General Framework for Cluster based Active Learning Algorithm

Prerna Mahajan,

Institute of Information Technology & Management,
Guru Gobind Singh Indraprastha University,
New Delhi, India

Dr. R. Kandwal,

India Meteorological Department,
Ministry of Earth Sciences & Science and Technology,
New Delhi, India.

Dr. Ritu Vijay

Department of Electronics,AIM & ACT
Bansthali University, Rajasthan,India

Abstract

This paper revisits the problem of active learning and decision making when the cost of labeling incurs cost and unlabeled data is available in abundance. In many real world applications large amounts of data are available but the cost of correctly labeling it prohibits its use. In many cases, where unlabeled data is available in abundance, active learning can be employed. In our proposed approach we will try to incorporate clustering into active learning algorithm and also data reduction is achieved through feature selection. The algorithm learns itself incrementally and will adjust clusters and select appropriate features as it explores more data points.

I. INTRODUCTION

In recent years there has been an explosive growth in data from Internet and other sources. This abundance of data requires an urgent need for computational theories and tools to assist humans in discovering information out of it. More sophisticated and structured algorithms are required to enable automated processing and reasoning. Machine learning is one of these fields. It is the study of computer algorithms that automatically improve through experience. They improve by becoming better at explaining observations, making decisions, or predicting outcomes. For example, machines can interpret human speech by learning from vocal recordings that have been annotated for words and sentences [14]. They can learn to drive a car after observing human driving behavior for a period of time [6]. They can even diagnose diseases by analyzing profiles of healthy vs. unhealthy patients [38].Generally, the learning methods used for information management tasks fall into two groups

• Unsupervised learning. The learning system is given a collection of “unlabeled” data. The goal is to organize aspects of the collection in some way. For example, clustering data points called instances into natural groups based on a set of observable features.
• Supervised learning. The learning system is given a collection of “labeled” instances, each denoted by the pair (h, y). The goal is to predict the label y for any new instance x, based on a set of features that describe it. When y is a real number, the task is called regression and when it is a set of discrete values, the task is called classification.

However, for many contemporary practical problems such as text classification, time series analysis, data stream analysis, Web mining and other real world domains there is often additional information available[13,46]; in particular, for many of these settings unlabeled data is often much cheaper and more plentiful than labeled data. As a consequence, there has recently been substantial interest in using unlabeled data together with labeled data for learning , since clearly, if useful information can be extracted from it that reduces dependence on labeled examples, this can be a significant benefit .
The active learning model is motivated by scenarios in which it is easy to amass vast quantities of unlabeled data (images and videos of the web, speech signals from microphone recordings, and so on) but costly to obtain their labels. It shares elements with both supervised and unsupervised learning. Like supervised learning, the

goal is ultimately to learn a classifier. But like unsupervised learning, the data come unlabeled. More precisely, the labels are hidden, and

each of them can be revealed only at a cost. The idea is to query the labels of just a few points that are especially informative about the decision boundary, and thereby to obtain an accurate classifier at significantly lower cost than regular supervised learning, [13, 32, 33, 39, 42, 43, 44, and 52]. Traditional machine learning methods focus on supervised methods to build classification models in various domains like finance, marketing and healthcare [34, 45]. In this paper we are focusing on unsupervised methods for cluster based Active learning.

II. ACTIVE LEARNING

In traditional supervised learning, Classifier must be trained on hundred or even thousands of labeled instances to predict optimal results. However in many real world domains obtaining labels are very difficult, time-consuming and expensive [4, 17, 53]

Active Learning requires small training data to build the initial classifier. Once the learner is built, it starts reading data from the set one at a time and tries to classify it. If the learner can classify the example confidently, it does not look for the help of the expert. It is only when it does not find enough confidence; it asks the expert (or the oracle) to provide the exact label for the current training data. In this way at each stage the learning model identifies especially useful additional data for labeling and updates itself. When successful, active learning methods reduce the number of instances that must be labeled to achieve a particular level of accuracy. The incentive in using active learning is mainly to expedite the learning process and reduce the labeling efforts required by the expert [12, 31, 32, 38, 40, 42, 43, and 51]

