OPEN ACCESS
Saudi Journal of Engineering and Technology
Abbreviated Key Title: Saudi J Eng Technol
ISSN 2415-6272 (Print) |ISSN 2415-6264 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com/sjeat

Review Article

Comparative Study of Some Supervised Machine Learning Algorithms for Information Retrieval

Kissinger Sunday*, Muhammad Bello Aliyu

Computer Science Department, Usmanu Danfodiyo University, Sokoto State, Nigeria

DOI: 10.36348/sjet.2020.v05i03.003 | **Received:** 04.03.2020 | **Accepted:** 13.03.2020 | **Published:** 30.03.2020

***Corresponding author:** Kissinger Sunday

Abstract

The volume and quality of online data has increased tremendously. Retrieval of such data relies so much on efficient methods. In recent times, information retrieval looks to the intelligence-based and inductive learning methods, such as genetic algorithm, neural networks and machine learning. Researchers however, have leverage on these newer techniques in order to enhance the retrieval capabilities and information processing of current information storage and retrieval systems. These methods provide various degrees of accuracy. But how effective are these methods and which of them is better suited for the information retrieval task? This paper investigates the efficiency of the selected algorithms: Artificial Neural Network, Support vector machine, and Genetic Algorithm, on designing the model for efficient and intelligent information retrieval. The selected algorithms were critically studied in line with the available matching models for information retrieval. Models like the Vector space model, Binary model, probabilistic models, Inverted Index, Latent semantic Analysis and the Latent Semantic Index models were respectively examined. The result from the experimentation from the respective algorithms shows that the neural network, in combination with Genetic algorithm or alone, performs better. However, it takes more time to execute.

Keywords: Information Retrieval, Machine Learning, Genetic Algorithm, Support vector machines.

INTRODUCTION

The quantity of data has been increasing these days as a result of the paradigm of Internet of things [1]. In the past three decade, there has been significant improvement in the development of information retrieval systems beyond its main goals of searching for important documents in a collection and text indexing. At present, information retrieval research spans through languages, systems architecture, filtering, document classification, modeling and categorization, user interfaces, data visualization, etc [2].

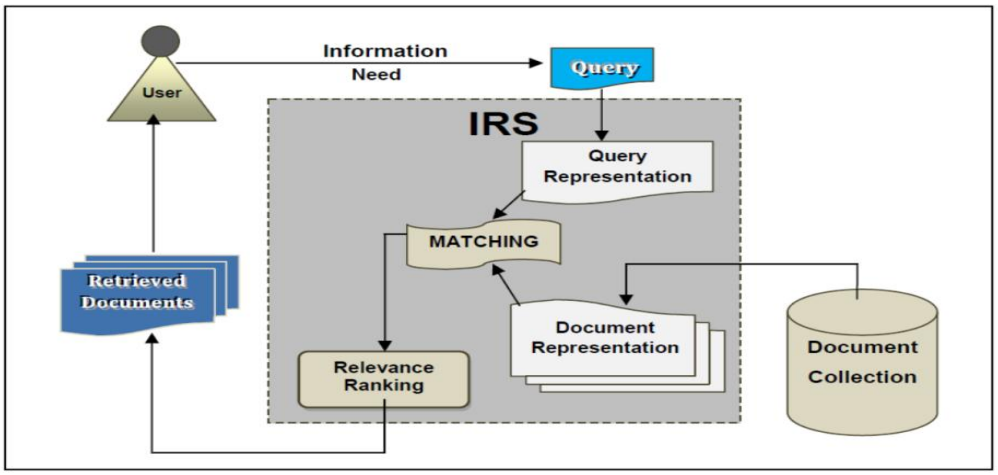

Fig-1: Information Retrieval Process [3]

Machine learning is the ability of machines to learn despite not been precisely and clearly programmed to learn [4]. There are many machine learning methods but all are classified under one (1) of four (4) categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning methods respectively. The objectives of machine learning classifiers entails enabling machines to perform predictions, building of clusters, extraction of association rules, or coming out with a decision from a known dataset.

