A text feature selection method based on category-distribution divergence

Yonghe Lu,* Wenqiu Liu, Xinyu He

School of Information Management, Sun Yat-sen University, Guangzhou, China

Received: July 11, 2015 Accepted: August 13, 2015 Online Published: September 7, 2015
DOI: 10.5430/air.v4n2p143 URL: http://dx.doi.org/10.5430/air.v4n2p143

Abstract

The purpose of this paper is to overcome the problem that traditional feature selection methods [such as document frequency (DF), Chi-square statistic (CHI), information gain (IG), mutual information (MI) and Odds ratio (OR)] do not consider the distribution of features among different categories. The work aims at selecting the features that can accurately represent the theme of texts and to improve the accuracy of classification. In this paper, we propose a text feature selection method based on Category-Distribution Divergence, and the degree of membership and degree of non-membership are introduced into CDDFS (feature selection based on category-distribution divergence). CDDFS is used as a filter which can filter the features having low degree of membership and high degree of non-membership. CDDFS is tested with five feature selection methods and three classifiers using the corpus of Sogou Lab Data, and experimental results show that this method performs better than other feature selection methods when using KNN, and close to CHI when using Rocchio algorithms and SVM at high dimensions. This research proposes the representativeness and distinguishability of feature for category, and the representativeness and distinguishability of feature for non-category. If a feature has good distinguishability and high representativeness, then this feature will be retained in feature selection.

Key Words: Text categorization, Category-distribution, Feature selection, Vector space model

1 Introduction

Text categorization is a key technique that aims at processing and organizing large amounts of text data, and it can solve some problems brought by the rapid growth of information. In text categorization, data dimension has a direct impact on the results and speed of categorization. For most categorization algorithms, the high dimensional data (especially the data with thousands even ten thousands dimensions) will make the classifier stop working due to excessive computing or consuming too many resources. [1–3] To overcome their shortcomings, many scholars proposed different modified methods. Liu et al. [6] put forward a kind of optimizing MI text feature selection method. This method can improve the efficiency of MI model. Considering that traditional Information Gain ignores the shortcoming of distributing information inside class and between classes, Guo and Liu [8] proposed modified IG by introduce the distribution information inside class and concentration information between classes. And the experimen-
We assume there are three features

2.1 The related theory

other categories. Lastly, we combine the degree of membership and degree of non-membership are introduced into CDDFS. Firstly, we calculate the degree of membership and degree of non-membership are introduced into CDDFS. Secondly, we calculate the degree of membership between a word and other categories. Lastly, we combine the degree of membership and degree of non-membership together.

2 Feature selection based on category-distribution divergence

2.2 Construct the function

We define that \( t_i \) is the i-th feature, \( (t_j) \) is “NOT” operation to \( t_i \). \( c_j \) is the j-th category, \( (c_j) \) is “NOT” operation to \( c_j \). \( N_{11} \) refers to the number of texts which contain feature \( t_i \) and belong to category \( c_j \). \( N_{10} \) refers to the number of texts which contain \( t_i \) and do not belong to \( c_j \). \( N_{01} \) refers to the number of texts which do not contain \( t_i \) but belong to \( c_j \). \( N_{00} \) refers to the number of texts which neither contain \( t_i \) nor belong to \( c_j \) (see Table 1).

Table 1: The Relationship among Parameters

<table>
<thead>
<tr>
<th>$t_i$</th>
<th>$N_{11}$</th>
<th>$N_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\overline{t}_i$</td>
<td>$N_{01}$</td>
<td>$N_{00}$</td>
</tr>
</tbody>
</table>

Based on the above reasons, we define \( M \) as the total number of categories and \( N \) as the total number of texts. The distinguishability of feature \( t_i \) to category \( c_j \) is computed as Equation (1).

\[
diff(t_i, c_j) = \log_2 \left( \frac{N_{11}}{N_{10} + 1} + 1 \right)
\]

(1)

The representativeness of \( t_i \) to \( c_j \) is computed as Equation (2).

\[
repr(t_i, c_j) = N_{11}
\]

(2)

Then the degree of membership of \( t_i \) to \( c_j \) is given by Equation (3).

\[
\text{belongpositive}(t_i, c_j) = \diff(t_i, c_j) \ast \repr(t_i, c_j) = N_{11} \log_2 \left( \frac{N_{11}}{N_{10} + 1} + 1 \right)
\]

(3)

