ACTIVE LEARNING IN TEXT CLASSIFICATION

YUYI GU and LINGJIA LI
Shanghai Jiao Tong University

In many real-world domains, supervised learning requires a large number of training examples. In this paper, we describe an active learning method that is classifier-agnostic and we have shown that it works across classifiers and feature representations. Our approach is similar to uncertainty sampling, where data with the least probability difference between the classifier giving the two most probable labels is queried. Our experiments are conducted in the text classification domain, which is characterized by a large number of features, many of which are irrelevant.

■

1. INTRODUCTION

The goal of *text classification* is to assign each document to the appropriate categories, based on the semantic content of the document. A knowledge engineering approach to text classification involves designing rules that correctly classify the documents. The goal is to develop **automatic** methods for text classification through the application of machine learning techniques. The amount of textual information that is available in electronic form has grown exponentially in recent years. Annotating documents for supervised learning has been a tedious, laborious, and time consuming task for humans. Given huge amounts of unlabeled documents, it is impractical for annotators to go over each document and provide a label. Automating the task of indexing, categorizing, and organizing these electronic documents will make it easier and cheaper for people to find relevant written materials.

Active learning refers to machine learning methods that allow the learning program to exert some control over the examples on which it learns [Cohn et al. 1994]. The purpose of this paper is to present results of experiments that demonstrate the effectiveness of active learning with uncertainty sampling, and also to analyze the sources of this effectiveness.

2. RELATED WORK

Active learning has long been a popular field that many researchers have contributed to it. "Active learning" in its most general sense refers to any form of learning where in the learning algorithm has some degree of control over the examples on which it is trained. One active learning approach is the membership query paradigm, in which the learner can construct new sets of inputs and request that the teacher provide their labels [Liere and Tadepalli 1997]. There have been some promising results in the active learning area. Cohn, Atlas, and Ladner developed the theory for an active learning method called selective sampling and then applied it to some small to moderate sized problems as a demonstration of the viability of this new approach[Cohn et al. 1994]. Lewis and Gale developed a method called uncertainty sampling, which is similar conceptually to selective sampling, but which is specifically meant for use in text categorization. Their method selects for labeling those examples whose membership is most unclear by using an approximation based on Bayes Rule, certain independence assumptions, and logistic regression. Since the method was developed for text categorization, it is able to handle noise as well as large numbers of features[Lewis and Gale 1994].

Dagan and Engelson use QBC stream-based sampling and vote entropy. In contrast, Andrew Kachites McCallumz et al. advocated density-weighted pool-based sampling and the KL metric.[McCallumzy and Nigamy 1998] They selected committee members using the Dirichlet distribution over classifier parameters, instead of approximating this with a Normal distribution.

Manali Sharma et al.(2015) present a simple and yet effective active learning approach that can incorporate rationales elicited from annotators into the training of any off-the-shelf classifier[Sharma et al. 2015]. They show that their simple approach is effective for multinomial naive bayes, logistic regression, and support vector machines.

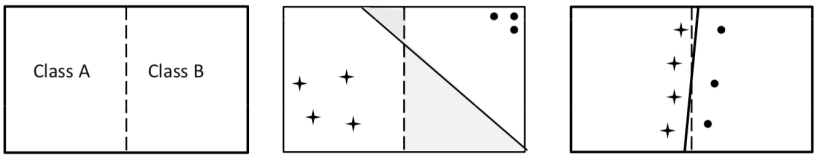
3. BRIEF OVERVIEW OF ACTIVE LEARNING

Active learning (sometimes called query learning or optimal experimental design in the statistics literature) is a sub field of machine learning and, more generally, artificial intelligence. The key hypothesis is that if the learning algorithm is allowed to choose the data from which it learns to be curious, if you will, it will perform better with less training.

