Asian Journal of Information Technology 5(8): 872-876, 2006
Medwell Online, 2006

Feature Selection for Efficient Text Categorization and Knowledge Discovery Using Classification Techniques

[1]Christy, A. and [2]P. Thambidurai
[1]Sathyabama Deemed University, Chennai, India
[2]Pondicherry Engineering College

Abstract: Text Categorization, which consists of automatically assigning documents to a set of categories deals with the management of huge number of features. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. It removes irrelevant, redundant or noisy data and brings immediate effects for data mining applications. In this study, we propose a filter system for feature set extraction, based on the similarity distance measure. Although past literatures have suggested that the use of features from irrelevant categories can improve the measure of text categorization, we believe that by incorporating only relevant feature can be highly effective. The experimental comparison is carried out between distance measure and four well-known classification techniques: C4.8, Multilayer perceptron, Least Mean Square and Linear Regression. The results also show that our proposed method can perform comparatively well with other classification measures, especially on a highly overlapped collection of topics and also it is found that C4.8 acts as a better classifier than other techniques.

Key words: Feature set extraction, filter, C4.8, precision, recall, information gain, etc

INTRODUCTION

As computers and database technologies advance rapidly, data accumulates in huge volumes in organizations. Data mining as a multidisciplinary joint effort from databases, statistics, machine learning, AI, expert systems plays the role of turning mountains of data into nuggets[1]. Inorder to use data mining tools effectively, data preprocessing is essential. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining.

Feature analysis deals with the methods for conditioning the raw data so that the information that is most relevant for classification and interpretation is enhanced and represented by a minimal number of features. Feature analysis consists of three major components: Nomination, selection and extraction. Feature Nomination (FN) refers to the process of proposing the original P features. Feature selection (FS) refers to choosing the best subset of S features (s<p) from the original p features. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. Feature selection is a process that selects a subset of original features. The optimality of a feature subset is measured by an evaluation criterion. As the dimensionality of the domain expands, the number of features N increases. Finding an optimal feature subset is shown to be NP-hard.

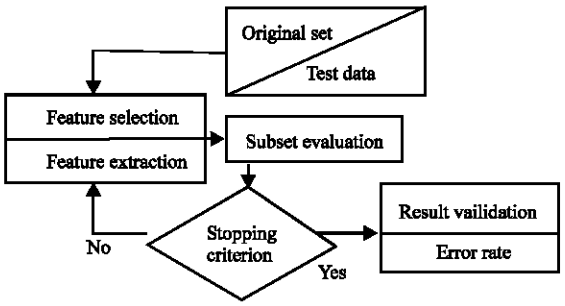

Fig. 1: Block diagram

Feature Extraction (FE) describes the process of transforming the original p-dimensional feature space into an s-dimensional space in some manner that "best" preserves or enhances the information available in the original p-space. The usual bench mark of feature quality is the empirical error rate achieved by a classifier on labeled test data. A typical feature selection process consists of 4 basic steps: Subset generation, Subset evaluation, Stopping criterion and Result validation. The possible block diagram is shown in Fig. 1.

Feature selection algorithms designed with different evaluation criteria are divided into three categories, the

Corresponding Author: Christy, A., Sathyabama Deemed University, Chennai, India

filter model, the wrapper model and the hybrid model. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm, but it is more computationally expensive than the filter model. The hybrid model takes advantage of the two models by exploiting their different evaluation criteria in different search stages.

Dash and Diu[2] divide the evaluation function into five categories: Distance, information, dependency, consistency and classifier error rate. The feature extraction method falls into two categories: One is based on the evaluation of individual features, the other is based on the evaluation of feature subsets. Information gain attribute ranking belong to the first category, whereas the Correlation based feature selection, Consistency-based feature selection and Wrapper subset selection fall into the second category[3].

Text categorization algorithms usually represent documents as bags of words and consequently have to deal with huge numbers of features. Prior studies found Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) to be the best performing algorithms for text categorization. Support vector machines are very robust even in the presence of numerous features and further observed that the multitude of text features are indeed useful for text categorization. To substantiate this claim, Joachims used a Naive Bayes classifier with feature sets of increasing size, where features were ordered by their discriminative capacity (using the information gain criterion) and then the most informative features were removed[4].

