Web News Classification Using Neural Networks Based on PCA

Ali Selamat, Hidekazu Yanagimoto and Sigeru Omatu
Division of Computer and Systems Sciences,
Engineering Department, Osaka Prefecture University,
Sakai, Osaka 599-8531, Japan.
aselamat@sig.cs.osakafu-u.ac.jp, {hidekazu,omatu}@cs.osakafu-u.ac.jp

Abstract: In this paper, we propose a news web page classification method (WPCM). The WPCM uses a neural network with inputs obtained by both the principal components and class profile-based features (CPBF). The fixed number of regular words from each class will be used as a feature vectors with the reduced features from the PCA. These feature vectors are then used as the input to the neural networks for classification. The experimental evaluation demonstrates that the WPCM provides acceptable classification accuracy with the sports news datasets.

Keywords: text categorization, WWW, profile-based neural networks, principal component analysis.

1. Introduction

Neural networks have been widely applied by many researchers to classify the text documents with different types of feature vectors. Wermeter has used the document title as the feature vectors to be used for document categorization [1]. Lam et al. have used the principal component analysis (PCA) method as a feature reduction technique to the input data for the neural networks [2]. However, if some original terms are particularly good when discriminating a class category, the discrimination power may be lost in the new vector space after using the Latent Semantic Indexing (LSI) as described by Sebastini [3]. Retrieval techniques based on dimensionality reduction, such as LSI, have been shown to improve the quality of the information being retrieved by capturing the latent meaning of words present in the documents. A limitation of PCA or LSI in supervised data is the characteristic variables that describe smaller classes tend to be lost as a result of the dimensionality reduction. Hence, the classification accuracy on the smaller classes can be degraded in the reduced dimensional space [4].

Here, we propose a web page classification method (WPCM), which is base on the PCA and class profile-based features (CPBF). Each web page is represented by the term frequency-weighting scheme. As the dimensionality of a feature vector in the collection set is big, the PCA has been used to reduce it into a small number of principal components. Then we combine the feature vectors generated from the PCA with the feature vectors from the class-profile which contains the most regular words in each class before feeding them to the neural networks for classification. We have manually selected the most regular words that exist in each class and weighted them using an entropy weighting scheme [5]. The Yahoo sports news web pages have been used for the classification purpose [6]. The PCA, TF-IDF, Bayesian and WPCM methods have been used as a benchmark test for the classification accuracy. The experimental evaluation demonstrates that the proposed method provides acceptable classification accuracy with the sports news datasets. The organization of this paper is as follows: The news classification using the WPCM is described in Chapter 2. The preprocessing of web pages is explained in Chapter 3. The comparisons of web pages classification using the PCA, TF-IDF, Bayesian, and WPCM methods are discussed in Chapter 4. In Chapter 5, the discussions of the classification results using different methods for the sports news web pages are stated. In Chapter 6, we will conclude the classification accuracy by using the WPCM compared with other methods.

2. Web Page Classification Method

The news web pages have different characteristics where the text length for each of them is variable. Also the structures of the pages are different in the tags usage (i.e., XML, html, SGML tags, etc.). Furthermore, a huge number of distinct words exist in those pages as there is no restriction on word usage in the news web pages discussed by Mase et al. [7]. The high dimensionality of the news web pages dataset has made the process of classification difficult. This is because there are many categories of news in the web news pages such as sports, weathers, politics, economy, etc. In each category there are many different classes. For example, the classes that exist in the business category are stock market, financial investment, personal finance, etc. Our approach is based on the sports news category of web pages. In order to classify the news web pages, we propose the WPCM which uses the PCA and CPBF as the input to the neural networks. Firstly, we have used the PCA algorithm to reduce the original data vectors to a small number of relevant features [8]. Then we combine these features to the CPBF before inputting them to the neural networks for classification as shown in Fig. 1.

