Directly Identify Unexpected Instances in the Test Set by Entropy Maximization

Chaofeng Sha¹, Zhen Xu¹, Xiaoling Wang², and Aoying Zhou²

¹ Department of Computer Science and Engineering
Fudan University
Shanghai 200433, China
{cfsha, xzhen}@fudan.edu.cn

² Shanghai Key Laboratory of Trustworthy Computing
Institute of Massive Computing
East China Normal University
Shanghai 200062, China
{xlwang, ayzhou}@sei.ecnu.edu.cn

Abstract. In real applications, a few unexpected examples unavoidably exist in the process of classification, not belonging to any known class. How to classify these unexpected ones is attracting more and more attention. However, traditional classification techniques can’t classify correctly unexpected instances, because the trained classifier has no knowledge about these. In this paper, we propose a novel entropy-based method to the problem. Finally, the experiments show that the proposed method outperforms previous work in the literature.

1 Introduction

For the present of unexpected instances in test set, the trained classifier can’t correctly classify them. In practice, this phenomenon is very common. For example, the evolution of species may cause that many new unknown or undefined species emerge. If new species are classified to some known or predefined old species, it is obviously unsuitable, even leads to terrible consequences. Take SARS(Severe Acute Respiratory Syndrome) as an example, a kind of new disease having broken out in south China and spread to other countries. Initially, SARS was considered as a kind of common flu, and hence the patients got the incorrect treatment. Ultimately, When treating SARS as a new virus for treatment, this kind of new virus was completely overwhelmed. Currently, identifying unexpected instances is an interesting topic in the data mining and machine learning communities.

With the absence of labeled unexpected (or negative) instances, traditional classification methods can’t be applied directly. The crucial point of this problem is how to find or generate the negative instances and put them into the training data set for learning. Once negative instances are obtained, the traditional classification approach is directly for the rest task.
In order to solve the above problem, we present an entropy-based approach to identify unexpected instances hidden in the test set. Experimentally, the proposed technique outperforms existing ones.

The main contributions of this paper are as follows:

- This is the first work of using entropy maximization techniques to identify directly unexpected instances in test set.
- Besides text data, the proposed method also can deal with other types of data (e.g. nominal).
- Even when the proportion of unexpected instances is very small, the performance remains consistent.

The rest of this paper is organized as follows: Section 2 presents related research works and Section 3 briefly introduces the proposed approach. The experiments are shown in Section 4 followed by the conclusion in Section 5.

2 Related Work

PU learning (learning from Positive and Unlabeled examples) gets more and more focus in recent years. Now several different approaches are proposed to this problem. For example, the outstanding method LGN [6] firstly generates artificial negative instances based on the different distribution of equal words within training and test set respectively, and then use Naïve Bayesian classifier for the rest. However, the assumption that every positive word has the same distribution is not easily satisfied in practice; Besides this, it only concerns text data, not other types.

Other approaches [1,5] also have similar shortcomings, e.g., when the percentage of unexpected instances is very small, these classifiers have very poor performances.

In this paper, however, a simple but effective approach is introduced. This new method directly classifies the unexpected instances hidden in test set using entropy, which firstly chooses negative instances and then realizes the rest classification. Besides text data, it also can apply to other types of data.

3 Our Approach

In this section, we propose the method named Naïve Bayesian classifier based on Entropy (NB-E) to find unexpected instances for classification.

3.1 Preliminaries

Entropy \( H(x) \). The entropy of the probability distribution \( P(x) \) on set \( \{x_1, \cdots, x_n\} \) is defined as [4]:

\[
H(x) = -\sum_{i=1}^{n} P(x_i) \log P(x_i)
\]
where \( n \) is the number of known or predefined classes, and \( P(x_i) \) is the class posterior probability of instance \( x \) belonging to the \( i \)th class.

The basic motivation is to use entropy as the direct measurement of confidence for correct classification, namely, the bigger the entropy \( H(x) \), the smaller the probability of correct classification, and vice versa. Therefore, the instances with the biggest entropy have nearly equal confidences to belong to any predefined class, i.e., the highest probability to be unexpected.