While it is possible to query the data randomly for labels, such an approach may not result in the best model, when each query is costly, and therefore, few labels will eventually become available. For example, consider the two class example of Figure 1. Here, we have a very simple division of the data into two classes, which is shown by a vertical dotted line, as illustrated in Figure 1(a). The two classes here are labeled A and B. Consider the case where it is possible to query only 7 examples for the two different classes. In this case, it is quite possible that the small number of allowed samples may result in a training data that is unrepresentative of the true separation between the two classes. Consider the case when an SVM classifier is used in order to construct a model. In Figure 1(b), we have shown a total of 7 samples randomly chosen from the underlying data. Because of the inherent noisiness in the process of picking a small number of samples, an SVM classifier will be unable to accurately divide the data space. This is shown in Figure 1(b), where a portion of the data space is incorrectly classified, because of the error of modeling the SVM classifier. In Figure 1(c), we have shown an example of a well chosen set of seven instances along the decision boundary of the two classes. In this case, the SVM classifier is able to accurately model the decision regions between the two classes. This is because of the careful choice of the instances chosen by the active learning process. An important point to note is that it is particularly useful to sample instances that provide a distinct view of how the different classes are separated in the data.

4. PROBLEM FORMULATION

Let D be a set of document-label pairs (x, y), where the label (value of y) is known only for a small subset $L \subset D$ of the documents: $L = \{(x, y)\}$ and the rest $U = D \setminus L$ consists of the



(a) Class Separation

(b) Random Sample with SVM Classifier

(c) Active Sample with SVM Classifier

Fig. 1. Motivation of active learning.

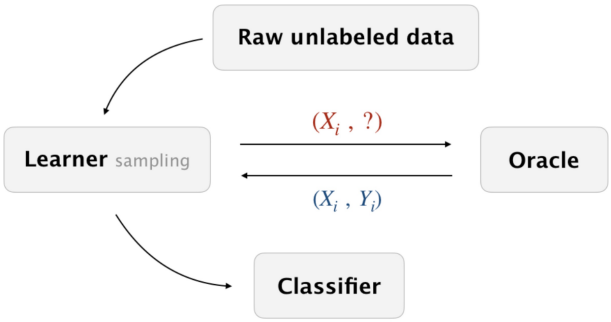


Fig. 2. Overall procedure in Active Learning

unlabeled documents: $U = \{(x, ?)\}$. We assume that each document x_i is represented as a vector of features (most commonly as a bag-of-words model with a dictionary of predefined set of phrases, which can be unigrams, bigrams, etc.):

$$x_i \triangleq \{f_1^i, f_2^i, ..., f_n^i\} \tag{1}$$

Each feature f_j^i represents the binary presence (or absence), frequency, or tf-idf representation of the word/phrase j in document x^i. Each label $y \in Y$ is discrete-valued variable

$$Y \triangleq \{y_1, y_2, ...y_l\} \tag{2}$$

Typical greedy active learning algorithms iteratively select a document $(x, ?) \in U$, query a labeler for its label y, and incorporate the new document (x, y) into its training set L. This process continues until a stopping criterion is met, usually until a given budget, B, is exhausted.

So the procedure of active learning, as is illustrated in Figure2, is first given some raw data, then the learner samples the point by a certain criteria, and asks the oracle of the label, then feeds the labelled data back to the classifier. After that, the classifier determines the decision boundary according to the given data.

5. PROPOSED METHODS

The key question in active learning algorithms is to design the precise strategies that are used for querying. At any given point, our goal is to select sample so as to **maximize the accuracy** of the classification process. As is evident from the discussion in the previous section, it is advantageous to use strategies, so that the contours of separation between the different classes are mapped out with the use of a small number of examples. Since the boundary regions are often those in which instances of multiple classes are present, they can be characterized by class label uncertainty or disagreements between different learners. However, this may not always be the case, because instances with greater uncertainty are not representative of the data, and may sometimes lead to the selection of unrepresentative outliers. This situation is especially likely to occur in data sets that are very noisy. In order to address such issues, some models focus directly on the error itself, or try to find samples that are representative of the underlying data. Therefore, we briefly introduce two types of models, heterogeneity and representativeness-based models.

