International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Big Data Analytics using Swarm Intelligence based Framework for Prediction on Datasets

C.Kalpana, B. Booba

Abstract— Data Analytics is a scientific as well as an engineering tool used to investigate the raw data to revamp the information to achieve knowledge. This is normally connected with obtaining knowledge from reliable information source and rapidity in information processing, and future prediction of the data analysis. Big Data analytics is strongly evolving with different features of volume, velocity and Vectors. Most of the organizations are now concentrating on analyzing information or raw data that are fascinated in deploying analytics to survive forthcoming issues and challenges. The prediction model or intelligent model is proposed in this research to apply machine learning algorithms in the data set. Then it is interpreted and to analyze the better forecast value of the study. The major objective of this research work is to find the optimum prediction from the medical data set using the machine learning techniques.

I. INTRODUCTION

Data Analytics is an engineering tool used to explore the raw data to revamp the information to achieve knowledge through the Computing methodologies. The complex decisions are formulated by collaborating the data with data analytics in order to face real world difficulties [1]. Real time varying and streaming data are analyzed by the process of data analytics and it is termed as Big Data Analytics. In big data, there is an exponential growth of data and it acts like frontier for business analysis and predicting, competition and innovation [2]. Consumer requirements, market trends, unfound correlations, future recommendations and hidden patterns of huge data are computed by data analytics which may assist the decision making process in critical situation [3].

New opportunities are identified by managing, handling, evaluating and analyzing the data by means of Big Data Analytics. It has a features like, service systems, smart decision making, advanced product development, cost reduction and efficient operations. Big data are represented using an advanced storage structures. This huge big data structures cannot be handled by the traditional systems of database. So, to manage huge data, it is necessary to find a new approach [4].

Manuscript published on November 30, 2019.
* Correspondence Author
C.Kalpana*, Research Scholar, VISTAS, Vel's University, Chennai, Tamil Nadu, India.
Dr. B. Booba, Professor, Dept. Of CSE VISTAS, Vel's University, Chennai,TamilNadu, India. (Email: {rkalpz, boobarajashekar}@gmail.com)

There are two classes of machine learning techniques namely, unsupervised and supervised learning. Accurate prediction of data is resulted by supervised learning and compact description is given by unsupervised learning. A machine learning technique has features like complex pattern analysis, Data analysis and classification, data clustering and data predictions or decisions. Complex datasets are learned by the Machine learning methodologies to make difficult decisions. High performance can be achieved by tuning the special features. In parallel programming framework, trained features defines the performance. From learning model, test features are computed to be exercised on dataset used for training [5-7].

On real time dataset, MapReduce programming model is used to implement the supervised learning techniques. Huge real time dataset is processed by tuning Machine Learning techniques. In order to predict, deep learning, pattern matching and recommendation systems, these Machine Learning are applied.

In this paper, section II discusses about swarm intelligence, section III discusses about firefly algorithm, section IV deals with features selection and reduction, experimental results are presented in section V and section VI concludes the work.

II. SWARM INTELLIGENCE

The numerical problems can be optimized by using swarm intelligent based metaheuristic algorithm which is termed as Artificial Bee Colony algorithm. The honey bees have intelligent behaviour in searching food source in the real world entity. This behaviour is used by the Artificial bee colony (ABC) system for optimizing through any data analysis problem. Social cooperation of honey bees are used to complete the task [11]. There are three categories of bees in ABC system. They are Scout bees, onlooker beers and employed bees. The food in and around the food sources are searched by employed bees and stored the information in his memory. This information are conveyed to the onlooker bees.

The good food sources are selected by the onlooker bees based on the information conveyed by employed bees. onlooker bees selects the source with high quality. Scout bees are taken from employed bees. The food source of the scout bees will abandoned and new food sources will be searched [9]. Swarm has employed bees in the first half and onlooker bees in second half. The number of onlooker or employed bees will communicate to number of solutions [10].

