JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

HYBRID ARTIFICIAL NEURAL NETWORK AND CASE BASED REASONING (HANNCBR) CLASSIFICATION MODEL FOR CLASSIFICATION OF AGRICULTURE DATA IN DIFFERENT AREA

[1] N. HARSHINI, [2] Dr. M. RATHAMANI
[1]Research Scholar, [2] Assistant professor,
[1, 2] Department of Computer Science,
[1, 2] Nallamuthu Gounder Mahalingam College,
[1, 2] Pollachi, Tamil Nadu, India.

Abstract – Lately the erratic weather conditions changes have prompted different polemics. The utilization of cropping designs that have been done from generation to generation regardless of environmental change and the environment is the reason. The advances in computing and information storage have given tremendous measures of data. The test has been to remove knowledge from this raw data that has lead to new strategies and techniques, for example, data mining that can connect the knowledge gap. This research expected to evaluate these new data mining techniques and apply them to soil nutrients and weather database to layout in the event that significant connections can be found. So in this paper we proposed Hybrid Artificial Neural Network and Case Based Reasoning (HANNCBR) Classification Model for conclude the best crop to be cultivated considering the factors, the soil's mineral substance proportions and weather patterns. The proposed Hybrid Classification Model (HANNCBR) is contrasted and Naïve Bayes and SVM. The proposed Hybrid Classification Model (HANNCBR) performs well other than existing methodologies.

Keyword: Data mining, soil nutrients and weather database, Hybrid Artificial Neural Network, Case Based Reasoning and Classification;

1. Introduction

Developing crops on land appropriate for them is significant for the legitimate development and yield of the crops. It now and again becomes confounding to conclude which crop is reasonable for the soil type and how much respect expect once planted. Additionally because of the changing climatic conditions, mineral items in the soil change and so soil quality. Since Tamil Nadu is a state clearly dynamic in the agriculture area, it makes it important to have a model created to view the best reasonable crop as filled in various pieces of the state to boost creation and supply. This undertaking solely focuses on crops that can be grown in Tamil Nadu by the soil and climatic conditions.

1.1 Data mining

Data mining is characterized as the process in which valuable information is extricated from the raw data. To gain fundamental knowledge extricating enormous measure of data is fundamental. This process of extraction is otherwise called misnomer. As of now in each field, enormous measure of data is available and analyzing entire data is truly challenging as well as it consumes a great deal of time. The prediction analysis is most helpful kind of data which is performed today. To play out the prediction analysis the examples needs to create from the dataset with the machine learning. The prediction analysis should be possible by social event historical information to create future patterns. Thus, the knowledge of what has happened already is utilized to give the best valuation of what will occur in future with prescient analysis.

Data mining is the knowledge discovering procedure by analyzing the potentially interesting and unknown patterns in huge volume of datasets from summing up the helpful information in different points of view. Separated knowledge from various data sets is utilized for better decision making at basic circumstances. Knowledge discover in database (KDD), frequently called data mining, Concentrate knowledge which are remembered for dataset and move that data in to human understandable structure to additional utilization is a target of the Data mining. Three regions are connected with the advancement of techniques in data mining. That is machine learning, artificial insight and statistics.

