International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 11830-11839

A Novel Weighted Probabilistic Based Gene-Disease Document Classification Model Using Hadoop Framework for Distributed Biomedical Repositories

Dr.B.R.S.Reddy
Professor, Department of Computer Science and Engineering
Ramachandra College of Engineering, Eluru Andhra Pradesh, India.

Narni.Siva Chintaiah
Assistant Professor, Department of Computer Science and Engineering
Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India.

Orcid Id: 0000-0003-2627-0909

Abstract

With the exponential growth of biomedical repositories and gene-disease databases, building a high dimensional ranking based classifier is an essential task for clinical decision making on distributed biomedical databases. Since years, a large number of works have been implemented to predict the gene related diseases by manually analyzing biomedical documents. This manual process is not only time consuming, but also inefficient on high dimensional features. Generally, classification techniques have been used to classify a large number of biomedical data for gene related disease prediction. Detection and annotation of gene-disease based biomedical documents require an extensive computational resource with high true positive rate. Currently, a large number of gene classification models have been developed on a single biomedical repository with limited dimensional space. As the size of the biomedical documents increases in distributed biomedical repositories, corresponding gene-disease entities and dimensionality also increases exponentially. Therefore, there is an essential need for automatic detection and classification of gene-disease documents on the distributed biomedical dataset using Hadoop framework. Experimental results proved that the proposed automatic gene-disease classification model has high computational efficiency in terms of memory, time and statistical analysis than the traditional models.

Keywords—Biomedical, Gene-Disease, MeSH.

INTRODUCTION

Biomedical documents play very important role in the process of medical decision making as well as in treatment of gene related diseases. It can also be stated that, biomedical documents are very essential for both healthcare professionals and for researchers. In the biomedical databases, the named entities (NEs) include genes, proteins, cells, drugs, chemicals, diseases, etc, which are frequently used in biomedical text for pattern analysis.Classification schemes are responsible for detecting and predicting several complex diseases by analysing biomedical documents for clinical decision making.

Initially, the document data give rise to features, and these features are evaluated in the process of document clustering. Mostly high-dimensional document space's hard to handle, pre-process and cluster due to large amounts of document sets. To improve the learning of the clustering algorithm, the numbers of samples are required to be learned according to its dimension. Conceptually, this document space is a sub-space of low dimensionality, and it is wrapped with ambient space. Due to this dimensionality issue, many dimension reduction methods were developed to resolve the above problem. The main objective of this method is, to decrease the document dimensions and enhance the performance and efficiency. Thus, through dimensionality reduction methods, dimensional feature spaces of high-dimensional documents are minimized so the conventional Clustering schemes are used to achieve the better clustering performance. Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) are two most used techniques for feature selection and dimension reduction [2]. These algorithms are implemented in the various fields such as pattern recognition, text mining and gene extraction and data analysis. In supervised machine learning algorithm, training data are required for the process of estimation or prediction.

Classification can be defined as a special kind of learning model which is responsible for categorization of different gene-disease datasets. These datasets are classified into set of finite or infinite classes. Apart from supervised and unsupervised machine learning approaches, there are two other machine learning techniques generally used for classification are: - regression and clustering. A learning function generally maps original data into their real-value

variable in the process of regression. This technique can estimate the predictive variable for every individual sample. Clustering is categorized under the category of unsupervised learning and here groups are formed according to the similarity of data items. The groups which are built in the process of clustering are known as clusters. Data items having high similarity are included in the same cluster, whereas data items having no similarity or least similarity are included in different clusters.

The classification scheme Bagging is categorized under a special kind of Bootstrap aggregation. Furthermore, the process of bagging also supports all characteristics of machine learning and meta-algorithm. Meta-algorithm can be defined as a specific algorithm which is developed for improvement of stabilization factor. Bagging has wide range of applications in the fields of statistical classification and regression. The process of bagging not only reduces variance, but also limits over fitting. Besides these, there exists another application of bagging classification i.e., decision trees. Some common factors are generally responsible for errors of machine learning algorithms, those are:- noise, bias and variance. Noise is generally defined as an error occurs by the target function. Biases are the targets which are not qualified to be learnt by the classification algorithms. Variance is the outcome of sampling process. The above mentioned classification approach reduces overall errors.

Boosting can be defined as a special kind of machine learning meta-algorithm. This algorithm has the prime objective of reducing bias significantly. Additionally, it is also responsible for decreasing variance. In other words, boosting is the process of transforming weak learners to strong learners. Weak learners are the learners which are very poorly correlated along with true classification. But, strong learners are strongly correlated with true classification.

Extreme Learning Machine can be defined as a single-hidden layer feed-forward neural network (SLFN) with learning model [4]. The traditional optimization approaches like gradient descent based back-propagation [5] evaluate weights and biases. The proposed technique is responsible for decreasing the training time effectively through random assignment of weights and biases. The above extended EL method results better efficiency and performance as compared to all traditional approaches. EL has wide range of applications in different domains like face recognition, human action recognition, landmark recognition and protein sequence classification, medical disease prediction [1, 3,4]. But, EL has two major issues, those are:- 1) This model has over fitting problem and the performance can't be predicted for unknown datasets. 2) This model is not applicable to binary classification and uncertain datasets.