Let ith solution in swarm is represented by $Y_i=\{ y_{i,1}, y_{i,2}.....,y_{i,F}\}$, dimension size is denoted by F. A new candidate solution Z_i is produced by every employed bee Y_i in neighbourhood of its next present position as follows:

$Zi,j = Yi,j + \Phi i,j(yi,j – ym,j)$

Where unsystematically selected candidate solution ($i \neq m$) is given by Y_m, its random dimension index selected from the set {1, 2, 3,….F} , and $\Phi_{i,j}$ is a random number within [-1,1] . A greedy selection is used after the generation of new candidate solution P_i. Update Y_i with P_i ; if the fitness value of Pi is better than that of its parent Y_i , otherwise keep Y_i unchangeable. Shake dances are used to share the information between employed bee and onlooker bee after the completion of search process by employed bees. This probabilistic selection is described as per the equation. The food source will be out of control, if the position cannot be improved within the cycle limit. If Yi is discarded source, a new food source will be identified by then the scout bee that will replace Yi as follows:

$y_{i,j}$ = lower bound $_j$ + R(0,1) (upper bound $_j$ – lower bound $_j$)

Where random number R(0,1)is depends on a normal distribution.

III. FIRE FLY ALGORITHM

Firefly is a Meta heuristic algorithm. Global optimization algorithms, soft computing techniques and computational intelligence has these meta heuristic algorithms as an important part. Multiple nature interacting agents are used by these algorithms. They are called as nature-inspired algorithms. Swarm intelligence based algorithms are the subset of meta heuristics. The swarm intelligence characteristics of biological agents are mimicked to develop SI- based algorithms. The biological agents may include birds, fish, humans and others.

Swarming behaviour of birds and fish is utilized by particle swarm optimization, flashing pattern of tropical fireflies is utilized by firefly algorithm and the brood parasitism of some cuckoo species is utilized by the cuckoo search algorithm. The multimodal and global optimization problems can be dealt efficiently by using these new algorithms. All meta heuristic algorithms requires a balance exploration and exploitation. Search landscape and algorithms optimization are also discussed. The effectiveness of the firefly calculation when contrasted with other irregular inquiry system are shown by discontinuous hunt procedure and numerical examinations. Fireflies are unisex. Fireflies will be pulled in by one another regardless of their sex.

- The attractiveness is corresponding to the brightness. Engaging quality and brightness decline when the separation increments. The firefly with less brilliance will move towards the brighter one. Firefly will move haphazardly, if there is no brighter one than a specific firefly.

- The scene of the target capacity characterizes the brightness of a firefly

IV. FEATURE SELECTION AND REDUCTION

Informative features and gene selection is a challenging task. Best features of the samples are searched by Bacterial Foraging Optimization (BFO) with HSIC to solve this issue. It is an activity shown by bacterial foraging behaviors which is named as “chemotaxis”. The gene and feature selection are done based on this behavior.

The features and genes are selected by rotating the flagella in counter clockwise direction. Due to this the organism “swims” which makes the bacterium to find most relevant cervical cancer risk factors by randomly “tumble” in a new direction and it swim again [8]. Best features in positive directions are found by the bacterium by making an changes between “swim” and “tumble”. To increase the prediction rate of cervical cancer the bacterium swings more frequently.

Bacterium moves from one features to search for more relevant other features or genes when the direction changes. It is called Tumbling. Complex combination of swimming and tumbling makes the Bacterial chemotaxis. Prediction rate of the cervical cancer can be increased by keeping those bacteria in highly concentrated places. The classical BFO system has chemotaxis, elimination-dispersal and reproduction mechanisms.

