A hybrid clustering and classification model for chemical code based medical disease prediction

Konda sreenu,
research scholar, dept of CSE, Acharya Nagarjuna University, Guntur,india.
Dr B.Raja Srinivasa reddy,
Professor, Dept of CSE, Sri vasavi institute of engineering and technology, india.

Abstract:
As the number of biomedical documents sets and medical datasets are increasing in size and dimensions, finding an essential key ICD based disease terms are difficultto extract in large training databases. Most of the traditional approaches use static ICD code extraction for the medical disease classification process. In this paper, a hybrid ICD-Disease clustering-based classification approach is designed and implemented on the large databases. In this work, a hybrid graph-based clustering algorithm is implemented in order to optimize the data clustering operation for the classification problem. Finally, a weighted neural network is applied on the clustered features for classification process. Experimental results show that the present model has high computational efficiency than the conventional models.
Keywords: neural network, clustering, classification, medical datasets.

1.Introduction

Most of the multi-layer perceptron (MLP) is made up of three layers-input layer, hidden layer, and output layer. The hidden unit nodes have nonlinear activation functions and linear activation functions are available at the output. Every original input layer is multiplied by a weight that is transferred to the other layers. Reinforcement learning (RIL) is a precursor to deep learning, replacing many traditional algorithms of learning. Reinforcement Learning (RIL) has evolved in company with other disciplines like game theory, operation research, multi-agent systems. The main difference between classical methods and RIL is that knowledge is not assumed by the later and it is never dependent on infallible solutions. With hardware requirements, the technical requirements for ANN-based machine learning are met. With the support of hardware enhancement, the deep modules inside ANN have transformed ANN into DNN (Deep Learning Neural Network).

Clustering is one of the vitaltoolsind at a mining and knowledge discovery . Dueto massive amounts of data collected in data bases,cluster analysis has beenrecentlybecomeahighlyactivetopicindatamining.MostlyK-meansclusteringalgorithmisusedtogroupsimilarobjectstogether.But, it requiresthenumberofclusterstobespecifiedinadvancewhichisconsideredtobeoneoftheproblemofthisalgorithm.Theabilitytoautomaticallyclustersimilaritemstogether, allows one to determine hidden similarities and key concepts. It also summarizes a large amount of information intoa small number of clusters. Therefore, biomedical literature has become more complicated for understanding. Thus, there is a necessity of more efficient approaches in order to extract biomedical information from vast numbers of resources. An appropriate mining approach is required to be implemented in order to discover different types of knowledge from biomedical literature. In biomedical texts we can find a degree of term variation. Apart from this, biomedical term can contain numbers, capital letters inside words, hyphens and different special characters

A group of researchers developed an advanced all path kernel technique in order to retrieve PPIs depending upon several lexical and syntactical features. In the subsequent time, approaches those depend upon deep neural network such as convolutional neural networks and recurrent neural networks have become more popular and widely accepted. In case of support vector machine, we have to select the neighboring word features, bag-of-words features, distance features, keywords features and shortest-path features. In case of CNN-based approach, sentence sequence and shortest dependency paths are the inputs. But in case of RNN-based approach, there is only one input that is sentence sequence. The word, part of speech, position and embeddings are considered as the input representation in case of CNN and RNN schemes. At last, the majority voting scheme is implemented on SVM, CNN and RNN schemes. Some other researchers combined RNN and CNN approaches in order to present an extended and advanced hybrid approach. We can mention here that, the inputs for this model are sentence sequences and SDPs produced from the dependency graph. The process of relation extraction is considered as the most important category of knowledge discovery. The most common and prime objective of all researchers is to detect relationships in between different biomedical concepts. Different numbers of approaches are implemented in the biomedical relation extraction process such as co-occurrence statistics, rule based techniques, pattern learning and classification . Some of the mostly used text mining processes are:- named entity recognition, text classification synonym and abbreviation extraction, relationship extraction and hypothesis generation. The named entity recognition process has the responsibility to detect particular names just like gene, protein, drug, chemical out of vast range of text. The process of text classification can be defined as a specific process that can automatically identify importance of a particular document. Apart from all of these, emphasis is also given on recognizing synonyms and term abbreviations. Another important process is the process of relationship extraction. In the process of relationship extraction emphasis is given on the detection of occurrences of a pre-specified relation just like associative, chemical or regulatory relations among various entities. Again, process of hypothesis generation has the responsibility to produce additional interesting relationships those are not directly mentioned before.

