Feature Selection and Reduction for Persian Text Classification

Zahra Robati
Department of Computer Engineering and Information Technology,
Shahrood, Iran

Morteza Zahedi
Department of Computer Engineering and Information Technology,
Shahrood, Iran

Najmeh Fayazi Far
Department of Computer Engineering and Information Technology,
Shahrood, Iran

ABSTRACT
With the rapid growth of the World Wide Web and increasing availability of electronic documents, the automatic text classification became a general and important machine learning problem in text mining domain. In text classification, feature selection is used for reducing the size of feature vector and for improving the performance of classifier. This paper improved Dominance which is a feature selection criterion and proposed Extended Dominance (E-Dominance) as a new criterion. E-Dominance is compared favorably with usual feature selection methods based on document frequency (DF), information gain (IG), Entropy, $\chi 2$ and Dominance on a collection of XML documents from Hamshahri2 which is a commonly used in Persian text classification. The comparative study confirms the effectiveness of proposed feature selection criterion derived from the Dominance.

General Terms
Text Classification, Feature Selection, Feature Reduction.

Keywords
Text Classification, E-Dominance, feature selection criterion

1. INTRODUCTION
Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the electronic documents. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification (supervised, unsupervised and semi supervised) and summarization. Text classification is a supervised task for which, given a set of categories, a training set of pre-classified documents is provided. Given this training set, the task consists in learning the class descriptions in order to be able to classify a new document in one of the categories [1-3]. However how these documented can be properly annotated, presented and classified. So it consists of several challenges, like proper annotation to the documents, appropriate document representation, dimensionality reduction to handle algorithmic issues [4,5] and an appropriate classifier function to obtain good generalization and avoid over-fitting. Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents are important for the research communities. The common feature selection criteria proposed in the literature [4-7] such as document frequency (*DF*), Entropy, information gain (*IG*), $\chi 2$ and Dominance consider the distribution of the documents containing the term between categories but the ratio of documents which contain the term on a certain class to all documents available on the same class is not taken to account. However, it should be considered that a term which is characteristic of a category must appear in majority of documents belonging to that category than other categories. For this reason, this article proposed the E-Dominance criterion. E-Dominance criterion is also compared with usual feature selection methods mentioned above on a collection of XML documents. Indeed, this method shows removing up to 94% of terms and can improve the classification accuracy measured by Recall. Experiments have been performed on Hamshahri2 database which is composed of news articles in Persian language.

The rest of this paper is organized as follows. Most commonly used feature selection criteria are presented in section 2. In section 3 methodology (feature representation, E-Dominance, classification) is described. Finally, experiments and achieved results are represented in section 5 and conclusion and future works are given in section6.

2. COMMONLY FEATURE SELECTION CRITERIA
In 4.2, proposed criterion compare with other usual feature selection methods for text categorization. These methods are document frequency (*DF*) is used in [10], Entropy [11], information gain (*IG*) is defined by [5], chi square ($\chi 2$) is used in [6], Dominance in [9]. These criteria consider the distribution of the documents containing the term between categories but the ratio of documents which contain the term on a certain class to all documents available on the same class is not taken to account. Table 1 shows these criteria with their formula (1-5). By using DF, only the terms that appear in a number of documents higher to a defined threshold, are selected. This threshold can be determined using a training set. With IG only the words for which the value of the criterion is the most important are considered as characteristics for a category. $\chi 2$ equals 0 when tj and ck are independent. On the contrary, tj is considered as a characteristic feature for ck if the value of $\chi 2(tj , ck)$ is high. The Entropy is minimal, equals 0, if the term tj appears only in one category. So this term might have a good discriminatory power in the categorization task. In spite of Dominance is maximal, equals 1, if the term tj appears only in one category and so this term might be a good discriminant term. These criteria are introduced in details in [11].

