A New Feature Selection Method in Fishery Information Processing

Jun Gu
College of Information Engineering
Dalian Ocean University
Dalian, China

Nan He
College of Information Engineering
Dalian Ocean University
Dalian, China

Abstract—**Fishery information processing can help fishery researchers obtain the needed information easily and quickly. The current information processing techniques have not solved the problem of high dimensional features in fishery information processing. In this paper, a feature selection method for fishery texts based on SVM-RFE was put forward in view of the characteristics of fishery texts. It removed the redundant information in text feature space and reduced the feature dimensions effectively. Three corpora were employed to verify the proposed method and the comparison with the traditional feature selection method was performed. The experimental results show that the method proposed in this paper can improve precision rate and recall rate of fishery information processing with the lower dimensional features, providing an effective way for fishery information processing.**

Keywords-fishery information processing; feature selection; precision rate; recall rate

I. Introduction

In recent years, fishery researchers have exposure to more and more electronic information due to the rapid development of Internet. The electronic information always exists in the form of text[1,2]. So, it becomes an urgent problem for fishery researchers to find the potentially valuable information efficiently in the face of such a large volume of text data. Text categorization technology is an effective way to solve this problem.

When current text categorization techniques are used to construct the text feature vectors, they always lead to high dimensionality and sparsity of the feature space[3,4,5,6]. Because of the ubiquitous synonyms and related words in Chinese texts, and in particular the speciality of the texts in the field of fishery, the text feature vectors which are constructed using traditional text feature selection methods, such as TF·IDF, information gain and mutual Information, etc, always own very huge dimensions. Consequently the following categorization algorithms are influenced in terms of the calculating time performance. In some worse cases, many categorization algorithms even cannot deal with these high dimensional feature vectors[7]. On the other hand, in these high dimensional feature vectors, many features are related to each other. So their effects for the categorization are similar, and the accuracy of text categorization does not improve due to the increase in dimension.

Recursive feature elimination method based on support vector machine (SVM-RFE) is a feature selection algorithm, which has been successfully applied to gene selection, remote sensing analysis, signal processing and many other fields[8, 9,10,11,12,13,14]. In this paper, to resolve the problem of high dimensionality in fishery information processing, a feature selection method based on SVM-RFE was proposed. The initial feature space was built up according to TF·IDF, and then SVM-RFE method was employed to reduce the feature dimensions. At last, the optimal value of feature dimensions was discussed on this basis. The method put forward in this paper lays a good experimental foundation for further improving the efficiency of fishery information processing.

II. Vector space model

In text feature extraction, text data are typically depicted using vector space model (VSM). The VSM method was put forward by Salton and his colleagues in 1970s, and it was successfully applied to the famous SMART text retrieval system[15, 16]. VSM views the text set as a vector space which is comprised of feature words. When VSM is used, the text matrix is built up and every feature item in the text vector represents the important degree of a feature word for this text.

The extractions of the feature items are mostly based on the frequencies of feature words. At present, TF·IDF is a widely accepted method. TF·IDF takes into account both the frequency of a feature word in one document and all documents. When a feature word appears in too many documents, it is deemed to be too general and not important. Contrarily, when the frequency of a feature word in one document is very high and that in other documents is very low, that feature word can be considered to be important. TF·IDF definition is defined as follows[17]:

$$TF \cdot IDF(t_i, d_j) = TF_{ij} \times \log(\frac{N}{DF_i}) \qquad （1）$$

Where t_i represents the ith feature word and d_j the jth document, and TF_{ij} represents the frequency of t_i in d_j. DF_i is the number of documents which contain the feature word t_i in the background corpus. N is the total number of documents in the corpus.

III. SVM-RFE

A．Support Vector Machine

Support vector machine (SVM) is a machine learning method based on statistics. It performs categorization by the establishment of the hyperplane described by the weight vector w and the error term[18,19,20], as shown in Fig. 1.