Active learning has been proposed in various forms [5, 6, 7, 8, 18, 30, 39, and 52].However the basic active learning algorithm is given in Figure 1.

start with a small labeled set and a large unlabeled set

repeat until some condition is met:

from the unlabeled set select the currently most interesting

example

query the expert/oracle/.. for the label

add the now-labeled example to the labeled set

Figure 1: General Active Learning Algorithm

III. CLUSTER BASED ACTIVE LEARNING

Most of the Active learning methods are supervised, that is, the learning algorithm induces a model that accurately predicts a label for some new instance. However an Active learning algorithm is unsupervised if its task is to simply organize a large amount of unlabeled data in a meaningful way Most importantly the supervised learner try to map instances using a predefined structure, whereas unsupervised learners exploits the inherent structure in the data to explore meaningful patterns. Most prevalent examples of unsupervised learning are clustering algorithms[40].Although unsupervised active learning seems to a bit hazy and clumsy, much research has been done on Active clustering approach in various application areas.

Hofmann et al have proposed an active clustering algorithm for proximity data, based on an expected value of information criterion [50].

Nguyen et al provides a formal framework that incorporates clustering into active learning. The algorithm first constructs a classifier on the set of the cluster representatives, and then propagates the classification decision to the other samples via a local noise model. The proposed model allows to select the most representative samples as well as to avoid repeatedly labeling samples in the same cluster. During the active learning process, the clustering is adjusted using the coarse-to-fine strategy in order to balance between the advantage of large clusters and the accuracy of the data representation [16].

Joshi et al used feature selection integrated with active learning approach in supervised machine learning algorithm. They demonstrated through various experiments that feature selection when integrated within active learning process yields inferior accuracy results because of their inherent properties and suggested that domain knowledge and prior distribution must be taken in account [13].

Raghavan et al studied the effect of feature selection and human feedback in active learning setting. They conducted experiments on a variety of text categorization tasks indicate that there is significant potential in improving classifier performance by feature reweighting, beyond that achieved via selective sampling alone (standard active learning) .They proposed an algorithm that interleaves labeling features and documents which significantly accelerates active learning[17].

Liu et al proposed the concept of active feature selection, and investigated a selective sampling approach to active feature selection in a filter model setting. They presented a formalism of selective sampling based on data variance, and apply it to a widely used feature selection algorithm Relief. Further, it is shown how it realizes active feature selection and reduces the required number of training instances to achieve time savings without performance deterioration [23].

Mallapragada et al emphasized on the importance of the choice of constraints is since improperly chosen constraints might actually degrade the clustering performance. They focused on constraint (query) selection for improving the performance of semi-supervised clustering algorithms. They presented 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 [29].

Campedel et al proposed a methodology that exploits objectively evaluated features followed by automatic selection. Different feature sets are highly redundant but relevant. In order to determinate the best features to be extracted, we propose a methodology based on automatic feature selection algorithms, applied in an unsupervised manner on a strongly redundant features set. It also demonstrates the usefulness of consensus clustering as a feature selection algorithm, allowing selected number of features estimation and exploration facilities [4].

Roth et al proposed a novel approach to combining *clustering* and *feature selection* .It implements a *wrapper* strategy for feature selection, in the sense that the features are directly selected by optimizing the discriminative power of the used partitioning algorithm. They present an efficient optimization algorithm with guaranteed local convergence property. The only free parameter of this method is selected by a re-sampling based stability analysis. Experiments demonstrate that the method is able to infer both meaningful partitions and meaningful subsets of features [48].

Huang et al exploits supervision in clustering using active learning. It utilizes Wikipedia to create a concept-based representation of a text document, with each concept associated to a Wikipedia article. It then exploits the semantic relatedness between Wikipedia concepts to find pair-wise instance-level constraints for supervised clustering, guiding clustering towards the direction indicated by the constraints. Approach is tested on three standard text document datasets. Empirical results show that basic document representation strategy yields comparable performance to previous attempts; and adding constraints improves clustering performance further by up to 20% [2].

