Indian Journal of Science and Technology, *Vol 9(22), DOI: 10.17485/ijst/2016/v9i22/92713, June 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Comparative Study on Various Data Mining Algorithms with Special Reference to Crop Yield Prediction

Hetal Patel* and Dharmendra Patel

Faculty of Computer Science and Applications, Charotar University of Science and Technology (CHARUSAT), Changa - 388421, Gujarat, India; hetalpatel.mca@charusat.ac.in, dharmendrapatel.mca@charusat.ac.in

Abstract:

Objectives: To compare different data mining algorithms with the same parameters on the 10fold cross validation test to predict the crop yield. **Methods/Analysis**: Different data mining classification algorithms like K-nearest Neighbor, K-means, Neural Network, Support Vector Machine, Case-based Reasoning, Decision Tree algorithm, etc. are applied for various application of agriculture domain. A comparative study is done by using J48, Naïve Bayes and Simple Cart algorithms to determine which classification algorithm is best fitted for crop prediction. **Findings**: In this study, this work reveals the superior performance of J48 classification algorithm with accuracy 89.33% for crop prediction than the other two classification algorithms Simple Cart and Naïve Bayes. **Novelty /Improvement**: This study first time demonstrates the application of different data mining classification techniques (as discussed above) in the domain of agriculture for yield prediction.

Keywords: Classification Algorithm, Crop Prediction, Data Mining, Decision Tree, J48

1. Introduction

Agriculture is being playing a significant role in the Indian Economy. Agriculture is the most important occupation for most of the Indian families and large percentage of population in India is reliant on agriculture for livelihood. Agriculture is the main source of income and employment for the majority of people in this world especially of rural areas. In above circumstance, the right selection of crop is absolutely necessary for better yield proportion so a system needs to be in place for doing correct prediction of yield.

Data Mining, termed as the process of discovering patterns from gigantic databases[1]. The Data Mining techniques are used to extract the precise and previously unknown patterns or information from huge volume of data[2,3]. In this study, various classification algorithms of data mining, particularly Simple Cart, J48, and Naïve Bayes algorithms of decision tree are explored for crop prediction. The utmost significant parameters in the selection of classification algorithm are accuracy, efficiency and error rate for proposed research work.

The main objective of using information unseen within the database provides the inspiration to the researcher in the area of agriculture for applying such techniques to do forecast for imminent trends of agricultural progressions. For the same, so many work is being done by employing various data mining techniques on agriculture database. Verheyen et al. had done the classification of the soil profile using fuzzy k-means of classification techniques with extra grades algorithm[4]. The author Bhargavi et al. has applied the Naïve Bayes algorithms of data mining for soil classification[5]. The k-means clustering algorithm is employed to classify plant, soil and residue using various color images[6]. The k-means algorithm is also explored for apple grading before marketing[7], to tracking out the

**Author for correspondence*

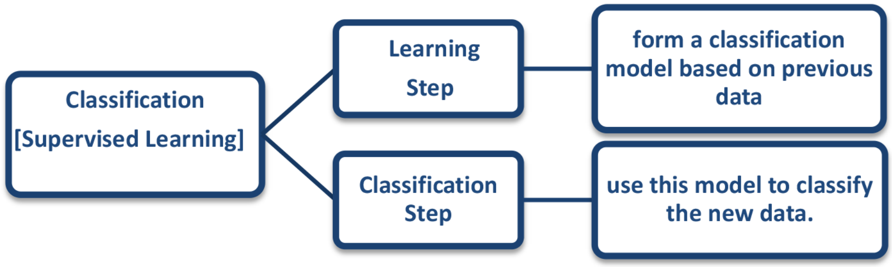

Figure 1. Classification technique.

