International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-11 Issue-4, April 2022

# Exploration on Document Taxonomy by Ganb Algorithm

**R. Sathish Babu**

***Abstract:*** ***In this research, we propose an integrated classification GANB algorithm that combines a feature extractor with a classifier to construct a classification model. The feature extractor automates the examination of raw pre-processed unstructured documents. Following feature extraction, categorization generates meaningful classes based on the supplied features. The study uses a genetic algorithm (GA) for feature extraction and Naïve Bayes(NB) for classification purposes. The simulation evaluates the suggested classification model's accuracy, sensitivity, specificity, and f-measure using various performance indicators. Over the Medline cancer datasets, the suggested GANB gets a higher classification rate than existing approaches.***

***Keywords*****:** ***Genetic Algorithm, Naïve Bayes, Feature Extraction, Classification***

## I. INTRODUCTION

To improve quality, compare efficacy research, evidence-based medicine, and set up solid genomic analysis phenotypes, where online data are increasingly important. Unfortunately, most online data implementations have been conceived to enable one-on-one interactions and not support aggregated data analysis as many secondary uses require. Consequently, efforts to clinical information must confront a few widespread data standards and large amounts of potentially helpful information stored as free, structured text [1]. Researchers have responded by inventing and using information retrieval (IR) technologies, including natural language processing (NLP), information extraction, and machine learning methods. Despite the fact that there have been over 20 years of experiments in IR performance, the complexity of the task and technical limitations to accessing clinical IR technologies have precluded broad application and translation. There exist several classification algorithm [2] – [5], where most of the methods depend on machine learning models like GA [6-10]. However, most of the ways stuck with the local optimum solution. The study proposes a document classification paradigm that includes stemming, stopword elimination, tokenization, filtering, and classification. Furthermore, the categorization of Medline document collections is improved by feature extraction utilizing evolutionary algorithms and Nave Bayes' classification goals.

Manuscript received on 27 March 2022.
Revised Manuscript received on 30 March 2022.
Manuscript published on 30 April 2022.
* Correspondence Author
**Dr. R. Sathish Babu***, Assistant Professor, Department of Computer and Information Science, Annamalai University, Annamalai Nagar, India. Email: sathishbabu.r8243@gmail.com



The main contribution of the work involves using a Genetic Algorithm for the feature extraction of documents and Naïve Bayes for the classification of cancer documents from Medline document datasets. The following is the outline for the paper: The proposed categorization model is shown in section 2. Section 3 assesses the entire project, while section 4 summarises the findings.

## II. PROPOSED METHOD

Figure 1 depicts the suggested method's design. The study first evaluates the Medline datasets and offers training data before allocating some of the newer cases for testing. The architecture first pre-processes the raw documents using a filtering procedure to remove any empty/redundant or null subsets. It then employs GA [6] - [10] feature selection to filter out the features relevant to the cancer datasets and exclude them. Finally, a classifier uses the NB [11-19] classifier to separate the papers into cancer and other groups.

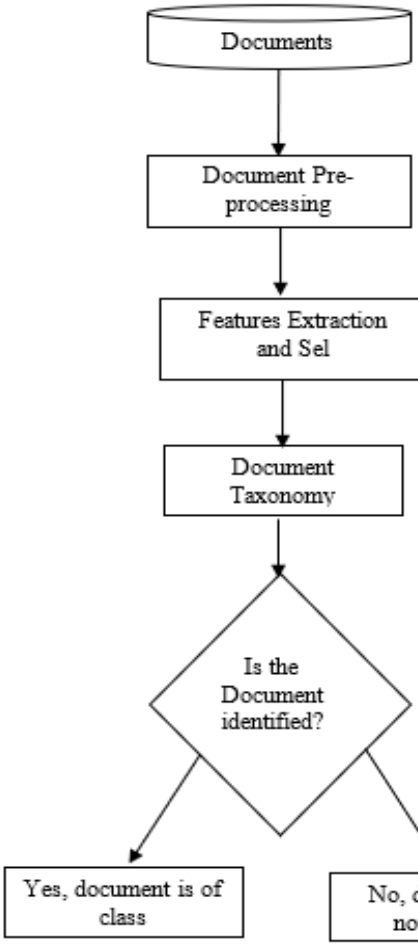


**Fig.1. Proposed Model**

*2.1 GA Algorithm*

GA is a meta-heuristic algorithm that is based on the Darwin theory of natural evolution, where the fittest survival by the method of natural selection. Genetic optimization essentially solves the high-quality solutions to these problems and the search issues depend on genetic operators like selection, crossover and modification.

