Dimension Reduction Based on Categorical Fuzzy Correlation Degree for Document Categorization

Qiang Li
Department of Computer Science and Technology
East China Normal University
Shanghai, China
liqiangecnu@gmail.com

Liang He*, Xin Lin
Department of Computer Science and Technology
East China Normal University
Shanghai, China
{lhe, xlin}@cs.ecnu.edu.cn

Abstract— High dimensionality of the feature space is a common problem in document categorization. Most of the features obtained through conventional feature selection algorithms such as IG are relevant and redundant. In this paper, a two-step feature selection method is proposed. At the first step redundancy analysis among original features based on categorical fuzzy correlation degree is applied to filter the redundant features with the similar categorical term frequency distribution. In the second step, conventional IG feature selection algorithm is adopted to select the final feature set for document categorization. Experiments dealing with the well-known Reuters-21578 and 20news-18828 corpuses show that the proposed method can eliminate redundant features with high fuzzy correlation degree between each other and obtain a compressed feature space where the dimension of feature space is dramatically reduced. The document categorization results on two corpuses show that the conventional IG feature selection algorithm can achieve a better document categorization performance on the compressed feature space and demonstrate the effectiveness of the proposed method.

Keywords- feature selection; fuzzy correlation degree; relevance; redundancy; document categorization

I. INTRODUCTION

Feature selection is the process of identifying and removing as much of the irrelevant and redundant information. In most existing document categorization algorithms, documents are represented as vector space model [2]. A major characteristic of this representation is the high dimensionality of the feature space, which costs much time or space for document categorization. Thus, it is important to develop the effective algorithms to recognize the key attributes and remove the useless features.

Now there are many feature selection methods for document categorization, such as document frequency (DF), term frequency (TF), mutual information (MI), χ² statistics (CHI), and information gain (IG), Expected Cross Entropy (ECE) etc. [1] [3] [4]. These conventional methods mostly adopt some kind of evaluation function to calculate the relevance between categories and individual features. For this reason, it merely has considered the relevance between feature and category, and ignored the relevance among features themselves, which leads to the redundancy in feature set.

In this paper, a novel algorithm based on fuzzy correlation degree is proposed to eliminate the redundant features which have similar categorical distribution between each others in original feature space. The proposed method makes full use of similar categorical term frequency distribution between pair of features to measure the redundancy of candidate features. Features with highly similar categorical distribution are identified by categorical fuzzy correlation degree (CFCD) and eliminated as redundant features. The difference from conventional feature selection is that a preselected feature set is obtained at first. Then IG algorithm is run to obtain the final feature set. Finally the document categorization experiments on Reuters-21578 corpus and 20Newsgroups-18828 corpus are adopted to validate the effectiveness of the proposed method.

The remainder of this paper is organized as follows. Section II briefly reviews well-known IG feature selection algorithm. Section III describes the proposed method in detail. In section IV the experimental results are presented using two English datasets. Finally, section V draws some conclusions and outlines the future work.

II. FEATURE SELECTION METHODS

In this section, we briefly describe well-known IG feature selection algorithm which are used to compare document categorization performance between original feature space and compressed feature space filtered by the proposed method in this paper.

IG feature selection algorithm assigns IG score to each of terms and all the terms are independently sorted according to the assigned score. Then, a predefined number of the best features are taken to form the final feature subset.

Information Gain (IG)

Information gain measures the amount of information obtained for category prediction by knowing the presence or absence of a term in a document. The information gain of term t is defined to be:
Calculation of Fuzzy Correlation Degree Based on Category (CFCD)

The proposed method implements 2-step dimension reduction to get the final feature set for document categorization.

Firstly the definition of Fuzzy Correlation Degree Based on Category (CFCD) is given.

Then the CFCD algorithm is proposed to identify and remove the redundant information among features according to the calculation of categorical fuzzy correlation degree between terms. A preselected feature set is obtained and the dimension of original feature space can be reduced greatly through the CFCD algorithm.

Finally the second dimension reduction is completed through conventional feature selection algorithms based on the preselected feature set obtained through CFCD algorithm.

A. Calculation of Fuzzy Correlation Degree Based on Category

It is very important in data statistical analysis to compute the correlation between variables or attributes. The correlation coefficients defined on crisp sets have been studied in many works of conventional statistics [11]. Usually the grades of membership functions of the two fuzzy sets are compared to see if there is any linear relationship between the two fuzzy sets. This paper defines a fuzzy set for each feature at first. Then a novel measurement criterion of fuzzy correlation degree between fuzzy sets based on categorical term frequency information is proposed to measure the correlation between features. Higher the fuzzy correlation degree between fuzzy sets, greater the redundancy between corresponding features.

