Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

A contemporary feature selection and classification framework for imbalanced biomedical datasets

Thulasi Bikku [a,*], Sambasiva Rao Nandam [b], Ananda Rao Akepogu [c]

[a] Department of CSE, Vignan's Nirula Institute of Technology and Science for Women, Palakaluru, A.P, India
[b] Principal, RITW, Hyderabad, Telangana, India
[c] Director of Academics & Planning, JNTUCEA, Ananthapuramu, India

ARTICLE INFO

Article history:
Received 21 March 2017
Revised 16 October 2017
Accepted 22 March 2018
Available online 29 March 2018

Keywords:
Biomedical data
Document clustering
Document classification
Bioinformatics
User recommended system

ABSTRACT

Due to the availability of a large number of biomedical documents in the PubMed and Medline repositories, it is difficult to analyze, predict and interpret the document's information using the traditional document clustering and classification models. Traditional document clustering and classification models were failed to analyze the document sets based on the user's keyword and MESH terms. Due to the large number of feature sets, conventional models, such as SVM, Neural Networks, Multi-nominal naïve bayes have been used as feature classification, where additional text filtering measures are typically used as feature selection process. Also, as the size of the document's increases, it becomes difficult to find the outliers using the document's features and MESH terms. Biomedical document clustering and classification is one of the essential machine learning models for the knowledge extraction process of the real-time user recommended systems. In this paper, we developed a novel biomedical document feature clustering and classification model as a user recommended system for large document sets using the Hadoop framework. In this model, a novel gene feature clustering with ensemble document classification was implemented on biomedical repositories (PubMed and Medline) using the MapReduce framework. Experimental results show that the proposed model has a high computational cluster quality rate and true positive classification rate compared to traditional document clustering and classification models.

1. Introduction

Clustering can be defined as the phenomenon of managing data objects into group of classes, which are disjoint in nature. Objects in a particular cluster share the same properties, whereas, object in different clusters share dissimilar properties. Document clustering are responsible for auto-grouping of documents into groups. Clustering is a type of unsupervised classification. Unsupervised classification is the process of classification which does not interact with previously defined classes and training sets. Bioinformatics, pattern recognition, image processing and data mining are some most common applications of clustering [1]. Data mining has an important tool for data analysis which is named as Cluster analyst.

The whole process of clustering can be split into two categories, these are: Hierarchical algorithm and Partition algorithm. Hierarchical algorithm decomposes the dataset into clusters gives rise to a child clusters in a hierarchical pattern, which is responsible for the generation of clusters named as a dendrogram. Each cluster gives rise to a child cluster. Partition algorithm is responsible for decomposition of datasets into smaller units in a step. Hierarchical algorithm can be further divided into two, those are: Agglomerative and Divisive clustering. Agglomerative clustering involves generation of singleton cluster and later it combines more clusters recursively [2]. Divisive clusters initiate with a single cluster and recursively divided into more clusters. This loop aborted when termination condition is satisfied.

There are two major demerits of clustering process, those are: 1. Calculating optimal cluster numbers. 2. Calculating relative goodness among two different clusters. Criterion and validation functions are essential to achieve all the advantages of clustering. In

[*] Corresponding author.
E-mail address: thulasi.jntua@gmail.com (T. Bikku).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

ELSEVIER Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2018.03.003
1110-8665/© 2018 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University.

the traditional clustering schemes, instances with similar properties are grouped under same cluster, whereas; instances having dissimilar properties are grouped under different category of clusters. These types of clustering schemes are known as hard clustering schemes. In case of soft clustering schemes, instances may be grouped in different clusters simultaneously.

Different clustering schemes are:- hierarchical schemes, vector quantization schemes, mixture density-based schemes, graph theory-based schemes, combinational search techniques-based schemes, fuzzy schemes, kernel-based schemes, and so on.

The most frequent applications of document classifications are noticed in the field of biomedicine and research domain. The process of classification is broadly categorized into two parts, those are: training and testing phase. Classification algorithm is responsible to create a classification model with the help of the training set. Later the model performance is evaluated in the testing phase. Many research works have been proposed since years in order to develop a classification algorithm with optimized performance. Some of the popular classification models [3] are briefly described below.

The Naive Bayes model is more efficient than that of the logistic regression scheme, nearest neighbor, decision tree and neural network, according to Receiver Operating Characteristic (ROC) curve, which is more significantly implemented in the research domain. The model is a simple, less parameterized and efficient one in terms of performance.

KNN is a special type of instance-based classification model, in which approximation function is computed locally and it continues until classification occurs. It is also termed as lazy learning because there is no need of training phase like other conventional approaches. These training data are tested in the testing phase. For large datasets, training data are split into smaller partial datasets. The most common application of KNN is implemented in density estimation process and parametric estimation process. Hidden Markov Model (HMM) is a type of statistical classification model, in which the whole model is taken as a Markov process and the Markov process is having unobserved or hidden states. This model is represented as a dynamic Bayesian model. Some applications of this model are: Face recognition and Detection process. The only flaw of this approach is it ignores state structures in super states.