Table 1- Most commonly used feature selection criteria

$\mathrm{DF}(f_i , c_j) = \mathrm{df}(f_i , c_j)$	(1)
$Entropy(t_i) = -\sum_{k=1}^{r} (tf_i^k) * \log_2 (tf_i^k)$, $tf_i^k = \frac{n_i^k}{\sum_{k=1}^{m} n_i^k}$	(2)

$$IG(t_i, c_k) = d(t_i, c_k) \log\left(\frac{d(t_i, c_k)}{d(t_i)d(c_k)}\right) + d(t'_i, c_k) \log\left(\frac{d(t'_i, c_k)}{d(t'_i)d(c_k)}\right) \quad (3)$$

$$\chi^2(t_i, c_k) = \frac{N.[d(t_i, c_k).d(t'_i, c'_k) - d(t'_i, c_k).d(t_i, c'_k)]}{d(t_i).d(t'_i) - d(c_k).d(c'_k)}) \quad (4)$$

$$Dominance(t_i, c_k) = \frac{d(t_i, c_k)}{d(t_i, c_k) + d(t_i, c'_k)}) \quad (5)$$

3. METHODOLOGY

3.1 Textual Document Representation

In text categorization, the vector space model (VSM) introduced by Salton *et al.* [7] is widely used as well for flat documents as for semi structured documents written in markup languages like HTML or XML. In this model, documents are represented as vectors which contain term weights. Given a collection D of documents, an index $T = \{t_1, t_2, ..., t_{|T|}\}$, where $|T|$ denotes the cardinal of T , gives the list of terms (or features) encountered in the documents of D. A document d_i of D is represented by a vector $\sim d_i = (w_{i,1}, w_{i,2}, ..., w_{i,|T|})$ where $w_{i,j}$ represents the weight of the term t_j in the document d_i. In order to calculate this weight, the TF.IDF formula can be used [6]:

$$Wi,j = TFIDF(t_i, d_j) = tf(t_i, d_j) * log\left(\frac{|D|}{|D(t_i)|}\right) \quad (6)$$

Where $tf_{i,j}$ is the number of occurrences of t_i in document dj, $|D|$ is the total number of documents in the corpus and $|\{dj : ti \in dj\}|$ is the number of documents in which the term ti occurs at least one time. In this article, we used VSM representation model and TFIDF method for weighing features.

In text classification domain, dimension of features is essential and effective problem. Even for limited collections, the dimension of the index can be exceedingly large. For example, in our experiments, total of number of features without reduction is 16893 unique words and the all terms belonging to this bag of words are not necessary and discriminant features. For this reason, non-useful words must be removed, in order to extract a subset T ′from T more suited for the categorization task. For that purpose, the local approach consists in filtering a specific subset for each category in such a way that the indexes used to represent documents belonging to different categories are not the same, while the global approach, adopted in this work, uses the same subset T ′extracted from T to represent all the documents of the collection [2,8]. This article introduces the E-Dominance criterion in order to select a subset T ′from T , providing a more efficient description of the documents.

3.2 Extended Dominance (E-Dominance) Criterion for Feature Selection

Formally, let $F=\{f_1, f_2, f_3, \ldots, f_M\}$ be the set of features associated with a collection; $C=\{c_1, c_2, \ldots, c_n\}$ be the set of categories that occur in a collection; $df(f_i, c_j)$ be the number of training documents associated with class c_j that contain f_i. Dominance was defined as follows [7]:

$$Dominance(f_i, c_j) = \frac{df(f_i, c_j)}{\sum_{k=1}^{n} df(f_i, c_k)}) \quad (7)$$

The smaller the number of distinct classes where a feature occurs, the higher the dominance. With this criterion, if a feature occurs only in one class, Dominance of this feature for this class is 1 and this feature will be discriminative. Using a threshold for dominance, we can select discriminative features. Although a word that occurs only in one or two documents, is discriminative feature, but probability of occurrence of this word in test documents will be low, too. Now, consider features with low probability of occurrence in test documents, will be selected as discriminant features. Although these features are discriminative, but a large number of them are not necessary and only increase size of feature vector. An efficient discriminative feature is a feature that occurs only in one class and in majority of documents of the same class. Then we add this property to dominance criterion with comprehensiveness factor of dominance, α_{ij}, that is defined as follows:

$$\alpha_{ij} = \frac{df(f_i, c_j)}{df(c_j)} \quad (8)$$

This factor indicates the ratio number of documents in class c_j that includes feature f_i, $df(f_i,c_j)$, to number of documents in class c_j, $df(c_j)$, and Extended Dominance (E-Dominance) is defined as follows:

$$E - Dominance(f_i, c_j) = \alpha_{ij} * Dominance(f_i, c_j) \quad (9)$$

If α_{ij} is equal to 1, then E-Dominance will be equal to real dominance. When α_{ij} is 1, feature f_i occurs in all of documents in class c_j then if this feature occurs only in this class, dominance will be 1, too. Now α_{ij} and dominance are 1, then E-Dominance will be 1. This mean the feature that has this value for E-Dominance, is discriminant and high frequency feature in this class and are selected for feature vector in feature selection step and by selection a threshold for E-Dominance, high performance features are selected.

