

Text Classification of Database of Genotypes and Phenotypes in Heart, Lung and Blood Studies

Dissertation

*Submitted in partial fulfillment of the requirement for the degree of
Master of Technology in Computer Engineering*

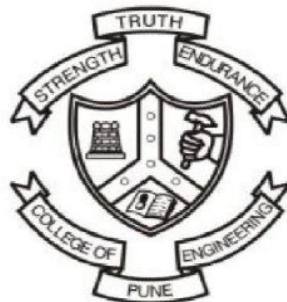
By

Suresh S. Kolekar

MIS No: 121222003

Under the guidance of

Prof. S. S. Kumbhar



Department of Computer Engineering and Information Technology

College of Engineering, Pune

Pune – 411005

June, 2014

**DEPARTMENT OF COMPUTER ENGINEERING AND
INFORMATION TECHNOLOGY,
COLLEGE OF ENGINEERING, PUNE**

CERTIFICATE

This is to certify that the dissertation titled

Text Classification of Database of Genotypes and Phenotypes in Heart, Lung and Blood Studies.

has been successfully completed

By

Suresh S Kolekar

MIS No: 121222003

and is approved for the partial fulfillment of the requirements for the degree of

Master of Technology, Computer Engineering

Prof. S. S. Kumbhar
Project Guide,
Department of Computer Engineering
and Information Technology,
College of Engineering, Pune,
Shivaji Nagar, Pune-411005.

Dr. J. V. Aghav
Head,
Department of Computer Engineering
and Information Technology,
College of Engineering, Pune,
Shivaji Nagar, Pune-411005.

June 2014

Acknowledgments

I express my sincere gratitude towards my guide Professor S. S. Kumbhar for his constant help, encouragement and inspiration throughout the project work also for providing me infrastructural facilities to work in. I also like to convey my sincere gratitude to Dr. J. V. Aghav (HOD), all faculty members and staff of Department of Computer Engineering and Information Technology, College of Engineering, Pune for all necessary cooperation in the accomplishment of dissertation. Last but not least, I would like to thank my family and friends, who have been a source of encouragement and inspiration throughout the duration of the project.

Suresh Sudam Kolekar

College of Engineering, Pune

Abstract

The database of genotypes and phenotypes (dbGaP) is the new database to store and distribute data from studies of genome wide association. dbGaP launch by National Library of Medicine (NLM) which is part of National Institutes of Health (NIH). Searching relevant studies of particular interest accurately and completely is challenging task due to keyword based search method of dbGaP Entrez system. For given queries, the dbGaP retrieval system returns several studies that are unrelated, and it is very difficult to find how particular studies are retrieved and why they come out in a particular sequence. Thus, users have to evaluate every study description carefully to find relevant studies, which is time consuming task.

Text mining is emerging research field which enable users to extract useful information from text documents and deals with retrieval, classification, clustering and machine learning techniques to classify different text document.

In this research, an empirical approach is proposed and implemented with K-nearest neighbor (KNN) and Radom forest machine learning algorithms to classify dbGaP study text in heart, lung and blood studies. It is evident from results that this text based classification outperforms conventional keyword based search of document retrieval system provided by dbGaP.

Keywords: Bioinformatics, Data Mining, Text Mining, database of Genotypes and Phenotypes.

Contents

Certificate	I
Acknowledgement	ii
Abstract	iii
List of Figures	Vii
List of Tables	Viii
1. Introduction	1
1.1 Introduction	1
1.2 Challenges in dbGaP Study Text Retrieval	2
1.3 Text Mining	2
1.4 Motivation	3
1.5 Problem Statement	3
1.6 Thesis Objective and Scope	3
1.7 Thesis Outline	3
2 Biological Prerequisite	4
2.1 Basic Biological System	5
2.2 Genetic Material	6
2.2.1 DNA	6

2.2.2	RNA	7
2.2.3	Gene	7
3	Literature Survey	9
3.1	National Center for Biotechnology Information	9
3.1.1	dbGaP	10
3.2	Challenging Issues in Existing System	10
3.2.1	Manual Performance Evaluation of dbGaP Search System	10
3.3	Supervised Learning	12
3.3.1	The K-Nearest Neighbor Algorithm	12
3.3.2	Decision Trees	13
3.3.3	Ensemble Learning	15
3.3.3.1	Bagging	15
3.3.3.2	Random Forest	16
3.3.3.3	Boosting	17
3.3.3.4	Stacking	17
4	Methodology	19
4.1	Overview of Text Classification	19
4.2	Document Preprocessing	20
4.2.1	Lexical Analysis	21
4.2.2	Stop Word Removal	21
4.2.3	Stemming	21
4.3	Weighting Scheme	21
4.4	Dataset Preparation	22
4.5	Classification Algorithm	22
4.5.1	K- Nearest Neighbor Algorithm	22

4.5.2	Random Forest Algorithm	23
4.6	Classification Performance Measures	24
5	Evaluations and Discussions	25
5.1	Dataset	25
5.2	System Requirement	25
5.3	Measures Used for Evaluation	26
5.4	Result of K-Nearest Neighbor Classifier	26
5.5	Result of Random Forest Classifier	27
5.6	Result Comparison	28
6	Conclusion and Future Scope	30
Appendix A		31
References		32

List of Figures

2.1	Cell Structure	5
2.2	DNA	7
2.3	Gene	8
3.1	Decision Tree	14
3.2	Random Forest	16
4.1	Flow of Text Classification	20
4.2	Flowchart of Random Forest Algorithm	24
5.1	Classification Model	26
5.2	Graphical Representation of performance result of KNN	27
5.3	Graphical Representation of performance result of RF	28

List of Tables

3.1	Label assigned to documents in dataset	10
3.2	Measurement definition of evaluation metrics	11
3.3	dbGaP keyword search result	12
5.1	Dataset	25
5.2	Performance result of KNN	27
5.3	Performance result of RF	27
5.4	Comparison table based on accuracy	28
5.5	Comparison table based on precision	28
5.6	Comparison table based on recall	29
5.7	Comparison table based on F-measure	29

Chapter 1

Introduction

1.1 Introduction

The database of genotypes and phenotypes (dbGaP) is the new database to store and distribute data from studies of genome wide association. dbGaP initiated by National Library of Medicine (NLM) which is part of National Institutes of Health (NIH). Genome wide association studies find relationship between particular genes and observable traits such as disease condition, weight and blood pressure. Relationship between phenotypes and genotypes gives information about genes that may be responsible for disease condition, which can be useful for better understanding the disease and for developing better diagnostic methods [2].

For the first time dbGaP, the database of Genotype and Phenotype, providing a central location for researcher to see all study documents and to analysis summaries of the measured variables in searchable web format.

The database contains phenotypic variables and statistical summaries of genetic information. Individual level data from dbGaP may be accessible if it is permitted by National Institutes of Health (NIH) Data Access Committee. The database is growing very fast. In 24 October 2013 dbGaP contained 402 studies and by 5 may 2014 there were 468 top- level studies.

