Combining Discovery and Maintenance Processor in Active Learning Query Processor Model

S.Kavitha[1], Mr.P.Senthil Kumar[2]

[1]PG student, Department of Computer Science and Engineering, S.Veerasamy Chettiar College of Engineering and Technology Puliangudi-627 855

[2]Assistant Professor, Department of Computer Science and Engineering, S.Veerasamy Chettiar College of Engineering and Technology Puliangudi-627 855

Abstract

Classification is the solution to many problems. Many real world problems contain uninteresting common classes along with interesting rare classes. The rare classes are often needed to be discovered while training a classifier. We propose an active learning method for scenarios with unknown, rare classes, where the problems of classification and rare class discovery need to be tackled jointly. To switch generative and discriminative classifiers, we used a multi-class generalization of unsupervised classification entropy. Classifier learning in the presence of undiscovered classes was achieved by formulating a new model driven by an adaptive mixture of new class seeking and multiclass entropy maximization. In our evaluation on nine data sets of widely varying domain, size, and dimension, our model was consistently able to adapt query criteria and classifier online as more data was obtained, thereby outperforming other contemporary approaches making less efficient use of their active query budget (notably non adaptively iterating over criteria, or sequentially applying discovery and then learning criteria). We therefore expect our approach to be of great practical value for many problems. Our active learning approach is also cheap compared to alternative active learning criteria. Our approach is also compatible with sub sampling techniques for pool-based active learning such as the "59 trick," which defines a constant time approximation to the full algorithm.

Index Terms— *Co ordination Registration, Knowledge Discovery, Data Mining, Discriminative Model Pair*

I.Introduction

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD),an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and updating. The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages: Selection Pre-processing, Transformation, Data Mining, and Interpretation/Evaluation. Data mining involves six common classes of tasks: Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation. Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis. Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

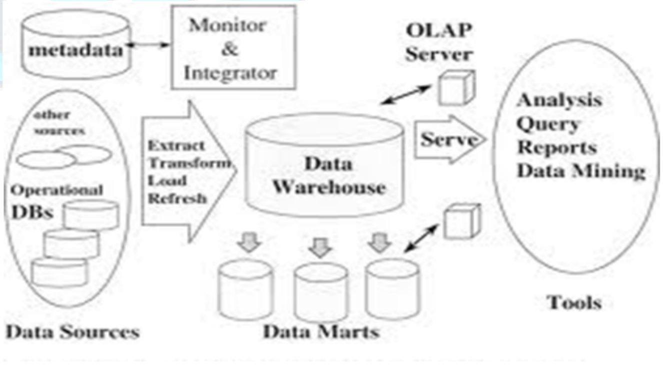

Fig 1. Data Mining Architecture

In [1], the empirical results to demonstrate 1) the effectiveness of the pKNN framework on large multi-class problems, 2) how active learning can guide the learning procedure to select critical examples to be labeled, and 3) the ability of the proposed framework to learn a good kernel function. Unfortunately, existing active learning methods for multi-class problems are inherently binary methods and do not scale up to a large number of classes. A probabilistic variant of the K-Nearest Neighbor method for classification that can be seamlessly used for active learning in multi-class scenarios. Unlike existing metric/kernel learning methods, our scheme is highly scalable for classification problems and provides a natural notion of uncertainty over class labels. The probabilistic nature of the formulation allowed us to seamlessly incorporate an active learning strategy into our framework.

A novel Markov Clustering Topic Model (MCTM) was introduced in [2] which builds on the strength of existing DBNs and PTMs, but crucially is able to overcome their drawbacks on accuracy, robustness and computational efficiency. In particular, the model makes two important novel contributions to LDA: (1) Hierarchical modeling, allowing simple actions to be combined into complex global behaviors; and (2) temporal modeling, enabling the correlation of different behaviors over time to be modeled. By introducing a Markov chain to model behavior dynamics, this model defines a DBN generalization of LDA. Learning from unlabeled training data is performed offline with Gibbs sampling; and a novel Bayesian inference algorithm enables dynamic scene understanding and behavior mining in new video data online and in real time.

