Informatics in Medicine Unlocked 16 (2019) 100188

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

A novel somatic cancer gene-based biomedical document feature ranking and clustering model

Thulasi Bikku[a,*], Radhika Paturi[b]

[a] Department of Computer Science and Engineering, Vignan's Nirula Institute of Technology and Science for Women, Palakaluru, Guntur, India
[b] Vignan's Nirula Institute of Technology and Science for Women, Palakaluru, Andhra Pradesh, India

ARTICLE INFO

Keywords:
Somatic cancer
Genomes
Bioinformatics
Feature selection
Feature ranking
Fuzzy clustering
Document clustering

ABSTRACT

Background: As the size of somatic genomes in biomedical repositories increases, it is essential to predict cancer related document sets using the machine learning models. Most of the traditional gene-based somatic cancer mining models are independent of somatic gene ranking and feature extraction due to high computational cost and memory for large datasets. A wide range of feature selection and feature extraction strategies are existing, and they are by and large generally utilized in various areas. Every one of these strategies plans to expel repetitive and irrelevant features from the trained datasets with the goal that the arrangement of new document data will be increasingly accurate. Data extraction is the activity of providing relevant data according to an information need from a collection of large resources of data

Results: Ranking consists of sorting the information offers according to some criterion, so that the "best" results appear in the top priority in the provided list. The mapping of somatic genomes and its equivalent words like synonyms to biomedical document ranking is intricate on vast biomedical document data sets. In order to overcome these limitations, a novel feature ranking based fuzzy clustering framework is designed and implemented on large biomedical databases

Conclusion: Experimental results are simulated with different cluster sizes and gene features for somatic document clustering. Experimental results proved that the present model has high computational cluster quality rate with document ranking for somatic gene-based document indexing.

1. Introduction

Machine learning (ML), a subset of Artificial Intelligence which is based on computational statistics and algorithms used for prediction-making and allows computers to learn automatically [1]. ML involves learning from different experiences and then using those experiences to predict the correct outcomes in the later stages of its use. To define it in a computational language we say that, "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Machine learning refers to algorithms that can make a computer learn from many examples. The basic idea is to extract a formal statistical model from the given examples and using it to predict the value or class of the target variable for an unseen example. If the value of the target variable is known for each example, it is called supervised machine learning [2]. Further, if the predicted variable is considered to be categorical, the task is known as classification and if the predicted variable is considered to be continuous, the task is known as regression. If the possible outcomes are limited to two, it is called binary classification. If the possible outcomes are above two, it is termed as the multi-class classification problem. The primary objective of a machine learning model is to correctly classify an instance. However, in many problem-domains like medicine, the classification is followed by critical decision making [27]. For example, take the case of using machine learning to predict whether a tumour is malignant or benign. In this problem, both the type of misclassifications i.e. false positives as well as false negatives are hazardous.

A false positive i.e. classifying a benign tumour as malignant will recommend not-required chemotherapy. On the other hand, a false negative i.e. classifying a malignant tumour as benign will lead to no treatment at all and let the disease advance further. In such problem domains, to develop a trust in the machine learning model, it is desirable that the outcome of the model should be understandable to a human expert. In other words, the human expert should be able to identify what features made the model predict a particular outcome for a given instance. So, a human-interpretable machine learning model is

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

https://doi.org/10.1016/j.imu.2019.100188
Received 1 April 2019; Received in revised form 3 May 2019; Accepted 4 May 2019
Available online 30 May 2019

basically one whose outcomes can be interpreted by a human expert. In machine learning, the features of the data are more, so the amount of features required to analyse the classification also grows exponentially [3]. Richard E. Bellman coined this phenomenon as the "curse of dimensionality", as the number of relevant features increases with the increase of dimensions when considering problems of enumeration on product spaces in dynamic optimisation [4]. To solve a large number of features in the datasets so as to analyse the scenario, feature selection has been extensively used in the machine learning applications. A classification model can be built by eliminating redundant or irrelevant features from the dataset and considering the high ranked or high priority features extracted based on the criterion by using feature selection strategies. The common approach to represent the high dimensional data into variables known as features, which preserves the information without losing the valuable information as shown in Fig. 1.

Microarray data sets are large and analyses the data by using variables and data points. This strategy is used to reduce the dataset into genomes, which can distinguish between the two cases or classes. For example, a data point can have 500,000 variables approximately and processing multiple data points is not an easy task, which high computational cost [5]. When the dataset dimensionality grows rapidly then it is very difficult to prove the result statistically due to the sparsity of the data in the dataset. The Biomedical Document parser identifies the structure of phrases and sentences in the XML and pdf formats of documents. Information about gene, protein and its pathways is analyzed using the MedScan toolkit [6]. The MedScan is a three-tier knowledge extraction system based on a biomedical document parsing model. In the first tier, the pre-processor module aimed to tag various biomedical MeSH terms using domain specific concepts. Pre-processor module reads the biomedical XML format of a MEDLINE abstract and parses into individual terms or sentences. In this module, a protein name dictionary is used as a training dataset to filter the protein names and to select the terms or sentences containing at least one gene-protein name. In the second tier, the natural language processor performs a set of semantic relationships between a term or sentence structures. It is based on context-free grammar and a lexicon parse tree for MEDLINE protein extraction. In the final tier, knowledge extraction engine acting as a domain knowledge filter for extraction key MeSH-based document information in the form of conceptual graph format. MedScan utilizes the ontology-based filter to select gene or protein semantic entities into the ontology tree structure [7]. The main limitations of the MedScan toolkit are: the efficiency of the natural language processor should be optimized by improving the size, quality and pre-processing algorithms. The volume of the biomedical gene, protein or disease entities can be increased several times by extending the ontology structure of high computational systems. Support vector machine optimisation aims at constructing a separation function for domain knowledge extraction. Each document is classified using the hyperplanes and feature vector space.

