COMPARATIVE STUDY OF FEATURE SELECTION METHOD OF MICROARRAY DATA FOR GENE CLASSIFICATION

NURULHUDA BINTI GHAZALI

UNIVERSITI TEKNOLOGI MALAYSIA

COMPARATIVE STUDY OF FEATURE SELECTION METHOD OF MICROARRAY DATA FOR GENE CLASSIFICATION

NURULHUDA BINTI GHAZALI

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

OCTOBER 2009

To my beloved Mummy and Abah…

Hazijun bt. Abdullah and Ghazali bin Sulong

My beloved sisters..

Nurhanani and Nur Hafizah

My beloved brother..

Ikmal Hakim

My brother-in-laws..

Saiful Azril and Faridun Naim

My beloved nieces..

Sarah Afrina and Sofea Alisya

My supervisor..

Assoc. Prof. Dr.Puteh Saad

and last but not least to all my supportive friends especially

Syara, Radhiah, Zalikha, Umi and Hidzir..

"Thank you for all the support and love given"

ACKNOWLEDGEMENT

In the name of Allah, Most Gracious, Most Merciful.

All praise and thanks be to Allah for His guidance that had lead me in completing this research. His blessings had given me strength and courage throughout this past year and had helped me overcome difficulties during this research period.

First and foremost, I would like to take this opportunity to express my sincere gratitude to those who had assisted me in finishing this research. To my dear supervisor, Assoc. Prof. Dr. Puteh Saad, thank you for all your supports and guidance in showing me the right path towards completing this research. I really appreciated your advices and motivations that you had given me within the period of this research.

My infinite thank you are dedicated to my loving and caring family, who had cherish me and give me full support in any kind. I am deeply appreciated for all the motivations and inspirations. Without them, it is impossible for me to finish my research.

And last but not least, an endless appreciation to all my fellow friends and classmates for all the supports and encouragements. Their friendships never fail to amaze me.

May Allah S.W.T bless them all and repay all of their kindness and sacrifices.

ABSTRACT

Recent advances in biotechnology such as microarray, offer the ability to measure the levels of expression of thousands of genes in parallel. Analysis of microarray data can provide understanding and insight into gene function and regulatory mechanisms. This analysis is crucial to identify and classify cancer diseases. Recent technology in cancer classification is based on gene expression profile rather than on morphological appearance of the tumor. However, this task is made more difficult due to the noisy nature of microarray data and the overwhelming number of genes. Thus, it is an important issue to select a small subset of genes to represent thousands of genes in microarray data which is referred as informative genes. These informative genes will then be classified according to its appropriate classes. To achieve the best solution to the classification issue, we proposed an approach of minimum Redundancy-Maximum Relevance feature selection method together with Probabilistic Neural Network classifier. The minimum Redundancy-Maximum Relevance feature selection method is used to select the informative genes while the Probabilistic Neural Network classifier acts as the classifier. This approach has been tested on a well-known cancer dataset which is Leukemia. The results achieved shows that the gene selected had given high classification accuracy. This reduction of genes helps take out some burdens from biologist and better classification accuracy can be used widely to detect cancer in early stage.

