A Review of Machine Learning Techniques using Decision Tree and Support Vector Machine

Madan Somvanshi
Department of Information Technology,
Pimpri Chinchwad College of engineering,Pune,India
Email: madansomvanshi11@gmail.com

Shital Tambade
Department of Information Technology,
Pimpri Chinchwad College of engineering,Pune,India
Email: shitaltambade1995@gmail.com

Pranjali Chavan
Department of Information Technology,
Pimpri Chinchwad College of engineering,Pune,India
Email: pranj9804@gmail.com

S.V. Shinde
Department of Information Technology,
Pimpri Chinchwad College of engineering,Pune,India
Email: swaatii.shinde@gmail.com

Abstract— In this paper, the brief survey of data mining classification by using the machine learning techniques is presented. The machine learning techniques like decision tree and support vector machine play the important role in all the applications of artificial intelligence. Decision tree works efficiently with discrete data and SVM is capable of building the nonlinear boundaries among the classes. Both of these techniques have their own set of strengths which makes them suitable in almost all classification tasks.

Keywords—classification; machine learning; decision tree; id3; support vector machine; kernel

I.INTRODUCTION

Statistically today we can see a huge amount of data scattered which again is collected to gain knowledge out of it. Knowledge is something that helps organizations to come out with certain results based on whatever data they have collected, and this result help them in improving their business and also helps in knowing the current market demands. The data collected comes from various sources like social networking sites, surveys, etc. All this data is in the digital form and is stored on some or the other databases, now these databases become the sources for knowledge recovery .Data warehousing, data mining are some terminologies used in this scenario. Data warehouse is the repository for storing the data; it is quite different as compared to our normal databases. It uses the star schema, etc as the design of actual databases, and is capable of storing much more data. In Data Mining certain techniques are used for extracting useful data from these data warehouses and then used for knowledge recovery. A prior step to it is to first organize these data properly, and categorize it. This is where the concept of classification and regression comes into picture.

II.MOTIVATION

The major challenges in Classification is the selection of proper attributes and parameters based on the datasets that would be classified, and also after a huge number of calculations and selection of parameters the accuracy in classification is not achieved, hence we are proceeding with the supervised learning techniques like support vector machine, decision tree algorithm and clustering techniques, so that we could provide a novel system that can classify the data more properly and accurately.

III.MACHINE LEARNING

Machine Learning is the ability of machines to learn, where a machine is built up using certain algorithms through which it can take its own decisions and provide the result to the user. Basically it is considered the subfield of Artificial Intelligence. Today Machine Learning is used for complex data classification and decision making [16]. In simple terms it is the development of algorithms that enables the system to learn, and to make necessary decisions. It has strong ties to mathematical optimization that delivers methods, theory and application domain to the field and, it is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Certain examples applications are Spam filtering, optical character recognition (OCR), Search Engines and Computer

Vision. Machine Learning methods and tasks are broadly divided into three categories as follows.

- Supervised Learning
- Un-Supervised Learning
- Reinforcement Learning

A. Supervised Learning

In this type of learning the system is provided with a sample inputs and it is mapped with the output. In this type of learning, each example is a pair consisting of an input object (basically a vector) and a desired output value (supervisory signal). A supervised learning algorithm analyses and studies the training data and produces an inferred function that can be used for mapping new examples. The optimal scenario will allow the algorithm to correctly determine the class labels for unseen instances. It is required by the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way. Approaches for Supervised Learning are Support Vector Machines, Decision Trees, etc. [1].

B. Un-Supervised Learning

In this type of learning the system is provided with some sample inputs but there is no any output present.Since there is no desired output over here categorization is done so that the algorithm differentiates correctly between the data sets. It is a task of defining a function to describe hidden structure from unlabelled data. Since samples or training sets given to the learner are unlabelled, there is no error to reward signal to evaluate a potential solution. In this way unsupervised learning differs from supervised learning and reinforcement learning. It is closely related to the problem of Density Estimation and statistics [1].