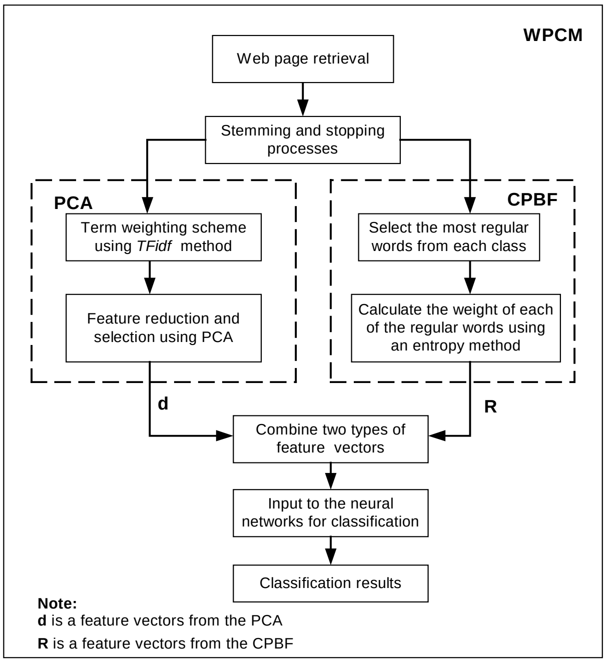
WPCM
Web page retrieval
Stemming and stopping processes
PCA
Term weighting scheme using *TFidf* method
Feature reduction and selection using PCA
CPBF
Select the most regular words from each class
Calculate the weight of each of the regular words using an entropy method
d
R
Combine two types of feature vectors
Input to the neural networks for classification
Classification results
Note:
d is a feature vectors from the PCA
R is a feature vectors from the CPBF

Fig 1 The classification process of a news web page using the WPCM method.

3. Preprocessing of Web Pages

The classification process of a news web page using the WPCM method is shown in Fig. 1. It consists of web news retrieval process, stemming and stopping processes, feature reduction process using our proposed method, and web classification process using error back-propagation neural networks. The retrieving process of sports news web pages has been done by our software agent during night-time [9]. Only the latest sports news web pages category will be retrieved from the Yahoo web server from the WWW. Then these web pages will be stored in the local news database. After that the stemming and stopping processes of the terms exist in each document will take place. Stopping is a process of removing the most frequent word that exists in a web page document such as 'to', 'and', 'it', etc. Removing these words will save spaces for storing document contents and reduce time taken during the search process. Stemming is a process of extracting each word from a web page document by reducing it to a possible root word. For example, the words 'compares', 'compared', and 'comparing' have similar meaning with a word 'compare'. We have used the Porter stemming algorithm to select only 'compare' to be used as a root word in a web page document [10]. Each term in the document will be represented as a term frequency. Term frequency TF_{jk} is the number of how many times the distinct word w_k occurs in document Doc_j where $k=1,...,m$. Doc_j is referring to each web page document that exist in the news database where $j=1,...,n$. The calculation of the terms weight x_{jk} of each word w_k is done by using a method that has been used by Salton which is given by $x_{jk}=TF_{jk}\times idf_k$ [10]. The document frequency df_k is the total number of documents in the database that contains a word w_k , and the inverse document frequency $idf_k=\log({}^{n}\!/_{df_k})$ where n is the total number of documents in the database.

4. Feature Reduction Using the PCA

Suppose that we have A, which is a matrix document-terms weight as below,

$$A=\begin{pmatrix} x_{11} & x_{12} & x_{1k} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{2k} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{nk} & \cdots & x_{nm} \end{pmatrix} \quad (1)$$