Similarly, the degree of non-membership of \( t_i \) to \( c_j \) (the de-
gree of membership of \( t_i \) to \( (c_j) \) is defined by Equation (4).

\[
\text{belong}_{\text{positive}}(t_i, c_j) = N_{10} \log_2 \left( \frac{N_{11}}{N_{10} + 1} + 1 \right) \tag{4}
\]

Considering the degree of membership and non-membership of \( t_i \) to \( c_j \), we can get the CDDFS with equation (5).

\[
\text{belong}(t_i, c_j) = N_{11} \log_2 \left( \frac{N_{11} + 1}{N_{10} + 1} + 1 \right)
- N_{10} \log_2 \left( \frac{N_{10}}{N_{11} + 1} + 1 \right) \tag{5}
\]

If \( \text{belong}(t_i, c_j) \leq 0 \), that means the degree of membership of \( t_i \) to \( c_j \) is lower than the degree of non-membership. Then the feature \( t_i \) should not be selected in feature selection. The evaluation function of CDDFS is computed by equation (6).

\[
\text{evaluation}_{\text{CDDFS}}(t_i, c_j) = \begin{cases} 
\log_2(\text{belong}(t_i, c_j) + 1) & \text{belong}(t_i, c_j) > 0 \\
0 & \text{belong}(t_i, c_j) \leq 0
\end{cases} \tag{6}
\]

### 3 Experimental results and analysis

#### 3.1 Experimental setting

To verify the validity of feature selection algorithm CDDFS, we compare it with DF,[19] CHI, IG, MI and OR. The data set is Sogou corpus of text categorization.[20] We select nine categories from the corpus, namely automotive, finance, IT, health, sports, tourism, education, recruitment and military. We select 200 texts for each category, and these texts are divided into training set and testing set according the ratio of 1:1. There are 900 texts in both training set and testing set respectively. We use open source package Lucene[21] to preprocess the text set, including Chinese word segmentation, word frequency statistic and so on. The document representation model is Vector Space Model (VSM), and we do experiments at 360, 720, \( \cdots \), 3600 dimensions respectively.

The value of Global evaluation function of feature \( t_i \) is computed by equation (7).

\[
\text{evaluation}(t_i) = \max_{j=1-M} \{ \text{evaluation}(t_i, c_j) \} \tag{7}
\]

Where \( M \) is the number of categories. The feature weight calculation method is traditional TF-IDF, and it is computed by equation (8).

\[
w_{id} = tf_{id} \log \frac{D}{df_i} \tag{8}
\]

Where \( tf_{id} \) is the number of feature \( i \) appearing in text \( d \). \( D \) is the text number of training set. \( df_i \) is the number of texts which contain feature \( i \) in training set.

In order to eliminate the influence of document’s length on categorization results, we use the cosine normalization.[22] It is calculated in equation (9).

\[
w_{id} = \frac{tf_{id} \log \frac{D}{df_i}}{\sqrt{\sum_{k=1}^{N-1} (tf_{id} \log \frac{D}{df_i})^2}} \tag{9}
\]

We select KNN, center-point method and SVM as the classifier. The \( k \) in KNN is 7, and the similarity is computed by equation (10).

\[
sim(d_i, d_j) = \cos \alpha = \frac{\sum_{k=1}^{n} (w_{ik}w_{jk})}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \sum_{k=1}^{n} w_{jk}^2}} \tag{10}
\]

Where \( \sim(d_i, d_j) \) is the similarity between text \( d_i \) and \( d_j \), \( w_{ik} \) is the \( k \)-th feature weight in text \( d_i \), \( n \) is the dimension number of feature vector.

With the Cross-validation method, we get the parameters in SVM. They are cost=8, gamma=0.38125. The evaluation criteria of categorization results are MR (macro-averaging recall), MP (macro-averaging precision) and MF (macro-averaging f-measure).

In order to verify the classification performance with CDDFS is significantly better than the ones using other feature selection methods, paired-sample T-test was used for significance test.

Before using paired-sample T-test, we set up two hypotheses. The first was the null hypothesis, which assumed that the mean of two paired samples were equal. The second hypothesis was an alternative hypothesis, which assumed that the means of two paired samples were significantly different. And we choose a significance level of 0.05, it meant there was a 5% chance of rejecting the null hypothesis when it was true.

#### 3.2 Experimental results

After using various feature selection methods and three different classifiers, we get the macro-averaging F1 of the categorization results, and they are shown in Figures 1-3 and Tables 2-4.

