An Effective and Robust Method for Short Text Classification

Victoria Bobicev  
Technical University of Moldova  
Studentilor, 7, Chisinau, Moldova  
vika@rol.md

Marina Sokolova *  
CHEO Research Institute  
401 Smyth Road, Ottawa, Ontario, Canada,  
msokolova@ehealthinformation.ca

Abstract

Classification of texts potentially containing a complex and specific terminology requires the use of learning methods that do not rely on extensive feature engineering. In this work we use prediction by partial matching (PPM), a method that compresses texts to capture text features and creates a language model adapted to a particular text. We show that the method achieves a high accuracy of text classification and can be used as an alternative to state-of-art learning algorithms.

Motivation

We focus on classification of texts with a high concentration of a specific terminology and complex grammatical structures. Those characteristics inevitably complicate standard feature engineering, which is done by language pre-processing (e.g., lemmatization, parsing) that is further complicated when the texts are short. Our goal is to avoid complex and, perhaps, error-prone feature construction by using a learning method that can perform reasonably well without preliminary feature engineering. We use prediction by partial matching (PPM), an adaptive finite-context method for text compression, that is a back-off smoothing technique for finite-order Markov models (Bratko et al. 2006). It obtains all information from original data, without feature engineering, is easy to implement and relatively fast. PPM produces a language model and can be used in a probabilistic text classifier.

The character-based PPM models were used for spam detection, source-based text classification and classification of multi-modal data streams that included texts. We opted to use the compression models for classification of terminology-intense data, e.g., medical texts. We applied PPM-based classifiers to the topic and non-topic classification of short texts, including classification of medical diagnosis. We built two versions of PPM-based classifiers, one calculating the probability of the next word and other calculating the probability of the next character. Our empirical results show that the PPM-based classifiers achieve a competitive accuracy of the short text classification.

PPM Classifier

PPM is based on conditional probabilities of the upcoming symbol given several previous symbols (Cleary and Witten 1984). The PPM technique uses character context models to build an overall probability distribution for predicting upcoming characters in the text. A blending strategy for combining context predictions is to assign a weight to each context model, and then calculate the weighted sum of the probabilities: \( p(\phi) = \sum_{i=1}^{m} q_i p_i(\phi) \), where \( q_i \) and \( p_i \) are weights and probabilities assigned to each order \( i \). PPM is a special case of the general strategy. The PPM models use an escape mechanism to combine the predictions of all character contexts of length \( \leq m \), where \( m \) is the maximum model order; the order 0 model predicts symbols based on their unconditioned probabilities, the default order −1 model ensures that a finite probability (however small) is assigned to all possible symbols. The PPM escape mechanism is more practical to implement than weighted blending. There are several versions of the PPM algorithm depending on the way the escape probability is estimated. In our implementation, we used the escape method C (Bell, Witten, and Cleary 1989).

Treating a text as a string of characters, a character-based PPM avoids defining word boundaries; it deals with different types of documents in a uniform way. It can work with texts in any language and be applied to diverse types of classification; more details can be found in (Bobicev 2007). We, however, built both word-based and letter-based PPM classifiers to compare their performance. Our utility function was: \( H_d^m = -\sum_{i=1}^{n} p^m(x_i) \log p^m(x_i) \), where \( n \) is the number of symbols in a text \( d \), \( H_d^m \) – entropy of the text \( d \) obtained by model \( m \), \( p^m(x_i) \) is a probability of a symbol \( x_i \) in the text \( d \). \( H_m^d \) was estimated by the modelling part of the compression algorithm. On the training step, we created PPM models for each class of documents; on the testing step, we evaluated cross-entropy of previously unseen texts using models for each class. The lowest value of cross-entropy indicates the class of the unknown text.

Empirical results

We applied our method on Newsgroups, clinical texts, and Reuters-21578. We tested the PPM models: word-based with orders 0, 1, 2 and letter-based with order 5. The results
of the word-based model with order 0 were unexpectedly much lower than those of other word-based models; we decided to skip it in further experiments. The results of the word-based model with order 1 were very close to those of order 2. We attribute the similarity to a short length of texts.