where x_{jk} is the terms weight that exist in the collection of documents. The definitions of j,k,l,m and n have been described in previous paragraph. There are a few steps to be followed in order to calculate the principal components of data matrix A. The mean of m variables in data matrix A will be calculated as $\bar{x}_k={}^{1}\!/_{n}\left(\sum_{j=1}^{n}x_{jk}\right)$. After that the covariance of matrix $S=\{s_{jk}\}$ is calculated. The covariance of s_{ik} is given by $s_{ik}={}^{1}\!/_{n}\sum_{j=1}^{n}\left(x_{ji}-\bar{x}_i\right)\left(x_{jk}-\bar{x}_j\right)$ where $i=1,\ldots,m$. Then we determine the eigenvalues and eigenvectors of the covariance matrix S which is a real symmetric positive matrix. An eigenvalue λ and nonzero vector e can be found such that, $Se=\lambda e$ where e is an eigenvector of S. In order to find a nonzero vector e the characteristic equation $|S-\lambda I|=0$ must be solved. If S is an $m\times m$ matrix of full rank, m eigenvalues $\lambda_1,\lambda_2,\cdots,\lambda_m$ can be found. By using $(S-\lambda I)e=0$, all corresponding eigenvectors can be found. The eigenvalues and corresponding eigenvectors will be sorted so that $\lambda_1\le\lambda_2\le\cdots\le\lambda_m$. The eigenvector matrix is represented $e = [u_1, u_2, \cdots, u_m]$. A diagonal nonzero eigenvalue matrix is represented as

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_m \end{pmatrix}. \tag{2}$$

In order to get the principal components of matrix S, we will perform eigenvalue decomposition that is given by $S = e\Lambda e^T$. Then we select the first $d \le m$ eigenvectors where d is the desired value, e.g., 100, 200, 400, 600 etc. The set of principal components is represented as $Y_1 = e_1^T x, Y_2 = e_2^T x, \cdots, Y_d = e_d^T x.$

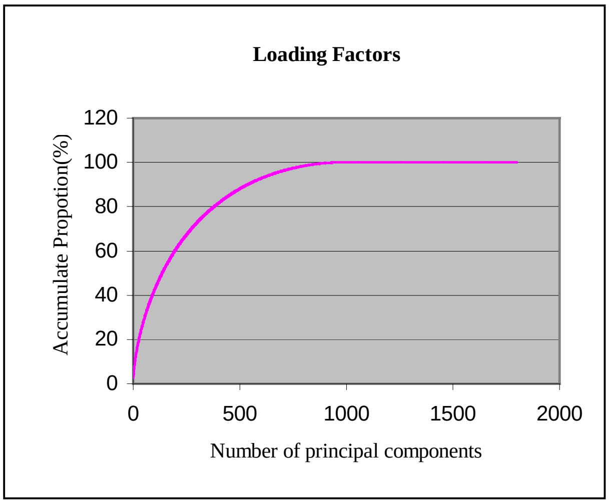

Fig. 2 Accumulated proportion of principal components generated by the PCA.

4.1. Feature Selection Using the CPBF

For the feature selection using class profile-based approach, we have manually identified the most regular words that exist in each category and weighted them using entropy weighting scheme before adding them to the feature vectors that have been selected from the PCA. For example, the words that exist regularly in a boxing class are 'Lewis', 'Tyson', 'heavyweight', 'fighter', 'knockout', etc. Then a fixed number of regular words from each class will be used as a feature vectors together with the reduced principal components from the PCA. These feature vectors are then used as the input to the neural networks for classification. The entropy-weighting scheme on each term is calculated as $L_{jk} \times G_k$ where L_j is the local weighting of term k and G_k is the global weighting of term k. The L_{jk} and G_k are given by

$$L_{jk} = \begin{cases} 1 + \log(TF_{jk}) & (TF_{jk} > 0) \\ 0 & (TF_{jk} = 0) \end{cases} \tag{3}$$

and

$$G_k = \frac{1 + \sum_{k=1}^{n} \frac{TF_{jk}}{F_k} \log \frac{TF_{jk}}{F_k}}{\log n} \tag{4}$$

where n is the number of document in a collection and TF_{jk} is the term frequency of each word in Doc_j. The F_k is a frequency of term k in the entire document collection. We have selected $R = 50$ words that have the highest entropy value to be added to the first d components from the PCA to be an input to the neural networks for classification.