IV. NEW CLUSTER BASED ACTIVE LEARNING APPROACH

The main aim of any cluster based active approach would be to generate or sample the unlabeled instances in such a way that they self- organize into small groups with minimal overlapping. In our proposed approach, we intend to combine the benefits of two approaches namely clustering and best feature selections to generate our active learner. Incorporating these two different kinds of supervision will bring a new unexplored dimension to the problem of Active Learning. Cluster approach is also more suitable in real world applications that need automated classification approaches which reduces effort for human annotation in more critical analysis (like fraud detection, disaster management etc) like the most informative cluster can be picked up for further analysis.

The basic idea is to selectively choose most discriminative features (based on information criterion) and also to pick up those, which are most uncertain and likely candidates to provide more information. The cluster will be assigned to a set of data points in the feature space based on some similarity measure like feature vector or cluster radius, density etc. The active approach works iteratively where in each round the learner actively selects a batch of unlabeled samples for training to improve the internal model (cluster adjustment, feature rejection) as quickly as possible. The main goals that we are hoping to achieve in this research are:

- Can we develop a framework for automatic or semi-automatic annotation of real world data sets where labeling is expensive?
- Would such a system be able provide formal evaluation of the cluster based proposed active learning methodology and be able to provide automated label suggestion with minimal user guidance?
- How we learn which attribute is most likely to be of interest, that is, most discriminative to the user for a specific task.
- Supervise the cluster labeling with Active learning approach with minimum data assistance.
- Extend the strands of Active Learning along with feature extraction in dynamic/run time scenario.

- Evaluate and compare the performance of our strategy in various kinds of data sets and extend it accordingly.

V. PROPOSED APPROACH

Instead of incorporating all the features defining an instance, in our feature selection approach for Active learning, at every incremental step, active learner will look for the next best feature to include for labeling it. This will result in the significant improvement in classifier performance. To identify which is the next best feature or which are the most valuable features, we must calculate the expected feature utility, for that various measures exist like information gain, ginni index etc. There are many benefits of reducing the number of features namely:

i. Maximize the accuracy of classifier while minimizing the associated costs.
ii. Reducing irrelevant and redundant features.
iii. Reduce the complexity of algorithm
iv. Reducing training data size and cost.
v. More robust and practical solution