We used the complete Newsgroup data, 19,974 texts that evenly belong to 20 topic categories. We report results for medical texts, classified among 20 given classes. Our results on Newsgroups are competitive with those of (Frank and Bouckaert 2006) that use Naive Bayes; see Table 1.

For medical diagnosis classification, we used 978 samples of chest x-ray and renal procedures from Medical NLP Challenge 2007. The texts contained 15 – 20 words, e.g.

**CLINICAL HISTORY:** Cough, congestion, fever. IMPRESSION: Increased markings with subtle patchy disease right upper lobe. Atelectasis versus pneumonia. Those texts had multiple labels (ICD-9-CM codes). The labels formed 94 categories with at least three positive examples (Reuters(96)), 105 categories with at least two positive examples (Reuters(105)). For Reuters(15), texts with only one label were selected from the whole test set. Consequently, 3 categories did not have test texts, thus, we calculated results for the remaining 12 categories. We resolved the multi-labelling problem as in experiments with medical texts.

For Newsgroups, we report some letter combinations and the words they appear in: ent, ati (patients, treatment); her (therapy); res (research, result); ect (infection, detect, effect) are frequent in medical texts and sparse otherwise; ver (government, university); thi (thing, think); sta (state, standard, started); tic (article, politics);uld (would, should, could) are sparse in medical texts and frequent otherwise.

### Discussion

We have presented a comparative study of classification of short text collections, a challenging text classification task. We applied PPM-based classifiers that do not require data pre-processing or feature engineering. The results of the experiments showed that PPM-based compression provides a reliable accuracy of text classification. The letter-based method performed slightly better than the word-based methods. A possible explanation is the quality of texts: texts are noisy and contain errors of different types that affect the word-based methods. Letter-based methods avoid these problems and, in general, better capture the characteristics of the text. It should be mentioned that the letter-based model is more compact and faster to build. For future work, we want to concentrate on PPM applications for multi-class classification. Another promising direction can be to test the method on other multi-labelled texts, especially medical texts, and explore different ways of multiple classification (several labels assigned) of unknown documents.

### References


---

**Table 1:** Classification of medical texts from Newsgroups.

<table>
<thead>
<tr>
<th>Compression</th>
<th>order</th>
<th>Prec</th>
<th>Recall</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>word-based</td>
<td>1</td>
<td>0.63</td>
<td>0.91</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.96</td>
<td>0.84</td>
<td>0.90</td>
<td>0.98</td>
</tr>
</tbody>
</table>

**Table 2:** The macroaveraging results for clinical texts and Reuters.

<table>
<thead>
<tr>
<th>Compression</th>
<th>order</th>
<th>Prec</th>
<th>Recall</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>word-based</td>
<td>1</td>
<td>0.33</td>
<td>0.45</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.36</td>
<td>0.42</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>letter-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Compression</th>
<th>order</th>
<th>Prec</th>
<th>Recall</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuters(15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>word-based</td>
<td>1</td>
<td>0.75</td>
<td>0.62</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.65</td>
<td>0.57</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>letter-based</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Reuters(96) |       |      |        |         |     |
| word-based  | 1     | 0.64 | 0.68   | 0.64    |     |
|             | 5     | 0.72 | 0.57   | 0.64    |     |
| letter-based|       |      |        |         |     |

| Reuters(105) |       |      |        |         |     |
| word-based  | 1     | 0.77 | 0.62   | 0.68    |     |
|             | 5     | 0.78 | 0.63   | 0.69    |     |
| letter-based|       |      |        |         |     |

---

1The data were provided by Computational Medicine Centre, Cincinnati Children’s Hospital, http://www.computationalmedicine.org/challenge

<table>
<thead>
<tr>
<th>Composition</th>
<th>order</th>
<th>Prec</th>
<th>Recall</th>
<th>F-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>word-based</td>
<td>1</td>
<td>0.63</td>
<td>0.91</td>
<td>0.77</td>
<td>0.93</td>
</tr>
<tr>
<td>letter-based</td>
<td>5</td>
<td>0.96</td>
<td>0.84</td>
<td>0.90</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 1: Classification of medical texts from Newsgroups.