1.2 Challenges in dbGaP Study Text Retrieval

As of 5 may 2014, 468 studies were available in dbGaP which include around 144716 phenotype variables. However, retrieving related studies correctly is become challenging issue, since phenotypic information of studies is stored in a non-standardized format. For given queries, the dbGaP retrieval system returns several studies that are unrelated, and it is very difficult to find how particular studies are retrieved and why they come out in a particular sequence. Thus, users have to evaluate every study description cautiously to determine relevant studies, which unnecessarily take lot of time when there are lot of studies to be retrieved. Text mining is the one of the popular research area in the field of automatic document retrieval system.

1.3 Text Mining

The age of information made it easy for humans to store huge amount of text documents. These are available on the internet, on corporate intranets and elsewhere. However, while amount of information is increasing day by day, but our ability to process and absorb this information remain constant.

Text mining is the process of finding unknown information from different text documents by automatic extraction of information. Text mining is also refers to extraction of interesting pattern from huge amount of text database for knowledge discovery.[16] Text mining applies analytical functions of data mining, natural language and information retrieval(IR) techniques.

Text mining is variation of data mining. The main difference between data mining and text mining is that in text mining the pattern are extracted from text of natural language rather than structural database.

Text mining, data mining and machine learning algorithms are in great demand in the field of bioinformatics. Text mining techniques applied to bioinformatics importantly involve methods like -

- **Classification** Text documents are arranged into groups of pre-labeled class. Learning schemes learn through training text documents and efficiency of these system is tested by using test text documents. Common algorithms include decision tree learning, naive Bayesian classification, nearest neighbour and neural network. This is called supervised learning.

- **Clustering** This is un-supervised learning method. Text documents here are unlabelled and inherent patterns in text are revealed through cluster formation. This can also be used as prior step for other text mining methods.

1.4 Motivation

While there is an increasing literature in the field of biomedical text categorization, a search of PubMed and Google Scholar revealed very less publications about text categorization applied to dbGaP. For the first time dbGaP, the database of Genotype and Phenotype, providing a central location for researcher to see all study documents and to analysis summaries of the measured variables in searchable web format. Searching related studies become challenging task due non standardization of phenotype variables. Due to this automatic document classification becomes tremendously important to provide accurate data retrieval, organize topics of interest for research, and streamline costs of data curation.

1.5 Problem Statement

The purpose of this research work is to improve the performance of dbGaP document retrieval system by utilizing supervised machine learning algorithms (Random forest and K- nearest neighbors) on dbGaP study text to identify heart, lung and blood studies.

1.6 Project Objective and Scope

- Evaluate the performance of dbGaP Entrez system which works on keyword based system
- Utilizing machine learning algorithm (KNN and RF) on dbGaP study text.
- Applying text preprocessing techniques to improve performance of classifier.
- Compare the performance KNN and RF classifier with existing dbGaP document retrieval system.

1.7 Project Outline

This thesis consists of six chapters. Chapter 1 presents introduction of thesis. Chapter 2 provides biological prerequisite for thesis. Chapter 3 presents detailed description of literature survey. In chapter 4, the methodology of thesis is discussed and presented. Chapter 5 gives comprehensive research description and result of classification models. Finally, chapter 6 presents the conclusion and future scope of research.

Chapter 02

Biological Prerequisite

Computational Biology- Computational biology includes the building and applying data analysis, theoretical strategies, computational simulation and mathematical modeling techniques to the study of biological, behavioral, and social systems.

There are several sub fields in biology that utilize computationally intensive methodologies that may expedite the processing of huge data, can improve accuracy of results etc.

- Computational Genomics
- Computational biomodelling
- Computational neuroscience
- Bio informatics

Bioinformatics- Bioinformatics deals with the study of methods which store, retrieve and analyze biological data. It extracts knowledge that is helpful in drug design and also development of software tools to process that knowledge. It also deals with databases, algorithms, information systems, artificial intelligence ,web technologies, information and computation theory, structural biology, soft computing, data mining, software engineering, image processing, , statistics, modeling and simulation, control and system theory , discrete mathematics and circuit theory. Following are sub branches within bioinformatics and computational biology:

- The development of tools and their implementation for efficient data access. These tools may also be used to manage big amount of data and retrieve necessary information.
- New algorithmic method's development aided with some new mathematical formulas. Statistical techniques to mine relationships between biological components. This basically involves retrieval of genetic sequences. Finding locus of gene prediction of structural and functional properties of proteins.

2.1 Basic Biological system

The human body is made up of trillions of cells which are building blocks of entire living things [13].

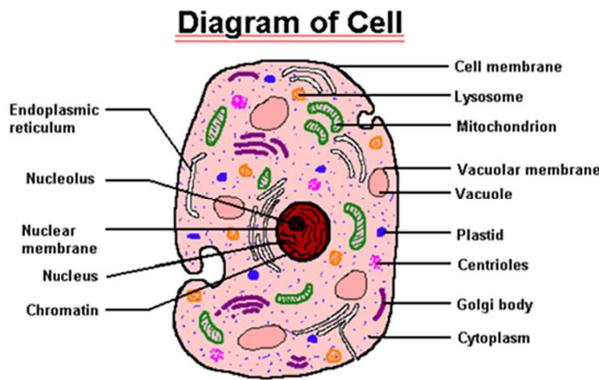


Figure 2.1: Cell Structure

Human cells made up of following parts:

- **Cytoplasm**

The cytoplasm contains cytosol which is jelly-like fluid and other structures that surround the nucleus.

- **Cytoskeleton**

The cytoskeleton is a network of long fibers that form the cell's structural framework. The cytoskeleton performs various functions, i.e., determining cell shape, involving in cell division, and allowing cells to move. Cytoskeleton directs the movement of organelles and other substances within cells.

- **Endoplasmic reticulum (ER)**

ER process and transport molecules to their particular destination which is created by the cell.

- **Golgi apparatus**

It packages molecules which are processed by the ER. It then transported out of the cell.

- **Lysosomes and peroxisomes**

These organelles are the recycling center of the cell. They absorb foreign bacteria that attack the cell, relieve the cell of toxic substances, and recycle worn-out cell components.

- **Mitochondria**

Mitochondria convert energy from food into a form that the cell can use. Mitochondria have their own genetic material, separate from the DNA in the nucleus, and can make copies of themselves.

- **Nucleus**

It serves as the cell's command center by sending directions to the cell to grow, mature, divide, or die. It is also house of DNA (deoxyribonucleic acid), the cell's hereditary material. It is surrounded by a membrane called the nuclear envelope, which protects the deoxyribonucleic acid and separates the nucleus from the rest of the cell.

- **Plasma membrane**

Plasma membrane disallows material to enter or leave as it separate cell from environment.

