A NEW CLASSIFICATION TECHNIQUE BASED ON HYBRID FUZZY SOFT SET THEORY AND SUPERVISED FUZZY C-MEANS

BANA HANDAGA

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

AUGUST 2013

DEDICATION

To my father ‘Paino’, May ALLAH (S.W.T) makes al-jannah to be his final residence.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Dr. Mustafa Mat Deris, my main supervisor, who has excellently directed and supported me during my PhD Program and especially in the writing up of my thesis. I would like to express my thanks to my parent for their vigorous support throughout the period of my study and their encouragement kindness made me feel at home during my study at UTHM. My sincere appreciation and thanks to my fellow PhD students for their true friendship and support. The same appreciations also go to the entire people in postgraduate office.

ABSTRACT

Recent advances in information technology have led to significant changes in today's world. The generating and collecting data have been increasing rapidly. Popular use of the World Wide Web (www) as a global information system led to a tremendous amount of information, and this can be in the form of text document. This explosive growth has generated an urgent need for new techniques and automated tools that can assist us in transforming the data into more useful information and knowledge. Data mining was born for these requirements. One of the essential processes contained in the data mining is classification, which can be used to classify such text documents and utilize it in many daily useful applications. There are many classification methods, such as *Bayesian*, *K-Nearest Neighbor*, *Rocchio*, SVM classifier, and Soft Set Theory used to classify text document. Although those methods are quite successful, but accuracy and efficiency are still outstanding for text classification problem. This study is to propose a new approach on classification problem based on hybrid fuzzy soft set theory and supervised fuzzy c-means. It is called Hybrid Fuzzy Classifier (HFC). The HFC used the fuzzy soft set as data representation and then using the supervised fuzzy c-mean as classifier. To evaluate the performance of HFC, two well-known datasets are used i.e., 20 Newsgroups and Reuters-21578, and compared it with the performance of classic fuzzy soft set classifiers and classic text classifiers. The results show that the HFC outperforms up to 50.42% better as compared to classic fuzzy soft set classifier and up to 0.50% better as compare classic text classifier.

ABSTRAK

Kemajuan terkini dalam teknologi maklumat telah membawa kepada perubahan penting dalam dunia hari ini. Menjana dan mengumpul data telah meningkat dengan pesat. Penggunaan popular Jaringan Sejagat (www) sebagai sistem maklumat global membawa kepada jumlah maklumat yang sangat banyak, dan ini mungkin adalah dalam bentuk dokumen teks. Ledakan pertumbuhan ini telah menjana keperluan segera bagi teknik-teknik baru dan alatan berautomatik yang boleh membantu kita dalam mentransformasi data kepada maklumat dan pengetahuan yang lebih berguna. Perlombongan data dilahirkan bagi keperluan ini. Salah satu proses penting yang terkandung di dalam perlombongan data adalah klasifikasi, yang boleh digunakan untuk mengklasifikasikan dokumen teks tersebut dan digunakan dalam pelbagai aplikasi kehidupan seharian. Terdapat pelbagai kaedah klasifikasi, seperti *Bayesian*, *K-Nearest Neighbor*, *Rocchio*, pengkelas SVM, dan Soft Set Theory yang digunakan untuk mengklasifikasikan dokumen teks. Walaupun kaedah tersebut boleh dikira sebagai sukses, tetapi ketepatan dan kecekapan masih belum jelas bagi permasalahan klasifikasi teks. Kajian ini adalah untuk mencadangkan satu pendekatan baru kepada permasalahan klasifikasi berdasarkan hibrid teori set lembut kabur dan c-min berselia kabur. Ia dipanggil Pengkelas Hibrid Kabur (HFC). HFC menggunakan set lembut kabur sebagai perwakilan data dan kemudiannya menggunakan c-mean berselia kabur sebagai pengkelas. Bagi menilai prestasi HFC, dua set data yang diketahui ramai digunakan iaitu, *20 Newsgroup* dan *Reuters-21578*, dan dibandingkan dengan prestasi pengkelas klasik Fuzzy Soft Set dan pengkelas klasik teks. Dapatan menunjukkan bahawa HFC melebihi performa sehingga 50.42% lebih baik berbanding dengan pengkelas Fuzzy Soft Set klasik dan 0.50% lebih baik dibanding pengkelas teks klasik.

TABLE OF CONTENTS

	TITLE		**i**
	DECLARATION		**2**
	DEDICATION		**3**
	ACKNOWLEDGEMENT		**4**
	ABSTRACT		**5**
	ABSTRAK		**6**
	TABLE OF CONTENTS		**7**
	LIST OF TABLES		**9**
	LIST OF FIGURES		**12**
	LIST OF ABBREVATIONS		**14**
CHAPTER 1	**INTRODUCTION**		**1**
	1.1	Background	1
	1.2	Problem Statement	5
	1.3	Research Objectives	6
	1.4	Contributions	6
	1.5	Research Scope	7
	1.6	Thesis Organization	7
	1.7	Chapter Summary	7
CHAPTER 2	**CLASSIFICATION AND SOFT SET THEORY**		**9**
	2.1	Introduction	9
	2.2	Datasets, observations, features	13
	2.3	Structured and Unstructured Data	15
	2.4	Classification	16

2.4.1 Binary Classification 19
2.4.2 Multi-class Classification 20
2.4.2.1 One-Against-All 20
2.4.2.2 All-Against-All 20
2.4.2.3 Error-Correcting Output-Coding (ECOC) 21
2.4.3 Multi-label Classification 21
2.5 The classification of text documents 29
2.5.1 Text (document) representation 31
2.6 Approaches to Text Classification 34
2.6.1 Rocchio Classifier 34
2.6.2 Bayesian Classifier 37
2.6.3 K-Nearest Neighbor Classifier 39
2.6.4 Support Vector Machines 42
2.7 Introduction to Soft Set Theory 45
2.8 Fuzzy C-means 47
2.9 Performance Measurement 48
2.9.1 Micro-average F-Measure 51
2.9.2 Macro-average F-Measure 52
2.10 Chapter Summary 52