ABSTRAK

Kemajuan terkini dalam bioteknologi, contohnya mikroarray, membolehkan tahap pengekspresan beribu-ribu gen diukur secara selari. Penganalisaan dari data mikroarray dapat memberikan pemahaman dan pengetahuan berkenaan fungsi sesuatu gen dan mekanisma pengaturannya. Penganalisaan ini adalah penting untuk mengenalpasti dan mengkelaskan penyakit-penyakit kronik terutama sekali penyakit kanser. Teknologi yang digunakan baru-baru ini dalam pengkelasan kanser adalah berdasarkan kepada maklumat dari pengekspresan gen berbanding kemunculan tumor itu secara fizikal. Walaubagaimanapun, tugas ini menjadi sukar kerana kewujudan pelbagai gangguan (noise) dalam pemprosesan data mikroarray dan juga jumlah bilangan gen yang sangat banyak. Oleh itu, ianya merupakan satu isu penting untuk memilih hanya sebilangan kecil gen daripada ribuan gen dalam data mikroarray dan ini dipanggil sebagai gen bermaklumat. Gen bermaklumat ini akan dikelaskan berdasarkan kelasnya yang sesuai. Untuk mencapai penyelesaian yang terbaik bagi permasalahan ini, kami mancadangkan pendekatan kaedah pemilihan gen iaitu 'minimum Redundancy-Maximum Relevance' bersama dengan pengkelas 'Probabilistic Neural Network'. 'minimum Redundancy-Maximum Relevance' digunakan untuk memilih gen-gen bermaklumat itu manakala 'Probabilistic Neural Network' bertindak sebagai pengkelas. Kaedah ini telah diuji ke atas sejenis penyakit kanser iaitu Leukimia. Keputusan eksperimen yang diperolehi sangat memuaskan dan ini dapat membantu kerja pakar-pakar biologi serta memberi harapan kepada masyarakat bagi mengesan kanser di peringkat awal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALL	-	Acute Lymphoblastic Leukaemia
AML	-	Acute Myeloid Leukaemia
ARFF	-	Attribute-Relation File Format
CSV	-	Comma-Separated Values
mRMR	-	Minimum Redundancy Maximum Relevance
PNN	-	Probabilistic Neural Network
DNA	-	Deoxyribonucleic Acid
k-NN	-	*k*-Nearest Neighbor
RNA	-	Ribonucleic Acid
mRNA	-	Messenger Ribonucleic Acid

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Every living organism has discrete hereditary units known as genes. Each gene provides some function or mechanism either by itself or it will combine with other genes that will eventually producing some property of its organism. Genome is a complete set of genes for an organism and is said as the ‘library” of genetic instruction that an organism inherits (Campbell and Reese, 2002). Each gene is made of deoxyribonucleic acid (DNA) molecule which consists of two long strands that tightly wound together in a spiral structure known as double helix (Amaratunga and Cabrera, 2004). Along each of these strands located various form of genes that differs by its sequences for each organism. This makes each organism unique and different from each other. The DNA molecule of an organism is located in a cell. A cell is the fundamental units of all living organism and it contains many substructure such as nucleus, cytoplasm and plasma membrane. The nucleus is where DNA is embedded. Genes in DNA is expressed by transferring its coded information into proteins that dwell in the cytoplasm. This process is called as gene expression (Russell, 2003). There are several experimental techniques to measure gene

expression such as expression vector, reporter gene, northern blot, fluorescent hybridization, and DNA microarray.

DNA microarray technology allows the simultaneous measurement of the expression level of a great number of genes in tissue samples (Paul and Iba, 2005). It yields a set of floating point and absolute values. Many explored on classification methods to recognize cancerous and normal tissues by analyzing microarray data. The microarray technology typically produces large datasets with expression values for thousands of genes (2000-20000) in a cell mixture, but only few samples are available (20-80) (Huerta *et al*.).

This study is focused on gene selection and classification of DNA microarray data in order to identify tumor samples from normal samples. Gene selection is a process where a set of informative genes is selected from the gene expression data in a form of microarray dataset. This process helps improve the performance of the classifier. On the other hand, classification is a process to classify microarray data in several classes that have its own characteristics. There are several techniques that have been used in gene selection such as ReliefF Algorithm, Information Gain, minimum Redudancy Maximum Relevance (mRMR) and Chi Square. For classification of microarray data, a few techniques have been applied in the bioinformatics field to classify the highly dimensional data. These techniques include Random Forest, Naïve Bayes and Probabilistic Neural Network (PNN).

The proposed method involved two stages where the first stage is the gene selection stage and the second one would be the classification stage. In gene selection method, the technique chosen is a technique called minimum Redundancy-Maximum Relevance (mRMR) feature selection and will be compared to three other method namely ReliefF, Information Gain and Chi Square. mRMR is a feature selection framework that was introduced by Ding and Peng in 2005. They supplement the maximum relevance criteria along with minimum redundancy criteria to choose additional features that are maximally dissimilar to already identified ones.