Feed-forward neural network can be considered as most commonly and widely implemented classification model. This method has one or more hidden layer(s) along with an output layer. The output layer is responsible for transmitting final response on the training dataset [2]. A large number of research works have been implemented in the field of Feed-Forward Neural Networks since years. This model has complex linear or nonlinear structure directly mapping from inputs. These structures are not appropriate for classical parametric constraints to manage large inputs in the traditional models. Another important feature of feed-forward neural network is the inter-dependency among the layers through parameter mapping. Single-Hidden-Layer Feed-Forward Networks (SLFNs) are treated as the most efficient and widely used feed-forward neural networks on small datasets.

In order to resolve the issues of traditional EL, weighted-EL approach is developed subsequently. The weights are increased gradually with respect to time in case of large sample size. In most of the feed forward ANN techniques, parameters of each and every layer is required to be tuned through several learning approaches. Gradient Descent-Based Approaches and Back-Propagation (BP) techniques are some important learning algorithms in feed forward neural networks [5]. The speed of learning models is very slow in case of feed forward neural networks compared to ANN. Because of better generalization capability and fast computational speed, EL approach is named as 'Extreme Learning Machine' (EL). A lot of problems are detected in case of conventional Gradient-Descent Algorithms like stopping criterion, learning rate, number of epochs and local minima [9].

Naive Bayes (NB) classification scheme is generally implemented in research domain. NB classification scheme is compared with some other pre-existing classification schemes (such as:- Logistic Regression, Nearest Neighbour, Decision Tree, Neural Network) and it is found to be most efficient and popular classification scheme. The above comparative analysis is performed by analysing and studying the Receiver Operating Characteristics (ROC) curve. NB classification scheme is simple, efficient as well as more effective classification technique which includes the features of machine learning. Another advantage of this technique is that, it requires less numbers of training datasets. There are not significant numbers of parameter present. Additionally, this approach is robust for unavailable and noisy data. Apart from these, the proposed technique is responsible for considering class condition independence and it also reduces the overall computation overhead.

The process of feature extraction has high significance in the field of classification. All features are divided into two groups, those are:-

1) According to the first group, features extraction using noisy attributes and contextual information.

2) The second group contains correlated features.

Traditional feature extraction models discard noisy features in order to decrease the high dimensional features to a lower

dimensional feature.

Let us assume Nmin and Nmax are minimum and maximum numbers of hidden neurons respectively, where N denotes the present value of hidden neurons. For each and every N, the average accuracy rate of EL through 10-fold cross-validation scheme is evaluated. At last, hidden neurons having maximum average accuracy is chosen as optimal. After selecting the optimal numbers of hidden neurons, the EL classifier is implemented in order to evaluate the classification accuracy by considering the outcomes of PCA and the outcomes are averaged later.

The training datasets used in this paper have a significant issue for any classification models as they have large number of feature space, ranging from 100 to 12000 features. The larger the feature space increases the search space and computational memory for disease prediction. Another crucial issue for handling the high dimensional features is the small sample size problem. The accuracy of the model employed will be reduced if the size of the training data is not sufficient relative to the feature space.

In the past, machine learning models used a single classification model to predict the test data using the training samples. However, multiple classifiers can be used to predict the same test data using the training samples. This process is known as ensemble learning. Ensemble classification has been successfully applied to different classification problems to improve the classification accuracy using the optimal feature selection measures.

Particle Swarm Optimization (PSO) is very popular optimization techniques in machine learning models. PSO is generally applied in the literature to adjust the initialization parameters of base classifiers in the ensemble learning models. The main objective of this paper is to optimize the traditional PSO parameters in the ensemble classification model in order to improve the accuracy and error rate. In the ensemble model, neural network is used as one of the base classifier and weights are initialized using the proposed PSO technique.

Related Works

F. Ö. Çatak developed an advanced classification strategy integrated with extreme learning machine in order to classify arbitrarily partitioned data [1]. Generally, machine learning approaches are considered to be efficient enough to analyse huge datasets. Due to its complexity, the data retrieval process has major challenges. In case of big data, the process of automatic classification is very complicated and hard. In the initial phase, datasets ensembles are constructed for big data. After successful completion of initial phase, ELM approach is implemented to construct weak learners. Finally, a group of weak learners can form a strong learner efficiently. The outcomes of above technique decreases the overall training time significantly. The accuracy and performance of this approach is compared with all previously existing traditional approaches.

B. Chandra et.al implemented a new statistical feature selection technique to classify gene expression data [2]. This proposed technique is known as Effective Range based Gene Selection technique. The prime objective of this approach is to perform the process of feature selection effectively along with the ranking of informative genes. The main workflow of this model is ; when a feature is able to discriminate a class more accurately, that feature is considered to have high priority as compared to other features. This proposed ERGS approach directly depends upon ranges of every individual class in case of a single feature. It eliminates the influence of outliers and classes having huge variance. No computationally efficient search method is mandatory for the above presented technique which is an exception in all traditional feature selection methods. Some other advantages of ERGS approach are:- It is very fast, simple to be applied with and never need any proper distribution assumption.

J. Chen et.al proposed a MapReduce-based extreme learning machine approach in order to analyse big data [3]. Many numbers of different ELMs are trained simultaneously and their outcomes are merged together by voting method. This can effectively enhance the overall classification performance. Apart from this, the above technique results very high efficiency as well as scalability whenever implemented in hadoop cluster. Some advantages of the above proposed approach are:- it is most accurate, efficient and scalable in case of big data analysis. Further future works can be carried out in order to extend the above proposed approach.

