Attribute-oriented Classification with Variable Importance using Random Forest Model

G. Rama Subba Reddy, Shaik Jaffar Hussain, K. Dinesh Kumar

Abstract*: In the present century, various classification issues are raised with large data and most commonly used machine learning algorithms are failed in the classification process to get accurate results. Datamining techniques like ensemble, which is made up of individual classifiers for the classification process and to generate the new data as well. Random forest is one of the ensemble supervised machine learning technique and essentially used in numerous machine learning applications such as the classification of text and image data. It is popular since it collects more relevant features such as variable importance measure, Out-of-bag error etc. For the viable learning and classification of random forest, it is required to reduce the number of decision trees (Pruning) in the random forest. In this paper, we have presented systematic overview of random forest algorithm along with its application areas. In addition, we presented a brief review of machine learning algorithm proposed in the recent years. Animal classification is considered as an important problem and most of the recent studies are classifying the animals by taking the image dataset. But, very less work has been done on attribute-oriented animal classification and poses many challenges in the process of extracting the accurate features. We have taken a real-time dataset from the Kaggle to classify the animal by collecting the more relevant features with the help of variable importance measure metric and compared with the other popular machine learning models.

***Keywords:* machine learning, decision trees, random forests, SVM, classification.**

I. INTRODUCTION

Many data scientists have done research on *data classification* over the demand to give a solution for the data classification problems, subsequently to proceed with the enhancement in future on different reasons. Primary reason is to identify the behavior of antisocial users of a community otherwise it leads to have negative impact [1]. Another reason is that the classification make sure they do research on world-wide social networks to have special awareness on deriving billions of users all over the world. One more reason is to analyze the media developed among social communities, also images, text, videos and sounds and to cluster the users based on their locations, friends' lists, and activities and so on. The primary objective of data classification is finding and allocating a predetermined class to an instance that is chosen, during the training of instances set with the provided class labels. Methods of classification are considered as the unique machine learning data-processing features [2] and permits performing multi-class data classification. Classification of data into predetermined classes are found as sentiment analysis or polarity analysis which represents the tone of emotional to a provided content and allocates the sentiment meaning whether it is positive or not. With the application of the sentiment analysis the each and every aspect in todays from products to services like healthcare, e-commerce, social media and some other considerable domains in which a user can give their feedback. Generally the companies always seek to gather user feedback on their products and services as well. Hence the enhanced concepts and methods of informatics engineering assist the novel thoughts which also include with sentiment analysis that illustrates the concepts of classification along with machine learning and collaborate with the user's community and their feedback, for example reviews about products and services [3].

1.1 Background

Data mining is the key among many machine learning applications. Often the users are likely to do mistakes while analyzing or during an attempt to build an association between many features. It is complex to them while solving particular issues. For such issues ML can be applied in order to solve those issues successfully. So it leads to increase the system's efficiency as well as machine designs. Each instance of any dataset which are used by the ML paradigms is represented with the help of similar feature set. Those features might be serial, binary or classified. If the instances are provided with predefined labels then it is known as supervised learning (refer Table 1), whereas the instances without labels are called as unsupervised learning [4]. The process of analyzing a rules set from the instances is known as Inductive machine learning, otherwise in general, generating a classifier for generalizing from the new instances. Figure 1 describes the procedure to apply supervised ML on real-time issue [5]. A critical step is used to specify the particular learning algorithm. Once the judgement on pre-testing is satisfied, then classifier is ready to use. The evaluation of classifiers is frequently done on the basis of accuracy in prediction. There exist at least 3 techniques for evaluating the accuracy of

Revised Manuscript Received on July 22, 2019.