3.3 Classification (KNN)

From last few years , the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Latent Semantic Analysis, Rocchio's Algorithm, Fuzzy Correlation and Genetic Algorithms etc. Normally supervised learning techniques are used for automatic text classification, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents. A commonly techniques, KNN is described below. We use this classifier in our experiments too.

The k-nearest neighbor algorithm (k-NN) [12] is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents. This method is an instant-based learning algorithm that categorized objects based on closest feature space in the training set [12]. The training sets are mapped into multi-dimensional feature space. The feature space is partitioned into regions based on the category of the training set. A point in the feature space is assigned to a particular category if it is the most frequent category among the k nearest training data. Usually Euclidean Distance is typically used in computing the distance between the vectors. The key element of this method is the availability of a similarity measure for identifying neighbors of a particular document [13]. The training phase consists only of storing the feature vectors and categories of the training set. In the classification phase, distances from the new vector, representing an input document, to all stored vectors are computed and k closest samples are selected. The annotated category of a document is predicted based on the nearest point which has been assigned to a particular category.

$$C = argmax_i \sum_{j=1}^{k} \text{sim}(D_j|\text{D}) * \delta(C(D_j), i) \quad (10)$$

Calculate similarity between test document and each neighbor and assign test document to the class which contains most of the neighbors (Figure 1).

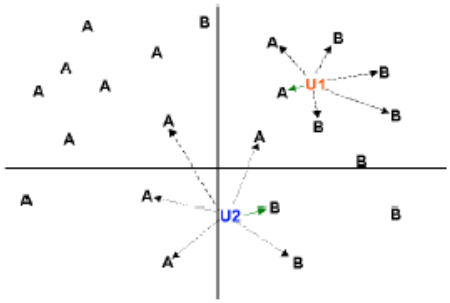

Figure 1: k-Nearest Neighbor

This method is effective, non-parametric and easy to implement. The k-nearest neighbor classification method is outstanding with its simplicity and is widely used techniques for text classification. This method performs well even in handling the classification tasks with multi-categorized documents. The major drawback of this method is it uses all features in distance computation, and causes the method computationally. Intensive, especially when the size of training set grows. Besides, the accuracy of k-nearest neighbor classification is severely degraded by the presence of noisy or irrelevant features.

In this paper we used BOW[1] model for feature vector representation create a n×m matrix, which n and m are number of features and number of documents, respectively. For preprocessing of texts, we used stop words removal for eliminate unnecessary features from feature vector. Selection of high frequency occurrence words, which are called stop words as discriminant features increases feature vector size and classification time. So stop word removal is essential step for text classification. In our experiments, each cell of feature vector indicates weight of feature in every document. Weights of features are calculated with TFIDF method. After stop words removal, 16893 are selected as initial features. By using a suitable threshold for TFIDF, numbers of initial features are reduced to 16430 words. In spite of applied feature reduction, feature vector size is still too large to classify text documents. So these features are reduced to 1041 by E-Dominance criterion. Then these features are classified using KNN classifier. Parameter K in KNN (k=7) was empirically selected so as to maximize the classification accuracy.

4. EXPERIMENTS

4.1 Database

This paper used a Persian text dataset, Hamshahri2, in experiments and selected randomly 500 documents for training and 250 documents for test. Table 2 shows the classes and number of selected documents from this dataset for train and test.

[1] Bag Of Word

Table 2- Classes and number of documents in experiments

Classes	Train Text	Test Text
Economic	100	50
Politic	100	50
Sport	100	50
Art	100	50
Scientific	100	50
Total	500	250

4.2 Results

To introduce a measure of accuracy and other performance measures, it should be noted that four predictions could be with the assumption of having a set of two classes of yes and no (See Table 3). True Positive (TP) and True Negative (TN) are correct classifications. False Positive (FP) happens when a sample which is truly negative is predicted as positive. Also, False Negative (FN) happens when a positive sample is predicted as negative. Therefore, the accuracy or overall success rate is the proportion of true results (both TP and TN) in the population (11). Also for the evaluation of learner, other parameters are used such as precision, recall, and F-Measure [14]. Recall parameter shows what proportion of positive classes the learner predicts correctly (12). Precision or positive predictive value is defined as the proportion of the true positives to all the positive results (both true positives and false positives) (13). F1-Measure considers both the precision and the recall of the test and is harmonic mean of them (14). Each of these measures will be calculate for each class separately and then mean of values for all classes will be reported as final value for related measures.