Huan Liu[1] has described a generalized filter, wrapper algorithm and Hybrid algorithm for feature selection and it has been reported that these algorithms can be implemented in text categorization. Elias F. Combarro[5] has introduced a set of linear measures for Feature selection in Text categorization. Narayanan K.[6] has shown a Categorical Descriptor Term (CTD) for text categorization, which is based on the classical term weighting scheme TFIDF. Evgeniy Gabrilovich[4] has developed a method to eliminate feature redundancy and shown that C4.5 is significantly superior to that of SVM by a narrow margin. Feature Subset Selection (FSS) refers to algorithms that select the most relevant features to the classification task, removing the irrelevant ones. Mauricio Kugler[7] has limited the methods belonging to the sequential selection algorithms, which will be applied for classification task using Support Vector Machines.

We have used the filter algorithm as shown in Table 1, in which the C4.8 algorithm is used for feature evaluation and Manhattan distance measure is used for feature extraction, where the Manhattan distance is found by the formula given below:

$$\text{dist}(t_i, t_j) = \sum_{h=1}^{k} \left|(t_{ih} - t_{jh})\right|$$

Table 1
Filter algorithm

Input: $D(F_0,F_1,\ldots F_{n-1})$ // a training data set with N features
S_0 // a subset from which to start the search
δ // a stopping criterion
Output: S_{best} // an optimal subset
begin
initialize: $S_{best} = S_0$;
γ_{best} = eval (S_0, D,A) ;
// evaluate S_0 by a classification algorithm A
do begin
S= generate(D) ; //generate a subset for evaluation
γ = eval (S,D,A);//evaluate the current subset S by A
if (γ is better than γ_{best})
$\gamma_{best} = \gamma$;
$S_{best} = S$;
end until (γ is reached);
return S_{best};
end;

We have used a classification algorithm for subset evaluation. For each generated subset S, it evaluates its goodness by applying the mining algorithm to the data with feature sunset S and evaluating the quality of mined results. Since mining algorithms are used to control the selection of feature subsets, the filter model tends to give superior performance as feature subsets are better suited to the predetermined mining algorithm.

MATERIALS AND METHODS

In our study, we have collected 200 documents related to image processing and information retrieval from www.computer.org. Initially, trained with a bag of words we have found the information gain based on its Entropy and gini index, where

$$\text{Entropy}(P) = -[p_1 \log(p_1) + p_2 \log(p_2) + \ldots + p_n \log(p_n)]$$

Info(T)=Entropy(P), where P is the probability of distribution of the partitions $C_1, C_2, \ldots C_n$

$$P = (|C_1| / T, |C_2| / T, \ldots |C_n| / T)$$

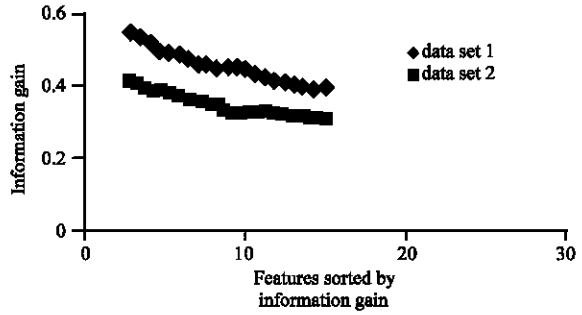

Fig. 2: Selected features of calculated based number

Gini index is a diversity measure from economics. It can be used to evaluate the goodness of a split.

Gini(T) = 1 - $\Sigma_j P_i^2$ where P_j is the relative frequency of class j in T.

In the filter model, we have removed the stop words have used the inflectional stemming algorithm to generalize the features. Based on the value of term frequency (tf), if a feature occurs in one more than one document, we have found the TFIDf as defined in[6]. The Manhattan distance is then calculated to find the similarity of the documents. The information gain is calculated based on the number of features selected and the results are shown in Fig. 2.

The classified results are verified using the classification techniques. The ID3 technique to building a decision tree is based on information theory and attempts to minimize the expected number of comparisons. The concept used to quantify information is called entropy. Entropy is used to measure the amount of uncertainty or surprise or randomness in a set of data. When all data in a set belong to a single class, there is no uncertainty. In this case case, the entropy is zero. The objective of the decision tree classification is to iteratively partition the given data set into subsets where all elements in each final subset belong to the same class. The C4.5 algorithm is the improved version of ID3 and it permits numeric attributes, deal sensibly with missing values and prune to deal with noisy data.