CHAPTER 3 HYBRID FUZZY SOFT SET AND FUZZY C-MEAN CLASSIFIER 53

3.1 Terminology **Error! Bookmark not defined.**
3.2 Soft Set Theory 53
3.3 Fuzzy Soft Set Theory 56
3.4 Similarity Measure of Two Fuzzy Soft Sets 59
3.5 The Degree of Membership a Fuzzy Soft Set to other Fuzzy Soft Set 61
3.6 Algorithm of Hybrid Fuzzy Classifier 67
3.7 Chapter Summary 75

CHAPTER 4 RESULTS AN DISCUSSION 76

4.1 Experiment 1: Classifying the Documents of the 20 Newsgroups dataset 77

	4.1.1	The Preprocessing of 20 Newsgroups Dataset	77
	4.1.2	Classification of the 20 Newsgroups dataset	85
	4.1.2.1	The Soft Set Classifier (based on comparison table)	87
	4.1.2.2	The Fuzzy Soft Set Classifier (FSSC)	88
	4.1.2.3	The Hybrid Fuzzy Classifier (HFC)	89
	4.1.2.4	The Rocchio Classifier	90
	4.1.2.5	The k-NN Classifier	90
	4.1.2.6	The Bayesian Classifier	91
	4.1.2.7	The Support Vector Machines (SVM) Classifier	92
	4.1.3	Performance Comparison between the Classifiers on the 20 Newsgroups Dataset	93
	4.2	Experiment 2: The Reuters-25178 Dataset	96
	4.2.1	The Document Distribution of the Reuters-21578 dataset	98
	4.2.2	Preprocessing of the Reuters-21578 dataset	98
	4.2.3	Classification of the Reuters-21578 dataset	101
	4.2.4	Performance Comparison between the Classifiers on the Reuters-21578 Dataset	105
	4.3	Chapter Summary	107
CHAPTER 5	**CONCLUSIONS AND FUTURE WORK**		**108**
	5.1	Research Summary	108
	5.2	Comparison between HFC and the others Soft Set Classifier Families	110
	5.3	Comparison between HFC and others Classic Text Classifiers	111
	5.4	Contributions of Research	111
	5.5	Future Works	112
	5.6	Chapter Summary	**Error! Bookmark not defined.**
REFERENCES			113
APPENDIX			124

LIST OF TABLES

2.1	The vertebrate data set.	17
2.2	The unseen vertebrate data set.	18
2.3	ECOC example.	21
3.1	Tabular representation of soft set $\boldsymbol{F_A}$.	54
3.2	A term document matrix using binary weights schemas	55
3.3	Table representation of *fs*-set $\boldsymbol{\Gamma_A}$	57
3.4	A term document matrix using *tf*-*idf* normalized weight schemas.	58
3.5	The table representation of $\boldsymbol{c\Gamma_A}$	59
3.6	Comparison table of three types of measures of similarity	60
3.7	The 30 samples document from three document classes in the 20 Newsgroups datasets	69
3.8	List of attributes (289) after removing the stop-words	71
3.9	The entry value of matrix $\mathcal{D}T$ (continue)	72
3.10	The sample of three cardinal set (the real size is 289×1).	73
3.11	The not-null attributes of the nine unknown class documents	74
3.12	The degree membership of test documents to the three document classes	74
4.1	The document distribution of the 20 Newsgroups dataset	78
4.2	The results of classification 20 Newsgroup dataset using SS classifier	87
4.3	The results of the classification 20 Newsgroups dataset using FSSC classifier	88
4.4	The results of classification 20 Newsgroups dataset using HFC-classifier	89
4.5	The results of classification 20 Newsgroups dataset using Rocchio classifier	90
4.6	The results of the classification 20 Newsgroups dataset using k-NN classifier	91

4.7	The results of the classification 20 Newsgroups dataset using Bayesian classifier	92
4.8	The results of classification 20 Newsgroups dataset using SVM Classifier	93
4.9	The comparison of performance between the soft-set classifier families on the 20 Newsgroups dataset	94
4.10	The comparison of performance between the HFC and the four classic classifiers on the 20 Newsgroups dataset	95
4.11	The document distribution of Reuters-21578 dataset	97
4.12	The result of classification Reuters-21578 dataset using SS classifier	102
4.13	The result of classification Reuters-21578 dataset using FSSC classifier	102
4.14	The result of classification Reuters-21578 dataset using HFC classifier	103
4.15	The result of classification Reuters-21578 dataset using Rocchio classifier	103
4.16	The result of classification Reuters-21578 dataset using Bayesian classifier	104
4.17	The result of classification Reuters-21578 dataset using k-NN classifier	104
4.18	The result of classification Reuters-21578 dataset using SVM classifier	105
4.19	Comparison between soft-set classifier families	105
4.20	Comparison between HFC and four classic text classifiers	106

LIST OF FIGURES

2.1 Data mining as a step in the process of knowledge discovery. 11
2.2 Architecture of a typical data mining system. 12
2.3 Classification as the task of mapping an input attribute set *x* into its class label *y*. 17
2.4 General approach for building a classification model 19
2.5 A decision tree for the concept buys computer 24
2.6 Fuzzy membership of income 29
2.7 Rocchio classification 35
2.8 The pseudo code of Rocchio algorithm: (a) training and (b) testing 36
2.9 The pseudo code of multinomial NB algorithm: (a) Training and (b) Testing 38
2.10 Cosine similarity. 40
2.11 The pseudo code of k-nearest neighbors algorithm: (a) Training and (b) Testing 41
2.12 SVM for a simple classification scenario: The solid line represents the separating hyperplane. Points on the dashed lines are the support vectors 43
2.13 The pseudo code of SVM algorithm: (a) Training and (b) testing 44
2.14 The unsupervised FCM algorithm 48
2.15 Contingency table for binary classification 48
2.16 A sample coincidence matrix for a two class classifier 50
2.17 A sample coincidence matrix for a three class classifier 50
3.1 The data distribution on an axis-*x* 62
3.2 The crisp membership function 63
3.3 The membership function in the Fuzzy C-Means 63
3.4 Matrix degree of membership (a) Crisp, (b) Fuzzy 64