G. Rama Subba Reddy, Dept. of C.S.E, Mother Theresa Institute of Engineering & Technology, Palamaner, Andrapradesh, India. Email: subbareddy1227@gmail.com

Shaik Jaffar Hussain, Dept. of C.S.E, Annamacharya Institute of Technology and Science, Rajampet, Andrapradesh, India. Email: jaffar.thebest@gmail.com

K Dinesh Kumar, SCSE, VIT University, Tamilnadu, India. Email: dineshkumar0904@gmail.com

classifier's. One of the techniques is classifying a training set by utilizing two thirds of training and the remaining is for performance estimation. Cross validation is other approach in which the training set is split into mutually exclusive and the subsets with the same size and for every subset, the classifier is trained on the other integrated subsets. The average error rate of every subset is the estimation of error rate of classifier. In cross validation, a Leave-one-out validation is considered as a typical case. Total test subsets comprises of one instance.

Table 1: Data in standard format for machine learning

Case	Feature 1	Feature 2	…	Feature n	Class
1	XXX	XXX	…	XXX	Good
2	XXX	XXX	…	XXX	Good
3	XXX	XXX	--	XXX	Bad

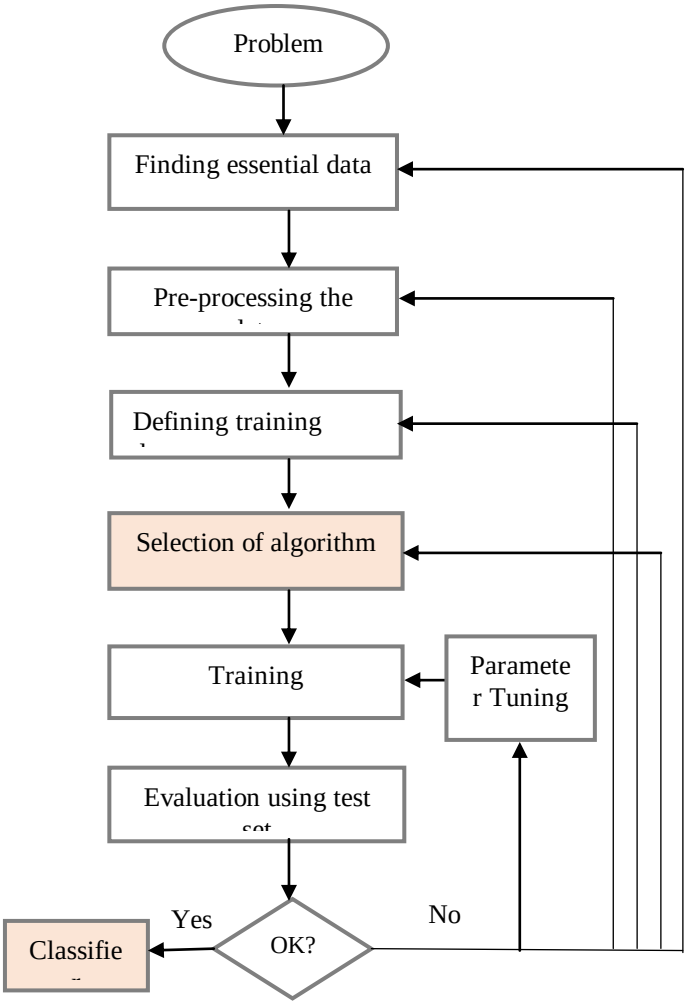


Figure 1: Supervised machine learning process

As per [6], the algorithms of ML specifically supervised that deals majorly with the classification comprising Linear Classifiers, Bayesian Networks, Quadratic Classifiers, Boosting, Naïve Bayes Classifier, Decision Tree, Perceptron, Random Forest (RF), Logistic Regression, K-Means Clustering, Neural networks, Support Vector Machine and so on. Linear models for the classification split the input vectors into the classes by using linear (hyperplane) decision boundaries [7] and Ordinary Least Squares (OLS) regression. Nevertheless with the logistic regression can predict the outcomes [8]. Logistic regression is a linear interpolation and is said as one of the mainly used mechanism in applied statistics and also in discrete data analysis [9]. Bayesian networks that contains directed acyclic graphs with the single parent node and many child nodes with the high assumption of independent nature between child nodes. Nevertheless, there has been compared naïve Bayes classifier with the state-of-the-art methods for induction of decision tree, learning based on instances and rule induction over the standard datasets and it was found that it is superior in some cases to the remaining learning methods. Bayes classifier is having an issue known as attribute-independence which was represented with Averaged One-Dependence Estimators [10]. Rest of the paradigms are based on the perceptron concept. The algorithm of Perceptron is utilized in learning from training instances batch by executing the algorithm recursively via training set till it identifies a prediction vector which is precise over all the training set [25]. These are the very recent techniques of supervised machine learning. Support Vector Machine (SVM) methods are closely associated to the classical multilayer perceptron neural networks. In [11], the researchers emphasized based on the significance of rule-based decision trees like method of classification. The two kinds of nondeterministic rules in decision tables are inhibitory rules and bounded rules. In the first