2011 International Conference on Information and Electronics Engineering
IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore

A Chi-Square-Test for Word Importance Differentiation in Text Classification

Phayung Meesad[1], Pudsadee Boonrawd[2] and Vatinee Nuipian[2,3] +

[1] Department of Teacher Training in Electrical Engineering
[2] Department of Information Technology
[3] Department of Institute of Computer and information Technology
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

Abstract. Text classification is the main issue in order to support searches of digital libraries and the Internet. Most approaches suffer from the high dimensionality of feature space, e.g. word frequency vectors. To overcome this problem, a new feature selection technique based on a new application of the chi square test is used. Experiments have shown that the determination of word importance may increase the speed of the classification algorithm and save their resource use significantly. By doing so, a success rate of 92.20% could be reached using documents of the ACM digital library.

Keywords: digital library, text classification, support vector machine, feature selection, Chi-Square-test

1. Introduction

Development and advanced research is necessary to help facilitate recommending expert systems, complex searches, summarizing results of retrieval theories and algorithms, also tools that make it easier for researchers to develop the next generation. A semantic system is considered an important part of the problem of Information Overload because data increases every day and information retrieval today uses keywords. But this cannot process the meaning of a word and its relationship to other words[1]. Therefore many researchers have tried to research semantic retrieval. Text mining is part of this research, data classification models are used to teach the computer automatically, but one thing to consider is the ambiguity caused by the classification of information provided [2]. Saengsiri [3] and Haruechaiyasak [6] used feature selection which is the principle of word frequency measuring Information Gain, Gain Ratio, Cfs, Document frequency and Chi-Square to select the frequency of terms and attributes because these can reduce resources and increase the speed of processing. Techniques for data classification that many researchers use are Decision Tree [4] Naïve Bayes [5] Support Vector Machine (SVM) [6] and Tammasiri [7] which applies Support Vector Machine and Grid in credit score. The axis adjustment kernel function with appropriate parameters should get the best results for data classification.

Thus, a major difficulty of text categorization is the high dimensionality of the feature space. Feature selection is an important step in text categorization to reduce the feature space. This research use feature selection methods such as Information Gain, Gain Ratio, Cfs, Document frequency, Chi-Square, Consistency and Filter to compare the based methods. After that, text classification is employed to create a new model.

2. A Review of Text Categorization and Feature Selection

Text categorization is the process of automatically assigning a text document into some predefined categories and building models. For the text domain, features are a set of terms extracted from document corpus. The document corpus must be analysed to determine the ambiguous words because those words create confusion in the classification. Documents are represented by keywords or indexes which are used for retrieval, also frequencies of words are implemented using the following principles.

+ Corresponding author. Tel.: + 66 86624-0179; fax: + 662-912-2019.
E-mail address: vtn@kmutnb.ac.th, pym@kmutnb.ac.th, pudsadee@kmutnb.ac.thh

2.1 Feature Selection

The main problem for text categorization is the high dimensionality of feature space. The feature set for a text document is a set of unique terms or words that occur in all documents. Feature selection is a method which reduces the number of attributes. The advantage of reducing the attribute list is the processing speed, which in turn gains higher performance. Saengsiri [3] and Haruechaiyasak [6] presented seven feature selection models. The feature selection methods are as follows.

- Chi-Square (χ^2): based on the statistical theory. It measures the lack of independence between the terms in the category [3]. Shown in equation 1.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

- Consistency: the set of attributes evaluated by the level of compatibility of a subset of attributes. Consistency of any subset can never be lower than that of the full set of attributes; hence the usual practice is to use this subset evaluator in conjunction with a Random or Exhaustive search which looks for the smallest subset with consistency equal to that of the full set of attributes.
- Filter: the methods are based on performance evaluation metric calculated directly from the data, without direct feedback from predictors that will finally be used on data with reduced number of features. Such algorithms are usually computationally less expensive than those from the first or the second group.
- Information Gain (IG): to find node impurity, this is the main idea to select the best split. Several concepts are GINI Index, Entropy and Misclassification error [8]. *INFO* based on Entropy measurement reduces because of the separate method. Entropy at a given node t is given in (2):

$$INFO_{Entropy\,(t)} = -\sum_{i} p(\frac{j}{t})\log_2 p(\frac{j}{t}) \quad (2)$$

$p\,(\frac{j}{t})$ is associated with frequency of category *j* at node *t*.

$$Gain = Entropy(t) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right) \quad (3)$$

INFO is shown in (3) the parent node, *t* is split into *k* partition; n_i is number of records in partition *i*. Nevertheless, bias splits can happen with large partitions.

