ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 6, Issue 1, January 2017

Classification Algorithms in Data Mining – A Survey

S.Ponmani, Roxanna Samuel, P.VidhuPriya

***Abstract*— Data Mining or Knowledge Discovery is the latest emerging trend in the information technology. It is the process of analyzing data from different perspectives and summarizing it into useful information. One of the function of data mining is classification, is a process of generalizing data sets based on different instances. There are various classification techniques which help as to group the data sets. Some the algorithms that this paper will be analyzing are Linear Regression, Multi Layer Perceptron, CART, J48, C4.5, ID3, Random forest and KNN.**

***Index Terms*— Data Mining, Classification, Linear Regression, Multi Layer Perceptron, CART, J48, C4.5, ID3, Random forest and KNN.**

I. INTRODUCTION

Data Mining is extraction of unknown information from huge data bases (Data Warehouse). It is a powerful new technology with great potential to help companies focus on the most important information. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [1]. These techniques can be implemented on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line.

II. CLASSIFICATION ALGORITHMS

Classification is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis [2]. Also called sometimes called a Decision Tree, classification is one of several methods intended to make the analysis of very large data sets effective. To create an effective set of classification rules which answers a query, makes decision based on the query and predicts the behavior. To begin with a set of training data sets are created with certain set of attributes or outcomes.The main objective of the classification algorithm is to mine, how that set of attributes reaches its conclusion.

***S.Ponmani**, Computer Science, Rajalakshmi Engineering College Chennai, India, 9444272578*
***Roxanna Samuel**, Computer Science, Rajalakshmi Engineering College, Chennai, India, 9710482686.*
***P. Vidhupriya**, Computer Science, Rajalakshmi Engineering College, Chennai, India, 9842955808.*

A. Linear Regression

[3,4] Linear regression is a linear model, that assumes a linear relationship between the input variables and an output variable. To learn or train the linear regression model, estimate the coefficients values used in the representation for the available data. When there is a single input variable, then the method is known as simple linear regression. When there are multiple input variables, then the method is known as multiple linear regressions.

Different techniques can be used to prepare or train the linear regression equation from data. Some of them are Ordinary Least square, Gradient descent, and Regularization. Among this Ordinary Least Squares is one of the most common techniques. It seeks to minimize the sum of the squared residuals. This means that given a regression line through the data, calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seek to minimize. Regularization methods are the extensions of the training of the linear model. These methods seek to both minimize the sum of the squared error of the model on the training data and also to reduce the complexity of the model. The pseudocode of the Linear Regression is as follows.

Algorithm: Linear Regression

1. Define a training set of the form
   S = {yi, xi1……, xip)},
   Where xi → p vector of
   Independent/Explanatory variable
   yi → dependent/Response
   variable
2. Use linear equation to set up the relationship between x,y
   The equation is of the form :
   $yi = \beta 1xi1 + ... + \beta pxip + \varepsilon i$, i = 1,...,n
   where βi --> regression coefficient
   εi --> error variable / noise
3. Estimate the value of regression coefficient using some technique. Eg: Ordinary Least Square
4. Use the trained linear model for prediction and estimation

B. Multilayer perceptrons

MultiLayer Perceptron technique was introduced by Werbos in 1974 and Rumelhart, McClelland, Hinton in 1986 also named feed forward networks. In machine learning,

MLP is a powerful learning algorithm. MLP is mainly used to solve non-linear problem with good quality solution[5]. It is suitable for regression and classification. MLPs are standard tool for establishing relationship between data in many real world problems in the absence of a parametric model[7]. In MLP the problem is converted into finite directed acyclic graph which contain n number of inputs, hidden and output nodes[6]. Here the parameters are measured based on weightage of each node and unlabeled patterns are estimated by using hidden layer concept.

Among the many approaches available, here minimization approach is used for adjusting the weight in the hidden layer. Minimization approach selects the minimum weight among the updated weight which is used for updating the weight of the hidden layer.

Algorithm: MuliLayer Perceptrons

1.Choose an initial weight vector w

2.Initialize the weight of the hidden layer using minimization approach

While error did not converge do

For all vector$\forall$(x,d) $\in$ D do

Apply x to network and calculate the network output

For all weights calculate the deviation in weight.