##### 3.2.1 The experiment with KNN classifier

From Figure 1 and Table 2, we can know that CDDFS has better classification performance than other feature selections when using KNN as classifier.

This experiment used six different feature selection methods. Table 2 describes \( M_{acF1} \) in different features dimensions for six feature selections. Table 3 shows the comparison of \( M_{acF1} \) by using paired T test.
Table 2: Experimental Results with KNN

<table>
<thead>
<tr>
<th>Dimension Number</th>
<th>OR</th>
<th>CHI</th>
<th>MI</th>
<th>IG</th>
<th>DF</th>
<th>CDDFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>360</td>
<td>0.762397</td>
<td>0.813257</td>
<td>0.577079</td>
<td>0.671722</td>
<td>0.307134</td>
<td>0.863770</td>
</tr>
<tr>
<td>720</td>
<td>0.742256</td>
<td>0.825844</td>
<td>0.675333</td>
<td>0.701573</td>
<td>0.352370</td>
<td>0.865793</td>
</tr>
<tr>
<td>1080</td>
<td>0.680397</td>
<td>0.828759</td>
<td>0.699871</td>
<td>0.740211</td>
<td>0.437025</td>
<td>0.869975</td>
</tr>
<tr>
<td>1440</td>
<td>0.680405</td>
<td>0.835206</td>
<td>0.722009</td>
<td>0.763695</td>
<td>0.420384</td>
<td>0.857111</td>
</tr>
<tr>
<td>1800</td>
<td>0.686937</td>
<td>0.828902</td>
<td>0.734031</td>
<td>0.757990</td>
<td>0.463555</td>
<td>0.853632</td>
</tr>
<tr>
<td>2160</td>
<td>0.697051</td>
<td>0.837103</td>
<td>0.746118</td>
<td>0.774204</td>
<td>0.505097</td>
<td>0.854226</td>
</tr>
<tr>
<td>2520</td>
<td>0.706218</td>
<td>0.836109</td>
<td>0.763378</td>
<td>0.787817</td>
<td>0.550143</td>
<td>0.853348</td>
</tr>
<tr>
<td>2880</td>
<td>0.711690</td>
<td>0.825478</td>
<td>0.790843</td>
<td>0.787653</td>
<td>0.564861</td>
<td>0.856722</td>
</tr>
<tr>
<td>3240</td>
<td>0.711022</td>
<td>0.817522</td>
<td>0.790695</td>
<td>0.779897</td>
<td>0.592975</td>
<td>0.856823</td>
</tr>
</tbody>
</table>

Figure 1: Experimental Results with KNN

Table 3 shows that all comparisons of $MacF_1$ seem to reject the null hypothesis, thus demonstrating the CDDFS is better than the other feature selections when using KNN as classifier.

Table 3: Comparison of $MacF_1$ by using paired $t$ test

<table>
<thead>
<tr>
<th></th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR vs. CDDFS</td>
<td>-18.876</td>
<td>.000</td>
</tr>
<tr>
<td>CHI vs. CDDFS</td>
<td>-8.745</td>
<td>.000</td>
</tr>
<tr>
<td>MI vs. CDDFS</td>
<td>-6.040</td>
<td>.000</td>
</tr>
<tr>
<td>IG vs. CDDFS</td>
<td>-8.064</td>
<td>.000</td>
</tr>
<tr>
<td>DF vs. CDDFS</td>
<td>-12.321</td>
<td>.000</td>
</tr>
</tbody>
</table>

3.2.2 The experiment with Rocchio Algorithms

As we can see from the Figure 2 and Table 4, when the dimensions are less than 2,520, CDDFS have better performance than other feature selections when using Rocchio algorithms. When the dimensions are more than 2,520, CDDFS is as good as CHI and its performance is better than OR, MI, IG, DF.