changes done in water quality[8], identifying the weeds from the crop fields[9]. The k-means algorithm of data mining can be used to predict the outcome of the fermentation in the early days of this process using[10]. The precipitations and weather variables are simulated on daily basis, is done by using k-nearest neighbor classification algorithm[11] and it is also performed well for estimating the soil water parameters and climate forecasting[12]. The neural network is being applied to predict the date of flowering and maturity of soybean and forecast of water resource's variables[13,14]. The classification of crop and the analysis of the climate change scenarios is done by support vector machine[15,16]. The prediction of soil fertility is done by using decision tree[17]. The investigation of agriculture land which is vanished, is done by clustering technique based on partitioning and hierarchical algorithms of data mining[18]. Also estimation of the crop yield analysis is done by k-means algotithm[19]. The author Ramesh et al. applied decision tree algorithm for predicting the fertility of soil[20]. The author analyzed the effect of traits inside the database to do soil classification using Naïve Bayes classifier[21]. The continuous soil profile classification is achieved using the 'fuzzy k-means which is explored by the authors22.

The article is organized into 5 sections. The Section 2 provides information about the classification algorithm for crop prediction. Experimental Results and Discussions are outlined in section 3. Finally, in section 4, the conclusion is mentioned.

2. Classification Algorithm for Crop Prediction

Classification which is also known as supervised learning or predictive modeling, is based on the nature of the information being extracted. Classification is a divided into two-step process. Fig. 1 shows that the first step which is known as learning step. In this step the classification model is build using the previously known data set. In the second step which is known as classification step, if the model's accuracy is adequate then deploy this model to classify the new data.

For classification, each decision tree algorithm has its particular process and all generates outcome of its classification without regardless of the outcome of rest classification algorithm. The tricky task is the selection of right algorithm as the type of data, how to retrieve the data, noisy data and time allotted to train the algorithm etc. effect on the performance and accuracy of classification algorithms.

Systems are going to build the diffident kind of classifiers using different tools of data mining. Such systems get input as group of cases which belongs to a small number of classes described by a fixed set of attributes. It generates a classifier which can be precisely forecast the class which belongs to newer class.

The Different types of attributes are numerical attributes, categorical attributes or mixed attributes. Numerical attributes are those attributes whose domain is numerical. Categorical attributes maybe either ordered or unordered. Different classification algorithms perform differently on different attribute's type and sizes. Following are the well-known classification techniques.

2.1 Decision Tree

Decision Tree is the one of the popular classification method that gives result in form of tree structure. Decision-tree is generally built by recursive partitioning. In this there is root and child of the tree. For the root of the tree, a single attribute split is chosen by using some criterion. For each child, the data is then divided according to the test, and the process repeats recursively. After built of the tree, a pruning step is executed, which reduces the tree size. In short, each node indicates a test on an attribute value and each branch indicates an outcome of test. It is widely used in the field of pattern recognition, machine learning and prediction[23]. Decision tree can easily be converted to classification tree[24]. It is very easy to understand and the provided result is worthy with small as well as large data. The data from different domain like Agriculture, Education, Medical, Diseases Analysis, Health Care, Medicine, Manufacturing, Production, Analysis of Financial, Fraud Detection and Astronomy etc. have been analyzed using Decision tree induction algorithms. The different decision tree algorithms are C4.5, ID3, CART, J48, NB Tree, REP Tree, Simple Cart and Naïve Bayes.

2.1.1 J48

The J48 algorithm was proposed by Ross Quinlan in 1993. The earlier versions are ID3 and C4.5. J48 is a classifier similar to C4.5 and C5.0. The classification tree generated is on the basis of the input attributes. The divide and conquer slant is use by decision tree. At each node, the testing of each attribute is done and the branches are prepared till leaf nodes are grasped to form a tree. The decision tree formed this way by using J48 algorithm, is an improved version of the C4.5[25]. The pruning method is of two type. In one which is known as sub tree replacement in which few sub trees are picked up and substituted by single leaves. In the second type, a node is proceeded upwards in the direction of root of the tree by replacing other nodes through the path. It has a negligible effect on decision tree models. For the study, J48 algorithm is used as it has more accuracy rate.