**Definition 3.0** fuzzy set for each feature

Given original feature set \( T = \{t_1, t_2, ..., t_n\} \), category set \( C = \{c_1, c_2, ..., c_m\} \), if there is a fuzzy set \( h(t_i) = \{F(t_i, c_j) \mid \forall c_j \in C\} \) for each feature \( t_i \), then \( F(t_i, c_j) \) represents the categorical fuzzy importance degree of feature \( t_i \) for category \( c_j \).

**Definition 3.1** categorical fuzzy importance degree (CFID)

\[
F(t_i, c_j) = CTF(t_i, c_j)
\]

\( F(t_i, c_j) \) : Fuzzy importance degree of feature \( t_i \) for category \( c_j \)

\( CTF(t_i, c_j) \) : Term frequency of term \( t_i \) in category \( c_j \), \( j = \{1, 2, ..., m\} \)

\( m \) : The number of categories in a document set.

**Definition 3.2** categorical fuzzy correlation degree (CFCD)

\[
CRT(t_i, t_j) = \frac{\sum_{k=1}^{n} \min\{F(t_i, c_k), F(t_j, c_k)\}}{\sum_{k=1}^{n} \max\{F(t_i, c_k), F(t_j, c_k)\}}
\]

\( CRT(t_i, t_j) \) : Categorical fuzzy correlation degree between term \( t_i \) and term \( t_j \).

Definition 3.1 proposes the definition of categorical term frequency which measures the fuzzy importance degree of term \( t_i \) for category \( c_j \). Intuitively CTF values reflect the categorical importance of each term through term frequency over various categories. The higher CTF value of term \( t_i \) in some category, the more important the term \( t_i \) to the category.

Definition 3.2 is based on the assumption that redundant features have similar categorical distribution. For term \( t_i \) and term \( t_j \), the quotient between sum of categorical minimum term frequency and sum of categorical maximum term frequency over categories of the two terms can reflect the fuzzy similarity of two terms. According to CFCD in the definition 3.2, there exists \( 0 \leq CRT(t_i, t_j) \leq 1 \). If term \( t_i \) and term \( t_j \) are more similar in categorical distribution, the \( CRT(t_i, t_j) \) value is greater. Given the two terms have the same categorical distribution, the value of \( CRT(t_i, t_j) \) is 1. If \( CRT(t_i, t_j) \) is 0, the two terms have an inverse categorical distribution. In this paper, \( CRT(t_i, t_j) \) values are adopted to measure Categorical Fuzzy Correlation Degree between pair of terms. For two terms, if they have higher value of CRT, the redundancy of the two features is stronger.

B. Proposed CFCD algorithm

Given a set of class labels \( C = \{c_1, c_2, ..., c_m\} \), a candidate feature set \( T = \{t_1, t_2, ..., t_n\} \), the main task of CFCD algorithm is to reduce the redundant features among candidate feature set \( T \) based on the definition 3.2.

CFCD algorithm achieves the first dimension reduction in this paper and gets the preselected feature set \( S \). The proposed CFCD algorithm can be realized as follows:

1) Initialize \( S = \emptyset \)
2) Find first feature \( t_i = t_1 \) in candidate feature set \( T \)
   \( S = \{t_i\} \), candidate feature set \( T = T \setminus \{t_i\} \).
3) Sequential selection of each feature in candidate feature set T.
   a) Greedy feature selection:
      For all \( t_j \in T \), Computation of \( \text{CRT}(t_i, t_j) \)
      If \( \text{CRT}(t_i, t_j) > \theta \), candidate feature set \( T = T \setminus \{ t_j \} \)
      Repeat until each feature \( t_j \) in \( T \) are computed.
   b) Find next feature \( t_i \) in \( T \), \( S = S \cup \{ t_i \} \) and candidate feature set \( T = T \setminus \{ t_i \} \).
   c) Repeat 3) until \( T = \emptyset \)

The argument \( \theta \) is the threshold of CFCD, which is determined through the experiments.

The key of CFCD algorithm lies in reducing the redundant features which has a larger CRT with the already selected features. After the first dimension reduction, the second dimension reduction is accomplished by conventional feature selection algorithm such as IG to obtain the final feature set from the preselected feature set \( S \). Finally, document categorization experiments are adopted to validate the effectiveness of the proposed algorithm.

IV. EXPERIMENTS

Some experiments are performed to compare the overall F1-micro and F1-macro performance of proposed method with classical algorithm such as DF, CHI and IG from 200 to 2000 numbers of features subsets. We used Naïve SVM classifiers in WEKA [7] to measure the performance of each feature selection algorithm. And we evaluated selected feature subsets using 3-fold Cross-Validation (CV) for two corpuses. Our document preprocessing included removing all non-alphabetic characters like full stops, commas, brackets, etc., lowering the upper case characters, ignoring all the words that contained digits or non alpha-numeric characters and removing words from a stop-word list. The number of features is reduced by merging various word forms into one distinct term by using Porter Stemmer.