C. Reinforcement Learning

Reinforcement learning is a sub domain of machine learning inspired by behaviourist psychology, dealing with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward. It is studied and used in many theories like game theory, control theory, operations research, information theory, swarm intelligence, statics and genetic algorithms [1].

This paper is majorly focusing on studding different algorithms of machine learning which would help system to accurately classify the data and would enable it for decision making in complex situations. The two main algorithms to study are decision tree algorithm and support vector machines.

IV.DECISION TREE FOR CLASSIFICATION

Data Mining is the huge domain to study which is all about extracting patterns, classifying huge and uncertain data, where data is of the heterogeneous forms like text, audio, video, etc. Sometimes the data provided may be incomplete, noisy, damaged, etc. Classification is one of the techniques to handle with this type of data. Decision tree can give the approximate solution to the data which is used in data mining and machine learning [3], [4].There are different algorithms which can be used neural network, etc. From these algorithms Decision trees in data mining for classification of data such as statistics neural network, etc. From these algorithms Decision trees is one of the most useful and powerful algorithm in data mining. It is able to handle number of input data like as nominal, numerical and alphabetical and this is the benefit of it [4].This algorithm can process the data which contains the missing values and errors. This type of contents may vary in number of platform and various packages of data. By using the decision rules decision tree are used to extract a data from large amount of available datasets. Decision tree simply classifies the data which can be easily stored and further it can be easily classified again. In this paper we describe different algorithms for classification of data using decision tree. The following example shows working of simple decision tree algorithm [4].

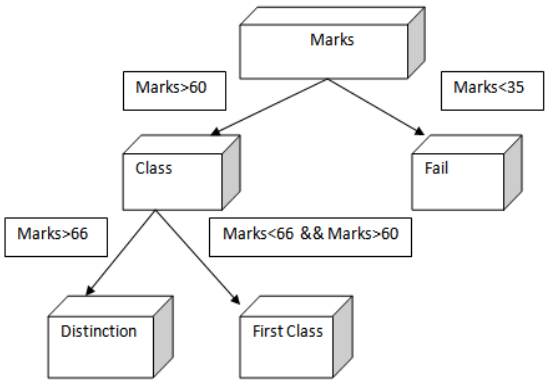

Figure 1: Example of Decision Tree

In above example simple decision tree is used in student database. Here we can easily classify the different categories of student based on their result. Hence we obtain different classes of student and also can easily get the count for the number of students in each class.

Approaches for Decision tree:

For generation of decision tree node it used information gain approach to determine the suitable property in decision tree. From the highest information gain we can select the attribute. There are different algorithm of decision tree in which ID3 is used to generate the decision tree. The ID3 is proposed in 1986 by QUINLAN which is more important algorithm in decision tree. Based on the information entropy the ID3 is created which is a supervised algorithm**.**ID3 was developed from several classes of sets from database**.** The attribute of the class the which classifies the other class is found by ID3 algorithms [2].

Different Algorithm of Decision Tree

1. C4.5

C4.5 is the decision tree algorithm which is used to classify the data. C4.5is next step of ID3 algorithm [4]. C4.5 uses information Gain Ratio as dividing criteria.

2. CART

The abbreviation of CART is Classification and Regression Trees which was developed by Breiman et al. CART is algorithm which actually works as binary tree where every internal node has exactly two outgoing edges. The splits are selected by using the Towing Condition and the generated tree is pruned by Cost-Complexity Pruning. An important feature of CART is it has the ability to obtained regression trees.

3. CHAID

Starting from early seventies researchers in applied statistics have developed several procedures for generating decision trees, such as: AID, MAID, THAID and CHAID. CHIAD (Chisquare-Automatic-Interaction-Detection) it is originally designed to handle nominal attributes only with respect to the target attributeai, CHAID finds the pair of values in Vi that is least significantly different for each attribute [3], [4].