2.1.2 Naïve Bayes

One of the most successful learning algorithms is Naive Bayes intended for text categorization which is based on the Bayes rule. The conditional independence between classes is the only assumption. Based on the rule, the algorithm attempts to estimate the conditional probabilities of classes given an observation[26]. using the joint probabilities of sample observations and classes, a simple probabilistic classifier which is based on applying Bayes' theorem having strong (naive) independence assumptions is a naive Bayes classifier. The Naive Bayes classifiers can be trained very well in a supervised learning setting which depend on the precise nature of the probability model. A benefit of the naive Bayes classifier is that it only wants a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification[27].

2.1.3 Simple Cart

The classification and regression tree algorithm was proposed by developed by Breiman et al., in 1984. The main use of CART was for data exploration and prediction. The construction of decision tree is done by using historical data set. For building up a decision tree, it is required to supply the learning sample which is a dataset of historical data having pre-allocated classes for all observations.

2.1.4 SVM

It is one of the supervised learning algorithm. It classifies the inputted data in two different classes. First of all, it makes a single or multiple hyper planes. The cases that outline the hyperplane are the support vectors. Then it chooses the hyper plane which gives the broadest way between the nearest points of the two classes. The main task of SVM is to maximize the space between the two classes to reduce the error when the given data are classified. The optimization is necessary which is only replays on the product of the pairs of sample data[32].

2.2 Role of Data Mining in Crop Prediction

In this study, a model is presented, which is shown in the Figure 2. The selection of data mining technique for crop prediction is very tedious because each and every algorithm has its own advantages and limitations. In the first step, the problem is identified. There is no such model which predicts what kind of plant should be cultivated so that there are maximum gains. The required data is collected from various reliable sources (Table 1) and converted into the required format. We have used different classification algorithm. The model is trained with various classification algorithms using J48, Naïve Bayes and Simple Cart algorithm. Finally, the original data is supplied to the model and generate the outcomes.

2.3 Data Source and Variables

The data were collected from different sources. The same data are transformed into attribute relation file format (ARFF). The various variables, its description and domain are shown in the Table 2 and Table 3.

3. Experimental Results and Discussions

The model is simulated using WEKA tool. Empirical results are listed in Table 4. For measuring the prediction accuracy of the three methods (J48, Naïve Bayes, Simple Cart), the authors used a 10-fold cross-validation procedure in which it divided the dataset into 10 mutually exclusive partitions using a stratified sampling technique. The instances used are 300. The comparison of algorithms with regard of confusion matrix is shown in the Table 5. The J48 algorithm is predicting better than the Naïve Bayes and Simple Cart as it takes less time to generate the output and it classified 280 instances correctly.

3.1 Assessment Criteria

In this experiment, we compared the performance of different classifier with each other in terms of accuracy, recall and precision which are explored by P. Suganya et al.[33]are as follows:

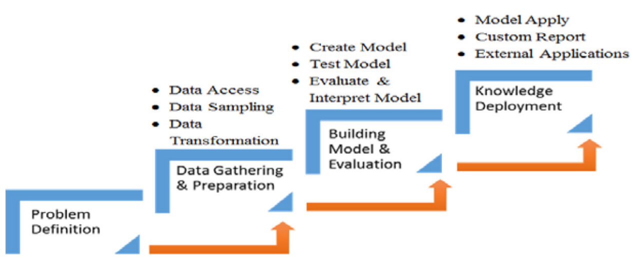

Figure 2. Model for crop prediction.