2. Background

The large datasets consist of "large p-number of features, n-number of samples, $p > n$" problem may have the issue of overfitting. A model which is over fitted can cause fluctuations for critical change in the information which can result in errors in the classification accuracy. These errors can also increase because of noisy and irrelevant features [8]. Noise in a dataset is the error in the variance of a measured variable, which can result from errors in measurements or normal variation. Feature selection is a procedure that picks M features as a subset from the total set of N features, based on an evaluation criterion. To reduce the features count so as reduce the dimensionality of the domain, the total number of features N decreases, so redundant and irrelevant features are removed. It is very difficult to find the best feature subset from the total features and related to feature selection problems have been considered as NP-hard. Feature selection is an effective research area in Computer Science. In statistical pattern recognition, Machine Learning (ML) and mining of data feature selection have been a prolific field of innovative work from the last few decades. The major issue in a wide range of areas is feature selection, particularly in forecasting, classifying documents, bioinformatics, and object recognition or in modelling of complex processes related to technology having datasets with a huge number of features. For a few applications, to analyse the data all the feature of the dataset needs to be considered, yet for some objective ideas, only a few subsets of relevant features of the dataset are required. The dimensionality of feature space can be reduced by selecting the best features, removes redundant, irrelevant, or noisy features from large datasets, the immediate effects after reduction are data mining algorithms are speeding up, data quality gets improved and thereof the performance of the model is improved, and mining results comprehensibility is increased. Feature selection algorithms can be categorised into three types, filters, wrappers and embedded [9]. Quality of selected features can be evaluated using Filters methods, irrespective of the classification algorithm, however, the wrapper methods evaluate the quality of the feature based on the application of a classifier. In the Embedded methods, the feature selection is done during the training phase, the features are noted and used in the testing phase. Focusing on relevant features and ignoring irrelevant ones is automated by a few classification algorithms like Decision trees, multi-layer perceptron (MLP) neural networks, having the input with strong regularization. Alternatively, some algorithms like the k-nearest neighbour algorithm that classifies the nearest training case, depending on the methods for feature selection to remove noisy features because they are not having feature selection provision by their own. The efficiency of a feature subset is important due to the availability of class information in data, further divided into supervised feature selection approaches and unsupervised feature selection approaches. Azar et al. proposed a method linguistic hedges neuro-fuzzy classifier with selected features (LHNFCSF) decreases the data features of the datasets, yet in addition improves the performance of the classifier by disposing of redundant features, noise-corrupted, or irrelevant dimensions [10]. This method results show that not only helps in decreasing the dimensionality of vast data collections yet additionally can accelerate the computational time, fast training time and learning ability of the model and simplify the tasks of classification. Lilleberg et al. proposed that the words and expressions of a document are changed over into a vector portrayal, word2vec adopts a completely new strategy on content classification [11]. In light of the supposition that word2vec brings additional semantic features that help in the content grouping. The viability of word2vec by demonstrating that tf-idf and word2vec consolidated can outperform tf-idf on the grounds that word2vec gives integral features to tf-idf. The experimental outcomes showed that the support vector machine performs well in document classification, particularly when semantic words are utilized as context-based features. Sheikhan et al., introduced a technique based on fuzzy grids–based association rules mining use in network applications for feature selection so as to detect misuses in the computer networks [12]. The main aim of this methodology is to find co-relationship between large datasets frequent itemsets and the system inputs to detect the relationship and eliminating the inputs which are redundant. A fuzzy ARTMAP neural system is utilized whose training datasets are enhanced by gravitational search algorithm to group the attacks. While picking ideal "feature subset measure modification" parameter, experimental outcomes demonstrate that the proposed framework performs better as far as recognition rate, false alert rate, and cost per model in the classification issues. Moreover, more than 8.4% decrease in computational intricacy after utilizing the decreased size list of features results. Song et al., proposed a two-step fast clustering-based feature selection algorithm (FAST), calculate the time required to identify a feature subset so as to improve the proficiency and quality [13]. In the initial step, by using graph-theoretic clustering methods relevant features are partitioned into clusters. In the second step, the most efficient features which are related to classes are identified from a cluster and form a feature subset. The subset of features in various clusters are generally

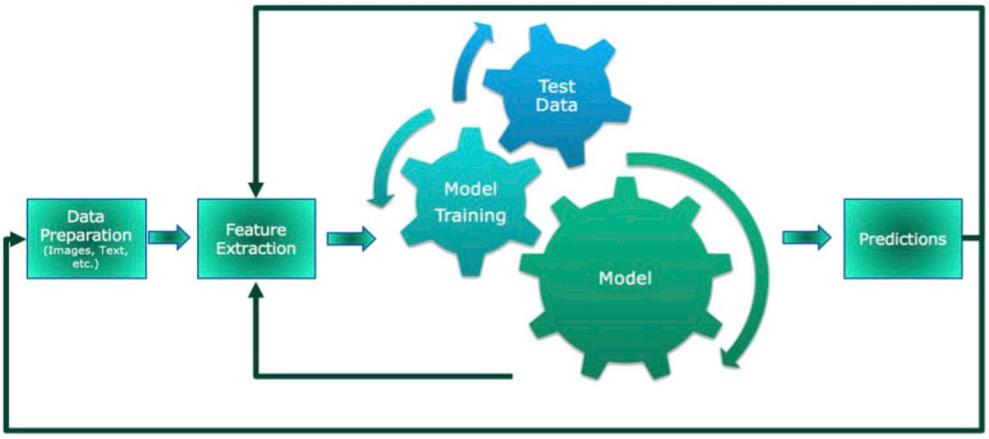

Fig. 1. Mechanism of machine learning.

