International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Discriminant Pearson Correlative Feature Selection based Gentle Adaboost Classification for Medical Document Mining

P.Poongothai,T.Devi

Abstract: ***This paper examines Discriminant Pearson Correlative Analysis Based Multivariate Gentle Adaboost Classification (DPCA-MGAC) and it is used to improve the performance of medical document mining with minimum time complexity. A large number of documents are collected from PubMed databases through the semantic-based search. Processes such as removing stop words, stemming, features identification, selection of features i.e., relevant keywords for document classification are carried out. The significant feature selection is carried out using DPCA, and with the selected features the documents are categorized into different classes using MGAC. This classification process combines the results of all weak learners and makes a strong classification in order to improve the precision of medical data mining and minimizes the false positive rate. Experimental evaluation has been performed using PubMed database.***

Keywords: Boosting, document classification, document collection, Text mining.

I. INTRODUCTION

Mining of biomedical information has become a challenging issue in the last few years as the growth of the webis dynamic in nature. With the several biomedical documents on the web, analyzing the text data related to the diabetics plays a significant task. Text data analysis is executed through various data mining techniques. Classification of diabetes-related documents becomes difficult to process since it has high dimensionality. It is reduced at the preprocessing stage to improve classification accuracy. A Graph-based Semantic Ranking Model (GSRM) has been introduced for retrieving the medical documents through the semantic search based classification [9]. But the false positive rate during the classification was not minimized.

A novel lexical approach based on k-nearest neighbor (LKNN) algorithm was designed for classifying the text documents in the biomedical domain [10]. But LKNN has more time complexity since it failed to use any feature selection. Multi-stage NLP system has been developed to categorize the heart disease risk document in diabetic patient's time [6].

Manuscript published on 30 September 2019
* Correspondence Author
P.Poongothai*, Ph.D. Research Scholar, Department of Computer Applications, Bharathiar University, Coimbatore, Tamil Nadu, India.
Dr.Prof.T.Devi, Professor and Head, Department of Computer Applications, Bharathiar University, Coimbatore, Tamil Nadu, India.

The Associative Rule integrated with Naive Bayes based Classification was presented for text categorization. But this model hadn't reduced the error during the classification [13]. A query expansion and ranking technique were introduced to retrieve the high-quality studies for the improvement of clinical strategy [3].

The technique failed to use semantic search for retrieval. The major issues are lack of accuracy, high false positive rate, more time complexity, failure to perform semantic search based classification. In order to overcome all these, a novel technique called DPCA-MGAC has been introduced.

II. RELATED WORKS

A semi-supervised spectral clustering method was introduced for grouping the MEDLINE Documents [9]. However, error rate was not considered in that for improving their performance. Nonnegative Matrix Factorization (NMF) and multi-view NMF method were introduced for clustering the clinical documents based on sample-feature matrices [14]. The method does not use patient's age, gender/demographical information, to increase the clustering performance. A hybrid Association Rules and Hidden Markov Model has been developed for classifying the biomedical texts along with their content [5]. The model has more time complexity because it does not train with less number of documents. The rule-based multi-pass sieve technique was introduced for categorizing the texts from PDF documents [4].

A hybrid feature selection approach based on Genetic Algorithm (GA) was introduced for choosing the relevant feature and minimizing the dimensionality [1]. But, document classification remains an open issue. A hybrid filter and wrapper selection method has been developed to select the keywords for text classification [12]. The method does not minimize the complexity at a required level. A hierarchical neural architecture was introduced for classifying the documents with minimum time consumption [7]. The classifier does not minimize the incorrect classification.

A regular expression discovery (RED) algorithm was developed for classifying the clinical text. The algorithm does not improve the classification performance [2]. A kernel-based extreme learning machine (ELM) was developed for classifying the clinical texts. But the method failed to address the clinical text classification time [8]. Semi-supervised learning approach was introduced for the classification of huge text data [15].

The approach does not use other machine learning algorithms for the bigger document set. The issues identified from the above said reviews are overcome by introducing a novel technique called, DPCA-MGAC.