When the decision tree is built, missing data are simply ignored. ie., the gain ratio is calculated by looking only at the other records that have a value for their attribute. It divides the data in to ranges based on the attribute values for that item that are found in the training sample. With subtree replacement, a subtree is replaced by a leaf node if this replacement results in an error rate close to that of the original tree, which study from bottom of the tree up to the root. The major advantage is C4.8 and its advanced version of the algorithms allows classification via either decision trees or rules generated from them.It can replace the left side of a rule by a simpler version if all records in the training set are treated identically.

Multilayer perceptron is a Classifier that uses backpropagation to classify instances. We have found that C4.8 significantly outperforms Support Vector Machine (SVM) and Multilayer perceptron, eventhough SVM is considered as a good classifier. When no feature selection is performed C4.8 constructs small decision trees compared to the other methods. When feature selection is optimized for each classifier C4.8 performs better than Multilayer perceptron, but less capable than that of Linear Regression and after applying 10-fold cross validation, again C4.8 performs best. The time complexity of C4.8 is slightly expensive than that of Linear regression while it remains better than all other classification methods and the results are tabulated in Table 2 and the

RESULTS

A number of feature selection techniques have been tested for test categorization in Information gain, Linear regression, $\chi 2$, Document frequency has been reported to be most efficient. We used classification accuracy as a measure of text categorization performance. We have classified the collection of texts into four categories and the Precision, Recall and F-measure values are calculated based on the C4.8 algorithm and the results are shown in measure of text categorization performance.Fig. 3 and the results after 10-folds cross validation is shown in Fig. 4. Inorder to find the optimized solutions, we have also performed the classification using genetic algorithms, in which at the error rate 0.95 for 25 generations, we have performed cross over at 4 different

Table 2: Classification analysis

		Mean absolute error		Root mean squared error	
Method	Correlation coefficient	Before cross validation	After 10 folds cross validation	Before cross validation	After 10 folds cross validation
Least median square	-	0.27	0.625	0.477	0.2887
Multilayer perceptron	-0.1113	0.1204	0.1861	0.2783	0.3719
Linear Regression	-0.0685	0.0983	0.1179	0.2564	0.31
Decision table	-	0.1185	0.1211	0.2818	0.2883
C4.8	-	0.1095	0.1305	0.2324	0.2833

Table 3: Comparative descriptives of some of features selected

Features	Mean	SD	SE	95% CI of mean
error	0.340	1.0595	0.1076	0.127 to 0.554
diffusion	0.134	0.8736	0.0887	-0.042 to 0.310
quantization	0.052	0.4176	0.0424	-0.033 to 0.136
quality	0.521	1.2310	0.1256	0.271 to 0.770
image	2.485	2.1752	0.2209	2.046 to 2.923
data	0.485	0.8674	0.0881	0.310 to 0.659
halfton	0.093	0.4805	0.0488	-0.004 to 0.190
embed	0.093	0.2916	0.0296	0.034 to 0.152
dot	0.124	0.5450	0.0553	0.014 to 0.234
filter	0.113	0.4300	0.0437	0.027 to 0.200
size	0.082	0.3119	0.0317	0.020 to 0.145
transmission	0.134	0.5886	0.0598	0.015 to 0.253
channel	0.103	0.5100	0.0518	0.000 to 0.206
knowledge	0.042	0.2009	0.0205	0.001 to 0.082
shape	0.247	1.0107	0.1026	0.044 to 0.451
block	0.031	0.2261	0.0230	-0.015 to 0.076
frequency	0.063	0.2833	0.0289	0.005 to 0.120