autonomous, the grouping based system of FAST has a high likelihood of delivering a subset of valuable and free features. To guarantee the productivity of FAST, we receive the proficient minimum-spanning tree (MST) grouping strategy. The proficiency and viability of the FAST calculation are assessed through an experimental examination of datasets. Bermejo et al., manages the issue of wrapper feature subset selection (FSS) in classification-oriented datasets with an extensive number of properties [14]. In high-dimensional datasets with a large number of factors, wrapper FSS turns into a relentless computational procedure on account of the measure of CPU time it requires. This depends on the blend of the Naïve Bayesian classifier with gradual wrapper FSS algorithms. Zou et al., proposed a Max-Relevance-Max-Distance (MRMD) feature ranking strategy, which adjusts exactness and dependability of ranking of features and forecast task [15]. The first is benchmark dataset with high dimensionality is image classification, while the second one is protein–protein communication expectation information, which originates from our past private research and has huge occasions. Sharma et al., proposed calculation first partitions qualities into subsets, the sizes of which are generally little, generally of size h, at that point chooses educational littler subsets of genes of size r < h from a subset and consolidations the picked genes with another gene subset of size r to refresh the quality subset. We rehash this procedure until all subsets are converted into one informative subset [16]. Wang et al., proposed online component choice in which an online learner contains a little and a fixed number of features keeps up a solitary classifier [17]. The key test of an online selection of features is the means by which to make a precise forecast for a case utilizing a few dynamic features. This is rather than the established setup of web-based realizing where every one of the features can be utilized for expectation. We endeavour to handle this test by considering sparsity regularization and truncation systems. Microarray classification of data order is a troublesome test for ML scientists because of its high number of features and the little sample sizes [18]. Feature selection has been before long thought about an accepted standard in this field since its presentation, and an immense number of selection of features strategies were used attempting to diminish the input dimensionality while improving the performance of classification [19]. Uysal et al. proposed a novel channel based probabilistic element determination strategy, to be specific distinguishing feature selector (DFS), for content grouping [20]. Trial results expressly demonstrate that DFS offers an aggressive act as for the previously mentioned methodologies as far as classification precision, measurement decrease rate and preparing time. AI calculations will, in general, be influenced by uproarious or noisy information. Clamour ought to be decreased however much as could reasonably be expected so as to keep away from superfluous unpredictability in the deduced models and improve the effectiveness of the model. The regular commotion can be separated into two kinds: (1) Attribute noise (2) Class noise. First one is brought about by mistakes in the characteristic qualities like wrongly estimated factors, missing qualities while the latter is brought about by tests that are named to have a place in more than one class as well as misclassifications. The computational cost increases with the increase in the dimensionality, typically exponentially [21]. To conquer this issue, it is important to figure out how to diminish the number of features in thought. Two methods are regularly utilized: (1) Feature subset determination. (2) Feature extraction. Cancer is among the leading causes of death worldwide accounting for more than 8 million deaths according to the World Health Organization. It is expected that the deaths from cancer will rise to 14 million in the next two decades. Cancer is not a single disease. There are more than 100 known different types of cancer and probably many more. The term cancer is used to describe the abnormal growth of cells that can, for example, form extra tissue called mass and then attack other organs. Microarray databases are a huge wellspring of genetic information, which, upon legitimate examination, could upgrade our comprehension of science and medicine. Numerous microarray tests have been intended to research the genetic components of malignant growth, and logical methodologies have been connected so as to arrange distinctive kinds of disease or recognize carcinogenic and

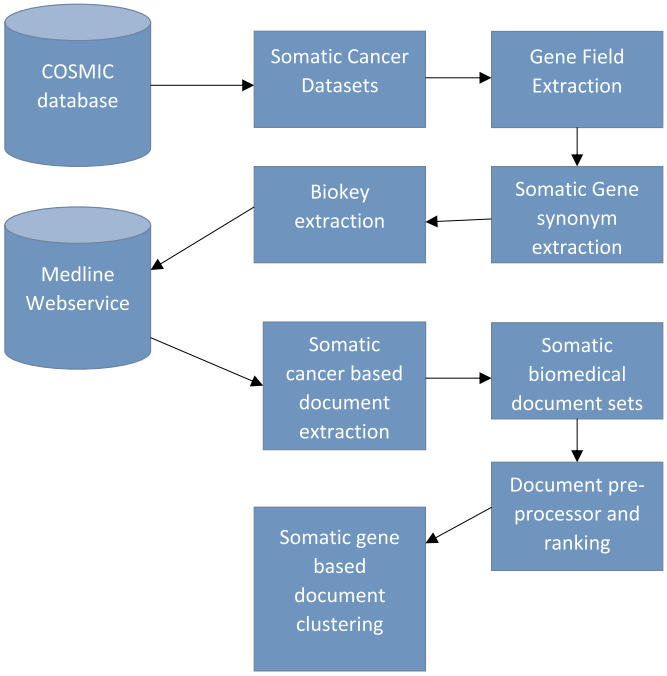

Fig. 2. Proposed model.

Table 1
Performance of the proposed somatic gene similarity measure on different somatic genes.