Table3 - Different scenarios for a two-class prediction

		Predicted classes	
		Yes	No
Actual class	Yes	TP	FN
	No	FP	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$F1 - measure = 2 * \frac{PR * R}{PR + R} \quad (14)$$

Best accuracy is achieved using E-Dominance criterion (proposed method) with threshold value 0.04. The number of selected features is 1041 from 16430 of initial features. Table 4 illustrates number of selected features from each class separately and ratio of selected features of each class to total

number of selected features. Total of number of selected features is 1041, but in table 4 is 1079, this difference is effect of overlapping of different classes. Features are selected from one or more classes.

Table4 - Number of selected features from each class

Classes	Number of selected features (A)	$\frac{A}{Total}$
Economic	280	27%
Politic	223	21%
Sport	213	20%
Art	147	14%
Scientific	216	21%
Total	1079	-

Figure 2 shows effect of E-dominance's threshold on the number of selected features. Lower thresholds select more features and best threshold is achieved 0.04 for best categorization.

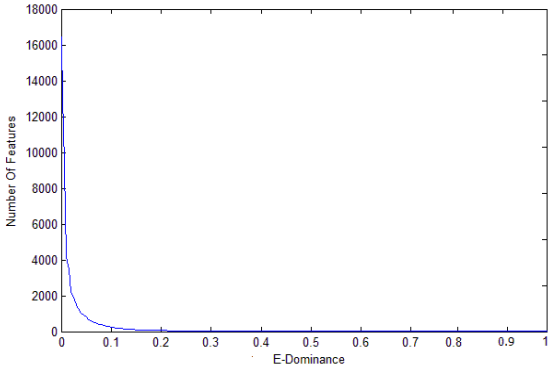

Figure 2: Effect of E-dominance's threshold on the number of selected features

And Figure 3 shows effect of the number of selected features on accuracy. With 1041 selected features, recall as accuracy is achieved 91.2% for proposed method. Without using this filtering for 16430 features, recall is 82%.

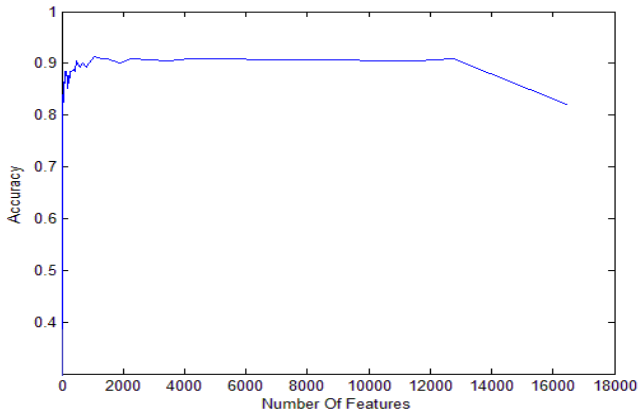

Figure 3: Effect of number of selected features on accuracy

Also figure 4 illustrates the effect of E-dominance's threshold on accuracy. With threshold value equals to 0.04, Recall is 91.2% for proposed method and for E-dominance higher than 0.7 is zero because number of selected features is zero.

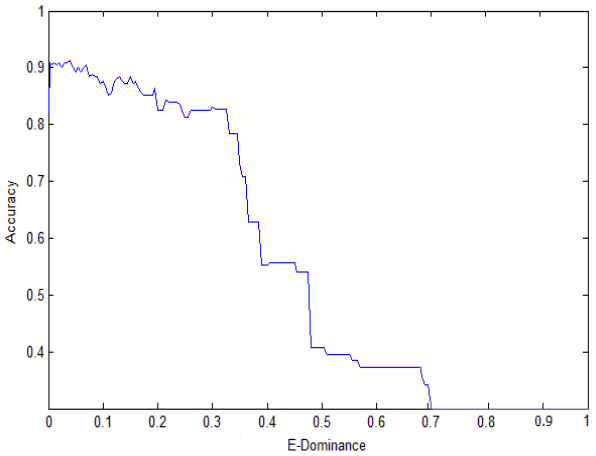

Figure 4: Effect of E-dominance's threshold on accuracy

Table 5 shows the value of performance measures for each class and mean of classes. The mean of recall and the mean of F1 are 0.912.