rule, the decisions exist on right side whereas in the second rule, few of the decisions exist on right side. On these two rules, there established two classification algorithms of polynomial time complexity and the comparison is being done among them. Then the practical works are done over a dataset of virtual data sets build to calculate the classifier's performance based on distinct metrics [12]. In [13], multiple logical analysis approaches are contrasted for the hypothetical target classification. The algorithms of classification are utilized for multiple activities such as spam filtering [14], music emotion classification [15], web page ranking calculation for web spam [16], feature-based mining of digital images [17], or annual crop classification [18], software defect detection [19], text classification [20].

II. DECISION TREES

A kind of directed, acyclic graph can be said as Decision tree. The nodes in the decision tree represents decisions with the help of square box, circular box is used to represent the random transitions or terminal nodes and the edges or branches with binary attributes such as true/false, yes/no denotes the possible paths from one node to another node. The decision tree whichever is utilized for the machine learning aspects do not comprises any random transitions. In order to utilize a decision tree during classification otherwise for regression, one can attain a row of information or gathered features and initiates at root node, later via every subsequent decision node to terminal node [21, 22]. The procedure is easy and not hard to interpret also. It allows the trained decision trees so that these might be utilized in selecting the variables otherwise usually for feature

engineering. Let us describe this clearly, assume that you wish to buy a car and to drive on random irregular road into a random forest. A dataset of distinct cars comprising three features are Car Drive Type (Categorical), Displacement (Numeric) and Clearance (Numeric). Below is an instance of learned decision tree which helps in taking decision. The decision tree representation for this is shown in Figure 2. The root or foremost node in a tree (only applicable when a tree is having a single root) is considered as a decision node which classifies a dataset with the help of a variable otherwise by a feature which yields good splitting measure calculated for every subset or class of that dataset generated from split. Decision tree realizes by dataset classification in recursive fashion at every node of decision from root node onwards (node by node way) based on the splitting measure. Terminal nodes are obtained during the classification of metric at wide range. Major splitting metrics are reducing the Gini Impurity (used by CART) or increasing the Information Gain (used by ID3, C4.5).

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (1)$$

From equation (1), Pi denotes the probability that an arbitrary tuple in D of class Ci.

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{X Info}(D_j) \quad (2)$$

Where, Info (D) represents an average data set essential in finding tuple's class label in D., |Dj|/|D| is taken as j^{th} partition weight and InfoA (D) is the predicted data used in tuple classification from D.