3.5	The mono-dimensional data initialized randomly	64
3.6	The degree of membership after running the algorithm	65
3.7	Pseudo code of HFC	68
3.8	The research framework	69
4.1	The document distribution of 20 Newsgroups dataset.	78
4.2	Python script 'preprocess.py' for document pre-processing	79
4.3	Running 'preprocess.py' script with IPyhton interface	79
4.4	A news from 20 newsgroups dataset for tokenize example	80
4.5	Tokens of the news from Figure 4.4	80
4.6	The distribution of terms in the collection.	81
4.7	The visualization of the non-zero elements of matrix TD.	81
4.8	The 119 terms in 'stop word'	82
4.9	(a) The terms distribution in the collection. (b) The visualization of the non-zero elements of matrix TD.	83
4.10	The pseudo code to create fuzzy soft set representation of document collection	83
4.11	The term distribution in the collection after the TF-IDF normalize.	84
4.12	Training and testing dataset split randomly	85
4.13	The pseudo code of creating two indicators class-label.	86
4.14	Comparison of F-Measure and cpu-times between HFC, SS, and FSSC on 20 Newsgroups dataset.	94
4.15	Comparison F-Measure and cpu-times between HFC, Rocchio, Bayesian, k-NN and SVM on 20 Newsgroups dataset	96
4.16	The document distribution of Reuters-21578 dataset	98
4.17	Running 'preprocessing_r.py' script with IPython interface	99
4.18	The terms distribution of the Reuters-21578 dataset	99
4.19	The visualization of the non-zero elements of matrix TD	100
4.20	The terms distribution in the collection after stemmed process	100
4.21	The terms distribution in the collection after TF-IDF normalizes	101

LIST OF ABBREVIATIONS

ACC	Accuracy
DF	Document Frequency
ECOC	Error Correcting Output Coding
FCM	Fuzzy C-Means
FN	False Negative
FP	False Positive
FSSC	Fuzzy Soft Set Classifier
HFC	Hybrid Fuzzy Classifier
IDF	Inverse Document Frequency
KDD	Knowledge Dicovery from Data
k-NN	K Nearest Neighbor
NB	Naïve Bayes
SSC	Soft Set Classifier
SVM	Support Vector Machine
TC	Text Classification
TDM	Term Document Matrix
TF	Term Frequency
TF-IDF	Term Frequency – Inverse Document Frequency
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
WWW	World Wide Web

LIST OF SYMBOLS

U	:	Initial universe.
$P(U)$	:	The power set of U.
$S(U)$	:	The set of all the soft sets over U.
$F(U)$	:	The set of all the fuzzy sets over U.
$FS(U)$	:	The set of all *fs*-sets over U.
$cFS(U)$	:	The set of all cardinal sets of *fs*-sets over U.
E	:	A set of parameters, and $A \subseteq E$.
F_A	:	A soft set.
f_A	:	A soft set approximation function.
μ_X	:	A membership functions of X.
Γ_A	:	A fuzzy soft set.
γ_A	:	A fuzzy approximate functions.
$c\Gamma_A$	:	A cardinal set of fuzzy soft set Γ_A.

CHAPTER 1

INTRODUCTION

1.1 Background

Recent advances in information technology have led to significant changes in today's world. The processes of generating and collecting data have been increasing rapidly. Contributing factors that lead to this include the computerization of business, scientific, and government transactions; the widespread use of digital cameras, publication tools, and bar codes for most commercial products; and advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems. In addition, popular use of the World Wide Web (www) as a global information system led to a tremendous amount of information. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can assist us in transforming the data into more useful information and knowledge (Han & Kamber, 2011).

Data mining was born for these requirements. Data mining refers to extracting or "mining" knowledge from large amounts of data. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD (Han & Kamber, 2011). Fayyad *et al*. (1996) has another view that is KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.

In computer science, data mining also called knowledge discovery in databases (KDD) is the process of discovering interesting and useful patterns and relationships in large volumes of data (Britanica, 2013).

In general, data mining tasks can be classified into two categories: descriptive and predictive (Han & Kamber, 2011). Descriptive mining tasks characterize the general properties of the data in the database. While predictive mining tasks perform inference on the current data in order to make predictions. In some cases, users may have no idea regarding what kinds of patterns in their data may be interesting, that could lead to searching for several other kinds of patterns in parallel. As such, it is important to have a system that can mine multiple kinds of patterns to accommodate different user expectations. Data mining functionalities consist of (a) concept or class description, (b) mining frequent patterns, associations, and correlations (c) classification and prediction (d) cluster analysis (e) outlier analysis and (f) evolution analysis.

1.2 Classification and Prediction

A bank officer needs analysis of her data in order to learn which loan applicants are “safe” and which are “risky” for the bank. A manager at computer shop needs data analysis to help guess whether a customer with given profile will buy a new machine. A researcher wants to analyze breast cancer data in order to predict which one of the three specific treatments a patient should receive. In all of these examples, the data analysis task is classification, where a model or classifier is constructed to predict categorical labels, such as “safe” or “risky” for the loan application data, “yes” or “no” label for the marketing data; or “treatment A”, “treatment B”, or “treatment C” for the medical data. These categories can be represented by discrete values, where the ordering among values has no meaning. For example, the value 1, 2, and 3 may be used to represent treatments A, B, and C, where there is no ordering implied among this group of treatment regimes.

Suppose that the marketing manager would like to predict how much a given customer will spend during a sale at computer shop. This data analysis task is an example of numeric prediction, where the model constructed predicts a continuous values function, or ordered value, as opposed to a categorical label. This model is a predictor. Regression analysis is a statistical methodology that is most often used for numeric prediction, hence the two terms are often used synonymously. For simplicity, when there is no ambiguity, we will use the shortened term of prediction to refer to numeric prediction.

The classification is the task of assigning objects to one of several predefined categories, and is one of the essential processes contained in the data mining. There are two forms of data analysis that can be used to extract models, whether describing data classes or to predict future data trends (Fayyad *et al*., 1996). Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions.

Basic technique for data classification consist of decision tree classifiers, Bayesian classifiers, Bayesian belief networks, rule-based classifiers, classification based on association rule mining, Back propagation classifier, support vector machine, k-nearest neighbors classifiers, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic techniques. Methods for prediction, including linear regression, non-linear regression, and other regression based models.