The existing key word search suffers from issues such as the negligence of the conceptual meaning of documents when retrieving data. Typically, when a user provides a query, the keywords from the query are extracted and then, the documents are subjected to a mining process for these keywords. A document is retrieved and deemed to be relevant to the query if only it possesses the keywords otherwise it is discarded. This mining technique is however not efficient as the conceptual meaning of documents is not taken into consideration. Direct evidence about the meaning of the document is not provided by individual words. Aside this, multiple words that may bear the same meaning is also neglected using this technique; making it dependent solely on documents containing the exact key words. Semantically relevant documents are discarded simply because they do not contain the key words.

Latent semantic indexing however has been widely employed to solve the above problem. In this method, patterns in the relationship between the terms used and the conveyed meaning is exposed using singular value decomposition. It follows a simple principle that words that have analogous meanings are used in the same context.

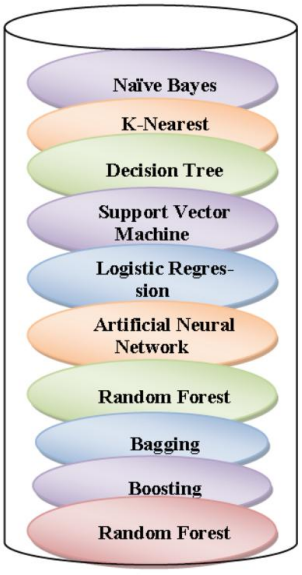

Fig-2: Machine Learning models used in Information Retrieval [5]

We tend to show the relevance of supervised machine learning algorithms for information retrieval in this study. We limited our ML algorithms to Artificial Neural Networks (ANN) and Support Vector Machine. This is because they are the most widely used ML algorithm in Information Retrieval. Furthermore we will look at a popular optimization approach for the ML algorithms such as Genetic Algorithm. Since the success of ML algorithms is determined by the type of Matching Function used; we will review the current model used in IR. We concluded by implementing the algorithms and giving our recommendation.

LITERATURE REVIEW

A survey on information retrieval and genetic algorithm was done by [6]. Their survey primarily focused on how genetic algorithm could be applied to the different areas of information retrieval. In order to boost information retrieval efficiency [7], carried out an analysis using the vector space model in which they discovered that documents which are high in similarity to query are been judged to be more relevant to a query and thus, will be the first to be retrieved. Their result shows that maximum precision was gotten by having 0.8 crossover and 0.01 mutation probability respectively, while 0.3 mutation and 0.8 crossover probability provided the maximum recall. Analysis of the various problems of current search engines was undertaking by [8]. In their work they suggest a new design with novel ideas in order to enhance the current state of web search engines. In order to enhance the proficiency of search engines; they advocated a mobile agents with an adaptive methods. A research carried out by [9] demonstrated that employing conventional approaches instead of heuristic search techniques leads to poor result in information retrieval systems. To further buttress their point, they designed and implemented a genetic algorithm then a hybrid genetic algorithm for information retrieval. The result of their research shows that both the designed algorithms outperform the classical approach for large data sets and the hybrid genetic algorithm yields the best performance in terms of solution quality and runtime. A research by [10] developed a novel approach for deep learning techniques and went ahead to represent the information retrieval techniques and models. Different information retrieval search techniques and indexing methods were also described. Finally they implemented their system using Hadoop (an open source framework for supporting distributed applications) for efficient retrieval of information. An incorporation of information extraction (IE), semantic technology (ST), natural language processing (NLP) are been used in [11] to bring to the fore a novel method for knowledge extraction from research documents. Keywords are extracted from the documents after it has been pre-processed using regular expression. In their work two triple-stores on sentences and words was used based on three type formats: Subject, Predicate and Object to extract useful information from the documents. After

the data must have been processed, they are then applied on triple-store data to extract knowledge. A support vector machine based classification document with a concept vector model was proposed by [12] In their research a novel approach for categorizing document based on support vector machine with a concept vector model was proposed in order to solve the drawbacks of classifying document using traditional approach which does not take into account the semantic relations among the documents/keywords. The results show that traditional term-based vector space model yields lower accuracy compared to using the concept vector model.