Biomedical repositories are large distributed systems in which each data center exchanges data throughout the network without centralized control. Each data node is directly linked to a large number of nodes within the overlay network. It is impractical to gather all the distributed data from the biomedical repositories into a centralized node or site and then perform the conventional data mining techniques. As the size of the biomedical data increases along with the available resources, traditional data mining models fail to find the efficient or optimized searching strategies on protein-protein interactions, gene-protein, and gene-disease associations. Also, traditional document classification models extract hidden patterns in the biomedical repositories to represent the document features in a concise format. Functional associations between genes and diseases are vital to enhancing high throughput in biomedical analysis. But still, there is no optimized technique developed for the detection of functional relationships among genes. Therefore, there is a need to find the functional relationship between genes and diseases using a big-data framework. The volumes of information are growing rapidly in different domains with the growth of distributed peers or networks. Document feature extraction is a reductive transformation of peer documents to generate a summary by selecting important information in the source document(s)[1].

Parallelism is a must to architectures based on DNN. Quick graphical process units, rapid RAM interactions are a major concern of this system. When a new hardware architecture is built, some custom chips must be created along with some vector processors. Specialized hardware and guaranteed communication across bandwidth are some other requirements in a DNN technical setup. The key objective of the new architecture is the concept of developing deep structure learning or hierarchical learning with generated features that shape other algorithms like twig join and N-list. Also, it should consider managing multiple sources of data. The solution would allow for a major deepening of the learning on data representation. [2] developed a novel method for finding the relationships between the gene sets. They used ontology structures to represent the relation between the genes and their properties. The probabilistic method is proposed based on the gene clusters and their properties to predict the new type of gene sets within the limited datasets. The major issues in this method include the false positive rate and error rate of the gene relationships and their neighbouring clusters. [3] proposed a novel approach to latent semantic indexing to cluster the datasets related to the genes. They used 50 biomedical documents to find the genes and with the specified number of clusters to form the related clusters. The proposed latent semantic model requires high computational memory and time as the size of the biomedical gene documents is exponentially increasing. [4] proposed a model for disease prediction in the mining and classification patterns. They used a microarray dataset to rank the pathways and to look for patterns related to the disease in limited data size. They used model random forest classification to filter and classify the patterns of the co-related disease in microarray datasets. For large datasets this model requires high computational storage and memory.

2.Related Works

[5] proposed a novel pattern of gene-based diseases using the model of metagraph construction. In this model, protein to protein interactions and keywords for biological genes are extracted using metagraphic model to find patterns for the disease. This model is restricted to small datasets with no more than 10k instances. [6] proposed a new model for the selection and classification of genes for diseases using the Phase Diagram approach. They used various datasets of the microarray genes for prediction and classification of diseases. PHADIA method is effective for datasets with limited instance space for micro arrays. [7] developed a model for detecting the gene pattern sequence in biomedical repositories. To find the correlation between the two sequences, they used genebank dataset as training data. Also, the proposed hidden markov model uses Genebank database to generate two or three states to represent the gene sequences. This model requires only two gene sequences, as each sequence increases in size or as the number of sequences increases, this model is not efficient for ranking. The above mentioned approaches involve maxi- mum entropy or support vector machines algorithm. The prime objective of this research work is to present an alternative approach for functionally annotating genes. This approach includesconstructionofefficientclassificationschemes,validationmodelsandgraphicalrepresentation of the outcomes. Apart from this, dimension reduction of the dataset is also another prime concern of this research work. The classification schemes are developed by considering the basic concepts of linear discriminant analysis approach. On the other hand, the validation models depend upon the concepts of statistical analysis and interpretation of theoutcomes.Multi-document clustering and feature extraction can be used to reduce inter-cluster variance, thus resolving this flaw. To remove knowledge duplication caused by

multiple original papers, this research considers a feature extraction technique as well as a key phrase clustering and pattern discovery approach.