S. Das et.al. developed an efficient classification strategy for cancerous disease detection [4]. In this technique, they integrated the basic concepts of traditional gene selection process along with decision tree technique. In case of disease related to genes prediction, high dimensional data may result reduced performance. Hence, this problem is considered as a major problem in all conventional classification algorithms. Correlation coefficients are the major building blocks used in the computation of gene dependency. Similarity coefficient is calculated through Jaccard Coefficient. In the subsequent stage, an advanced gene similarity matrix is formed. Additionally, each gene is allotted a priority rank according to its importance. Gene having highest priority is accepted as the most important gene among all other genes in that particular set.

H. I. Elshazly, et.al. developed an advanced and new classification scheme in order to evaluate the performance of biomedical data [5]. In this piece of research work, the overall performance analysis of two ensemble classifiers is carried out. They conducted their experiment in the evaluation phase on five medical datasets.

X. Fei et.al. developed a large scale parallelized text classification scheme in order to process large scale clinical data [6]. They integrated the MapReduce method for text classification solution. The prime objective of this model is to enhance the overall accuracy and efficiency of multi-class problem in large scale clinical data of TCM. They implemented a parallel processing of MapReduce in order to predict N binary classifiers. They also suggested introducing an advanced framework for an intelligent reasoner. Intelligent reasoner has the responsibility of decision making. Further future research works can be carried out in order to detect gene-disease prediction and learning associated factors.

H. Hu analysed and studied different patterns in disease classification using random forest approach [7]. In this approach, they proposed a new method to analyse the relationships among gene patterns in various disease-relevant classification techniques. There are two important characteristics of this technique are :- 1) The RRF approach is capable of providing robust selection pathways. These classifiers are responsible for performing as the best classifiers. Chances of errors are least in this case. 2) When FIM approach is implemented, genes can be detected out of different pathways. Interactions among various pathways are detected by this proposed technique. Generally, frequent co-occurrence of genes may point to specific degree of association. Direct correlation interaction can't be represented by the above co-occurrence. In future, this technique can be modified and extended in order to achieve better accuracy and performance.

D. Kiela, et.al. Presented a new technique for unsupervised discovery of information structure in biomedical documents [8]. For the process of text classification, Information structure (IS) is considered as most efficient technique. It has wide range of applications in Biomedical Text Mining. This approach is most useful for faster, accelerating and most time consuming process. The discipline of biomedicine is influenced greatly by domain variation and this makes the whole process of classification too expensive. Hence, IS approach can't be implemented effectively in the domain of biomedicine. The overall performance of many unsupervised approaches from PubMed are analysed and compared with the above presented technique. The proposed multilevel weighted graph clustering approach results 0.70 F-scores in case of most IS approaches. It can be showed that by the process of evaluation, the introduced technique outperforms all other existing techniques.

M. Kumar et.al. developed a MapReduce-based Proximal Support Vector Machine classification technique in order to classify microarray data [9]. Since few years, microarray-based classification approach is considered as a significant problem for classification approaches. This technique has one severe disadvantage that is, issue of dimensionality. Numbers of different MapReduce-based approaches are developed in order to choose relevant features. MapReduce based proximal support vector machine (mrPSVM) technique is implemented in order to carry out the process of classification more efficiently. Such approaches have wide range of applications in Hadoop framework. Here, all the microarray-based feature selection approaches are studied and analysed. In future, this method can be extended with an integration of Spark framework. Some other algorithms like ANN, Deep learning and decision tree based approaches can be integrated with Hadoop framework in future and its effect can be analyzed on high dimensional datasets.

F. Lin, et.al. Designed an advanced high-performance multiclass classification framework [10]. The above proposed framework uses cloud computing architecture. Here, the traditional multiclass classification scheme is responsible for the integration of genetic algorithm with SVM. Generally, it needs huge amount of computing resources. mRNA dataset with 14 tumor types was included in the evaluation process of classification framework. It generally uses GEP (Gene Expression Profiles) to predict cancer disease at a very early stage. They performed series of experiments in the evaluation process. On increasing numbers of servers from 1-10, the training time of classifiers are decreased significantly. Additionally, the classification accuracy is enhanced to 94% which is better as compared to other classical approaches.

H. Liu, et.al. developed a grouping-based ensemble gene selection approach in order to classify microarray gene data [11]. In this paper, an advanced three-stage ensemble gene selection approach is proposed which basically uses a grouping method for classification. This approach usually selects more numbers of genes than that of other pre-existing gene selection approaches. There also exists difficulty in computing the optimal value of parameter t at an early stage. Further future researches can be done in order to overcome the above mentioned problems. Additionally, this technique can be evaluated with more number of datasets.

Y. Liu,et.l. introduced parallelization of back propagation neural network in MapReduce and Spark [12]. This approach is known as Parallelized Back Propagation Neural Network (PBPNN). In order to enhance the accuracy of traditional classification schemes, PBPNN applies ensemble approaches like bootstrapping and majority voting. The process of bootstrapping is responsible for maintaining original data information in sub-dataset. Majority voting produces strong classifier which directly depends on aggregation of weak classifiers. In the evaluation phase, numbers of experiments are carried out and it is demonstrated that this approach performs better as compared to BPNN. In case of various distributed computing platforms, the efficiency of algorithm is computed. In between various platforms,

I. Palit et.al. proposed a scalable and parallel boosting technique integrated with MapReduce [13]. They evaluated and analysed performances of presented approaches

implemented in a parallel distributed MapReduce framework. Further future works can be done in order to implement different data partitioning techniques which can enhance the classification performance. Furthermore, the above technique can be extended to be integrated with multi-resolution boosting models.

