ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

RNN based Batch Mode Active Learning Framework

Gaurav Maheshwari and Vikram Pudi
gaurav_m@students.iiit.ac.in, vikram@iiit.ac.in
Centre for Data Engineering, IIIT Hyderabad, India

Abstract. Active Learning has been applied in many real world classification tasks to reduce the amount of labeled data required for training a classifier. However most of the existing active learning strategies select only a single sample for labeling by the oracle in every iteration. This results in retraining the classifier after each sample is added which is quite computationally expensive. Also many of the existing sample selection strategies are not suitable for the multi-class classification tasks. To overcome these issues, we propose an efficient batch mode framework for active learning using the notion of influence sets based on Reverse Nearest Neighbor, which is applicable for multi-class classification as well. To demonstrate the effectiveness of our technique, we compare its performance against existing active learning techniques on real life datasets. Experimental results show that our technique outperforms existing active learning methods significantly especially on multi-class datasets.

1 Introduction

With rapid technological advancement in recent years, we have witnessed an explosive growth in the amount of data available to us. In many practical classification tasks we have a large amount of unlabeled data, but labeling this data is often very costly or time-consuming. Active Learning is a popular technique for overcoming this labeling bottleneck, which aims at reducing the amount of labeled data required for training a supervised classifier to achieve satisfactory performance. The key idea behind active learning is that a machine learning algorithm can obtain greater accuracy with fewer labeled samples if it is allowed to choose the data from which it learns [1]. An active learner obtains the label of an unlabeled sample from an oracle (expert) by asking queries. An active learner reduces the amount of labeled samples required for effective learning by selecting only the *most informative samples* for labeling by the oracle.

One of the major issues with the existing active learning approaches is that they select only a single sample for labeling in each iteration. This results in retraining of the model after a labeled example is added, which becomes computationally expensive. Another major problem is that many suggested approaches work well for the binary classification tasks but are not extensible or do not perform well for multi-class classification problems. Also many of the existing Active Learning techniques are applicable only for some specific classification model and hence couldn't be applied to other classification algorithms. To overcome the above issues, we propose a new framework for active learning based on RNN influence set that is able to select a batch of unlabeled examples simultaneously for labeling efficiently. The proposed algorithm also handles multi-class classification task.

The rest of the paper is organized as follows :- Section 2 presents work done in this field, Section 3 describes RNN and presents motivation for our approach. Section 4 describes our batch mode active learning framework. Section 5 presents the proposed algorithm. Section 6 describes the experiments performed and the results of our empirical studies. In Section 7, we conclude by offering our observations as well as suggestions for potential future work.

2 Related Work

Uncertainty sampling [2] is the most commonly used Active Learning technique which selects those samples for labeling by the oracle, whose predicted class labels are the least certain. The most widely used uncertainty measure considers samples which lie closest to the classification boundary as the most uncertain ones, since their predicted class labels are more likely to be

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

incorrect than other samples. A major drawback of the uncertainty sampling technique is that since it selects those samples that are closest to the decision boundary, it is often prone to selecting outliers for labeling.

Density weighted sampling techniques on the other hand try to select informative points for labeling by oracle that are representative of the underlying data distribution. Representative sampling [3] is a density based sampling technique which uses the k-means algorithm to cluster the samples lying within the margin of a Support Vector Machine classifier trained on the current labeled set. The samples at cluster centers are the ones selected for labeling. A general behavior observed in density weighted sampling technique is that they improve accuracy in the initial phase at a rapid rate. However, after the initial gains, they exhibit very slow additional learning, while uncertainty sampling has lower learning rate during the initial phase but it improves as the number of labeled samples increase and gradually outperforms the density based sampling techniques[4]. SVM based sampling techniques have also been proposed that select samples that reduces the version space the most [5].