One of the most important methods in data mining and information exploration is clustering. Cluster analysis has recently become a hot topic in data mining due to large quantities of data obtained in databases. To group related objects together, the K-means clustering algorithm is commonly used. However, one of the problems with this algorithm is that it allows the number of clusters to be determined in advance. The ability to automatically group objects that are similar helps one to discover secret similarities and key concepts. In addition, it condenses a vast volume of data into a smaller number of clusters. As a result, interpreting biomedical literature has become more difficult. As a result, more efficient methods are needed to extract biomedical knowledge from large amounts of data. In order to discover various types of information from biomedical literature, an effective mining method must be introduced. There is a degree of term variance in biomedical texts. Aside from that, a biomedical term can include numbers, capital letters inside words, hyphens, and other special characters.

Document clustering is a method of categorizing text documents into hierarchical clusters or categories, with documents in one cluster being similar and documents in other clusters being different. It's one of the most important aspects of text mining. According to Liping (2005), the development of the internet and analytical processes has paved the way for various clustering techniques. Text mining, in particular, has grown in importance, and it entails a variety of tasks, such as the development of granular taxonomies, document summarization, and so on, in order to extract higher-quality information from text. Most systems exhibit heterogeneous data stream uncertainty (Charu Aggarwal et al 2003). However, the latest methods for clustering heterogeneous unpredictable data streams are unsatisfactory in terms of clustering performance. Guo-Yan Huang et al. (2010) proposed an ambiguity-based clustering approach for heterogeneous data streams. The H-UCF frequency histogram aids in the tracking of categorical statistic characteristics. Initially, the latest technique is proving to be more useful in terms of clustering efficiency than UMicro, since it constructs' n' clusters using a K-prototype algorithm.

[8] reported on a new biomedical language learning techniques which use the Unified Modelling Language (or the Unified Modelling Language, as it is known in North America) . Biological knowledge is created on a continuous basis in terms of how it flows from one generation to the next. Therefore, with each passing day, the amount of medical literature also increases. Ontologies have the duty to serve as the lingua franca. Thus, these books are excellent sources of biomedical information, conceptual understanding. In this study, an advanced technique is applied to generate biomedical ontologies. In terms of presentation, their method is a full use of natural language processing as well as well as evolutionary approaches various concepts and relationships are presented Furthermore, they have developed a new approach in which they use UMLS to merge different concepts as an addition to their previously developed algorithm. Furthermore, they also applied different strategies to find and exploit themes and character types by means of OWL. These archetypes can guide the extraction of the complete knowledge[9]. Finally, they sought to incorporate axioms into their system. Additional research can be done to improve the above model.

[10] found successful in semi-supervised autoencoders in semi-supervised learning applications. The responsibility to generate different types of knowledge like protein-protein interactions, drug interactions, and drug-drug interactions are part of the biological literature[11-14]. The objective of biomedical extraction is to automatically extract various relations in text relevant to the biomedical discipline.

3. Proposed Model

A large part of traditional extraction-based medicine does on supervised machine learning. Consequently, all of the mentioned strategies and techniques are grounded in labeled data. However, there are enormous quantities of biomedical text records in the PubMed that remain unlabeled. In order to alleviate the burden of labeling, computational approaches are preferable. A complex semi-supervised approach uses a variational autoencoder to automate the entire medical relation process. Three things are needed to understand people: classes, encoding, and decoding. This approach utilized new and extremely advanced text mining and network analysis methods to find genes that operate at high altitudes This project's aim is to detect high altitude sickness genes in every functional association This paper claims to have found the gene networks that govern these ailments. Concepts pertaining to gene interaction and analysis have been given in a straight forward manner, rather than in literature. First, co-occurring gene pairs are extracted from the MEDLINE database by a mining algorithm. in the next phase, each and every gene pair has its own weight is combined according to their co-occurrence. At long last, various statistical measures are applied to an efficient ICD network in order to search for new associations. In the future, full text posts and text mining could resolve current and future model problems for this strategy. The amount of biomedical information that can be absorbed is great, since it will assist the scientists to comprehend the complete design and execution of biochemical mechanisms. Moreover, it will also teach you about genes and proteins within gene networks. In the future, researchers can make new models out of protein-protein interactions using this one to add to it. Again, documents within a specific context can be referred to as clusters. All classical document clustering methods ignore the notions of topic and concept distribution. It has the obligation to make the most of the collected sum of documents' Furthermore, this framework makes use of the semantic that data processing consists of the discrimination of terms. It also educates various contexts.