This research focused on classification problem, and selects four basic classification techniques to compare with proposed technique, implemented in text classification problem. These four basic text classification techniques are as follows:

(i). Bayesian classifiers (Domingos & Pazzani, 1997; Duda *et al*., 2000; Langley *et al*., 1992; Ordonez & Pitchaimalai, 2010; Rish, 2001)

(ii). K-Nearest Neighbor classifiers (Dasarathy, 1991; Duda *et al*., 2000; S. Jiang *et al*., 2012; Qiao et al., 2010)

(iii). Rocchio classifier (specific for text classifier) (Miao & Kamel, 2011; Rocchio, 1971)

(iv). Support vector machines (Boser *et al*., 1992; Cortes & Vapnik, 1995; Joachims, 1998; Pan *et al*., 2012; Scholkopf *et al*., 1999; Sullivan & Luke, 2007; Tong & Koller, 2002; Vapnik, 1998; Yu *et al*., 2003)

Each technique typically suits a problem better than others (Fayyad *et al.*, 1996). Thus, there is no universal data-mining method, and choosing a particular algorithm for a particular application is something of an art. In practice, a large portion of the application effort can go into properly formulating the problem (asking the right question) rather than into optimizing the algorithmic details of a particular data-mining method (Langley & Simon, 1995).

1.3 How does classification work?

Data classification is a two-step process (learning step and classification step). The first step that is the learning step, where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels.

A tuples, X, is represented by n-dimensional attribute vector, $X = \{x_1, x_2, \dots, x_n\}$, depicting n measurements made on tuple from n database attributes, respectively, $A_{1,} A_2, \dots, A_n$. Each tuple, X, is assumed to belong to a predefined class as determined by another database attribute called the class label attribute. The class label attribute is discrete valued and unordered. It is categorical in that each value serves as a category or class. The individual tuples making up the training set are referred to as training tuples and are selected from database under analysis. In the context of classification, data tuples can be referred to as samples, examples, instances, data points, or objects.

Because of the class label of each training tuple is provided, this step is also known as **supervised learning**. It contrasts with **unsupervised learning** (or clustering), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.

In the second step, the model is used for classification. A test set is used, made up of test tuples and their associated class labels. These tuples are randomly selected from the general data set. They are independent of the training tuples, meaning that they are not used to construct the classifier. In other word, tuples in the test set must be different from the tuples in the training set.

Classification methods can be compared and evaluated according to the following criteria, (i). Accuracy: The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data. Similarly, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

(ii). Speed: This refers to the computational costs involved in generating and using the given classifier or predictor.

(iii). Robustness: This is the ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.

(iv). Scalability: This refers to the ability to construct the classifier or predictor efficiently given large amounts of data.

(v). Interpretability: This refers to the level of understanding and insight that is provided by the classifier or predictor. Interpretability is subjective and therefore more difficult to assess

1.4 Problem Statement

In 1999, the concept of soft set theory as a mathematical tool for dealing with uncertainties has initiated by (D. Molodtsov, 1999), which has been further developed by (P. K. Maji *et al*., 2003). The soft set theory is different from traditional tools for dealing with uncertainties, and further it is free from the inadequacy of the parameterization tools of those theories (D. A. Molodtsov, 2004). The soft set theory has a rich potential for applications in several directions, few of which had been shown by Molodtsov in his pioneer work (D. Molodtsov, 1999).

At present, work on the soft set theory is progressing rapidly both in theoretical models and applications. As for practical applications of soft set theory, great progress has been achieved. The soft set theory can be applied to solve the decision-making problem (F. Feng *et al*., 2010, 2012; P. K. Maji *et al*., 2002; Roy & Maji, 2007), parameter reduction (Herawan *et al*., 2009; Ma et al., 2011), data clustering (Qin, Ma, Zain, *et al*., 2012), data analysis under incomplete information (Qin, Ma, Herawan, *et al*., 2012; Zou & Xiao, 2008), the combined forecasting (Xiao *et al*., 2009), and association rules mining (Herawan & Deris, 2010).

An example of the application of soft set theory for classification is proposed by (Mushrif *et al*., 2006). They used the soft set theory to classify images texture based on application soft set theory on decision-making problem. A soft set classifier based on similarity measure between the two generalized fuzzy soft sets has reported by (Majumdar & Samanta, 2010). In their work, they provided an example on how the similarity between the two generalized fuzzy soft sets used to detect whether an ill person is suffering from a certain disease.

Although both methods are quite successful for classification, low accuracy and efficiency when applied to text classification is the problem. The writing of this thesis has a purpose to propose a new approach on classification problem based on hybrid fuzzy soft set theory and supervised fuzzy c-means. This new approach is expected to improve the accuracy and the efficiency of classification in text classification problem.

1.5 Research Objectives

The objectives of this research are:

(i). To propose new classification technique based on hybrid fuzzy soft set theory and fuzzy c-means.

(ii). To develop an algorithm based on the proposed technique as in (a).

(iii). Applying the algorithm that develop in (b) on text classification problem.

(iv). To compare the algorithm with the existing algorithm based on efficiency and accuracy performance metrics.

1.6 Contributions

The main contributions of this study are in the area of data mining, the detail of these contributions is as follows:

(i). Extend the area application of soft set theory. The study has introduced a new algorithm for classification based on fuzzy soft set theory.

(ii). Introduce a new algorithm of classification for text classification problem. Applying the proposed algorithm to classify text document that has performance outperform as compare to the previous soft set classifiers and the classic text classifiers, based on efficiency and accuracy performance metrics.

(iii). Introduce a new hybrid algorithm of classification. The proposed algorithm is a hybrid fuzzy algorithm, which is consist of fuzzy soft set theory and supervised fuzzy c-means.

1.7 Research Scope

This study focus on developing the new approach to classify text document based on hybrid fuzzy soft set theory and Fuzzy C-means. Test case will be done using two well-known datasets that are the Reuter-21578 dataset for unevenly distributed dataset, and the 20 Newsgroups for evenly distributed dataset. Comparison will be done on the two groups of classifier. The first group will be used to compare the proposed algorithm with the other two soft set classifiers such as soft set classifier based on decision making-problem and soft set classifier based on similarity between two fuzzy soft sets. The second group will be used to compare the proposed algorithm with the four classic text classifiers, such as k-NN, Rocchio, Bayesian, and Support Vector Machine (SVM).

