New Method Of Feature Selection For Persian Text Mining Based On Evolutionary Algorithms

Akram Roshdi
Department of Computer, Islamic Azad University, Khoy Branch,Iran
Akram.roshdi5@gmail.com

Abstract

Today, with the increasingly growing volume of text information, text classification methods seem to be essential. Also, increase in the volume of Persian text resources adds to the importance of this issue. However, classification works which have been especially done in Persian are not still as extensive as those of Latin, Chinese, etc. In this paper, a system for Persian text classification is presented. This system is able to improve the standards of accuracy, retrieval and total efficiency. To achieve this goal, in this system, after texts preprocessing and feature extraction, a new improved method of feature selection based on Particle Swarm Optimization algorithm (PSO) is innovated for reducing dimension of feature vector. Eventually, the classification methods are applied in the reduced feature vector. To evaluate feature selection methods in the proposed classification system, classifiers of support vector machine (SVM), Naive Bayes, K nearest neighbor (KNN) and Decision Tree are employed. Results of the tests obtained from the implementation of the proposed system on a set of Hamshahri texts indicated its improved precision, recall, and overall efficiency. Also, SVM classification method had better performance in this paper.

Keywords: *Feature vector, classification, support vector machines, Feature Extraction, Dimensions Reduction.*

1. Introduction

In text classification, before using any method, it is important to convert texts into a suitable display form. After such a conversion, feature selection algorithms are applied and texts are classified using the selected features. In this study, an improved model based on Particle Swarm Optimization algorithm is used to select the text feature. In fact, in this study, the mutation operator (based on genetic algorithm) is employed to improve the speed and accuracy of convergence of Particle Swarm Optimization algorithm. Methods and strategies used in the course of this investigation are given in the following sections.

2. Previous works

Due to complex structural problems of Persian language, fewer studies have been undertaken in the field of text mining in Persian than other languages. Classifying Persian texts is one of the areas affected by these problems and few works have been done in this regard so that, compared to the previous works in other languages, this issue seems to be a rather new research area. In the field of text classification, previous studies are, essentially, based on two-class classification technique. Some training methods in accordance to two-class strategy are Naive Bayes, K nearest neighbor (KNN), support vector machine (SVM), Decision Tree, and etc. Most researches on text classification, were designed, carried out and tested on English articles. A number of training techniques for text classification have been also implemented on other European languages such as German, Italian and Spanish. Some other techniques have been conducted on Asian languages such as Chinese and Japanese. In confrontation with complex structures and large data volume, statistical methods do not work. Since this issue has a number of features, some feature selection methods must be used which can reduce the number of features. Therefore, feature selection is one of the most important steps in text classification. For feature selection, several methods exist. Recently, researchers pay much attention to the evolutionary algorithms in the feature selection. Evolutionary algorithms can be used for image processing tasks. For example, genetic algorithm is used for face recognition system. Naive Bayes and K nearest neighbor (KNN), along with genetic algorithm, are used for to remove difficult-to-learn data. Three combinatory methods have been investigated on five Standard English data. The results indicate the non-relevant data elimination. In all these papers, it has been shown that the efficiency in evolutionary algorithms for feature selection is more than traditional statistical methods. Persian language due to its complex structures, in accordance to other languages, has few studies. Classification of Persian texts is among the fields which are affected by this problem. A list of studies on the Persian language is presented in [1, 2]. An automated technique for the detection of stop words in Persian language is presented [3]. In the paper [4], a new method for root detection with a bottom-up strategy is presented for Persian language. Test results show that this method has good flexibility. In Persian documents, statistical information of documents is used to retrieve and rank the documents. The efficiency of various information retrieval models, such as vector space model, possible space model, the language model [5], fuzzy model [5], is evaluated on a set of Persian documents.

3. Classifying text documents

Text classification includes several steps: preprocessing phase prepares documents for classification process that is similar to many natural language processing issues, in which irrelevant tags and words are removed. Indexing phase uses a weighting scheme to weigh statements in the text. The following classification step is selecting appropriate feature space among the statements in documents, which is a vital step; also, system precision largely depends on selection keys that show the document.

In the next classification step, documents are divided into training and testing parts; the former is used to train the system to recognize different patterns of classes and the latter is used for system evaluation. Classification process depends upon the applied algorithms [6]. Figure 1 shows a general overview of classification steps.

classification steps.