S. Sumathipala, et.al. developed a protein name classification technique which is based upon new set of features [14]. This approach is capable of measuring probabilistic stability of syntagmatic structure of words just like Unit hood and Protein hood. There exist some major drawbacks in this approach, those are:- 1) As compared to single term, proteinhood is very effective in case of compound terms. 2) Sufficient knowledge is required to detect general suffixes and keywords those are essential in the process of protein recognition. On discarding suffixes, performance of the system is degraded. In future, more emphasis can be given on automatic learning of common suffixes, prefixes and keywords.

J. Zhai, et.al. studied and analysed the features of MapReduce and ensemble learning algorithm [15]. In this research work, a new classification scheme was developed in order to classify imbalanced huge datasets. This technique has two major advantages, those are :- 1) It is responsible for expanding the learning region of positive class instances. 2) Also, it is responsible for classification of imbalanced huge datasets. It can be concluded that, the above technique is very much feasible and performs better as compared to other techniques like SMOTE-Vote, SMOTE-Boost and SMOTE-Bagging. The above proposed approach has good performance, speed up as well as scale up. In future, additional works can be carried out in order to classify imbalanced huge datasets having multiple classes.

PROPOSED APPROACH

Data Preparation Phase

Biomedical documents, phrases, sentences are used in the feature extraction to extract the main features of the original documents. The graph based feature extraction generates the feature extraction by extracting phrases or sentences from the set of key peer nodes of the overlay network. Extracting the key phrases or sentences from the set of documents can be obtained by computing the ranking scores for each phrase or sentence and then selecting the highest scored phrases or sentences as shown in Figure 1.

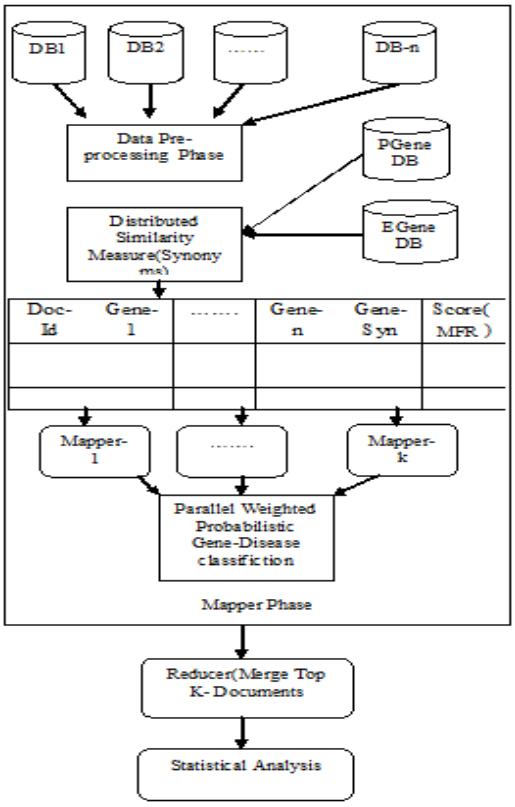

Figure 1: Proposed Model

Identifying the biomedical terms

Unified Medical Language System (UMLS) can be defined as a medical database that contains the superset of all biomedical terms. Thus, for detection of a biomedical term, UMLS database is searched. We have selected terms out of the simplified text files and a searching operation is performed in UMLS database.

In the data preparation phase, user specific PubMed and Embase documents are extracted along with disease types. In this phase, 1 million gene terms are extracted to find the relevant documents from the PubMed repository. NLP and text mining approaches have been applied to gene-disease based clinical documents to find the relevant contextual features.

The pseudo code for gene based PubMed data preparation in hadoop Mapper phase is given below:

Input: User query, Disease list, Gene DB

Output: Gene-Disease based Document sets.

Procedure: Map(TextList DiseaseList,Text Gene-DB, Text Query)

Connection con=PubMed(URL);

If(con!=null) then

Load GeneDB=getGeneNames();

Load DiseaseList=getDiseaseList();

end if

else

Check connection;

For each user – defined disease categories DC[] do
if (DC[i] ∈ DiseaseList)
then
getDocSet1[] = PubMed(DC[i]);
getDocSet 2[] = Embase(DC[i]);
getDocSet[] = CombineSets(getDocSet1, getDocSet2);

for(i = 0; i < getDocSet.length; i++)
do
Apply Stemming(getDocSet).
Apply Stopword_removal(getDocSet)
Remove Non-Functional Characters
Tokens[]=Tokenization(getDocSet,);

TrainGenes[] = Extract(getDocSet)
for(j = 0; j < TrainGenes.length; j++)
do
for(k = 0; k < geneDB.length; k++)
do
if (TrainGenes[j] == geneDB[k])
then

Add(PEDlist < –Map(GeneDList[j], Sim(U, V))
/ / Pubmed and Enbase Gene list with similarity score.
end if
done
done
Add GeneDocs(getDocSet[i], PEDlist,SimList)

GeneDlist[j] = TrainGenes[j];

$$Sim(U,V) \leftarrow \frac{\max |\mu_{pgeneDB}(k) - \mu_{pGeneDlist}(j)|}{\min |\sigma^2_{pgeneDB}(k) - \sigma^2_{pGeneDlist}(j)|}.P$$

where $P = \frac{Prob(TrainGenes[j] \cup getDocSet[i])}{Prob(geneDB[k] \cap TrainGenes[j])}$

Data Pre-processing Phase

In the data pre-processing phase, each document from the GeneDocs dataset is filtered to find the relevant features in the gene-disease based documents. Also, gene synonyms are discovered to each gene-type using the gene-synonym identifier. In this phase, gene-disease and gene-synonym-disease document sets are extracted in Mapper phase for document ranking process.