1.8 Thesis Organization

The thesis is organized into six different chapters. Chapter 1 provides the background and describes what motivated the researcher to introduce the new algorithm for text classification using soft set theory. Chapter 2 will explains the foundations of basic theory of soft set, fuzzy soft set, and text classification. Next, Chapter 3 will describes the new algorithm to classify text document based on fuzzy soft set theory and supervise fuzzy c-means. After that, Chapter 4 will reports the experimental results and discussion, which then tabulate and compare its findings to other research work. Finally, Chapter 5 will conclude and propose future work.

1.9 Chapter Summary

Recent advances in information technology have led to significant changes in today's world. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can assist us in transforming the data into more useful information and knowledge. The classification is the task of assigning objects to one of several predefined categories, and is one of the essential processes contained in the data mining. There are two forms of data analysis that can be used to extract models, whether describing data classes or to predict future data trends. Although classic methods are quite successful for classification, low accuracy and efficiency when applied to text classification is the problem. Objective of this research is to propose new classification technique based on hybrid fuzzy soft set theory and fuzzy c-means.

Some important terms related to this study include the following:

(i). **Data mining** is a process to extracting or “mining” knowledge from large amounts of data.

(ii). **Knowledge Discovery from Data** (KDD) is the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.

(iii). **Classification** is task of assigning objects to one of several predefined categories, and is one of the essential processes contained in the data mining. There are two models of classification, (a) classification model when the model is used to predict categorical labels, (b) prediction model when the model is used to predict a numerical.

(iv). **Supervised learning** is a learning process when the class label of each training tuple is provided, otherwise is **unsupervised learning**.

(v). **Soft set theory** is as a theory proposed by Molodtsov to deal with uncertainty problem that work with binary features.

(vi). **Fuzzy soft set theory** is a extended version of soft set theory to work with fuzzy number of features.

(vii). **Fuzzy c-means** is a data mining technique to data clustering.

CHAPTER 2

CLASSIFICATION AND SOFT SET THEORY

This chapter describes some basic theories, which will be used as a basis for classification proposed in this research. This includes soft-set theory, classic classification based on soft set theory, fuzzy set theory, fuzzy soft set theory, and fuzzy C-means.

2.1 Introduction

Machine learning, knowledge discovery in databases (KDD) and data mining are three terms that often appear associated with data processing and classification. They have similarities and differences. The similarities between them relate to the two fundamental facts:

(i). All of them develop methods and procedures to process data, and

(ii). Any data processing algorithm or procedure may belong to any.

The differences are in the different perspectives. The difference in perspectives does not affect the procedures but it affects the choice between them in the interpretation of concepts and results (Mirkin, 2011).

REFERENCES

Airoldi, E. M., Cohen, W. W., & Feinberg, S. E. (2004). Bayesian models for frequent terms in text. *In Proceedings of the CSNA & INTERFACE Annual Meetings*.

Aktas, H., & Cagman, N. (2007). Soft sets and soft groups. *Information Sciences*, *177*(13), 2726–2735.

Ali, M. I., Feng, F., Liu, X., Min, W. K., & Shabir, M. (2009). On some new operations in soft set theory. *Computers and Mathematics with Applications*, *57*(9), 1547–1553. Retrieved from http://www.sciencedirect.com/science/article/pii/S0898122108006871

Ali, S., & Smith, K. A. (2006). On learning algorithm selection for classification. *Applied Soft Computing*, *6*(2), 119–138. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V.

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, *1*, 113–141. JMLR.org. Retrieved from http://dx.doi.org/10.1162/15324430152733133

Aly, M. (2005). *Survey on Multiclass Classification Methods*. Technical report, California Institute of Technology, California, USA.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. (Be. Baeza-Yates, R A and Ribeiro-Neto, Ed.) (p. 513). New York: Addison Wesley. Retrieved from http://web.simmons.edu/~benoit/LIS466/Baeza-Yateschap01.pdf

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191–203. Retrieved from http://www.sciencedirect.com/science/article/pii/0098300484900207

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144–152). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/130385.130401

Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Jackel, L. D., LeCun, Y., *et al*. (1994). Comparison of classifier methods: a case study in handwritten digit recognition. *Pattern Recognition, 1994. Vol. 2 - Conference B: Computer*

Vision Image Processing., Proceedings of the 12th IAPR International. Conference on (Vol. 2, pp. 77 –82 vol.2).

Britanica, C. (2013). Data mining. *Encyclopedia Britannica Online*. Retrieved January 10, 2013, from http://global.britannica.com/EBchecked/topic/1056150/data-mining

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining Knowledge Discovery*, *2*(2), 121–167. Hingham, MA, USA: Kluwer Academic Publishers. Retrieved from http://dx.doi.org/10.1023/A:1009715923555

Cagman, N, Enginoglu, S., & Citak, F. (2011). Fuzzy Soft Set Theory and Its Applications. *Iranian Journal of Fuzzy Systems*, *8*(3), 137–147. Retrieved from http://journals.usb.ac.ir/Fuzzy/en-us/JournalNumbers/Articles58/

Cagman, Naim, & Enginoglu, S. (2010a). Soft set theory and uni-int decision making. *European Journal of Operational Research*, *207*(2), 848–855. Retrieved from http://www.sciencedirect.com/science/article/pii/S0377221710003589

Cagman, Naim, & Enginoglu, S. (2010b). Soft matrix theory and its decision making. *Computers and Mathematics with Applications*, *59*(10), 3308–3314. Retrieved from http://www.sciencedirect.com/science/article/pii/S0898122110001914

De Campos, L. M., & Romero, A. E. (2009). Bayesian network models for hierarchical text classification from a thesaurus. *International Journal of Approximate Reasoning*, *50*(7), 932–944. Retrieved from http://www.sciencedirect.com/science/article/pii/S0888613X08001746

Chen, D., Tsang, E. C. C., Yeung, D. S., & Wang, X. (2005). The parameterization reduction of soft sets and its applications. *Computers and Mathematics with Applications*, *49*(5-6), 757–763. Tarrytown, NY, USA: Pergamon Press, Inc.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naive Bayes. *Expert Systems with Applications*, *36*(3, Part 1), 5432–5435. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417408003564

Connor, M., & Kumar, P. (2010). Fast Construction of k-Nearest Neighbor Graphs for Point Clouds. *Visualization and Computer Graphics, IEEE Transactions on*, *16*(4), 599–608.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. Hingham, MA, USA: Kluwer Academic Publishers. Retrieved from http://dx.doi.org/10.1023/A:1022627411411

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, *13*(1), 21–27.

