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

Mr. Suresh S Kolekar
Department of Computer Engineering and Information Technology
College of Engineering
Pune, India
*e-mail: suresh.kolekar123@gmail.com*

Mr. Satish S Kumbhar
Department of Computer Engineering and Information Technology
College of Engineering
Pune, India
*e-mail: satish116@gmail.com*

***Abstract*—**The database of Genotypes and Phenotypes (dbGaP) was developed by National Center for Biotechnology Information (NCBI) to archive and distribute the results of various studies that have examined the interaction of genotype and phenotype. It is public repository for individual level phenotype, exposure, genotype, sequence data and the associations between them. Searching relevant studies of particular interest accurately and completely is challenging task due to keyword based search method of dbGaP Entrez system. 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. This paper surveys of text classification, process of text classification and different text classifications algorithms to classify studies of database of Genotypes and Phenotypes.

***Keywords-****Text classification;Machine Learning;database;genome-wide association studies;GWAS.*

*****

## I. INTRODUCTION

The National Library of Medicine (NLM), part of the National Institutes of Health (NIH), announces the introduction of dbGaP, a new database designed to archive and distribute data from genome wide association (GWA) studies. GWA studies explore the association between specific genes and observable traits, such as blood pressure and weight, or the presence or absence of a disease or condition (phenotype information). Connecting phenotype and genotype data provides information about the genes that may be involved in a disease process or condition, which can be critical for better understanding the disease and for developing new diagnostic methods and treatments. dbGaP, the database of Genotype and Phenotype, for the first time will provide a central location for interested parties to see all study documentation and to view summaries of the measured variables in an organized and searchable web format.

As of today, dbGaP contained 468 studies, including more than 144716 phenotype variables. However, retrieving relevant studies accurately and completely is challenging issue, because phenotypic information related to studies is often stored in a non-standardized format. For particular queries, the dbGaP Entrez system returns several studies that are irrelevant, and it does not make clear how particular studies are selected and why they appear in a particular order. Thus, users have to review each study description carefully to determine relevant studies, which can become a laborious and time-consuming task when there are many studies to be retrieved.

Text classification is an important part of text mining in the field of document retrieval. Current research of text classification aims to improve the quality of text representation and develop high quality text classifier. Machine learning techniques have been actively explored for text classification. Among these are Naive bayes classifier, K-nearest neighbor classifiers, support vector machine, neural networks.

## II. CHALLENGING ISSUES IN EXISTING SYSTEM

dbGaP Entrez system returns several studies that are not always relevant for particular queries. Consequently, users have to review each study description carefully to retrieve relevant studies, which can become a laborious and time-consuming task when there are many studies to be retrieved.

### *A. Manual Performannce 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 I, and were assigned manually depending on its relevance to the document.

 ISSN: 2321-8169
1078– 1080

TABLE I. LABELS ASSIGNED TO DOCUMENTS IN DATASET

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

Evaluation metrics used were accuracy, precision, recall, and F-measure . The F-measure is the harmonic mean between precision and recall.

Table II represents measurements definition of evaluation metrics

TABLE II. 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)<br>Precision = TP/(TP+FP)<br>Recall = TP/(TP+FN)<br>F-measure = 2*Precision*Recall/(Precision+Recall) | | |

*1) 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

*2) Evaluation Metrics for Lung Keyword*: dbGaP Entrez system for heart 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

*3) Evaluation Metrics for Blood Keyword*: dbGaP Entrez system for heart 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 III. DBGAP KEYWORD SEARCH RESULT

| | ***Heart*** | ***Lung*** | ***Blood*** |
|---|---|---|---|
| **Accuracy** | 0.72 | 0.71 | 0.42 |
| **Precision** | 0.19 | 0.13 | 0.08 |
| **Recall** | 0.74 | 0.82 | 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.

## III. PROPOSED SOLUTION

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.