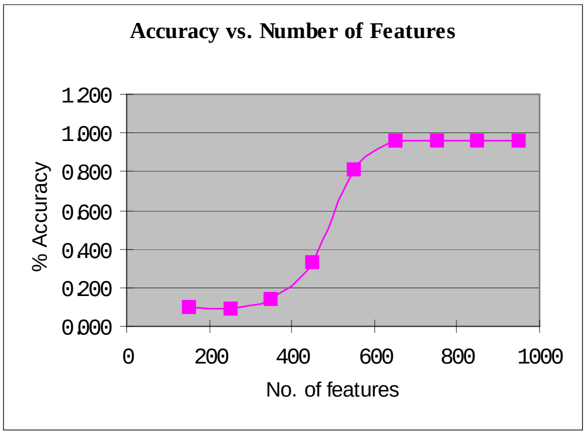

Fig. 3 The classification accuracy with a combination of the PCA and CPBF feature vectors.

4.2. Input Data to the Neural Networks

After the preprocessing of news web pages, a vocabulary that contains all the unique words in the news database has been created. We have limited the number of unique words in the vocabulary to 1,800 as the number of distinct words is big. Each of the word in the vocabulary represents one feature vector. Each feature vector contains the document-terms weight. The high dimensionality of feature vectors to be as an input to the neural networks is not practical due to poor scalability and performance. Therefore, the PCA has been used to reduce the original feature vectors $m = 1800$ into a small number of principal components. In our case, we have selected a few values of d (e.g., 100, 200, 300, 400, and 500, 600) together with $R = 50$ features selected from the CPBF approach because this parameter performs better for web news classification compared to other parameters to be input to the neural networks. The loading factor graph for the accumulated proportion of eigenvalues is shown in Fig. 2. The value of $d = 50$ contributes 81.6% of proportions from the original feature vectors. The neural network parameters for the NN-PCA and the WPCM methods are shown in Table 1. We have selected 600 features from the PCA and 50 features from CPBF methods. We have found that this combination is the best in getting the accuracy of the classification as shown in Fig. 3.

Table 1 The error back-propagation neural networks parameters that have been used in the experiments.

NN Parameters	Values
Learning rate (η)	0.005
Momentum rate (α)	0.001
Number of iteration (t)	1000
Means square error (MSE)	0.005

The parameters of the error-backpropagation neural networks are shown in Table 1. The numbers of inputs to the neural networks are 650 with 50 hidden layers and 12 output layers. The try and error approach has been used to select the appropriate value of hidden layers which indicate the highest value of document classification accuracy. The numbers of neural networks outputs are based on the number of classes exist in the sports news database as shown in Table 2.

4.3. The TF-IDF Method

The TF-IDF classifier is based on the relevant feedback algorithm using the vector space model [11]. The algorithm represents documents as vectors so that the documents with similar contents have similar vectors. Each component of a vector corresponds to a term in the document, typically a word. The weight of each component is calculated using the term frequency inverse document frequency-weighting scheme (TF-IDF) which tries to reward words that occur many times but in few documents. To classify a new document Doc', the cosines of the prototype vectors with corresponding document vectors are calculated for each class. Doc' is assigned to the class which its document vector has the highest cosine. Further description on the TF-IDF measure is described by Joachim [12].

4.4. The Bayesian Classifier

For statistical classification, we have used a standard Bayesian classifier. When using the Bayesian classifier, we have assumed that term's occurrence is independent of other terms. We want to find a class cs that gives the highest conditional probability given a document Doc'. $w_k^m = w_1, w_2, \ldots, w_m$ is the set of words representing the textual content of the document Doc' and k is the term number where $k = 1,2,\ldots,m$. The classification score is measured by

$$P(cs) = \prod_{k=1}^{m} P(w_k | cs) \qquad (5)$$

where $P(cs)$ is the prior probability of class cs, and $P(w_k | cs)$ is the likelihood of a word w_k in a class cs that is estimated on a labeled training document. A given web page document is then classified in a class that maximizes the classification score. If the scores for all the classes in the sports category are less than a given threshold, then the document is considered unclassified. The Rainbow toolkit has been used for the classification of the training and test news web pages using the Bayesian and TF-IDF methods [13].