A study by [13] proposed an adaptive genetic algorithm to improve the operation of Information Retrieval System. The data set used for information retrieval emanated form IEEE abstract papers of 2010-2015. Relevance feedback is further optimized by utilizing various operators such as mutation and crossover coupled with variable probabilities. However, in the traditional GA which uses fixed values of those, during execution the values still remain unchanged. The simulated results from the algorithm show that its performance is superior to the existing algorithm by using it in Distributed Environment. A study by [2] investigated new Machine Learning techniques of Information Retrieval such as ant colony algorithm, artificial neural network, genetic algorithm, and differential evaluation. They believe the application of soft computing can make an information retrieval system more potent. A model to rank document which comprises two separate deep neural networks was proposed by [14]. This model uses both local representation and learned distributed representation to march both query and document. These networks however, were collectively trained as part of a single neural network. According to their hypothesis, marching with traditional local representations could complement matching with distributed representations hence combining the two approaches is greatly desirable. Their results show that the combination greatly outperforms traditional baselines and other models that have been proposed recently on neural networks and also performs better in terms of ranking a web page on individual neural network.

MACHINE LEARNING

Machine learning is a branch of artificial intelligence that entails the programming of systems in order to facilitate automatic learning from data [15]. This learning process tends to be better with experience. In this case, learning can mean the understanding and recognition of the input data and be able to take favourable decisions based on the data supplied.

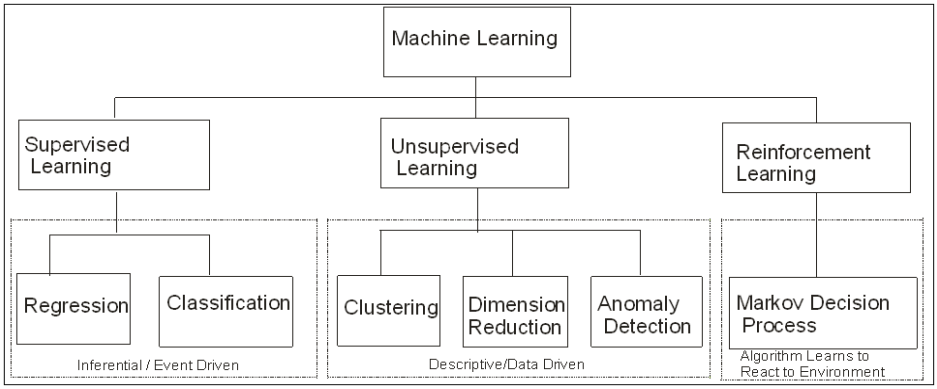

Fig-3: Types of Machine Learning [15]

One of the goals of machine leaning is to come up with an algorithm that is capable of replacing or supplementing experts involved in knowledge engineering. Engaging the power of learning algorithms to automate the processes of information retrieval such as user modeling and document classification can serve as palliatives to the workload of information and the inconsistency introduced by human error will be considerable reduced.

Supervised Machine Learning

The main objective of supervised learning is to ensure that input to output data are mapped using rules with labels, description, targets or desired outputs attached to the learning data. This learning data is known as a labeled data and is then used to label new data with output that is unknown. For example when building a system for image classification containing say a school, a pet, a ship or even a person; what we first do is to collect data set of school, pet, ship and person each with its own label. When the training starts, an image is shown to the machine and then the machine produces an output in the form of a vector of scores. An objective function that measures the error between the desired pattern of scores and the output scores is computed. This error however is reduced by the modification of the internal adjustable parameters by the machine. The input-output function of the machine is defined by the adjustable parameters (weights) which are real numbers that can be seen as 'knobs'. A deep learning system typically has more of these adjustable weights and labelled examples say hundreds of millions used in training the machine.