- **Ribosome**

Ribosome process the cell's genetic instructions to create proteins. Ribosomes can float freely in the cytoplasm or it can be connected to the endoplasmic reticulum.

2.2 Genetic Material

Genetic material found in the cytoplasm and nuclei. Genetic material plays a fundamental role in determining the nature and structure of a cell. For instance, the genetic material of a cell can be a part of a gene, deoxyribonucleic acid molecule or a genome of an organism.

2.2.1 DNA (Deoxyribonucleic acid)

DNA, or deoxyribonucleic acid, is the hereditary material in almost all organisms. Almost every cell in a person's body has the same DNA. Most DNA is found in the

cell nucleus (nuclear DNA), but a small amount of DNA can also be found in the mitochondria (mtDNA)

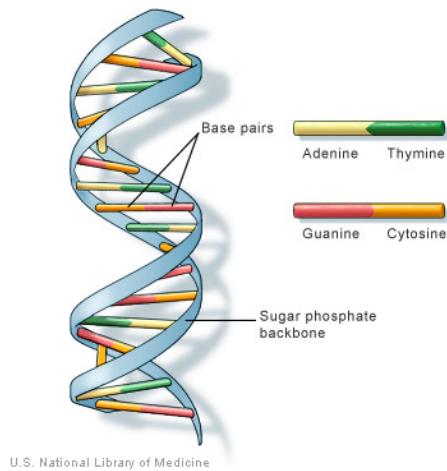


Figure 2.2 DNA

The information stored in DNA is contains four chemical bases:
adenine (A), guanine (G), cytosine (C), and thymine (T).

2.2.2 RNA (Ribonucleic acid)

It is single stranded molecule which takes part in protein synthesis

Types of RNA:

mRNA: This is also known as messenger RNA. It present in nucleus carries protein coding information during protein synthesis

tRNA: This is also called Transfer RNA . tRNA carries amino acids to site of protein synthesis

rRNA: This is also called ribosomal RNA. rRNA is structural component of ribosome
It activates site of protein synthesis

2.2.3 Gene

A gene made up of deoxyribonucleic acid which acts as instructions to form molecules which known by proteins.In humans, genes size vary from a few hundred DNA bases to more than 2 million bases.

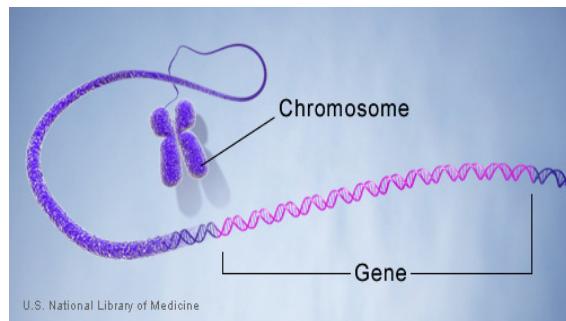


Figure 2.3 Gene

Everyone has two copies of each gene, one taken from each parent. All people have same genes, but a small number of genes are slightly different between people.

Chapter 03

Literature Survey

This chapter covers the prior work and related papers in the field of bioinformatics, data mining and text mining which are useful in defining basic idea of this work. First section of this chapter will cover the background and prior work in the field of bioinformatics, second will cover literature survey related to different machine learning algorithms which are useful in area of text mining.

3.1 National Center for Biotechnology Information (NCBI)

The National center for biotechnology information provides resource for molecular biology. It is multi-disciplinary research group formed by National Institutes of Health and the National Library of Medicine in 1998. The idea behind creation of NCBI was to understand molecular mechanisms that affect human health and disease. The goal of NCBI was to create and maintain public databases, design software to examine genomic data and to promote research in computational biology.

Due to widespread use of internet, NCBI became increasingly aware of pure biological research. This was evident as various specialized databases were created by the NCBI. Database of genotypes and phenotypes (dbGaP) created by NCBI to store and distribute data from studies of genome wide association [2].

3.1.1 dbGaP

The technological advancement and weakening prices for genotyping availing investigators new opportunities to do increasingly convoluted analyses of genetic associations alongside phenotypic and illness characteristics. The important candidates for such genome expansive association studies (GWAS) are continuing large-scale cohort and clinical studies that amassed affluent sets of phenotype data. To prop investigator admission to data from these initiatives at the Nationwide Institutes of Condition (NIH) and elsewhere, the Nationwide Center for Biotechnology Data (NCBI) has crafted a database of Genotypes and Phenotypes (dbGaP) alongside stable identifiers that make it probable for published studies to debate or quote the main data in a specific and uniform way.

To protect the privacy of study subjects, dbGaP accepts merely de-identified data and needs investigators to go across an approval procedure in order to admission individual-level phenotype and genotype datasets. Synopsis phenotype and genotype data, as well as discover documents, are obtainable lacking restriction [2].

3.2 Challenging Issues in Existing System

dbGaP study retrieval system returns several studies that are not related for given queries. Thus, users must evaluate every study description cautiously to determine relevant studies, which unnecessarily take lot of time when there are a lot of studies to be retrieved [6].

3.2.1 Manual Performance Evaluation of dbGaP Search System

Four hundred and Sixty eight studies were available in dbGaP on May 1, 2014. Each title and abstract was manually reviewed and annotated into heart, lung, blood, and other categories.

Here four different labels assigned to each documents. These labels are shown in table 3.1, and were assigned manually depending on its relevance to the document.

dbGaP : 468 studies			
Heart	Lung	Blood	Other
39	23	37	369

TABLE 3.1 LABELS ASSIGNED TO DOCUMENTS IN DATASET

Evaluation metrics used were accuracy, precision, recall, and F-measure .Table 3.2 represents measurements definition of evaluation metrics.

	Correct label	Incorrect label
Assigned label	True Positive(TP)	False Positive(FP)
Not assigned label	False negative(FN)	True Negative(TN)
Accuracy = $(TP+TN)/(TP+TN+FP+FN)$		
Precision = $TP/(TP+FP)$		
Recall = $TP/(TP+FN)$		
F-measure= $(2*Precision*Recall)/(Precision + Recall)$		

TABLE 3.2 MEASUREMENTS DEFINITION OF EVALUATION METRICS

- Evaluation Metrics for Heart Keyword

dbGaP Entrez system for heart query return 29 true positive,306 true negative,123 false positive and 10 false negative studies, which gives ,

Accuracy=0.72

Precision = 0.19.

Recall = 0.74.

F- Measure=0.30

- Evaluation Metrics of Lung Keyword

dbGaP Entrez system for lung query return 19 true positive,313 true negative,132 false positive and 04 false negative studies, which gives ,

Accuracy = 0.71.

Precision = 0.13.

Recall = 0.82.