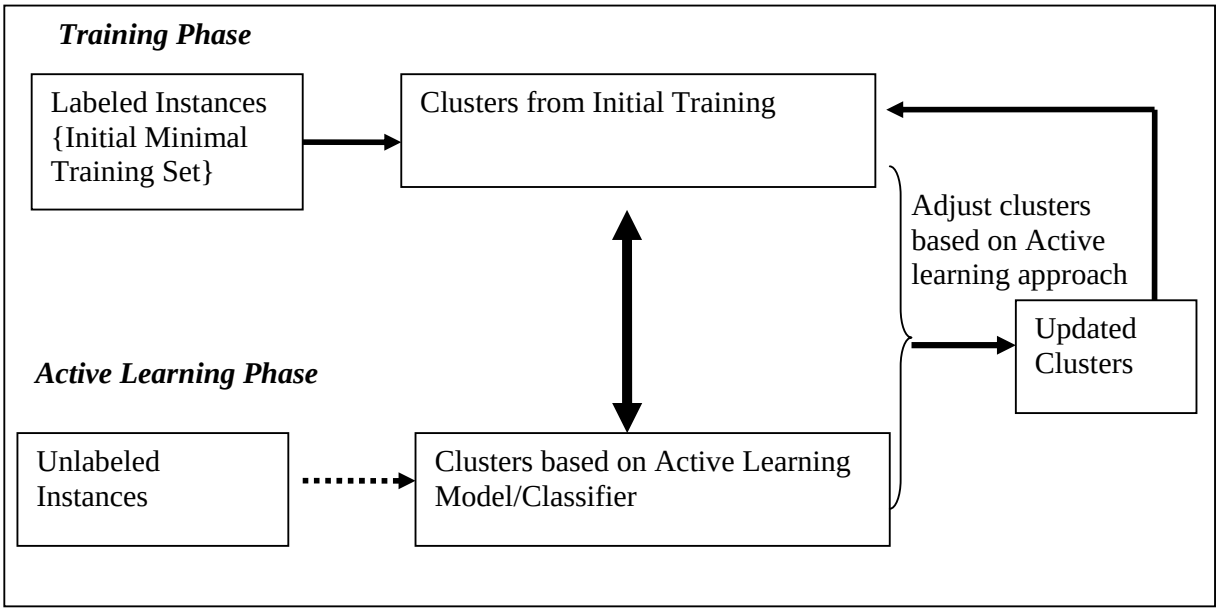

Figure 2: Proposed Active Learning Approach

To achieve this we are proposing a cluster based active learning approach and to identify proper clusters we are aiming for the best defining features for sample data set. For cluster definition we will be extending feature selection strategies and to further adjust clusters (for incremental learning), we will try to reduce the number of features by eliminating (based on our evaluation criteria) them till no further improvement is achieved. We will also compare our approach with different existing cluster based learning approaches.

VI. PROPOSED METHODOLOGY

Phase1-**Training Phase**: We identify dataset of sample size S, we will label initial dataset D with feature set F and its value vector , initially we will assume that features can take only binary value{0,1}. So for every labeled data point we will have a feature vector of 0,1s. We will group data-points with most similar feature vectors. The distance between features can be calculated using spectral clustering or some similar kernel function. While clustering we will try to set some threshold for dissimilarity and grouping of data points in same or different clusters.

Phase 2-(i) **Active learning Phase**: To incorporate active learning, we will employ learning algorithm which will pick up an unlabeled instance, algorithm will identify its feature vector and based on the previous learning, it will compare the feature vector for new instance with feature vectors for clusters, if it fit any we will label it, otherwise we will see how much distance exist and label it with the clusters where the similarity is maximum.

Phase 2-(ii): **Adjust clusters**, after a batch of say n unlabeled instances, we will again check the performance of clusters (by identifying some appropriate measure for it like uncertainty, density etc), if it is degraded then we will remove feature which is causing max no of feature dissimilarity. An overall idea of proposed research work is depicted in Figure 2.

VII. CONCLUSION AND FUTURE WORK

In our research we are trying to exploit the benefits of Active learning which maximizes the accuracy of the learned hypotheses while minimizing the amount of labeled training data and detect and ask expert to label only most informative examples.

We propose to show that our cluster based Active learning approach can be effectively used for the following

i. Reduce the number of training examples required to learn an accurate model, which is all the more important in real world setting where labeling incurs a cost and unlabeled data is in abundance.
ii. Clustering to achieve faster data reduction and will achieve broad classification, we will also employ feature selection method to restrict to the cluster quality and numbers.
iii. Combining both approaches for improved results and accurate model with lower costs.
iv. We plan to implement our framework to validate our assumptions on various datasets and compare the performance of our method with other existing approaches on datasets from various domains.