This can expand the representative power of the feature subset and help improves their generalization properties. The classification problem will be handled by Probabilistic Neural Network (PNN) technique. PNN has been widely used in solving classification problems. This is because it can categorize data accurately (Nur Safawati Mahshos, 2008). Both techniques will be assessed on a bench mark cancer dataset which is Leukemia (Golub *et al,* 1999).

1.2 Background of the Problem

Cancer is a killer disease to everyone worldwide. There are at least 100 different types of cancer that has been identified. Traditionally cancer is diagnosed based on the microscopic examination of patients' tissue. This kind of diagnosis may fail when dealing with unusual or atypical tumors. Currently, cancer diagnosis is based on clinical evaluation and also referring to medical history and physical examination. This diagnosis takes a long time and might however limit the finding of tumor cell especially in early tumor detection (Xu and Wunsch, 2003). If tumor cell is found in its critical stage, then it might be too late to cure the patient.

Thus, classification for cancer diseases has been widely carried out for the past 30 years. Unfortunately, there has been no general or perfect approach to identify new classes or assigning tumors to known classes. This happens because there are various ways that can cause cancer and too many types of cancer that sometimes difficult to distinguish. By depending on morphological appearance of tumors, it is hard to discriminate between two similar types of cancer (Golub *et al,* 1999).

In order to overcome the above issues, a new technique based on cancer classification has been introduced. The technique employs an advanced microarray technology that measures simultaneously the expression level of a great number of genes in tissue samples. Nevertheless, this technique contributes to a new problem whereby there exist a numerous number of irrelevant genes or overlapping of genes. Hence, selection and classification must be done in order to select the most significant genes from a pool of irrelevant genes and noises.

Nowadays, there are a lot of selection and classification techniques that has already been studied and developed to help in better classification of microarray data. Among these techniques, there are a few that gives promising result such as mRMR, ReliefF, Information Gain and Chi Square for gene selection and PNN classification. mRMR is chosen as the primary technique for gene selection since this technique are proposed originally for gene selection (Ding and Peng, 2003). The advantage of this technique is it focuses on redundancy of genes together with the relevance of genes. Unlike other techniques; ReliefF (Kononenko, 1994), Information Gain (Cover and Thomas, 1991) and Chi Square (Zheng *et al*, 2003), they were firstly proposed only for general feature selection, rather than genes. For comparison, these four techniques are used to select genes in order to measure the performance.

As for classification, the technique chosen in this research is Probabilistic Neural Network (PNN) classifier. PNN has been use in many studies of feature classification (Pastell and Kujala, 2007; Shan *et al*, 2002). These studies have proved that PNN yield better result in classification accuracy compared to other existing classifiers. Thus, this research combines a few feature selection methods together with PNN classifier to classify microarray data according to its classes.

1.3 Problem Statement

The challenging issue in gene expression classification is the enormous number of genes relative to the number of training samples in gene expression dataset. Not all genes are relevant to distinguish between different tissue types (classes) and introduced noise (Liu and Iba, 2002) in the classification process and thus it drowns out the contribution of the relevant genes (Shen *et al*, 2007). On top of that, a major goal of diagnostic research is to develop diagnostic procedures based on inexpensive microarrays that have adequate number of genes to detect diseases. Hence, it is crucial to recognize whether a small number of genes will be sufficient enough for gene expression classification.

1.4 Objectives of Research

The aim of this research is to select a set of meaningful genes using a minimum Redundancy-Maximum Relevance feature selection technique and to classify them using Probabilistic Neural Network. In order to achieve aim, the following objectives must be fulfilled:

1. To select a set of meaningful genes using Minimum Redundancy-Maximum Relevance (mRMR), Information Gain, ReliefF and Chi Square.
2. To evaluate the effectiveness of feature selection method using Probabilistic Neural Network (PNN) classifier.
3. To compare the performance of mRMR as feature selection method using PNN, Random Forest, and Naïve Bayes classifiers.