V. RESULTS AND DISCUSSIONS

In machine learning algorithm, the parameters like Logarithmic loss, Area under curve are used to calculate the performance. The algorithms like SVM and ANN are used for performance comparison. Performance of algorithms are compared in terms of parameters like recall, precision, Accuracy, F-measure. To evaluate the performance of the proposed technique synthetic data is used. Various real-world dataset are used to evaluate performance of feature selection algorithm. The experimentation is done in MATLAB environment and experimental results are discussed in this section. Hospitals in 'Universitario de Caracas' at Caracas are used to collect the dataset of cervical cancer. 32 risk factors are used to represent the dataset. The factors includes patient's habits, historic medical records and demographic information and they are shown in Table.1. Hinselmann, Biopsy, Schiller and Cytology are target variables. Colposcopy using acetic acid is referred to as Hinselmanns test and using iodine are referred as Schillers test, cytology and biopsy. Few patients may not answer some questions due to their privacy. So dataset must be pre-treated in order to deal with values that are missing. Such a data is called imbalanced data. In the pre- treatment process, oversampling is applied. Due to the lack in values available, risk factor 27 and 28 are removed.

Precision

Precision is proportion of the quantity of significant records recovered to the absolute number of immaterial and important records recovered. It is communicated in rate. The understanding between a few judgments of a similar amount is given by exactness. The estimation of division from genuine worth and its disperse is named as accuracy.

Precision or Positive Predictive Accuracy =TP/(TP +FP)

True Positive (TP): Quantity of right expectations that an occurrence is certain.

False Positive (FP): Quantity of wrong expectations that an example is certain.

False Negative (FN): Quantity of wrong expectations that an example is negative.

True Negative (TN): Quantity of right expectations that an occurrence is negative

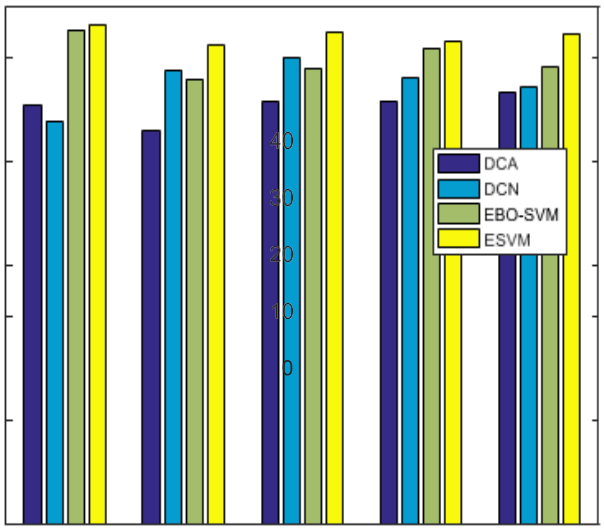

Fig.5.1 Comparison of Precision

The comparison of precision values of SVM, ANN and ANFIS Classification algorithms is shown in Figure 5.1. The ANFIS approach produces high va;ue of precision value when compared to other methods.

RECALL

Recall is proportion of number of pertinent records recovered to all out number of important records in database. It is communicated in rate. Recall is additionally called as the genuine positive rate or affectability. The division of important material returned by the hunt is characterized by a factual measure called recall

Sensitivity or recall =TP/(TP +FN)