Support Vector Machine is a construction-based classification model. It also follows the idea of statistical learning [4]. This algorithm achieves enhanced performance as compared to other existing approaches. With the help of hyperplanes, SVM defines decision boundaries. It also distinguishes data points of different classes, which can solve both linear and nonlinear classification problems. A mapping function is invoked for mapping of the original data points from the input space to a high-dimensional or infinite-dimensional feature space through a kernel function. Relevance Vector Machine (RVM) is an enhanced version of SVM. This model provides better performance rate when compared with other models including SVM. It emphasizes on sparsity and compressed sensing. This approach uses subsets of the training data of SVM which are fewer than that of support vectors.

In decision trees classification models, divide-and-conquer is followed to create a tree, in which instances are combined with other attributes. Decision tree contains nodes and a leaf node with the output of this algorithm is either true or false. Both the path and nodes are considered to form pre-conditions so constraints are formed by the path from root to leaf. Tree pruning helps to discard unwanted preconditions and redundancy. One of the best among classification algorithms is Random forest algorithm classifies the huge amount of data with high accuracy. In decision trees the models are generated by creating a large number of decision trees. The result contains modal class which is predicted by individual trees. It uses a concept of merging weak learners to form strong learners [5].

Genetic Algorithms (GA) are categorized as evolutionary and stochastic algorithm which gives an optimal solution. Crossover and mutation operations are performed to encode candidate solutions. To create offspring, solutions are selected according to the fitness function. The initial population is randomly constructed; every candidate solution is evaluated to gain a fitness score at each generation. In hierarchical clusters, decomposition occurs and smaller datasets are formed with respect to the different hierarchical pattern. Clusters form child clusters or leaf clusters are constructed which are known as Dendograms. The process of document classification plays a vital role in recent days. For decades the vast amount of research has been carried out to propose a new integrated approach combing document clustering and classification to achieve optimized performance rate.

Hadoop is a reliable, distributed and scalable open source software framework that is suitable for scalable and parallel programming on a large amount of data. This framework is designed and implemented to automatically divide large data into small segments, each of which can be executed on any cluster node and to handle node failures efficiently. The Hadoop framework contains two essential components: Hadoop Distributed File system and MapReduce. HDFS can store large data and partition these data into smaller blocks of 64 MB each. MapReduce mechanism consists of two sub-phases: Map and Reduce. The input partitioned chunks are given to mapper phase for parallel processing. The output of the mapper phase is given to reduce phase to generate the output. Since a single individual machine has restricted an amount of processing power and enhancing its processing power will be more expensive. To improve the performance of the single machine processor, an efficient Map-Reduce framework is used to develop large-scale applications, which are independent of hardware and distributed environment.

The main contribution of the proposed model is included, the gene based features clustering and classification, which is responsible for construction of predictive classification models. In this model, feature based relational gene clusters are used to find the probability of a document belonging to a specific gene related clusters. If a document has a low similarity index in one cluster, then it is grouped in another cluster which is having a high similarity index. There is an effective index which helps in extraction of related documents from previous index. The proposed method reduces the estimated cost of clustering and classification.

2. Related works

Rojcek [1] developed a new technique to achieve uncontrolled fuzzy clustering and fast fuzzy classification model on limited dataset. They implemented the concept of KMART neural network for document classification, which is an innovative approach used to integrate clustering with fuzzy classification. Clustering as well as classification algorithms share common initial weights. This uncontrolled system is based on two basic methods, those are: 1. involves plasticity, 2. involves stability. This approach can be implemented on classification of labelled ones with limited feature set. These methods are formed without affecting the pre-defined structure (stability) of the training set. This model was tested on small datasets and the algorithm shows better performance than that of existing conventional approaches for document clustering and classifications with limited instances and dimensions. The main limitations in this model include high misclassification rate and less true positive rate on the high dimensional datasets.

Aïtelhadj et al. [2] proposed a new technique in order to cluster XML documents. The main objective is to share common structures to multiple Hadoop structures. The proposed model was executed in two steps: In the first step, XML documents are classified using the features automatically and these structured documents are pre-processed as a model for classification. XML documents with common structure are more frequently cluster the structure of the same query. They evaluated on the real and synthetic data and implemented on both XML and semi-structured documents. As the size of the documents and clusters are increasing, mean cluster error rate and the classification rate decreasing. Additional large feature set documents are added in order to increase contents, so that search engine processing will be improved for query processing.