*Selection*: Selection of two best chromosomes through the selection operator will lead to the reproduction. The best chromosomes for reproduction are therefore selected more often.

*Crossover*: This operator brings together the two-selected chromosome to produce offspring that holds the parent chromosomal features. Alocus in both the chromosome and bits are exchanged to produce two descendants.

*Mutation*: The reduction in local optimum solution for the problem is lowering the transformation process. This operator modifies the newly generated offspring by switching random from 0 to 1 bits. This is where the likelihood of $P(m)$ mutation is set at low (as 0.05).

The phases of GA's algorithm for addressing optimization problems for document classification are as follows:

Step 1. Population or feature set*X*generation with chromosomes *P*
Step 2. Compute the fitness values *f*for entire chromosomes
Step 3. OffspringProduction *X*.
Step 4. Perform the selection and crossover
Step 5. Perform the mutation operation.
Step 6. Find the *rand* value
Step 7. Fix mutation probability $P(m)$ as 0.05.
Step 8. If $rand<P(m)$
Step 9. $X=X+(P(m)*rand)$
Step 10. End
Step 11. Update current population with new ones.
Step 12. Continue till stopping criterion.

*2.2 Naïve Bayes Classifier*

The Naive Bayes Approach (NBA) is a transparent classifier that gathers data in both numerical and textual formats. When compared to various spare allocation systems, it is trivial to actualize and computationally inefficient. One of the most significant barriers to the classifier's performance, when lineaments are tightly coupled, is that it executes incorrectly. As a result, examining the frequency of phrase contingency in the feature vector is a mistake. The Naive Bayes category method's duct disadvantage and similarly low type of entire are equivalent to other discernment procedures NBA has made significant progress in a variety of exploratory projects (Friedman, Sahami, McCallum and Nigam, Mladenic, Craven, Nigam ) was used as a yardstick in others (Twycross and Cayzer 2002), (Yang et al. 2002).

$$c_{NB} = arg\max_{c_i \in C} P(C_i) \prod_i P\left(\frac{f_i}{c_i}\right) for\ all\ c_i \in C \quad (1)$$

Using an NBA, the operation starts by creating a probability distribution based on the set of guidance papers. This distribution is called a prior distribution. When a recent test testimony is taken into account, equation (1) is used to divide it, and the previous dispensation is remodeled into a posterior distribution. According to supposition, NBA has the least error among all the alternative classifiers. Nonetheless, according to a convention, it is impossible to consistently estimate appearance and class conditional independence. NB classifies the features referring to Bayes theorem with naïve independence assumption [15]. The NB classifier is accurate and speedy in classifying the features. The formula of NB is given as follows, where it is defined as a simplification model for classifying the documents.

$$P(V_i) = \frac{P(A|B)P(B)}{P(A)}$$

The Bayes formula is thus given as:

$$V_{map} = \arg_s \max \frac{P\left\langle V_j \middle| a_1, a_2, ..., \middle| V_j \right\rangle P(V_j)}{P(a_1, a_2, ..., a_n)}$$
$$v_j \in V$$

The value of $P(a_1,a_2,\ldots..,a_n)$ over individual $V_j$ is considered similar, where the value is omittedsuch that the Eq.(2) is modified as follows:

$$V_{map} = \arg_s \max_{v_j \in V} \left(V_j \middle| a_1, a_2, ..., a_n\right) P(V_j)$$

In each word of documents $<a_1,a_2,a_n>$ is considered independent, where $(V_{j|a1,a2,an})$ $P(V_j)$ in following equation is represented as.

$$\left(V_j \middle| a_1, a_2, ..., a_n\right) P(V_j) = \prod P\left(a_i \middle| V_j\right)$$

Therefore theequation is defined as:

$$V_{map} = \arg_s \max_{v_j \in V} \prod P(V_j) P\left(a_i \middle| V_j\right)$$