Table 2 The number of documents that have been used for classification.

Class No.	Class Name	Number of documents
1	baseball	569
2	boxing	210
3	cycling	80
4	football	550
5	golf	456
6	hockey	856
7	motor-sports	405
8	rugby	52
9	skiing	233
10	soccer	261
11	swimming	169
12	tennis	255
	Total	**4096**

5. Experiments

We have used a web page dataset from the Yahoo sports news as shown in Table 2. The types of news in the database are baseball, boxing, cycling, football, golf, hockey, motor-sports, rugby, skiing, soccer, swimming and tennis. The total of documents are 4096. For the training, we have selected randomly 1000 documents from different classes. The rest of the documents are used as the test sets. For the PCA approach, we have selected 600 principal components and input them directly to the neural networks for classification. The PCA, TF-IDF, Bayesian, and WPCM classification methods are evaluated using the standard information retrieval measures that are precision, recall, and F1. They are defined as follows

$$precision = \frac{a}{(a+b)} \qquad (6)$$

$$recall = \frac{a}{(a+c)} \qquad (7)$$

$$F1 = \frac{2}{\left(\frac{1}{precision} + \frac{1}{recall}\right)} \qquad (8)$$

where the values of *a, b,* and *c* are defined in Table 3. The relationship between the system classification and the expert judgment is expressed using four values as shown in Table 4.

Table 3 The definitions of the parameters *a, b,* and *c* which are used in Table 4.

Value	Meaning
a	The system and the expert agree with assigned category.
b	The system disagrees with the assigned category but the expert did.
c	The expert disagrees with the assigned category but the system did.
d	The system and the expert disagree with the assigned category.

Table 4 The decision matrix for calculating the classification accuracies.

Expert	System	
	Yes	No
Yes	*a*	*b*
No	*c*	*d*

Table 5 The classification results using the PCA method.

Class No.	Precision (%)	Recall (%)	F1 (%)
1	89.82	100.00	94.64
2	84.75	100.00	91.74
3	71.60	72.50	72.05
4	97.06	99.00	98.02
5	100.00	97.00	98.48
6	90.83	99.00	94.74
7	97.09	100.00	98.52
8	100.00	100.00	100.00
9	99.01	100.00	99.01
10	71.26	67.00	67.07
11	82.62	95.00	92.23
12	69.44	75.00	72.12
Average	**87.79**	**92.04**	**89.89**

Table 6 The classification results using the WPCM method.

Class No.	Precision (%)	Recall (%)	F1 (%)
1	97.14	68.00	80.00
2	86.96	100.00	93.02
3	98.68	93.75	96.15
4	86.09	99.00	92.09
5	89.91	98.00	93.78
6	94.29	99.00	96.59
7	84.03	100.00	91.32
8	97.62	100.00	98.80
9	89.29	100.00	94.34
10	73.91	68.00	70.83
11	90.09	100.00	94.79
12	96.15	100.00	98.04
Average	**90.35**	**93.81**	**91.65**

Table 7 The classification results using the F1 measure for the TF-IDF and Bayesian methods.

Class No.	TF-IDF (%)	Bayesian (%)
1	83.04	83.04
2	88.89	90.48
3	100.00	100.00
4	79.39	78.79
5	86.86	81.75
6	68.03	86.89
7	56.46	59.49
8	50.00	75.00
9	91.30	89.86
10	88.46	80.77
11	64.00	56.00
12	80.26	90.79
Average	**78.06**	**81.07**