5.1 Heterogeneity Models

Heterogeneity-based models[Aggarwal 2014] attempt to sample from regions of the space that are either more heterogeneous, or dissimilar to what has already been seen so far. Examples of such models include uncertainty sampling, query-by-committee, and expected model change. All these methods are based on sampling either uncertain regions of the data, or those that are dissimilar to what has been queried so far. These models not only look at the heterogeneity behavior of the queried instance, but also the effect of its addition on the performance of a classifier on the remaining unlabeled instances.

5.1.1 *Uncertainty Sampling*. The simplest and most commonly used query framework is uncertainty sampling. In this framework, the learner attempts to label those instances for which it is least certain how to label. The criteria is especially relevant for k-ary classification. It is the entropy measure or the gini-index.

$$G(\overline{X}) = 1 - \sum_{i=1}^{k} {p_i}^2 \tag{3}$$

If the predicted probabilities of the k classes are $\{p_1, ..., p_k\}$, respectively, based on the current set of labeled instances, then the entropy measure $En(\overline{X})$ is defined as follows:

$$En(\overline{X}) = \sum_{i=1}^{k} p_i log(p_i) \tag{4}$$

Larger values of the entropy indicate greater uncertainty. Therefore, this objective function needs to be maximized.

5.1.2 *Margin-based Sampling.* We have found that the criterion for the least confident strategy only considers information about the most probable label. Thus, it effectively throws away information about the remaining label distribution. To correct for this, a different multi-class uncertainty sampling variant called *margin-based sampling* has been used.

$$x_M^* = argmin_x P_\theta(\hat{y_1}|x) - P_\theta(\hat{y_2}|x) \tag{5}$$

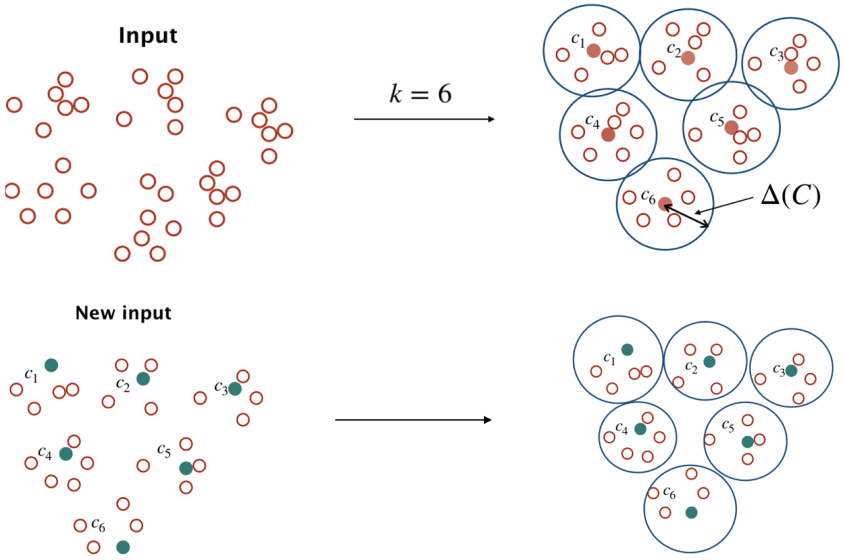


Fig. 3. Brief illustration on the procedure of K-center sampling ($k = 6$)

where $\hat{y_1}$ and $\hat{y_2}$ are the first and second most probable class labels under the model, respectively. Margin-based sampling aims to correct for a shortcoming in least confident strategy, by incorporating the posterior of the second most likely la- bel.

5.2 Representativeness-Based Models

Representativeness-based models attempt to create data that is as representative as possible of the underlying population of training instances. For example, density-based models are an example of such scenarios. In these cases, a product of a heterogeneity criterion and a representativeness criterion is used in order to model the desirability of querying a particular instance. Thus, these methods try to balance the representativeness criteria with the uncertainty properties of the queried instance.

As is shown in the exploded graph Figure 3, if every time the learner will choose, lets say, 6 points to query. This algorithm will set k equal to 6 and performs the k-center greedy algorithm. After 6 points have been selected, the unlabeled data us removed, and procedure is repeated. During the whole process, we only feed the data to the classifier, instead of asking the classifier its preference of data.