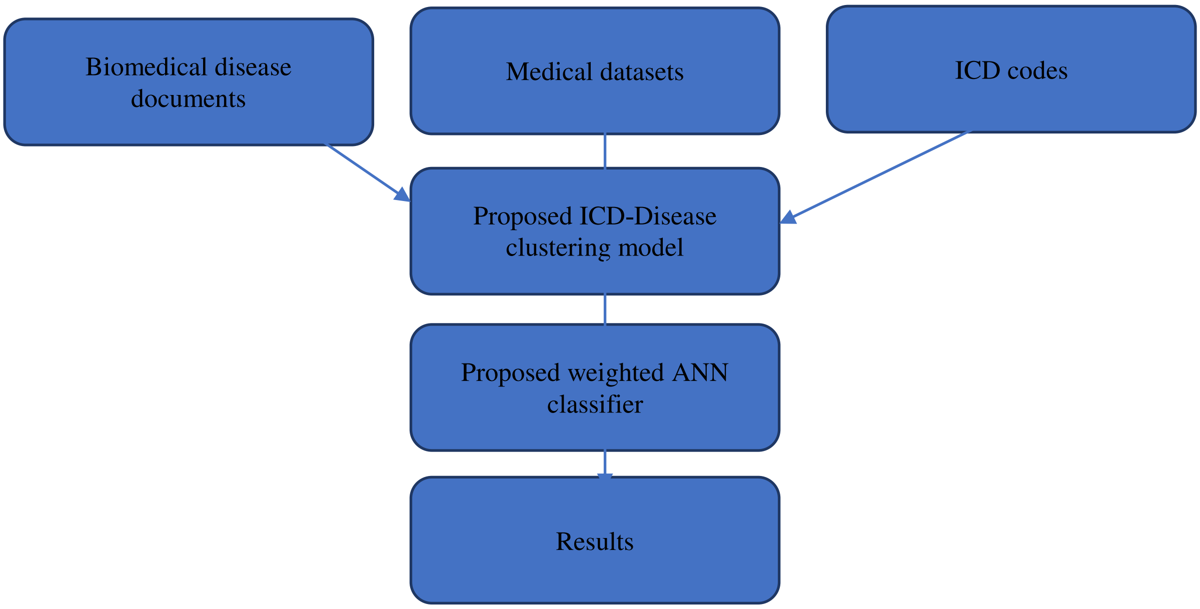

Biomedical disease documents
Medical datasets
ICD codes
Proposed ICD-Disease clustering model
Proposed weighted ANN classifier
Results

Figure 1: Proposed Model

Algorithm 1: Graph clustering based classification

Input :,ICD and Disease classification

Procedure:

```
Initialization of graph paramters.
Get initial clusters using optimized kmean similarity measure on the graph nodes
$\Pr(v_i . v_j)$ = probability that features of $v_i, v_j$ vertices appear in the same document.
$\Pr(v_i / v_j) = \frac{n(v_i, v_j)}{n(v_j)}$
$n(v_i, v_j)$: Number of documents where both features in $v_i, v_j$ matches.
$n(v_j)$: Number of documents which contain features $v_j$.
$ChiSim = \frac{\Pr(v_i . v_j)^2}{\Pr(v_i).\Pr(v_j)}$
Cluster-set[]=Kmeans(G);
For each peer or node $p_k$ in cluster i
Do
For each document $D_{C_i, \vec{\mathbf{d}}_j}$ in cluster i of $p_k$
For each phrase $ph_m$ in $D_{C_i, \vec{\mathbf{d}}_j}$ // m phrases
Do
Features[]=Features( $ph_m$ ); // Disease releted ICD codes
$v_1$ = Features [0];// initialize first term in vertex
If $v_1$ is not in Graph G
Add $v_1$ to Graph G.
Endif
For each term Features [id] // id=2,3….len(terms)
Do
```