4. ID3

ID3 stands for Iterative Dichotomiser 3.It is proposed in 1986 by QUINLAN.It is very important and simple decision tree algorithm. This algorithm does not apply any pruning. The ID3 algorithm uses information gain to decide the dividing attribute. Given a collection of possible outcomes, Entropy is nothing but the uncertain data present in the data set and it is measured by equation

$$\text{Entropy}(S) = -\sum p(x)\log 2p(x) (1)$$

Where, S is dataset for which entropy is calculated X is set of classes in dataset $P(x)$ is proportion/probability of the number of element in class X to the number of element in the set S .When $I(S) = 0$ then the dataset is perfectly classified i.e. all element in S are of the same class.In other words how much uncertainty in S is reduced after splitting set S on attribute A is given by the equation

$$\text{IG}(S) = \text{I}(S) - \sum p(t) * \text{I}(t)$$

Where, $I(S)$ is entropy of dataset T is subset created from splitting S by attribute $A.\,P(t)$is proportion/probability of the number of element in class to the number of element in the set $S.\,I(t)$ is entropy of subset t. The basic algorithm steps are as follows [4].

Algorithm for ID3 is as follows

1. Select all attributes from the different levels of decision tree nodes.
2. Calculate the information growth for every attribute.
3. Use the information gain as the attribute selection criteria/measures and choose the attribute with largest information gain to decide the root node of the decision tree.
4. Branches of the decision tree are calculated by the different information gain values of the nodes.
5. Build the decision tree nodes and branches recursively till a particular dataset of the instances belongs to the same group.

This algorithm does not give the proper attributes so it is not suitable for attribution [3], [4].

- It chooses the attributes based on their occurrence instead of their importance.
- It does not deal with noisy data sets properly.
- It over-fits the tree to the training data hence unreliable.
- It creates complex trees without pruning unnecessarily.

Advantages of Decision Tree:

All advantages can be studied from [14], [15].

- It classifies unknown records very fast.

- In the presence of redundant attributes decision tree work very good.
- Decision trees are somewhat strong in the presence of noise if the methods likes over fitting are provided.

Disadvantages of Decision Tree:

All disadvantages can be studied from [14], [15].

- In the construction of a decision tree not applicable data effects badly.
  – E.g. ID numbers
- Any small changes in the data can change the overall look of decision tree.
- A sub-tree in decision tree can be make a copy many times.

Applications of Decision Tree:

All applications can be studied from [14], [15].

- Medicine: Decision tree is most useful in diagnostics of various diseases. Also for heart and sound diagnosis [14].
- Intrusion Detection: Decision tree use for generating genetic algorithm to automatically generated rule for an intrusion detection expert [14], [15].
- Image processing: For grouping of 3D features in aerials images using decision tree classifiers [14].

V. SUPPORT VECTOR MACHINE

SVM is into picture since 1992, when there was a need of classification and regression tools based on some predictions. It is introduced by Vapnick, Guyon and Boser in COLT-92. For separating any data we define certain classes and depending on the complexity of the datasets we define it as the linear or nonlinear classification. SVM can just be defined as a prediction tool wherein we search for a particular line or decision boundary termed as hyperplane which easily separates out the datasets or classes, hence it avoids the extra over fit to the data. It uses hypothesis space of a linear space into an high dimensional feature space. It is also capable of classifying the nonlinear data where it uses kernel functions. We will discuss all this things further [5].

1. Neural Networks

Today Neural Networks are used in almost all fields of classification and regression and it contributes more in Artificial Intelligence. Here we have neurons and these neurons are responsible for building up a network i.e. grouping of similar datasets or classes of it, hence it was applied on both supervised an unsupervised learning which initially showed good results but later as the number of nodes were increased the complexity went on increasing and hence it was bit difficult to study. There we conclude that for small number of nodes neural networks are best suited. SVM recovers these disadvantages and can be applied on huge datasets as well. Neural Networks are simple and it may also use multilayer perceptron (MLP) properties where again the MLP's uses recurrent and feed forward networks. Properties of MLP include the approximating of continuous nonlinear functions which again may not provide the accurate results [6], [16], [7].