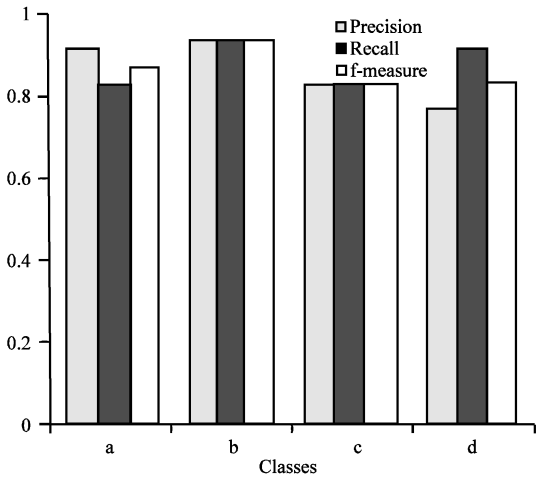

Fig. 3: F-measure values

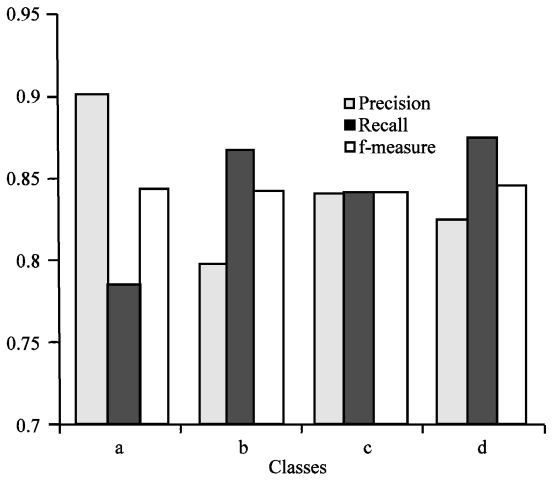

Fig. 4: 10-foldscross vailidation

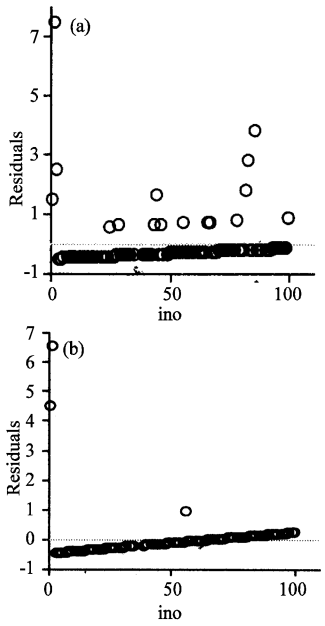

Fig. 5: Squared errors of the slope lines

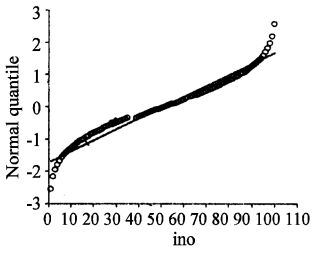

Fig. 6: Continouos summary

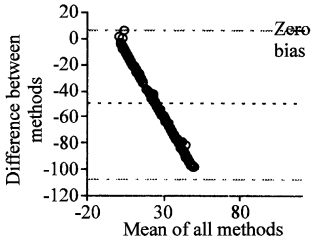

Fig. 7: Bias plots

comparative descriptives for some of the selected features are show in Table 3. poins (0.99, 0.8,0.7 and 0.6) with the mutation value 0.01 and we have found that the best recall value is found at the crossover point 0.7 as 0.55102. In the Linearity test, the squared error of the Intercept lies in the range 0.0454-0.2457 while the squared error of the slope lines between 0.0008-0.0042, which is very much tolerable as shown in Fig. 5. The continuous summary in Fig. 6 and Bias plots shown in Fig. 7 shows the good performance of the classifier.

CONCLUSION

In this study, we have shown the extraction of features using the filter model for text categorization and the classification of the features extracted using the well known classification techniques. The error rate in case of all the classification algorithms have shown the expected value lying close to the actual value. Also, we have shown the Linear Regression performing close to C4.8.

REFERENCES

1. Huan Lu and Lei Yu, 2005. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE transactions on Knowledge and Data Engin., 17: 491-502.
2. Dash, M. and H. Liu, 1997. Feature selection for classification", Intelligent Data Analysis: An Intl. J., 1: 131-156.
3. Guangzhi Qu, Salim Hariri and Mazin Yousif, 2005. A New dependency and Correlation Analysis for Features. IEEE Transactions on Knowledge and Data Engin., 17: 1199-1207.
4. Evgeniy Gabrilovich and Shaul Markovitch, 2004. Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5. Proceedings of the 21st Intl. Conference on Machine Learning, Ban®, Canada.
5. Elias F. Combarro, Elena Montanes, Irene Diaz, Jose Ranilla and Richardo Mones, 2005. Introducing a family of Linear Measures for Feature Selection in Text Categorization. IEEE transactions on knowledge and Data Engin., 17: 1223-1232.
6. Narayanan, K., 2004. An Empirical Study of Feature Selection for Text Categorization based on Term Weightage. Web Intelligence. WI 2004. Proceedings. IEEE /WIC/ ACM Intl. Conference, pp: 599-602.
7. Mauricio Kugler, Kazuma Aoki, Susumu Kuroyanagi and Akira Iwata. Feature Subset Selection for Support Vector Machines using Confident Margin.