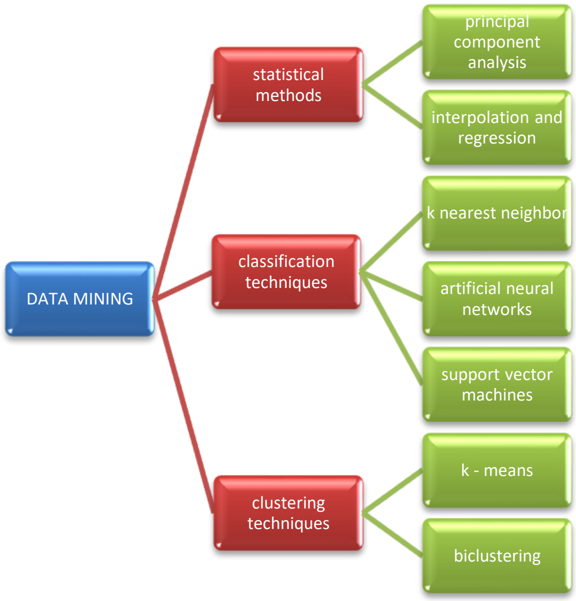


Figure 1.Representation of classification of data mining techniques

1.2 Importance of Data Mining:

Data mining is the significant procedure for assortment of data's in different structures among the data gathered during the time spent data mining incorporates research data, survey data, organization data, competitive data and social media, for example, whatsapp, Facebook. A few stages are engaged with analyzes on selected set of data where the process includes of filtering, transformation, testing, modeling, visualization and documentation is ready and the outcome is outputted (or) the data is put away as needs be in data warehouse or databases. To propose a smart agriculture should predict the yield of crop based on the water, texture of soil and climate. It is fundamental for country to fabricate a huge creation of natural crops. So by applying data mining techniques for agriculture can diminish the expense of food production and further develops efficiency which experiences in more prominent dynamic process in business world.

1.3 Five Major Elements in Data Mining:

- Bring the data and load the data to change onto the warehouse system.
- Store and utilize the data in the database system.
- Make accessible to get to data for researchers, IT professionals and for different organizational analytics.
- Inspect the expected data utilizing suitable software's.
- Plan the data's educate regarding table or graph to address data in a helpful configuration.

1.4 Classification

Classifications problems plan to recognize the qualities that demonstrate the groups to which each case has a place. This pattern can be utilized both to understand the existing data and to predict how new instances will act. For example the organization might need to predict whether the client will answer positively or negatively for the new lunching product. Data mining makes classification models by examining currently grouped data and inductively tracking down a predictive pattern. These existing data might come from a historical database, for example, individuals who have previously gone through a specific medical treatment. Research might come from a trial in which an example of the whole database is tried in real world and the outcomes used to make a classifier.

2. Literature Survey

2.1 Integrating heterogeneous agriculture information using naive Bayes and FCA

Kanchana et.al proposed integrating heterogeneous agriculture information using naive Bayes and FCA. Naive Bayes is utilized in this paper which utilizes probability to classify real and discrete data. The Formal Concept Analysis is utilized in mapping the results of the Naive Bayes and the given data. Agriculture is one of the fields that poor person yet took full advantage of the capability of technology. It is one field where technology and techniques ought to be applied and assist farmers who with giving us fundamental requirements, the food. There are many elements that influence the development of a crop. Every one of the variables must be analyses prior to effective money management on a specific crop. The undertaking expects to utilize data mining method to help farmers. A pattern is analyzed from the enormous data sets and the ongoing production is anticipated based on probability utilizing Naive Bayes algorithm.

2.2 Research of Support Vector Machine in Agricultural Data Classification

Duan et.al proposed The Research of Support Vector Machine in Agricultural Data Classification. As a famous machine learning algorithm, SVM has been broadly utilized in many fields like information retrieval and text classification somewhat recently. In this paper, SVM is introduced to characterize the agricultural data. SVM can plan original input data into a high dimensional component space to look for a different hyper plane, and then it can perform classification by using the constructed N-dimensional hyper plane that preferably disengages the data. SVM has been generally utilized in different fields and it can obtain superior performance in many real world classification applications, for example, image retrieval, cancer recognition, text classification and credit scoring. SVM is introduced to arrange the agricultural data.

2.3 Decision Tree Classifier

Decision Tree is one of the most utilized, practical methodologies for supervised learning. It tends to be utilized to settle both Regression and Classification tasks with the last option being placed more into down to earth application. It is a tree-structured classifier with three kinds of nodes. At long last, the Leaf Nodes address the result. This algorithm is extremely valuable for solving decision-related problems. Decision tree regression sees features of an object and trains a model in the design of a tree to foresee data in the future to convey significant persistent result.

2.4 Random Forest Classifier

A Random Forest is an ensemble technique equipped for performing both regression and classification tasks with the utilization of numerous decision trees and a way called Bootstrap and Aggregation. enerally, it has numerous decision trees as base learning models. Randomly perform line sampling and have sampling from the dataset forming test datasets for each model.

The basic steps involved in Random Forest algorithm is as follows:

Stage 1: Start selecting the random examples from the given training dataset.