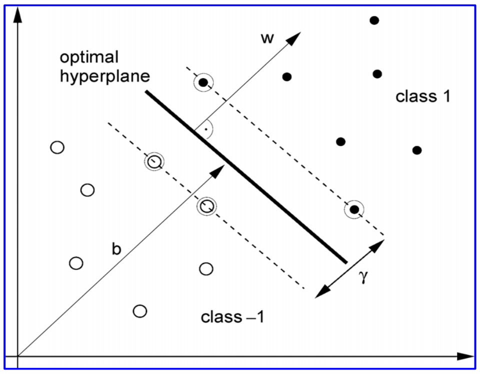

Fig.1 SVM hyperplane

In SVM method, the training set is composed of s samples, and each of these samples contains x_i, the feature vector of the ith sample, and the corresponding class label y_i:

$$x_1,y_1),(x_2,y_2),\ldots,(x_s,y_s)\in R^N\times\{-1,1\} \tag{2}$$

A specific hyperplane, which normal vector is w, can be found out by learning. In the testing stage, the label of a new feature vector x can be predicted by projecting x in the w direction：

$$f(x)=w\cdot x+b \tag{3}$$

Symbolic of the projection represents the predicted category label. Under the linearly inseparable situation, the soft margin SVM is constructed according to the maximum margin criterion:

$$\text{Minimize}\quad \frac{1}{2}\|w\|^2+C\sum_i\xi_i$$
$$\text{Subject to}\quad y_i(wx_i+b)\geq 1-\xi_i \tag{4}$$

Where C represents the penalty factor, and ξ_i represents relaxation factor. The value of ξ_i must be more than zero. The above formula is converted to the Lagrange formula and calculated for its maximum, and then the expression of the weight vector w is obtained as follows:

$$w=\sum_{i=1}^{N_s}y_i\alpha_i x_i \tag{5}$$

Where 0≤α_i≤C, $\sum_{i=1}^{s}\alpha_i y_i$=0, N_s represents the number of the support vectors. Substituting equation 5 into equation 3, then we can reach the following optimal decision function:

$$f(x)=\sum_{i=1}^{N_s}y_i\alpha_i(x\cdot x_i)+b\,. \tag{6}$$

For the nonlinear problem, we need the nonlinear transformation to translate it to a linear problem of another feature space. And then the optimal hyperplane can be constructed and the corresponding optimal decision function is:

$$f(x)=\sum_{i=1}^{N_s}y_i\alpha_i K(x\cdot x_i)+b \tag{7}$$

Where $K(x\cdot x_i)$ is the kernel function which meeting the Mercer condition.

B. SVM-RFE

Recursive feature elimination(RFE) is a dimension reduction method proposed by Guyon and his colleagues to obtain the optimal feature subset under some certain feature rank criterion. The SVM-RFE method is the expansion of the feature rank criterion in SVM for RFE[21]. It is a backward sequence reduction algorithm based on the maximum interval principle in SVM. In RFE method, SVM is used to be the classifier to execute the recursive elimination of features, and the ranking criteria is drawn up according to the trained SVM parameters.

The specific calculating steps are:

a) Name the training sample matrix as $\boldsymbol{X_0}$=[$x_1,x_2,\ldots,x_l$]^T, and the corresponding categories tags as $\boldsymbol{y}$=[$y_1,y_2,\ldots,y_l$]^T. Initialize the current feature vector $\boldsymbol{s}$=[1,2,…,k]^T and feature ranking vector $\boldsymbol{r}$=[ ];

b) Get the new training sample matrix $\boldsymbol{X}$=$\boldsymbol{X}_0$(：, s) according to the rest features, and train the classifier $\boldsymbol{\alpha}$= SVMtrain($\boldsymbol{X}$, $\boldsymbol{y}$), and calculate the weight vector $w=\sum_{i=1}^{N_s}y_i\alpha_i x_i$;

c) Calculate the ranking criteria c_i=$\|w_i\|^2$, and search the feature which gets the smallest ranking score, getting *f=argmin(c)*;

d) Update the feature ranking vector ***r*** =[***s****(f)*，***r***], and eliminate the feature which has the minimum score , getting ***s*** = ***s****(1: f-1，f+1:length(****s****))*;

e) If ***s*** =[ ], go to step 6, else go to step 2;

f) Output the feature ranking vector ***r***.