Mappers	Somatic Gene	AvgSimilarity	Time (ms)
#2	ATT,CTT,GTT,TAT	0.956	2535
#4	TCG,GAT,TGT,TTA	0.975	1957
#6	AGG,CCC,CGG,GGC	0.984	1793
#8	GCA,ACG,CCG,GCC	0.985	1475
#10	ATG,CCT,GGT,GTA	0.978	1285

noncancerous tissue. Over the most recent ten years, ML methods have been explored in microarray information investigation. A few methodologies have been attempted so as to (i) recognize harmful and noncancerous tests, (ii) order distinctive sorts of malignant growth, and (iii) distinguish subtypes of disease that may advance forcefully [22]. Every one of these examinations is looking to create organically important translations of complex datasets that are adequately fascinating to drive follow-up experimentation. The final section is about using prior knowledge in combination with a feature extraction or feature selection method to improve classification accuracy and algorithmic complexity. The Feature extraction (FE) extracts the relevant features from the dataset without losing the originality of the dataset, the correlation among the objects, covers the latent structure. Irrelevant features of the datasets, which do not covers cover latent structure are less effective.

3. Proposed methods

In the proposed model the Collected Data is the first and foremost step, secondly, we pick the model, third, we train the model and then finally we test the model. By simply adding the data to the model it will start recognizing patterns for us. Now the data sets can come in the whole bunch of different forms namely csv, pdf, holograms, and images but should be in a well-known format which we can parse from. The dataset then has different features according to our requirement. The sign of good features is that they should be simple for one. Secondly, they should be independent and third they should be informative. The data obtained from the medical documents, which is raw data, so pre-processing should be done. The pre-processing may contain (1) removal of Stop words, there many complicated words in English, the most common words used in the sentences are for e.g., are, a, all, an, the, how, who, whom, be, and other common and generally used words having less weight while clustering the documents are known as delimiters or stop words. The stop words do not contain any standards or rules, they are completely based on the user, who are going to identify the stop words in their documents. It winds up important to expel them as they typically spread a considerable part of a report or document and thus increase the quantity of subset of features superfluously as well as misdirect and break down execution of the fundamental classification technique. In this technique, stop words are evacuated as per the stop word list which contains 571 words, which is accessible at the website1 of Journal of Machine Largening Research [23].

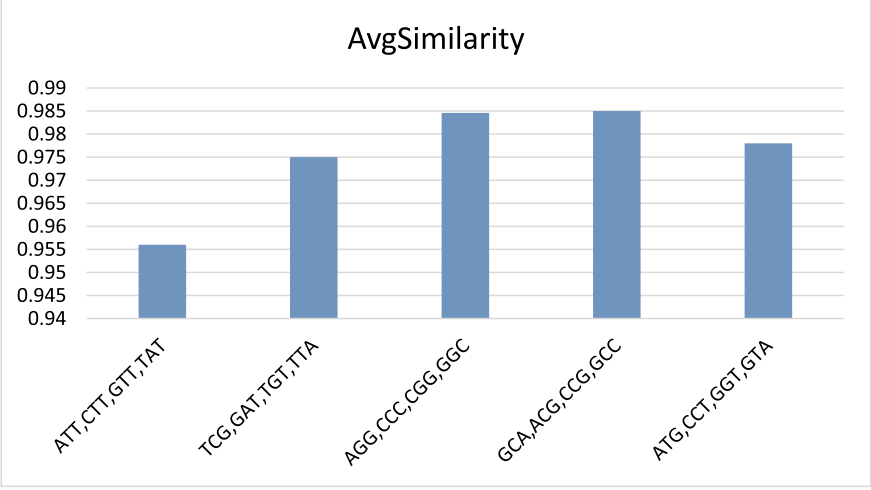

Fig. 3. Somatic gene sequences and its similarity measures.

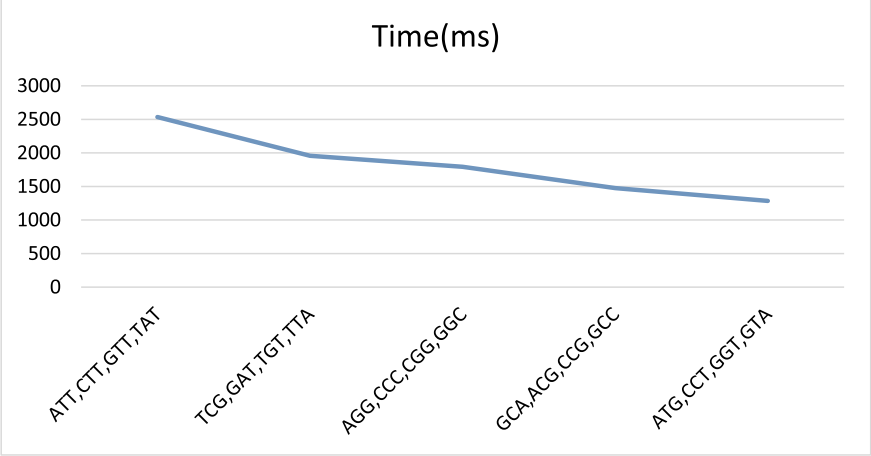

Fig. 4. Somatic gene sequences and its runtime (ms) computation for different mappers.

(2) Starting from the elimination of suffix elimination and producing the word stems comes under Stemming [24]. Adding suffix to the root word like jump, jumped and jumping, here jump is the root word, the suffixes liked ed, ing are added, those words to be considered as same words.

These type of words to be considered as a single word, considering the root word and added to the dictionary, so as to reduce the storage and processing time (3) Tokenization is a procedure that a document or report gets split into autonomous terms known as tokens. The size of the token differs from a solitary term (unigram) to a successive arrangement of n-terms (n-grams). In this technique, we utilize single terms for the representation of documents [25]. (4) Term weighting tells about the vector representation of terms in the bunch archives, the representation of vectors demonstrate the utilized to delineate reports into a minimized vector portrayal. In the vector portrayal, each report operation {p = 1; 2; 3; ...n} and term tf {f = 1; 2; 3; … s} present in the accumulation is considered for portrayal. Different term weighting plans have been proposed in the writing to delineate substance into a numeric arrangement in which frequency-inverse document frequency (tfidf) is frequently used [26]. A vital approach of the proposed model is that it incorporates the multi-faceted functionality of feature selection, extraction, grouping, determination and dimensionality reduction analysis. In the proposed system, novel somatic cancer genes are used to find and extract the clustered and ranked documents from the large biomedical documents sets. In the present system, somatic cancer genes are taken from the COSMIS website to index and map the essential key biomedical features in the document feature extraction, feature ranking and clustering schemes.