Category detection is an emerging area of machine learning that can help address this issue using a" human-in-the-loop" approach. In this interactive setting, the algorithm asks the user to label a query data point under an existing category or declare the query data point to belong to a previously undiscovered category. The goal of category detection is to bring to the user's attention a representative data point from each category in the data in as few queries as possible. In a data set with imbalanced categories, the main challenge is in identifying the rare categories or anomalies; hence, the task is often referred to as rare category detection. A rare category detection based on hierarchical mean shift. A hierarchy is created by repeatedly applying mean shift with an increasing bandwidth on the data. The main ad-vantage of this methodology over existing approaches is that it does not require any knowledge of the dataset properties such as the total number of categories or the prior probabilities of the categories. HMS approach discovers all the categories in the datasets used in our experiments in much fewer queries than existing approaches such as Interleave and NNDM.

Most existing active learning studies assume that all classes are known a priori. Active learning is a special case of semi-supervised machine learning in 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/teacher 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 (e.g., conflict and ignorance) with adaptive, incremental learning policies in the field of Online machine learning.

The rest of the paper is organized as follows. Section II describes the proposed methodology in detail. Section III deals with the experimental results obtained. Section IV covers the conclusion and future works of the paper.

II.PROPOSED METHODOLOGY

Recently there have been a number of works that explicitly focus on the rare class discovery problem. We address joint discovery and classification by adaptively balancing multiple criteria based on their success both at discovery and improving classification. Specifically, we propose to build a generative-discriminative model pair because as we shall see, generative models naturally provide good discovery criteria and discriminative models naturally provide good classifier learning criteria. As a second contribution, we note that depending on the actual supervision cost and sparsity of rare class examples, the availability of labels will vary across data sets and classes. Given the nature of data dependence in generative and discriminative models (in which generative models are often better with very little data; and discriminative models are often better asymptotically) the better classifier will vary across both the data set and the stage of learning. We address this uncertainty by proposing a classifier switching algorithm to ensure the best classifier is selected for a given data set and availability of labels. Evaluation on a batch of vision and UCI data sets covering various domains and complexities shows that our approach consistently and often significantly outperforms existing methods at the important task of simultaneous discovery and classification of rare classes.

Incremental GMM Estimation

The constant time incremental agglomerative algorithm is used for online GMM learning. For the first n=1…N training points observed with the same label y, a model P(Xly) is incrementally built for y using kernel density estimation with Gaussian kernels and weight w_n=1/n *d is the dimension of x

$$p(\mathbf{x}|y) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \cdot \sum_{n=1}^{N} w_n \exp -\frac{1}{2}((\mathbf{x}-\mathbf{x}_n)^T\Sigma^{-1}(\mathbf{x}-\mathbf{x}_n)).$$

After reaching some maximal number of Gaussian Kernels N_{max}, we merge two existing Gaussian kernels i and j by moment matching. The components to be matched are chosen by selecting the pair (Gi, Gj) in terms of the Kullback – Leibler divergence.

SVM

We use a predefined SVM approach with RBF kernels, treating multi class classification as a set of 1-Vs-all decisions.

Adapting Active Query Criteria:

We have to first analyze how to adaptively combine the query criteria online for discovery and classification. The following algorithm involves probabilistically selecting a query criterion according to some weights and then sampling the query point from the distribution. The weights are adapted based on the discovery and classification performance of our active learner at each iteration.

Adaptive Selection of Classifiers

Though the Generative GMM classifier has better initial performance and the discriminative SVM classifier has better asymptotic performance, the best classifier varies with data set and active learning. Because of lack of data, it is not easy to determine reliability using cross validation. We choose a simpler but more robust approach with switches as the final classifier at the end of each iteration to the one with higher MCE, aiming to perform as a better classifier. After each training iteration, we compute multi class classification entropies over the train set U.

The process of multi class posterior estimation for SVMs requires cross validation and is inaccurate with limited data. At each iteration, to compute the uncertainty criterion, a posterior of the classifier determined to be more reliable by MCE is used, instead of using the discriminative model posterior.

Query Strategies

Algorithms for determining which data points should be labeled can be organized into a number of different categories:

-Uncertainty sampling: label those points for which the current model is least certain as to what the correct output

-Query by committee: a variety of models are trained on the current labeled data, and vote on the output for unlabeled data; label those points for which the "committee" disagrees the most.

-Expected model change: label those points that would most change the current model.

-Expected error reduction: label those points that would most reduce the model's generalization error.

Generative Model

A generative model is a model for randomly generating observable data, typically given some hidden parameters. It specifies a joint probability distribution over observation and label sequences. Generative models are used in machine learning for either modeling data directly (i.e., modeling observations drawn from a probability density function), or as an intermediate step to forming a conditional probability density function. A conditional distribution can be formed from a generative model through Bayes' rule.

