & McCallum, A. (2007). Simple, robust, scalable semi-supervised learning via expectation regularization. The 24th International Conference on Machine Learning.

[25] Arora, E. Nyberg, and C.P. Rose.´ Estimating annotation cost for active learning in a multi-annotator environment. In Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing, pages 18–26. ACL Press, 2009.

[26] M.F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In Proceedings of the Conference on Learning Theory (COLT), pages 45–56. Springer, 2008.

[27] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance-weighted active learning. In Proceedings of the International Conference on Machine Learning (ICML), pages 49–56. ACM Press, 2009.

[28] A. Carlson, J. Betteridge, R. Wang, E.R. Hruschka Jr, and T. Mitchell. Coupled semi-supervised learning for information extraction. In Proceedings of the International Conference on Web Search and Data Mining (WSDM). ACM Press, 2010.

[29] S. Dasgupta and D.J. Hsu. Hierarchical sampling for active learning. In Proceedings of the International Conference on Machine Learning (ICML), pages 208–215. ACM Press, 2008.

[30] S.Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In Advances in Neural Information Processing Systems (NIPS), volume 20, pages 353–360. MIT Press, 2008.

Ms.A.Savithamani, M.Phil Scholar, Department of Computer Science, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore-49, TamilNadu, India.

Mr.M.Mohanraj, Assistant Professor, Department of Computer Applications, Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore-49, TamilNadu, India.