The overall framework of the proposed model is shown in Fig. 2. In Fig. 2, somatic cancer datasets are taken from the COSMIS web repository. These cancer datasets contain different types of fields to detect somatic cancer using the gene sets. Initially, all the fields relevant to the somatic genes are extracted from the COSMIC, the Catalogue of Somatic Mutations in Cancer is the world's biggest and most exhaustive asset for investigating the effect of substantial transformations in human malignancy, which raw datasets as a trained dataset to the model. The synonyms of the genes are identified using the synonym database. The Gene Field Extraction is used to study disease causes, which is very useful in the field of bioinformatics diagnostics and drugs. It is also essential for carrying out forensic science, sequencing genomes, detecting bacteria and viruses in the environment and for determining paternity. Here the model extracts the synonyms and acronyms of genes, it will be helpful to extract the relevant documents. The documents are clustered based on the ranking algorithm as shown in Fig. 2. Somatic cancer genes and its synonyms are used to find the document feature extraction and ranking process. Finally, the ranked feature documents are pre-processed to cluster the somatic gene documents using the fuzzy-based clustering model.

Algorithm 1. Somatic Gene Data Extraction

Input: COSMIC dataset CD, synonym training dataset SD.
Step 1: Read COSMIC database CD.
Step 2: Somatic Genes SG [] =null;
For each record r in CD
Do
SG[i]=FindGeneField(CD[j]); j is gene field index in COSMIC dataset.
Done
Step 3: Filtering somatic genes based on cancer type.
For each gene g_i in SG[]
Do
If(g_i=True)
Then
G[i]=g_i
End if
Done
Step 4: Find the similarity between somatic gene to the synonym geneset SD.
For each gene g_m in G[]
Do
Somatic synonym geneset =SSG[] =Sim(g_m,SD)

$$SSG[] = Sim(g_m, SD)$$

$$Sim(g_m, SD_j) = Max\{\frac{Prob(g_m / SD_j)}{|SD| . Prob(g_m)}\};, m = 1, 2... |G|$$

$$j = 1, 2... |SD|$$

$$Add\ SSGD[m]=\{Sim(g_m, SD_j), g_m, SD_j\}$$

done

Step 5: Assign somatic synonym genes as bio keyterms.
For each biomedical document from Medline
Do
For each sg_i in SSGD []
Do
If($Sim(g_m, SD_j) > T$)
Then
Bio key terms BKT[i]={ g_m, SD_j }
End if
End for
End for

In the above algorithm, somatic genes are extracted from the COSMIC website. Each record in the COSMIC data is processed to find the gene index value. If the gene index of the COSMIC data is true then gene-term and its synonyms are identified in the training gene set. Probabilistic-based gene similarity is used to find the context similarity between the gene and its synonyms. Finally, biomedical somatic genes are filtered using the given threshold T.

Algorithm 2. Somatic gene document extraction in Medline repository

Input: Biokey terms BKT[] as somatic genes g_m and synonynm genes SD.
Procedure:
Step 1: Connect Medline webservice using biokey terms as an argument.
Step 2: Connection c=Medline();
If (c!=null)
Then
For each biokey term k_i in BKT[]
Do
Extract medline documents using the biokey BKT[i]
Biomedical documents BD[]=C(k_i)
done
End if
Step 3: Pre-processing each document in the BD[] .
For each document d_i in BD[]
Do
Dt_i=Tokenize(d_i)
Sdt_i=stemming (Dt_i)
PBD[]=stopwordremoval(Sdt_i)
Done
Step 4: Somatic gene term based document feature ranking is computed on the pre-processed biomedical documents.
For each pre-processed biomedical document PBD[]
Do
FBD[i]=F_rank(PBD[i],BKT[j]);

$\theta_i <- PBD[i]$

$\theta_j <- BKT[j]$

$\phi_{TP}(\theta_i, \theta_j)$ defines the number of documents in PBD[i] that contains biokey term BKT[j]

$\phi_{FP}(\bar{\theta}_i, \theta_j)$ defines the number of documents not in PBD[i] that contains biokey term BKT[j]

$\phi_{TN}(\theta_i, \bar{\theta}_j)$ defines the number of documents in PBD[i] that does not contains biokey term BKT[j]

$\phi_{FN}(\bar{\theta}_i, \bar{\theta}_j)$ defines the number of documents not in PBD[i] that does not contains biokey term BKT[j]

$$\phi_{rank_1} = \lambda(\phi_{TP}(\theta_i, \theta_j) + \phi_{FN}(\bar{\theta}_i, \theta_j)$$

$$\text{Where } \lambda \in \left(\frac{1}{2\pi\sigma_{fg}} e^{-\left(\frac{fg-\mu_{fg}}{\sigma_{fg}}\right)}, \min(\phi_{TP}(\theta_i, \theta_j), \phi_{FN}(\bar{\theta}_i, \theta_j))\right)$$

$$\phi_{rank_2} = \lambda_2(\phi_{FP}(\bar{\theta}_i, \theta_j) + \phi_{TN}(\theta_i, \bar{\theta}_j)$$

$$\text{Where } \lambda_2 \in \left(\frac{1}{2\pi\sigma_{fg}} e^{-\left(\frac{fg-\mu_{fg}}{\sigma_{fg}}\right)}, \min(\phi_{TN}(\theta_i, \bar{\theta}_j), \phi_{FP}(\bar{\theta}_i, \theta_j))\right)$$

$$F\_rank(PBD[i], BKT[j]) = \frac{-\phi_{rank_1}.Prob(BKT[j]/PBD[i])}{\max\{\phi_{rank_1}, \phi_{rank_2}\}} \log\left(\frac{Prob(BKT[j]/PBD[i])}{\max\{\phi_{rank_1}, \phi_{rank_2}\}}\right)$$