A common approach toward extending algorithms for batch mode active learning is to select *k most informative* samples as decided by the learning algorithm. However these approaches don't take into consideration the correlation among the selected samples. Hoi et al have proposed a framework for batch mode active learning that applies the Fisher information matrix to measure the overall informativeness for a set of unlabeled examples [6]. Brinker has proposed a framework for incorporating diversity in Active Learning with SVM [7]. However their approach is not extensible for multi-class classification task.

3 Reverse Nearest Neighbors

3.1 Definition

Conceptually reverse nearest neighbor query is the inverse of the nearest neighbor query. In a given dataset the reverse nearest neighbor set of point x is defined as the set of data points which considers x as their nearest neighbor. Similarly the k-reverse nearest neighbor set of x is defined as the set of data points which contains x in their k-nearest neighbor set. Let S be the set of points in the dataset. *RNN(x)* is formally defined as :-

$$RNN(x) = \{ r \in S \mid \forall p \in S : \mathrm{d}(r, x) <= \mathrm{d}(r,p) \} \quad (1)$$

Similarly *kRNN(x)* is defined as :-

$$kRNN(x) = \{ r \in S \mid x \in k\mathrm{NN}(r) \} \quad (2)$$

3.2 Motivation for using RNN

In this paper we propose a sampling framework that uses a RNN based metric to measure the informativeness of a sample. Our motivation in using RNN based approach for selecting informative samples lies in the observation that RNN set of a sample capture its *influence* on the dataset, i.e., RNN set of a sample is symbolic of its influence on other samples in the dataset [8]. Larger the size of RNN set for a sample denotes that it has a larger influence on other samples while the samples having smaller RNN sets have less influence on other samples. Thus by selecting a sample with large RNN set for labeling, we can infer the labels of a large no. of other samples correctly which will help in increasing the accuracy of the classifier. Also RNN influence set of samples are independent of the class labels and is thus easily applicable to multi-class setting also for selecting informative samples. Also the proposed framework requires only the class probabilities of the top 2 predicted classes for a sample to determine its informativeness and thus can be applied to a whole range of classification algorithms.

4 Batch Mode Active Learning

Our batch mode active learning framework selects a batch of most informative samples from the dataset that should be labeled by the oracle to significantly improve the accuracy of the classifier. The factors on which our batch mode selection strategy depends are described below.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

4.1 Uncertainty

It is one of the most important factors to be considered while selecting a sample for labeling. Samples that have high uncertainty w.r.t current classifier form suitable candidates for labeling by the oracle, since knowing the true label would help the model discriminate more effectively between them and thus improving the performance of the classifier. We use a slightly modified version of the uncertainty measure proposed by Joshi et al. [9] that considers uncertainty of a sample as the difference between posterior probabilities of the best and the second best predictions. The uncertainty of a sample x is defined below as:-

$$Uncertainty(x) = 1 - (P(y_1|x) - P(y_2|x)) \quad (3)$$

where, y1 and y_2 are the classes with the largest and second largest posterior class probabilities.

If the difference between the two best class predictions is small, it means that the model is more confused on the sample and thus it should have high uncertainty. This measure is equivalent to the entropy-based method in binary classification, but in multiclass setting the above method has shown remarkable improvements on several benchmark datasets [9].

4.2 Density

Density measure is another important factor to be considered while selecting a sample for labeling. A sample with high density or representativeness will have influence on a large number of elements in the dataset and hence selecting such point for labeling will improve the classifier substantially. We propose a RNN based measure that estimates the density of sample x in the dataset as:-

$$Density(x) = \frac{1+ | \text{RNNunlabel}(x) |}{1+ | \text{RNNlabel}(x) |} \quad (4)$$

where, $\text{RNN}_{\text{label}}(x)$ is the set of labeled samples in *kRNN(x)*
$\text{RNN}_{\text{unlabel}}(x)$ is the set of unlabeled samples in *kRNN(x)*.