Gain Ratio (GR) : technique improves the problem of *INFO*. The structure of method is created by using to-down design. GR was developed by Quinlan in 1986 and based on evaluation of information theory. Generally, probability, ($P(v_i)$), is to answer v_i, then the information (2) of the answer is given by [9]. *SplitINFO* is presented in (4) to resolve bias in *INFO*.

In (4), *INFO* is adapted by using the entropy of the partitioning (*SplitINFO*). Thus, higher Entropy partitioning is adjusted.

$$SplitINFO = -\left(\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n} \right) \quad (4)$$

$$GainRatio = \frac{\Delta INFO}{SplitINFO} \quad (5)$$

- Document frequency (DF): is a number of terms which occur in a document. The value can be calculated for each term from a document corpus. All unique terms that have document frequency in training set less than some predefined threshold were removed [6].
- Cfs: is the measurement process which determines high correlation of dimensionality subset with class and ignores relation among them. Therefore, irrelevant features are reduced and power features [3].

2.2 Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances.

- Decision trees: tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) [9].
- The Naive Bayes (NB): algorithm was first proposed and used for text categorization task by D. Lewis (1998) [10]. NB is based on the bayes' theorem in the probabilistic frame work. The basic idea is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. NB algorithm makes the assumption of word independence, i.e., the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category.
- Support Vector Machine (SVM): is the machine learning algorithm introduced by Vapnik [11]. SVM applies in credit scoring [7]. SVM is based on the structural risk minimization with the error-bound analysis. SVM models are a close cousin to classical multilayer perceptron neural networks. Using a kernel function, SVM's are an alternative training method for polynomial, radial basis function shown in equation (6) (7).

  Polynomial function kernel: (SVMP)

$$k(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^d \tag{6}$$

  Radial basis function kernel (SVMR)

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2) \tag{7}$$

3. Experiment and Discussion

To evaluate the proposed methodology, experimental simulations were performed. Abstract data from ACM Digital Library [12] Domain Information System were used. The data consisted of 1,099 documents from 2009-2010. The data was pre-processed to obtain only data needed. The text analysis component converts semi structured data such as documents into structured data stored in a database. The fields are divided into title, author, abstract, keywords etc. Ambiguity words are considered to be part of the confusion matrix. A confusion matrix is a visualization tool typically used in supervised learning. Each column of the matrix represents the instances in a predicted class. The LexTo program [13], perform text processing and keywords selection, remove stop words and stemming.

WEKA, an open-source machine learning tool, was used to perform the experiments. WEKA has many data mining tools to be employed. In this study, Decision Tree, Naïve Bayes, BayesNet, Support Vector Machine, which are classification mechanisms, were used for judgement in feature selection process. The performance metrics to evaluate the text categorization used in the experiments were accuracy precision, recall and F-measure. The selected algorithms were training with the 10-fold cross validation technique. The feature selection for classification model in Fig. 1 and the experimental results are summarized in Table 1 below.

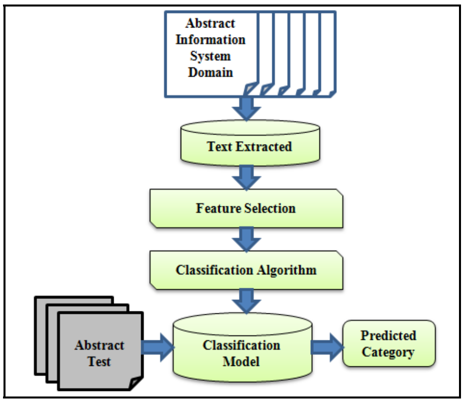


Fig. 1: Feature Selection for Classification Model.

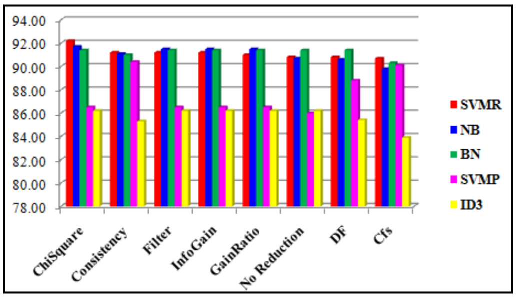


Fig. 2: Shows F-Measure values for feature selection and classification.