5.1. Results

The classification results using the PCA and the WPCM methods are shown in Table 5 and 6. The average for precision, recall, and F1 measures using the PCA approach are 87.79%, 92.04%, and 89.89% respectively. In comparison with the WPCM approach, the precision, recall, and F1 measures are 90.35%, 93.81%, and 91.65% respectively. Furthermore, we have compared the WPCM with the TF-IDF and Bayesian methods using the F1 measure. They are 78.06% and 81.07% for the TF-IDF and Bayesian methods as shown in Table 7. This indicates that if the feature vectors are selected carefully, the improvement of web sports news classification using the combination of the PCA and CPBF in the WPCM approach will increase the classification accuracy. The rugby and cycling classes (class number 8 and 3) contain a small number of documents in the dataset. The F1 measure using the PCA approach for the rugby and cycling classes are 100% and 73% respectively, as shown in Table 5. But for the WPCM approach the classification accuracies for both classes are 98.80% and 96.15% respectively. These indicate that the WPCM approach has been able to classify the documents correctly although the number of documents representing the class is small. Also, we have overcome the limitation of the PCA in supervised data where the characteristic variables that describe smaller classes tend to be lost as a result of the dimensionality reduction by using the WPCM approach. Furthermore, the classification accuracy on the small classes can be improved although they have been reduced into a small number of principal components.

6. Conclusions

We have presented a new approach of web news classification using the WPCM. The feature selection for the WPCM is based on the combination of the PCA and CPBF. The comparison of classification accuracy using the PCA, WPCM, TFIDF, and Bayesian approaches have been presented in this paper. The WPCM approach has been applied to classify the Yahoo sports news web pages. The experimental evaluation with different classification algorithms demonstrates that this method provides acceptable classification accuracy with the sports news datasets. Also, we have overcome the limitation of the PCA in supervised data where the characteristic variables that describe smaller classes tend to be lost as a result of the dimensionality reduction by using the WPCM approach. Furthermore, the classification accuracy on the small classes can be improved although they have been reduced into a small number of principal components. Although the classification accuracy using the WPCM approach is high in comparison with other approaches, the time taken for training is relatively long compared with other methods.

7. References

[1] S. Wermeter, Neural network agents for learning semantic text classification, Information Retrieval, Vol. 3. No. 2, p. 87-103, 2000.

[2] S. L.Y. Lam and D. L. Lee, Feature reduction for neural network based text categorization, Proceedings of the 6th International Conference on Database Systems for Advanced Applications 19 - 22 April, Hsinchu, Taiwan, 1999.

[3] F. Sebastini, Machine learning in automated text categorization, ACM Computing Surveys, 34(1), 2002.

[4] G. Karypis and E.H. Sam, Concept indexing: a fast dimensionality reduction algorithm with applications to document retrieval & categorization, CIKM 2000 (2000).

[5] S. T. Dumais, Improving the retrieval of information from external sources, Behavior Research Methods, Instruments and Computers, 23(2), 229-236, 1991.

[6] Yahoo Web Pages, (http://www.yahoo.com), 2001.

[7] H. Mase and H. Tsuji, Experiments on automatic web page categorization for information retrieval system, Journal of Information Processing, IPSJ Journal, Feb. 2001, pg. 334-347, 2001.

[8] R. Calvo, M. Partridge, and M. Jabri, A comparative study of principal components analysis techniques, In Proc. Ninth Australian Conf. on Neural Networks, Brisbane, QLD., pp. 276-281, 1998.

[9] A. Selamat, S. Omatu, H. Yanagimoto, Information retrieval from the internet using mobile agent search system, International Conference of Information Technology and Multimedia 2001 (ICIMU), University Tenaga Nasional (UNITEN), Malaysia, August 13-15, 2001.

[10] Salton & McGill, Introduction to modern information retrieval, New York, McGraw-Hill, USA, 1983.

[11] S. Jones, Karen, and Peter Willet, Readings in information retrieval, San Francisco: Morgan Kaufmann, USA, 1997.

[12] T. Joachims, Probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, Proceedings of International Conference on Machine Learning (ICML), 1997.

[13] A. K. McCallum, Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, (http://www.cs.cmu.edu/~mccallum/bow), 1996.