F- Measure=0.22

- Evaluation Metrics of Blood Keyword

dbGaP Entrez system for blood query return 22 true positive,174 true negative,257 false positive and 15 false negative studies, which gives ,

Accuracy = 0.42

Precision = 0.08

Recall = 0.59

F-measure = 0.14

	Heart	Lung	Blood
Accuracy	0.72	0.71	0.42
Precision	0.19	0.13	0.08
Recall	0.74	0.84	0.59
F-measure	0.30	0.22	0.14

Table 3.3 dbGaP keyword Search result

Results of the manual keyword search method of dbGaP demonstrate the opportunity for improvement in accuracy, precision, recall, and F-measure.

3.3 Supervised Learning

Machine learning algorithms were categorized into two groups supervised learning algorithm and unsupervised learning algorithm. Supervised learning algorithm take a set of labeled training data, each with a feature vector and a class label. Most algorithms produce a classifier model that maps an object to a class label. Unsupervised learning algorithms use set of unlabeled objects and works without the class information of sample. In this work, the focus is on the application of supervised learning algorithms to classify the dbGaP text study in order to compare and evaluate the classification performance of these algorithms with performance of existing document retrieval system of dbGaP.

3.3.1 The k-Nearest Neighbor Algorithm

The k-nearest neighbor (KNN) algorithm [15] belongs to category of instance-based learners which is simple and one of important machine learning algorithms. Instance-based learners are also called lazy learner algorithm because it delays actual generalization process until classification is performed. There is no model building process. Instance-based learners do not abstract any information from the training data during the learning phase. Learning is merely a question of encapsulating the training data.

KNN is works based on principle that instances within dataset generally exist in close proximity to other instances within dataset that have similar properties[15].If the

objects are tagged with a classification label then objects are classified by a taking majority vote of their neighbors and it assigns to most common class amongst its k-nearest neighbors. K is small odd positive number and correct classification is prior known. The objects can be n-dimensional points within n-dimensional instance space where each point corresponds to one of the n features which describe objects. The distance of object is calculated by using distance metric, for example the Manhattan distance or the Euclidean distance [20].KNN is highly susceptible to noise in the training data due to high degree of local sensitivity. Thus the value of K influences the performance of KNN algorithm. The optimal choice of k is a problematic issue, but cross validation can be used to reveal optimal value of k for objects within training set.

General evaluation KNN is simple algorithm. It is quite good and comparable to the performance achieved with rule based learners and decision tree. KNN models are not prone to over fitting and It can be applied to incremental learning strategies because KNN does not build a classification model, Newly classified instances can easily be added to training set.

Relevant Literature There are several literature that survey the application of KNN for classification tasks. for example, a recent survey article by Jiang et al. [18] and the book by Mitchell [17], that proposed several improvements in KNN algorithms for classification. A publication on weighted KNN classification based on symbolic features was published by Cost and Salzberg [19].To produce real-valued distances from features, distance tables are calculated. The authors state that their technique is comparable or superior to standard KNN in three different application domains and has advantages in training speed and simplicity.

3.3.2 Decision Trees

Decision trees also known as classification trees. It learns from set of independent instances by applying ‘Divide and conquer’ approach. Decision trees are designed its node contain attribute test conditions to classify instances which have different characteristics. Decision trees branches leads to those classification and leaf node represent respective class. Constructing optimal decision trees is an NPcomplete problem; heuristics are used for constructing optimal trees. It select those features that best divide the training data to partition the records into smaller subsets. The important step is how to determine which feature to split on. There are different

feature evaluations techniques are available to find for optimal splitting feature. These techniques are derived from information theory and most of them are based on Shannon's entropy.

Information gain is one of technique which can also be applied as feature ranking method. It is one of most widely used splitting criteria. By expanding tree nodes that contribute to largest gain in average maximize the global mutual information. Rules derived from distance measures calculate the separability and discrimination between classes. Gini index of diversity one of the popular distance measures, a measure of the inequality of a distribution, the Kolmogorov-Smirnov distance.

The decision tree algorithm works as below: First, it selects an attribute to create root node and create one branch for each value of this attribute. This divides the training set into subsets, one for every value of the attribute. Then, it repeats the process recursively for each branch. If at any particular time all records at a node have the same classification then stop developing that part of the tree.

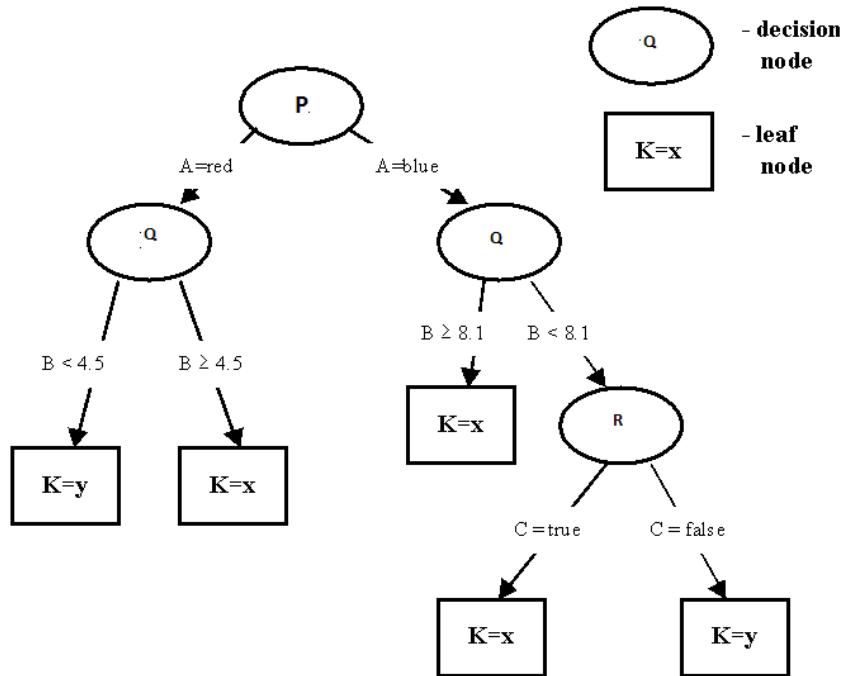


Figure 3.1 Decision Tree

Three major types of decision tree algorithms are used in practice. C4.5 is well known algorithm which are extension of ID3 algorithm.C5.0 decision tree algorithm is the successor of C4.5.These all introduced by Quinlan [15]. ID3, C4.5 and C5.0 trees use

information gain criteria to split and can handle both, numeric and categorical attribute. C4.5 can deal with missing value. C5.0 is the commercial successor of C4.5 which is faster and more memory efficient than C4.5 and support boosting.

General evaluation The clarity of decision trees is one their most important characteristics, so domain experts can easily understand the principle of the decision tree. It is very easy to understand why a certain object is classified to belong to a specific class. Decision trees are most extensively researched machine learning method, which deal with any kind of input data. They can also handle missing values because the information that attribute values are missing for specific objects can be processed by most decision tree algorithms. The learning process of decision trees is quite fast compared to other learning methods like neural networks or support vector machines.