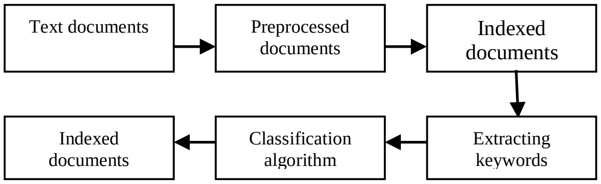

Fig.1 An overview of classification steps

4. Feature extraction methods

To apply text classification and feature extraction methods, an appropriate structure must be selected for displaying documents. The simplest and most conventional document display method is creating a feature space using all the words of a document. In this space of features, after removing special words and sometimes doing the etymology, a list of all words is made in the text and every document is displayed in different ways based on the words existing on the list and the weight of these words in the document. Table 1 shows text document display method in the form of vectors [7]. As seen in this table, the set { F1,…,Fn} indicates space of features and F_i shows a word which has been used at least once in the text. Moreover, the set { D1,…,Ds } shows a set of documents displayed in this vector display method. Each Wij value shows the weight of word Fj in document Di. In fact, there are different criteria for word weighing.

Table 1: Display method of vectors of text documents

words space				
F_N	-	F_2	F_1	
W_{1N}	-	W_{12}	W_{11}	D_1
W_{2N}	-	W_{22}	W_{21}	D_2
-	-	-	-	-
W_{SN}	-	W_{S2}	W_{S1}	D_S

For example, TF weighting scheme studies the number of feature repetition in a document, while TF-IDF studies feature repetition rate in other documents to determine its rate in the document. TFC also involves document length in its feature weight. All of these criteria try to determine the information importance of each feature in each document so that the classifier can classify documents in different classes based on the similarities of these scores. Texts cannot be directly evaluated by classifiers and are thus converted into appropriate display forms. As shown in Equation 1, text dj is usually displayed as a vector of word weight [8]:

$$d_j = \{ w_{1j}, w_{2j}, \ldots, w_{|T|j} \} \qquad (1)$$

In Equation 1, T is a set of words (features) that is repeated at least once in at least one text of the training set and 0≤wij≤1 indicates the weight allocated to the i^{th} word in the j^{th} text. Word weights are obtained using TFIDF normalized function. This method was first proposed in data recovery. Then, they were used in document classification for weighting features [9].

TFIDF method is one of the most conventional feature weighting methods which is obtained from the combination of TF- and IDF-based methods as follows [9].

$$wkj = \frac{tfidf(t_k d_j)}{\sqrt{\sum_{s=1}^{|T|}\left(tfidf(t_s d_j)\right)^2}} \qquad (2)$$

where

$$tfidf(t_k, d_j) = \#(t_k, d_j).Log\frac{|Tr|}{\#Tr(t_k)} \qquad (3)$$

In Equation 3, $|T_r|$ is the number of texts in the training set; the number of times word t_k is repeated in text d_j is # (t_k, d_j). T_r is training set and $|T_r|$ is its length. Also, # Tr (t_k) is the number of training set texts, in which word t_k occurs [8].

5. Evaluation criteria

In general, feature selection methods are applied as a step before the learning of classifiers. Effect of feature selection methods and document display methods is obtained using efficiency measurement of each of the classifiers. In order to measure efficiency, standard definitions of precision (P), recall (R), and Fβ function are used[10].

$$\mathrm{P} = \frac{\text{number of correctly found classes}}{\text{total number of found classes}} \qquad (4)$$

$$\mathrm{R} = \frac{\text{number of correctly found classes}}{\text{total number of correct classes}} \qquad (5)$$

$$F\beta = \frac{(\beta^2+1)*\mathrm{P}*\mathrm{R}}{\beta^2*\mathrm{P}+\mathrm{R}} \qquad (6)$$

P shows precision of classifications and R shows completion rate of the found set. High value of both of these factors indicates a high level of classification method; however, higher precision value is usually accompanied by reduced recall rate and high recall rate usually involves reduced precision. Depending on the application, sometimes high precision is important and sometimes higher recall is more optimal. Therefore, to attribute different weights to precision and recall, Fβ criterion can be used; considering the application and importance of each of these two factors (precision and recall), different weights can be assigned to them. In many academic researches, F1 criterion is used that gives equal weights to both of these factors. In this paper, this criterion was used for various evaluations.

6. Materials and methods

In this article, second version of Hamshahri's standard dataset was used. This body of text included more than 318 pieces of news between 1996 and 2007. The proposed algorithm is an improved model based on Particle Swarm Optimization algorithm for feature selection in text. In fact, in this paper, to improve the convergence speed and accuracy of Particle Swarm Optimization algorithm, the mutation operator (in accordance to genetic algorithms) is used. The proposed method has been named PM (Proposed Method). The proposed method has the appropriate global search (moving particles using particle swarm algorithm relations) and local search (due to the mutation applied on particles) at the same time. By adding the mutation operator to the Particle Swarm Optimization algorithm, the likelihood of escaping from local minimum increased greatly and this leads to likelihood enhancement of achieving an optimal solution. For increased accuracy, retrieval and overall efficiency of classification algorithms, several pre-processing methods such as document indexing and extra words deletion are used in training course construction. For feature selection, three methods of genetic algorithm and particle swarm optimization algorithm and the proposed method are used. Finally, 4 methods of classification as SVM, Naive Bayes, KNN and Decision Tree are used for document classification. Also, for preprocessing and classification stages, C# and MATLAB software were used, respectively. Afterward, precision, recall, and overall efficiency of classification algorithms were evaluated using the testing set.
Figure 2 shows the framework of the proposed classification system.