16. $v_{id} = Features[\mathrm{id}]$
17. $v_{id-1} = Features[\mathrm{id}-1]$
18. For each biomedical document PBD[i]
    Do
19. If($((v_{id}, v_{id-1}) \in PBD[i])$
20. then

$$\text{Weighted edge feature rank} = \mathrm{WEFR}((\mathrm{v_{id}}, \mathrm{v_{id-1}})) = \frac{-\mathrm{Pr\,ob(BKT[j]/PBD[i])}}{\max\{\mathrm{ICD}_{\mathrm{rank}_1}, \mathrm{ICD(Disease)}_{\mathrm{rank}_2}\}} \log(\frac{\mathrm{Pr\,ob(BKT[j]/PBD[i])}}{\min\{\mathrm{ICD}_{\mathrm{rank}_1}, \mathrm{ICD(Disease)}_{\mathrm{rank}_2}\}})$$

21. $e_{id} = (v_{id-1}, v_{id}, WEFR((v_{id}, v_{id-1})))$
22. If $v_{id} \notin$ G
23. Then
24. Add v_{id} to G
25. End if
26. 7: Construct the filtered top k-clusters FC[k].
27. 8: For each ranked feature FC[i] do
28. Check the distance metric >0
29. If(dist(SC[i],C[k])>0)
30. Then
31. Classify the instance SC[i].
32. End if
33. 9: done
34. Using the standard deviation of the class labels, the T-statistical weighting measure is used to find the variation in the gene characteristics. It is basically the ratio of the class label to the maximized standard deviation.

$$\mathrm{W1} = \frac{\mu_P - \mu_N}{\sqrt{\max\{\sigma_P^2/|P|, \sigma_N^2/|N|\}}} \quad \text{-----(1)}$$

where μ_P is the mean of the positive cluster class samples

μ_N is the mean of the negative cluster class samples.

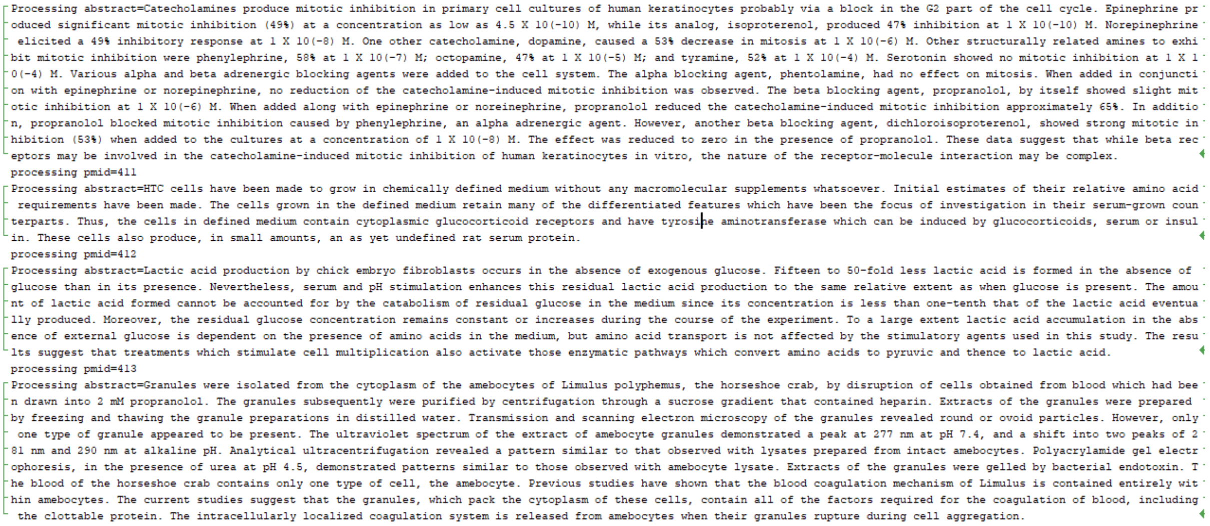
$$\mathrm{W2} = \frac{\mu_P - \mu_N}{\sqrt{\max\{\sigma_P^2/|P|, \sigma_N^2/|N|\}}} \quad \text{-----(2)}$$

35. It is the maximization of feature correlation, hybrid t-test and hybrid SNR ratio. This measure of ranking is used to select the optimum functionality of the binary class in each cluster.
36. Weights W[]=Max{W1,W2}
37. Defining the input, hidden and output layers to neural network algorithm.