Relevant Literature There are several studies available that provide comprehensive surveys on the application of decision tree algorithms in the fields of data mining, text mining and machine learning. Breslow and Aha [22] reviewed methods of tree simplification to improve the comprehensibility of the trees.

3.3.3 Ensemble Learning

Ensemble methods combined the predictions of multiple classifiers into a single learning model. Several classifier models called weak or base are trained and their results are usually combined by voting or averaging process. The concept behind ensemble learning has been taken with real life. When critical decisions have to be taken, opinions of several experts are taken into account rather than depending on a single opinion. Ensembles are more accurate in many cases than the individual classifiers, but it is not always useful to combine models. Ideal ensembles combine classifiers with high accuracy which differs as much as possible. If each classifier makes different mistakes, the total error rate will be decreased, if the classifiers are identical, a combination become useless because the results remain unchanged.

3.3.3.1 Bagging

The basic principle behind bootstrap aggregating, i.e. bagging is to aggregate predictions of several weak learner fitted to bootstrap samples of the original dataset through majority. Bootstrapping refers to re sampling with replacement from the original sample of instances and a weak learner is some pre-defined standard learning

algorithm. The term weak learner is widely used across the literature, even though weak learners are not bad classifiers.

Advantages of bagging are its ability to minimize variance and to avoid model over fitting. It is an intuitive and easy to implement approach and can be used with any learning algorithm as weak learner. Bagging has shown very good performance in several studies and the empirical fact that bagging improves the predictive performance of several learning algorithms is widely documented

3.3.3.2 Random Forest

A random forest is a modification of bagging that combine the bagging approach with a random sub-sampling method. While bagging works with any algorithm as weak learner, random forests are ensembles of classification trees. The commonly used growing algorithm for the single decision trees used within the random forest algorithm is Just like bagging, random forests select instances randomly with replacement (bootstrap), but unlike bagging, random forests also sample attributes for each tree. The trees are grown to maximal depth and each tree performs separate classification. Then each tree assigns a vector of attributes to a class and the random forest chooses the class of majority vote over all trees.

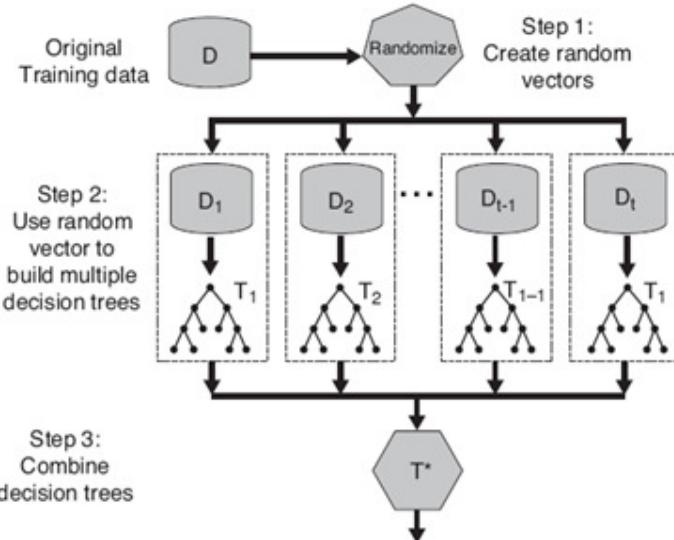


Figure 3.2 Random Forest

Similar to bagging, random forests are easily comprehensible and can reduce the variance of the prediction accuracy, but due to the sampling of attributes, the learning process of random forests is usually faster. Random forests can handle a very large number of input variables, and even when a large portion of attribute values is missing; it is often able to maintain the desired accuracy. Moreover, it is possible to measure attribute importance as well as the similarity between attributes.

3.3.3.3 Boosting

In the ideal bagging situation, all models complement each other, each being a specialist in a part of the domain where the other models do not perform very well. Boosting combines different learning models that complement each other, and applies certain re-sampling strategies in order to get the most informative strategic data. Instances that were classified incorrectly by some classifiers are taken into the training set more often, i.e., boosting encourages new models to become experts for instances handled incorrectly by earlier models.

On the one hand, boosting is similar to bagging because both methods combine models of the same weak learner and use voting to combine outputs of individual models. On the other hand, boosting differs from bagging because it is an iterative process where each new model is influenced by the performance of the models built previously. Moreover, boosting weights a model's contribution by its performance rather than giving equal weight to all models.

3.3.3.4 Stacking

Although Stacking is older than bagging and boosting it is not widely used. Unlike bagging and boosting, stacking combines different learning algorithm models. Instead of voting or averaging the results of different models or choosing the best model, stacking uses the concept of a meta-learner. Stacking tries to learn which weak learners are the reliable ones, using meta-learner algorithm to find the best way to combine the outputs of the base learners. If applied correctly, the meta-learner can be more efficient than a standard voting or averaging procedure. Since stacking does not rely only on one specific weak learner but on a combination of various weak learners, it is less prone to over fitting, produces a more general model of the data, and has been shown to be consistently effective for applications in several data mining domains.

General Evaluation of Ensemble learning Ensemble methods improve accuracy by combining the results of a set of different learning model. However, Ferri [18] introduces two important drawbacks of ensemble methods:

- Memory requirement is more to store set of multiple hypotheses.
- The comprehensibility of a single hypothesis is lost.

Relevant Literature of Ensemble Learning Various studies on random forests have shown that ensembles of trees improve the performance of decision trees [23]. Svetnik et al. [24] have combined random forests with a feature selection algorithm based on measuring the importance of single features, and successfully applied this combination to the task of QSAR-modeling

Chapter 04

Methodology

4.1 Overview of the Text classification

We present our novel approach for text classification of dbGaP text Study in to heart, lung and blood studies. The Figure shows the flow of text classification process. We have 468 dbGaP study text with pre-defined class. This model processes each document individually and finds the keywords for each document. It removes stop words from each document by applying stop word removal algorithm. Each keyword from text document converts into its root form by applying stemming algorithm on each document. These pre-processed study text given as input to K-nearest neighbor and random forest classifier.