The proposed measure estimate the density of an unlabeled sample x to be the ratio of its k-reverse nearest neighbor in unlabeled and labeled set. The reason for this strategy is that the sample selected for labeling should be dissimilar to the other selected examples while it should be similar to most of the unselected examples. Initially all the samples in $kRNN(x)$ are unlabeled i.e. they belongs to the $\text{RNN}_{\text{unlabel}}(x)$, but as labeling occurs some of these samples will get labeled and they will move from $\text{RNN}_{\text{unlabel}}(x)$ to $\text{RNN}_{\text{label}}(x)$. The measure takes into consideration that the informativeness of sample x should decrease if elements belonging to its *kRNN* set get labeled and they move to training set.

4.3 Diversity

The key idea behind our approach to select a batch of samples for labeling is that the selected samples should be as diverse from each other as possible so that they all provide unique information to the classification model. We propose a method to measure similarity between two samples based on their *kRNN* set as follows:-

$$Similarity(x, y) = \frac{1 + (| kRNN(x) \cap kRNN(y) |)}{\sqrt{(1+ | kRNN(x) |) * (1+ | kRNN(y) |)}} \quad (5)$$

Similarly we define the diversity between 2 samples x and y as:-

$$Diversity(x,y) = 1 - Similarity(x,y) \quad (6)$$

The proposed similarity measure is based on the reasoning that large size of $kRNN(x) \cap kRNN(y)$ implies that lot of samples consider both x and y among their k-Nearest Neighbors. Larger the extent of overlap between the influence sets of x and y, larger is the degree of their representing common samples in the dataset and hence larger is the probability of them providing similar information to the classification model.

Procedure 1 contains the method for selecting a batch of diverse samples from an input set S and an initial element contained in the batch. We apply a greedy approach to select a batch of samples such that it maximizes the diversity among those samples. In each iteration, one

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

sample is selected from the input set S to be added to the batch such that its minimum diversity from any other sample in the batch so far, is maximum among all the elements in set S.

Procedure 1: **Select Diverse**

Input: Sample s, Input Set S and batch-size b

$T = \{s\}$
min = {1}*|S|
for i = 1 to b-1 do
  max= 0
  for j = 1 to |S| do
    if (Diversity(T_i , S_j) < $\min_j$)
      $\min_j$ = Diversity(T_i , S_j)
    if ($\min_j$ > max)
      max=$\min_j$
      t=j
  end for
  $T = T \cup S_t$
  Remove S_t from S and $\min_t$ from min
end for
return T

5 Algorithm

Our batch mode active learning algorithm uses a new RNN-Uncertainty measure for determining the informativeness of a sample. The RNN-Uncertainty measure combines the margin based uncertainty measure with the RNN influence set based density measure to select samples that are highly uncertain as well as representative of the samples in the dataset. The samples having larger RNN-Uncertainty value are more informative. The RNN-Uncertainty measure for an unlabeled sample x is defined as:

$$\textit{RNN-Uncertainty}(x) = \textit{Uncertainty}(x) * \textit{Density}(x) \quad (7)$$

The idea behind our approach is to select a batch of samples that have high RNN-Uncertainty measure, such that diversity among them is maximized. The algorithm initially forms a set of size λ times the batch size of samples, having the largest RNN-Uncertainty values. From this set, the algorithm finally selects a batch of those samples for labeling which maximize the diversity among the elements in the batch. This is done so that we can select highly informative samples from the dataset that are also highly diverse from each other. The algorithm is presented below.