Perform one update step of the minimization approach

End for

End while

C. Cart

Cart (Classification and Regression Trees and also designated as C&RT) algorithm is based on data enquiry and prediction algorithm. Cart was developed in the year early 1980's. It generate classification or regression trees, whether the variable is categorical or numerical.. It can also handle missing values effectively. Cart is different from other Hunt based algorithm. Binary tree generated by cart is referred as Hierarchical Optimal Discriminate analysis (HODA)[10].

Cart uses binary decision tree algorithm that recursively partition data into two homogeneous subsets. Here, we are using splitting rule [11] for constructing a classification tree. Splitting attribute can be done using Gini index splitting measure. All splits are selected using the twoing criteria and the result tree is pruned by the method of Cost-complexity pruning. It also work with missing values after tree growth [12].An important feature of CART is its ability to generate regression trees [9]

Algorithm:CART[8]

1. A tree building is done using recursive splitting rule
2. Build a "maximal" tree based on some 'stopping rule': a "maximal" tree has been produced, which probably greatly overfits the information contained within the learning dataset.
3. Optimal tree selection: the tree which fits the information in the learning dataset, but does not overfit the information, is selected from among the sequence of pruned trees.

D. J48

As J48 is an open source Java implementation of the C4.5 decision tree algorithm [15][16][18]. A decision tree is actually a predictive machine-learning model [19], which decides the dependent variable (ie, Target value) based on various attributes of the available training data set. The internal nodes of a decision tree denotes varied attributes, the connecting branches of various nodes give us the likely values of the attributes and the terminal node states the classification of the dependent variable.

J48 Decision Tree Classifier uses two phases [13][19].

1. Tree construction
   a. Starts with the whole data set at the root
   b. Check the attribute of the data set and partition them based on the following cases
      Case I: - If the attribute value is clear and has a target value, then it terminates the branch and assigns the value as Target value (classification)
      Case II: - If the attribute, gives the highest information, then continue till we get a clear decision or run out of attributes.
      Case III:- If we run out of attributes or we are presented with ambiguous result, then assign the present branch as target value.
      Case IV:- ignore missing values[14][17].
2. Tree Pruning
   Identify and remove branches that reflects noise and outliers to reduce classification errors [13][20].

```
Algorithm:J48
INPUT:D//Training data
OUTPUT:T//Decision Tree
DTBUILD (*D)
{
        T= ∞;
        T= Create root node and label with splitting
attribute:
        T= Add arc to root node for each split predicate
and label;
        For each arc do
                D = Database created by applying
        splitting Predicate to D;
                If stopping point reached for this path,
        then
                        T'= create leaf node and label
                with appropriate class;
                Else
                        T'= DTBUILD(D);
        T= add T' to arc;
}
```

E. C4.5

Decision tree algorithms are easy to understand and implement. C4.5 is used for constructing decision tree. ID3 and C4.5 uses Shannon entropy. From the constructed decision tree decision rules can be formulated. In [21] they have implemented C4.5 using weka tool with freely available data sets. It is concluded that C4.5 works efficiently with noisy and missing data. Also it is concluded that c4.5 provides scalability. In[22], they have worked out ID3 and C4.5 using weather data set. In [23], they have constructed efficient C4.5 by using three strategies: Quick sort, Counting Sort and RainForest algorithm. They have applied the above strategies to find local threshold in case of continuous data. In [24] student performance is predicted by applying C4.5

```
Algorithm: C4.5
Let A={ a1,a2,a3,..,an }be the attributes that need to be
classified.
While(∀ ai ∈ same class C)
{
   Compute information gain for each attribute;
   Droot=max_gain(ai) where i=1,2,3,..,n
   A=A-ai
}
```

F. ID3

ID3(Iterative Dichotomiser 3) is a classification algorithm that are used for the classification of untrained data by constructing decision tree. It is a Greedy algorithm[27]. Given a set of attributes, ID3 selects one of the attribute as the root with the help of information gain. i.e. attribute with the highest information gain is selected as root. It may convert the continuous attribute to discrete attribute. In [25], ID3 is implemented for weather database. Student performance is evaluated[26] using ID3.

Algorithm is best suited for [27]

1. Instance is represented as attribute-value pairs.
2. Target function has discrete output values.
3. Attribute values should be nominal.

Ways of handling missing attribute [27]

1. Considering the missing value as the new value of the attribute.
2. Leaving the instance with the missing value.
3. Replacing the missing value with some common value.