III. DPCA-MGAC FOR TEXT DOCUMENT MINING

The proposed DPCA-MGAC technique executes both preprocessing and feature selection for reducing dimensionality of the data set before the classification. The DPCA-MGAC technique also improves the document classification with the selected feature set. The overall architecture of the DPCA-MGAC technique is demonstrated in fig. 1.

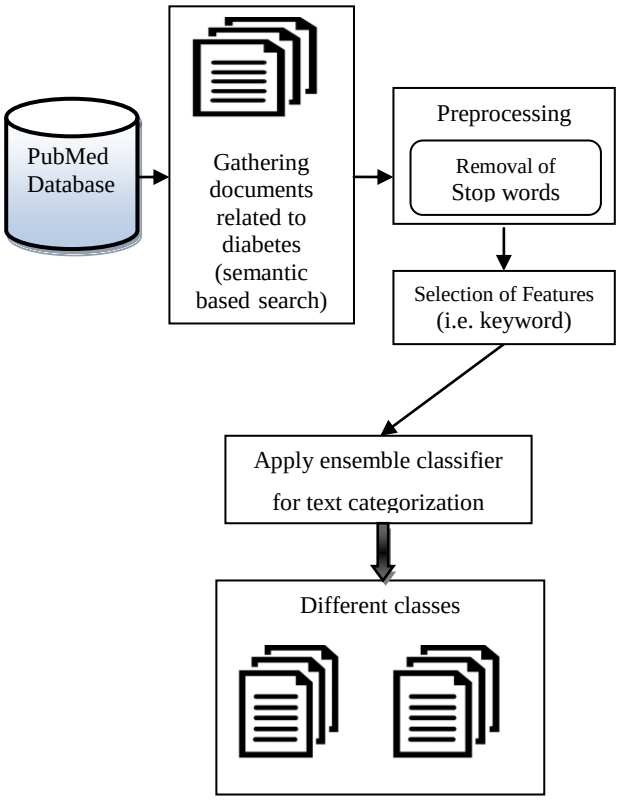

Figure 1. Architecture diagram of the DPCA-MGAC technique

Fig. 1 illustrates an architecture diagram of the proposed DPCA-MGAC technique for categorizing the text documents as treatment i.e., medicine for men, women and also children.

A. Document collection

The first process in the DPCA-MGAC technique is the document collection. Semantic-based search is carried out using a query on 'diabetes and the semantic relation' on the collected documents. Semantic search is used to improve search accuracy of diabetes-related documents. Semantic search systems include the context of search, a variation of words and synonyms for those words

$$D = \{doc_1, doc_2, doc_3, \ldots . doc_n\} \in R^D \qquad (1)$$

In the above equation (1), D denotes a document set$\{doc_1, doc_2, doc_3, \ldots . doc_n\}$, R^D denotes a pubmed database. Each document has 'm' features.

$$doc_1 = \{t_1, t_2, t_3, \ldots t_m\} \qquad (2)$$

In the above equation (2), doc_1denotes a document which contains the set of features $t_1, t_2, t_3, \ldots t_m$ (i.e. keywords). The dimensionality is equal to the number of the keywords which is reduced by the feature selection in the given documents.

B. Preprocessing

Preprocessing comprises unstructured data and it is not feasible to classify the documents directly using data mining techniques. Preprocessing is essential in text mining and in performing, the documents are converted into a list of features or keywords by performing stop words removal and word stemming process are performed [11].Stop words are the words which occur repeatedly in the documents and they do not provide any meaning within the document. Stop words are "and", "are", "this" and so on. These words are not used in the text document classification. Therefore, these words are removed from the given documents. Word stemming is the process of lessening the words to their word base form. For example, if the word ends with 'ed', 'ing', 'ly', and these parts are removed and stem words are retained within the documents.