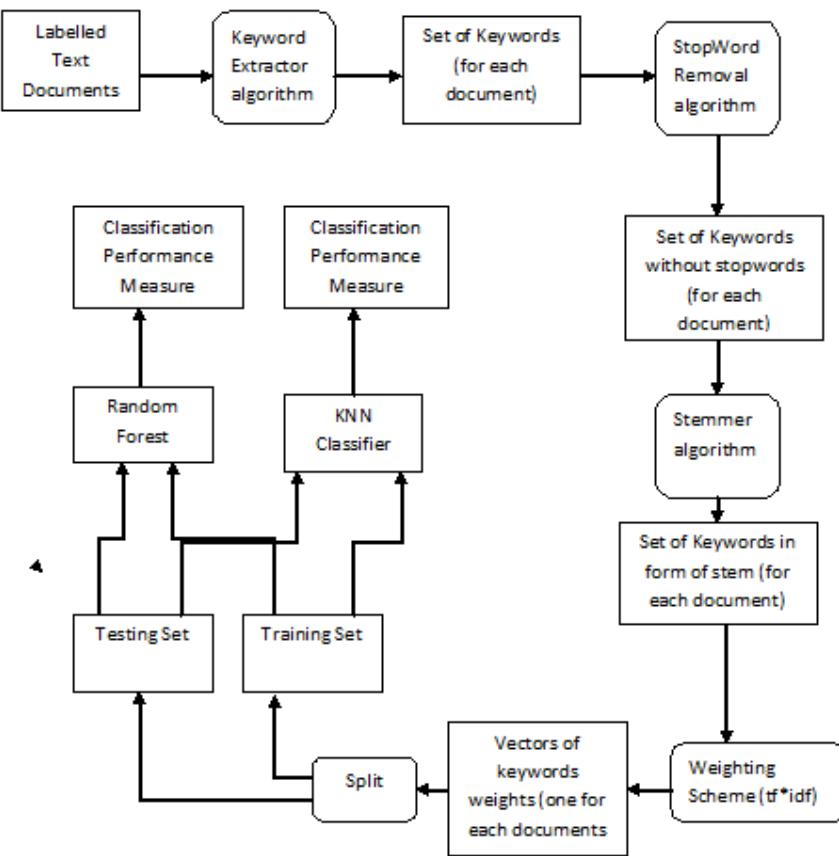


Figure 4.1 Flow of text classification

4.2 Document Preprocessing

Document pre-processing is the process of absorbing a new text document into text classification system.

The purpose of document preprocessing

- Represent the document efficiently by removing useless keywords.
- Improve retrieval performance.

Document pre-processing includes following stages:

1. Lexical analysis
2. Stop word elimination
3. Stemming

4.2.1 Lexical analysis

Lexical analyzer extracts keywords from text document by using tokenizer. It determines words from text documents. Lexical analysis separates the input alphabet into characters (the letters a- z) and separators (space, newline, tab).

Lexical analysis removes digits, punctuation marks because these are useless for making decision in text classification.

4.2.2 Stop Word Elimination

In the context of text classification stop words referred as useless symbols. So these stop words have to remove from text document in order to improve the performance of text classifier. Stop words include articles, prepositions, conjunctions, pronouns and possibly some verbs, nouns, adverbs.

Stop word elimination improves the size of the indexing structures.

4.2.3 Stemming

In information retrieval system morphological variants of words have similar semantic interpretations and can be considered as equivalent. For this purpose number of stemming Algorithms have been designed, which reduce a word to its root form. Thus, document is represented by stems rather than by the original words which helps to reduce dictionary size. The meaning of "fishing", "fished", "fisher" and "fish" is same in context of information retrieval system. A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish".

4.3 Weighting Scheme

tf-idf, short for term frequency-inverse document frequency used as weighting factor in text mining. It reflects how word is important to a document. The value of tf-idf rising as the number of times a word appears in the document.

term frequency-inverse document frequency is combination of two terms:

- **Term Frequency**

The term frequency is defined as number of occurrences of the term t_i within particular documents d_j .

$$tf_{i,j} = n_{i,j}$$

Where $n_{i,j}$ is number of occurrence of term t_i in document d_j .

To prevent bias for larger documents, term frequency often normalized as shown below.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- **Inverse Document Frequency**

The inverse document frequency is defined as total number of documents divide by number of documents containing the term t_i and taking logarithm of quotient.

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

To avoid divide by zero error, we can use $1 + |\{d : t_i \in d\}|$.

4.4 Dataset Preparation

The dataset can be split in following ratio:

50:50, 60:40, 66.7:33.3, 70:30 and 80:20 where first number denotes training set and second number denotes testing set. A 60:40 ratio means 60% of original dataset used as training set and remaining 40% used as testing set. Minimum training set must be 50% because training set with less than 50% would not sufficient for accurate model. 66.7:33.3 is widely used ratio to construct accurate model.

4.5 Classification algorithm

In classification step, the documents are split into training and testing documents, the training documents are used to train the system to identify different categories, the testing documents are used to evaluate the system. There are different text classification algorithm each have its own advantages and disadvantages. In this work we have implemented K-nearest neighbor and random forest algorithm on dbGaP study text.

4.5.1 K-nearest Neighbor (KNN) algorithm

K-nearest neighbor algorithm is from lazy classifiers group which works based on distance measures. Distance measure used can be Euclidean distance or correlation score. Let t_x be training instance. Correlation score $R(t_y, t_x)$ is calculated with formula-

$$R(t_y, t_x) = \frac{\sum_{j=1}^m (t_{y_j} - \bar{t}_y)(t_{x_j} - \bar{t}_x)}{\sqrt{\sum_{j=1}^m (t_{y_j} - \bar{t}_y)^2} \sqrt{\sum_{j=1}^m (t_{x_j} - \bar{t}_x)^2}}$$

Where \bar{t}_x and \bar{t}_y are the means of training and test tuples respectively. This correlation score is used as similarity score between training and test sample. In KNN algorithm optimal k neighbors are chosen for voting. The correlation score are

calculated for these neighbors and respective class predicted by majority of these and assigned to test instances in query.

4.5.2 Random forest algorithm

Random forest is the ensemble method specifically designed for decision tree classifiers. Random forest as name suggests grows many decision trees which are called as ensemble of unpruned decision trees. Each base classifier of random forest classify new vector to particular class. Forest chooses the classification having most votes over all trees in forest.

Random forest introduce two sources of randomness: “Bagging” and “Random input vectors”

- Each tree is grown using a bootstrap sample of training data
- At each node, best split is chosen from random sample of mtry variables instead of all variables

Each tree is grown as below: Suppose T is the number of cases in the training set. Randomly samples T cases with replacement. This sample is used as training set for growing the tree. If M is input variables and m is less than M then at each node m variables are selected at random out of the M and choose best split on these m attribute.