Chan and Chong [3] gave emphasize on non-text based classification scheme and identified the most common issues in the classification model. They classified hypertext-based non-textual information for the purpose of classifying unknown documents. In the traditional approaches of classification, many search engines processes only text-based documents and unable to retrieve the graphic or diagrammatic documents. In this paper, a new algorithm for non-textual diagrammatic document classification was implemented on document sets to analyze the similarity index of the documents. Various feature vectors for classification model are presented and tested. This method analyzes and compared with the other methods (like HAC) and showed that the proposed approach is far better than that of others in terms of performance. The main issue in the non-textual based classification scheme is, this model failed to process large number of document sets for clustering and classification.

Curtis et al. [4] proposed a novel unsupervised model to handle huge volume of documents, also it is quite hard for the algorithms to retrieve and classify massive documents. This model is based on supervised learning, but the idea of unsupervised algorithm is more feasible and effective one for large scale data. A large number of clustering schemes have been developed in the last few years using unsupervised learning. In this model, a novel two-threshold method for hierarchical decomposition of features was implemented on un-structured datasets with missing values and class labels. It discards all the clusters that do not belong to the same cluster. The main problems in this model include the updating of static threshold and parameter initialization in each iteration.

Dai et al. [5] analyzed various real-world applications and identified limitations of uncertainty and imbalanced nature. Heterogeneous data extraction and classification from different sources is very costly and time taking process. The conventional machine learning approaches do not perform effectively due to imbalance and uncertainty. In this work, they categorized the labelled data in a separate domain and termed it as in-domain data, whereas, the unlabeled data are categorized under out-of-domain data. Their main objective was to extract the knowledge from in-domain and implement it on out-of-domain. They proposed a Co-Clustering-based Classification scheme (CoCC) to improve the clustering and the classification rate in between in-domain and out-of-domain. They performed empirical analysis and evaluated their work and analyzed that this model can't handle massive documents with dynamic parameter pruning.

Diaz-Valenzuela et al. [6] identified the major issue of semi-supervised approaches and proposed a novel auto-supervised model. In order to overcome the limitations of semi-supervised approaches, they proposed a new method with auto-generation of data structure. Partition based clustering is used to detect the relationships features, among various instances. In the document clustering model, this approach is validated and evaluated in two stages. The first stage of this algorithm contains a k-means algorithm which helps to cluster must-link or cannot-link constraints, and the subsequent stage uses these constraints with input data in semi-supervised clustering. The major problem identified in this work includes dynamic clustering measures need to be used to enhance the fuzzy constraints and to improve the performance rate.

Hachenberg and Gottron [7] developed a new approach to achieve structural classification and clustering by an innovative template based feature extraction technique. This technique is used to find the locality-based hashing and compression. Compression schema depends on locality-based hashing and document's tag; this approach is tested on 13,000 documents with limited dimensions. In this model, runtime complexity does not depend on the size of the training set. Traditionally, Medical document classification algorithms based on the Medical Subject Heading (MeSH) are used to classify document indexing and making a filter based search engine for biomedical repositories. MeSH, a "controlled vocabulary corpus ", was developed by the National Library of Medicine (NLM) to organize document key terms in a hierarchical MeSH tree structure that cover the various domains such as medicine, dentistry, nursing, pre-clinical terms and health-care system [8]. Each MeSH tree structure facilitates document term searching from the root level to the leaf level. As the size of the MeSH term descriptors increases, corresponding hierarchical tree structures are also growing exponentially and it is difficult to discover important patterns using traditional document classification models.

Ke [9] proposed a new technique for automated text classification with least document feature representation. They implemented a Least Information Theory (LIT) for document feature extraction. Shannon entropy is enhanced for non-linear relation between information and uncertainty. LIT is a type of information-based method to weight probabilistic distribution. There are two weight measures used for classification, those are: LI Binary and LI Frequency, which are evaluated independently in the early stage. The researcher experimented and evaluated this method with three benchmark collections, which provides significant performance (LIB $*$ LIF) as compared to other methods (TF $*$ IDF).

Lin and Chen [10] developed a new genre classification scheme for musical documents with different melodic patterns consisting of keywords, statistical and low-level features. The proposed classification depends on correlation analysis of musical patterns grouped by appropriate clustering is implemented. Performance is increased remarkably by smoothing methods. They considered five different types of musical patterns (such as jazz, lyric, rock, classical) are transformed into symbolic patterns and achieved 70.67% of accuracy. The patterns are transformed into symbolic patterns. In biomedical document classification, K-nearest neighbor technique is a special type of machine learning model based on the document feature extraction and key phrase extraction; it does not build a training model until the new instances are classified. The basic idea of the KNN model is to classify a new document based on the similarity measure between the document feature vector and the training document set. After training the model, it gets K documents as majority voting based on the highest similarity measure.