In themeantime, the $P(v_j)$ is found w.r.t. training data, where the value is given as:

$$P(V_j) = \frac{Doc}{|D|}$$

$$P\left(W_k \middle| V_j\right) = \frac{N_{K+1}}{n + |Vocabulary|}$$

## III. RESULTS AND DISCUSSIONS

The study uses the proposed classification method to retrieve documents from a vast online database in this section. In addition, the researchers used cancer-related documents from the Medline database to test the system's performance. The study's success is measured using a variety of measures, including accuracy, sensitivity, specificity, F-measure, and geometric mean.

**Table.1. Clustering Accuracy**

| Documents | GA | NB | GA-NB |
|---|---|---|---|
| 100 documents | 0.812 | 0.827 | 0.857 |
| 500 documents | 0.751 | 0.764 | 0.792 |

**Table.2. Sensitivity**

| Documents | GA | NB | GA-NB |
|---|---|---|---|
| 100 documents | 0.801 | 0.812 | 0.849 |
| 500 documents | 0.735 | 0.752 | 0.781 |

**Table.3. Specificity**

| Documents | GA | NB | GA-NB |
|---|---|---|---|
| 100 documents | 0.792 | 0.804 | 0.817 |
| 500 documents | 0.741 | 0.742 | 0.774 |

**Table.4. F-measure**

| Documents | GA | NB | GA-NB |
|---|---|---|---|
| 100 documents | 0.782 | 0.792 | 0.801 |
| 500 documents | 0.741 | 0.756 | 0.762 |

Table 1 shows the classification accuracy over Medline datasets between the proposed GA-NB and other existing methods. The results of classification accuracy shows that the proposed method achieves higher rate of accuracy on all document classification than existing methods.

Table 2 shows the sensitivity over Medline datasets between the proposed GA-NB and other existing methods. The results of sensitivity shows higher rate of sensitivity on all document classification.

Table 3 shows the specificity over Medline datasets between the proposed GA-NB and other existing methods. The results of specificity on Medline datasets shows higher specificity on all document classification.

Table 4 shows the F-measure over Medline datasets between the proposed GA-NB and other existing methods. The results of F-measure shows high f-measure value on all document classification.

## IV. CONCLUSIONS

The paper develops an integrated classification model with pre-processing, feature extractor and classifier for document classification. The feature extraction using genetic algorithm and classification purpose for Naïve Bayes improves the classification of Medline document datasets. The simulation shows that the proposed GANB classification model obtains higher accuracy, sensitivity, specificity and f-measure. To arrive at a universal framework for document organisation and classification, a study plan has been presented. Taking into account all of the preceding categories, the framework will employ various contextual characteristics while taking domain expertise into account to determine which features or combinations of features are most appropriate. Using classification accuracy, precision, sensitivity, specificity analysis, and datasets from the machine learning repository, the proposed GANB methods are compared to those of existing classifiers. The experimental results show that the proposed methods have produced impressive results in classifying real-life challenging multi-class problems.