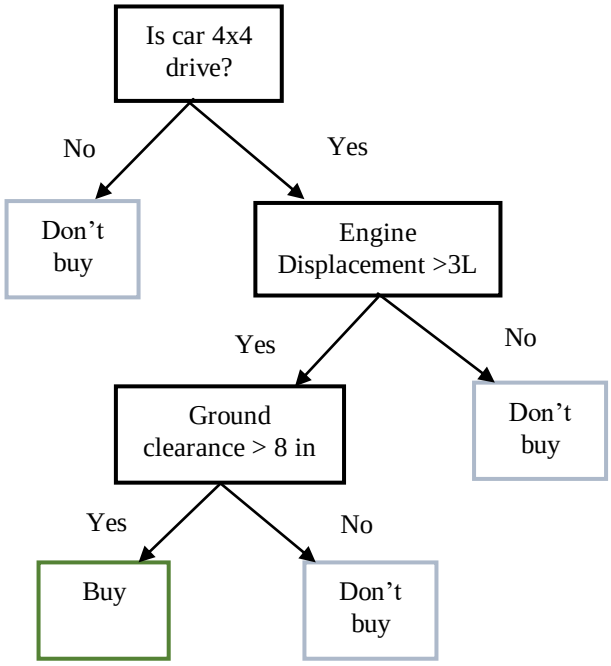


Figure 2: Decision Tree model for classification

III. RANDOM FOREST MODEL

Random Forest (RF) is a set of decision trees that are not pruned. Random forests are frequently utilized when there are having huge training datasets and vast input variables. A classifier known as random forest model comprises of various decision trees and class results is a class mode generated by single trees [23]. Every tree is generated based on the Algorithm 1.

Algorithm 1: Decision Tree generation in RF

Step 1: 'N' is the No. of training cases, and; 'M' is the count of variables.
Step 2: 'm; are utilized in determining a decision node in a tree; m<M.
Step 3: Select a training set to such tree by selecting N times by replacing from complete existing N training sets (say sample of bootstrap). Utilize the remaining sets to predict the error of tree, by guess its classes.
Step 4: For every node in a tree, the random selection of m variables is done based on decision node. Evaluate the best classification considering those m variables of training set.
Step 5: Every tree is totally created and is not done any pruning over it (as done in the generation of classifier of a normal tree).

The benefits with random forests when compared with the other classification paradigms are: a) for distinct data sets, it yields best accurate classifier. b) It manages huge set of the input variables. c) It predicts the variables significance by determining the classification. d) It creates an interior unbiased estimation of generalization error as a forest building outputs. e) It also includes a best approach to predict the skipped information and manages preciseness when high quantity of data is skipped [24]. The complete set of data is divided in to subsets, for each subset corresponding decision trees are generated. This process is shown in Figure 3 and the decision forest representation is shown in Figure 4.

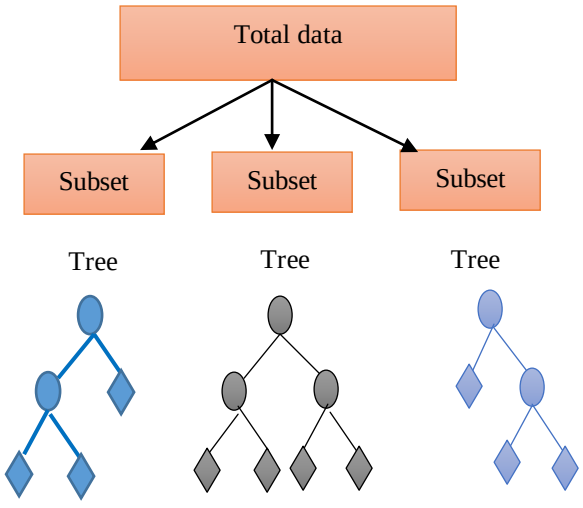


Figure 3: Random forest with multiple decision trees

Random forest is a predictor contains a set of randomized related regression trees $\{r_n(x, \Theta_m, D_n), m \geq 1\}$, where $\Theta_1, \Theta_2 \ldots$ are the outcomes of Θ, a randomizing variable. Those random trees are integrated to constitute aggregated regression estimate:

$\overline{r_n}$ (X,D_n) =E_Θ [r_n (X, Θ, D_n)] (3)

Here E_Θ refers to prediction of random metric, considering X and dataset D_n. In the below, to minimise the notation we are neglecting the dependencies of sample estimates and say for instance, use $\overline{r_n}$ (X) than $\overline{r_n}$ (X,D_n).

1. At every node, X coordinate is equal to $X^{(1)}$,…, $X^{(d)}$ is chosen with j-th feature with the $p_{nj} \in (0,1)$ probability.