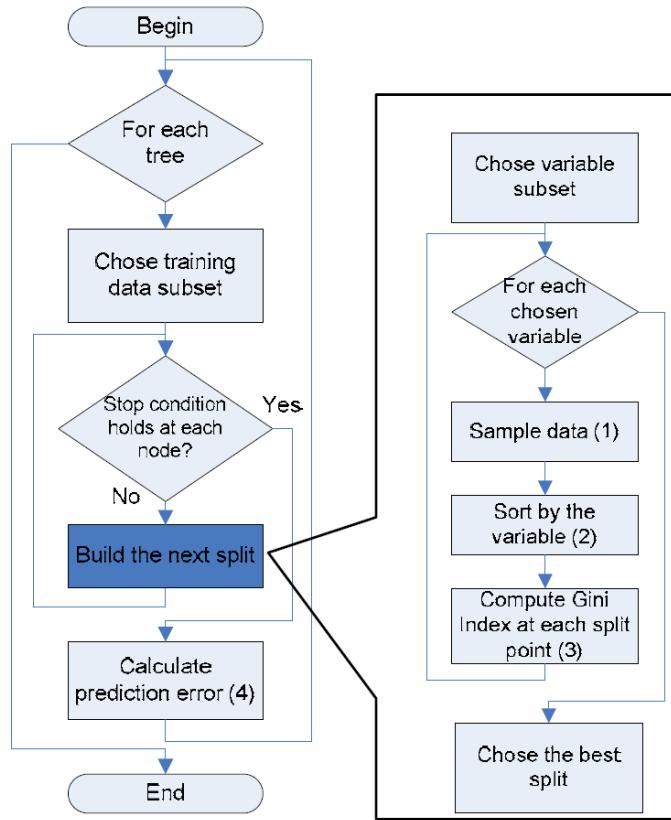


Figure 4.2 Random forest algorithm(Flow Chart)

4.6 Classification Performance Measures

Once classification model constructed, It can be used to predict class of unseen objects. The true positives (TP) and true negatives (TN) are correct classification, i.e., an object of class ‘yes’ is correctly predicted as ‘yes’. When the outcome of an object of class ‘no’ is incorrectly predicted as ‘yes’, this is called a false positive(FP) and false negative (FN) occurs when an object is incorrectly predicted as ‘no’ when it actually belongs to class ‘yes’.

Based on these four possible predictions, various performance metrics can be computed to compare quality of the classification results.

- *The overall accuracy* is the total number of correct classifications (TP +TN) divided by total number of classification (TP+TN+FP+FN).
- *The true positive rate* is TP divided by total number of positives (TP+FN).
- *The false positive rate* is FP divides by total number of negatives (FP+FN).

Chapter 05

Evaluations and Discussions

In this chapter we briefly present the result of k-nearest neighbor and random forest algorithm. We compare the result of KNN and RF algorithm with existing system of dbGaP study text retrieval system.

5.1 Dataset

For experimentation, we have manually collected and evaluated all available study description from dbGaP .There are 468 studies available in dbGaP. Out of that 39 are from heart studies, 23 are from lung studies, 37 are from blood studies and 369 are from other than heart, lung and blood studies.

dbGaP : 468 studies			
Heart	Lung	Blood	Other
39	23	37	369

Table 5.1 dataset

5.2 System Requirement

- Hardware Processor Intel i3, Ram 2 GB, Storage 320 GB, Frequency 2.20 GHz.
- Operating System Windows 7 Ultimate 32 bit.

- Programming Language Implementation is purely in C# language.
- IDE Programming is done in visual studio IDE environment.

5.3 Measures used for Evaluation

For classification performance evaluation, true positives, true negatives, false positives, and false negatives used to compare the results of the classifier under test.

This is illustrated as in figure 5.1:

		actual class (expectation)	
		tp (true positive)	fp (false positive)
predicted class (observation)	Correct result	Unexpected result	
	fn (false negative)	tn (true negative)	Missing result Correct absence of result

figure 5.1 Classification Model

$$\bullet \quad \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\bullet \quad \text{Precision} = \frac{Tp}{Tp+FP}$$

$$\bullet \quad \text{Recall} = \frac{TP}{TP+FN}$$

$$\bullet \quad \text{F- measure} = \frac{2*Precision*Recall}{precision+Recall}$$

5.4 Result of K-nearest Neighbor classifier

To present result, we applied KNN algorithm on the 400 studies of dbGaP text study. Here 66.7:33.3 ratios applied on 400 study text to split dataset into training set and testing set. Out of 400 study text 364 are correctly classified and 36 are incorrectly classified. Overall accuracy of KNN classifier on 400 studies is 91%.

	Heart	Lung	Blood	Other
Accuracy	0.95	0.98	0.98	0.92
Precision	0.76	0.84	0.90	0.93
Recall	0.68	0.73	0.79	0.96
F-measure	0.72	0.78	0.84	0.95

Table 5.2 Performance result of KNN

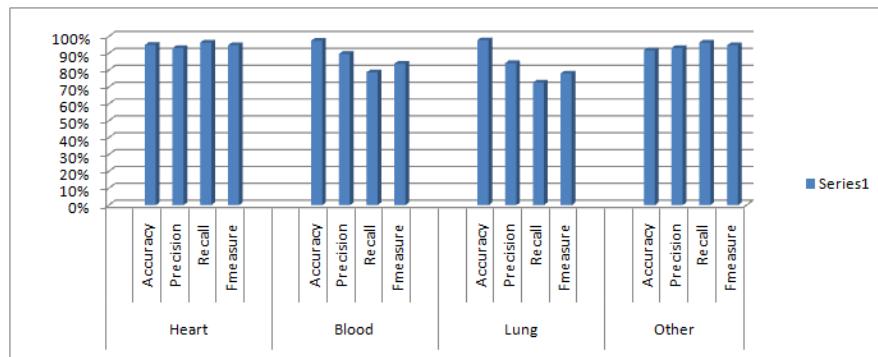


Figure 5.2 Graphical representation of performance result of KNN

5.5 Result of Random Forest Classifier

To present result for random forest classifier, we applied algorithm on the 400 studies of dbGaP text study. Here 66.7:33.3 ratios applied on 400 study text to split dataset into training set and testing set. Out of 400 study text 352 are correctly classified and 48 are incorrectly classified. Overall accuracy of KNN classifier on 400 studies is 88%.

	Heart	Lung	Blood	Other
Accuracy	0.96	0.94	0.97	0.89
Precision	0.84	0.44	0.95	0.91
Recall	0.69	0.64	0.64	0.94
F-measure	0.76	0.52	0.76	0.93

Table 5.3 Performance result of RF

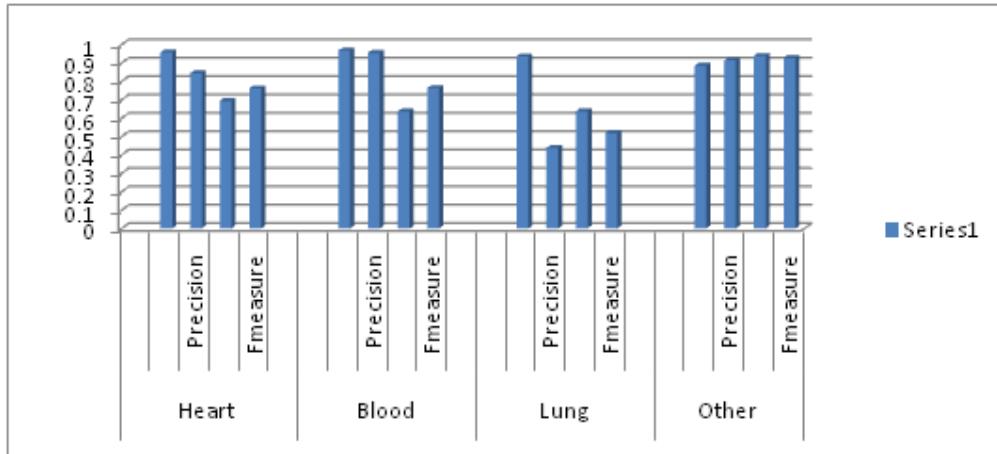


Figure 5.3 Graphical representation of performance result of RF

5.6 Result comparison

As expected comparison of our classifier with existing system of dbGaP study text retrieval show significant improvement. Performance of KNN and RF classifier compare with existing dbGaP system on the basis of accuracy, precision, recall and F-measure.