These are evaluated by implementing ID3 using weather database. In [28],[29], ID3 is implemented using Harved Charvat Entropy and performance is evaluated using UCI Machine Learning Repository.

```
Algorithm:ID3
Let A={ a1,a2,a3,..,an }be the attributes that need to be
classified.
While(∀ai ∈ same class C)
{
   Droot=attribute that best classifies(ai);
   A=A-ai
}
```

G. Random Forest

Random forests or Random decision forests are an ensemble learning method. It is used mainly used to solve classification, regression problems and also other problems. Random forest is one of the accurate learning algorithm. The basic concept of the algorithm is to build many small decision-tree and then merging them to form a forest. It is computationally easy and cheap process to build many such small and weak decision trees. So such decision trees can be formed in parallel and then it can be combined to form a single and strong forest. The algorithm for random forests uses the common technique of bootstrap bagging. [31] Given a training set S = {(x1, y1),……, (xn, yn)}, bagging repeatedly (B times) selects a random sample from the training set and construct trees to fit these samples. This procedure leads to better performance that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. The psuedocode [30] for the Random Forest is as follows.

```
Algorithm: Random Forest
Precondition:
      A training set S = {(x1, y1),……, (xn, yn)}, where xi →
input vector, yi → class
      F → Features,
      B → number of trees in forest
Outcome: H → a forest of B trees
function RandomForest(S, F)
   H ← Ø
   for i ϵ 1, . . . . , B do
      S (i)  ← A bootstrap sample from S
       hi  ← RandomizedTreeLearn(S (i), F)
       H  ← H υ {hi}
   end for
   return H
end function
function RandomizedTreeLearn(S , F)
   At each node:
      f ← very small subset of F
      Split on best feature in f
   return The learned tree
end function
```

The pseudocode works as follows: for each tree in the forest, select a bootstrap sample S(i) from the training set S. Then construct a decision tree for the selected sample. The pseudocode for constructing the decision works as follows: for each node of the tree, select a very small subset of features f from F where F is the complete set of features. It is computationally expensive process to decide which features to be selected for the decision tree learning. But by narrowing the set of features, the learning process becomes very fast. [3] Then to classify a new object, apply the input vector to each

tree in the forest and each tree will give a particular class as output. The forest will choose the class with the most votes.

H. K-Nearest Neighbor (KNN)

KNN classifier is an instance-based learning Algorithm which is based on a distance function for pairs of observations, such as the Euclidean distance or Cosine. In this paradigm, k nearest neighbors of a training data is computed first. Then the similarities of one sample from testing data to the k nearest neighbors are aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class.[32]
KNN is an *non parametric lazy learning* algorithm. That is a pretty concise statement. When you say a technique is non parametric , it means that it does not make any assumptions on the underlying data distribution. This is pretty useful , as in the real world , most of the practical data does not obey the typical theoretical assumptions made (eg gaussian mixtures, linearly separable etc) . A Precise information about the algorithms discussed in this paper are given in Table 1.

Algorithm:KNN
Assign random weight wi to each instance xi in the training set
Divide the number of training examples into N sets
Train the weights by cross validation
For every set Nk in N, do
Set Nk = Validation Set
For every example xi n For every example xi in N such that x in N such that xi does not belong
to does not belong to Nk do
Find the K nearest neighbors based on the Euclidean distance
Calculate the class value as
$\sum$ wk X xj,k where j is the class attribute
If actual class != predicted class then apply gradient descent
Error = Actual Class – Predicted Class
For every Wk
Wk = Wk + α X Error
Calculate the accuracy as
Accuracy = (# of correctly classified examples / # of examples in Nk) X 100 [1]

TABLE 1 APPLICATIONS, PROS AND CONS OF VARIOUS CLASSIFICATION TECHNIQUES

S.No	Algorithm	Applications	Advantages	Disadvantages
1	Linear Regression	1.Epideiology 2.Finance 3.Economics 4.Environmental Science	1. Linear regression implements a statistical model that, when relationships between the independent variables and the dependent variable are almost linear, shows optimal results.	1. Linear regression is often inappropriately used to model non-linear relationships. 2. Linear regression is limited to predicting numeric output. 3. A lack of explanation about what has been learned can be a problem.
2	Multi Layer Perceptron	1. Neural Network-Architecture Selection. 2.Ocean Surveillance and remote sensing. 3.Change Detection-Dynamic Behaviour	1. Used for solving non-linear problem by converting it to linear problem. 2. It gives best quality solution.	1. Requires a help of other algorithms. 2. Number of hidden layers in the network is user dependent.
3	CART	1.Financial sector 2.Agriculture 3.Biomedical Engineering	1.In-built features that deal with missing attributes. 2.Variable selection can be done automatically	1. Data is sorted at every node to determine the best splitting point. 2. The linear combination splitting criteria is used during the regression analysis.