The flowchart of SVM-RFE method is shown in Fig. 2.

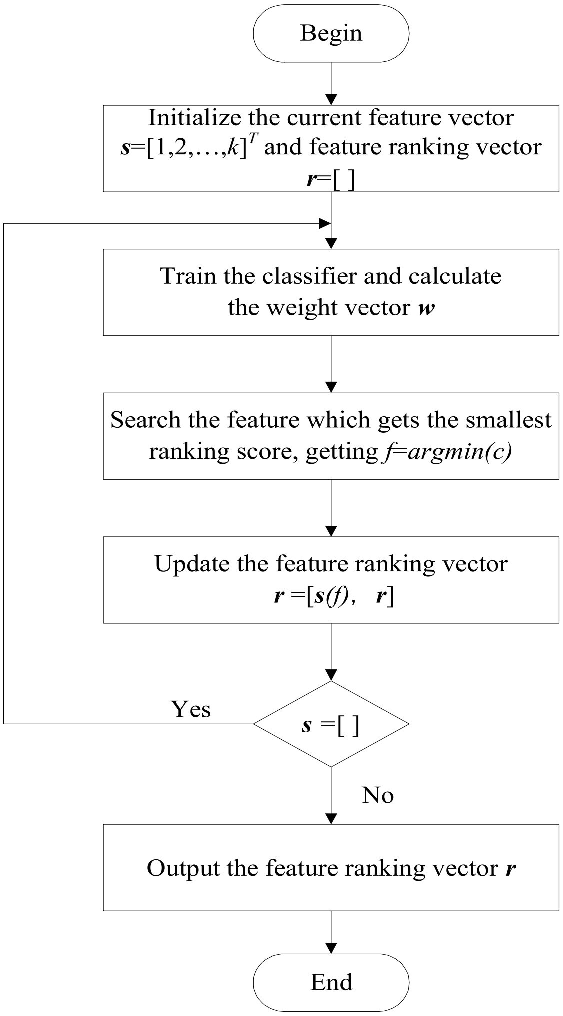
Begin
Initialize the current feature vector $\boldsymbol{s}=[1,2,\ldots,k]^T$ and feature ranking vector $\boldsymbol{r}=[\,]$
Train the classifier and calculate the weight vector $\boldsymbol{w}$
Search the feature which gets the smallest ranking score, getting $f=argmin(c)$
Update the feature ranking vector $\boldsymbol{r}=[\boldsymbol{s}(f),\ \boldsymbol{r}]$
Yes
$\boldsymbol{s}=[\,]$
No
Output the feature ranking vector $\boldsymbol{r}$
End

Fig.2 The flow chart of SVM-RFE

IV. Experiments and Results Analysis

A. The Corpora

Considering there are no public corpora for the specific fishery area, three corpora were employed to verify the proposed method in this paper:

a) Binary categorization for agriculture and non-agriculture was executed on the public corpus provided by Fudan University Natural Language Processing Group. The corpus is composed of about 20 text categories, including agriculture, environment, art, economy, electronic, etc;

b) Binary categorization for fishery and non-fishery was executed on the adjusted corpus from the public one provided by Fudan University Natural Language Processing Group. Fishery texts were handpicked from the public corpus in the light of fishery correlation;

c) Binary categorization for fishery and non-fishery was executed on the corpus composed of the papers in fishery and non-fishery downloaded from Chinese National Knowledge Infrastructure(CNKI). These papers were converted to the text type firstly.