VIII. REFERENCES

[1] A.L.Blum and P. Langley, "Selection of relevant features and examples in machine learning", *Artificial Intelligence*, 97:245–271, 1997.
[2] A Huang, M David , E Frank & I Witten, "Clustering Document with Active Learning using Wikipedia", 2008
[3] A.Y. Ng., "On feature selection: learning with exponentially many irrelevant features as training examples", In Proceedings of the Fifteenth International Conference on Machine Learning, 404–412, 1998.
[4] Burr Settles, "Active Learning Literature Survey", Computer Sciences Technical Report, 1648, University of Wisconsin–Madison. 2009.
[5] Campedel ,M, Kyrgyzov I and Maˆıtre, H., "Consensual clustering for unsupervised feature selection-Application to SPOT5 satellite images indexing", Journal of Machine learning Research,4,48-59,2008.
[6] C.. Urmson et al. Autonomous driving in urban environments: Boss and the urban challenge. Journal of Field Robotics, 25(8), 425-466, 2008.
[7] D.Cohn, Atlas, L., and Ladner, R.., "Improved generalization with active learning", Machine Learning, 15:201-221, 1994.
[8] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models", Journal of Artificial Intelligence Research, 4:129-145, 1996.
[9] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with mixture models- In Multiple model approaches to modeling and control", Taylor and Francis, 1997.
[10] D.Angluin, "Queries and concept learning", Machine Learning, 2(4):319{342, 1988.
[11] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In Proceedings of the Seventeenth Annual ACM-SIGR Conference on Research and Development in Information Retrieval,pages 3-12, 1994.
[12] E.. Leopold and Kindermann J. Text categorization with support vector machines."How to represent texts in input space", Machine Learning, 46:423–444, 2002.
[13] E.Xing, M.Jordan and R. Karp, "Feature selection for high- dimensional genomic microarray data", In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.
[14] G, Schohn & D.Cohn, "Less is more: Active learning with support vector machines" In *Proceedings of 17th International Conf on Machine Learning*, 839-846, Morgan Kaufmann, CA.
[15] G.Tur, D. Hakkani- Tur, and R.E.Schapire, "Combining active and semi-supervised for spoken language understanding", Speech Communication, 45(2):171–186, 2005.
[16] H.Joshi X. Xu, "Using Active learning with Integrated Feature selection", Technical Report UALR02,2006.
[17] H. Liu and H. Motoda, "Feature Selection for Knowledge Discovery and Data Mining" Boston: Kluwer Academic Publishers, 1998.
[18] H Nguyen and A Smeulders, "Active learning using pre clustering", In *Proceedings of 21st International Conference on Machine learning*, Banff, Canada,79,2004
[19] H Raghavan, O Madani and R Jones, "Interactive feature selection", IJCAI-19, 841-846, 2005.
[20] H.Seung, M.Opper and H.Sompolinksy, "Query by committee", In *Proceedings of the Fifth ACM Workshop on Computational Learning Theory*, 287-294,1992.
[21] J.G.Dy and C.E.Brodley, "Feature subset selection and order identification for unsupervised learning", In *Proceedings of the Seventeenth International Conference on Machine Learning*, 47–254, 2000.
[22] J.G.Dy and C.E. Brodley, "Visualization and interactive feature selection for unsupervised data", In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 360–364, 2000.
[23] K.S. Ng and H. Liu., "Customer retention via data mining", AI Review,14(6):569 – 590, 2000.
[24] K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM", Machine Learning, 39:103–134, 2000.
[25] Liu H & Yu.L, "Efficient feature selection via analysis of relevance and redundancy", Journal of Machine Learning research, 5, 1205-1224, 2004.
[26] L. Talavera, "Feature selection as a preprocessing step for hierarchical clustering", In *Proceedings of International Conference on Machine Learning*, ICML'99, 389–397, 1999.
[27] M.A .Hall, Correlation Based Feature Selection for Machine Learning, P.hd thesis, University of Waikato, Dept. of Computer Science, 1999.
[28] M. Dash and H. Liu., "Feature selection for classification", Intelligent Data Analysis: An International Journal, 1(3):131–156, 1997.
[29] M.Dash and H. Liu., "Feature selection for clustering", In *Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining*, PAKDD-2000,. Kyoto, Japan, 110–121, Springer-Verlag, 2000.