4.Experimental results

Experimental results are simulated in java environment with ICD training data and medical disease datasets. Following are the input sample datasets and ICD codes extraction screenshots.

Sample Data in xml format:



```
Processing abstract=Catecholamines produce mitotic inhibition in primary cell cultures of human keratinocytes probably via a block in the G2 part of the cell cycle. Epinephrine pr
oduced significant mitotic inhibition (49%) at a concentration as low as 4.5 X 10(-10) M, while its analog, isoproterenol, produced 47% inhibition at 1 X 10(-10) M. Norepinephrine
 elicited a 49% inhibitory response at 1 X 10(-8) M. One other catecholamine, dopamine, caused a 53% decrease in mitosis at 1 X 10(-6) M. Other structurally related amines to exhi
bit mitotic inhibition were phenylephrine, 58% at 1 X 10(-7) M; octopamine, 47% at 1 X 10(-5) M; and tyramine, 52% at 1 X 10(-4) M. Serotonin showed no mitotic inhibition at 1 X 1
0(-4) M. Various alpha and beta adrenergic blocking agents were added to the cell system. The alpha blocking agent, phentolamine, had no effect on mitosis. When added in conjuncti
on with epinephrine or norepinephrine, no reduction of the catecholamine-induced mitotic inhibition was observed. The beta blocking agent, propranolol, by itself showed slight mit
otic inhibition at 1 X 10(-6) M. When added along with epinephrine or noreinephrine, propranolol reduced the catecholamine-induced mitotic inhibition approximately 65%. In additio
n, propranolol blocked mitotic inhibition caused by phenylephrine, an alpha adrenergic agent. However, another beta blocking agent, dichloroisoproterenol, showed strong mitotic in
hibition (53%) when added to the cultures at a concentration of 1 X 10(-8) M. The effect was reduced to zero in the presence of propranolol. These data suggest that while beta rec
eptors may be involved in the catecholamine-induced mitotic inhibition of human keratinocytes in vitro, the nature of the receptor-molecule interaction may be complex.
processing pmid=411
Processing abstract=HTC cells have been made to grow in chemically defined medium without any macromolecular supplements whatsoever. Initial estimates of their relative amino acid
 requirements have been made. The cells grown in the defined medium retain many of the differentiated features which have been the focus of investigation in their serum-grown coun
terparts. Thus, the cells in defined medium contain cytoplasmic glucocorticoid receptors and have tyrosine aminotransferase which can be induced by glucocorticoids, serum or insul
in. These cells also produce, in small amounts, an as yet undefined rat serum protein.
processing pmid=412
Processing abstract=Lactic acid production by chick embryo fibroblasts occurs in the absence of exogenous glucose. Fifteen to 50-fold less lactic acid is formed in the absence of
glucose than in its presence. Nevertheless, serum and pH stimulation enhances this residual lactic acid production to the same relative extent as when glucose is present. The amou
nt of lactic acid formed cannot be accounted for by the catabolism of residual glucose in the medium since its concentration is less than one-tenth that of the lactic acid eventua
lly produced. Moreover, the residual glucose concentration remains constant or increases during the course of the experiment. To a large extent lactic acid accumulation in the abs
ence of external glucose is dependent on the presence of amino acids in the medium, but amino acid transport is not affected by the stimulatory agents used in this study. The resu
lts suggest that treatments which stimulate cell multiplication also activate those enzymatic pathways which convert amino acids to pyruvic and thence to lactic acid.
processing pmid=413
Processing abstract=Granules were isolated from the cytoplasm of the amebocytes of Limulus polyphemus, the horseshoe crab, by disruption of cells obtained from blood which had bee
n drawn into 2 mM propranolol. The granules subsequently were purified by centrifugation through a sucrose gradient that contained heparin. Extracts of the granules were prepared
by freezing and thawing the granule preparations in distilled water. Transmission and scanning electron microscopy of the granules revealed round or ovoid particles. However, only
 one type of granule appeared to be present. The ultraviolet spectrum of the extract of amebocyte granules demonstrated a peak at 277 nm at pH 7.4, and a shift into two peaks of 2
81 nm and 290 nm at alkaline pH. Analytical ultracentrifugation revealed a pattern similar to that observed with lysates prepared from intact amebocytes. Polyacrylamide gel electr
ophoresis, in the presence of urea at pH 4.5, demonstrated patterns similar to those observed with amebocyte lysate. Extracts of the granules were gelled by bacterial endotoxin. T
he blood of the horseshoe crab contains only one type of cell, the amebocyte. Previous studies have shown that the blood coagulation mechanism of Limulus is contained entirely wit
hin amebocytes. The current studies suggest that the granules, which pack the cytoplasm of these cells, contain all of the factors required for the coagulation of blood, including
 the clottable protein. The intracellularly localized coagulation system is released from amebocytes when their granules rupture during cell aggregation.
```