Nam and Quoc [11] proposed a hybrid approach by merging cluster-centric filter feature selection with frequency-centric filter feature as a Frequency Cluster Feature Selection (FCFS). Due to this integration, FCFS chooses MeSH terms, so that it is frequently appears inside each category. They chose datasets from two distinguished domains, which are: news and medicine databases. The main limitations of this work include an enhancement of the intra-category feature of FCFS, is gained by optimization of closeness of documents.

Shruti and Shalini [12] proposed a new algorithm known as Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm

(FRECCA) for identifying interrelated sentence clusters, which uses fuzzy clustering for this method. It compares the chosen sentences and calculates a similarity index. Further works may be done in this algorithm by extending it to gain better performance than that of FRECCA. Genetic algorithm takes input data sets according to random classes. After processing the resulting solution must contain strongly interlinked clusters.

Hadoop Distributed File System is managed by a master process known as NameNode, it decides which chunk is assigned to which server and all information is maintained sequentially. In slave process, DataNodes interacts with the operating system by logically dividing data into fix sized regions. All these regions contain HFiles which are capable of storing datasets located in columns sequentially. When a particular region exceeds the minimum size limit, it is divided into two parts. The operations like integration and decomposition are carried out through background processes and never affect the normal behavior. These approaches permit HBase to build a random-access database over a file system which supports sequential reading of data. ElasticSearch (ES) can be defined as a software framework which depends upon Lucene and provides distributed as well as real-time search on large datasets. An index is a special kind of data structure which stores the HDFS data in an efficient format. The Lucene indexing model provides a better solution for building efficient and scalable indexes.

Motivation of the proposed model:

- Traditional models failed to classify the high dimensional documents with different initialization parameters.
- Traditional bio-medical document classification models are evaluated on the limited datasets with limited memory constraints and runtime factors.
- Gene similarity based biomedical document classification models were tested on static training gene-sets without symbolic gene-names.
- As the size of the uncertainty and noisy documents increases, the accuracy and true positivity of the traditional models decreases due to memory constraints and lack of optimal feature selection measures.

3. Proposed model

The proposed model concentrated on finding gene features identification with the document clustering and classification shown in Fig. 1. This model finds quality gene-disease information from the unstructured data. By merging rich document representations, it resolves textual classification issues in machine learning. Gene based vector representation helps as a measure of semantics. The proposed model optimizes the gene based features clustering and classification, which is responsible for construction of predictive classification models. In this model, feature based relational gene clusters are used to find the probability of a document belonging to a specific gene related clusters. If a document has a low similarity index in one cluster, then it is grouped in another cluster which is having a high similarity index. There is an effective index which helps in extraction of related documents from previous index. The proposed method reduces the estimated cost of clustering and classification.

In this proposed model, medical documents are collected from each data source as structured or unstructured format. In this system, a novel gene based document's relationships are analyzed using Map-Reduce framework to discover the textual patterns from a large collection of medical document sets.

3.1. Gene feature selection measures

Gene-disease feature selection is performed using the proposed feature selection methods as:

Gene mutual information measure

Gene based document features are extracted using the gene mutual information measure. This measure is implemented in the Mapper phase. As the size of the biomedical documents increases, this measure finds the filtered candidate sets with highest probabilistic measure. This measure takes clustered documents as input and finds the topmost features from inter and intra clustered document sets.

Gene mutual information can be computed using the following formula.

$$\mathrm{GeneMI\ (GMI)} = \max\left\{\mathrm{Prob(GDW_i/WVD_j)}\cdot\log\sum\sum\left(\frac{\mathrm{Prob(GDW_i/Clusters[m])}}{\mathrm{Prob(Clusters[m])}}\right)\right\} \tag{1}$$

$\mathrm{Prob(GDW_i/Clusters[])}$ is the probability of the $\mathrm{GDW_i}$ weighted genes that exist in the given documents clusters.

Gene Chi-square measure

Gene based chi-square measure computes the relationship between the biomedical document clusters. This measure also evaluates the dependency between the gene term to the MeSH term of the clustered documents.

$$\mathrm{GCHI} = \frac{\mathrm{Prob(GDW_i/Clusters[m])}\left[\max\{\mathrm{Prob(GDW_i/WVD_j)}\}.\max\{\mathrm{Prob(GDW_i/\overline{WVD_j})}\}\right]}{\prod\mathrm{Prob(GDW_i)}\times\prod\mathrm{Prob(Clusters[m])}} \tag{2}$$

where $\mathrm{Prob(GDW_i/Clusters[m])}$, the probability of the gene is related feature vector that appear in the mth cluster.