As an extreme example, there are three documents \( x_a, x_b \) and \( x_u \) belonging to three different classes respectively. After preprocess, the term vectors are \( x_a \triangleright \{w_1, w_2\} \), \( x_b \triangleright \{w_3, w_4\} \) and \( x_u \triangleright \{w_5, w_6\} \). Here, \( x_a \) and \( x_b \) are training instances for two positive classes \( C_1 \) and \( C_2 \), and \( x_u \) as the negative example hidden in test set.

Then, \( C_1 \) and \( C_2 \) have the same prior probability, namely, \( P(C_1) = P(C_2) = \frac{1}{2} \); Subsequently, by the naïve bayesian formula, the class conditional probabilities are as follows:

\[
P(w_5|C_1) = P(w_6|C_1) = P(w_5|C_2) = P(w_6|C_2) = \frac{1}{6};
\]

Then, the posterior probabilities of \( x_u \) are listed as follows:

\[
P(C_1|x_u) = P(C_2|x_u) = \frac{1}{2}
\]

Since \( H(x_u) \) is biggest, the conclusion is made that \( x_u \) has higher probability of being unexpected or negative.

### 3.2 Naïve Bayesian Classifier Based on Entropy (NB-E)

As a highly effective classification technique, Naïve Bayesian classifier\(^2\) is employed within this paper.

Based on the property of entropy, the proposed approach is presented in Algorithm[1] which mainly has the following four steps:

- From Line 3 to 5, TF-IDF technique is applied to do feature selection for document data;
- The entropy of the class posterior probabilities of any test instance, is computed and sorted by ascending order(Line 7 to 12). For any positive class, the top \( k \) instances are removed from test set and are labeled as the corresponding positive class, namely, every positive class has more \( k \) new instances for learning.(Line 14 to 17);
- From Line 19 to 24, by descending order, every instance is sorted based on the entropy. The biggest \( n \) instances chosen from test set are added into training set(labeled as “-”). At the same time, all positive training instances are viewed as one class (denoted as “+”);
- From Line 25 to 31, there are only two different classes within training set. Finally, the multinomial Naïve Bayesian classifier is directly applied classification. Finally, the hidden unexpected examples are returned (Line 32).
Algorithm 1. NB-E()

**Input:** training set $P$, testing set $U$, negative percentage $\alpha$, threshold parameter $\delta$, choice factor $n$;

**Output:** unexpected instances set $U_e$;

1. $U_e = \phi$;
2. $C = \{c_1, \ldots, c_n\}$ is the set of classes appearing in the training set;
3. if $P$ is the document data then
   - Using TF-IDF technique to extract representative attribute words for every class;
   - Build a naïve Bayesian classifier (NB) with the training set $P$;
4. for each instance $d_i \in U$ do
   - Use NB to classify $d_i$;
   - $H(d_i) = - \sum_{c_j \in C} P(c_j | d_i) \log P(c_j | d_i)$;
   - $j \leftarrow \text{argmax}_{c_j \in C} P(c_j | d_i)$;
   - $B_j = B_j \cup \{d_i\}$;
5. end
6. $k \leftarrow \lfloor \delta \times \text{argmin}_{c_j \in C} |B_j| \rfloor$, $\delta \in (0, 1)$ and the default value is 0.8;
7. for each class $c_i \in C$ do
   - Sort all the instances in $B_i$ according to the ascending relation of the entropy of every instance’s posterior probability;
   - Remove the first $k$ instances from $U$ and throw them into $P$ as the instances of $c_i$;
8. end
9. Build a new Naïve Bayesian classifier (NB) with $P$; /* here, $P$ is updated */
10. for each instance $d_i \in U$ do
   - Use NB to classify $d_i$;
   - $H(d_i) = - \sum_{c_j \in C} P(c_j | d_i) \log P(c_j | d_i)$;
11. end
12. Rank the entropy of all instances in $U$ by the descending order and choose the top $n$ instances and insert into $P$ as the instances of unexpected class “-”;
13. Merge all positive classes in $P$ and view them as a whole positive class “+”;
14. Build a new Naïve Bayesian classifier (NB) with $P$; /* here, $P$ is updated */
15. for each instance $d_i \in U$ do
   - Use NB to classify $d_i$;
   - if $P(-|d_i) > P(+|d_i)$ then
     - $U_e = U_e \cup \{d_i\}$;
16. end
17. end
18. return $U_e$;
4 Experimental Evaluation