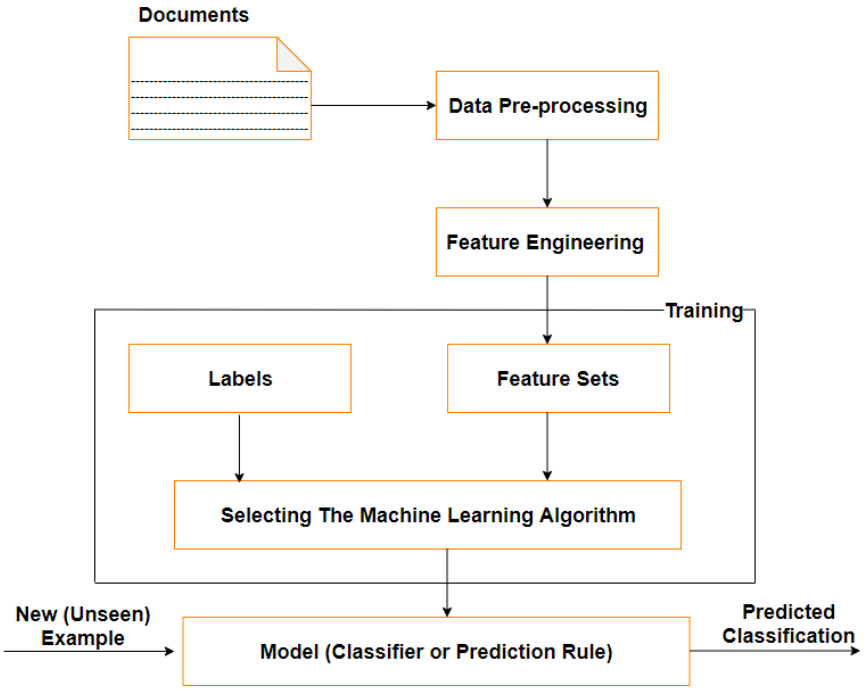

Fig-4: Processes of Supervised Machine Learning [16]

We have two categories of supervised learning algorithm; these are Classification and Regression. Supervised learning task include the categorization of e-mails into two (spam and not-spam), recognition of voice, labeling WebPages based on their content amongst a host of others.

Artificial Neural Network (ANN)

This network model deals with the connection of nodes otherwise known as neurons together in order to form network of nodes. This model is mostly referred to as a neural network since it interconnects artificial neurons together. It uses a connectionist approach to process information based on a model of computation. An ANN that is adaptive tends to change its structure as a result of the information that flows through the network.

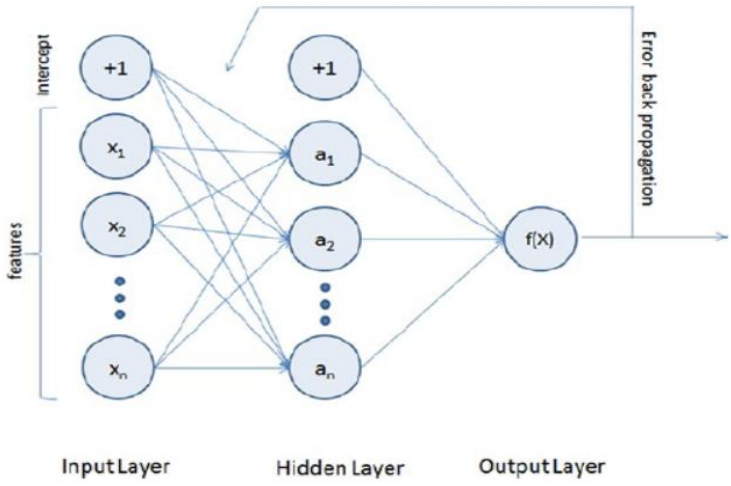

Fig-5: Artificial Neural Network [15]