Table 4: Experimental Results with Rocchio Algorithms

<table>
<thead>
<tr>
<th>Dimension Number</th>
<th>OR</th>
<th>CHI</th>
<th>MI</th>
<th>IG</th>
<th>DF</th>
<th>CDDFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>360</td>
<td>0.812725</td>
<td>0.842178</td>
<td>0.558861</td>
<td>0.746689</td>
<td>0.307942</td>
<td>0.866788</td>
</tr>
<tr>
<td>720</td>
<td>0.778864</td>
<td>0.851801</td>
<td>0.711011</td>
<td>0.795018</td>
<td>0.353030</td>
<td>0.867351</td>
</tr>
<tr>
<td>1080</td>
<td>0.720103</td>
<td>0.848168</td>
<td>0.763095</td>
<td>0.774038</td>
<td>0.420384</td>
<td>0.857111</td>
</tr>
<tr>
<td>1440</td>
<td>0.719991</td>
<td>0.854049</td>
<td>0.763378</td>
<td>0.774204</td>
<td>0.420384</td>
<td>0.857111</td>
</tr>
<tr>
<td>1800</td>
<td>0.728231</td>
<td>0.859293</td>
<td>0.787387</td>
<td>0.819956</td>
<td>0.477501</td>
<td>0.857714</td>
</tr>
<tr>
<td>2160</td>
<td>0.732612</td>
<td>0.859613</td>
<td>0.807248</td>
<td>0.829748</td>
<td>0.497927</td>
<td>0.865251</td>
</tr>
<tr>
<td>2520</td>
<td>0.735075</td>
<td>0.857378</td>
<td>0.815842</td>
<td>0.829748</td>
<td>0.497927</td>
<td>0.865251</td>
</tr>
<tr>
<td>2880</td>
<td>0.728048</td>
<td>0.862398</td>
<td>0.818616</td>
<td>0.816466</td>
<td>0.597405</td>
<td>0.857228</td>
</tr>
<tr>
<td>3240</td>
<td>0.733851</td>
<td>0.860097</td>
<td>0.828310</td>
<td>0.822511</td>
<td>0.617944</td>
<td>0.860439</td>
</tr>
<tr>
<td>3600</td>
<td>0.734742</td>
<td>0.863981</td>
<td>0.829789</td>
<td>0.822929</td>
<td>0.636366</td>
<td>0.859246</td>
</tr>
</tbody>
</table>

This experiment used six different feature selection methods. Table 4 is the value of $MacF_1$ in different features dimensions for six feature selections. Table 5 shows the comparison of $MacF_1$ by using paired $T$ test.

Table 5 shows that all comparisons of $MacF_1$ seem to reject the null hypothesis, thus demonstrating the CDDFS is better than the other feature selections when using Rocchio algorithms as classifier.

Table 5: Comparison of $MacF_1$ by using paired $t$ test

<table>
<thead>
<tr>
<th></th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR vs. CDDFS</td>
<td>-13.458</td>
<td>.000</td>
</tr>
<tr>
<td>CHI vs. CDDFS</td>
<td>-2.319</td>
<td>.046</td>
</tr>
<tr>
<td>MI vs. CDDFS</td>
<td>-3.488</td>
<td>.007</td>
</tr>
<tr>
<td>IG vs. CDDFS</td>
<td>-5.924</td>
<td>.000</td>
</tr>
<tr>
<td>DF vs. CDDFS</td>
<td>-9.932</td>
<td>.000</td>
</tr>
</tbody>
</table>

3.2.3 The experiment with SVM classifier

As we can see from the Figure 3 and Table 6, CDDFS offers a better performance than other feature selections except CHI when using SVM. Although CHI has the best performance, CDDFS is easier to understand and its calculation is simpler than CHI.
This experiment used six different feature selection methods. Table 6 is the value of $MacF_1$ in different dimensions for six feature selections. Table 7 shows the comparison of $MacF_1$ by using paired $t$ test.

Table 7 shows that the comparisons of $MacF_1$ by CHI vs. CDDFS do not reject the null hypothesis at a significance level of 0.05, thus demonstrating that using CDDFS cannot improve the classification performance with SVM. But the comparisons of $MacF_1$ by OR vs. CDDFS, MI vs. CDDFS, IG vs. CDDFS, DF vs. CDDFS seem to reject the null hypothesis, thus demonstrating that using CDDFS can improve the classification performance when using SVM as classifier.

4 Conclusions

In this paper, we analyze the distinguishability and representativeness of feature $t$ for category $C$. And then, we propose the degree of membership and degree of non-membership. At last, we propose the text feature selection method based on category-distribution divergence.

The new method is tested separately with five feature selection algorithms and three classifiers using Sogou corpora. The results show that CDDFS clearly offers a better performance than other five feature selection methods in the majority of tested cases. However, there are some limitations in our research. Firstly, the experiments only used the Sogou corpora. The other corpora like Reuters or 20 Newsgroups are not being used. Secondly, the experiments only verified the effectiveness in feature selection except the analysis of mathematics principles.

In future work, we would like to use other corpora to test the effectiveness of CDDFS and focus on the analysis of mathematics principles of this method.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No. 71373291).

References