Generative models contrast with discriminative models, in that a generative model is a full probabilistic model of all variables, whereas a discriminative model provides a model only for the target variable(s) conditional on the observed variables. Thus a generative model can be used, for example, to simulate (i.e. generate) values of any variable in the model, whereas a discriminative model allows only sampling of the target variables conditional on the observed quantities. Despite the fact that discriminative models do not need to model the distribution of the observed variables, they cannot generally express more complex relationships between the observed and target variables. They don't necessarily perform better than generative models at classification and regression tasks.

The following are the modules of the proposed method. It includes,

1. Coordination registration
2. User registration
3. Author-active learning process
4. Discovery and evaluation process
5. Quantitative performance analysis

A) Coordination Registration

In this module the coordinator can able to access the whole details of the databases which are being stored in it. The coordinator can able to view the overall process of the group details, marks and progress. The coordinator has the full authorization over the databases so that he/she can able to modify the data which are stored in it.

The system architecture of the proposed methodology is shown below.

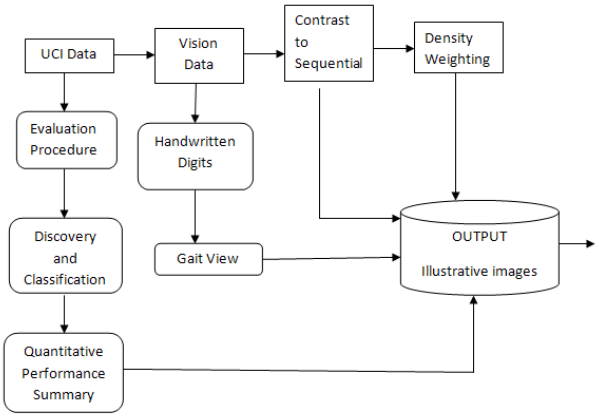

Fig 2. System Architecture of the Proposed System

B) User Registration

The user registration is the examiner those who attend the exams in it. The user should register the student name, user id, password, qualification, select the group type they want to attend, mail, security question with its answers so that if they forget the username or password they can able to retrieve from answering the security question. Then submit to the data to store the information in the database. After the user registration is completed the user can able to start the examination.

C) Author-Active Learning Process

An author can perform three process group result, task assigning, task valuation. In the group result process the author view the results of the marks obtained by various examiners in various type of groups. An author can modify the tasks assigned for the group members in various groups. The author can also able to update or modify the repository which are stored. The author assigns various tasks to the individual's users in the various groups.

Active learning is traditionally applied to classification, as we do, but can also be applied to regression (MacKay, 1992). MacKay (1992) is actually concerned with experiment design, where one selects the most informative scientific experiments to run with limited time/budget, an area closely related to active learning. Transfer learning is related in some situations, by virtue of handling the relationship between known and unknown classes. An example of this is Lee and Grauman (2010), which uses the relationship between known classes and unknown classes to automatically infer the unknown classes, ready for human verification followed by further learning. Reinforcement learning (Kaelbling et al., 1996) is also closely related to the presented kind of active learning, via the exploration-exploitation problem.

Stopping Conditions:

Active learning is concerned with limited resources - the fact that it takes time/money/energy to provide ground truth information for a classification algorithm. Eventually the querying has to stop. Three common options can be considered: - Query budget: A fixed number of queries are performed. -Sufficient performance: Enough queries are performed for classification performance to surpass a threshold. It can be estimated using n fold cross validation once enough queries have been performed to get an accurate enough estimate. - Cost-benefit analysis. In many situations misclassification can have a directly attributed cost, as can providing further labeled exemplars - the total cost can then be minimized. To exemplify a widget factory may have a classifier to detect faulty products, alongside a given defect rate. The defect rate multiplied by the false negative rate of the classifier will give the percentage of faulty products sent to customers - multiply this by the sales projections and the cost of handling a return and you obtain the money wasted by the classifiers mistakes.

The false positive rate should also be factored in, in terms of throwing out usable widgets. Given the cost in employee time to train the classifier we can now work out at what point the cost of further training exceeds the value obtained (For a given product lifespan.), and hence when to stop training. Complex effects can exist, e.g. sending customers faulty products can generate bad publicity, making sales a function of the classifiers false probability rate.

The choice of scheme is scenario specific however, and as such we will not be exploring it further. However, by presenting results to a deep enough query count the above stopping conditions are implicitly represented using graphs of inlier rate against query count. Query budget is represented by seeing which is highest after a given number of queries (a vertical line), whilst performance is given by which algorithm crosses a given threshold first (a horizontal line). A cost benefit analysis is often represented by a straight line at an angle set by the relative costs of failure and further training. More sophisticated cost-benefit models can generate an arbitrary curve.