Support Vector Machine (SVM)

SVM is one of the Supervised Machine Learning Classifiers that is commonly used in information extraction from unstructured text. SVM has proved to be efficient and effective for a range of diverse classification task such as Information Extraction (IE) [17]. The most valued parameters of the SVM implementation include the cost and the uneven margins. Variables that are both continues and dependent can easily be predicted using the concept of SVM. In SVM, the objective here is to construct a hyperplane that divides the two classes optimally such

that the margin is maximal between the hyperplane and the observations. The possibility of having different hyperplanes is illustrated in figure 5 below. The objective of SVM however is to locate the one which is capable of producing a high margin.

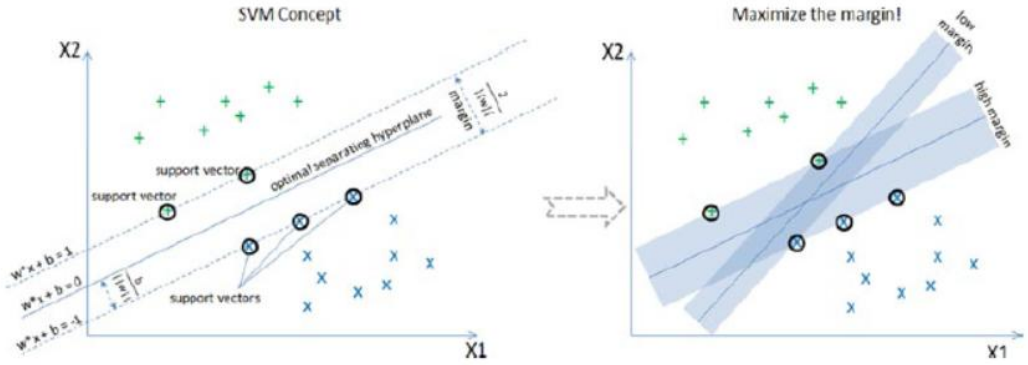

Fig-6: Support Vector Machine [15]

Review of Retrieval Models

The model of an Information Retrieval System (IRS) is characterised by defining both the query representation and the document characterises of the model [3]. In this research, we analyze the suitability of IR models and their various drawbacks.

The Boolean model

This is the widely used model in commercial IR system which is based on the theory of mathematics in which index terms are used to represent document. A query is composed of logical connectives and index terms. When a document satisfies the logical formula representing the query it is automatically retrieved as been relevant by the IR system In the Boolean model, there is no indication on the terms that are more valuable than others (weights are either 1 or 0) [18]

Advantages

1. It has a clean and clear formalization technique.
2. It is not difficult to implement.

Disadvantages

1. Since it works on exact marching of text, too many or very few documents may be retrieved.
2. The ranking of output is not easy.
3. There is great difficulty in controlling the number of documents that will be retrieved.

The Vector Space model

In this model, each index term is associated with an index value known as the index term weight which expresses the significance that the term has in building the content of the document based on spatial interpretation of both the queries and the document. This significantly improves query results as compared to the Boolean model.

Evaluation of the document's relevance to the query is achieved through the similarity measures between the query and document's representation.

Advantages

1. The association of an index value on each index term has been known to significantly improve the retrieval performance.
2. Partial matching is enabled in this model.
3. The degree of similarity between documents and query enables the sorting of retrieved documents.

Disadvantage

1. In some cases the performance of this model is affected as a result of the assumption placed on terms that are mutually independent.

The Probabilistic model

In this model, documents are ranked based on the decreasing order of their probability of been evaluated in relevance to the information need of the user. Much use of formal theories of statistics & probability in order to at least estimate or evaluate the probability of relevance has been put forward by past and present research.

Advantages

1. The recursive and iterative nature of the algorithm enables the initial guess of the parameters and hence it then tries to significantly improve the initial guess in order to obtain the final ranking of the relevant probabilities.