5.2.1 *Hierarchical Sampling.* A hierarchical clustering of the unlabeled points is constructed so that some pruning of it is weakly informative of the class labels. Suppose it is possible to prune the cluster tree to m leaves (m unknown) that are fairly pure in the labels of their constituent points. Then, after querying just O(m) labels, the learner will have a fairly accurate estimate of the labels of the *entire* data set.

6. EXPERIMENTS

In this section we first describe the settings, datasets, and classifiers used for our experiments. Then, we present the results comparing the learning curves achieved with learning on different data sets and sampling approaches.

Comments	Label
我们家的小豆包开始上学啦！在购物车里待一年的书包终于到手了，不错哦	书包
这款面料不错，版型也很好，柔软有坠感。洗过穿了，没褪色也不起球。很喜欢这条阔腿裤，真的很修身。这种裤子又很百搭，而且很有气质。果断拍下了，和我想象中的一样好，已经收藏店铺了。	阔腿裤
T恤有点长，是想要的有点厚度的那种，满意鸭	T恤
第二双鞋收到，质量真是不错，谁说便宜没有好货啊，我感觉做工真心不错。服务更周到！	运动鞋

Fig. 4. Taobao review

Comments	Label
只是给你们说下，就不要发两件出来了，我只要一件的	1
辛苦了	4
你门家用不了红包	16

Fig. 5. E-commerce customer queries

6.1 Classifiers

We feed labeled data to several simple classifiers, such as Naive Bayes classifier, Support Vector Machines, logistic regression classifier.

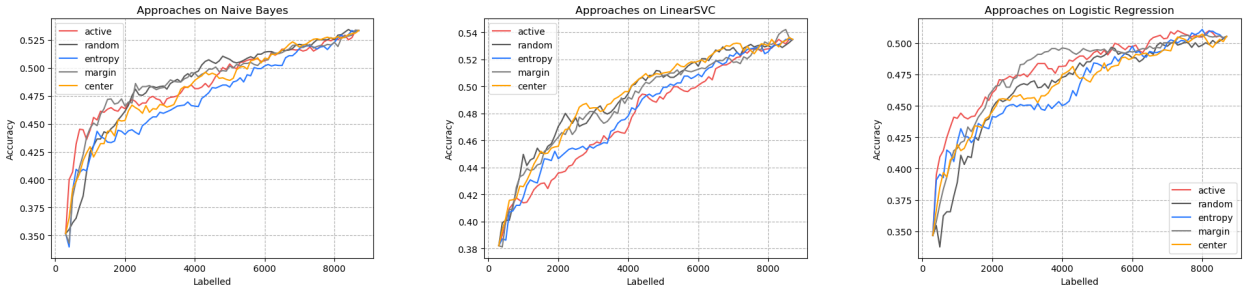


Fig. 6. Results on Taobao review data set

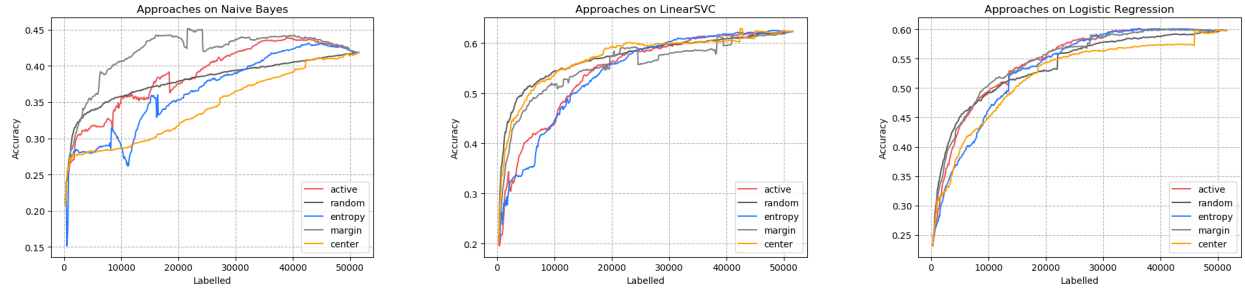


Fig. 7. Results on E-commerce customer queries data set

(1) Naive Bayes classifiers are a family of simple *probabilistic classifiers* based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

(2) A Support Vector Machine model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

(3) Logistic regression is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

6.2 Data Sets

The downstream task for our active learning algorithms is text classification. It is performed on two data sets. One is crawled Taobao reviews of different item categories, which is illustrated in Figure 4. The other one is provided E-commerce customer queries of different intents, which is illustrated in Figure 5. Different from common works, we put our emphasis on a multi-class classification tasks, which is trikier than previous works.