Accuracy

	dbGaP search System	KNN Classifier	RF Classifier
Heart	0.72	0.95	0.96
Lung	0.71	0.98	0.94
Blood	0.42	0.98	0.97

Table 5.4 Comparison table based on accuracy

Precision

	dbGaP search System	KNN Classifier	RF Classifier
Heart	0.19	0.93	0.84
Lung	0.13	0.84	0.44
Blood	0.08	0.90	0.95

Table 5.5 Comparison table based on precision

Recall

	dbGaP search System	KNN Classifier	RF Classifier
Heart	0.74	0.96	0.69
Lung	0.84	0.73	0.64
Blood	0.59	0.79	0.64

Table 5.6 Comparison table based on recall

F-measure

	dbGaP search System	KNN Classifier	RF Classifier
Heart	0.30	0.95	0.76
Lung	0.22	0.78	0.52
Blood	0.14	0.84	0.76

Table 5.7 Comparison table based on F-measure

Chapter 06

Conclusion and Future Scope

Although relatively small, the number of studies in dbGaP is rapidly increasing. We demonstrated that using k-nearest neighbor and random forest document classifier yields better document retrieval results than the keyword-based search currently available in dbGaP. Result analysis shows that performance of K-nearest neighbor classifier is better than random forest classifier on basis of accuracy, precision, recall and f-measure. It is evident from result analysis that this K-nearest neighbor classifier for dbGaP study text outperforms random forest classifier and conventional keyword based search of document retrieval system provided by dbGaP.

The proposed and implemented text classifier shows satisfactory performance in classifying dbGaP study text. In future work, dimensionality reduction techniques can be incorporated in order to further improve performance of text classifier. Due to high dimensionality of text documents, dimension reduction is usually performed before applying to classification algorithm. Feature extraction and dimensionality reduction can combine in single step using principal component analysis, canonical correlation analysis or linear discriminate analysis.

Appendix A

Paper Publication

Publication 1: Text Classification of Database of Genotypes and Phenotypes in Heart, Lung and Blood Studies.

Publication 2: Text Studies Classification of Database of Genotypes and Phenotypes using K- Nearest Neighbor Algorithm.

Following table shows acceptance and publication status.

Publication	Conference/Journal	Status
Publication 1	IJARCSSE journal	Accepted
Publication 1	IJRITCC journal	Published
Publication 2	IJSR journal	Accepted

References

- [1] Witten IH, Frank E, Hall MA. “Data Mining: Practical Machine Learning Tools and Techniques.” 3rd ed. Burlington, MA: Morgan Kaufmann; 2011.
- [2] Mailman MD, Feolo M, Jin Y, et al. “The NCBI dbGaP database of genotypes and phenotypes.” *Nat Genet*. 2007;39(10):1181–6.
- [3] Wei Q, Collier N. “Towards classifying species in systems biology papers using text mining.” *BMC Res Notes*. 2011;4(1):32.
- [4] Yang YaP, J. “A Comparative Study on Feature Selection in Text Categorization.” Proceedings of ICML-97, 14th International Conference on Machine Learning. 1997:412–20.
- [5] Kraft P, Zeggini E, Ioannidis JP. “Replication in genome-wide association studies.” *Stat Sci*. 2009;24(4):561–73.
- [6] Suresh S Kolekar,Satish S Kumbhar, “The text classification of database of genotypes and phenotypes in heart, lung and blood studies.” IJRITCC 2014 2(5):1078-1080.
- [7] Trieschnigg D, Pezik P, Lee V, de Jong F, Kraaij W, Rebholz-Schuhmann D. “MeSH Up: effective MeSH text classification for improved document retrieval.” *Bioinformatics*. 2009;25(11):1412–8.
- [8] Donaldson I, Martin J, de Bruijn B, et al. “PreBIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine.” *BMC Bioinformatics*. 2003;4:11.
- [9] SHI Yong-feng, ZHAO, “Comparison of text categorization algorithm” Wuhan university Journal of natural sciences. 2004.
- [10] David D. Lewis and Marc Ringuette, “A comparison of two learning algorithms for text categorization”, Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, US 1994.

- [11] Joachims, T. "Text Categorization with Support Vector Machines: Learning with many relevant features", europeon conference on machine learning pp 143-151, 1998
- [12] Sebastiani.F, "Machine Learning in Automated Text Categorization", ACM Computing Survey. pp.1-47, 2002.
- [13] National Library of Medicine (NLM). Genetic Home Reference, Handbook <http://www.nlm.nih.gov/handbook.html>.
- [14] S.N.Sivanandam, S. N. Deepa "Principles of Soft Computing"
- [15] Thomas M. Cover and P. E. Hart. "Nearest Neighbor Pattern Classification." IEEE Transactions on Information Theory, 13(1):21-27, 1967.
- [16] Moty Ben-Dov, Ronen Feldman, "Text Mining and Information Extraction, The Data Mining and Knowledge Discovery" Handbook 2005, 801-831, Springer.
- [17] Tom M. Mitchell. "Machine Learning." McGraw-Hill Education (ISE Editions), 1998.
- [18] Liangxiao Jiang, Zhihua Cai, DianhongWang, and Siwei Jiang. "Survey of Improving K-Nearest Neighbor for Classification." Fuzzy Systems and Knowledge Discovery, Fourth International Conference on, 1:679-683, 2007.
- [19] Scott Cost and Steven Salzberg. "A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features." Mach. Learn., 10(1):57-78, 1993.
- [20] Kilian Q Weinberger, John Blitzer, and Lawrence K. Saul. "Distance Metric Learning for Large Margin nearest Neighbor Classification." Advances in Neural Information Processing Systems, 18:1473-1480, 2006.
- [21] Ross J. Quinlan. "C4.5: Programs for Machine Learning." Morgan Kaufmann, 1993
- [22] Leonard A. Breslow and David W. Aha. "Simplifying Decision Trees: A Survey." Knowl.Eng. Rev., 12(1):1-40, 1997.
- [23] Leo Breiman. Random Forests." Machine Learning", 45(1):5-32, 2001.
- [24] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. "A Classification and Regression Tool for Compound Classification and QSAR Modeling." J. Chem. Inf. Comput. Sci., 43(6):1947-1958, 2003.