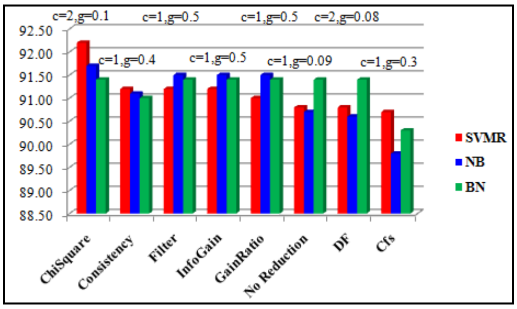
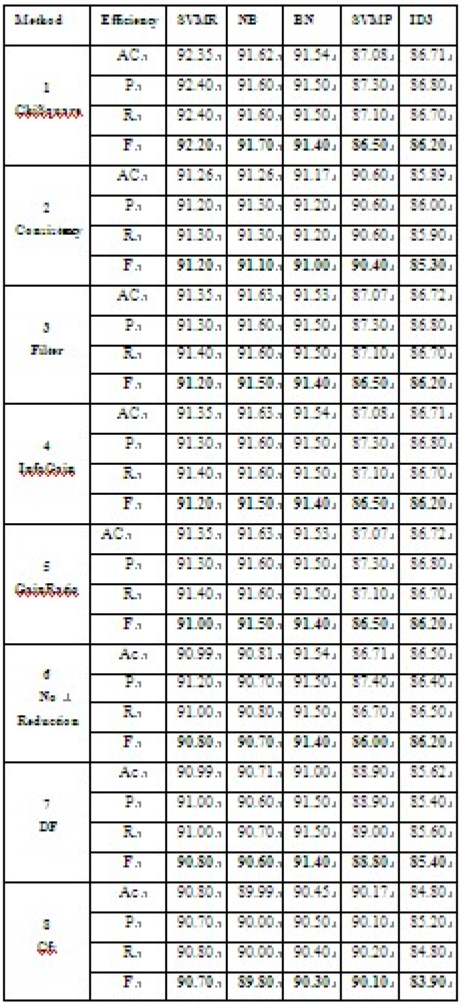


Fig. 3: High performance of NB,BN and SVM Kernel Function Radial basis function for based C and gamma.

Table 1: Text Classification Evaluation Results



Method	Efficiency	SVMR	NB	BN	SVMP	ID3
1 ChiSquare	AC.	92.35	91.62	91.54	87.08	86.71
	P.	92.40	91.60	91.50	87.30	86.80
	R.	92.40	91.60	91.50	87.10	86.70
	F.	92.20	91.70	91.40	86.50	86.20
2 Consistency	AC.	91.26	91.26	91.17	90.60	85.89
	P.	91.20	91.30	91.20	90.60	86.00
	R.	91.30	91.30	91.20	90.60	85.90
	F.	91.20	91.10	91.00	90.40	85.30
3 Filter	AC.	91.35	91.63	91.53	87.07	86.72
	P.	91.30	91.60	91.50	87.30	86.80
	R.	91.40	91.60	91.50	87.10	86.70
	F.	91.20	91.50	91.40	86.50	86.20
4 InfoGain	AC.	91.35	91.63	91.54	87.08	86.71
	P.	91.30	91.60	91.50	87.30	86.80
	R.	91.40	91.60	91.50	87.10	86.70
	F.	91.20	91.50	91.40	86.50	86.20
5 GainRatio	AC.	91.35	91.63	91.53	87.07	86.72
	P.	91.30	91.60	91.50	87.30	86.80
	R.	91.40	91.60	91.50	87.10	86.70
	F.	91.00	91.50	91.40	86.50	86.20
6 No Reduction	Ac.	90.99	90.81	91.54	86.71	86.50
	P.	91.20	90.70	91.50	87.40	86.40
	R.	91.00	90.80	91.50	86.70	86.50
	F.	90.80	90.70	91.40	86.00	86.20
7 DF	Ac.	90.99	90.71	91.00	88.90	85.62
	P.	91.00	90.60	91.50	88.90	85.40
	R.	91.00	90.70	91.50	89.00	85.60
	F.	90.80	90.60	91.40	88.80	85.40
8 Cfs	Ac.	90.80	89.99	90.45	90.17	84.80
	P.	90.70	90.00	90.50	90.10	85.20
	R.	90.80	90.00	90.40	90.20	84.80
	F.	90.70	89.80	90.30	90.10	83.90