4. Results

All the empirical studies are performed in the Windows 7 environment on a machine having i7 core processor and 4 GB RAM. All feature selection methods are coded in Java and final clusters of the documents using k-means have been obtained using MAP/REDUCE. Experimental results are simulated on different real-time somatic features taken from the cosmic repository. Results are simulated on Hadoop framework using Amazon AWS server. Somatic cancer features are extracted to find the gene sets and its related synonyms from the PubMed and Medline repositories. Experimental

results are compared to the traditional gene similarity measures and the somatic document ranking models using the Hadoop framework. DNA Sequencing implies the combinations of four chemical substances form DNA molecule known as “bases”. This conveys researchers about the sort of genetic data in specific DNA sequence. In the DNA sequence, the four substance bases dependably bond with a similar accomplice to frame “base sets.” Adenine (A) dependably combines with thymine (T); cytosine (C) dependably matches with guanine (G). This blending is the reason for the component by which DNA particles are duplicated when cells isolate, and the matching likewise underlies the strategies by which most DNA sequencing tests are finished. For making and maintaining a human being about 3 billion base pairs are required in the human genome shown in DNA Sequential Taxonomy 1.

DNA Sequential Taxonomy 1: Relational Somatic Gene patterns

```
SeqContent = ATT
|  mutAss = neutral
|  |  isFlanking < 0.83
|  |  |  isFlanking < 0.01
|  |  |  |  dbSNP = true : false (27.89/1.34) [1.02/0.68]
|  |  |  |  dbSNP = false
|  |  |  |  |  pattern = CG : false (0/0) [0/0]
|  |  |  |  |  pattern = CA : false (0/0) [0/0]
|  |  |  |  |  pattern = CT : false (0/0) [0/0]
|  |  |  |  |  pattern = TA : true (1.68/1) [0.68/0]
|  |  |  |  |  pattern = TC
|  |  |  |  |  |  inExAct = true : false (11.41/1.1) [0.34/0.34]
|  |  |  |  |  |  inExAct = false : true (5.03/1.68) [1.48/0.14]
|  |  |  |  |  pattern = TG : true (4.87/0.48) [0/0]
|  |  |  isFlanking >= 0.01
|  |  |  |  CNT < 0.5
|  |  |  |  |  fre < 0.01
|  |  |  |  |  |  VAF < 40.3 : true (4.56/0.97) [0.85/0]
|  |  |  |  |  |  VAF >= 40.3 : false (8.28/1.42) [0.97/0.85]
|  |  |  |  |  fre >= 0.01 : false (10.08/0) [0.28/0]
|  |  |  |  CNT >= 0.5 : false (19.37/0) [2/0]
|  |  isFlanking >= 0.83
|  |  |  pattern = CG : false (0/0) [0/0]
|  |  |  pattern = CA : false (0/0) [0/0]
|  |  |  pattern = CT : false (0/0) [0/0]
|  |  |  pattern = TA : true (7.08/0.91) [2.75/0]
|  |  |  pattern = TC
|  |  |  |  dbSNP = true : false (23.19/1) [1.13/0.75]
|  |  |  |  dbSNP = false
|  |  |  |  |  inExAct = true : false (7.2/0.75) [0.38/0.38]
|  |  |  |  |  inExAct = false : true (10.87/2.75) [0.53/0.16]
|  |  |  pattern = TG : false (7.86/3.44) [0/0]
|  mutAss = low
|  |  pattern = CG : false (0/0) [0/0]
|  |  pattern = CA : false (0/0) [0/0]
|  |  pattern = CT : false (0/0) [0/0]
|  |  pattern = TA : true (16.62/7.31) [1/0]
|  |  pattern = TC
|  |  |  polyphen = benign : false (38.05/11.63) [1.75/0.53]
```

```
| | | polyphen = probably : false (16.35/7.14) [0.75/0.23]
| | | polyphen = possibly : false (17.61/5.46) [0.81/0.24]
| | pattern = TG : false (21.54/9.93) [2/1]
| mutAss = medium
| | pattern = CG : false (0/0) [0/0]
| | pattern = CA : false (0/0) [0/0]
| | pattern = CT : false (0/0) [0/0]
| | pattern = TA
| | | polyphen = benign
| | | | VAF < 21.08 : true (3.24/0.89) [0.45/0]
| | | | VAF >= 21.08 : false (6.37/0.45) [0/0]
| | | polyphen = probably : true (5.93/2.21) [0.28/0]
| | | polyphen = possibly : true (5.93/2.21) [0.28/0]
| | pattern = TC : false (55.09/12.95) [5.24/0]
| | pattern = TG : true (8.19/1.47) [2/0]
| mutAss = high
| | pattern = CG : false (0/0) [0/0]
| | pattern = CA : false (0/0) [0/0]
| | pattern = CT : false (0/0) [0/0]
| | pattern = TA : true (2.06/1.03) [0/0]
| | pattern = TC : false (6.39/2.12) [1.03/0]
| | pattern = TG : true (2.15/0.06) [1/0]
| mutAss = stopgain : false (2.12/1.05) [0.01/0]
| mutAss = stoploss : false (0/0) [0/0]
SeqContent = CTT
| mutAss = neutral
| | pattern = CG : true (0/0) [0/0]
| | pattern = CA : true (0/0) [0/0]
| | pattern = CT : true (0/0) [0/0]
| | pattern = TA
| | | VAF < 10 : true (6.52/1) [0/0]
| | | VAF >= 10 : false (16.57/6.52) [3/1]
| | pattern = TC
| | | isFlanking < 0.74
| | | | isFlanking < 0.01
| | | | | VAF < 43.58 : true (12.52/1.88) [1.91/0.19]
| | | | | VAF >= 43.58 : false (11.2/1.72) [3.36/0]
| | | | isFlanking >= 0.01
| | | | | VAF < 40.28
| | | | | | dbSNP = true : false (2.24/0) [2.13/0]
| | | | | | dbSNP = false : true (5.3/1) [0.48/0]
| | | | | VAF >= 40.28 : false (8.43/0.48) [0.24/0]
| | | isFlanking >= 0.74
```

```
| | | | inExAct = true : false (11.99/3.8) [0.61/0]
| | | | inExAct = false : true (14.41/2.3) [1.8/0]
| | pattern = TG
| | | isFlanking < 0.79
| | | | polyphen = benign
| | | | | inExAct = true : false (7.69/0.48) [1/0]
| | | | | inExAct = false : true (11.91/1.77) [0.37/0]
| | | | polyphen = probably : false (1.02/0.3) [0.02/0.02]
| | | | polyphen = possibly : true (4.72/1.1) [1.09/0]
| | | isFlanking >= 0.79 : true (27.96/5.2) [3.79/0]
| mutAss = low
| | pattern = CG : true (0/0) [0/0]
| | pattern = CA : true (0/0) [0/0]
| | pattern = CT : true (0/0) [0/0]
| | pattern = TA : true (38.97/7.49) [4/0]
| | pattern = TC
| | | isFlanking < 0.39
| | | | VAF < 42
| | | | | inExAct = true : false (9.91/1.94) [0.48/0]
| | | | | inExAct = false : true (13.55/0) [0.65/0]
| | | | VAF >= 42
| | | | | inExAct = true : false (12.24/0.65) [1/0]
| | | | | inExAct = false
| | | | | | polyphen = benign : true (3.16/1.18) [0.22/0]
| | | | | | polyphen = probably : false (3.16/0.98) [0.22/0.22]
| | | | | | polyphen = possibly : false (3.16/0.98) [0.22/0.22]
| | | isFlanking >= 0.39
| | | | VAF < 39.21 : true (15.42/1.42) [1.62/0.26]
| | | | VAF >= 39.21 : false (9.39/2.42) [1.35/0.35]
| | pattern = TG
| | | polyphen = benign : true (44.39/10.15) [7.38/0.47]
| | | polyphen = probably
| | | | VAF < 44.52 : true (27.62/1.06) [4.03/0.35]
| | | | VAF >= 44.52 : false (5.68/2.13) [1/0]
| | | polyphen = possibly : true (16.17/3.24) [2.96/0.17]
| mutAss = medium
| | pattern = CG : true (0/0) [0/0]
| | pattern = CA : true (0/0) [0/0]
| | pattern = CT : true (0/0) [0/0]
| | pattern = TA : true (35.51/7.26) [3/0]
| | pattern = TC
| | | VAF < 34.73
| | | | dbSNP = true : false (5/2) [0.63/0]
| | | | dbSNP = false : true (32.57/3.31) [1/0]
| | | VAF >= 34.73 : false (19.94/8) [3/0]
| | pattern = TG : true (78.57/10) [2.31/0]
| mutAss = high
| | fre < 0.01 : true (20.83/1.26) [0.09/0.04]
| | fre >= 0.01 : false (3.09/1) [0.04/0]
| mutAss = stopgain : true (4.16/0.06) [0.02/0.02]
| mutAss = stoploss : true (1.04/0.02) [1.01/0]
```