Table 1 Average ICD term context similarity on different training Documents

Medlinesize	BioNER	Naivebayes	SVM	RF	Proposed Model
#50	0.676	0.856	0.763	0.89	0.979
#25	0.766	0.832	0.785	0.78	0.971
#75	0.788	0.787	0.767	0.91	0.985
#100	0.846	0.898	0.897	0.945	0.964

Table 1 representS the average ICD disease context similarity accuracy on different Medline document sets. From the figure , it is clear that proposed model has high preprocessing rate compare to traditional model in terms of clustering and classification process.

Table 2 :Runtime(secs) comparison of Medline preprocessing models on Medline Datasets

Medline size	BioNER	Naivebayes	SVM	RF	Proposed Model
#25	244	249.5	218.9	239.9	212
#50	256	252.2	324.8	318.7	23
#75	453	253.2	419.1	448.9	289.7
#100	532	485.7	572.1	593.6	301.9
#125	684	543.5	729.5	679.3	387.3

Table 2 describes the runtime comparison of proposed model to the existing models in terms of milli-secs. From the table , it is clear that proposed model has less time complexity compared to traditional models.

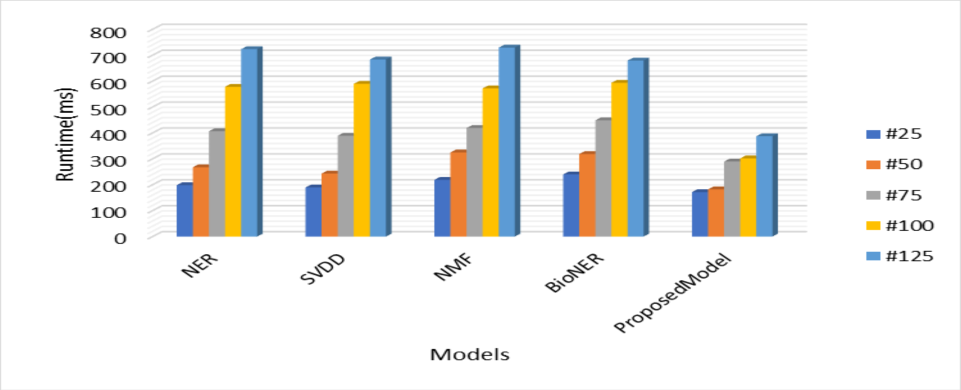


Figure 2:Performance of runtime measure on the bimedical ICD entity sets.

Figure 2 describes the runtime comparison of proposed model to the existing models in terms of milli-secs. From the figure , it is clear that proposed model has less time complexity compared to traditional models.

Table 3: Performance comparison of the proposed model to the existing models in terms of gene or protein related documents count.