The training text set and the test text set were Chinese word segmented using the ICTCLAS system provided by Institute of Computing Technology of Chinese Academy of Science, and then the VSM model was built up. Feature items of the feature vector corresponding to every text file were calculated by use of TF·IDF method.

B. Feature Selection and Categorization

SVM-RFE method was employed to carry out the feature selection for the high dimensional feature vectors which were produced during the above steps, and the parameter *C* is set to 100. Considering the time-consuming of SVM-RFE, multiple features were eliminated in every recursive process in order to improve the computation efficiency. The number of features to eliminate decreased as the number of remained ones reduced. The categorization accuracies with specific dimensions were calculated by use of the SVM classifier and the 10-fold cross-validation method. The dimension value with which the precision rate achieved its maximum was chosen to be the optimal value of feature dimensions of SVM-RFE method. And the got feature set corresponding to the optimal value of feature dimensions was supposed to be the optimal feature set. Fig. 3 showed the investigation results of the first corpus in the dimensions of less than 3000.

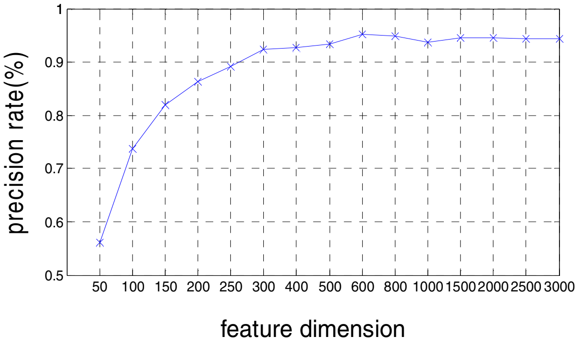

Fig.3 Precision rates with different feature dimensions of the first corpus

We saw that after the value of feature dimensions reached a certain value, precision rates of text categorization did not continue to improve with the increase of feature dimensions. This appearance indicated that too much feature dimensions could not always improve precision rates of text categorization. Moreover, it was also displayed in Fig. 3 that precision rate reached the maximum when the value of feature dimensions was 600. So, we chose 600 to be the optimal value of feature dimensions, and recorded the first 600 features in the feature ranking list corresponding to this optimal value. Similarly, the optimal values of feature dimensions for the second and third corpora were got to be 500 and 200 respectively, as shown in Fig. 4 and Fig. 5, and the corresponding optimal feature sets were obtained.

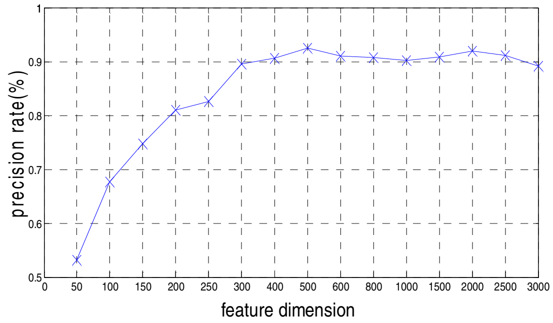

Fig.4 Precision rates with different feature dimensions of the second corpus

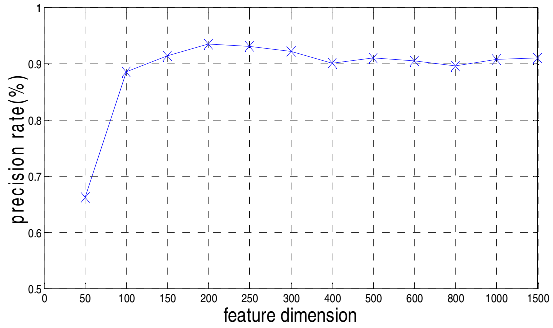

Fig.5 Precision rates with different feature dimensions of the third corpus

According to the above calculated optimal feature sets, a linear SVM classifier was trained using the training text set. Then the trained classifier model was employed to test the testing text set. The categorization results were assessed using two universal general indexes: precision rate and recall rate[22]. In order to demonstrate the superiority of the proposed method, we compared the experimental results with TF·IDF method. The detailed assessments of three corpora were listed in table I, table II and table III.