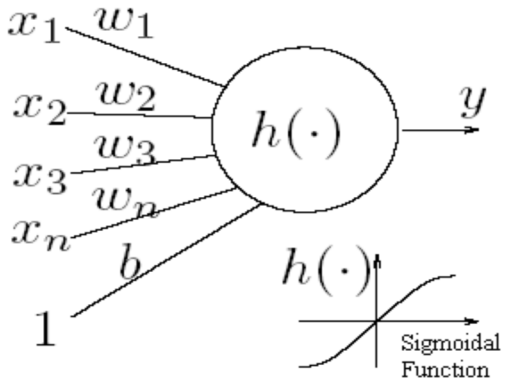

Figure2: Simple Neural Network (Taken Vikramaditya Jakkula 2013) [6], [7].

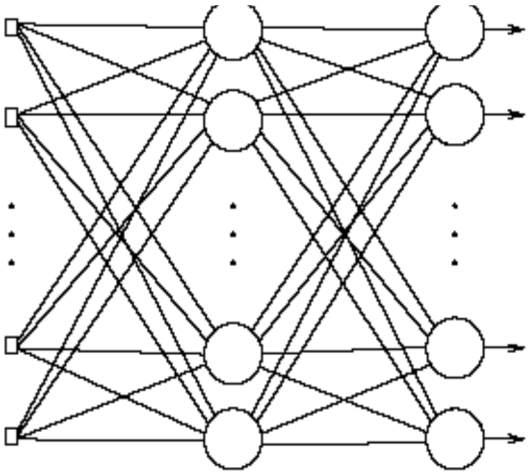

Figure 3: MultilayerPerception (Taken Vikramaditya Jakkula 2013) [6], [7].

2. SVM for Linear Classification

Support Vector Machine is used for classification and Regression. It is a novel strategy of separating the samples by just drawing a decision boundary known as hyper plane in case of linear classification. Now here in the below figure 4 we can see that for classification we have many decision

boundaries, which are capable of classifying the dataset, but the question is that which hyper plane should be selected such that it will be optimal? Here we require a hyper plane that is justice to both the categories of samples which means out of all the hyper planes or decision boundaries only one of them has to be selected. For selection of hyper plane we follow the below steps.

1. Define a function such that it will generate the required hyperplane i.e. boundary in between the different datasets.
2. Next step is to select a hyper plane and calculate its distance from both the sides of the datasets.
   i. If the distance which is calculated is maximum on both the sides as compared to the previous hyperplane then select this hyperplane as the new decision boundary.
   ii. Mark the samples which are close to the hyperplane as the supporting vectors. (helps in selection of decision boundary)
3. Repeat step 2 until best hyperplane is found.

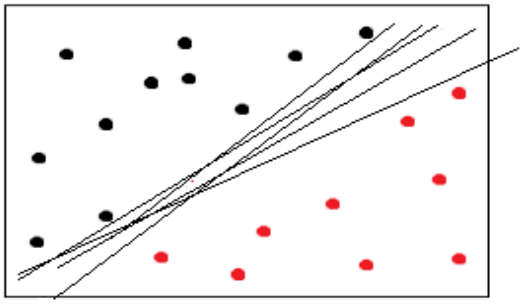

Figure 4: Hyperplane

Problem of Maximum Margin

We have seen that for selection of hyperplane we need to solve the maximum margin problem that is the distance between the decision boundary and the supporting vectors hence the below figure 5 represents the solution for it. Expression for Maximum margin is given as [5], [8].

$$margin = \arg\min_{x \epsilon D} d(x) = \arg\min_{x \epsilon D} \frac{|x - w + b|}{\sqrt{\sum_{i=1}^{d} w_t^2}}$$

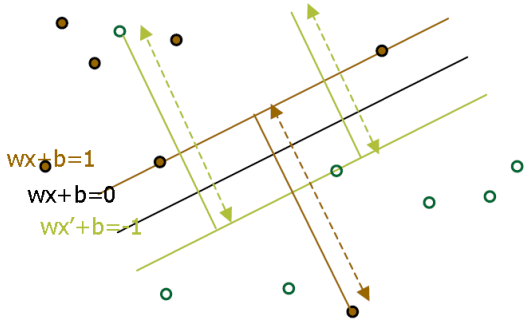

Figure 5: Maximum margin problem and solution (Taken Vikramaditya Jakkula 2013) [1].