Table 1. Data source

Sr. No	Data Source
1	https://india.gov.in/topics/agriculture/crops
2	http://farmer.gov.in/FarmerHome.aspx
3	https://ikhedut.gujarat.gov.in/
4	http://aggie-horticulture.tamu.edu/vegetable/guides/organic-vegetable-production-guide/organic-crop-production-requirements/
5	http://www.accuweather.com/en/in/anand/188164/weather-forecast/188164
6	http://shc.aau.in/home/soil
7	http://agritech.tnau.ac.in/agriculture/crop_production_varieties.html
8	http://www.gujenvis.nic.in/PDF/soil.pdf
9	http://fert.nic.in/node/1452
10	http://nfsm.gov.in/

Table 2. Comparison of different classifiers

Sr. No	Variable Name	Description	Domain
1	soil_type	Type of Soil	{scl,sltc,cl,cltc,c,s, sltsc,scltc,ds}
2	Soil_fert	Fertility of soil	{high,low, medium,nil}
3	season_type	Season type like Kharif, Ruby and Summer	{k,r,s}
4	weather	Weather condition	{hot,windy,rainy, cloudy}
5	wind_speed	Speed of wind	{high,low,avg}
6	moisture_per	Moisture percentage	{high,low,nil}
7	water_source	Source of water	{irrigation,well, cannal}
8	water_cost	Cost of water	{high,low,nil}
9	tech_used	Technology used	{gh,other,nil}
10	fertilizer	Type of fertilizer	{l,g,n}
11	pesticide	Type of pesticide like solid or liquid	{s,l}
12	fert_type	Type of fertilizer like organic or inorganic	{org,inorag}
13	(Response Variable) crop_type	Crop predicted	{cereals,vegetables, flowers,fruits,pulses}

a) Accuracy – It is the proportion of correctly classified data (true positives & true negatives).

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}$$

b) Precision –It is the percentage of selected data items that are correct (true).

$$Precision = \frac{T_P}{T_P + F_P}$$

c) Recall (Sensitivity) – It is the percentage of correct data items that are selected.

$$Recall = \frac{T_P}{T_P + F_N}$$

Table 3. Attributes and its description

Sr. No	Attribute	Description
1	{scl,sltc,cl,cltc,c,s,sltsc,scltc,ds}	Type of Soil:sandy clay loam(SCL), silt loam to clay(sltc), clay like(cl), clay loam to clay(cltc), clay(c), sand(s), sandy loam to sandy clay loam(sltsc), sandy clay loam to clay(scltc), no definite structure (ns)
2	{high,low,medium,nil}	Fertility of soil
3	{k,r,s}	Season type like Kharif(k), Ruby(r) and Summer(s)
4	{hot,windy,rainy,cloudy}	Weather condition
5	{high,low,avg}	Speed of wind
6	{high,low,nil}	Moisture percentage
7	{irrigation,well,cannal}	Source of water
8	{high,low,nil}	Cost of water
9	{gh,other,nil}	Technology used: Green House(gh)
10	{l,g}	Type of fertilizer: Liquid(l), Granular(g)
11	{s,l}	Type of pesticide like solid or liquid
12	{org,inorag}	Type of fertilizer like organic or inorganic
13	{cereals,vegetables,flowers,fruits,pulses}	Crop predicted

Table 4. Comparison of different classifiers

<table>
<tr><th>Technique</th><th>Instances</th><th>Number of Leaves</th><th>Size of the tree</th><th>Confusion Matrix</th><th>Time taken to build model</th></tr>
<tr><td>J48</td><td>300</td><td>21</td><td>27</td><td>a b c d e <-- classified as
30 0 0 0 10 | a = cereals
0 30 0 0 0 | b = vegetables
0 0 30 0 0 | c = flowers
0 0 0 38 2 | d = fruits
0 0 0 20 140 | e = pulses</td><td>0.01 Seconds</td></tr>
<tr><td>Naïve Bayes</td><td>300</td><td>-</td><td>-</td><td>a b c d e <-- classified as
30 0 0 0 10 | a = cereals
0 30 0 0 0 | b = vegetables
0 0 30 0 0 | c = flowers
0 0 0 20 20 | d = fruits
0 0 0 10 150 | e = pulses</td><td>0.03 Seconds</td></tr>
<tr><td>Simple Cart</td><td>300</td><td>8</td><td>15</td><td>a b c d e <-- classified as
31 0 0 0 9 | a = cereals
0 30 0 0 0 | b = vegetables
0 0 30 0 0 | c = flowers
0 0 0 38 2 | d = fruits
6 0 0 20 134 | e = pulses</td><td>0.06 Seconds</td></tr>
</table>