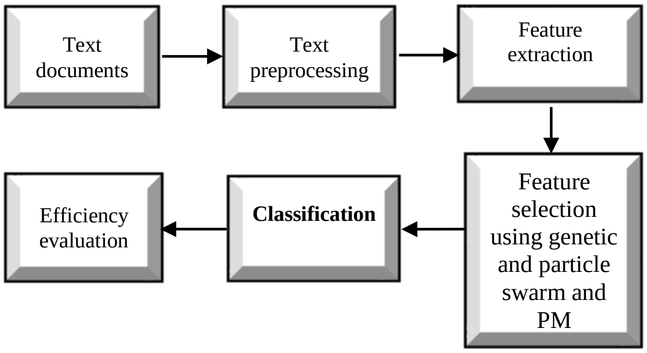

Fig.2 Framework of the proposed classification system

7. Results and discussion

Table 2, 3, 4 and 5, respectively, the comparison between efficiency, precision, retrieval and overall efficiency and SVM, Naive Bayes, KNN and Decision Tree are given.

Table2: Comparing efficiency (precision, recall, and F1 measurement) using SVM

Percentage of Features (%)	GA			PSO			PM		
	P	R	F	P	R	F	P	R	F
5	76.54	76.54	76.54	76.40	74.68	75.53	77.75	76.58	77.16
10	77.24	75.95	76.59	75.25	74.05	74.65	78.83	76.58	77.69
15	77.92	77.85	77.88	78.18	77.22	77.69	79.11	79.28	79.19
20	78.49	76.58	77.52	78.42	77.85	78.13	80.01	80.38	80.19
25	76.54	75.95	76.24	79.09	77.85	78.46	80.92	79.75	80.33
30	79.50	74.68	77.01	82.13	81.65	81.89	82.28	82.28	82.28
35	81.44	80.38	80.91	81.21	81.01	81.11	83.83	83.54	83.68
40	81.26	79.75	80.50	83.72	83.54	83.63	84.63	82.28	83.44
45	81.02	79.75	80.38	83.77	83.54	83.66	84.01	84.38	84.19

Table3: Comparing (precision, recall, and F1 measurement) using naive Bayesian

Percentage of Features (%)	GA			PSO			PM		
	P	R	F	P	R	F	P	R	F
5	72.14	70.25	71.19	70.73	70.26	70.49	69.74	69.62	69.68
10	72.40	67.72	69.98	73.03	70.25	71.61	73.79	71.52	72.64
15	75.65	71.52	73.53	73.18	71.52	72.34	75.88	74.05	74.95
20	72.04	68.35	70.15	73.73	72.78	73.25	74.47	72.78	73.62
25	74.55	72.78	73.66	74.75	72.15	73.43	75.51	71.52	73.98
30	75.04	69.62	72.23	74.83	72.15	73.47	77.62	73.95	75.74
35	76.17	71.52	73.77	76.32	72.15	74.18	78.98	75.29	77.09
40	77.51	73.42	75.41	76.21	72.78	74.45	77.99	76.01	76.99
45	76.34	68.99	72.48	75.17	71.52	73.77	77.52	74.60	76.03

Table4: Comparing efficiency (precision, recall, and F1 measurement) using KNN

Percentage of Features (%)	GA			PSO			PM		
	P	R	F	P	R	F	P	R	F
5	70.69	69.62	70.15	75.64	74.05	74.84	77.79	75.95	76.86
10	75.69	72.15	73.88	76.50	75.32	75.90	78.83	76.58	77.69
15	76.67	75.32	75.99	78.32	77.22	77.76	79.37	78.48	78.92
20	75.64	74.05	74.84	80.03	79.11	79.57	79.99	77.85	78.90
25	78.33	77.22	77.77	79.04	77.22	78.12	80.92	79.75	80.33
30	79.86	77.85	78.84	80.43	79.11	79.76	82.28	82.64	82.28
35	81.25	79.78	80.49	81.13	80.38	80.76	82.60	81.75	82.17
40	81.97	79.75	80.84	81.92	80.38	81.14	83.65	83.11	83.38
45	79.80	79.11	79.46	81.01	81.84	82.10	82.94	82.28	82.61