Stage 2: Next, this algorithm will construct a decision tree for each example using the decision tree algorithm. Then for every decision tree a result is obtained.

Stage 3: Next voting will be performed for each outcome that is predicted.

Stage 4: Now select the most voted result as the final prediction result.

3. Proposed Methodology

This research plans to recommend the appropriate crop that can be grown in any district of Tamilnadu considering the climatic conditions gathered using Open Weather API and soil ripeness subtleties from the client. Research additionally intends to predict the yield of the particular crop to assist the cultivators with staying prepared. Classification algorithms are implemented to conclude the best crop to be developed considering the variables, the soil's mineral content proportions and weather conditions. The qualities of the gathered datasets are shown beneath.

	N	P	K	temperature	humidity	ph	rainfall
count	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000	2200.000000
mean	50.551818	53.362727	48.149091	25.61624	71.481779	6.469480	103.463655
std	36.917334	32.985883	48.149091	25.616244	22.263812	0.773938	54.958389
min	0.000000	5.000000	5.000000	8.825675	14.258040	3.504752	20.211267
25%	21.000000	28.00000	20.000000	22.769375	60.261953	5.971693	64.551686
50%	37.000000	51.000000	32.000000	25.598693	80.473146	6.425045	94.867624
75%	84.250000	68.000000	49.000000	28.561654	89.948771	6.923643	124.267508
max	140.000000	145.000000	205.000000	43.675493	99.981876	9.935091	298.560117

Table 1.Summary statistics of the data

3.1 Hybrid Artificial Neural Network and Case Based Reasoning (HANNCBR)

The block diagram of the Proposed Model is shown in Figure 2. The block diagram has two important trained machine learning systems. The First is Artificial Neural Network system which involved the Back propagation algorithm for its training. The Second one is Case Based System which utilized the K-Nearest Neighbor Algorithm for its training. In Complete dataset 80% are utilized for training and the remaining 20% utilized for testing reason. In the trying period the new test data will be gone through the trained ANN and the CBR Systems.

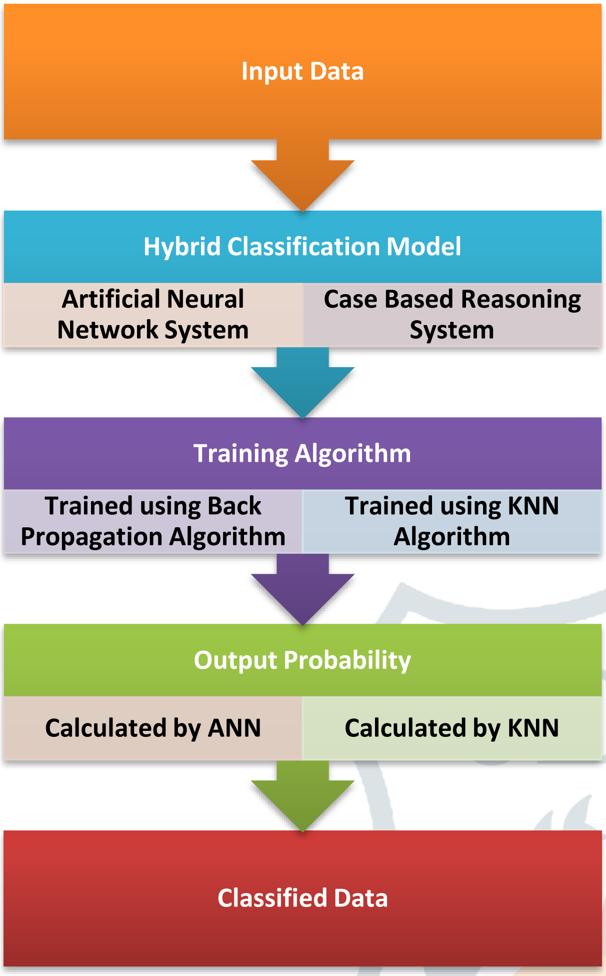

Input Data
Hybrid Classification Model
Artificial Neural Network System
Case Based Reasoning System
Training Algorithm
Trained using Back Propagation Algorithm
Trained using KNN Algorithm
Output Probability
Calculated by ANN
Calculated by KNN
Classified Data

Figure 2.Proposed Hybrid Classification Model (HANNCBR) Workflow

3.2 Artificial Neural Network Structure

The Artificial neural Network utilized the Multi layer feed forward Network architecture and it involved the Back propagation Algorithm for Training. The network has a single hidden layer, an input layer and a output layer. The input layer has 8 input nodes. The hidden layer has 5 neurons and the output layer has a single neuron. Sigmoid function is utilized in the hidden layer and in the output layer. Squared Error is utilized as cost function for error calculation to change the network weight.