4.1 Experimental Dataset

For the evaluation, we mainly use two kinds of representative data collections – 20 newsgroups\(^1\) and uci letter\(^2\). Here, several parameters are set as follows:

- \(\alpha\) is the proportion of negative instances hidden in test set with respect to positive ones. In order to simulate various situations, \(\alpha\) has 8 different values, namely, \(\{5\%, 10\%, 15\%, 20\%, 40\%, 60\%, 80\%, 100\%\}\);
- \(n\) is the number of selected likely negative instances. As different cases, \(n\) has 2 rows of different values, i.e., \(\{5, 10, 15, 20, 30, 30, 30, 30\}\)(denoted as NB-E) and a constant 10(denoted as NB-E\(^+\)) respectively.

Additionally, as to the number of positive sub-classes, there are mainly 2 different assumptions: 2 positive sub-classes existing in the training set and 3 ones.

![F-score with different values for \(\alpha\) in document experiments](a) \(F\)-score with different values for \(\alpha\) in document experiments

![F-score with different values for \(\alpha\) in nominal experiments](b) \(F\)-score with different values for \(\alpha\) in nominal experiments

Fig. 1. The comparison results with different values for \(\alpha\) in 2 and 3-classes experiments

4.2 Experimental Results

As for document data, we mainly compare NB-E with LGN\(^\dagger\), which has the best performance among former approaches, however, previous works are never done for nominal data, hence no comparison here.

Fig.1 shows that as to document data, NB-E outperforms LGN, especially when the number fo positive sub-classes increases; Both at document and nominal data, NB-E has consistent performances, without being influenced by the number of positive sub-classes too.

Fig.2 shows that the values of parameter \(n\) has little influence on the final results, only if not too big or too small; On the other hand, because the number attribute words of document data is greater than the one of nominal data, document data have higher accuracy and F-score values.

\(^1\) http://people.csail.mit.edu/jrennie/20Newsgroups

\(^2\) http://archive.ics.uci.edu/ml/datasets.html
Fig. 2. The comparison results with different values for \( n \) in 2 and 3-classes experiments

5 Conclusion

To identify unexpected instances hidden in test set, we propose a novel entropy-based approach called NB-E, which is simple but very useful. Meanwhile, the experiments prove that this approach has excellent performances, such as high F-score and accuracy at document and nominal data.

There are still some places to be improved, however, e.g., much more datasets are evaluated; The cluster techniques are applied to improve the accuracy of the chosen negative instances; Besides these, many other classification methods are used together, inclusive of logic regression, SVM and ensemble methods, etc.

Acknowledgement. This work is supported by NSFC grants (No. 60673137 and No. 60773075), National Hi-Tech 863 program under grant 2008AA01Z1470967 and Shanghai Leading Academic Discipline Project(Project NumberB412).

References

1. Liu, B., Dai, Y., Li, X., Lee, W., Yu, S.: Building Text Classifiers Using Positive and Unlabeled Examples. In: IJCAI 2003 (2003)
2. Györfi, L., Györfi, Z., Vajda, I.: Bayesian decision with rejection. Problems of Control and Information Theory 8, 445–452 (1978)
3. Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, Heidelberg (1996)
4. Cover, T., Thomas, J.: Elements of information theory. Wiley Interscience, Hoboken (1991)
5. Li, X., Liu, B., Lee, W., Yu, S.: Text Classification by Labeling Words. In: AAAI 2004 (2004)
6. Li, X., Liu, B., Ng, S.: Learning to identify unexpected instances in the test set. In: IJCAI 2007 (2007)
7. Guo, Y., Greiner, R.: Optimistic active learning using mutual information. In: IJCAI 2007 (2007)