Input: Gene-Synonym DB(GENETAG-DB), GeneDocs document sets.

Output: Distributed Gene-Synonym Documents

Procedure:

// PubMed Gene-DB annotation
For each gene1 in PubMed DB
Do
For each gene2 in Embase DB
Do
Add(PEGENE-DB,gene1,gene2);
Done
Done
For each gene synonym in distributed PEGENE -DB
Do
For each GDoc in GeneDocs
Do
Geneslist[i]=GDoc;
For each gene token gt in Geneslist [i]
Do
getGSynonym[]=GeneSynonym(gt);
PGSynDocs[]=PubMed(Url(disease,getGSynonym[]);
EGSynDocs[]=Embase(Url(disease,getGSynonym[]);

Distributed Probabilistic Mutual information in textual information theory, is a measure to check the mutual independence and strongly correlated gene features among different distributed data repositories.

D_i = traindata[]; / / i = 1 to | traindata |

$$ProbabilisticFeatureRank(PFR) = Cr.\sum_{j=1} \log\left(\frac{Pro(\frac{D_i}{PgetGSynonym[j]})}{Pro(\frac{D_i}{EgetGSynonym[j]})}\right).Cond\,Prob$$

$$Cond\,Prob = (PgetGSynonym[j] \cap EgetGSynonym[j]);$$

$$Cr = \prod_{j=1}^{|getGSynonym[j]|} Correlation(D_i, PgetGSynDocs[j], EgetGSynonym[j]);$$

$Pro(D_i \cap getGSynonym[j])$ is the common probability of the D_i in the given gene synonym list.

$Pro(D_i \cup getGSynonym[j])$ is the sum of all probability of D_i to the given gene synonym list.

Pro(D_i) is the probability of the gene synonym documents D_i in the given Total documents list.

If(PFR>0.75)
Then Add GESynDocList(MFR,PgetGSynDocs, EgetGSynDocs);
end if
Done
Done

Parallel weighted probabilistic gene-disease classification Model

Automatic gene to disease classification is the widely used classification model in single repository, due to its simplicity, parameter estimation, and efficiency. However, they do not compete with statistical learning models on more than two biomedical repositories due to complex structures and high

dimensionality. Generally, naïve Bayesian and multinomial naïve Bayesian are applicable on large datasets such as News corpus, spam text etc, due to their high computational time and storage space. In order to improve the efficiency of these models, a parallelized version of the probabilistic naïve Bayes model was developed on a many-core GPUs. In this enhanced version, the main issues on high dimensional data include difficult to handle noisy and NULL values with high computational accuracy. Data classification using parallel naïve Bayes requires high features along with weighted features as training data. Let $W = \{w_1, w_2, ...w_n\}$ denotes the disease based weighted feature vector and $D = \{d_1, d_2, ...d_m\}$ represent the training GeneDocs dataset.

A biomedical document d(i) to be classified and assigned to a gene category by applying hybrid weighted probabilistic gene prediction model as:

$$\text{GeneClassify}\,\text{Prob}(c(k), d_{(i)}) = \prod_{j=1}^{N} w_{ij} . \frac{\text{Prob}(c(k)) * \text{Prob}(d_{(i)} / c(k))}{\text{Prob}(d_{(i)})}$$

Where $\text{Prob}(d_{(i)} / c(k))$ denotes likelihood estimator and $\text{Prob}(c(k))$ denotes prior probability

The weight of the gene feature within the document is computed as (1)

$$W_{ij} = \max\{\frac{f_{Pd}(i,j)}{F_{Pd}(i,j)} . \log(\frac{|D|}{f_{Ptf}}), \frac{f_{Ed}(i,j)}{F_{Ed}(i,j)} . \log(\frac{|D|}{f_{Etf}})\}$$

Where $f_{Pd}(i,j)$ represents the total occurrences of the i^{th} gene feature in j^{th} document of PuMed repository.

Where $f_{Ed}(i,j)$ represents the total occurrences of the i^{th} gene feature in j^{th} document of Embase repository.

$F_{Pd}(i,j)$ represents the total number of gene features occur in j^{th} document of PubMed repository.

$F_{Ed}(i,j)$ represents the total number of gene features occur in j^{th} document of Embase repository.

f_{Ptf} : represents the total number of documents in which i^{th} gene feature occurs at least once in PubMed repository.

f_{Etf} : represents the total number of documents in which i^{th} gene feature occurs at least once in Embase repository.

|D|: Total number of biomedical documents in both PubMed and Embase.