TABLE I. RESULTS OF THE FIRST CORPUS

Method	**TF·IDF**		**SVM-RFE**	
	Agriculture	***Non-agriculture***	***Agriculture***	***Non-agriculture***
Precision rate	85.1	85.7	95.2	94.2
Recall rate	83.4	87.3	93.3	95.9

TABLE II. RESULTS OF THE SECOND CORPUS

Method	**TF·IDF**		**SVM-RFE**	
	Fishery	***Non-fishery***	***Fishery***	***Non-fishery***
Precision rate	86.1	88.8	92.4	94.4
Recall rate	86.9	88.2	93.7	93.5

TABLE III. RESULTS OF THE THIRD CORPUS

Method	**TF·IDF**		**SVM-RFE**	
	Fishery	***Non-fishery***	***Fishery***	***Non-fishery***
Precision rate	86.5	86.4	93.5	92.3
Recall rate	84.1	88.5	91.0	94.5

As we saw from table I, our method completed the binary categorization for agriculture and non-agriculture satisfactorily based on the public corpus. This results illustrated the proposed method was applicable for text categorization. At the same time, it was shown in table II and table III that our method also achieved the binary categorization for fishery and non-fishery satisfactorily in the light of the adjusted fishery corpora.

The proposed method could reach the higher precision rates than TF·IDF method with the reduced feature dimensions. It was because that SVM-RFE method removed the correlation between retained features through the elimination of features which made the feature criterion function changed minimally in each iteration for feature selection. This principle impelled SVM-RFE remove the redundant information from the feature space effectively. While the traditional TF·IDF method only took into account the frequency of a feature word in one document and all documents, without considering the correlation between feature words. Therefore, the constructed high dimensional feature space had not improved the precision rate of text categorization because of the increasing feature dimensions. On the contrary, it leaded to the over-fitting phenomenon and the humble text categorization performance. The experimental results show that fishery texts have the higher dimensional feature space, and we can effectively reduce the dimension of features and improve precision rate and recall rate through feature selection by utilizing of the proposed method.

V. CONCLUSIONS

Fishery texts have their specialty and have less common words with the ones in other areas. The text feature space constructed by traditional feature selection methods has the shortage of high dimensionality and sparsity. A feature selection method based on SVM-RFE was put forward in this paper. It provided the method to determine the optimal value of feature dimensions, removing the correlation between the features, and then decreased the feature dimensions of fishery texts. Three corpora were used to test our method and the comparison with the traditional method was implemented. The experimental results indicate that the proposed method can effectively improve the categorization performance for fishery texts, being worth further researching in the area of fishery information processing.

REFERENCES

[1] D. Lu and Q. M. Yu, “The model of SOA-based fishery information resources integration,” Control & Automation, vol. 26, pp. 28-30, 2010.(in Chinese)

[2] S. M. Zhang, H. Zhang and W. Fan, “Design and implementation of ocean catch data input and analysis module: an example of Chilean jack mackerel fishery,” Journal of Dalian Ocean University, vol. 26, pp. 162-167, 2011.(in Chinese)

[3] R. G. Rossi, T. P. Faleiros and A. A. Lopes, “Inductive model generation for text categorization using a bipartite heterogeneous network,” 2012 IEEE 12th International Conference on Data Mining. Brussels, Belgium, 2012, pp. 1086 – 1091.

[4] S. J. Lee and J. Y. Jiang, “Multi-label text categorization based on fuzzy relevance clustering,” IEEE Transactions on Fuzzy Systems, vol. PP, pp. 1-16, 2013.

[5] S. Abdul-Rahman, S. Mutalib and N. A. Khanafi, “Exploring feature selection and support vector machine in text categorization,” 2013 IEEE 16th International Conference on Computational Science and Engineering. Australia: Sydney, 2013, pp. 1101 – 1104.