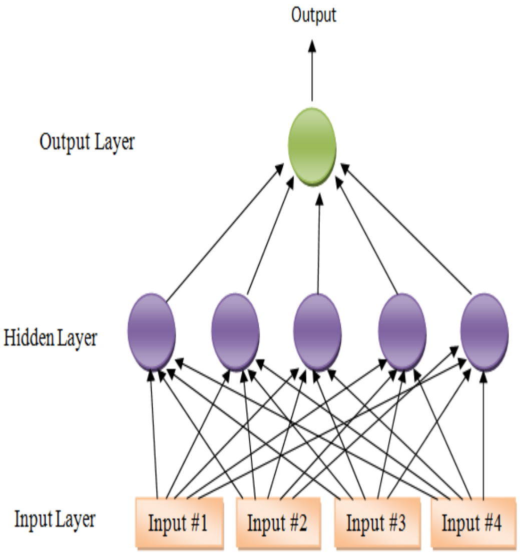


Figure 3.Multilayer feed forward network

Can be utilized to solve complicated problems

Algorithm

Input: dataset D, learning rate, network.

Step 1: *Start the process*

Step 2: *Get the input.*

Step 3: *Measure the input.*

Step 4: *Each input sent to network must be weighted i.e. multiplied by some random value between -1 and +1.*

Step 5: *Sum all the weighted input.*

Step 6: *The output of network is produced by passing that sum through the activation function.*

Step 7: *End the process.*

Output: a trained neural network.

3.3 Case-Based Reasoning

Case-based thinking is a methodology for taking care of issues by using previous encounters. It involves retaining a memory of past problems and their solutions and using it to take care of new problems. When presented with problem case base reasoned searches its memory of previous cases and attempts to find a case that has a similar problem specification as the current case. In the event that the reasoned cannot find an identical case in its case base, it will attempt to find a case or cases in the case base that most intently match the current query case. The CBR involved the K-NN Algorithm for training. The k-nearest neighbor (k-NN) method is a data mining procedure viewed as among the super five methods for data mining. In this research consider every one of the qualities in training set as a different dimension in some space, and take the worth an observation has for this trademark to be its coordinate in that dimension, so getting a bunch of points in space. Research can then consider the comparability of two points to be the distance between them in this space under some suitable metric. The manner by which the algorithm concludes which of the points from the training set are comparative enough to be considered when choosing the class to predict for a new observation is to pick the k nearest data focuses to the new perception, and to take the most widely recognized class among these. To this end it is known as the k Nearest Neighbors algorithm.

The implementation of algorithm can be noted as below:

Step 1: *Start the process.*

Step 2: *Load the data.*

Step 3: *Introduce K to your picked number of neighbors.*

Step 4: *For every model in the data*

> ***Step 4.1:*** *Calculate the distance between the query model and the ongoing model from the data.*
>
> ***Step 4.2:*** *Add the distance and the index of to an ordered collection.* ϖ

Stage 5: *Sort the arranged collection of distances and indices from smallest to biggest (in ascending order) by the distances.*

Stage 6: *Pick the primary K entries from the arranged collection.*

Stage 7: *Get the labels of the chose K passages.*

Stage 8: *If regression, return the mean of the K labels.*

Stage 9: *If classification, return the method of the K labels.*

Stage 10: *End the process.*

Proposed Hybrid Classification Model (HANNCBR) Algorithm

Stage 1: *Start the process.*

Stage 2: *Divide the preprocessed total dataset into two by 80% and 20% and named it as Training Dataset D1 and Testing Dataset D2.*

Stage 3: *Train the Artificial Neural Network System using the Back propagation Algorithm on Training Dataset D1.*