Disadvantages

1. The programming and building of probabilistic models can sometimes be very hard and complex.
2. The simplifying assumptions of this model such as the independence between documents and terms is however unrealistic. Probabilistic models also require several unrealistic simplifying assumptions, such as independence between terms as well as between documents.

Inverted Index

In this model which is also a data structure is composed of posting list which is associated with each

term that appears on the collection, enabling it to quickly find the location of all appearances of that query term in the collection. Retrieval engines make heavy use of inverted index for retrieving documents in our today's world. Objects to be retrieved are generally termed "documents" even though they may be pdf, webpages, or even code fragments. The inverted index containing the query terms are scored with respect to some models used for ranking; this is however done by the query engine when given a query by the user. The engine also takes into consideration features like term proximity, attributes of the terms, hyperlink structure as well as term matches.

Advantages

1. Inverted index allows speedy search of documents.
2. Developing this model is not difficult.
3. It is the most widely used data structure in document retrieval systems, especially in search engines.

Disadvantages

1. The high cost of maintenance coupled with a large storage overhead makes it less attractive to use.

Latent Semantic Index (LSI)

This model involves the projection of documents and queries into a latent semantic space with fewer dimensions as compared to the original space. Hence LSI technique seeks to reduce this dimension by representing existing objects in high – dimensional space into a low dimensional space for the purpose of visualization.

It uses a mathematical method known as single value decomposition to a word by document matrix.

Merits

1. The documents and queries are well represented in the new dimensions.
2. The LSI model is good at retrieving documents that employ a different vocabulary on topics that tend to be the same.
3. Words that possess multiple meaning (polysemy) are easily handled by this model. The effect of having polysemous words is that the precision of a search is reduced significantly thus hampering the retrieval process.

Demerits

1. The compact nature of the SVD representation makes it generally unsuitable for some task.
2. The drawback of the LSI model is that, to efficiently retrieve relevant document from the representation, every query has to be compared with the document in the collection. This however, goes a long way in slowing down the rate at which relevant documents are retrieved.
3. The design of the LSI model takes into consideration normally distributed data without considering count data. This however, has great setback because a term by document matrix is composed of count data.

Experiment

As depicted in Fig-4, the first task is to obtain the data for use in the experiment. The rest of the processes depend on the data.

Data

The dataset for the experiment involves a Coventry university dataset. The dataset was gathered been a part of the research work on software engineering subjects from Coventry University, UK. The dataset took into consideration a broad view of what constitute software engineering subjects including the design, development and innovation of software. The dataset is a collection of retrieved data using the technique reported in [19]. The dataset is labelled with two (2) clearly marked levels: positive (relevant) and negative (irrelevant). The subjects include: Natural language processing, Machine learning, software testing and software cost estimation. The data is collected from a scholarly search engines. We used the binary search strings defined in [19]. The collection consists of 2,400 documents (data instances). Relevant judgements on documents that are related to all queries have been made by human judges. In total, there are 24,000 query-document pairs (instances).

Data Pre-processing

The data is mainly text hence, it requires pre-processing to make the data suitable for the machine learning experiment. In our experiment, we cleaned and transformed the data. The following pre-processing activities were performed on the data: tokenization, removing unnecessary tags, removing stop words, stemming and lemmatization.

Feature Engineering

The query and document is used to determine each instance which comprises vector of features. The standard features that are often used in retrieving document were used in this research [20]. These features are: the length of the document (DL), inverse document frequency (IDF), term frequency (TF) and their various combinations. We reduced the effects of large numbers by taking log on the feature values. However, based on preliminary experiments, the results tend to be unchanged. We removed the stop words and we conducted stemming in retrieval and indexing. We went further to generate feature vectors from the query-document pairs. The data was then split into two sets: training set and test set with the ratio of 80:20. The test set had 480 while the training set had 1,920 queries

Training

In the information retrieval task, the main point is to rank the retrieved information (documents) using the selected algorithms. Given the query, *Q*, and

the document *D*, the training involves the computation of relevance score for the documents. Some works such as [21] used three (3) tier classification for the documents: relevant, partially relevant and irrelevant. In our work however, we used two (2) tier classification, indicated by positive (+1) for the relevant documents and negative (-1) for irrelevant documents. This arrangement conforms to the SVM classification architecture where a separation hyperplane divides the classes into 2.