Each document \( d_j \) is represented by the normalized TFIDF vector \( g_j = \{ g_{ij1}, g_{ij2}, \ldots, g_{ijn} \} \) on selected feature subset [5] [6], \( g_{ij} \) represents vector weight value of the \( i^{th} \) term \( t_i \) in document \( d_j \), given by

\[
g_{ij} = \frac{t_{ij} \log(\frac{|D|}{DF(t_i)}) + 0.01)}{\sqrt{\sum_{i=1}^{n} (g_{ij} \log(\frac{|D|}{DF(t_i)}) + 0.01))^2}}
\]

Here \( t_{ij} \) is the frequency of the \( i^{th} \) term in document \( d_j \), \( DF(t_i) \) is document frequency of term \( t_i \), \( |D| \) is the number of documents in corpus.

A. Data set

In the experiments, the commonly used Reuters-21578 benchmark corpus, ModApte version [9] and the CMU’s 20Newsgroups-18828 dataset [10] are used to evaluate all considered algorithms.

Reuters-21578 benchmark corpus is a collection of 10,788 documents from the Reuters financial newswire service, partitioned into a training set with 7769 documents and a test set with 3019 documents in 135 classes related to economics. We used only 9645 documents merely attributed to one category in those 20 classes for which there exist at least one training and one testing documents. The resulting vocabulary size was 20032 terms by document preprocessing. The distribution of the Reuters-21578 corpus is shown in Table I.

The 20Newsgroups-18828 corpus is a collection of 18,828 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups and duplicates removed. We used 9535 documents in those commonly used 10 classes. The resulting vocabulary size was 44438 terms by document preprocessing. The distribution of the 20Newsgroups-18828 corpus is shown in Table II.

<table>
<thead>
<tr>
<th>category</th>
<th>files_num</th>
<th>category</th>
<th>files_num</th>
</tr>
</thead>
<tbody>
<tr>
<td>acq</td>
<td>2347</td>
<td>jobs</td>
<td>51</td>
</tr>
<tr>
<td>coffee</td>
<td>122</td>
<td>livestock</td>
<td>54</td>
</tr>
<tr>
<td>corn</td>
<td>9</td>
<td>lumber</td>
<td>14</td>
</tr>
<tr>
<td>cotton</td>
<td>26</td>
<td>money-fx</td>
<td>629</td>
</tr>
<tr>
<td>crude</td>
<td>490</td>
<td>money-supply</td>
<td>164</td>
</tr>
<tr>
<td>earn</td>
<td>3949</td>
<td>reserves</td>
<td>56</td>
</tr>
<tr>
<td>gas</td>
<td>34</td>
<td>ship</td>
<td>192</td>
</tr>
<tr>
<td>gold</td>
<td>114</td>
<td>sugar</td>
<td>133</td>
</tr>
<tr>
<td>grain</td>
<td>498</td>
<td>trade</td>
<td>423</td>
</tr>
<tr>
<td>interest</td>
<td>318</td>
<td>wheat</td>
<td>22</td>
</tr>
</tbody>
</table>
TABLE II. 20NEWSGROUP-18828 CORPUS DISTRIBUTION TABLE

<table>
<thead>
<tr>
<th>category</th>
<th>files_num</th>
<th>category</th>
<th>files_num</th>
</tr>
</thead>
<tbody>
<tr>
<td>alt.atheism</td>
<td>799</td>
<td>sci.crypt</td>
<td>991</td>
</tr>
<tr>
<td>comp.sys.mac.hardware</td>
<td>961</td>
<td>sci.electronics</td>
<td>981</td>
</tr>
<tr>
<td>misc.forsale</td>
<td>972</td>
<td>soc.religion.christian</td>
<td>997</td>
</tr>
<tr>
<td>rec.autos</td>
<td>990</td>
<td>talk.politics.guns</td>
<td>910</td>
</tr>
<tr>
<td>rec.sport.baseball</td>
<td>994</td>
<td>talk.politics.mideast</td>
<td>940</td>
</tr>
</tbody>
</table>

B. Evaluation Criteria

Evaluation criteria F (macro-recall) and F (micro-recall) are adopted to evaluate the document categorization performance based on various feature selection algorithms [8].