Table5 – Result of classification with performance measures

Classes	Recall	Precision	F1	Accuracy
Economic	0.94	0.87	0.904	0.96
Politic	0.94	0.887	0.913	0.964
Sport	0.90	0.882	0.891	0.956
Art	0.92	0.939	0.929	0.972
Scientific	0.86	1	0.925	0.972
Mean	0.912	0.916	0.912	0.965

And finally the table 6 illustrates the other feature reduction methods are compared with proposed method. As it can be shown in table 6 Dominance feature is the best feature using recall measure. But its features dimension is 9866 and reduces only 40% of features. Proposed criterion, E-Dominance has best feature reduction rate and reduces 93.66% of features. DF and $\chi 2$ have next ranks in feature reduction's rate. So proposed method, E-Dominance, has the best performance between these criteria.

Table 6 – Result of comparison of proposed criteria and other criteria in text classification

Measures	Recall	Number of selected features	Features reduction's rate
DF	0.90	2133	87.01%
Entropy	0.888	14586	0.99%

Info-Gain	0.90	8267	49.68%
$\chi 2$	0.90	2948	81.84%
Dominance	0.916	9866	39.95%
E-Dominance	0.912	1041	93.66%

5. CONCLUSIONS

The purpose of this study is to present a criterion which is able to reduce initial features into high performance features. The proposed method could recognize the class of documents with 91.2% accuracy by using E-Dominance criterion for feature selection. The proposed criterion selects features which occur in fewer classes and majority of documents in these classes. This criterion will be able to reduce features with 93.6% feature reduction rate. In comparison with other criteria, this criterion has more reduction and best accuracy. In this paper, the single word features is used as a feature vector. In future work, co-occurrence features with two words, will be the feature vector. If the number of single feature in initial feature vector be n, the number of co-occurrence features with two words will be n*(n-1)/2. In other words, the lower the number of individual features, the final feature vector size will be smaller. So using feature selection criterion presented in this paper, the initial features are reduced for creating final feature vector with lower dimension.

6. REFFRENCES

[1] Y. Yang and X. Liu. A re-examination of text categorization methods. In SIGIR'99: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 42–49, 1999.

[2] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34: pages 1–47, 2002.

[3] J. S. Ronen Feldman. The text mining handbook: Advanced approaches to analyzing unstructured data. Cambridge University Press, Cambridge, 2007.

[4] E. Wiener, J. O. Pedersen, and A. S. Weigend. A neural network approach to topic spotting. In SDAIR'95: Proceedings of the 4th Symposium on Document Analysis and Information Retrieval, pages 317–332, 1995.

[5] M. F. Caropreso, S. Matwin, and F. Sebastiani. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In A. G. Chin, editor, Text Databases and Document Management: Theory and Practice, pages 78–102. Idea Group Publishing, Hershey, US, 2001.

[6] H. T. Ng, W. B. Goh, and K. L. Low. Feature selection, perceptron learning, and a usability case study for text categorization. In SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, pages 67–73, 1997.

[7] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11): pages 613–620, 1975.

[8] B. C. How and W. T. Kiong. An examination of feature selection frameworks in text categorization. In AIRS'05: Proceedings of 2nd Asia information retrieval symposium, pages 558–564. Lecture notes in computer science, 2005.

[9] F. Figueiredo, L.R., T. Couto, T. Salles, M. A. Goncalves, W. MeiraJr. Word co-occurrence features for text classification, Information Systems, 36, pages 843–858, 2011.

[10] E. Wiener, J. O. Pedersen, and A. S. Weigend. A neural network approach to topic spotting. In SDAIR'95: Proceedings of the 4th Symposium on Document Analysis and Information Retrieval, pages 317–332, 1995.

[11] C. Largeron, C.M., M. Gery, Entropy based feature selection for text categorization, ACM Symposium on Applied Computing, TaiChung : Taiwan, Province Of China, 2011.

[12] V. Tam, A. Santoso and R Setiono. A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization, Proceedings of the 16th International Conference on Pattern Recognition, pages 235–238, 2002.

[13] Eui-Hong (Sam) Han, George Karypis, Vipin Kumar. Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification, Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA, 1999.

[14] G.R. Dunlop. A rapid computational method for improvements to nearest neighbor interpolation, Computers& Mathematics with Applications 6(3), pages 349-353, 1980.