Cunningham, P., & Delany, S. J. (2007). *k-Nearest Neighbour Classifiers*. University College Dublin and Dublin Institute of Technology.

Dasarathy, B. V. (1991). *Nearest neighbor (NN) norms: nn pattern classification techniques*. (B. V Dasarathy, Ed.). IEEE Computer Society Press.

Dietterich, T. G., & Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *J. Artif. Int. Res.*, *2*(1), 263–286. USA: AI Access Foundation. Retrieved from http://dl.acm.org/citation.cfm?id=1622826.1622834

Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, *29*(2-3), 103–130. Hingham, MA, USA: Kluwer Academic Publishers. Retrieved from http://dx.doi.org/10.1023/A:1007413511361

Douglas, D. A., Covington, M. A., Covington, M. M., & Covington, C. A. (2009). *Dictionary of Computer and Internet Terms, Tenth Edition* (10th ed.). New York: Barron's Educational Series, Inc.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification (2nd Edition)*. Wiley-Interscience.

Dumais, S. T., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In G. Gardarin, J. C. French, N. Pissinou, K. Makki, & L. Bouganim (Eds.), *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management* (pp. 148–155). Bethesda, US: ACM Press, New York, US. Retrieved from http://robotics.stanford.edu/users/sahami/papers-dir/cikm98.pdf

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, *3*, 32–57.

Eyheramendy, S., Lewis, D. D., & Madigan, D. (2003). On the Naive Bayes Model for Text Categorization. *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37–54. American Association for Artificial Intelligence. Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230

Feng, F., Jun, Y. B., Liu, X., & Li, L. (2010). An adjustable approach to fuzzy soft set based decision making. *Journal of Computational and Applied Mathematics*, *234*(1), 10–20.

Feng, F., Li, Y., & Cagman, N. (2012). Generalized uni-int decision making schemes based on choice value soft sets. *European Journal of Operational Research*, *220*(1), 162–170. Retrieved from http://www.sciencedirect.com/science/article/pii/S0377221712000355

Feng, G., Guo, J., Jing, B.-Y., & Hao, L. (2012). A Bayesian feature selection paradigm for text classification. *Information Processing and Management*, *48*(2), 283–302. Retrieved from http://www.sciencedirect.com/science/article/pii/S0306457311000811

Friedman, J. H. (1996). *Another approach to polychotomous classification*. Retrieved from http://www-stat.stanford.edu/~jhf/ftp/poly.ps.Z

Garcia, V., Debreuve, E., & Barlaud, M. (2008). Fast k nearest neighbor search using GPU. *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on* (pp. 1–6).

Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hechenbichler, K., & Schliep, K. (2006). Weighted k-nearest-neighbor techniques and ordinal classification. *Discussion Paper 399, SFB 386*.

Herawan, T., & Deris, M. M. (2010). A soft set approach for association rules mining. *Knowledge-Based Systems, In Press*, , -.

Herawan, T., Rose, A. N. M., & Mat Deris, M. (2009). Soft Set Theoretic Approach for Dimensionality Reduction. In D. Slezak, T. Kim, Y. Zhang, J. Ma, & K. Chung (Eds.), *Database Theory and Application* (Vol. 64, pp. 171–178). Springer Berlin Heidelberg.

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, *13*(2), 415–425.

Hu, R. (2011). *Active Learning for Text Classification*. Dublin Institute of Technology.

Jiang, J.-Y. (2011). *Feature Reduction and Multi-label Classification Approaches for Document Data*. National Sun Yat-sen University.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, *39*(1), 1503–1509. Retrieved September 26, 2012, from http://www.sciencedirect.com/science/article/pii/S0957417411011511

Joachims, T. (1998). Text Categorization with Support Vector Machine. *Proceedings of the European Conference on Machine Learning*. Springer-Verlag.

Katakis, I., Tsoumakas, G., & Vlahavas, I. (2008). Multilabel Text Classification for Automated Tag Suggestion. *Proceedings of the ECML/PKDD 2008 Discovery Challenge*.

Kim, S.-B., Han, K.-S., Rim, H.-C., & Myaeng, S. H. (2006). Some Effective Techniques for Naive Bayes Text Classification. *Knowledge and Data Engineering, IEEE Transactions on*, *18*(11), 1457–1466.

Koller, D., & Sahami, M. (1997). Hierarchically Classifying Documents Using Very Few Words. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 170–178). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=645526.657130

Kong, Z., Gao, L., & Wang, L. (2009). Comment on "A fuzzy soft set theoretic approach to decision making problems". *Journal of Computational and Applied Mathematics*, *223*(2), 540–542. Retrieved from http://www.sciencedirect.com/science/article/pii/S0377042708000162

Kong, Z., Gao, L., Wang, L., & Li, S. (2008). The normal parameter reduction of soft sets and its algorithm. *Computers and Mathematics with Applications*, *56*(12), 3029–3037. Retrieved from http://www.sciencedirect.com/science/article/pii/S0898122108004483

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. *Proceedings of the tenth national conference on Artificial intelligence* (pp. 223–228). San Jose, California: AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=1867135.1867170

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communication ACM*, *38*(11), 54–64. New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/219717.219768

Larkey, L. S., & Croft, W. B. (1996). Combining classifiers in text categorization. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 289–297). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/243199.243276

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of ECML-98, 10th European Conference on Machine Learning* (pp. 4–15). Chemnitz, DE: Springer Verlag, Heidelberg, DE. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.8397

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, *48*(2), 354–368. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167923609002097

Lopes, C., Cortez, P., Sousa, P., Rocha, M., & Rio, M. (2011). Symbiotic filtering for spam email detection. *Expert Systems with Applications*, *38*(8), 9365–9372. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417411003228

Ma, X., Sulaiman, N., Qin, H., Herawan, T., & Zain, J. M. (2011). A new efficient normal parameter reduction algorithm of soft sets. *Computers and Mathematics with Applications*, *62*(2), 588–598. Retrieved from http://www.sciencedirect.com/science/article/pii/S0898122111004366

Madjarov, G., Gjorgjevikj, D., & Deroski, S. (2012). Two stage architecture for multi-label learning. *Pattern Recognition*, *45*(3), 1019–1034. Retrieved from http://www.sciencedirect.com/science/article/pii/S0031320311003487

Maji, P., Biswas, R., & Roy, A. (2001). Fuzzy soft sets. *Journal of Fuzzy Mathematics*, *9(3)*, 589–602.