[30] M.Dash., H.Liu and J.Yao, "Dimensionality reduction of unsupervised data", In *Proceedings of the Ninth IEEE International Conference on Tools with AI*, ICTAI'97, 532-539, Newport Beach, California,1997.
[31] Mallapragada P, Jin R & Jain A, "Active Query selection for semi supervised clustering" ,ICPRIEEE, 1-4,2008.
[32] McCallum, A., & Nigam, K., "Employing EM and pool-based active learning for text classification" In *Proceedings. of 15th Intl. Conf. on Machine Learning* ,ICML-98,. Madison, WI: Morgan Kaufmann.,1998.
[33] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning", In *Proceedings. of the 23rd International Conference on Machine Learning*, 2006.
[34] M.-F. Balcan, E. Even-Dar, S. Hanneke, M. Kearns, Y. Mansour, and J.Wortman, "Asymptotic active learning" In *NIPS Workshop on Principles of Learning Problem Design*, 2007.
[35] M. Berry and G. Lino, "Data Mining Techniques: For Marketing, Sales, and Customer Support", John Wiley and Sons, Inc., 1997.
[36] M.Seeger, "Learning with labeled and unlabeled data", Technical report, Edinburgh University, 2001.
[37] O.. Chapelle, J Weston and B. Schölkopf, "Cluster kernels for semi- supervised learning", Advances in Neural Information Processing Systems, 2002.
[38] O.. Mangasarian, W.N. Street, and W. Wolberg, "Breast cancer diagnosis and prognosis via linear programming", Operations Research, 43(4):570–577, 1995.
[39] P.Langley, "Selection of relevant features in machine learning", In *Proceedings of the AAAI Fall Symposium on Relevance*, 140–144. AAAI Press, 1994.
[40] R.Agrawal and R. Srikant , "Fast algorithms for mining association rules", In *Proceedings of Int. Conf. of Very Large Data Bases*, 487–499, Santiago, Chile,September 1994.
[41] R. Castro and R. Nowak, "Minimax bounds for active learning", In *Proceedings of the 20th Conference on Learning Theory*, 2007.
[42] R. Duda, P. Hart, and D. Stork, "Pattern Classification", Wiley-Interscience, Neural Information Processing Systems (NIPS), volume 15, 2003.
[43] R.Liere and P. Tadepalli, "Active learning with committees for text categorization" In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 91-596, 1997.
[44] S. Dasgupta, A. Kalai, and C. Monteleoni, "Analysis of perceptron-based active learning" In *Proceedings of the 18th Conference on Learning Theory*, 2005.
[45] S. Dasgupta, D. Hsu, and C. Monteleoni, "A general agnostic active learning algorithm", Technical Report CS2007-0898, Department of Computer Science and Engineering, University of California, San Diego, 2007.
[46] S. Hanneke. Teaching dimension and the complexity of active learning. In Proceedings of the 20th Conference on Learning Theory, 2007
[47] S. M. Weiss and C. A. Kulikowski, Computer Systems that Learn, Morgan Kaufmann, 1991.
[48] Shen, X, & Zhai, C., "Active feedback- UIUC TREC- 2003 HARD experiments", The 12th Text Retrieval Conference, TREC, 2003.
[49] T. Hofmann and J.M. Buhmann, "Active data clustering", In Advances in Neural Information Processing Systems (NIPS), volume 10, 528—534, Morgan Kaufmann, 1998.
[50] Tang, M., Luo, X., & Roukos, S. , "Active learning for statistical natural language parsing", In *Proceedings of the Association for Computational Linguistics*, 40th Anniversary Meeting, Philadelphia, PA,2002.
[51] Tong S, Active Learning: Theory & Applications, Ph.D Dissertation, 2001.
[52] Tong, S. & Koller,D., "Support vector machine active learning with applications to text classification", Journal of Machine learning Research,2001,45-66.
[53] V.Roth & T.Lange, "Feature selection in clustering problems", Advances in Neural Information Processing Systems, 16, 2004.
[54] Xu Z, Yu K, Tresp, V, Xu, X & Wang,J., "Representative sampling for text classification using support vector machines", *25th European Conf. on Information Retrieval Research*, ECIR, Springer,25-32,2003
[55] X. Zhu. Semi-Supervised Learning with Graphs,Ph.D thesis, Carnegie Mellon University, 2005a.
[56] Y.Freund, H. Seung, E.Shamir and N.Tishby, "Selective sampling using query by committee algorithm", Machine Learning, 28:133-168, 1997.
[57] Y.Kim ,W.Street and F. Menczer, "Feature selection for unsupervised learning via evolutionary search", In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 365-369, 2000.
[58] Zhang, C., & Chen, T., "An active learning framework for content-based information retrieval", IEEE transactions on multimedia, 4, 260–268, 2002.