1.5 Scope of Research

The scope of study is stated as below:

- mRMR, ReliefF, Information Gain and Chi Square is utilized for gene selection.
- PNN technique is used for gene expression classification.
- Leukemia microarray dataset is used for testing (Data source: Weka Software Package, *http://www.cs.waikato.ac.nz/ml/weka/*)
- 10-fold cross validation is utilized to perform the validation.
- The tools used are Matlab, Knime, Weka and IOS GeneLinker

1.6 Importance of the Study

This study is carried out to aid in classification of cancer diseases. Cancer diseases are lethal to human. Several methods have been conducted to detect this deadly disease. Unfortunately, the time taken is too long to confirm that someone has the disease. This is due to the symptoms that can only be seen after a very long time and by the time, cancer level has reached a critical stage.

Common examination of patients require weekly checkup to precisely identify the presence of the disease. Due to the long term of examination, the disease might get more critical without exact cure or treatment. The advanced technology of microarray lessens the burden among medical staffs. The microarray of human genes can be used to detect cancer diseases earlier.

Despite the fact that microarray technology is said has the capability to solve the problems, but unfortunately this technology requires an excellent technique to select only the best subset of all genes to give enough information about a particular cancer disease. This is due to the overwhelming number of genes produce by microarray in a few sample sizes.

Thus, by doing this research, the best approach can be achieved to solve the problems in gene selection and classification. The idea was to apply the minimum Redundancy-Maximum Relevance feature selection technique (compared with other feature selection techniques) together with Probabilistic Neural Network to give a tremendous result in a short time. This research provides knowledge in the field of bioinformatics and it gives benefit in medical area. Apart from that, it helps saving human life by detecting cancer disease in early stage.

REFERENCES

Ahlers, F.J., Carlo, W.D., Fleiner, C., Godwin, L., Mick, Nath, R.D., Neumaier, A., Phillips, J.R., Price,K., Storn, R., Turney,P., Wang, F., Zandt, J.V., Geldon, H., Gauden, P.A. Differential Evolution. (accessed May 20, 2009). http://www.icsi.berkeley.edu/~storn/code.html

Alon, U., Barkai, N., Notterman, D.A., *et al.* (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*. Vol 96: 6745-6750

Amaratunga. D. and Cabrera, J. (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. New Jersey, USA: Wiley Inter-Science. 8-10

Babu, B.V. and Chaturvedi, G. Evolutionary Computation Strategy for Optimization of an Alkylation Reaction. *Birla Institute of Technology and Science*.

Babu, B.V. and Sastry, K.N.N (1999). Estimation of Heat Transfer Parameters in a Trickle-bed Reactor using Differential Evolution and Orthogonal Collocation. *Elsevier Science*.

Balasundaram Karthikeyan, Srinivasan Gopal, Srinivasan Venkatesh and Subramanian Saravanan. (2006). PNN and its Adaptive Version – An Ingenious Approach to PD Pattern Classification Compared with BPA Network. *Journal of Electrical Engineering.* Vol 57: 138-145.

Bi, C., Saunders, M. C. and McPheron, B. A. (2007). Wing Pattern-Based Classification of the *Rhagoletis pomonella* Species Complex Using Genetic Neural Networks. *International Journal of Computer Science & Applications.* Vol 4: 1-14

Breiman, L., 2001. Random Forests. Mach. Learn. 40, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont.

Breiman, L., 2003. RF/tools—A class of two eyed algorithms. In: SIAM Workshop, http://oz.berkeley.edu/users/breiman/siamtalk2003.pdf.

Campbell, N.A. and Reece, J.B. (2002). *Biology*. Sixth edition. San Francisco: Benjamin Cummings.

Comma-separated Values. Wikipedia. (accessed June 15, 2009). http://en.wikipedia.org/wiki/Comma-separated_values.

Cover, T., and Thomas, J. (1991). *Elements of Information Theory*. New York :John Wiley and Sons.

Data Preparation. Encyclopedia.com, (accessed October 6, 2009). http://www.encyclopedia.com/doc/1O11-datapreparation.html

Díaz-Uriarte R, Alvarez de Andrés S. (2006). Gene Selection and Classification of Microarray Data using Random Forest. *BMC Bioinformatics*. 2006 Jan 6;7:3.

Ding, C. and Peng, H. (2003). Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Proceedings of the Computational Systems Bioinformatics*.