3.2. Proposed MapReduce algorithm

```
Input: Medline Source MS, Pubmed Source PS.
 $MD\{D^1_{MD}, D^2_{MD}, D^3_{MD}, \ldots D^p_{MD}\}$represents medline training documents
 $PD\{D^1_{PD}, D^2_{PD}, D^3_{PD}, \ldots D^q_{PD}\}$represents PubMed trianing documents,
 $\varphi$ denotes threshold
 VD : Vector Document set
Output: Gene based Entity Identification and Clustering
// Mapper Phase
Mapper $(MD\{D^1_{MD}, D^2_{MD}, D^3_{MD}, \ldots D^p_{MD}\}, PD\{D^1_{PD}, D^2_{PD}, D^3_{PD}, \ldots D^q_{PD}\}, \lambda, Q)$
1. Let $M[] = Mapper(nodes(), D_i)$
2. $D = (D^1_{MD}, D^2_{MD}, D^3_{MD}, \ldots D^p_{MD}, D^1_{PD}, D^2_{PD}, D^3_{PD}, \ldots D^q_{PD})$
3. For each document in D do
       $VecDoc(D_i, VD_i) = Doc2Vec(D_i)$;//Document to vector representation
   Done
```

```
4. For each Vector v in $VD_i$ do
      Represent each document in the form of <VecDoc($D_i$),tf, itf, W>
      //Replace missing values
    $\mathrm{if}(v == \phi)$
    then
    For each term t in  $\mathrm{VecDoc}(D_i)$
    do
    $\mathrm{simv}[] = \mathrm{Prob}(t/VD) * \mathrm{Prob}(t)$
    done
    $\mathrm{VecDoc}(v) = \max\{\mathrm{simv}[]\};$
    end if
      $VecWgt = \log(\sum \mathrm{VecDoc}(D_i)/(tf)) * itf$; //vector weight,
  term frequency and inverse term frequency
      //Weighted Vector Document
      $WVD_i$(1…VecDoc[i].length) $\leq$ VecDoc($D_i$),tf,itf,VecWgt >
      Done
5. // Computing the similarity between the Gene documents.
    Let Gene Database and Gene Synonyms are represented as
    $G[] =$ GeneDB Corpus
    $Syn[] =$ SynDB Corpus
    for each weighted vector term $WVD_i(t)$ in $WVD_i$ do
    for each gene keyword j in G[] do
    if($WVD_i(t)$==G[j])then
    //compute gene weight as
    Genew= $(\mathrm{prob}(tf_i \cap G[j])/\mathrm{prob}(tf_i)) * WVD_i(\mathrm{VecWgt})$
    Selected Gene with weights
    GW < $WVD_i$,t,Gene weight>=< $WVD_i$,t,Genew>
    else
    continue;
    end if
    end for
6. for each weighted vector term $WVD_i(t)$ in $WVD_i$ do
   // Document wise total genes weight
   TGW= $\sum$Genew/N;
   if(TGW > $\varphi$) // Total genes weight must satisfy the threshold
   Then
     // Extract synonyms of the high weighted gene
     for each synonyms s in Syn[] do
     SynList < $WVD_i$(t),S[ ]>=<$WVD_i$(t),Syn[$WVD_i$(t)]>
     SGenew $(\mathrm{prob}(s_i \cap G[j])/\mathrm{prob}(s_i)) * WVD_i(\mathrm{VecWgt})$
     Done
     $\mathrm{NewGenew} = \mathrm{TGW} + \sum \mathrm{SGenew}/N$
     $\mathrm{GDW} < WVD_i, \mathrm{TotalGenesWeight(TGW)} >=< WVD_i, \mathrm{NewGenew} >$
   else
     Continue;
end for
7. $\mathrm{Clusters}[] = \mathrm{AgglomerativeClustering}(WVD_i, \mathrm{GDW}, G[], \mathrm{Syn}[])$
8. $\mathrm{Reducer} < WVD_i, \mathrm{GDW}, G[], \mathrm{Syn}[], \mathrm{Clusters} >$
   Select top gene features using the Gene based Chi-square
   Measure, Gene Based Mutual Information to the Clustered Gene data.
9.For each cluster instances $GWD_i$ in the mapper $M_i$ do
   Build the modified multi-nominal naïve Bayesian tree to each $GWD_i$.
   for each attribute $A_i$ in the $GWD_i$ do
   Compute the attribute selection measure using the integrated feature selection measure as:
   $A_{max} = \mathrm{Max}\{\mathrm{GeneMI}, \mathrm{GCHI}\}$
   Let $A_{max}$ be the attribute with the maximum attribute selection measure.
   Construct the MNB tree using the maximum attribute $A_{max}$.
   Return the Mapper $M_i$ tree to the Reducer phase Reduce < $M_i$, return MNB($GWD_i$)>.
   done
   done
```