Table5: Comparing (precision, recall, and F1 measurement) using decision tree classifier

Percentage of Features (%)	GA			PSO			PM		
	P	R	F	P	R	F	P	R	F
5	69.45	68.99	69.22	74.19	72.78	73.48	75.98	74.05	74.51
10	72.18	67.72	69.88	76.77	74.68	75.71	77.02	75.95	76.48
15	73.63	66.46	69.86	77.77	74.05	75.86	78.41	75.32	76.83
20	74.55	72.78	73.66	75.10	73.42	74.25	76.56	75.95	76.26
25	75.82	74.68	75.25	78.23	75.32	76.74	78.41	75.32	76.83
30	75.13	71.52	73.28	78.83	76.58	77.69	79.01	78.48	78.74
35	76.71	76.58	76.65	79.62	74.05	76.74	80.03	79.11	79.57
40	74.94	74.68	74.81	78.33	77.22	77.77	78.46	73.42	75.85
45	74.12	74.12	74.12	78.08	72.15	75.00	78.60	74.68	76.59

According to the obtained results, it can be seen that SVM classifier had better performance than all other classifiers, as indicated by the tables given in the previous section. After SVM, KNN classifier had better performance than the two others. Figure 3 also shows the same issue.

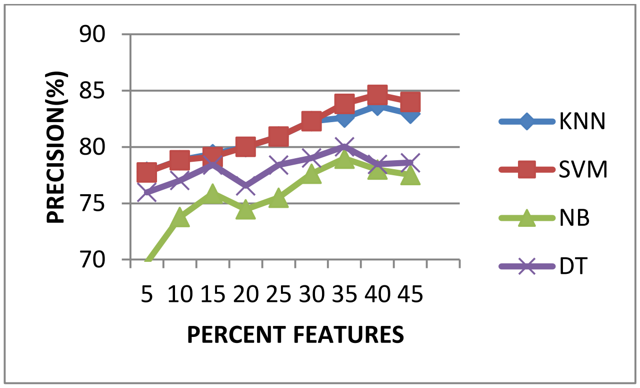

Fig.3 Comparing precision SVM ,KNN,DT and NB With proposed method

8. Conclusions

In this paper, a brief investigation was done about feature extraction methods and types of classification methods. All experiments were conducted on Persian documents of Hamshahri (citizen) standard data. Experimental results show that the proposed method has a good performance and the precision, retrieval and efficiency of support vector machine classifier enjoy better function.

In future work, to improve the precision, retrieval and overall efficiency, we intend to generalize the proposed method for text categorization with more classes (for example 10 or 20 classes) using more features extraction. This is certainly a great step in increasing the accuracy of the researches in the field of information retrieval in Persian language. In the following, by examining other evolutionary algorithms and comparing them with the proposed algorithm for feature selection and providing a combinatory method for feature selection, we intend to increase the accuracy and efficiency of classification algorithms.

References

[1]. M .Aci, And , M . Avci,” A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm”,Elsevier, 2010,vol. 37, p.5061–5067.

[2]. M. shamsfard ,”processing Persian text: past finding and future challenges”, Tehran universitypress, 2007.

[3]. A.yoosofan and M. zolghadri,”an automatic method for stopword recognition in Persian language”, amirkabir university press, 2005.

[4]. M.Aljaly and O.frieder,” improving the retrieval effectiveness via light stemming approach”, journal of information science,2004,vol. 158, pp. 69-88.

[5]. A.Unler and A. MuratA,” maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification”,Elsevier,2011, No. 181, P. 4625–4641.

[6] M.saleh,” list of dissertations on Persian language and computers”, Tehran university press, 2007.

[7] M. shamsfard ,”processing Persian text: past finding and future challenges”, Tehran university press,2007.

[8]. M. Litvakand and. M. last ,”classification of web documents using concept extraction from ontologies”,Proceedings of the 2nd international conference on Autonomous intelligent systems: agents and data mining, Russia, , 2007, pp. 287-292.

[9]. V. Gupta, and S. lehal,” a survey of text mining technique and applications”,journal of emerging technologies in web intelligence,2009,vol.1, no.1.

[10]. A. Sharma , Sh. Dey,” Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis”, International Journal of Computer Applications on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012.

Akram Roshdi received his B.Sc. in computer engineering from Shabestar University of Applied Science And Technology, Tabriz, Iran, in 2003, and his M.Sc. in Computer Engineering from Islamic Azad University, Shabestar branch, Tabriz, Iran, in 2013. She is Currently a PhD student at Department of Engineering, Qom branch, Islamic Azad University. His research interests include wireless networks, cloud computing and Data mining.