Algorithm:

Until training true positive of the gene improves in the document do

For each gene in the document, do

$\text{Cpredict} = \text{GeneClassify}\,\text{Prob}\left(\text{gene} - \text{disease},\ \text{document}\,(i)\right);$

if(Cpredict > Thres) / / Thres : user-defined threshold

then

addPositive(gene− disease, document(i));

else

add(Negative(gene− disease, document(i));

end if

done

return Mapper(k,v) ← (document-id,{gene-disease,genedocs,Cpredict)

Experimental Results

Experimental results are performed on biomedical repositories such as Medline, PubMed and Embase etc.The class accuracy rate A, F-Measure rate F and Recall Rate R measures are used to find the performance of the proposed algorithm on the document classification algorithm.

Accuracy Rate $A(D_c, F_i) = | D_c \cap F_i | / | D_c |$

Recall Rate $R(D_c, F_i) = | D_c \cap F_i | / | F_i |$

F-Measure Rate= $\sum_{i=1}^{k} | F_i | . \varphi(F_i) / \sum_{i=1}^{k} | F_i |$

where

$$\varphi(F_i) = \max_{i=1}^{k} (2.A(D_c, F_i).R(D_c, F_i) / (A(D_c, F_i) + R(D_c, F_i))$$

In this experimental study, a filtered based probabilistic gene-disease prediction model using hadoop framework is implemented on different biomedical datasets such as MEDLINE, Embase and PubMed repositories. The configuration of Amazon AWS server contains 10 CPU cores and 20 GB RAM to master and slave nodes. Large numbers of gene or protein MESH terms are used to classify the topmost document patterns from millions of records. Finally, Hadoop based framework is used to test the performance of hybrid ensemble classifier on biomedical disease documents.

ROC and F-measure are the commonly used performance metrics in classification models. However, due to noise and imbalanced problems, traditional receiver operating characteristics (ROC) and F-measure may not be a good choice. In this ensemble model, a novel phase wise accuracy measures such as geometric mean (GM) and Sum True positive rate are used to evaluate the performance of the proposed model to the traditional ensemble models.

Sample Gene/Protein Preprocessing and Matching Patterns

RPS11|AtCg00750;Chloroplast 30S ribosomal protein S11;CsCp075;Grc000081;9311100, Nip102,

PA102;PSC0809;PS158;ENSANGG00000009494;40S ribosomal protein S11;QnpA 10190;QnpA10190;CG8857;S11;anon MMS23;anonMMS23;anon fast evolving 1D2;anonfastevolving1D2;clone 23;fa91c09;fb34b11;MGC64491;cRPS11;rpsK;DDB0230027; ribosomal protein S11;ZFP318|D530032D06Rik; TZF;

AT4G25880|F14M19.160; F14M19_160;

B6|CG3100;NIP1-1|ZmNIP1 1;ZmNIP11;

SCO3608|SC66T3.19c;TRI5|TOX5;Trichodiene synthase;Sesquiterpene cyclase;TOX 5;FG03537;RP11-109G10.3|CTGLF4;LOC439975;METR|BUsg_030;HTH type transcriptional regulator metR;HTHtype transcriptional regulator metR;Z5349, ECs4758;b3828;HI1739;STY3595, t3333;STM3964;STMD1.26;SF3906, S_3849;ECK3822;JW3804;PSPTO4180;RS01772;

SCO4075|SCD25.11c;NPP-17|F10G8.3;Nucleoporin 17;Nucleoporin17;Nuclear pore complex protein 17;CeRAE1;

AT4G26490|M3E9.80;M3E9_80;PSPTO_5062|PSPTO5062;

PRSSL1|UNQ782;GLGL782;Df2;protease, serine like 1;protease, serinelike 1;LOC410682|GB16244;

AT4G32710|F4D11.90;F4D11_90;D6WSU116E|A130095H06;C530005J20Rik;POG1|YIL122W;LOC552513|GB18849;

AT1G64560|F1N19.29;F1N19_29;ACS|Z5668, ECs5051;Acetyl coenzyme A synthetase;Acetylcoenzyme A synthetase;Acetate CoA ligase;AcetateCoA ligase;Acyl activating enzyme;c_5064;b4069;STY4473, t4181;STM4275;YPO0253,y0510, YP0406;EG11448;acsA;yfaC;facS;

BC052883|Gm1065;MGC60753;

AT5G50110|MPF21.12;MPF21_12;

1;GMP1;Sentrin;fb74c02;zeh0670;MGC89967;UBL1;PIC1;SENP2;SMT3;SMT3C;SMT3H3;MGC128420;MGC103203;SMTP3;MGC109561;SMALL UBIQUITIN LIKE MODIFIER 1;SMALL UBIQUITINLIKE MODIFIER 1;UBIQUITIN LIKE 1;UBIQUITINLIKE 1;SMT3, YEAST, HOMOLOG 3;SMT3 suppressor of mif two 3 homolog 1;

LOC725151|GB15566;LOC409026|GB15575;

SCO6921|SC1B2.27c;DDBDRAFT_0188533|DDB0188533;

Gene- Disease Context Similarity using gene/protein Synonyms

HG-4036-HT4306=>synonyms are Retinoblastoma1

Gene-Disease Context-Similarity =>0.4166666666666667

HG-4051-HT4321=>synonyms are Choline Acetyltransferase

Gene-Disease Context-Similarity =>0.23500000000000001

HG-4052-HT4322 => synonyms are Glutamate Ionotropic Receptor 1

Gene-Disease Context-Similarity =>0.31698028673835127

HG-4058-HT4328 => synonyms are "Oncogene Aml1-Evi-1, Fusion Activated"