Stage 4: *Train the Case Based Reasoning System utilizing K-Nearest Neighbor Algorithm on Preparing Dataset D1.*

Stage 5: *Then ascertains the combined mean value from the outputs of the ANN and CBR systems. Based on the determined mean value it shows the output.*

Stage 6: *Utilize the Testing dataset D2 to ascertain the classification performance of the Proposed System.*

Stage 7: *Stop the process.*

4. Results Discussion

4.1 Correlation between different features

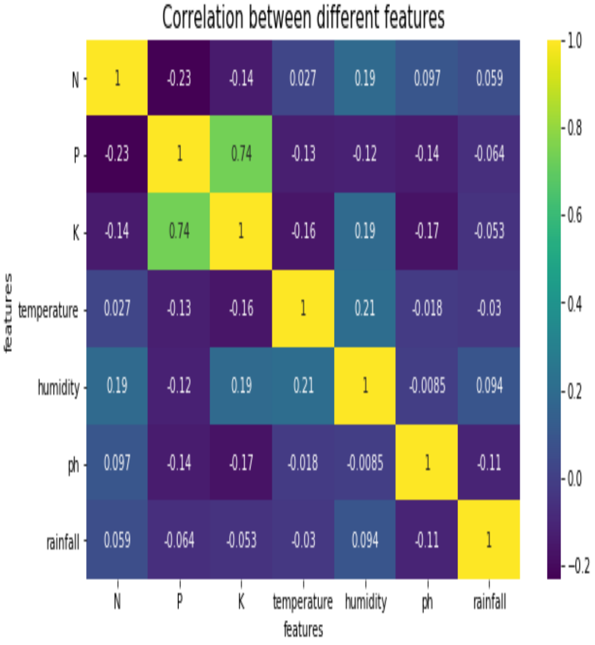
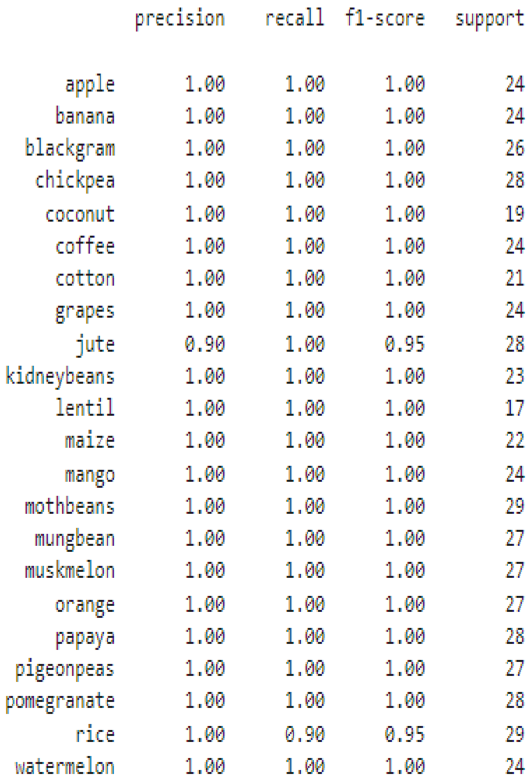


Figure 4.Correlation between different features

According to research proposed method algorithm first research constructed independently an Artificial Neural Network and a Case Based Reasoning System for the preprocessed Dataset. Research partitioned the complete dataset 80% of the training dataset and 20% testing dataset. Research utilized a similar 80% training dataset to train both the ANN and KNN Systems and a similar testing dataset used to test the classification performance of the ANN and KNN Systems. Research took the average probability value based on the probability performance from the ANN and KNN systems.



```
              precision    recall  f1-score   support

       apple       1.00      1.00      1.00        24
      banana       1.00      1.00      1.00        24
   blackgram       1.00      1.00      1.00        26
    chickpea       1.00      1.00      1.00        28
     coconut       1.00      1.00      1.00        19
      coffee       1.00      1.00      1.00        24
      cotton       1.00      1.00      1.00        21
      grapes       1.00      1.00      1.00        24
        jute       0.90      1.00      0.95        28
 kidneybeans       1.00      1.00      1.00        23
      lentil       1.00      1.00      1.00        17
       maize       1.00      1.00      1.00        22
       mango       1.00      1.00      1.00        24
   mothbeans       1.00      1.00      1.00        29
    mungbean       1.00      1.00      1.00        27
   muskmelon       1.00      1.00      1.00        27
      orange       1.00      1.00      1.00        27
      papaya       1.00      1.00      1.00        28
  pigeonpeas       1.00      1.00      1.00        27
 pomegranate       1.00      1.00      1.00        28
        rice       1.00      0.90      0.95        29
  watermelon       1.00      1.00      1.00        24
```