Maji, P. K., Biswas, R., & Roy, A. R. (2003). Soft set theory. *Computers and Mathematics with Applications*, *45*(4-5), 555–562. Retrieved from http://www.sciencedirect.com/science/article/pii/S0898122103000166

Maji, P. K., Roy, A. R., & Biswas, R. (2002). An application of soft sets in a decision making problem. *Computers and Mathematics with Applications*, *44*(8-9), 1077–1083. Retrieved from http://www.sciencedirect.com/science/article/pii/S089812210200216X

Majumdar, P., & Samanta, S. K. (2008). Similarity Measure Of Soft Sets. *New Mathematics and Natural Computation (NMNC)*, *4*(01), 1–12.

Majumdar, P., & Samanta, S. K. (2010). Generalised fuzzy soft sets. *Computers & Mathematics with Applications*, *59*(4), 1425–1432. Tarrytown, NY, USA: Pergamon Press, Inc.

Majumdar, P., & Samanta, S. K. (2011). On Similarity Measures of Fuzzy Soft Sets. *International Journal Advance Soft Compututing Application*, *3*(2). Retrieved from http://www.i-csrs.org/Volumes/ijasca/vol.3/vol.3.2.4.July.11.pdf

Manning, C. D., Raghavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval (on line edition)*. Cambridge University Press, Cambridge, England.

McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *Workshop on Learning for Text Categorization* (pp. 41–48). AAAI Press.

Miao, Y., & Kamel, M. (2011). Pairwise optimized Rocchio algorithm for text categorization. *Pattern Recognition Letters*, *32*(2), 375–382. Retrieved September 26, 2012, from http://www.sciencedirect.com/science/article/pii/S0167865510003223

Mirkin, B. (2011). Data analysis, mathematical statistics, machine learning, data mining: Similarities and differences. *Advanced Computer Science and Information System (ICACSIS), 2011 International Conference on* (pp. 1–8).

Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill.

Molodtsov, D. (1999). Soft set theory: First results. *Computers and Mathematics with Applications, 37*(4-5), 19–31. Retrieved from http://www.sciencedirect.com/science/article/pii/S0898122199000565

Molodtsov, D. A. (2004). *The theory of soft sets*. Moscow: URSS Publishers.

Morelos-Zaragoza, R. H. (2006). *The art of error correcting coding*. Wiley Interscience.

Mushrif, M., Sengupta, S., & Ray, A. (2006). Texture Classification Using a Novel, Soft-Set Theory Based Classification Algorithm. In P. Narayanan, S. Nayar, & H.-Y. Shum (Eds.), *Computer Vision - ACCV 2006* (Vol. 3851, pp. 246–254). Springer Berlin / Heidelberg.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. *Mach. Learn.*, *39*(2-3), 103–134. Hingham, MA, USA: Kluwer Academic Publishers. Retrieved from http://dx.doi.org/10.1023/A:1007692713085

Nogueira, T. M., Rezende, S. O., & Camargo, H. A. (2010). On the use of fuzzy rules to text document classification. *Hybrid Intelligent Systems (HIS), 2010 10th International Conference on* (pp. 19–24).

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques* (p. 180). Springer-Verlag Berlin Heidelberg.

Ordonez, C., & Pitchaimalai, S. K. (2010). Bayesian Classifiers Programmed in SQL. *Knowledge and Data Engineering, IEEE Transactions on*, *22*(1), 139–144.

Pan, S., Iplikci, S., Warwick, K., & Aziz, T. Z. (2012). Parkinson's Disease tremor classification: A comparison between Support Vector Machines and neural networks. *Expert Systems with Applications*, *39*(12), 10764–10771. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417412004629

Pawlak, Z. (1982). Rough sets. *International Journal of Parallel Programming*, *11*(5), 341–356. Springer Netherlands. Retrieved from http://dx.doi.org/10.1007/BF01001956

Pei, D., & Miao, D. (2005). From soft sets to information systems. *Granular Computing, 2005 IEEE International Conference on* (Vol. 2, pp. 617 – 621 Vol. 2).

Perez-Diaz, N., Ruano-Ordas, D., Mendez, J. R., Galvez, J. F., & Fdez-Riverola, F. (2012). Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification. *Applied Soft Computing*, *12*(11), 3671–3682. Retrieved from http://www.sciencedirect.com/science/article/pii/S1568494612002748

Porter, M. F. (1997). Readings in information retrieval. In K. Sparck Jones & P. Willett (Eds.), (pp. 313–316). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=275537.275705

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Retrieved from http://snowball.tartarus.org/texts/introduction.html

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. Retrieved from http://www.sciencedirect.com/science/article/pii/S1751157709000108

Qiao, Y.-L., Lu, Z.-M., Pan, J.-S., & Sun, S.-H. (2010). Fast k-nearest neighbor search algorithm based on pyramid structure of wavelet transform and its application to texture classification. *Digital Signal Processing*, *20*(3), 837–845. Retrieved from http://www.sciencedirect.com/science/article/pii/S1051200409001894

Qin, H., Ma, X., Herawan, T., & Zain, J. M. (2012). DFIS: A novel data filling approach for an incomplete soft set. *International Journal of Applied Mathematics and Computer Science*, *22*(4), 817–828. Retrieved from http://www.amcs.uz.zgora.pl/?action=paper&paper=651

Qin, H., Ma, X., Zain, J. M., & Herawan, T. (2012). A novel soft set approach in selecting clustering attribute. *Knowledge-Based Systems*, (0), -. Retrieved from http://www.sciencedirect.com/science/article/pii/S0950705112001712

Raghavan, V. V, & Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, *37*(5), 279–287.

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *In Proceedings of the Twentieth International Conference on Machine Learning* (pp. 616–623).