Equation (1) (2) (3) shows the definition of precision, recall and F1 measure for a given category $c_i$.

\[
P_{ci} = \frac{TP_i}{(TP_i + FP_i)} \]  
\[
R_{ci} = \frac{TP_i}{(TP_i + FN_i)} \]  
\[
F_{ci} = \frac{2P_{ci}R_{ci}}{(P_{ci} + R_{ci})} \]

$p_{ci}$: Precision of category $c_i$ ;

$r_{ci}$: Recall of category $c_i$ ;

$f_{ci}$: F1 metric for a given category $c_i$ ;

$TP_i$: Attributed number of documents to category $c_i$ correctly;

$FP_i$: Attributed number of documents to category $c_i$ wrongly;

$FN_i$: Number of documents isn’t assigned to category $c_i$ but they are in category $c_i$ ;

We can compute two average values F (Micro-average) and F (Macro-average) over all categories to assess the overall performance of each classifier.

F (Macro-average) can be computed by equations (4).

\[
F(Macro-average) = \frac{\sum_{i=1}^{m} F_{ci}}{m} \]

Equations (5) (6) (7) are adopted to compute F (Micro-average).

\[
P_{micro} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FP_i)} \]  
\[
R_{micro} = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FN_i)} \]  
\[
F(Micro-average) = \frac{2P_{micro}R_{micro}}{P_{micro} + R_{micro}} \]

From the above definitions, F (Micro-average) value is mainly influenced by document categorization results on common categories which include most of the documents in corpuses, and F (Macro-average) is sensitive to rare ones. We use both metrics to compare different algorithms on Reuters-21578 corpus and 20Newsgroup corpus.

C. Experiment results and analysis

Document categorization experiments are used to validate the effectiveness of the proposed method. The novel method is compared with other classical feature selection through F (Micro-Average) and F (Macro-Average) metrics illustrated above.

In this paper, the argument $\theta$ is the threshold of $\text{CRT}(t_i,t_j)$, which is determined through the experiments.

When the threshold is determined as 0.95, the preselected features of Reuters-21578 corpus are reduced from the original 20032 features to 4668 features. For 20Newsgroup corpus, the size of feature space is reduced from the original 44438 to 9262. Table III and table IV is the experimental F (Macro-average) and F (Micro-average) comparison result of Reuters-21578 corpus and 20Newsgroup corpus on SVM classifier between the proposed method CFCD+IG and IG algorithm in 200 to 1000 features while $\theta=0.95$. From table III and table IV, while the same number of features are selected, the F1-measure of CFCD+IG has been enhanced compared with IG. For example, when the selected 200 features, F (Micro-average) on Reuters-21578 corpus is increased from 0.837 to 0.853 and F (Micro-average) on 20Newsgroup corpus is increased from 0.603 to 0.645. The experimental comparison result of CFCD+IG and IG on two corpuses shows that the proposed CFCD algorithm can not only preserve the features with categorical distinctive characteristics but also remove the plenty of redundant features which have the similar categorical distribution.

The proposed method fulfills dimension reductions for two times. By redundancy analysis of CFCD algorithm, the feature space of two corpuses is reduced greatly. Plenty of features with similar categorical term frequency distribution are eliminated which are ignored in conventional feature
selection algorithms. Then IG algorithm completes the second dimension reduction.

Document categorization experimental results demonstrate the effectiveness of the novel method and shows that the combination of conventional IG feature selection algorithm and the proposed redundancy analysis method can obtain more representative features which have less categorical distribution redundancy between each other and enhance the performance of document categorization.

<table>
<thead>
<tr>
<th>Method</th>
<th>CFCD+IG</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of Feature Space</td>
<td>4668</td>
<td>20032</td>
</tr>
<tr>
<td>Selected Features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200 F (Micro-average) F (Macro-average)</td>
<td>0.853 0.489</td>
<td>0.837 0.442</td>
</tr>
<tr>
<td>400 F (Micro-average) F (Macro-average)</td>
<td>0.884 0.582</td>
<td>0.884 0.584</td>
</tr>
<tr>
<td>600 F (Micro-average) F (Macro-average)</td>
<td>0.899 0.651</td>
<td>0.892 0.613</td>
</tr>
<tr>
<td>800 F (Micro-average) F (Macro-average)</td>
<td>0.903 0.674</td>
<td>0.902 0.665</td>
</tr>
<tr>
<td>1000 F (Micro-average) F (Macro-average)</td>
<td>0.901 0.670</td>
<td>0.902 0.671</td>
</tr>
</tbody>
</table>

V. CONCLUSION

In this paper, a two step document categorization selection method is presented. At the first step, a filtering approach through CFCD measure is used for redundancy analysis. The CFCD algorithm reduces the high dimensionality of feature space by removing redundant features with similar categorical term frequency distribution and passes a proper preselected feature set to the next step. The feature selection computational time in second step is reduced. The experiments on two English corpuses in section IV validate the proposed method. The reduced feature set provides better accuracy and less computation time than the original data set. In the future, we plan to combine the novel method with semantic redundant analysis through semantic dictionary to optimize the feature space further.

ACKNOWLEDGMENT

This work is supported by the national science & technology support program (2012BAH74F02).

REFERENCES