Table 2
Comparison of the present somatic gene rank to the existing ranking measures on the selected somatic genes.

GeneRankingModel	SomaticGenes	AvgRank	Runtime (ms)
SVMRank	ATT,CTT,GTT,TAT	0.879	4231
DistanceGeneRanker	TCG,GAT,TGT,TTA	0.924	3985
NaiveRanker	AGG,CCC,CGG,GGC	0.947	3574
ProposedRanker	GCA,ACG,CCG,GCC	0.987	3105

Table 1, describes the performance of the present somatic gene similarity measure for the selection of the genes in the real-time biomedical repositories. From the table, it is observed that different somatic gene sequences are used to find and extract the somatic genes from the repository using the proposed similarity measure.

Fig. 3, describes the performance of the present somatic gene similarity measure for the selection of the genes in the real-time biomedical repositories. From the table, it is observed that different somatic gene sequences are used to find and extract the somatic genes from the repository using the proposed similarity measure.

Fig. 4, describes the performance of the present somatic gene similarity measure for the selection of the genes in the real-time biomedical repositories. From the figure, it is observed that different somatic gene sequences are used to find and extract the somatic genes from the repository using the proposed similarity measure. Here, different mappers are used to finding and extract the genes from the biomedical repository. As shown in the figure, runtime (ms) is computed to each mapper in the proposed model.

Table 2, describes the comparison of the proposed gene ranking model to the existing gene ranking measure on the selected somatic cancer gene sequences. From the table, it is noted that the present method has high computational average rank for the selection of somatic genes in the large biomedical repositories. Also, the average runtime (ms) in the proposed model is less compared to the traditional ranking measures.