4	J48	1.Medical 2.Iris 3.weather 4.Bank data set 5.Soil Fertily	1. All datas are examined and categorized 2. Larger programs are split into more than one class	1. Ignores missing values 2. computation is slower
5	C4.5	Test clinical data	1. It produces the accurate result. 2. It takes the less memory to large program execution. 3. It takes less model build time. 4. It has short searching time.	1. Empty branches. 2. Insignificant branches. 3. Over fitting.
6	ID3	Weather	1. Works well for data set without continuous data and missing values.	1.ID3 favours attribute with many values because such attributes will have high information gain. 1. It leaves some attribute and constructs the decision tree. 2.
7	Random forest	1.Remote Sensing 2.Medical	1.Can handle large set of data with high dimensionality 2. Useful in the case of missing data	over1. Fit for some datasets with noisy classification /regression tasks. 2. Classifications made by random forests are difficult to interpret.
8	KNN	1.Text Mining 2.Agriculture 3.Medicine 4.Finance	1. It is an easy to understand and easy to implement classification technique. 2. Training is very fast. 3. Robust to noisy training data. 4.It is particularly well suited for multi- modal classes	1. It is sensitive to the local structure of the data. 2. Memory limitation. 3. Being a supervised learning lazy Algorithm i.e., runs slowly.

I. CONCLUSION

This paper deals with various classification techniques used in data mining and a study on each of them. Data mining is a wide area that integrates techniques from various fields including machine learning, artificial intelligence, statistics and pattern recognition, for the analysis of large volumes of data. Classification methods are typically strong in modeling interactions Each of these me

thods can be used in various situations as needed where one tends to be useful while the other may not and vice-versa. These classification algorithms can be implemented on different types of data sets like data of patients, financial data according to performances. Hence these classification techniques show how a data can be determined and grouped when a new set of data is available. Each technique has got its own pros and cons as given in the paper. Based on the needed Conditions each one as needed can be selected On the basis of the performance of these algorithms, these algorithms can also be used to detect the natural disasters like cloud bursting, earth quake, etc

REFERENCES

[1] Lei Zhang and Fengchun Tian "Performance Study of Multilayer Perceptrons in a Low-Cost Electronic Nose" IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, VOL. 63, NO. 7, pp1670-1679 , July 2014.

[2] Raúl Vicen-Bueno, Rubén Carrasco-Alvarez, María Pilar Jarabo-Amores, José Carlos Nieto-Borge, and Enrique Alexandre "Detection of Ships in Marine Environments by Square Integration Mode and Multilayer Perceptrons" IEEE TRANSACTIONS ON