## REFERENCES

1. Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, *24*(1), 198-208.
2. Jia, W., Zhao, D., & Ding, L. (2016). An optimized RBF neural network algorithm based on partial least squares and genetic algorithm for classification of small sample. *Applied Soft Computing*, *48*, 373-384.
3. Szenkovits, A., Meszlényi, R., Buza, K., Gaskó, N., Lung, R. I., & Suciu, M. (2018). Feature selection with a genetic algorithm for classification of brain imaging data. In *Advances in feature selection for data and pattern recognition* (pp. 185-202). Springer, Cham.
4. Qiao, Z., Zhang, Q., Dong, Y., & Yang, J. J. (2017, October). Application of SVM based on genetic algorithm in classification of cataract fundus images. In *2017 IEEE International Conference on Imaging Systems and Techniques (IST)* (pp. 1-5). IEEE.
5. Li, H., Yuan, D., Ma, X., Cui, D., & Cao, L. (2017). Genetic algorithm for the optimization of features and neural networks in ECG signals classification. *Scientific reports*, *7*, 41011.
6. Kaabi, H., Jabeur, K., & Ladhari, T. (2018). A genetic algorithm-based classification approach for multicriteria ABC analysis. *International Journal of Information Technology & Decision Making*, *17*(06), 1805-1837.
7. Beheshti, I., Demirel, H., Matsuda, H., & Alzheimer's Disease Neuroimaging Initiative. (2017). Classification of Alzheimer's disease and prediction of mild cognitive impairment-to-Alzheimer's conversion from structural magnetic resource imaging using feature ranking and a genetic algorithm. *Computers in biology and medicine*, *83*, 109-119.
8. Ibrahim, K. A. (2018). Feature Selection based on Genetic Algorithm for Classification of Mammogram Using K-means, k-NN and Euclidean Distance. *Tikrit Journal of Pure Science*, *22*(9), 106-112.
9. Rahmadani, S., Dongoran, A., & Zarlis, M. (2018, March). Comparison of Naive Bayes and Decision Tree on Feature Selection Using Genetic Algorithm for Classification Problem. In *Journal of Physics: Conference Series* (Vol. 978, No. 1, p. 012087). IOP Publishing.
10. Motieghader, H., Najafi, A., Sadeghi, B., & Masoudi-Nejad, A. (2017). A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. *Informatics in Medicine Unlocked*, *9*, 246-254.
11. David, S., Ranjithkumar, K., Rao, S., Baradwaj, S., & Sudhakar, D. (2018). Classification of Massive Data Streams Using Naïve Bayes. *IAETSD Journal for Advanced Research in Applied Sciences*, *5*(4), 208-215.
12. Saritas, M. M., & Yasar, A. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88-91.
13. Slamet, C., Andrian, R., Maylawati, D. S. A., Darmalaksana, W., & Ramdhani, M. A. (2018, January). Web scraping and Naïve Bayes classification for job search engine. In *IOP Conference Series: Materials Science and Engineering* (Vol. 288, No. 1, p. 012038). IOP Publishing.
14. Akella, J. O., & Akella, L. Y. (2018, November). Sentiment Analysis Using Naïve Bayes Algorithm: With Case Study. In *2018 3rd International Conference on Inventive Computation Technologies (ICICT)* (pp. 1-4). IEEE.
15. Pranav, A., & Chauhan, S. (2015). Efficient focused web crawling approach for search engine. *International Journal of Computer Science and Mobile Computing*, *4*(5), 545-551.
16. Yuvaraj, N., & Vivekanandan, P. (2013, February). An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data. In 2013 International Conference on Information Communication and Embedded Systems (Icices) (pp. 761-768). IEEE.
17. Sivaram, M., Mohammed, A. S., Yuvaraj, D., Porkodi, V., Manikandan, V., & Yuvaraj, N. (2019, February). Advanced Expert System Using Particle Swarm Optimization Based Adaptive Network Based Fuzzy Inference System to Diagnose the Physical Constitution of Human Body. In International Conference on Emerging Technologies in Computer Engineering (pp. 349-362). Springer, Singapore

18. Yuvaraj, N., Raja, R., & Dhas, C. (2018). Analysis on Improving the Response Time with PIDSARSA-RAL in ClowdFlows Mining Platform. EAI Endorsed Transactions on Energy Web, 5(20).
19. Natarajan, Yuvaraj and Raja, R.Arshath and Kousik, Dr. N.V. and Johri, Prashant, Improved Energy Efficient Wireless Sensor Networks Using Multicast Particle Swarm Optimization (February 21, 2020). Available at SSRN: https://ssrn.com/abstract=3555764

## AUTHORS PROFILE

**Dr. R. Sathish Babu,** is currently an Assistant Professor at the Department of Computer Information Science, Annamalai University. He completed his undergraduate degree in 1999 at the university of madras, his post-graduate degree in M.C.A. at bharathidasan university, and his master of philosophy in distinction at bharathidasan university. He obtained his Ph.D. in computer application at Annamalai University. He has 20 years of experience in teaching and is involved in research activities. His specialization areas include Data Mining, Software Engineering, Big Data Analysis, and Software testing. He develops the online public access catalog for the library and many software for acquisition and reports for management. He guided many students for M.Phil., M.C.A and M.Sc Computer Science. He has published 4 research papers in International Journals.