Gene-Disease Context-Similarity =>0.3920940170940171

HG-406-HT406 => synonyms are "P97 Antigen, Melanoma-Specific"

Gene-Disease Context-Similarity =>0.2351190476190476

HG-4068-HT4338 => synonyms are Phosphoprotein Tal2

Gene-Disease Context-Similarity =>0.33735380116959063

HG-4073-HT4343 => synonyms are Cytosolic Acetoacetyl-Coenzyme A Thiolase

Gene-Disease Context-Similarity =>0.3091124661246612

HG-4074-HT4344 => synonyms are Rad2

Gene-Disease Context-Similarity =>0.0

HG-4102-HT4372 => synonyms are N-Ethylmaleimide-Sensitive Factor

Gene-Disease Context-Similarity =>0.31502525252525254

HG-4114-HT4384 => synonyms are Olfactory Receptor Or17-209

Gene-Disease Context-Similarity =>0.3996913580246913

HG-4126-HT4396 => synonyms are Zinc Finger Protein Hzf4

Gene-Disease Context-Similarity =>0.3680555555555556

HG-4128-HT4398 => synonyms are "Anion Exchanger 3, Cardiac Isoform"

Gene-Disease Context-Similarity =>0.22685185185185186

HG-4144-HT4414 => synonyms are Zinc Finger Protein Hzf6

Gene-Disease Context-Similarity =>0.3680555555555556

HG-415-HT415 => synonyms are "Lectin, Galactoside-Binding, Soluble, 2"

Gene-Disease Context-Similarity =>0.3777584204413473

HG-825-HT825 => synonyms are "Guanine Nucleotide-Binding Protein, Alpha 12"

Gene-Disease Context-Similarity =>0.37422360248447206

HG-830-HT830 => synonyms are Potassium Channel (Gb:L02750)

Gene-Disease Context-Similarity =>0.3281335522714833

HG-831-HT831 => synonyms are Potassium Channel (Gb:L02752)

Gene-Disease Context-Similarity =>0.2372742200328407

Table 1: Performance of the proposed probabilistic gene-disease classification model on different distributed document sets.

Algorithm	1-lakh (Docs)	2-lakh (Docs)	3-lakh (Docs)	5-lakh (Docs)	10-lakh (Docs)
Hierarchical MDT	0.697	0.698	0.706	0.687	0.709
SVM	0.739	0.712	0.7987	0.719	0.748
Naïve Bayes	0.716	0.799	0.724	0.709	0.799
Neural Networks	0.796	0.794	0.779	0.749	0.756
PSO+ NaïveBayes	0.871	0.832	0.889	0.9171	0.924
Proposed Method	**0.952**	**0.9598**	**0.9624**	**0.9698**	**0.9719**

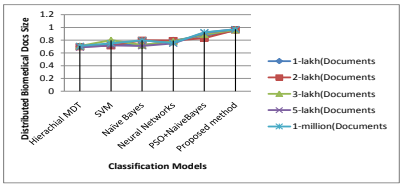

Figure 2: Performance of the proposed probabilistic gene-disease classification model on different distributed document sets**.**

Table 1 and Figure 2 represent the classification efficiency of the proposed gene-disease classification to the traditional models on distributed biomedical repositories. From the table 1, it is clear that the as the size of the document sets increases the accuracy and true positivity of the gene-disease prediction in the documents increases.

Table 2: Comparison of Gene-Disease Prediction Score in traditional and proposed similarity measure**.**

Top K-Relational Gene-Disease Prediction Score w.r.t Threshold				
Data Size	**Thres =0.5**	**Thres =0.75**	**Thres =0.8**	**Thres =0.9**
Hierarchical MDT	0.824	0.8165	0.8663	0.8988
SVM	0.798	0.824	0.897	0.904
Naïve Bayes	0.897	0.899	0.907	0.917
Neural Networks	0.879	0.889	0.923	0.938
PSO+NaïveBayes	0.879	0.882	0.904	0.921
Proposed Method	**0.926**	**0.9289**	**0.9319**	**0.9547**

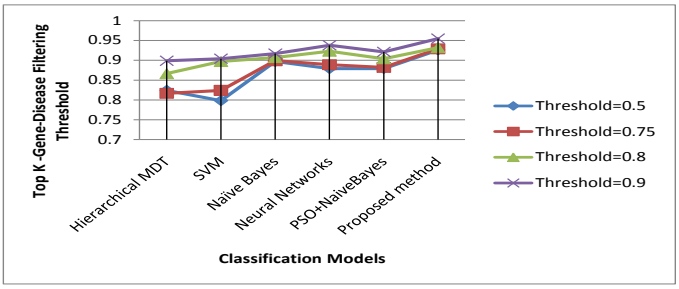

Figure 3: Comparison of Gene-Disease Prediction Score in traditional and proposed similarity measure**.**

Table 2 and Figure 3 illustrate the performance of the gene-disease prediction rate using proposed similarity measure on distributed documents sets. Here, a novel probabilistic similarity measure is used in proposed classification and traditional classification models for comparative analysis using Hadoop framework. From the results, it is clear that the proposed model has high computational prediction rate compared to the traditional models.

Table 3: Comparison of the proposed classification model and traditional models in terms of memory and time.