6.3 Data Preprocessing

When it comes to preprocessing procedure, we first segment our comments into words, and remove the stopwords. After that, we use our word list to calculate TF-IDF[Joachims 1996]. In Word2Vec[Le and Mikolov 2014], we feed the word list to get word vectors, then calculate the sentence vector by averaging all the word vectors in the sentences.

6.4 Results

6.4.1 *Results on Taobao Review.* Results using different classifiers on Taobao reviews are shown in Figure 6. The X-axis is the amount of labelled data, while the Y-axis is the accuracy of the classifier. Since our goal is to make the decision boundary as accurate as possible using the fewest labeled points. The beginning point is the same since we set the same initial labelled data. The ending point is the same since when all the points have been sampled, it becomes a supervised problem. So we can simply evaluate the performance of different models by area under curve(AUC). The larger the area under the curve, the better the result. The red line denotes the uncertainty sampling method, the blue one is the entropy sampling approach, the gray one is the margin-based sampling, and the black one is our baseline, random sampling. From the left of Figure 6 to right, we use three different classifiers, nave bayes, linear SVC and logistic regression. Although proposed method doesn't have a big margin over the baseline, we can expect they share some similar patterns. For example, we can assume the gray line which is the margin-based approach is better than other methods.

6.4.2 *Results on E-commerce customer queries.* On our E-commerce customer queries, as shown in Figure 7, from left to right we use three different classifiers, nave bayes, linear SVC and logistic regression. Although the gray one makes a big margin over others on Naive Bayes, it performs poorly on linear SVC. In other words, it is not a consistent result. To explain this case, we do some analysis on our data.

6.5 Analysis

To further analyze the inconsistent results achieved on E-commerce queries, we compare its train and test data distribution with Taobao

