International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Text Studies Classification of Database of Genotypes and Phenotypes using K-Nearest Neighbor Algorithm

Kolekar Suresh S[1], Kumbhar Satish S[2]

[1, 2] Department of Computer Engineering and Technology, College of Engineering, Pune, India

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) 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.

1. Introduction

1.1 dbGaP

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).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.

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. 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 neighbor 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.

2. Challenging Issues in Existing System

dbGaP study retrieval system returns several studies that are not related for particular 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.

2.1 Manual Performance Evaluation of dbGaP search system [6]

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 I, and were assigned manually depending on its relevance to the document.

Table 1: Labels Assigned to Documents on dataset

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

- Evaluation metrics used were accuracy, precision, recall, and F-measure .Table 2 represents measurements definition of evaluation metrics.

Table 2: 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)		

- 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

Table 3: dbGaP keyword Search result

	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

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

3. Methodology

Result of manual keyword search shows that performance of dbGaP Entrez system is very poor. Proposed solution is to improve study retrieval in the context of the dbGaP database by using machine learning algorithms for text classification.

3.1 Overview of Text Classification Process

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 convert into its root form by applying stemming algorithm on each document. These pre-processed study text given as input to K-nearest neighbor classifier.

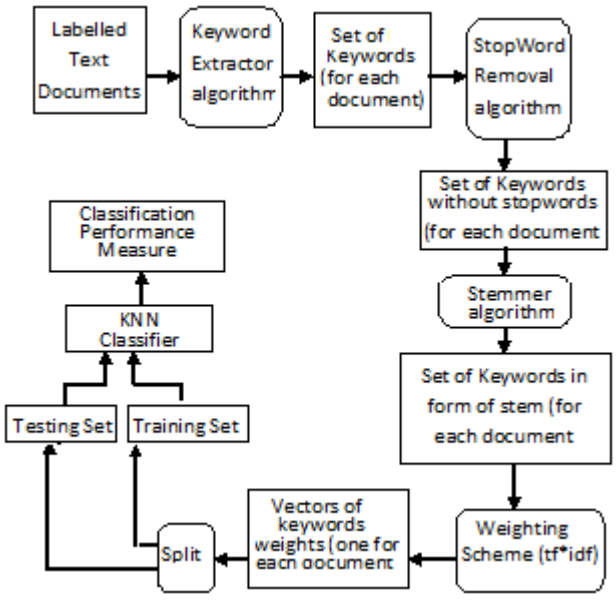

Figure 1: Flow of text classification

3.2 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:

- Lexical analysis
- Stop word elimination
- Stemming

3.2.1 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.

3.2.2 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.

3.2.3 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".

3.3 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 ass 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.

3.4 Classification algorithm

In classification step, the documents are are split into training and testing documents, the training documents are used to train the system to indentify 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 algorithm on dbGaP study text.

3.4.1 K-nearest Neighbor (KNN) algorithm

K-nearest neighbor algorithm is from lazy classifiers group which works based on distance measures [15]. 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:

$$R(t_y,t_x) = \frac{\sum_{j=1}^{m}(t_y - \bar{t_y})(t_x - \bar{t_x})}{\sqrt{\sum_{j=1}^{m}(t_y - \bar{t_y})^2}\sqrt{\sum_{j=1}^{m}(t_x - \bar{t_x})^2}} \quad (1)$$

Where $\bar{t_x}$ and $\bar{t_y}$ are the means of training and test tuples respectively. This R (t_y,t_x) 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. Experimentation and Results

In this chapter we briefly present the result of k-nearest neighbor algorithm.

4.1 Dataset Used for experimentation

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.

Table 4: Dataset

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

4.2 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.

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision = $\frac{Tp}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F- measure = $\frac{2*Precision*Recall}{precision+Recall}$

4.3 Result of K-nearest Neighbor classifier

To present result, we applied KNN algorithm on the 468 studies of dbGaP text study. Here 66.7:33.3 ratios applied on 468 study text to split dataset into training set and testing set. Performance of KNN algorithm over dbGaP study text is given as follows.

Table 5: Result of KNN classifier

	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

Overall accuracy of KNN classifier on 468 studies is 91%

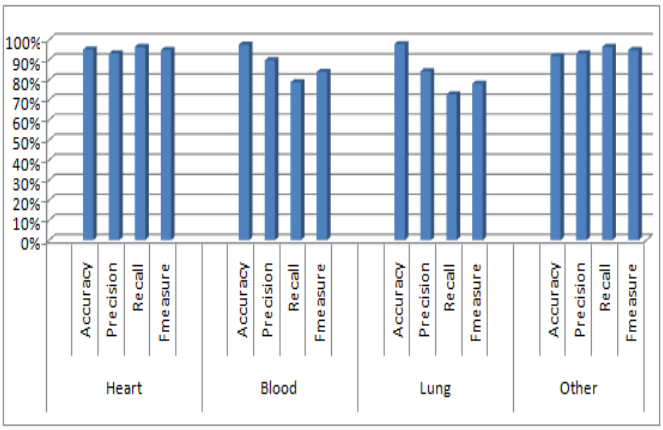

Figure 2: Graphical Evaluation of Performance of KNN classifier

5. Conclusion and future scope

dbGaP Entrez system works based on keyword based search system with poor performance. By using appropriate text classifier performance of dbGaP search system improves significantly. It is evident from results that KNN classifier for text based classification outperforms 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.

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). UMLS Metathesaurus FactSheet.2012.http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.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.

Author Profile

Mr Suresh S Kolekar received a B. Tech from Dr. Babasaheb Ambedkar Technological University,Raigad..He is currently M. Tech student at College of Engineering, Pune, India.

Mr Satish S Kumbhar received a M. Tech from College of Engineering, Pune. He is currently working as Assistant Professor at College of Engineering, Pune, India.