C. DPCA analysis based Features selection

After preprocessing, the relevant features for the classification are chosen from text files. The stemmed words are given as the input for feature selection and reduction. As a result, the classification time gets minimised. The proposed technique DPCA is used to discover a linear combination of features from the feature set. Consider the number of features in the feature space,

$$f = \{t_1, t_2, t_3, \ldots t_m\} \qquad (3)$$

In the above equation (3), f denotes a features set $t_1, t_2, t_3, \ldots t_m$. The DPCA distinguishes relevant features among the several features using linear discriminant vector. The analysis uses a fisher criterion rule that is defined as the ratio between the class scatter and within the class scatter. The scatter is also used to find whether the features are correlated within the classes. The fisher criterion is formulated as follows,

$$F_s = \frac{\sigma_{between}}{\sigma_{within}} = \frac{v^T c_b(t_i) w}{v^T c_w(t_i) w} \qquad (4)$$

In the above equation (4), F_s denotes a Fisher separation criterion function based on,$\sigma_{between}$, $\sigma_{within}$where they denote the variance between and within the subset respectively. v^T denotes a discriminant vector to project the features into any one of the subset based on optimal projection direction 'w'. $c_b(t_i)$ denotes a between class scatter, $c_w(t_i)$ denotes a within the class scatter and is computed based on the mean value of subset and features.

$$c_b(t_i) = \sum m * (\mu_s - \mu_t)(\mu_s - \mu_t)^T \qquad (5)$$

In the above equation (5),$c_b(f_i)$ represents a class scatter between the classes, $'m'$ denotes a number of features, μ_f denotes a mean

value of the features, μ_s represents a mean value of subset. T denotes a transpose matrix. The scatter within the subset is computed as follows:

$$c_w(f_i) = \sum\sum(t_i - \mu_s)(t_i - \mu_s)^T \quad (6)$$

In the above equation (6), $c_w(f_i)$ represents a class scatter within the subset and t_i denotes features, μ_c represents a mean of the class. T denotes a transpose of a matrix. The mean value of the feature and subset is computed as follows,

$$\mu_s = \frac{1}{n^k}\sum_{i=1}^{m} t_i \quad (7)$$

In the above equation (7), μ_s denotes a mean value of feature subset, n^k denotes a number of features in k^{th} subset. Mean of the feature set is computed as follows,

$$\mu_t = \frac{1}{n} * \sum_{t_i \in F} t_i \quad (8)$$

In the above equation (8),μ_t represents mean value of the feature set, 'n' denotes a number of features (t_i) ina feature vector space$F$. Therefore, the projection vectors in the separation function minimize the variance and maximize the correlation inside the class. The correlation inside the class is measured to discover redundancy among the selected features. The proposed technique uses the Pearson product-moment correlation coefficient for finding the correlation between the features inside the subset. It is formulated as follows,

$$\rho\left(t_{1,}t_2\right) = \frac{n*\sum t_1 t_2 - (\sum t_1)(\sum t_2)}{\sqrt{[n*\sum {t_1}^2 - (\sum t_1)^2][n*\sum {t_2}^2 - (\sum t_2)^2]}} \quad (9)$$

In the above equation (9), ρ denotes correlation coefficient. $'n'$ represents a number of features in the subsets. $\sum t_1 t_2$ denotes a sum of the product of paired score of two features. ${t_1}^2$ denotes a squared score of the feature t_1 and ${t_2}^2$ denotes a squared score of t_2. The correlation coefficient $'\rho'$ provides either "-1" or "+1", and if the coefficient provides '+1 (high score), then the features within the subset are correlated whereas '-1'indicates negative correlation between two features inside the subsets. Based on the positive correlation between the features, the redundancy features are eliminated and high scored features such as age, gender and gestational diabetes are selected for further classification. As a result, time complexity gets minimized.

D. Multivariate Gentle Adaboost for text document classification

The DPCA-MGAC technique categorizes the documents as treatment (medicine) for men, women and children with the selected features using Multivariate Gentle Adaboost technique. It is a machine learning ensemble meta-algorithm to increase the accuracy of the statistical classification. It is exploited to convert weak learners into strong ones by combining the several classifications with randomly created training sets. Multivariate kernelized support vector machine classifier is used as a weak learner for document categorization. Multivariate is often called as the weak learner classifies the documents with more than one feature.