INSTRUMENTATION AND MEASUREMENT, VOL. 60, NO. 3, pp 712-724 Mar 2011.
[3] Matteo Pardo and Giorgio Sberveglieri "Remarks on the Use of Multilayer Perceptro for the Analysis of Chemical Sensor Array Data"IEEE SENSORS JOURNAL, VO 4, NO. 3, pp 355-363., JUNE 2004
[4] Roger J. Lewis "An Introduction to Classification and Regression Tree (CART) Analysis" UCLA Medical Center Torrance, California Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.
[5] Sonia Singh Priyanka Gupta "COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY" International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682 Vol.27, No.27, Jul 2014.
[6] Mr. Brijain R Patel, 2Mr. Kushik K Rana "A Survey on Decision Tree Algorithm Fo Classification" International Journal of Engineering Development and Research © 201 IJEDR | Volume 2, Issue 1 | ISSN: 2321-9939
[7] Hardeep Kaur* Harpreet Kaur "Proposed Work for Classification and Selection of Be Saving Service for Banking Using Decision tree Algorithms" International Journal Advanced Research in Computer Science and Software Engineering Research Paper
[8] Milija Suknovic · Boris Delibasic ·Milos Jovanovic · Milan Vukicevic ·Dragana Becejski-Vujaklija · Zoran Obradovic "Reusable components in decision tree induction Algorithms" Received: 18 February 2009 / Accepted: 5 February 2011 © Springer-Verlag 2011
[9] Anshul Goyal and Rajni Mehta, "Performance Comparison of Naïve Bayes and J4 Classification Algorithms", International Journal of Applied Engineering Researc Vol.7, No.11(2012).
[10] Tina R. Patil, Mrs S S Sherekar, "Performance Analysis of Naïve Bayes and J4 Agorithm for Data Classification", International Journal of Computer Science An Applications, Vol.6, No.2, Apr 2013, pp. 256-261.
[11] S.Singaravelan, D.Murugan and R.Mayakrishnan," Analysis of Classification Algorithms J48 and Smo on Different Datasets", World Engineering & Applied Science Journal, Vol.6, No.2, 2015, pp.119-123.
[12] Rohit Arora and Suman, "Comparative Analysis of Classification Algorithms o Different Datasets using WEKA", International Journal of Computer Application Vol.54, No.13, Sept 2012, pp.21-25.
[13] Sunita Joshi, Bhuwaneshwari Pandey and Nitin Joshi, "Comparative Analysis of Naïv Bayes and J48 Classification Algorithms", International Journal of Advanced Researc in Computer Science and Software Engineering, Vol.5, No.12, Dec 2015, pp.813-817.
[14] Jay Gholap, "Performance Tuning of J48 Algorithm for Prediction of Soil Fertility" Asian Journal of Computer Science and Information Technology, 2:8(2012) 251-252.
[15] Bhuvaneswari .T, Prabaharan. S and Subramaniyaswamy. V,"An Effective Prediction Analysis Using J48", ARPN Journal of Engineering and Applied Sciences, Vol.10, No.8, May 2015, pp.3474-3480.
[16] Gaganjot Kaur and Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications, Vol.98, No.22, July 2014, pp.13-17.
[17] Harvinder Chauhan, Anu Chauhan "Implementation of decision tree algorithm c4.5", International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013
[18] Badr HSSINA, Abdelkarim MERBOUHA,Hanane EZZIKOURI,Mohammed ERRITALI "A comparative study of decision tree ID3 and C4.5", (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications
[19] Salvatore Ruggieri "E Æcient C4.5"
[20] Ping Gu and Qi Zhou "Students performance prediction based on Improved C4.5 decision tree algorithm",
[21] Rupali Bhardwaj , Sonia Vatta "Implementation of ID3 Algorithm", Volume 3 Issue 6, June 2013, ijarcsse.
[22] Hitarthi Bhatt et al "Use of ID3 Decision Tree Algorithm for Placement Prediction", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4785-4789 [database of students details].
[23] Anand Bahety, Department of Computer Science.
[24] "Extension and Evaluation of ID3 – Decision Tree Algorithm"
[25] Nishant Mathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal, "The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree", International Journal of Information and Electronics Engineering, Vol. 2, No. 2, March 2012
[26] Sandeep Kumar, Prof. Satbir Jain "Intrusion Detection and Classification using Improved ID3 Algorithm of Data Mining" International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 5, July 2012[UCI Machine Learning Repository]
[27] Nancy.p ,Dr.GeethaRamani R, "A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data", International Journal of Computer Applications (0975 – 8887) Volume 32– No.8, October 2011
[28] http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf
[29] https://en.wikipedia.org/wiki/Random_forest
[30] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.
[31] http://databases.about.com/od/datamining/g/classification.htm.
[32] https://en.wikipedia.org/wiki/.
[33] http://machinelearningmastery.com/linear-regression-for-machine-learning

Authors

S.Ponmani is M.E., (software Engineering) graduate . She has 15 y experience in teaching. She is Currently working as a assistant profe Rajalakshmi Engineering college. Her area of interest is Data mining

Roxanna Samuel is M.E.,(Computer Science and Engineering)graduate. She has eight years of experience in teaching. She is currently working as a assistant professor in Rajalakshmi Engineering college. Her area of interest is Datamining , Security.

P.Vidhupriya is M.E.,(Computer Science and Engineering). She has 3years of industrial experience and around 3 years of teaching experience. She is currently working as a assistant professor in Rajalkshmi Engineering College. Her area of interest is Machine learning , Data mining.