D) Discovery and Evaluation Process

Active learning is a process whereby students engage in activities, such as reading, writing, discussion, or problem solving that promote analysis, synthesis, and evaluation of class content. Cooperative learning, problem-based learning, and the use of case methods and simulations are some approaches that promote active learning.

The dataset consists of the following classification problems:

glass: Infer glass type given its chemical contents, for forensic investigation. Features include chemical properties and how it breaks.

Ecoli: Predict which part of a cell contains a protein localization site, for E.coli.

Segment: Labeling regions from images of outdoor scenes, with labels such as grass, path and sky. Input is a small patch of pixels; output is the label for the centre pixel of the patch.

Page blocks: Classifying regions from document scans, e.g. as text, picture or graphic. Features include color ratios and measures of texture.

Cover type: Predicting forest cover type given geographic information, such as elevation and soil type.

Thyroid: Determining the disease that a thyroid has given observed and measured properties.

Wine quality: Predict the quality of Portuguese wine given various chemical properties. Strictly speaking this is a quantized regression problem.

Letters: Recognizing handwritten letters from the English alphabet. Input is images of each letter.

Shuttle: Infer the state of part of the space shuttles propulsion system, given various sensor readings, as relating to the Challenger disaster.

kdd99: Data set used for the 3rd Knowledge Discovery and Data Mining Tools Competition uses asimulation of a military network with the goal being to detect intrusions given tcp dump data. The original data set included multinomial attributes, which have been concatenated as part of the feature vector, hence the high dimensionality of the problem.

Gait: Inferring the quantised walking direction from aligned silhouettes that have been averaged over multiple frames (input is a greyscale image), as in Han and Bhanu (2006). This data set was sampled to be imbalanced, such that each class is half the size of the next larger.

Digits: Recognising the handwritten digits, 0 to 9, given images of the digits. This data set was sampled to be imbalanced, such that each class is half the size of the next larger.

E) Quantitative Performance Analysis

In this module we can able to find out the top scorer for the various groups. We can also view them in group wise designation. We can also able to view the overall performance of the various groups.

Algorithm: Active Learning For Discovery and Classification

```
Input: Initial labeled L and unlabeled U samples.
Classifiers {fc}, query criteria {Qk}, weights w.
1) Build unconditional GMM p(x) from L U u
2) Estimate _ by cross-validation on p(x)
3) Train initial GMM and SVM classifiers on L
Repeat as training budget allows:
1) Compute query criteria plik (i) and punc (i)
2) Sample query criteria to use k ~Multi (w)
3) Query point i*~ pk (i), add (xi*, yi*) to L
4) Update classifiers with label i*
5) Update query criteria weights w
6) Compute entropies Hgmm and Hsvm
7) If Hgmm > 'Hsvm: select classifier fgmm(x)
8) Else: select fsvm(x)
```

III.EXPERIMENTAL RESULTS

The method was tested on seven datasets from the UCI repository. These data sets were selected because they contained multiple classes in naturally unbalanced proportions, thereby representing real discovery and classification problems. In every case, we started with one labeled point from the largest class and the goal was to discover and learn to classify the remaining classes Table 1 summarizes the properties of each data set. Performance was evaluated by two measures at each active learning iteration: 1) the percentage of distinct classes in the training data set discovered and 2) the average classification accuracy over all classes. Note that in contrast to (1), this accuracy measure ensures that ability to classify each rare class is weighted equally with the majority class despite the fewer rare class points. The standard approach to quantitatively summarizing the (time varying) performance of active learning algorithms is to compute the area under their classification curve (AUC) during learning. Of the comparison models, there is no consistent best performer with G/G, S/S, S/GSmix, and S/GSonline performing best on three, one, two, and one data sets, respectively. Moreover, each model performs poorly (last or second to last) on at least one data set. This supports our earlier insight that a big challenge of this problem is the strong data set dependence of the ideal query criterion.