Ding, C. and Peng, H. (2005). Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *Journal of Bioinformatics and Computational Biology*. Vol 3: 185-205.

Discretization of Continuous Features, Wikipedia, (accessed Oct 23, 2009) http://en.wikipedia.org/wiki/Discretization_of_continuous_features

DNA. Wikipedia, (accessed May 11, 2009). http://en.wikipedia.org/wiki/DNA

Dudoit, S. and Gentleman, R. (2003). Classification in Microarray Experiment.

Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: *Machine Learning. Proceedings of the Thirteenth International Conference*. pp. 148–156.

Golub, T.R., Slonim, D.K, Tamayo, P., *et al*. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. Vol 286: 531-537

Huerta, E.B., Duval, B., Hao, J.K. A hybrid GA/SVM approach for gene selection and classification of microarray data.

Jamain, A. and Hand, D. J. (2005). The Naïve Bayes Mystery: A Classification Detective Story. *Pattern Recognition Letters*. Vol 26: 1752-1760

Jin, X., Xu, A., Bie, R., and Guo, P. (2006). Machine Learning and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles. In Li, J. *et al.* (Eds) *Data Mining for Biomedical Applications.* (pp: 106-115). Berlin Heidelberg: Springer-Verlag

Kim, Y.B. and Gao, J. (2006). A New Hybrid Approach for Unsupervised Gene Selection. *IEEE Explorer*.

Kohavi, R and John, G.H. (1997). Wrappers for Feature Subset Selection.

Kononenko, I. (1994). Estimating Attributes: Analysis and Extensions of Relief. *Proceedings of the European Conference on Machine Learning*. Springer-Verlag New York. 171-182

K-nearest_neighbor_algorithm, Wikipedia, (accessed May 21, 2009) en.wikipedia.org/wiki/K-nearest_neighbor_algorithm

Lakshminarasimman, L. and Subramanian, S. (2008). Applications of Differential Evolution in Power System Optimization. *Advances in Differential Evolution.* Vol 143: 257-273.

Langdon, W.B. (2005). Evolving Benchmarks. *The 17th Belgian-Dutch Conference on Artificial Intelligence*. 365–367.

Liu, J. and Iba, H. (2002). Selecting Informative Genes Using a Multiobjective Evolutionary Algorithm. *Proceedings of the 2002 congress*. 12-17 May. 297-302

Mendes, S.P., Pulido, J.A.G., Rodriguez, M.A.V., Simon, M.D.J., Perez, J.M.S. (2006). A Differential Evolution Based Algorithm to Optimize the Radio Network Design Problem.

Mishra, S.K. (2006). Global Optimization by Differential Evolution and Particle Swarm Methods: Evaluation on Some Benchmark Functions. *MPRA Paper* 1005. 7 November 2007.

Mitchell, T.M. (1997). *Machine Learning.* New York : McGraw-Hill

Muhammad Faiz bin Misman (2007). *Pembangunan Program Selari Menggunakan Message Passing Interface(MPI) Pada Teknik Gabungan Algoritma Genetik dan Mesin Sokongan Vektor.* Universiti Teknologi Malaysia: Tesis Sarjana Muda

New Gene Selection Method. The Medical News, July 8, 2004 (accessed May 14, 2009). http://www.news-medical.net/news/2004/07/08/3157.aspx

Nur Safawati binti Mahshos (2008). *Pengecaman Imej Kapal Terbang Dengan Menggunakan Teknik Rangkaian Neural Radial Basis Fuction Dan Rambatan Balik.* Universiti Teknologi Malaysia: Tesis Sarjana Muda

Nurulhuda binti Ghazali (2008). *A Hybrid of Particle Swarm Optimization and Support Vector Machine Approach for Genes Selection and Classification of Microarray Data.* Universiti Teknologi Malaysia: Tesis Sarjana Muda

Park, H., and Kwon, H-C. (2007). Extended Relief algorithms in instance-based Feature Filtering. *Sixth International conference on Advanced Language Processing and Web Information Technology.* 123-128

Pastell, M.E. and Kujala, M. (2007). A Probabilistic Neural Network Model for Lameness Detection. *American Dairy Science Association*. Vol 90: 2283-2292.