Training data = *{R, X}*

Where R is the relevance score, and X is the feature representation of (D,Q) pairs, given by

R = {-1, +1} ... is D relevant to Q or not.
X = feature representation of (D,Q) pairs
The optimal combination of the features

The score for each of the document is the weighted combination of features given by

$\sum_i WiXij$ ..... equ. 2

During ranking, an instance is created from a query-document pair when SVM is applied to document retrieval. We define each feature as a function of document and query. For example the determination of the feature of term frequency is as a result of the number of times the query term appears in the document. We then combine all the instances from the queries in the training. No difference in treatments exists toward the instances from the different queries and between instance pairs from the different rank pairs.

In order to rank SVM for IR, we generate each instance x from a query-document pair and followed by a label with one rank (from the 2 possible ranks defined above).

RESULTS & FINDINGS

The SVM has achieved an accuracy of 79%. This means about 4 out of every 5 of the relevant documents were identified by the SVM. In comparison to the Genetic Algorithm, the following fig shows the accuracies difference.

The accuracy of a model is the fraction of the predictions which the algorithms got right. For binary classification, we calculate the accuracy in terms of positive and negative predictions as depicted in the equation below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$ ..... equ. 3.

Where:

TP = number of the correctly identified positive instance (True Positive).
FP = number of the wrongly identified positive instances (False Positive)
FN = number of the wrongly identified negative instances (False Negative)
TN = number of the correctly identified negative instances (True Negative)

After the training, the model was validated with the validation set. The table 1.0 below shows the result in comparison with the rest of the methods.

Table-1: Comparing Neural Network, Support Vector Machine & Genetic Algorithm (we use the * symbol to represent the worst and ** symbol to represent the best)**

	Neural Network (NN)	**NN+GA**	**Support Vector Machine (SVM)**	**SVM+GA**
Accuracy in general (%)	72	77	69	71
The learning speed considering the number of instances and the attributes	**	*	*****	***
classification speed	**	*	*****	***
Tolerance to values that are missing	***	**	****	***
Tolerance to attributes that are irrelevant.	**			
Tolerance to attributes that are redundant	****	***	***	*
Tolerance to attributes that are highly interdependent.	***	*	****	**
How it deals with discrete/binary/continuous attributes	****	***	*****	***
Tolerance to noise				
Dealing with danger of overfitting	**	*	**	**
Attempts for learning incrementally	*****	***	*****	***
How transparent is the knowledge?	**	**	***	**
How model parameters are handled?	***	**	***	****

CONCLUSION

Estimating the accuracy of the candidate algorithms on the problem is one of the simplest approaches to undertake when faced with the decision of choosing the most accurate algorithm on the classification problem. In our case, the neural network performs better than the other algorithms. However, it takes more time to execute.

The idea of using multiple classifiers is proposed as a novel area for the improvement of the performance of individual classifiers. The goal of integrating algorithms is to generate precise and accurate system results. All learning algorithms tend to have many different parameters and configurations to be adjusted in order to achieve optimal performance on a dataset. This feature will no doubt make it suitable to some problem types better than others. In this work, we combined the neural network and genetic algorithm as well as support vector machines and the genetic algorithm. The result of that combination is an improved result in both accuracy and other factors. However, there is increase in time needed to execute the classification as well.