		ICD-Disease Related Documents			
Documents size	Bioner	BAYESIAN RANKING	NMF	IDR	Proposed
#5k	485	574	746	757	1043
#10k	2874	3874	5833	5893	6987
#15k	7973	8944	9763	10723	13878
#20k	8828	10883	13781	15872	18567
#50k	20848	25873	28774	34788	43847

Table 3, describes the number of relevant document extraction using the ICD and Disease patterns . From the table, it is observed that the traditional models have less document extraction process based on ICD-disease compared to the proposed model.

5.Conclusion

In this paper, a hybrid cluster-based classification model is designed and implemented on the medical databases. Since, most of the conventional approaches are difficult to find the ICD codes in the medical databases, it is necessary to integrate the ICD codes in the biomedical medical datasets. In this paper, a hybrid clustering model is implemented on the training medical datasets. These clustered ICD codes are given to the training medical datasets for classification problem. Experimental results show that the present model has high computational efficiency than the conventional approaches on large medical datasets.

References

[1] S. Belciug and F. Gorunescu, “Learning a single-hidden layer feedforward neural network using a rank correlation-based strategy with application to high dimensional gene expression and proteomic spectra datasets in cancer detection”,Journal of Biomedical Informatics 83 (2018) 159–166.

[2] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez and F. Herrera , A review of microarray datasets and applied feature selection methods, Information Sciences 282 (2014) 111–135.

[3] V. Bolón-Canedo, N. Sánchez-Maro˜no and A. Alonso-Betanzos, Distributed feature selection: An application to microarray dataclassification, Applied Soft Computing 30 (2015) 136–150.

[4] S. H. Bouazza, K. Auhmani, A. Zeroual and N. Hamdi, Selecting significant marker genes from microarray data by filter approach for cancer diagnosis, Procedia Computer Science 127 (2018) 300–309.

[5] S. Chormungea and S. Jena, Correlation based feature selection with clustering for high dimensional data,Journal of Electrical Systems and Information Technology, 2017.

[6] R. Dash, A Two Stage Grading Approach for Feature Selection and Classification of Microarray Data using Pareto based Feature Ranking Techniques: A Case Study, Journal of King Saud University - Computer and Information Sciences.

[7] M. Ghosh , S. Begum , R. Sarkar , D. Chakraborty and U. Maulik, Recursive Memetic Algorithm for Gene Selection in Microarray Data, Expert Systems with Applications.

[8] S. Guo, D. Guo, L. Chen and Q. Jiang, A L1-regularized feature selection method for local dimension reduction on microarray data, Computational Biology and Chemistry 67 (2017) 92–101.

[9] Bangare S.L., Pradeepini G., Patil S.T. (2017),’Brain tumor classification using mixed method approach’,2017 International Conference on Information Communication and Embedded Systems, ICICES 2017,(),PP.-.

[10] Potharaju S.P., Sreedevi M. (2017),’A novel M-cluster of feature selection approach based on symmetrical uncertainty for increasing classification accuracy of medical datasets’,Journal of Engineering Science and Technology Review,10(6),PP.154-162.

[11] Potharaju S.P., Sreedevi M. (2017),’A novel clustering based candidate feature selection framework using correlation coefficient for improving classification performance’,Journal of Engineering Science and Technology Review,10(6),PP.38-43.

[12] Rehman S.N., Hussain M.A. (2017),'Glaucoma classification based on contourlet transform',Indian Journal of Public Health Research and Development,8(3),PP.106-108.

[13] Patra B., Bisoyi S.S. (2018),'CFSES Optimization Feature Selection with Neural Network Classification for Microarray Data Analysis',Proceedings - 2nd International Conference on Data Science and Business Analytics, ICDSBA 2018, (),PP. 45-50

[13] Lakshmi Prasanna P., Rajeswara Rao D. (2018),'Text classification using artificial neural networks',International Journal of Engineering and Technology(UAE),7 (0),PP. 603-606

[14]Kolli S., Sreedevi M. (2018),'Prototype analysis of different data mining classification and clustering approaches',ARPN Journal of Engineering and Applied Sciences,13 (9),PP. 3129-3135