Table 5. Comparison of detailed accuracy measure

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
J48	0.75	0	1	0.75	0.857	0.998	cereals
	1	0	1	1	1	1	vegetables
	1	0	1	1	1	1	flowers
	0.95	0.077	0.655	0.95	0.776	0.94	fruits
	0.875	0.086	0.921	0.875	0.897	0.939	pulses
	0.893	0.056	0.912	0.893	0.896	0.953	Weighted Avg.
Naïve Bayes	0.75	0	1	0.75	0.857	0.832	cereals
	1	0	1	1	1	1	vegetables
	1	0	1	1	1	1	flowers
	0.5	0.038	0.667	0.5	0.571	0.919	fruits
	0.938	0.214	0.833	0.938	0.882	0.911	pulses
	0.867	0.119	0.867	0.867	0.861	0.92	Weighted Avg.
SimpleCart	0.775	0.023	0.838	0.775	0.805	0.981	cereals
	1	0	1	1	1	1	vegetables
	1	0	1	1	1	1	flowers
	0.95	0.077	0.655	0.95	0.776	0.94	fruits
	0.838	0.099	0.924	0.838	0.879	0.954	pulses
	0.877	0.055	0.892	0.877	0.879	0.935	Weighted Avg.

where TP is true positive rate, TN is true negative rate, FP is false positive rate and FN is false negative rate.

d) Kappa statistics – It measure the agreement of prediction with the true class.

$$K = \frac{P_A - P_{CE}}{1 - P_{CE}}$$

where P_A indicates the percentage agreement i.e. between classifier and actual truth. P_{CE} indicates the chance agreement. If value of k is one, then it indicates that the agreement is faultless. If value of K is zero, then it means there may be a chance of fault or agreement. The Kappa statistics in Fig. 4, we can see that, each classifier produces K value greater than 0, that means each classifier is doing better than chance of agreement. As per the fig.3, J48 classifier has higher kappa statistics i.e. 0.841 in comparison with other two classifiers for the agriculture database.

Classification accuracy as per the equation (1), is seen in the Fig.4. The accuracy of J48 is greater than the other two i.e. 89.33%. Also it is cleared that all algorithm is having more than 80% accuracy but J48 is performing well the said database.

Table 5 shows the results from the classification experiments using the said datasets. The ROC area for all type of crop is higher with J48 algorithm. So again J48 is more effective than the other two algorithms.

4. Conclusions and Future Scope

In our research work, we have equated the efficiency of the classification algorithms named J48, Simple Cart and Naïve Bayes in terms of various evaluation parameters like false positive rate, true positive rate, Recall, Precision, ROC Area and F-Measure. It helps in construction of effectual crop prediction system. It is perceived by the generated result that the best predictor algorithm of decision tree is J48 with 89.33% accuracy on given data set, however Simple Cart and Naïve Bayes give 85.66% and 82.66% accuracies, respectively. This comparative study for crop prediction applying a huge dataset having a 10-fold cross-validation provides a perception about the prediction capacity of employed data mining algorithms.

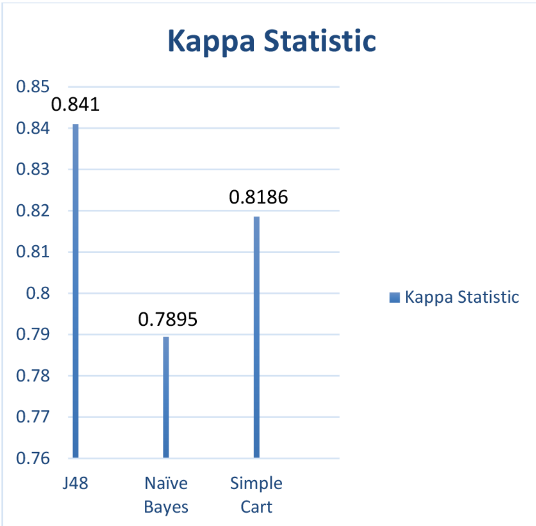

Figure 3. Kappa statistics of J48, Naïve Bayes and Simple Cart.