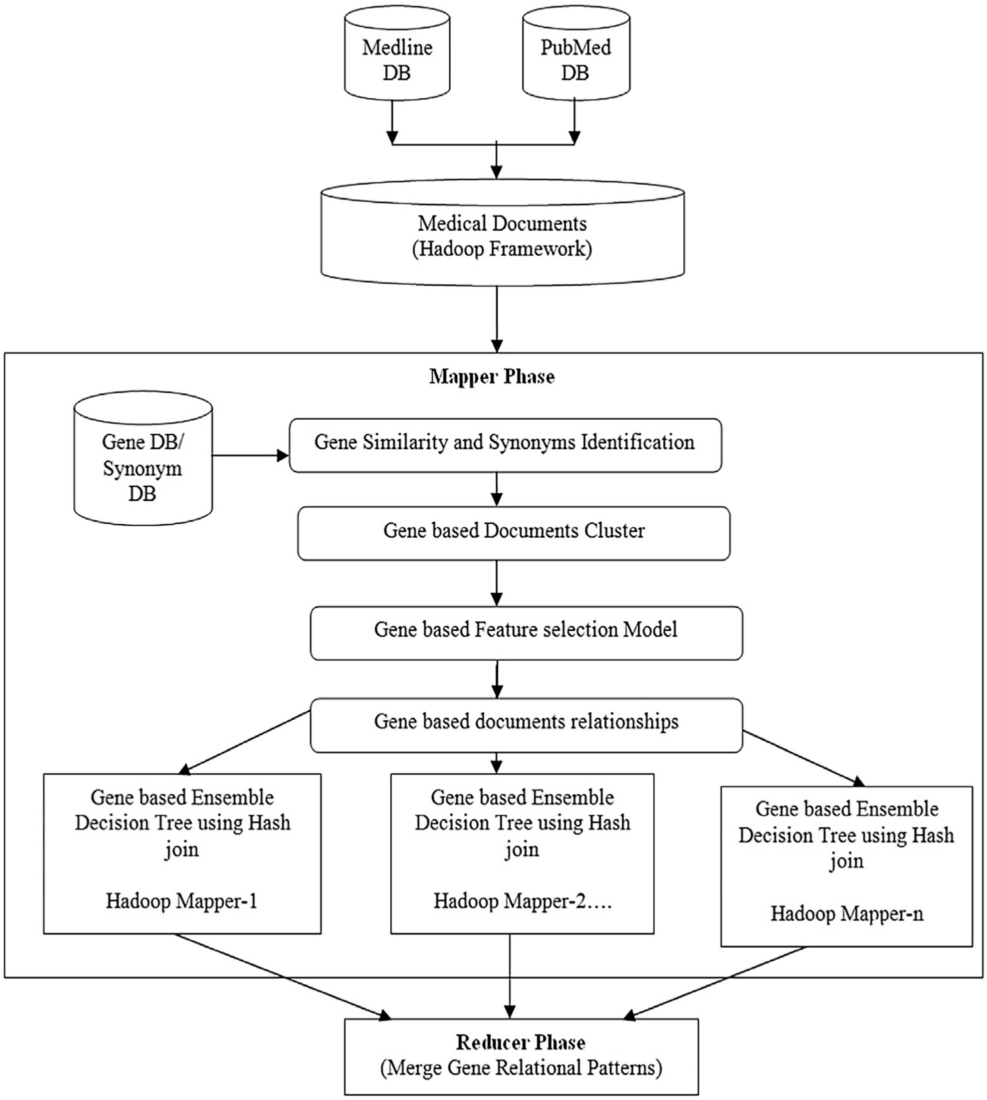
Medline DB
PubMed DB
Medical Documents (Hadoop Framework)
Mapper Phase
Gene DB/ Synonym DB
Gene Similarity and Synonyms Identification
Gene based Documents Cluster
Gene based Feature selection Model
Gene based documents relationships
Gene based Ensemble Decision Tree using Hash join Hadoop Mapper-1
Gene based Ensemble Decision Tree using Hash join Hadoop Mapper-2….
Gene based Ensemble Decision Tree using Hash join Hadoop Mapper-n
Reducer Phase (Merge Gene Relational Patterns)

Fig. 1. Proposed model.

In the Mapper phase, each biomedical document is processed to find the sparsity and null values. Initially, training dataset D is prepared using the Medline and PubMed repositories. Each document in the training data is represented in vector format. Here, document gene, disease and MeSH terms are extracted using gene term entity discovery procedure. For each document in the vector document set, term similarity and weights are computed in vector format as weighted vector document sets WVD. In the next step, the similarities between the Gene documents are computed using gene database and gene-synonyms extraction method. Documents with highest gene-disease patterns are extracted for document clustering and feature selection procedure. Agglomerative cluster method is used to find inter and intra cluster gene documents for feature selection measures. Here, proposed gene mutual information and chi-square feature selection measures are used to find the gene dependencies on the disease patterns. Finally, documents with highest gene-disease document patterns are extracted as candidate sets to Reducer phase.

```
// Reducer Phase
Reducer < Mapper-M[i], Patterns[]>
  1. for each M[i] in the list do Apply HashJoin operation and display the top k patterns in the pattern list Done
  2. In the reducer phase of the Hadoop framework, HashJoin operation is used to find the top k patterns from the gene-disease
     document patterns as decision patterns.
```

4. Experimental results

In this experimental study, we have applied the proposed framework on Medline and PubMed repositories [13,14] .We employ the current Apache Hadoop framework with Amazon AWS server. The configuration of the Amazon AWS Server contains

10–50 cluster nodes with 10 CPU cores and 24 GB RAM to each mapper node.