5. Conclusion

Feature classification and selection is one of the main challenges of mining medical data used in bioinformatics. With the immense measure of information accessible, the improvement of effective classifiers with high prescient execution is required for big data applications such as bioinformatics, DNA sequencing etc. Feature extraction is a basic undertaking in bioinformatics that plans to arrange genomes into a predefined specific class dependent on named information. Two noteworthy issues that ought to be taken care of legitimately for effective and strong mining of biomedical data are managing high-dimensional component space and accomplishing high-performance accuracy. The exploratory outcomes demonstrate that feature choice and group technique mix can be helpful for accomplishing better prescient execution. The proposed method describes the performance of the present somatic gene similarity measure for the selection of the genes in the real-time biomedical repositories.

Acknowledgements

We thank our supervisors for their continuous suggestions and feedback, we thank our college management especially Dr.Lavu Rathaiah Garu for supporting us, who put their faith in us. This work is mainly based on biomedical data extraction from the corpus, so we thank the referees in this wide area. We thank each and everyone who supported us to accomplish this work successfully. This material has not been published in whole or in part elsewhere and not funded by any organization or committee.

Abbreviations

ML	Machine learning
XML	eXtensible Markup Language
MEDLINE	Medical Literature Analysis and Retrieval System Online
MeSH	Medical Subject Headings
MLP	Multi-Layer Perceptron
LHNFCSF	Linguistic Hedges Neuro-Fuzzy Cclassifier with Selected Features
Fuzzy ART	Fuzzy logic and Adaptive Resonance Theory
FAST	Fast clustering-based feature selection algorithm
MST	Minimum-Spanning Tree
FSS	Feature Subset Selection
MRMD	Max-Relevance-Max-Distance
DFS	Distinguishing Feature Selector
FE	Feature Extraction
COSMIC	Catalogue of Somatic Mutations in Cancer
BKT	Bio-Key Terms
DNA	DeoxyriboNucleic Acid

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2019.100188.

References

[1] Michie D, Spiegelhalter DJ, Taylor CC. Machine learning. Neural and Statistical Classification 1994;13.
[2] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Emerging artificial intelligence applications in computer engineering 2007;160:3–24.
[3] Lee WS, Liu B. August. Learning with positive and unlabeled examples using weighted logistic regression. ICML, vol. 3. 2003. p. 448–55.
[4] Rust J. Using randomization to break the curse of dimensionality. Econometrica: Journal of the Econometric Society 1997:487–516.
[5] Quackenbush J. Computational genetics: computational analysis of microarray data. Nat Rev Genet 2001;2(6):418.
[6] Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics 2003;19(13):1699–706.
[7] Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Briefings Bioinf 2005;6(3):239–51.
[8] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.
[9] Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, de Schaetzen V, Duque R, Bersini H, Nowe A. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE ACM Trans Comput Biol Bioinform 2012;9(4):1106–19.
[10] Azar Ahmad Taher, Aboul Ella Hassanien. Dimensionality reduction of medical big data using neural-fuzzy classifier. Soft computing 2015;19(4):1115–27.
[11] Lilleberg Joseph, Zhu Yun, Zhang Yanqing. Support vector machines and word2vec for text classification with semantic features. 2015 IEEE 14th international conference on cognitive informatics & cognitive computing (ICCI* CC). IEEE. 2015.
[12] Sheikhan Mansour, Sharifi Rad Maryam. Gravitational search algorithm–optimized neural misuse detector with selected features by fuzzy grids–based association rules mining. Neural Comput Appl 2013;23(7–8):2451–63.
[13] Song Q, Ni J, Wang G. A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Trans Knowl Data Eng 2013;25(1):1–14.
[14] Bermejo P, Gámez JA, Puerta JM. Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. Knowl Based Syst 2014;55:140–7.
[15] Zou Q, Zeng J, Cao L, Ji R. A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing 2016;173:346–54.
[16] Sharma A, Imoto S, Miyano S. A top-r feature selection algorithm for microarray gene expression data. IEEE ACM Trans Comput Biol Bioinform 2012;9(3):754–64.
[17] Wang J, Zhao P, Hoi SC, Jin R. Online feature selection and its applications. IEEE Trans Knowl Data Eng 2014;26(3):698–710.
[18] Bolón-Canedo V, Sánchez-Marono N, Alonso-Betanzos A, Benítez JM, Herrera F. A review of microarray datasets and applied feature selection methods. Inf Sci 2014;282:111–35.
[19] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015;2015:198363. 13 pages [1-13].
[20] Uysal AK, Gunal S. A novel probabilistic feature selection method for text classification. Knowl Based Syst 2012;36:226–35.
[21] Zhu X, Wu X. Class noise vs. attribute noise: a quantitative study. Artif Intell Rev 2004;22(3):177–210.

[22] Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. CanDrA: cancer-specific driver missense mutation annotation with optimized features. PLoS One 2013;8(10):e77945.
[23] Silva C, Ribeiro B. July. The importance of stop word removal on recall values in text categorization. Proceedings of the international joint conference on neural networks, 2003, vol. 3. IEEE; 2003. p. 1661–6.
[24] Singh J, Gupta V. A systematic review of text stemming techniques. Artif Intell Rev 2017;48(2):157–217.
[25] Mullen LA, Benoit K, Keyes O, Selivanov D, Arnold J. Fast, consistent tokenization of natural language text. J. Open Source Softw 2018;3:655.
[26] Nalisnick E, Mitra B, Craswell N, Caruana R. Improving document ranking with dual word embeddings April Proceedings of the 25th international conference companion on world wide webInternational World Wide Web Conferences Steering Committee; 2016. p. 83–4.
[27] Bikku T, Nandam SR, Akepogu AR. A contemporary feature selection and classification framework for imbalanced biomedical datasets. Egyptian Informatics Journal 2018 Nov 1;19(3):191–8.