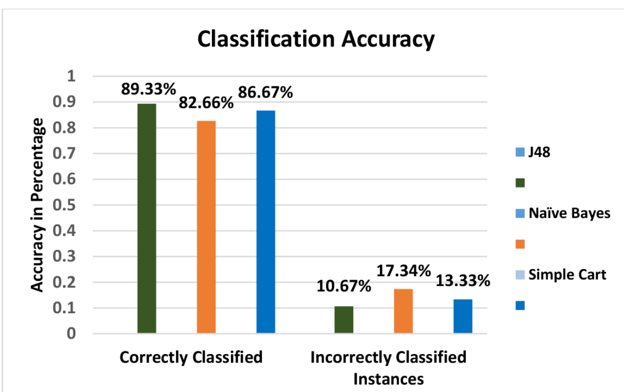

Figure 4. Comparison of correctly and incorrectly classifiers.

Accuracy of the model shows that there is a large scope in Data Mining to design classification techniques.

5. Acknowledgement

The authors would like to thank Dr. Pritpal Singh for providing his valuable inputs to improve this paper. The authors also thank Charotar University of Science and Technology (CHARUSAT) for providing necessary resources to accomplish this research.

6. References:

1. Chen MS, Han J, Yu PS. Data mining: an overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering. 1996 Dec; 8(6):866–83.
2. Fayyad UM, Piatetsky-Shapiro G, Smyth P. Knowledge discovery and data mining: towards a unifying framework. American Association for Artificial Intelligence. 1996 Aug 2; 96:82–88.
3. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996 Mar 15; 17(3):37–54.
4. Verheyen K, Adriaens D, Hermy M, Deckers S. High-resolution continuous soil classification using morphological soil profile descriptions. Geoderma. 2001 Apr 30; 101(3):31–48.
5. Bhargavi P, Jyothi S. Applying naive bayes data mining technique for classification of agricultural land soils. International Journal of Computer Science and Network Security. 2009 Aug; 9(8):117–22.
6. Meyer GE, Neto JC, Jones DD, Hindman TW. Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images. Computers and Electronics in Agriculture. 2004 Mar 31; 42(3):161–80.
7. Leemans V, Destain MF. A real-time grading method of apples based on features extracted from defects. Journal of Food Engineering. 2004 Jan 31; 61(1):83–9.
8. Klise KA, McKenna SA. Water quality change detection: multivariate algorithms. Optics and Photonics in Global Homeland Security II; 2006 May 5.
9. Tellaeche A, BurgosArtizzu XP, Pajares G, Ribeiro A. A vision-based hybrid classifier for weeds detection in precision agriculture through the Bayesian and Fuzzy k-Means paradigms. Innovations in Hybrid Intelligent Systems; 2007. p. 72–79.
10. Urtubia A, Pérez-Correa JR, Soto A, Pszczólkowski P. Using data mining techniques to predict industrial wine problem fermentations. Food Control. 2007 Dec 31; 18(12):1512–7.
11. Rajagopalan B, Lall U. A k-nearest-neighbor simulator for daily precipitation and other weather variables. Water resources research. 1999 Oct 1; 35(10):3089–101.
12. Mucherino A, Papajorgji PJ, Pardalos PM. Data mining in agriculture. Springer Science & Business Media; 2009 Sep 22.
13. Elizondo DA, McClendon RW, Hoogenboom G. Neural network models for predicting flowering and physiological maturity of soybean. Transactions of the ASAE. 1994; 37(3):981–8.
14. Maier HR, Dandy GC. Neural networks for the prediction and forecasting of water resources variables: a review of