Amazon web services EC2, has three different types of storage sizes such as small, medium, large and this type include all kinds of cloud resources. These instance types are available for several zones. Nowadays, cloud computing technology has become widespread in many application domains. In most of the cases, the customers are unaware of their cloud usage. Small and medium scale industries are migrating towards cloud computing due to its faster access to their applications and decreased the amount of infrastructure cost. The process of cloud computing can be considered as a business model in which the computation services are sold and rented. The major concern of cloud computing technology is to provide different cloud services according to user's need with a nominal amount of charges. Amazon Elastic Compute Cloud (Amazon EC2) provides variable-sized computation capacity in the cloud. EC2 provides simple and easy web-scale computation for developers.

Amazon EC2 supports all major operating systems, including RedHat Linux, Windows server, SuSE Linux, Ubuntu, Fedora, Debian, Cent OS, Gentoo Linux, Oracle Linux, and FreeBSD. Amazon plans to include several additional operating systems to EC2 instances in future. For successful operation of this work, a 64-bit Ubuntu server 12.04 is implemented. Generally, AMI uses t1.large, as it supports both 32-bit as well as 64-bit operating system.

In Table 1, the Hadoop cluster nodes, document size and its statistical computations are illustrated. From the table, an Avg gene specifies the average number of genes identified in the biomedical document sets, ranked gene documents specifies the ranking of the gene to disease patterns and its relational synonyms. From Table 1, it is observed that the proposed model has high average ranked discovery patterns and relational genes on the large training datasets.

Table 2, illustrates the Hadoop cluster nodes, documents size and its performance analysis. From Table 2, an Average genes specifies the average number of genes identified in the biomedical document sets. AvgTime specifies the average time taken by the Hadoop cluster nodes in Hadoop framework. From Table 2, it is observed that proposed model has high average gene discovery patterns and less average time on the large training datasets.

Table 1
Relational gene documents and its statistics with different hadoop cluster nodes.

Hadoop cluster nodes	Documents (×100000)	Avg genes (×1000)	Ranked gene docs	Relational genes (×1000)
5	5	1.6	0.57	0.142
10	10	2.1	1.46	0.251
15	15	3.8	1.96	0.835
20	20	4.16	2.43	0.976
25	25	4.89	3.17	1.75

Table 2
Average gene documents and its time computation with different hadoop cluster nodes.

Hadoop cluster nodes	Documents (×100000)	Avg genes (x1000)	Avg time (secs)
5	5	1.6	52
10	10	2.1	61
15	15	3.8	79
20	20	4.16	86
25	25	4.89	102

Table 3
Average ranked gene documents and its quality cluster rate with different Hadoop cluster nodes.

Hadoop cluster nodes	Documents (x100000)	Ranked gene docs	Cluster rate
5	5	0.57	0.965
10	10	1.46	0.975
15	15	1.96	0.986
20	20	2.43	0.975
25	25	3.17	0.972

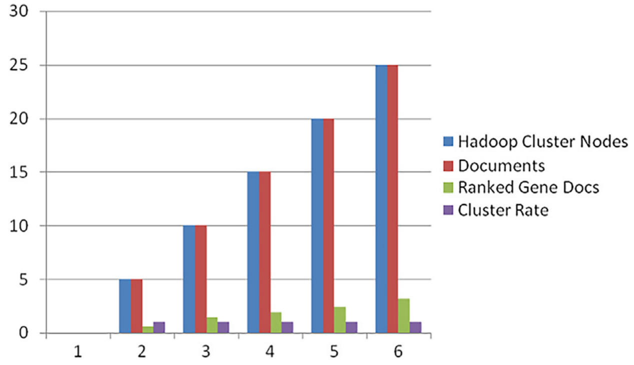

Fig. 2. Average ranked gene documents and its quality cluster rate with different Hadoop cluster nodes.

Table 4
Document clustering and classification rate of the proposed model to the traditional models.