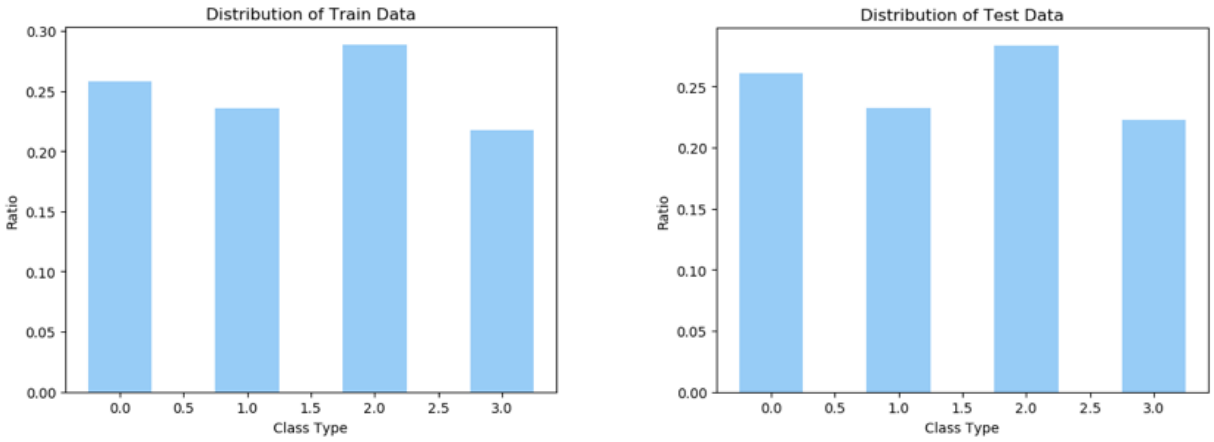


Fig. 8. Distribution of train and test data on Taobao Review

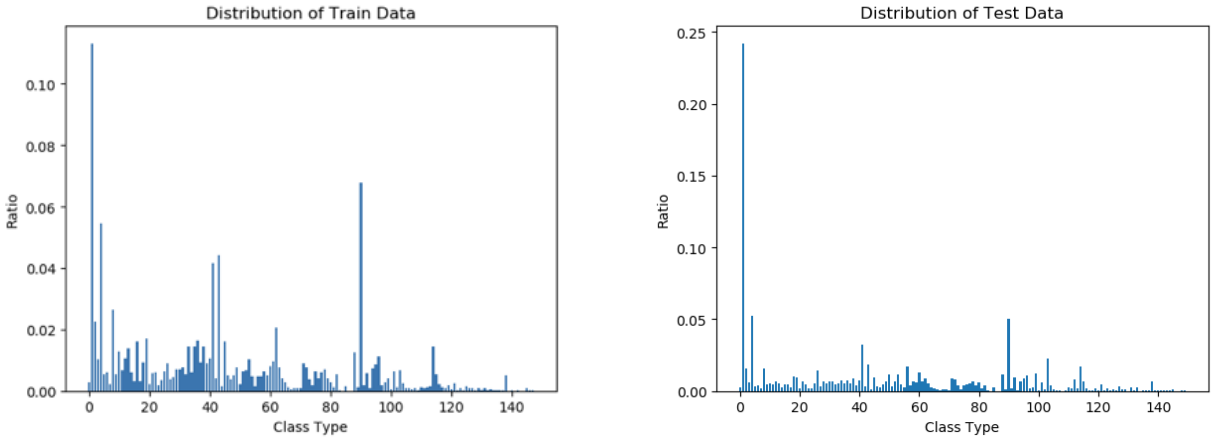


Fig. 9. Distribution of train and test data on E-commerce customer queries

data distribution. Figure 8 is the normalized distribution of our Taobao review data. As is shown on the figure, the distribution of the test and train data is totally the same. However, as is shown in Figure 9, the train data and test data of our E-commerce queries is imbalanced, which is not as ideal as Taobao review data. Nonuniformity between train data and test data may be the main reason of the inconsistency in our experiment result.

7. CONCLUSION

We have implemented several proposed models on our multi-class text classification tasks and also designed a procedure to improve it. Results have shown that it does work on most of the cases, but it cannot perform equally well using different classifiers on imbalanced data sets.

To further improve our models, we have come up with some new ideas. Since previous approaches are general methods, we can go deeper and extract some features of our specific tasks. For example, we can use zero shot learning to solve the imbalanced data distribution problem. The relation between different classes, and transfer learning methods can also be used to adapt the model trained on classes with more wealthy data to classes with less data. In addition, since classifiers employed now are mainly applied in binary classification tasks, thus interfering with its performance in multi-class classification tasks. Meanwhile, they are too simple to extract higher semantic structure. Therefore, a more complicated classifier model like LSTM with attention will be leveraged to do the task.

ACKNOWLEDGMENTS

We are grateful to the following people for resources, discussions and suggestions: Prof. Li Jiang, Prof. Kenny Zhu and teacher assistants Chaoqun chu and Xiaoyi Sun. We also thank E-commerce company, who provide queried labeled data.

REFERENCES

Charu C Aggarwal. 2014. *Data classification: algorithms and applications*. CRC Press.

David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning* 15, 2 (1994), 201–221.

Thorsten Joachims. 1996. *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* Technical Report. Carnegie-mellon univ pittsburgh pa dept of computer science.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc., 3–12.

Ray Liere and Prasad Tadepalli. 1997. Active learning with committees for text categorization. In *AAAI/IAAI*. 591–596.

Andrew Kachites McCallumzy and Kamal Nigamy. 1998. Employing EM and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning (ICML)*. Citeseer, 359–367.

Manali Sharma, Di Zhuang, and Mustafa Bilgic. 2015. Active Learning with Rationales for Text Classification. In *North American Chapter of the Association for Computational Linguistics – Human Language Technologies*. http://www.cs.iit.edu/~ml/pdfs/sharma-naaclhlt15.pdf