TABLE I
UCI DATASET PROPERTIES

DATASET	N	D	NC	S%	L%
ECOLI	325	7	8	1.42	42
PAGE BLOCKS	5473	10	5	0.5	90
GLASS	214	10	6	4	36
COVER TYPE	5000	10	7	3.6	25
SHUTTLE	20000	9	7	0.01	98
THYROID	7200	21	3	2.5	92

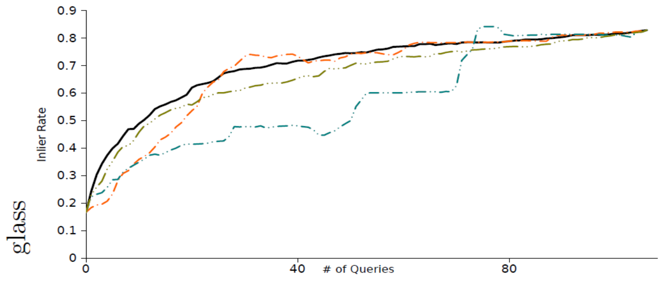

Fig 3. Graph of Inlier Rate Against Number of Queries

IV. Conclusion and Future Work

The algorithm for active learning to classify a priori undiscovered classes based on adapting two query criteria and choosing classifiers. To switch generative and discriminative classifiers, we used a multi-class generalization of unsupervised classification entropy. Classifier learning in the presence of undiscovered classes was achieved by formulating a new model driven by an adaptive mixture of new class seeking and multiclass entropy maximization. In our evaluation on nine data sets of widely varying domain, size, and dimension, our model was consistently able to adapt query criteria and classifier online as more data was obtained, thereby outperforming other contemporary approaches making less efficient use of their active query budget (notably non adaptively iterating over criteria, or sequentially applying discovery and then learning criteria). We therefore expect our approach to be of great practical value for many problems. Our active learning approach is also cheap compared to alternative active learning criteria. Our approach is also compatible with sub sampling techniques for pool-based active learning such as the "59 trick," which defines a constant time approximation to the full algorithm.

There are various interesting questions for future research including: further theoretical analysis and grounding of the joint discovery-classification problem and algorithms introduced here; how well our fusion methods generalize to other generative-discriminative pairs and query criteria; and how to create tighter coupling between the generative and discriminative classifiers. A final key goal is to generalize some of the contributions we have discussed in this paper to the domain of online—rather than pool-based active learning, which is a more natural setting for some practical problems where online real-time classification is required and new classes may appear over time.

References

[1] P. Jain and A. Kapoor, "Active Learning for Large Multi-Class Problems," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 762-769, 2009.

[2] T. Hopedale's, S. Gong, and T. Xiang, "A Markov Clustering Topic Model for Behavior Mining in Video," Proc. IEEE Int'l Conf. Computer Vision, 2009.

[3] P. Vatturi and W.-K. Wong, "Category Detection Using Hierarchical Mean Shift," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 847-856, 2009.

[4] Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classi_ers. Proc. Conf. on Research and Development in Information Retrieval.

[5] Huang, G. B., Ramesh, M., Berg, T., and Learned Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

[6] Settles, B. (2009). Active learning literature survey. Technical Report 1648, Uni. of Wisconsin-Madison

[7] Angluin, D. (1988). Queries and concept learning. Machine Learning, 2(4):319{342.

[8] Blackwell, D. and MacQueen, J. B. (1973). Fergusondistributions via Polya urn schemes. Annals of Statistics, 1(2):353{355.

[9] Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classi_ers. Computational Learning Theory, 5:144{152

[10] Breiman, L. (2001). Random forests. Machine Learning, 45(1):5 - 32.

[11] Cohn, D., Atlas, L., and Ladner, R. (1994). Improving generalization with active learning. Machine Learning, 15(2):201-221

[12] Culotta, A. and McCallum, A. (2005). Reducing labeling e_ort for structured prediction tasks. Proc. Nat.Conf. Arti_cial Intelligence, pages 746-751

[13] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. J. American Statistical Association, 90(430):577 – 588

[14] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1(2):209 – 230

[15] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics,7(2):179 - 188.

[16] Guillaumin, M., Verbeek, J., and Schmid, C. (2009). Is that you? metric learning approaches for face identification. ICCV.

[17] Haines, T. S. F. and Xiang, T. (2011). Active learning using dirichlet processes for rare class discovery and classification. BMVC.

[18] Han, J. and Bhanu, B. (2006). Individual recognition using gait energy image. Pattern Analysis and Machine Intelligence, 28(2):316 - 322.

[19] He, J. and Carbonell, J. G. (2007). Nearest-neighbor based active learning for rare category detection. Neural Information Processing Systems, 21.Ho, T. K. (1995). Random decision forests. Proc. Document Analysis and Recognition, 1:278 - 282.

[20] Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2):85-126.