4.2 Precision

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Dataset	Naïve Bayes	SVM	Proposed HANNCBR Model
Dataset 1	67	82	85
Dataset 2	69	79	88
Dataset 3	71	76	90
Dataset 4	76	72	93
Dataset 5	80	70	96

Table 4.Comparison table of Precision

The Comparison table 4 of Precision Values explains the different values of existing Naïve Bayes, SVM and proposed HANNCBR Model. While comparing the Existing algorithm and proposed HANNCBR Model, provides the better results. The existing algorithm values start from 67 to 80, 70 to 82 and proposed HANNCBR Model values starts from 85 to 96. The proposed method provides the great results.

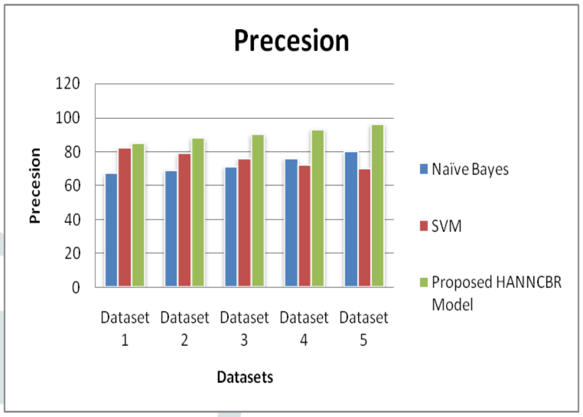


Figure 5.Comparison chart of Precision

The Figure 5 Shows the comparison chart of Precision demonstrates the existing Naïve Bayes, SVM and proposed HANNCBR Model. X axis denote the Dataset and y axis denotes the Precision ratio. The proposed HANNCBR Model values are better than the existing algorithm. The existing algorithm values start from 67 to 80, 70 to 82 and proposed HANNCBR Model values starts from 85 to 96. The proposed method provides the great results.

4.3 Recall

$$Recall = \frac{True\ Positives}{True\ Positives\ + False\ Negatives}$$

Dataset	Naïve Bayes	SVM	Proposed HANNCBR Model
Dataset 1	0.76	0.82	0.84
Dataset 2	0.79	0.78	0.87
Dataset 3	0.82	0.75	0.90
Dataset 4	0.85	0.72	0.94
Dataset 5	0.87	0.70	0.97

Table 5.Comparison table of Recall

The Comparison table 5 of Recall Values explains the different values of existing Naïve

Bayes, SVM and proposed HANNCBR Model. While comparing the Existing algorithm and proposed HANNCBR Model, provides the better results. The existing algorithm values start from 0.76 to 0.87, 0.70 to 0.82 and proposed HANNCBR Model values starts from 0.84 to 0.97. The proposed method provides the great results.

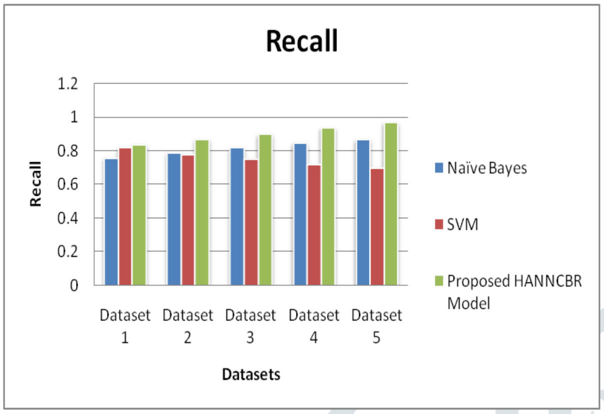


Figure 6.Comparison chart of Recall

The Figure 6 Shows the comparison chart of Recall demonstrates the existing Naïve Bayes, SVM and proposed HANNCBR Model. X axis denote the Dataset and y axis denotes the Recall ratio. The proposed HANNCBR Model values are better than the existing algorithm. The existing algorithm values start from 0.76 to 0.87, 0.70 to 0.82 and proposed HANNCBR Model values starts 0.84 to 0.97. The proposed method provides the great results.