REFERENCES

1. Farhan, L., Kharel, R., Kaiwartya, O., Hammoudeh, M., & Adebisi, B. (2018). Towards green computing for Internet of things: Energy oriented path and message scheduling approach. *Sustainable Cities and Society*, *38*, 195-204.
2. Kausar, M. A., Nasar, M., & Singh, S. K (2013). Information Retrieval using Soft Computing: An Overview. *International Journal of Scientific & Engineering Research*, 388-395.
3. Crestani, F., & Pasi, G. (1999). Soft Information Retrieval: Applications of Fuzzy set theory and Neural Network. *Semantic Scholar*.
4. Mohammad, M., Khan, M. B., & Bashier, E. B. (2017). *Machine Learning Algorithm and Application.* New York: CRC Press Taylor and Francis Group.
5. Sunday, K., Ocheja, P., Hussain, S., Oyelere, S. S, Balogun, O. S., & Agbo, F. J. (2020). Analyzing Student Performance in Programming Education using Classification Techniques. *International Journal of Emerging Technologies in Learning (IJET*), 15(2).
6. Kausar, M. A., Nasar, M., & Singh, S. K. A. (2013). Detailed Study on Information Retrieval using Genetic Algorithm. *Journal of Industrial and Intelligent Information*.
7. Klabbankoh, B., & Pinngern, O. (1999). Applied genetic algorithms in information retrieval. *IJCIM*.
8. Koorangi, M., & Zamanifar, K. A. (2007). distributed agent based Web search using a genetic algorithm. *International Journal of Computer Science and Network Security*, 65-76.
9. Drias, H., Khennak, I., & Boukhedra, A. A. (2009). Hybrid Genetic Algorithm for Large Scale Information Retrieval . *IEEE*.
10. S. Kanimozhi, A., & Devi, A. B. A. (2018). Novel Approach for Deep Learning Techniques Using Information Retrieval from Big Data. *International Journal of Pure and Applied Mathematics*, 601-606.
11. Upadhyay, R., & Fujii, A. (2016). Semantic Knowledge Extraction from Research Documents. *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, 439-445.
12. Deng, S., & Peng, H. (2006). Document Classification Based on Support Vector Machine Using A Concept Vector Model. *Proceeding of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.*
13. Mitkal, P., & Gore, D. (2016). Improving the Performance of Information Retrieval System using AGA in Distributed Environment. *International Journal of Innovative Research in Computer and Communication Engineering*.
14. Bhaskar, M., Fernando, D., & Nick, C. (2017). Learning to Match using Local and Distributed Representation of Text for Web Search. *International World Wide Web Conference Committee.* Austrialia: ACM.
15. Swamynathan, M., & Karnataka, B. (2017). Mastering Machine Learning with Python in Six Steps; A Practical Implementation Guide to Predictive Data Analytics Using Python.
16. Mohammad, B. A., Iqbal, R., James A., & Nkantah, D. (2019). Convolutional Neural Network for Core Sections Identification in Scientific Research Publication. In International Conference on Intelligent Data Engineering and Automated Learning; Springer, Cham, 265-273.
17. Aljamel, A., Osman, T., Acampora, G., Vitiello, A., & Zhang, Z. (2018). Smart Information Retrieval: Domain Knowledge Centric Optimization Approach. *IEEE Xplore*, 4167-4183.
18. Habibi A. L., Mahdavi F., & Ghomi V. A. (2009). Boolean Model in Information Retrieval for Search Engines. *In International Conference on Information Management and Engineering*, 101:385-389.
19. Muhammad, B. A. (2017). Efficiency of Boolean Search strings for Information Retrieval. *American Journal of Engineering Research (AJER)*,6(11): 216-222.
20. Nallapati, R. (2004). Discriminative models for information retrieval. Proceedings of the 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval; 64-71.
21. Cao, Y., Xu, J., Liu, T. Y., Li, H., Huang, Y., & Hon, H. W. (2006). Adapting Ranking SVM to Document Retrieval. In Proceedings of the 29th annual International ACM SIGIR conference on Research and development in Information Retrieval; 186-193.