The strong classification results minimize the false positive rate and improve the true classification. Initially, a number of weak learners such as kernelized SVC are constructed and classifies the documents using the optimal separating hyperplane. The kernelized support vector classifier categorizes all the documents on one side of the decision boundary as belonging to one class and the remaining documents on the other side belonging to the other class. These two classes are obtained through the separating hyperplane which is mathematically represented by equation 10

$$\alpha.x_i - k = 0 \quad (10)$$

Where, x_i denotes training samples such as documents, k represents a bias and α represents a normal weight of the feature. All the training samples are linearly separable in given dimensional space. The two marginal hyperplanes are expressed by following two equations,

$$\mathrm{M}_1 = \alpha.x_i - k \geq 0\,, \forall\, x_i \text{ of class 1} \quad (11)$$

$$M_2 = \alpha.x_i - k \leq 0 \quad \forall\, x_i \text{ of class 2} \quad (12)$$

The distance among the two marginal hyperplanes M_1 and M_2 is as large as possible. The distance is computed using the formula given in (13)

$$D = \frac{2}{\|\alpha\|} \quad (13)$$

Where, D denotes a distance between the two marginal hyperplane M_1 andM_2, α denotes a normal vector of the feature. The distance between the planes is maximized while minimizingα. Based on the hyperplane and the marginal hyperplane, the classifier categorizes the documents using the following mathematical eqn. (14),

$$\mathrm{h_i}(x_i) = sgn\sum \alpha y_i k(doc_i, c_i) \quad (14)$$

Where, $\mathrm{h_i}(x_i)$ denotes output of a kernelized classification results, "sgn" represents whether the predicted classification output is positive or negative. The positive result of the classifier provides the higher similarity between the documents whereas negative result denotes less similarity. $k(doc_i, c_i)$ denotes a kernel function that measures the similarity between any pair of documents. α denotes a weight of the features depending on the number of occurrences within the document. Based on the weight value, the classifier performs semantic search and categorizes the documents into a different class. The semantic search in the document classification uses the alternative word for that selected feature and produces more relevant documents results. The semantic relations between the features in the documents are categorized into a particular class. The kernelized SVC uses the Gaussian kernel function to measure the similarity. The set of documents are assigned to a particular class depending on their similarity a kernel function is expressed as follows:

$$k\left(doc_i, c_j\right) = exp\left(-\frac{1}{2\,\sigma^2}\left\|doc_i - c_j\right\|^2\right) \quad (15)$$

Where, $k\left(doc_i, c_j\right)$ represents a kernel function,σ denotes a free parameter. $\left\|doc_i - c_j\right\|^2$represents a squared euclidean distance between the documentsdoc_i and class c_j . The support vector classifier uses kernel function for correctly classifying the documents as treatment (medicine) for men and women. As a result, the corresponding documents are assigned to the particular class. After the classification**,** the weak classifier output has some training loss resulting in minimization of the classification accuracy and increases the false positive rate. In order to minimize these kinds of issues during the classification, the gentle adaboost technique has been applied. Boosting is a machine learning technique to create strong classifier by combining the outputs of all weak learners.

$$Y = \sum_{i=1}^{n} \mathrm{h_i}(x_i) \quad (16)$$

In the above equation (16),Y represents the strong classification results, $\mathrm{h_i}(x_i)$ denotes an output of weak learners. Then the weight assigned to each weak learner is expressed as follows:

$$\omega_i \rightarrow h_1(x_i), h_2(x_i), \ldots . h_n(x_i) \quad (17)$$

where, ω_i denotes a weight assigned to the weak learner $\mathrm{h_i}(x_i)$. Gentle adaboost classifier minimizes the training loss i.e. error of the model to avoid over-fitting. Over-fitting is the modeling error that occurs when the function directly fits into a smaller amount of data, but it is difficult to process more data with minimum error. Therefore, the training loss of the each weak classifier is computed as follows:

$$L = (y_i - \mathrm{h_i}(x_i))^2 \quad (18)$$

where, L denotes a training loss for the weak learners. y_i denotes a actual output of the each classifier and $\mathrm{h_i}(x_i)$ denotes a predicted output. Based on the error calculation, the initial weight of each weak learner is updated as follows:

$$\omega'' = \frac{\omega_i \mathrm{e}^{-y_i h_i(x_i)}}{\gamma} \quad (19)$$

Where, ω'' represents updated weight for weak classifier andω_i denotes an original weight of base learner. Here, $h_i(x_i)$ denotes the classification results of i^{th}base classifier, γ represents the normalization factor. In conventional adaboost, the adjustment coefficient is used for updating the weight of each weak learner. But in case of the gentle adaboost, there is no need to use any adjustment coefficient for classification. As a result, the weak learner with less error rate exhibits perfect medical document classification performance which further minimizes the false positive rate and also reduces the time complexity. Finally, the gentle adaboost classifier uses gradient descent function to find the weak learner with minimum error.

$$Y = \arg\min L\{\mathrm{h_i}(x_i)\} \quad (20)$$

where, $arg\ min$ is an argument minimum function and 'L' represents an error of the weak learner, Y represents strong classification results. The obtained strong classifier output to improve the documents classification accuracy i.e. precision and minimizes the incorrect classification i.e. false positive rate. The algorithmic process of the proposed DPCA-MGAC technique is described.

Input: Number of documents $doc_1, doc_2, doc_3, \ldots . doc_n$
Output : Improve document classification accuracy
1. Perform semantic search for Multivariate Adaboost Text Classification algorithm
2. Collect doc_n related to diabetes from R^D
3. Perform preprocessing to remove stop and stem words
4. **For each** doc_n
5. Extract features $t_1, t_2, t_3, \ldots t_m$
6. Measure class scatter of $c_b(t_i)$ and $c_w(f_i)$
7. Find the correlation $\rho\left(t_1, t_2\right)$
8. Select the relevant features and remove redundancy
9. **End for**
10. **For all** doc_n with selected features $t_1, t_2, t_3, \ldots t_m$
11. Construct weak KSVC learners $\mathrm{h_i}(x_i)$
12. Combine all weak KSVC learners results $\sum_{i=1}^{n} \mathrm{h_i}(x_i)$
13. Initialize the similar weight $\omega_i \rightarrow \mathrm{h_i}(x_i)$
14. Calculate training loss L for the each weak classifier $\mathrm{h_i}(x_i)$
15. Update the weight ω''
16. Select best weak learner $arg\ min\ L\{h_i(x_i)\}$
17. Obtain the strong classification results
18. **End for**
19. **End**

Algorithm 1: Discriminant Pearson Correlative Analysis Based Multivariate Gentle Adaboost Classification

Algorithm 1 clearly describes the document classification with the selected features. Initially, the proposed technique performs the semantic search to collect the documents related to diabetes. After that, the preprocessing is carried out to remove the stop words and stem words in the given documents. This helps to reduce the time complexity. The DPCA-MGAC technique constructs several weak learners to classify the documents with the selected features. DPCA-MGAC technique improves classification accuracy for mining the diabetes-related documents in medical set. As a result, the proposed technique improves the precision and minimizes the false positive rate.

IV. EXPERIMENTAL EVALUATION

An experimental evaluation of the proposed DPCA-MGAC technique and existing methods namely Graph-based Semantic Ranking Model (GSRM) and LKNN (Lexical KNN) algorithms are implemented using MATLAB and PubMed database as input [9], [10].

V. RESULTS AND DISCUSSION

The results and discussion of proposed DPCA-MGAC technique and existing methods are described with various performance metrics such as precision, false positive rate and time complexity. With the help of these parametric analysis, the comparison between three methods namely DPCA-MGAC technique, GSRM and LKNN is performed [9], [10]. The comparison results of three methods are explained in following section.

A. Impact of precision

The precision is measured as the ratio of the numbers of documents that are correctly classified into different classes to the total number of collected documents. It is measured in terms of percentage (%).