[6] R. G. Rossi, T. de Paulo Faleiros and A. de Andrade Lopes, “Inductive model generation for text categorization using a bipartite heterogeneous network,” 2012 IEEE 12th International Conference on Data Mining. Belgium:Brussels, 2012, pp. 1086 – 1091.

[7] Z. Y. Xiong, L. L. Fu and Y. F. Zhang, “Mixed method of feature reduction based on concept mapping in text classification,” Computer Engineering and Applications, vol. 48, pp. 166-169, 2012. (in Chinese)

[8] Y. Bazi and F. Melgani, “Toward an optimal SVM classification system for hyperspectral remote sensing images,” IEEE Transactions on Geoscience and Remote Sensing, vol. 44, pp. 3374-3385, 2006.

[9] L. Yu, Y. Han and M.E. Berens, “Stable gene selection from microarray data via sample weighting,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, pp. 262 – 272, 2012.

[10] J. Canul-Reich, L. O. Hall and D. Goldgof, “Filtering for improved gene selection on microarray data,” 2010 IEEE International Conference on Systems Man and Cybernetics. Turkey:Istanbul, 2010, pp. 3250 – 3257.

[11] N. Longepe, P. Rakwatin and O. Isoguchi, “On the use of support vector machines for land cover analysis with L-band SAR data,” 2010 IEEE International Geoscience and Remote Sensing Symposium. USA: Honolulu, 2010, pp. 3263 – 3266.

[12] J. T. Atkinson, R. Ismail and M. Robertson, “Mapping bugweed (Solanum mauritianum) infestations in pinus patula plantations using hyperspectral imagery and support vector machines,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, pp. 17-28, 2014.

[13] Y. S. Park, T. I. Netoff and K. K. Parhi, “Reducing the number of features for seizure prediction of spectral power in intracranial EEG,” 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers. USA: Pacific Grove, 2012, pp. 770 – 774.

[14] B. L. Koley and D. Dey, “Selection of features for detection of Obstructive Sleep Apnea events,” 2012 Annual IEEE India Conference. Japan: Kochi, 2012, pp. 991 – 996.

[15] G. Salton, A. Wong and C. S. Yang, “On the specification of term values in automatic indexing,” Journal of Documentation, vol. 29, pp. 351-372, 1973.

[16] P. Nedungadi, H. Harikumar and M. Ramesh, “A high performance hybrid algorithm for text classification,” 2014 Fifth International Conference on the Applications of Digital Information and Web Technologies. India: Bangalore, 2014, pp. 118 – 123.

[17] X. F. Wang, R. F. Wang and S. G. Zhang, “An effective unsupervised feature computing model,”. Journal of Jilin University(Science Edition), vol. 48, pp. 79-84, 2010. (in Chinese)

[18] R. Baly and H. Hajj, “Wafer classification using support vector machines,” IEEE Transactions on Semiconductor Manufacturing, vol. 25, pp. 373-383, 2012.

[19] Y. Wang, E. W. M. Ma and T. W. S. Chow, “A two-step parametric method for failure prediction in hard disk drives,” IEEE Transactions on Industrial Informatics, vol. 10, pp. 419 – 430, 2014.

[20] R. Phan and D. Androutsos, “Robust semi-automatic depth map generation in unconstrained images and video sequences for 2D to stereoscopic 3D conversion,” IEEE Transactions on Multimedia, vol. 16, pp. 122 – 136, 2014.

[21] I. Guyon, J. Weston and S. Barnhill, “Gene selection for cancer classification using support vector machines,” Machine Learning, vol. 46, pp. 389-422, 2002.

[22] J. Sarmah, A. K. Barman and S. K. Sarma, “Automatic assamese text categorization using WordNet,” 2013 International Conference on Advances in Computing, Communications and Informatics. India: Mysore, 2013, pp. 85 – 89.