Paul, T.K and Iba, H. (2005). Gene selection for classification of cancers using probabilistic model building genetic algorithm. *BioSystems*. Vol 82: 208-225

Peng, H., Long, F. and Ding, C. (2005). Feature Selection Based on Mutual Imformation: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pettern Analysis and Machine Intelligence.* Vol 27: 1226-1238.

Peng, H. (2005). mRMR (minimum Redundancy Maximum Relevance Feature Selection).(accessed June 1, 2009). http://penglab.janelia.org/proj/mRMR/index.htm

Principal Component Analysis, Wikipedia, (accessed May 14, 2009) http://en.wikipedia.org/wiki/Principal_components_analysis

Russell, P.J. (2003). *Essential iGenetics.* San Francisco: Benjamin Cummings. 226-265

Savitch, W. (2006). *Problem Solving with C++*. Sixth edition. USA: Pearson International Edition.

Shan, Y., Zhao, R., Xu, G., Liebich, H.M. and Zhang, Y. (2002). Application of Probabilistic Neural Network in the Clinical Diagnosis of Cancers based on Clinical Chemistry Data. *Analytica Chimica Acta*. 77-86.

Shen, Q., Shi, W.-M., Kong, W., Ye, B.-X. (2007). A Combination of Modified Particle Swarm Optimization Algorithm and Support Vector Machine for Gene Selection and Tumor Classification. *Talanta*. Vol 71: 1679-1683

Shena, M. (2003). *Microarray Analysis*. New Jersey: John Wiley & Sons, Inc.

Specht, D.F. (1990). Probabilistic neural networks. *Neural Networks.*Vol 3 : 110-118.

Storn, R. and Price, K. (1997). Differential Evolution – A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces. *Journal of Global Optimization.*

Suzila binti Sabil (2007). *Aplikasi Prinsip Analisis Komponen dan Rangkaian Neural Perceptron untuk Mengkelaskan Data Kanser Usus*. Universiti Teknologi Malaysia: Tesis Sarjana Muda

Vapnik, V. N. (1995). *The nature of Statistical Learning Theory*. Springer, New York

Vasan, A. and Raju, K.S. Optimal Reservoir Operation Using Differential Evolution. *Birla Institute of Technology and Science.*

Visen, N. S., Paliwal, J., Jayas, D.S. and White, N.D.G. (2002). Specialist Neural Networks for Cereal Grain Classification. *Biosyst. Eng.*Vol 82:151–159.

Xu, R. and Wunsch, D. C. (2003). Probabilistic Neural Networks for Multi-class Tissue Discrimination with Gene Expression Data . *Proceedings of the International Joint Conference on Neural Network.* Vol 3: 1696-1701

Xue, F. (2004). *Multi-objective Differential Evolution: Theory and Applications.*Rensselaer Polytechnic Institute: Doctor of Philosophy.

Yang, Z., Yang, Z., Lu, W., Harrison, R.G., Eftestøl, T. and Steene, P.A. (2005). A Probabilistic Neural Network as the Predictive Classifier of out-of-hospital Defibrillation Outcomes. *Resuscitation*. Vol 64:31–36.

Yousefi, H., Handroos, H. and Soleymani, A. (2008). Application of Differential Evolution in System Identification of a Servo-hydraulic System with a Flexible Load. *Elsevier*. Vol 18: 513-528

Yuan, S.-F. and Chu, F.-L. (2007). Fault Diagnostics based on Particle Swarm Optimization and Support Vector Machines. *Mechanical Systems and Signal Processing*. Vol 21: 1787-1798

Zhang, L.-X., Wang, J.-X., Zhao, Y.-N. and Yang, Z.-H. (2003). A Novel Hybrid Feature Selection Algorithm: Using ReliefF Estimation for GA-Wrapper Search. *Proceedings of the Second International Conference on Machine Learning and Cybernetics*. Xi'an. 380-384.

Zheng, Z., Srihari, R. and Srihari, S. (2003). A Feature Selection Framework Text Filtering. *Proceedings of the Third IEEE International Conference on Data Mining*.