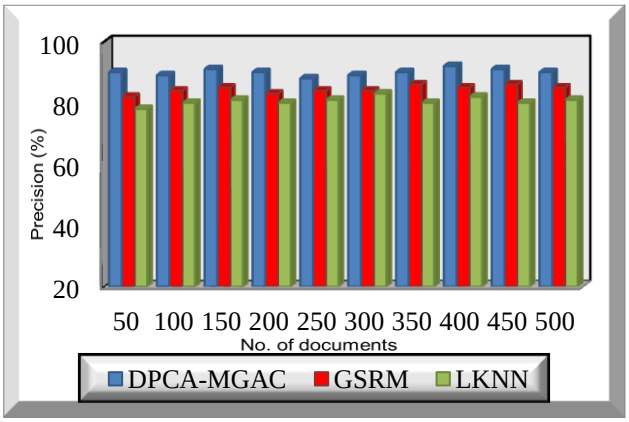

Figure 2 performance results of precision

Fig. 2 illustrates the performance results of precision with respect to a number of documents where the numbers of documents related to diabetes are taken in 'x' direction and the performance results of precision are shown in 'y' direction. As a result, the two-dimensional graphical views of the precision with three different methods DPCA-MGAC technique, and existing GSRM and LKNN are clearly illustrated [9], [10]. It clearly depicts that the precision is considerably increased using DPCA-MGAC technique than the other two existing methods.

Consider the number of documents taken from PubMed database. The proposed DPCA-MGAC technique initially performs the semantic search to gather the number of documents related to diabetes. Based on the search, the semantically related documents are collected. Totally ten different runs are carried out with the number of documents. After performing the ten runs, the precision result of the proposed technique is compared with two existing methods. Then the average comparison results show that the performance enhancement of the DPCA-MGAC technique. As a result of the comparison, the precision is increased by 6% and 12% when compared to existing GSRM and LKNN respectively [9], [10].

B. Impact of False positive rate

The false positive rate is defined as the ratio of number of documents that are incorrectly classified to the total number of documents.

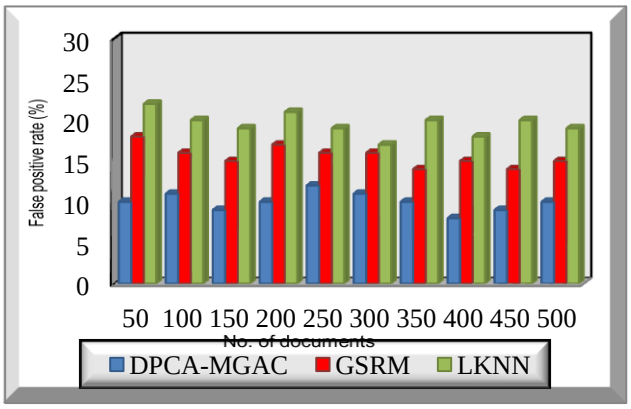

Figure 3 Performance results of false positive rate

As shown in fig. 3, the performance results in false positive rate versus a number of documents are illustrated. The above two-dimensional graphical representation clearly illustrates the false positive rate of three methods namely DPCA-MGAC technique, GSRM and LKNN [9], [10]. The performance result of the false positive rate is significantly minimized by DPCA-MGAC technique. This improvement of the proposed DPCA-MGAC technique is obtained by applying the boosting classifier.

The gentle AdaBoost classifier combines all the base learners i.e multivariate kernelized support vector classifier. After classification, the boosting technique combines the entire weak learners and the weight is assigned to the results of base learners. The loss between the actual and predicted results is computed for each base leaner. The weight of each base learner is appropriated based on the error value. Based on the updated weight, the DPCA-MGAC technique uses gradient descent function to determine the best classification results with minimum error rate. This helps to minimize the incorrect classification. But the existing methods use a single classifier and also failed to compute the error for document categorization and hence, they do not minimize the incorrect classification. This problem is resolved by using an ensemble technique in the DPCA-MGAC. The comparison between the proposed and existing methods shows that the DPCA-MGAC technique significantly minimized the metric or parametric by 36% when compared to existing GSRM. In addition, the performance of DPCA-MGAC technique improves the metric or parametric by 48% when compared to other existing methods and LKNN.