Rifkin, R., & Klautau, A. (2004). In Defense of One-Vs-All Classification. *J. Mach. Learn. Res.*, *5*, 101–141. JMLR.org. Retrieved from http://dl.acm.org/citation.cfm?id=1005332.1005336

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI-01 workshop on "Empirical Methods in AI"*. Retrieved from http://www.intellektik.informatik.tu-darmstadt.de/~tom/IJCAI01/Rish.pdf

Rocchio, J. (1971). Relevance feedback in information retrieval. In Gerard Salton (Ed.), *The Smart Retrieval System-Experiments in Automatic Document Processing* (pp. 313–323). Prentice Hall.

Roy, A. R., & Maji, P. K. (2007). A fuzzy soft set theoretic approach to decision making problems. *Journal of Computational and Applied Mathematics*, *203*(2), 412–418. Retrieved from http://www.sciencedirect.com/science/article/pii/S0377042706002160

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian Approach to Filtering Junk E-Mail. *Learning for Text Categorization: Papers from the 1998 Workshop*. Madison, Wisconsin: AAAI Technical Report WS-98-05. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.1254

Salton, G, Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communication ACM*, *18*(11), 613–620. New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/361219.361220

Salton, Gerard, & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, *24*(5), 513–523. Tarrytown, NY, USA: Pergamon Press, Inc. Retrieved from http://dx.doi.org/10.1016/0306-4573(88)90021-0

Schneider, K.-M. (2003). A comparison of event models for Naive Bayes anti-spam e-mail filtering. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1* (pp. 307–314). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.3115/1067807.1067848

Schneider, K.-M. (2004). On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification. In J. Vicedo, P. Martinez-Barco, R. Munoz, & M. Saiz Noeda (Eds.), *Advances in Natural Language Processing* (Vol. 3230, pp. 474–485). Springer Berlin / Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-30228-5_42

Scholkopf, B., Bartlett, P., Smola, A., & Williamson, R. (1999). Shrinking the tube: a new support vector regression algorithm. *Proceedings of the 1998 conference on Advances in neural information processing systems II* (pp. 330–336). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=340534.340663

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/505282.505283

Singh, S. R., Murthy, H. A., Gonsalves, T. A., Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature Selection for Text Classification Based on Gini Coefficient of Inequality. *Proceedings of the Fourth International Workshop on Feature Selection in Data Mining, June 21st, 2010, Hyderabad, India* (Vol. 10, pp. 76–85).

Soucy, P., & Mineau, G. W. (2001). A Simple KNN Algorithm for Text Categorization. In N. Cercone, T. Y. Lin, & X. Wu (Eds.), *Proceedings of ICDM01 IEEE International Conference on Data Mining* (pp. 647–648). IEEE Computer Society Press, Los Alamitos, US. Retrieved from http://portal.acm.org/citation.cfm?id=645496.757723&coll=GUIDE&dl=GUIDE&CFID=91696714&CFTOKEN=59964605#

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680.

Sullivan, K. M., & Luke, S. (2007). Evolving kernels for support vector machine classification. *Proceedings of the 9th annual conference on Genetic and* *evolutionary computation* (pp. 1702–1707). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1276958.1277292

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Addison-Wesley Companion Book Site.

Tan, S. (2005). Neighbor-weighted K-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications*, *28*(4), 667–671. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417404001708

Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, *2*, 45–66. JMLR.org. Retrieved from http://dx.doi.org/10.1162/153244302760185243

Tsoumakas, G., & Katakis, I. (2009). Multi-Label Classification: An Overview. *Database Technologies: Concepts, Methodologies, Tools, and Applications,* 309–319.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. *In Data Mining and Knowledge Discovery Handbook* (pp. 667–685).

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

Vidhya, K. A., & Aghila, G. (2010). A Survey of Naive Bayes Machine Learning approach in Text Document Classification. *International Journal of Computer Science and Information Security*, *7*(2), 206–211.

Wan, C. H., Lee, L. H., Rajkumar, R., & Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. *Expert Systems with Applications*, *39*(15), 11880–11888. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417412003120

Webopedia. (2012). Structure Data. *Webopedia*. Retrieved December 12, 2012, from http://www.webopedia.com/TERM/S/structured_data.html

Widyantoro, D. H., & Yen, J. (2000). A fuzzy similarity approach in text classification task. *Fuzzy Systems, 2000. FUZZ IEEE 2000. The Ninth IEEE International Conference on* (Vol. 2, pp. 653 –658 vol.2).

Wu, D., Vapnik, V. N., & Bank, R. (1998). Support Vector Machine for Text Categorization. *Learning*, 1–16. State University of New York at Buffalo. Retrieved from http://portal.acm.org/citation.cfm?id=1048295

Xiao, Z., Gong, K., & Zou, Y. (2009). A combined forecasting approach based on fuzzy soft sets. *Journal of Computational and Applied Mathematics, 228*(1), 326–333. Retrieved from http://www.sciencedirect.com/science/article/pii/S0377042708005001

Yang, X., Lin, T. Y., Yang, J., Li, Y., & Yu, D. (2009). Combination of interval-valued fuzzy set and soft set. *Computers and Mathematics with Applications*, *58*(3), 521–527. Retrieved from http://www.sciencedirect.com/science/article/pii/S0898122109003228

Yang, X., Yu, D., Yang, J., & Wu, C. (2007). Generalization of Soft Set Theory: From Crisp to Fuzzy Case. In B.-Y. Cao (Ed.), *Fuzzy Information and Engineering* (pp. 345–354). Springer Berlin Heidelberg.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 42–49). New York, NY, USA: ACM.

Yang, Y., & Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 412–420). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=645526.657137

Yu, H., Yang, J., & Han, J. (2003). Classifying large data sets using SVMs with hierarchical clusters. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 306–315). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/956750.956786

Yukinawa, N., Oba, S., Kato, K., & Ishii, S. (2009). Optimal Aggregation of Binary Classifiers for Multiclass Cancer Diagnosis Using Gene Expression Profiles. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, *6*(2), 333–343. Los Alamitos, CA, USA: IEEE Computer Society Press. Retrieved from http://dx.doi.org/10.1109/TCBB.2007.70239

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353.

Zou, Y., & Xiao, Z. (2008). Data analysis approaches of soft sets under incomplete information. *Knowledge-Based Systems*, *21*(8), 941–945. Retrieved from http://www.sciencedirect.com/science/article/pii/S0950705108001056