3. F-Measure

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

Dataset	Naïve Bayes	SVM	Proposed HANNCBR Model
Dataset 1	0.79	0.84	0.96
Dataset 2	0.77	0.81	0.95
Dataset 3	0.76	0.79	0.93
Dataset 4	0.73	0.78	0.91
Dataset 5	0.70	0.75	0.90

Table 6.Comparison table of F -Measure

The Comparison table 6 of F -Measure Values explains the different values of existing Naïve Bayes, SVM and proposed HANNCBR Model. While comparing the Existing algorithm and proposed HANNCBR Model, provides the better results. The existing algorithm values start from 0.70 to 0.79, 0.75 to 0.84 and proposed HANNCBR Model values starts from 0.90 to 0.96. The proposed method provides the great results.

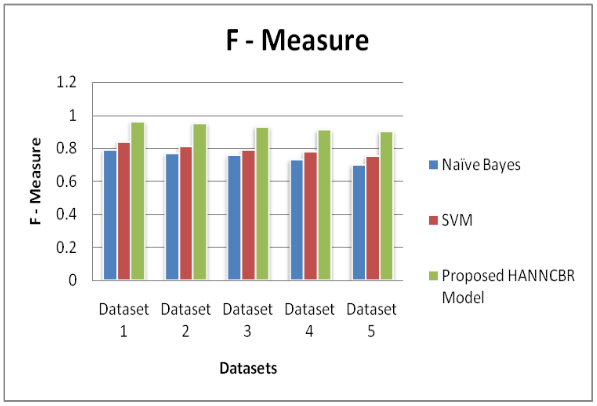


Figure 6.Comparison chart of F -Measure

The Figure 6 Shows the comparison chart of F -Measure demonstrates the existing Naïve Bayes, SVM and proposed HANNCBR Model. X axis denote the Dataset and y axis denotes the F - Measure ratio. The proposed HANNCBR Model values are better than the existing algorithm. The existing algorithm values start from 0.70 to 0.79, 0.75 to 0.84 and proposed HANNCBR Model values starts from 0.90 to 0.96. The proposed method provides the great results.

4. Accuracy

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ preditions}$$

Dataset	Naïve Bayes	SVM	Proposed HANNCBR Model
Dataset 1	84	78	97
Dataset 2	81	75	95
Dataset 3	79	72	92
Dataset 4	76	70	90
Dataset 5	72	68	88

Table 7.Comparison table of Accuracy

The Comparison table 7 of Accuracy Values explains the different values of existing Naïve Bayes, SVM and proposed HANNCBR Model. While comparing the Existing algorithm and proposed HANNCBR Model, provides the better results. The existing algorithm values start from 72 to 84, 68 to 78 and proposed Enhanced PCA values starts from 88 to 97. The proposed method provides the great results.

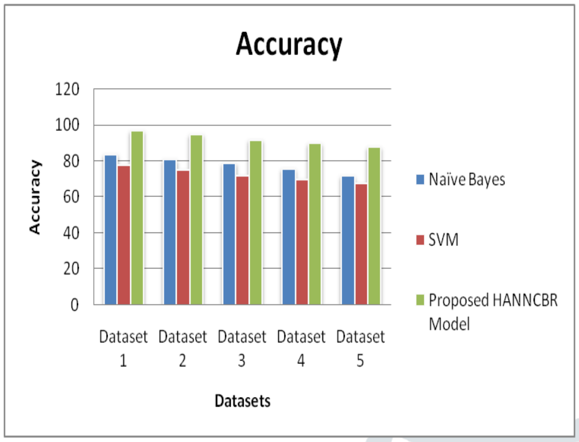


Figure 7.Comparison chart of Accuracy

The Figure 7 Shows the comparison chart of Accuracy demonstrates the existing Naïve Bayes, SVM and proposed Enhanced PCA. X axis denote the Dataset and y axis denotes the Accuracy ratio. The proposed HANNCBR Model values are better than the existing algorithm. The existing algorithm values start from 72 to 84, 68 to 78 and proposed HANNCBR Model values starts from 88 to 97. The proposed method provides the great results.