C. Impact of time complexity

Time complexity is defined as the amount of time required to classify the document depending on selected features.

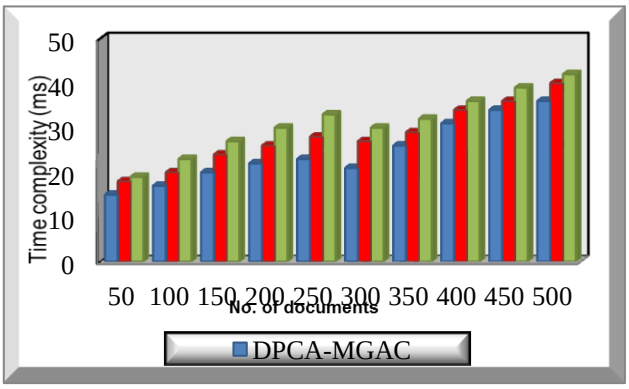

Figure 4 Performance results of time complexity

Fig. 4 depicts the performance results of time complexity versus a number of documents. The graphical results clearly demonstrate that the time complexity of the proposed DPCA-MGAC is considerably minimized than the existing methods. This is because, the DPCA-MGAC initially collects the number of documents related to diabetes from the PubMed database. Then the preprocessing is performed to remove the stop words as well as stem words. In the preprocessing step, the XML documents are converted into a text document. After that, the features are extracted from the documents. Each document contains many features and for categorizing the documents, the significant features from the documents are selected using discriminant pearson correlative analysis. In this analysis, the features that frequently occur in the document are identified through the discriminant vector. This vector separates the feature space into different subsets. After that, the redundancies among the features are determined by the correlation measure. Pearson correlation is used to find the more relevant features and as a result, the features related to Diabetes are selected. The document classification is performed with the selected features and their semantic words for classification resulting in minimizing the complexity.

Consider the 50 documents, the time for classification using DPCA-MGAC technique is 15ms whereas time complexity of other existing methods are 18ms and 19ms respectively. The above performance results clearly illustrate that the time complexity of the DPCA-MGAC technique is minimized by 14% and 22% when compared to existing GSRM and LKNN respectively. The above results and discussions clearly show that the DPCA-MGAC technique considerably improves the document classification with minimum time and false positive rate.

VI. CONCLUSION

An efficient technique called "Discriminant Pearson Correlative Analysis based Multivariate Gentle Adaboost Classification (DPCA-MGAC)"has been introduced for improving the effectiveness or efficiency text document categorization. DPCA-MGAC technique collects the semantically relevant documents on diabetics from the medical database for further processing. Then, the preprocessing is carried out to remove the unwanted words from the text documents and the relevant features are chosen for improving the classification process. Finally, the classification is done by applying a Multivariate Gentle Adaboost Classification. The boosting technique uses the Kernelized SVM classifier as a weak learner to categorise the documents into two distinct classes as a treatment for men and women with selected features. After classification, the training loss of each weak leaner is calculated to make strong classification results and the strong classification result reduces the incorrect classification. Finally, the quality of DPCA-MGAC technique and existing methods are analyzed with parameters such as precision, false positive rate and time complexity. The experimental results show that the DPCA-MGAC technique improves the precision with minimum false positive rate and time complexity when compared to existing methods.

ACKNOWLEDGEMENTS

The authors thank the University Grant Commission for providing Rajiv Gandhi National Fellowship to support this research by research fellowship.