Algorithm 1 : **RNN based Batch Mode Active Learning**

Input : Initial labeled set L, Unlabeled set U, Independent Test Set T
Classifier C, Batch Size b, λ and k

Initialization: Calculate the k-reverse nearest neighbor set for every element in $L \cup U$

repeat
1: Train the classifier C on the examples in L.
2: Use Classifier C to label the unlabeled examples in U.
3: S = set of λ*b samples from U having largest RNN-Uncertainty values.
4: s be the sample in U having the highest RNN-Uncertainty measure
5: P = Select Diverse(s,S,b)
6: Label the elements in P
7: Augment L with the elements of P and remove them from U
until stopping criteria is met

6 Experiments

6.1 Experimental Setup

To evaluate the performance of the proposed RNN-Uncertainty sampling technique we have compared it with a baseline random instance selection technique, a non-batch myopic uncertainty sampling algorithm which selects the most uncertain sample and a batch-mode

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

active learning method representative sampling. We conducted experiments on following real life datasets available from UCI Machine Learning repository: *Diabetes*, *German*, *Heart*, *Hepatitis*, *Ionosphere*, *Sonar*, *Eye*, *Vowel and Letter Recognition*. We used WEKA toolkit for the preprocessing and classification task. We used Naive Bayes classifier as the base learner for Uncertainty and RNN-Uncertainty techniques, while we used LibSVM module available in WEKA to get the SVM classifier required for representative sampling. In all our experiments, we start with a very small labeled set containing one sample for each class. We randomly selected 2/3 of the remaining instances as the unlabeled set while using the remaining instances as testing set. The batch size used is 10 and we have set λ=2 and k=15 for computing *k*RNN set in our RNN-Uncertainty algorithm. All the results reported are averaged over 10 times repetition. For every dataset, we have plotted graph of accuracy (in %) vs. number of labeled samples for all sampling techniques for easier comparison. All the experiments were performed on 1.66Ghz intel core 2 duo processor with 2 GB of RAM.

6.2 Experimental Results

Figures 1-6 show the comparison result on binary datasets. As the results show, the performance of the proposed RNN-Uncertainty algorithm is better than the remaining techniques on all datasets. Also the performance of Random Sampling is significantly lower than the rest of the sampling algorithms which signifies the importance of active learning paradigm. On *Sonar* and *Ionosphere* dataset, the RNN-Uncertainty algorithm clearly outperforms the rest of the algorithms by a significant amount. On both these datasets, Representative algorithm shows good performance during the initial stages, but RNN-Uncertainty algorithm quickly overtakes it. On *Diabetes* dataset also, the performance of Representative sampling is better than the rest initially, but it saturates quickly. On *German* dataset, RNN-Uncertainty and Uncertainty sampling strategies have the best performance. On *Heart* dataset the performance of Uncertainty and Representative algorithms is almost similar while the RNN-Uncertainty algorithm has slightly better performance during the entire active learning process. Figures 7-9 show the comparison results on the multi-class datasets. Since Representative Sampling selects informative samples by clustering the points lying inside the boundary of SVM hyperplane, it is not easily extensible to multi-class setting where there are multiple separating hyperplanes and hence it is not included in our experiments on multi-class datasets. In all the experiments, RNN-Uncertainty algorithm easily outperforms the rest of the techniques over the entire range of the active learning process by a significant amount.