2. At every node, once the selection of coordinate is done, split will be at the middle of selected side.

Decision Forest

-

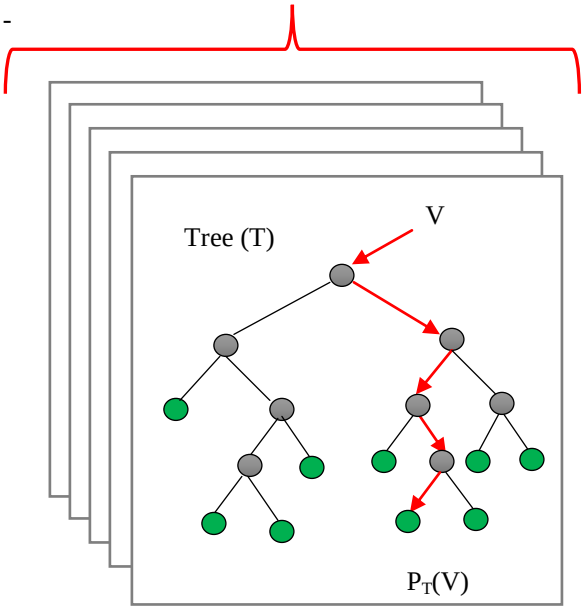


Figure 4: Random forest Model

IV. RESULTS AND DISCUSSION

In this section, we presented our experimental analysis on classification of animals using random forest algorithm and compared the result with classification result of the decision trees, SVM and K-NN algorithms. We have collected the dataset from the Kaggle, which contains 101 animals from the zoo. We considered 6 classes, with 16 types of attributes. The main goal of this dataset is accurate classification of the animal by considering the given attributes. We have shown a random sample set of animals i.e. 10, and attributes 9 in Figure 5.a among the total of 101 animals and 16 attributes. The class type and the label of animals in each class is shown in Figure 5.b and the total number of animal in each class is shown in Figure 5.c. For instance, crow, duck, skimmer are the class of Bird, and flea, housefly, and honeybee are the class of Bug.

The input data is clean without any missing values or outliers and we considered 80% as training data and 20% as test data. Logistic regression will not work with properly with probabilities, here we need at least 10 observations per a single feature. On the other hand, naïve Bayes is feature independent, and we have high correlated features it is not optimal to choose this algorithm. Random forest algorithm is giving the results with 100% accuracy. We also compared this with other models such as K-NN, SVM, and Decision trees. The accuracy of the other models are shown in Figure 6. The importance of each feature corresponding to the animals are shown in Figure 7. Here, the model is able to train with 14 features to get better accuracy, but we got 99% accuracy with 8 (almost half of the original) features among total of 14 features. The features legs, toothed, backbone, tail, breaths, feathers, milk, and aquatic are considered. It has given one misclassification and hence we have taken 10 features in another experiment to get the 100% accuracy. The learning curve for different training examples along with training and validation accuracy is shown in Figure 8. The green color curve denotes the validation score and the orange curve denotes the training score.

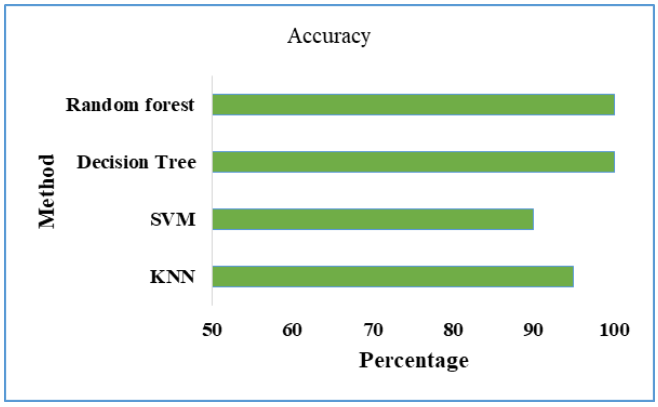


Figure 6: Comparison of accuracy with other models.