REFERENCES

1. Abdullah Saeed Ghareb, Azuraliza Abu Bakar, Abdul Razak Hamdan, "Hybrid feature selection based on enhanced genetic algorithm for text categorization", Expert Systems with Applications, Elsevier, Vol. 49, 2016, pp.31-47.
2. Duy Duc An Bui and Qing Zeng-Treitler, "Learning regular expressions for clinical text classification", Journal of the American Medical Informatics Association, 21(5), 2014, pp. 850–857.
3. Duy Duc An Bui, Siddhartha Jonnalagadda, Guilherme Del Fiol, "Automatically finding relevant citations for clinical guideline Development", Journal of Biomedical Informatics, Elsevier, Vol. 57, 2015, pp. 436–445.
4. Duy Duc An Bui, Guilherme Del Fi a, Siddhartha Jonnalagadda, "PDF text classification to leverage information extraction from publication reports", Journal of Biomedical Informatics, Elsevier, Vol. 61, June 2016, pp.. 141-148.
5. Huda Ali Al-qozani and Khalil saeed Al-wagih, "A Hybrid Approach of Association Rule & Hidden Markov Model to Improve Efficiency Medical Text Classification", International Journal of Computer Applications Technology and Research 7(2), 2018, pp. 45-52.
6. Jay Urbain, "Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models", Journal of Biomedical Informatics, Elsevier, Vol. 58, 2015, pp. S143–S149.
7. Jianming Zheng, Yupu Guo, Chong Feng and Honghui Chen, "A Hierarchical Neural-Network-Based Document Representation Approach for Text Classification", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Vol. 2018, March 2018, pp. 1-10.
8. Paula Lauren, Guangzhi Qu, Feng Zhang, Amaury Lendasse, "Discriminant document embeddings with an extreme learning machine for classifying clinical narratives", Neurocomputing, Vol. 277, 2018, pp. 129-138.
9. Qing Zhao, Yangyang Kang1, Jianqiang Li, Dan Wang, "Exploiting the semantic graph for the representation and retrieval of medical documents", Computers in Biology and Medicine, Elsevier, Vol. 101, 2018, pp. 39–50.
10. Rajni Jindal and Shweta Taneja, "A Lexical Approach for Text Categorization of Medical Documents", Procedia Computer Science, Elsevier, Vol. 46, 2015, pp. 314 – 32.
11. SagarImambi.S, Sudha. T, "A novel feature selection method for classification of medical documents from pubmed", International Journal of Computer Applications, Vol. 26, July 2011, pp. 29-33.
12. Serkan Gunal, "Hybrid feature selection for text classification", Turkish Journal of Electrical Engineering & Computer Sciences, 20(2), 2012, pp. 1296-1311.

13. Wa'el Hadi, Qasem A. Al-Radaideh, Samer Alhawari, "Integrating associative rule-based classification with Naïve Bayes for text classification", Applied Soft Computing, Elsevier, Vol. 69, August 2018, pp. 344-356.
14. Yuan Ling ,Xuelian Pan, Guangrong Li, Xiaohua Hu, "Clinical Documents Clustering Based on Medication/Symptom Names Using Multi-View Nonnegative Matrix Factorization", IEEE Transactions on Nano Bioscience, 14(5), 2015, pp. 500 – 504.
15. Zewen Xu, Jianqiang Li, Bo Liu, Jing Bi, Rong Li, Rui Mao, "Semi-supervised learning in large scale text categorization", Journal of Shanghai Jiaotong University (Science), Springer, 22(3), 2017, pp. 291–302.
16. ftp://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/

AUTHOR PROFILE

Ms. P. Poongothai received her B.Sc(CS) from Velalar College for Women, Erode in 2008, Master of Computer Applications from Vellalar Ccollege of Engineering &Technology, Erode in 2012 and M.Phil in Computer Science from Bharathiar University, Coimbatore in 2014. Her area of interest is Data Mining and Concurrent Engineering.

Dr. T. Devi received Master of Computer Applications from P.S.G. College of Technology, Coimbatore in 1987 and Ph.D. from the University of Warwick, United Kingdom in 1998. She is presently heading the Department of Computer Applications, Bharathiar University, Coimbatore. Prior to joining Bharathiar University, she was an Associate Professor in Indian Institute of Foreign Trade, New Delhi. Her current research includes Software Engineering, Product Introduction, Technical Process Management and Concurrent Engineering. She has contributed more than 150 research papers in various national / international conferences / journals.