5. Conclusion

Agriculture is an intuitive space which is typically transformed starting with one generation then onto the next generation. Data mining involves the systematic analysis of enormous data sets, and data mining in agricultural datasets is exciting and modern research region. Exact information in characterizing crops relies upon climatic, geographical, biological and different factors. These are vital inputs to generate characterization and prediction models in data mining. Classifying agricultural data is a very monotonous undertaking for its utilization in reality applications. In this paper, a new algorithm called hybrid machine learning algorithm HANNCBR is proposed which is made by similar study of individual algorithms to attain exact classification of cotton data set. The proposed functionalities improvise the presentation when contrasted with individual algorithms and it helps with breaking down the order to the most limit specification.

References

1. Kaushik Bhagawati, AmitSen, Kshitiz Kumar Shukla, Rupankar, Bhagawati, Application and scope of data mining in agriculture, International Journal of Advanced Engineering Research and Science, 3(7) 2016,66-69
2. R. S. Kodeeshwari, K. Tamil Ilakkiya ,Different Types of Data Mining Techniques Used in Agriculture - A Survey International Journal of Advanced Engineering Research and Science (IJAERS), ISSN: 2349-6495(P) | 2456-1908(O)
3. Namita Mirjankar, Smitha Hiremath , Application of Data Mining In Agriculture Field ,International Journal of Computer Engineering and Applications, iCCSTAR-2016, Special Issue, May2016.
4. Hetal Patel, Dharmendra Patel,A Brief survey of Data Mining Techniques Applied to Agricultural Data International Journal of Computer Applications (0975 – 8887) Volume 95– No. 9, June 2014 6.
5. Ramesh Babu Palepu, Rajesh Reddy Muley, An analysis of agricultural soils by using Data Mining techniques, 2017 IJESC Volume 7 Issue no 10
6. Alberto Tellaeche1, Xavier-P. Burgos Artizzu, Gonzalo Pajares, and Angela Ribeiro, A Vision-Based Hybrid Classifier for Weeds Detection in Precision Agriculture Through the Bayesian and Fuzzy k-Means Paradigms, Advances in Soft Computing, vol 44
7. Yang, C.C., Prasher, S.O. Landry, J.A. and Ramaswamy, H.S.: Development of an Image Processing System and a Fuzzy Algorithm for Site-specific Herbicide Applications. Precision Agriculture, 4 (2003) 5–18.
8. Verheyen K, Adrianens M, Hermy S Deckers. High resolution continuous soil classifcation using morphological soil profle descriptions. Geoderma. 2001;101:31–48.
9. Gonzalez-Sanchez Alberto, Frausto-Solis Juan, Ojeda-Bustamante W. Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res. 2014;12(2):313–28.

10. Pantazi XE, Moshou D, Alexandridis T, Mouazen AM. Wheat yield prediction using machine learning and advanced sensing techniques. Comput Electron Agric. 2016;121:57–65.
11. Veenadhari S, Misra B, Singh D. Machine learning approach for forecasting crop yield based on climatic parameters. In: Paper presented at international conference on computer communication and informatics (ICCCI-2014), Coimbatore. 2014.
12. Rahmah N, Sitanggang IS. Determination of optimal epsilon (Eps) value on DBSCAN algorithm to clustering data on peatland hotspots in Sumatra. IOP conference series: earth and environmental. Science. 2016; 31:012012.
13. Forbes G. The automatic detection of patterns in people's movements. Dissertation, University of Cape Town. 2002.
14. Ng RT, Han J. CLARANS: A Method for Clustering Objects for Spatial Data Mining. In: IEEE Transactions on Knowledge and Data Engineering. 2002; 14(5).
15. Gleaso CP. Large area yield estimation/forecasting using plant process models. Paper presentation at the winter meeting American society of agricultural engineer's palmer house, Chicago, Illinois. 1982; 14–17.