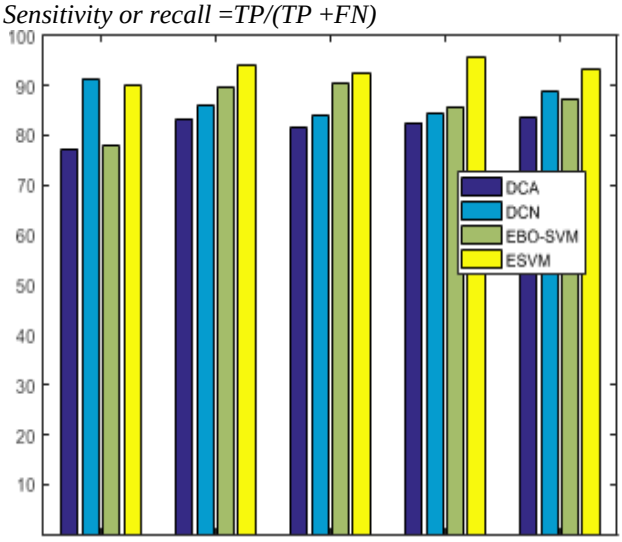

Fig.5.2 Comparison of Recall

The comparisons of recall values of ANN, SVM and ANFIS Classification algorithms is shown in Figure 5.2. The ANFIS approach produces high value of recall value when compared to other methods. The precision and recall value are used to compute the returned-measure

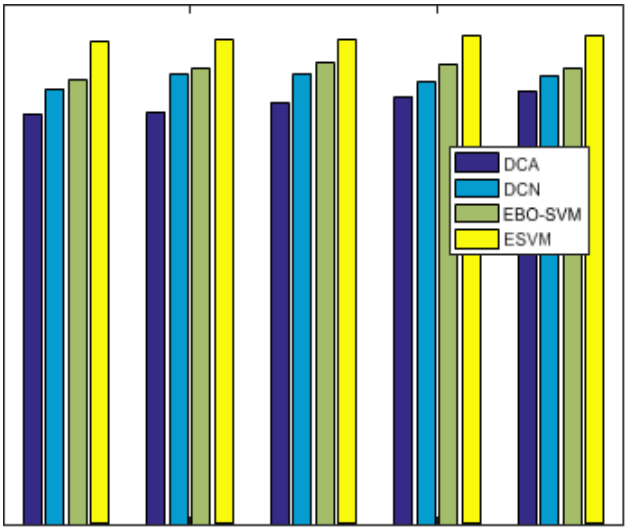

Fig.5.3 Comparison of F-measure

The comparison of F-Measure of SVM, ANN and ANFIS Classification algorithms is shown in Figure 5.3. The ANFIS produces high value of F- Measure when compared to other algorithm as shown by the results.

ACCURACY

The relationship between experimental value to theoretical (—true) value of a quantity is defined as the accuracy. Present difference expresses it. Accuracy is a ratio of true results to the total number of cases examined. It is given by,

Total Accuracy =(TP + TN)/(TP +TN + FP + FN)

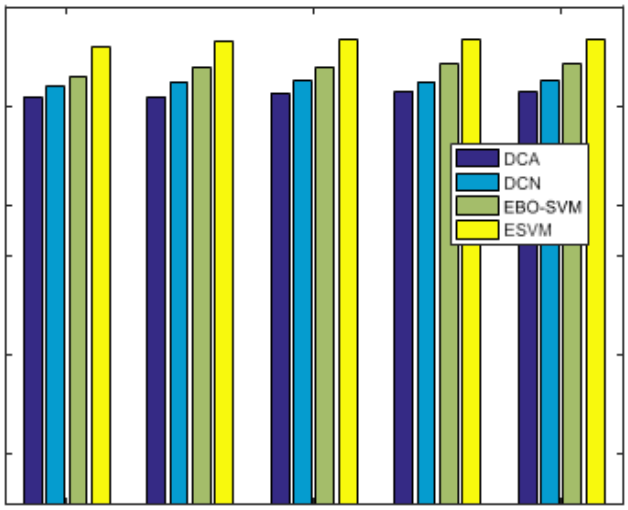

Fig.5.4 Comparison of accuracy

The comparison of accuracy of ANN, SVM and ANFIS Classification algorithms are shown in figure 5.4. The result shows that ANFIS algorithm produces high accuracy when compared to other algorithms.