modelling issues and applications. Environmental modelling & software. 2000 Jan 31; 15(1):101–24.
15. Camps-Valls G, Gómez-Chova L, Calpe-Maravilla J, Soria-Olivas E, Martín-Guerrero JD, Moreno J. Support vector machines for crop classification using hyperspectral data. Pattern Recognition and Image Analysis. 2003 Jun 4; 2652:134–41.
16. Tripathi S, Srinivas VV, Nanjundiah RS. Downscaling of precipitation for climate change scenarios: a support vector machine approach. Journal of Hydrology. 2006 Nov 15; 330(3):621–40.
17. Gholap J. Performance tuning of J48 Algorithm for prediction of soil fertility. Asian Journal of Computer Science and Information Technology. 2012; 2(8).
18. Megala S, Hemalatha M. A novel datamining approach to determine the vanished agricultural land in Tamilnadu. International Journal of Computer Applications. 2011; 23(3):1–6.
19. Ramesh D, Vardhan BV. Data mining techniques and applications to agricultural yield data. International Journal of Advanced Research in Computer and Communication Engineering. 2013 Sep; 2(9):3477–80.
20. Ramesh V, Ramar K. Classification of agricultural land soils: a data mining approach. Agricultural Journal. 2011; 6(3):82–6.
21. Patel H, Patel D. A Brief survey of data mining techniques applied to agricultural data. International Journal of Computer Applications. 2014 Jan 1; 95(9):1–3.
22. Sharma AK, Sahni S. A comparative study of classification algorithms for spam email data analysis. International Journal on Computer Science and Engineering. 2011 May; 3(5):1890–5.
23. Kohavi R. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. Association for the Advancement of Artificial Intelligence. 1996 Aug 2; 96:202–07.
24. Bahramirad S, Mustapha A, Eshraghi M. Classification of liver disease diagnosis: a comparative study. 2013 Second International Conference on Informatics and Applications (ICIA); 2013 Sep 23. p. 42–6.
25. Venkatesan E, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. Indian Journal of Science and Technology. 2015 Nov 7; 8(29):1–8. DOI: 10.17485/ijst/2015/v8i1/84646.
26. Suma VR, Renjith S, Ashok S, Judy MV. Analytical study of selected classification algorithms for clinical dataset. Indian Journal of Science and Technology. 2016 Mar 22; 9(11):1–9. DOI: 10.17485/ijst/2016/v9i11/67151.
27. Patil TR, Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International Journal of Computer Science and Applications. 2013 Apr; 6(2):2561–61.
28. Saadati M, Bagheri A. Mining children ever born data; classification tree approach. Indian Journal of Science and Technology. 2015 Nov 14; 8(30):1–7. DOI: 10.17485/ijst/2015/v8i30/90251.
29. Jenicka S, Suruliandi A. Comparative study of classification algorithms with modified multivariate local binary pattern texture model on remotely sensed images. 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai: Tamil Nadu; 2011 Jun 3. p. 848–52.
30. Cios KJ, Liu N. A machine learning method for generation of a neural network architecture: A continuous ID3 algorithm. IEEE Transactions on Neural Networks. 1992 Mar; 3(2):280–91.
31. Venkatesan N, Arasan KA, Muthukumaran S. An ID3 algorithm for performance of decision tree in predicting student's absenteeism in an academic year using categorical datasets. Indian Journal of Science and Technology. 2015 Jul 1; 8(14):1–5. DOI: 10.17485/ijst/2015/v8i14/72730.
32. Joachims T. Making large scale SVM learning practical. Universität Dortmund; 1999 Oct 29.
33. Suganya P, Sumathi CP. A novel metaheuristic data mining algorithm for the detection and classification of parkinson disease. Indian Journal of Science and Technology. 2015 Jul 1; 8(14):1–9. DOI: 10.17485/ijst/2015/v8i14/72685.