Algorithm	φ = 0.5	φ = 0.7	φ = 0.8	φ = 0.9
K-modes	83	81.65	83.23	82
Knn + K-means	68.45	74	76.34	76.89
MEAD	87.89	86.34	85.67	87
Neural Networks	85.89	79.05	74.56	77.12
GA + SVM	74	75.78	76.45	79.45
Multiclass ensemble Model	93.45	94.16	95.24	96.74
Proposed Model	94.62	96.25	97.45	97.95

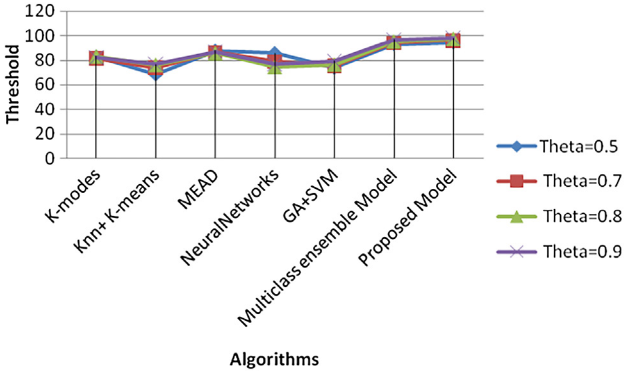

Fig. 3. Performance analysis of proposed model to existing models.

Table 3, illustrates the Hadoop cluster nodes, documents size and its performance analysis. From Table 3, a ranked gene document specifies the ranking of the gene to disease patterns and its relational synonyms in the biomedical document sets. Cluster quality rate specifies the cluster rate in terms of correctness of the gene terms in the document in the hadoop framework. From Table 3, it is observed that proposed model has high cluster rate on the large training datasets.

From Fig. 2, a ranked gene document specifies the ranking of the gene to disease patterns and its relational synonyms in the biomedical document sets. Cluster quality rate specifies the cluster rate in terms of correctness of the gene terms in the document in the hadoop framework.

From Table 4 and Fig. 3, proposed model achieved (~97.5%) accuracy than the traditional ensemble model on different node configurations and threshold with Accuracy on X-axis and different models on Y-axis. As the size of the gene pattern threshold increases from 0.5 to 0.9 values, proposed model has high computational accuracy as compared to traditional models.

5. Conclusion

Biomedical document clustering and classification is one of the essential machine learning models for the knowledge extraction process of the real-time user recommended systems. As the amount of information in the biomedical repositories increases, many organizations are facing the unprecedented issues of how to process the available volumes of data efficiently. In this paper, a novel feature selection based document clustering and classification model was implemented with the multiple integrated biomedical databases such as Medline and PubMed repositories. This model is used as a user recommended system on larger document sets using the Hadoop framework. Experimental results show that the proposed model has a high computational cluster quality rate (~96%) and true positive classification rate (~98%) compared to traditional document clustering and classification models. The computational complexity of the Mapper phase is O(nlogn) and Reducer phase is O(logn). In future, this work can be extended to protein clustering and classification using the Hadoop framework.

References

[1] Rojcek M. System for fuzzy document clusterng and fast fuzzy classification. In: 15th IEEE international symposium on computational intelligence and informatics; 2014. p.39–42.
[2] Aïtelhadj A, Boughanem M, Mezghiche M, Souam F. Using structural similarity for clustering XML documents; 2011. p. 109–139.
[3] Chan SW, W Chong M. Unsupervised clustering for nontextual web document classification. Decis Supp Syst 2004:377–96.
[4] Curtis D, Kubushyn V, Yfantis EA, Rogers M. A hierarchical feature decomposition clustering algorithm for unsupervised classification of document image types. In: Sixth international conference on machine learning and applications; 2007. p.423–8.
[5] Dai W, Xue G, Yang Qi, Yu Y. Co-clustering based classification for out-of-domain documents. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2007. p. 210–9.
[6] Diaz-Valenzuela I, Loia V, Martin-Bautista MJ, Senatore S, Vila MA. Automatic constraints generation for semisupervised clustering: experiences with documents classification. Soft Comput. 2016;20(6):2329–39.
[7] Hachenberg C, Gottron T. Locality sensitive hashing for scalable structural classification and clustering of web documents. In: Proceedings of the 22nd ACM international conference on information & knowledge management. ACM; 2013. p. 359–63.
[8] Jiang S, Lewris J, Voltmer M, Wang H. Integrating rich document representations for text classification. In: IEEE systems and information engineering design conference (SIEDS '16)"; 2016. p. 303–8.
[9] Ke W. Least information document representation for automated text classification. In: Proceedings of the American Society for Information Science and Technology 49.1; 2012. p. 1–10.
[10] Lin B, Chen T. Genre classification for musical documents based on extracted melodic patterns and clustering. In: Conference on technologies and applications of artificial intelligence; 2012. p. 39–43.
[11] Nam LN, Quoc HB. A combined approach for filter feature selection in document classification. In: IEEE 27th international conference on tools with artificial intelligence; 2015. p. 317–24.
[12] Shruti S, Shalini L. Sentence clustering in text document using fuzzy clustering algorithm. In: International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT); 2014. p. 1473–6.
[13] https://www.ncbi.nlm.nih.gov/pubmed/.
[14] https://www.nlm.nih.gov/bsd/pmresources.html.