Methods	Samp les	Precision(%)	Recall(%)/ Sensitivity(%)	Specifici ty (%)	F-measu re (%)	Accura cy (%)
DCA	50	80.95	77.27	85.71	79.07	82.00
	100	76.08	83.33	81.03	79.54	82.00
	150	81.42	81.42	83.75	81.42	82.66
	200	81.63	82.47	82.69	82.47	83.00
	250	83.20	83.52	81.66	83.52	82.80
DCN	50	77.77	91.30	77.77	84.00	84.00
	100	87.50	85.96	83.72	86.72	85.00
	150	90.12	83.90	87.30	86.90	85.33
	200	86.00	84.31	85.71	85.14	85.00
	250	84.39	88.80	81.03	86.54	85.20
EBO-SVM	50	95.45	77.77	95.65	85.71	86.00
	100	86.00	89.58	86.53	87.75	88.00
	150	88.09	90.24	85.29	89.15	88.00
	200	91.83	85.71	91.57	88.67	88.50
	250	88.13	87.39	89.31	87.76	88.40
ESVMI	50	96.42	90.00	95.00	93.10	92.00
	100	92.45	94.23	91.66	93.33	93.00
	150	94.80	92.40	94.36	93.59	93.33
	200	93.10	95.57	90.80	94.32	93.50
	250	94.77	93.38	93.86	94.07	93.60

VI. CONCLUSION

Data Analytics is normally connected with obtaining knowledge from reliable information source and rapidity in information processing, and future prediction of the data analysis. Big Data analytics is strongly evolving with different features of volume, velocity and Vectors. In this work, images are segmented, features are extracted and images are classified using swarm and firefly algorithms. Experimentation is conducted using medical image dataset. The experimental results are shown by the proposed methodology is drastically good compared with the existing system.

REFERENCES

1. Zheng C. H., Zhang. L, Ng V. T. Y., Shiu S. C. K., and Huang D. S.,"Molecular pattern discovery based on penalized matrix decomposition," IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 8,no. 6, pp. 1592–1603, 2011.
2. Zheng C. H., Zhang L., Ng V. T. Y., Shiu S. C. K., and Huang D. S., "Metasample-based sparse representation for tumor classification," IEEE/ACM Trans. Comput. Biol. Bioinformat., vol. 8, no. 5, pp. 1273–1282, 2011.
3. Chen, R., Shi, Y.H., Zhang, H., Hu, J.Y. and Luo, Y., 2018. Systematic prediction of target genes and pathways in cervical cancer from microRNA expression data. Oncology letters, 15(6), pp.9994-10000.

4. Claesen, M., Smet, F. D., Suykens, J. A., & Moor, B. D. (2014). EnsembleSVM: A library for ensemble learning using support vector machines. Journal of Machine Learning Research, 15, 141–145.
5. Creasman WT and Miller DS: Adenocarcinoma of the uterine corpus. In: Clinical Gynecologic Oncology. Elsevier, Philadelphia, PA, pp141-174, 2012.
6. Huang D. S. and Yu H. J., "Normalized feature vectors: A novel alignment- free sequence comparison method based on the numbers of adjacent amino acids," IEEE/ACM Trans. Comput.Biol. Bioinformat., vol. 10, no. 2, pp. 457–467, 2013.
7. Deng, S.P., Zhu, L. and Huang, D.S., 2016. Predicting hub genes associated with cervical cancer through gene co-expression networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 13(1), pp.27- 35.
8. DiLeo, M.V., Strahan, G.D., den Bakker, M. and Hoekenga, O.A., 2011. Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. PLoS One, 6(10), p.e26683.
9. Fatlawi, H.K., 2007. Enhanced Classification Model for Cervical Cancer Dataset based on Cost Sensitive Classifier. Int. J. Comput. Tech, 4, pp.115- 120.
10. Chen H., Zhu Y., and Hu K., "Cooperative bacterial foraging optimization," Discrete Dynamics in Nature and Society, vol. 2009, no. 815247, pp.1-17, 2009.
11. Han, H., Wang, W.Y. and Mao, B.H., 2005, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International Conference on Intelligent Computing (pp. 878-887). Springer, Berlin, Heidelberg.