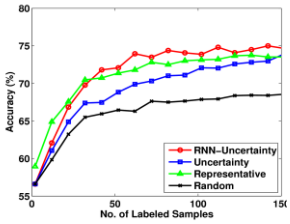

Figure 1 Diabetes

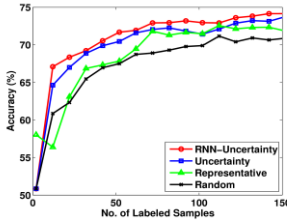

Figure 2 German

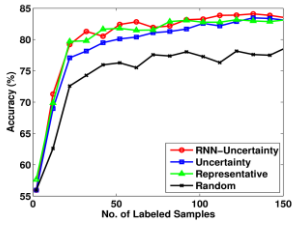

Figure 3 Heart

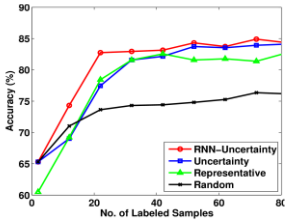

Figure 4 Hepatitis

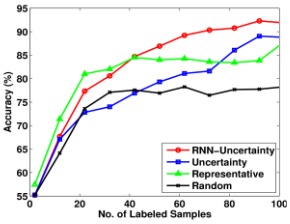

Figure 5 Ionosphere

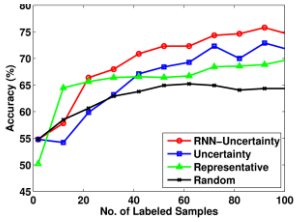

Figure 6 Sonar

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

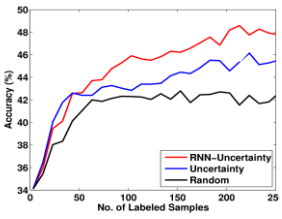

Figure 7 Eye

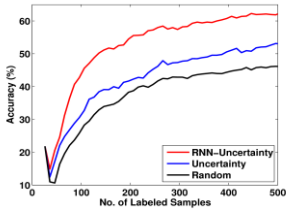

Figure 8 Letter

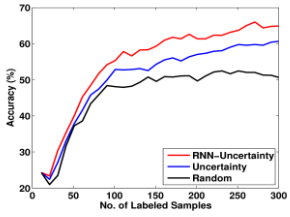

Figure 9 Vowel

Apart from this, we also measured time required for the active learning process on *Vowel* dataset which contains 990 instances. Table 1 contains the time required for the active learning process to complete for various batch sizes on vowel dataset. Increase in batch size results in less time for the active learning process to complete. The results indicate that time required for the active learning process is almost inversely proportional to the batch size. The better performance of RNN-Uncertainty algorithm in comparison to the existing techniques can be attributed to its hybrid approach which give due importance to all the significant factors affecting the informativeness of samples

b	1	5	10	20	50
time (sec)	452	95	50	27	14

Table 1. Time required for active learning on vowel dataset

7 CONCLUSIONS & FUTURE WORK

Our experiments show that proposed RNN based active learning framework outperforms previous approaches in active learning. The proposed method shows good performance on both binary as well as multi-class datasets. The proposed RNN-Uncertainty selection strategy is an efficient method to select batches of new training examples requiring only a small amount of additional computational time. The good performance of this measure can be attributed to the incorporation of both representative as well as uncertainty measures in deciding the informativeness of a sample along with the inclusion of diversity measure to avoid selecting similar samples for labeling. The main contribution of this paper is the development of an efficient batch mode active learning framework that can be applied to a large category of classifiers. An interesting observation from our experiments is that the accuracy of the proposed algorithm increases initially as no. of labeled samples increases. However after sufficient labeled samples have been added, further increase in no. of labeled samples no longer result in increase in accuracy of the classifier and sometimes even degrade its performance. Our future work would focus on investigating the reason behind the above observation and finding suitable criteria for stopping active learning process.

8 REFERENCES

[1] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[2] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In SIGIR '94, pages 3−12. Springer-Verlag New York, Inc., 1994.

[3] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang. Representative sampling for text classification using support vector machines. In ECIR '03, pages 393−407. Springer-Verlag, 2003.

[4] P. Donmez, J. G. Carbonell, and P. N. Bennett. Dual strategy active learning. In ECML '07, pages 116−127. Springer-Verlag, 2007.

[5] S. Tong and D. Koller. Support vector machine active earning with applications to text classification. In Journal of Machine Learning Research, pages 999−1006, 2000.

[6] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In ICML '06, pages 417−424. ACM, 2006.

[7] K. Brinker. Incorporating diversity in active learning with support vector machines. In ICML'03, pages 59−66. AAAI Press, 2003.

[8] F. Korn and S. Muthukrishnan. Influence sets based on reverse nearest neighbor queries. In, SIGMOD '00, pages 201−212. ACM, 2000.

[9] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR'09*, pages 2372−2379, 2009.