From Table 1, we can see that χ^2 method for feature selection had the best performance. Measurement of classification was undertaken via F-measure. The results are as follows: SVMR = 92.20%, NB = 91.70%, BN = 91.40%, SVMP = 90.40% and ID3 = 86.20%, respectively. The result matched with the study given in Saengsiri [3] and Haruechaiyasak [6].

Fig. 2 SVMR classification used feature selection evaluate by F-Measure. The result are as follows ChiSquare = 92.20%, Consistency, Filter, InfoGain = 91.20%, GainRatio = 91.00% , No Reduction = 90.80, DF = 90.80% and Cfs = 90.70%.

The Optimal values of this parameter are adjusted by C and gamma parameters shown in Fig. 3.

4. Conclusions and Future Works

In this paper, Chi-Square-Test with the best classification model is proposed to overcome the high dimensionality of feature space. Data used in the experiments came from the ACM Digital Library, Domain Information System, during 2009-2010, which comprised of 1,099 documents. Searching used keywords or indexes to represent the document.

The experiments show that the proposed method improves the performance of text categorization techniques using Chi-Square (χ^2) for feature selection with the F-measure of 92.20%. The best classification model is based on Support Vector Machine with radial basis function (SVMR). Feature selection can reduce the number or features while preserve the high performance of classifiers.

Future work, to further test our approach we can increase the number of datasets and its number of patterns to see if this has any positive or negative results.

5. Acknowledgements

Choochart Haruechaiyasak Human Language Technology Laboratory (HLT) National Electronics and Computer Technology Center (NECTEC) Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand helped for provide Lexto and suggest.

6. References

[1] S. Saeneapayak, "A Development of Knowledge Warehouse Prototype for Knowledge Sharing Support: Plant Diseases Case Stusy," Inforamtion Technology, Faculty of Computer Engineering, Kasetsart University, 2005.

[2] K. Thongklin, S. Vanichayobon and W. Wett, "Word Sense Disambiguation and Attrbute Selection Using Gain Ratio and RBF Neural Network," IEEE Conference Innovation and Vision for the Future in Computing & Communication Technologies (RIVF' 08)**,** 2008.

[3] P. Saengsiri, P. Meesad, S. Na Wichian and U. Herwig, "Comparison of Hybrid Feature Selection Models on Gene Expression Data," IEEE International Conference on ICT and Knowledge Engineering, 2010, pp.13 -18.

[4] Ko. Youngjoong and Seo. Jungyun, "Using the Feature Projection Technique Based on a Normalized Voting Method for Text Classification," Information Processing & Management. Vol. 40, pp.191-208, 2004.

[5] K. Canasai and J. Chuleerat, "Thai Text Classification based on NaïveBayes," Faculty of Computer Science. Kasetsart University, 2001.

[6] C. Haruechaiyasak, W. Jitkrittum, C. Sangkeettrakarn, and C. Damrongrat, "Implementing News Article Category Browsing Based on Text Categorization Technique," The 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI-08) workshop on Intelligent Web Interaction (IWI 2008), 2008, pp.143-146.

[7] D. Tammasiri, and P.Meesad, "Credit Scorring using Data Mining based on Support Vector Machine and Grid," The 5th National Conference on Computing and Information Technology, 2009, pp.249-257.

[8] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining," Addison Wesley, 2006, pp.150-163.

[9] Quinlan, J. R, "Induction of Decision Trees," Machine Learning 1(1), 2006, pp.81-106.

[10] D. Lewis. Naive bayes at forty: The independence assumption in information retrieval. Proc. of European Conf. on Machine Learning, pages 4–15, 1998

[11] V. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York, 1995.

[12] The ACM Portal is published by the Association for Computing Machinery. Copyright 2009-2010 , Inc. Available online at http://portal.acm.org/portal.cfm

[13] LexTo : Thai Lexeme Tokenizer Available online at http://www.sansarn.com/lexto/