### *A. Text Classification Algorithm*

Database of Genotypes and Phenotypes are rich with hidden information which are important for intelligent decision making. Classification algorithms can be used to extract models to describe important data classes. Text classification applies supervised, unsupervised and semi supervised methods to classify text. There are several methods that can be used to classify text such as Naive Bayes Classifier, Decision Trees, K Nearest Neighbor (KNN), Artificial Neural Networks (ANN), and Support Vector Machine (SVM). Some techniques are described in sub section A.

*1) Naïve Bayes Classifier :*

Naïve Bayes classifier is traditional approach for text classification. It is supervised machine learning algorithm which learns training examples in priori probability given unseen examples. It is probabilistic classifier based on applying Bayes theorem with independent assumptions. In Text classification [8] performance of Naïve Bayes is very poor when features are dependent on each other. It is simple and fast classifier.

*2) Decision Tree :*

Decision tree is tree structure classifier where each node is either leaf node or decision node. Leaf node indicates the value of class attribute of examples and decision node specifies some test to be carried out on single attribute value with one branch and sub tree for each possible outcome of the test. Decision tree classification is simple to understand and interpret even for non-expert user as compared to other decision support tools [9].

*3) K Nearest Neighbor(KNN):*

KNN algorithm is one of the simplest algorithm among all machine learning algorithms. It is a type of lazy learning, or instance based learning.KNN assign weight to neighbors contributions. Nearer neighbors contribute more to the average than the more distant ones. The neighbors are taken from set of objects for which object or class property value in known. This can be use as training set, though no explicit training step is required.KNN is simple and competitive algorithm with Support Vector Machine for text classification [11].

*4) Artificial Neural Networks(ANN) :*

Artificial neural networks are computational models inspired by human brain.ANN is applicable to machine learning as well as pattern recognition. Artificial neural networks are systems of interconnected "neurons" which can compute values from inputs.ANN provide better solution for those problems which cannot be solved sequentially or by sequential algorithms. Among the various algorithms provided by artificial neural networks Backpropagation is very popular algorithm. A back propagation neural network is a multilayer and feed-forward neural network which consists of input layer, a hidden layer and an output layer [13].

*5) Support Vector Machine :*

Support vector machine is the machine learning algorithm used for classification of linear and non linear data. It uses non linear mapping to transform training data into higher dimension and then it search for linear optimal separating hyper plane. This algorithm initially applied to text classification by Joachim in 1998[10].He compare SVM with KNN and NB.

## IV. CONCLUSION

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. This paper survey on text classification for dbGaP studies. This survey focused on challenging issue in existing system and explored the text classification algorithms to construct the text classifier.

## REFERENCES

[1] Mailman MD, Feolo M, Jin Y, et al. *The NCBI dbGaP database of genotypes and phenotypes.* Nat Genet. 2007;39(10):1181–6.

[2] Wei Q, Collier N. *Towards classifying species in systems biology papers using text mining.* BMC Res Notes. 2011;4(1):32.

[3] 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.

[4] Kraft P, Zeggini E, Ioannidis JP. "Replication in genome-wide association studies." Stat Sci. 2009;24(4):561–73.

[5] Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques.* 3rd ed. Burlington, MA: Morgan Kaufmann; 2011.

[6] 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.

[7] 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.

[8] SHI Yong-feng, ZHAO, "Comparison of text categorization algorithm" Wuhan university Journal of natural sciences. 2004.

[9] 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.

[10] Joachims, T. "Text Categorization with Support Vector Machines: Learning with many relevant features", europeon conference on machine learning pp 143-151, 1998

[11] Sebastiani.F, "Machine Learning in Automated Text Categorization", ACM Computing Survey. pp.1-47, 2002.

[12] National Library of Medicine (NLM). UMLS Metathesaurus FactSheet.2012.http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html.

[13] S.N.Sivanandam, S. N. Deepa "Principles of Soft Computing"