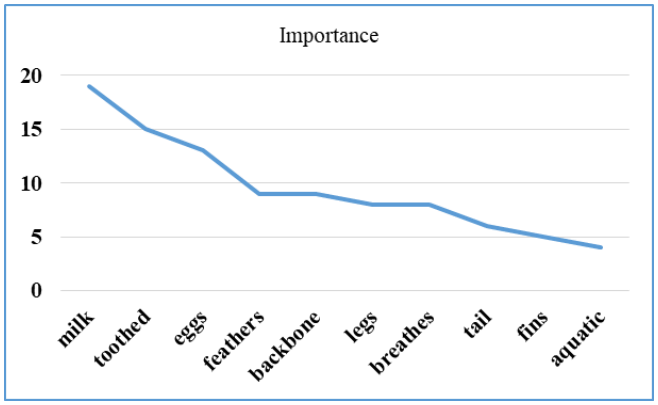


Figure 7: Importance of features selection

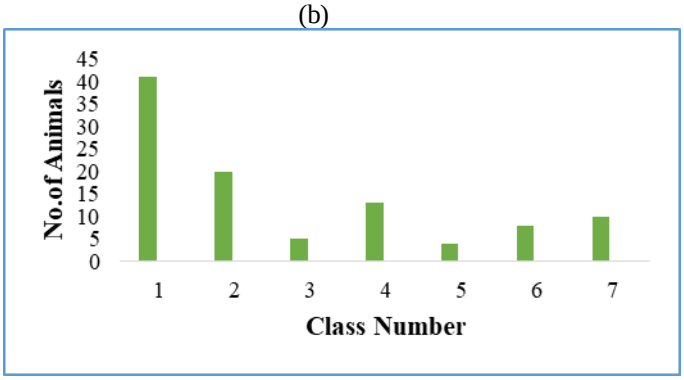
Figure 8: ROC curve for Training and Validation sets.

Animal	hair	feathers	eggs	milk	toothed	backbone	breathes	tail	domestic	class type
aardvark	1	0	0	1	1	1	1	0	0	**1**
crab	0	0	1	0	0	0	0	0	0	**7**
crayfish	0	0	1	0	0	0	0	0	0	**7**
crow	0	1	1	0	0	1	1	1	0	**2**
deer	1	0	0	1	1	1	1	1	0	**1**
dogfish	0	0	1	0	1	1	0	1	0	**4**
dolphin	0	0	0	1	1	1	1	1	0	**1**
dove	0	1	1	0	0	1	1	1	1	**2**
duck	0	1	1	0	0	1	1	1	0	**2**
elephant	1	0	0	1	1	1	1	1	0	**1**

(a)

Class Number	Class type
1	Mammal
2	Bird
3	Reptile
4	Fish
5	Amphibian
6	Bug
7	Invertebrate

(b)



(c)

Figure 5: Random sample data from the zoo dataset. a. attributes b. class type c. No. of animals of each type

V. CONCLUSION

Machine learning is one of most exiting technology among which, classification problems are placed a key role in many applications and poses numerous challenges. Random forest is one of the ensemble classifier used to classification and regression. It uses the decision trees algorithm as a base classifier. It consists of multiple trained classifiers to classify and to generate new instances. In this paper, we presented a brief review about decision tree and random forest algorithm. Later, we described the current and recent work on random forest algorithm along with its application. Then, we applied this algorithm on a real time zoo data set and compared the results with other machine learning models. It does not require many binary classifiers to evaluate the results of multi-class problems. Though random forest is giving accurate results, it takes more time and need huge amount of labelled information compared with other classification techniques. In future, we can extend our work to reduce the time complexity of random forest method by taking a real-time dataset.