Computational results of Proposed and Existing Models			
Algorithm	**Average Memory (MB)**	**Average Time (ms)**	**Top Gene-Disease Patterns (%)**
Hierarchical MDT	2765	15723	27.89
Naïve Bayes	2846.23	18576	24.65
Neural Networks	2788.56	18876.9	27.88
PSO+NaiveBayes	2619.77	17853.8	24.19
Proposed Method	**2343.34**	**15878.8**	**21.98**

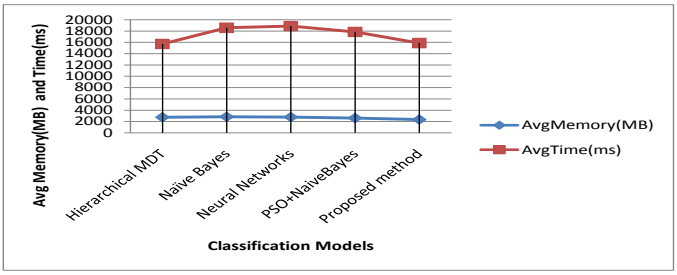

Figure 4: Comparison of the proposed classification model and traditional models in terms of memory and time.

Table 3 and Figure 4 describe the performance results of proposed distributed gene classification model in terms of memory, time and top-k document set percentage. From the table, it is clear that the proposed model gets high classification accuracy, time and memory parameters for high dimensional datasets using the Hadoop framework.

Conclusion

Raw medical abstracts are large in number and it is very difficult to process manually. Detecting and predicting diseases at early stage is very much essential for both healthcare professionals and patients. As the amount of information in the biomedical repositories increases, document preprocessing, ranking, classification and scalability have become a major issues for distributed databases. In this paper, a novel probabilistic gene-disease based document classification model is implemented on the multiple integrated biomedical databases such as MEDLINE, Embase and PubMed repositories. This model is used as a user recommended system on the large document sets using the Hadoop framework. Experimental results show that the proposed model has a high computational classification rate(~96%) and prediction rate(~95%) compared to traditional document classification models. In future, this work can be extended to gene-disease classification and gene clustering model using Hadoop framework for distributed repositories.

REFERENCES

[1] F. Ö. Çatak, "Classification with boosting of extreme learning machine over arbitrarily partitioned data", "Springer Conf. on Soft Comp.", 2015.

[2] B. Chandra and M. Gupta, "An efficient statistical feature selection approach for classification of gene expression data ", "Journal of Biomedical Informatics 44 (2011)", pp. 529-535, 2011.

[3] J. Chen, G. Zheng and H. Chen, "ELM-MapReduce: MapReduce Accelerated Extreme Learning Machine for Big Spatial Data Analysis", "10th IEEE International Conference on Control and Automation (ICCA)", pp. 400-405, 2013.

[4] S. Das and A. K. Das, "Gene selection and decision tree based classification for cancerous sample detection ", "Int. J. Biomedical Engineering and Technology, Vol. 21, No. 1", pp.1-14, 2014.

[5] H. I. Elshazly, A. M. Elkorany, A. E. Hassanien and A. T. Azar, "Ensemble classifiers for biomedical data: performance evaluation", pp. 186-189, 2013.

[6] X. Fei, X. Li and C. Shen, "Parallelized Text Classification Algorithm for Processing Large Scale TCM Clinical Data with MapReduce ", "Proceeding of the 2015 IEEE International Conference on Information and Automation Lijiang, China, August 2015", pp.1983-1986, 2015.

[7] H. Hu, "Mining patterns in disease classification forests", "Journal of Biomedical Informatics 43 (2010)", pp. 820–827, 2010.

[8] D. Kiela, Y. Guo, U. Stenius and A. Korhonen, "Unsupervised Discovery of Information Structure in Biomedical Documents", "Bioinformatics Advance Access published November 18, 2014", pp. 1-8, 2014.

[9] M. Kumar and S. K. Rath, "Classification of Microarray using MapReduce based Proximal Support Vector Machine Classifier", "Preprint submitted to Elsevier", pp. 1-35, 2015.

[10] F. Lin, C. Shen, C. Liu, H. Lin, C. F. Huang, C. Kao, F. Lai and J. Lin, "High-Performance Multiclass Classification Framework Using Cloud Computing Architecture", "Journal of Medical and Biological Engineering", pp.795-802, 2015.

[11] H. Liu, L. Liu and H. Zhang, "Ensemble gene selection by grouping for microarray data classification", "Journal of Biomedical Informatics", pp. 81-87, 2010.

[12] Y. Liu, L. Xu and M. Li, "The Parallelization of Back Propagation Neural Network in MapReduce and Spark", "International Journal of Parallel Programming", 2016.

[13] I. Palit and C. K. Reddy, "Scalable and Parallel Boosting with MapReduce", "IEEE Transactions On Knowledge And Data Engineering, VOL. 24, NO. 10, October 2012", pp. 1904-1916, 2012.

[14] S. Sumathipala, K. Yamada and M. Unehara, "Protein Named Entity Classification with Probabilistic Features Derived from GENIA Corpus and MEDLINE", "SCIS&ISIS 2014, Kitakyushu, Japan, December 3-6, 2014", pp. 1257-1261, 2014.

[15] J. Zhai, S. Zhang and C. Wang, "The classification of imbalanced large data sets based on MapReduce and ensemble of ELM classifiers", "International Journal of Machine Learning and Cyber Security", 2015.