REFERENCES

1. Madsen, Kristoffer H., et al. "Perspectives on machine learning for classification of Schizotypy using fMRI data." Schizophrenia bulletin 44.suppl_2 (2018): S480-S490.
2. Chang, Chih-Wei, and Nam T. Dinh. "Classification of machine learning frameworks for data-driven thermal fluid models." International Journal of Thermal Sciences 135 (2019): 559-579.
3. Tabak, Michael A., et al. "Machine learning to classify animal species in camera trap images: applications in ecology." Methods in Ecology and Evolution 10.4 (2019): 585-590.
4. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3-24.
5. Osisanwo, F. Y., et al. "Supervised machine learning algorithms: classification and comparison." International Journal of Computer Trends and Technology (IJCTT) 48.3 (2017): 128-138.
6. Taiwo, O. A. (2010). Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang (Ed.), ISBN: 978-953-307-034-6, InTech, University of Portsmouth United Kingdom. Pp 3 – 31.
7. Elder, J. (n.d). Introduction to Machine Learning and Pattern Recognition. Available at LASSONDE University EECS Department York website:

http://www.eecs.yorku.ca/course_archive/2011-12/F/44045327/lectures/01%20Introduction.pdf

8. Newsom, I. (2015). Data Analysis II: Logistic Regression. Available at: http://web.pdx.edu/~newsomj/da2/ho_logistic.pdf
9. Logistic Regression pp. 223 – 237. Available at: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12 .pdf
10. Hormozi, H., Hormozi, E. & Nohooji, H. R. (2012). The Classification of the Applicable Machine Learning Methods in Robot Manipulators. International Journal of Machine Learning and Computing (IJMLC), Vol. 2, No. 5, 2012 doi: 10.7763/IJMLC.2012.V2.189pp. 560 – 563.
11. Delimata, Paweł, et al. "Comparison of some classification algorithms based on deterministic and nondeterministic decision rules." Transactions on rough sets XII. Springer, Berlin, Heidelberg, 2010. 90-105.
12. Ketkar, Nikhil S., Lawrence B. Holder, and Diane J. Cook. "Empirical comparison of graph classification algorithms." 2009 IEEE Symposium on Computational Intelligence and Data Mining. IEEE, 2009.
13. Espinosa-Ortega, Fabricio, et al. "Comparison of autoantibody specificities tested by a line blot assay and immunoprecipitation-based algorithm in patients with idiopathic inflammatory myopathies." Annals of the rheumatic diseases 78.6 (2019): 858-860.
14. Li, Tong, et al. "Differentially private Naive Bayes learning over multiple data sources." Information Sciences 444 (2018): 89-104.
15. Rachman, Fika Hastarita, Riyanarto Sarno, and Chastine Fatichah. "Music emotion classification based on lyrics-audio using corpus based emotion." International Journal of Electrical and Computer Engineering 8.3 (2018): 1720.
16. Li, Yuancheng, Xiangqian Nie, and Rong Huang. "Web spam classification method based on deep belief networks." Expert Systems with Applications 96 (2018): 261-270.
17. Nie, Xiushan, et al. "Robust image fingerprinting based on feature point relationship mining." IEEE Transactions on Information Forensics and Security 13.6 (2018): 1509-1523.
18. Vuolo, Francesco, et al. "How much does multi-temporal Sentinel-2 data improve crop type classification?." International journal of applied earth observation and geoinformation 72 (2018): 122-130.
19. Bennin, Kwabena Ebo, Jacky W. Keung, and Akito Monden. "On the relative value of data resampling approaches for software defect prediction." Empirical Software Engineering24.2 (2019): 602-636.
20. Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." arXiv preprint arXiv:1801.06146 (2018).
21. Salloum, Said A., et al. "A survey of Arabic text mining." Intelligent Natural Language Processing: Trends and Applications. Springer, Cham, 2018. 417-431.
22. Denisko, Danielle, and Michael M. Hoffman. "Classification and interaction in random forests." Proceedings of the National Academy of Sciences 115.8 (2018): 1690-1692.
23. Bernard S, Heutte L, and Adam S, (2012): Dynamic Random forests, Pattern Recognition Letters, 33, 15801586
24. Biau, GÃŠrard. "Analysis of a random forests model." Journal of Machine Learning Research 13.Apr (2012): 1063-1095.
25. Kumar, K.D. et al, Prediction methods for effective resource provisioning in cloud computing: A Survey". Multiagent and Grid Systems, 14(3), 2018