After solving the problem the maximum margin problem is reduced to

$$\boldsymbol{margin} = \frac{\mathbf{2}}{||\boldsymbol{w}||}$$

An error in the dataset is nothing but some of the samples may be present in the other category of training set.Figure 6 illustrates the representation of errors, here we have represented the two training sets i.e. red and black (this is only for clearing the concept). Here we can see that the red training sets are present across the decision boundary i.e. in the black training set hence this is termed as an error and the decision boundary fails here. We need to recover this error, this means that the decision boundary need not be always a straight line, it may follow any curve also.

$$\min_{f,\varepsilon_i} ||f||_k^2 + c\sum_{i=1}^{l} \varepsilon_i \quad y_i f(x_i) \geq 1 - \varepsilon_i$$
$$for\ all\ i \varepsilon_i \geq 0$$

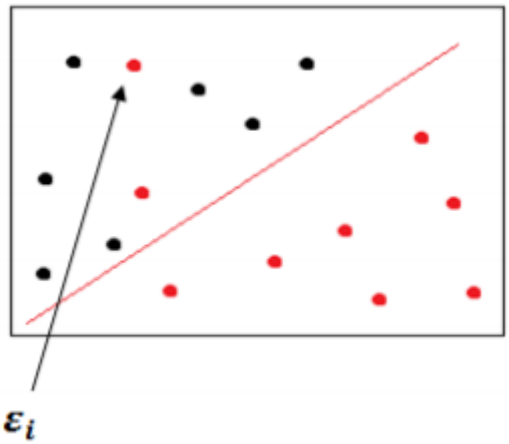

Figure 6: Representation of error

3. SVM for Non-linear classification

Support Vector Machine works excellently with linear classification. For Non Linear Classification it makes use of the Kernel function because it now requires a huge feature space for classifying the data.

3.1. Soft Margin Classifier

Always it is not necessary that we will get an exactly separable line that divides the data. However the hyperplane is capable of classifying the datasets where we may get curved decision boundaries but this might not be desirable if the datasets consists of noise into it. So it will be always better for the smooth boundary to neglect some data points than be curved or loops, around the outliers and here we use the slack variables. So we have,

$$y_i(w'x + b) \geq 1 - Sk[9][10].$$

Now this allows a particular point to be at small distance on the wrong side of the hyper plane without breaking or violating the constraint. Here we may get a huge slack hence Lagrangian variable is introduced which deals with this huge slack

$$\min L = w'w - \sum\lambda k(yk(w'xk + b) + sk - 1) + \alpha\sum sk$$

Here α is reduced and hence it allows more data to lie onto the opposite or wrong side of the hyperplane and can be treated as outliers therefore resulting in smoother decision boundary [10].

3.1.2 Kernel

For linear data, a separating hyperplane can be used for classifying it. However as discussed earlier it is not necessary to have linear data all the time sometimes nonlinear data has to be classified where is separating hyperplane won't work easily hence we need a special function known as the kernel function to map the nonlinear data to high dimensional feature space. Hence the new mapping is now linearly separable [1].Below figure 9 illustrates it [7], [11], [12].

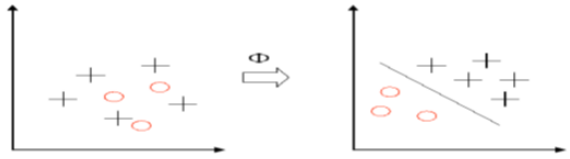

Figure 7: Use of Kernels [7], [11], [12]

The Mapping function defined by kernel is:

$$K(x, y) = \Phi(x).\Phi(y)$$

3.1.3. Feature Space

Transformation of data into feature space makes it easy to classify nonlinear data and define a similarity measure based on the dot product. In non-linear classification the data sets are present anywhere which cannot be classified using hyperplane, hence the data has to be transformed into high dimensional feature space, if the feature space is chosen correctly, pattern recognition can be easy [1].

$$< x1.x2 > \leftarrow K(x1, x2) = < \Phi(x1).\Phi(x2) >$$

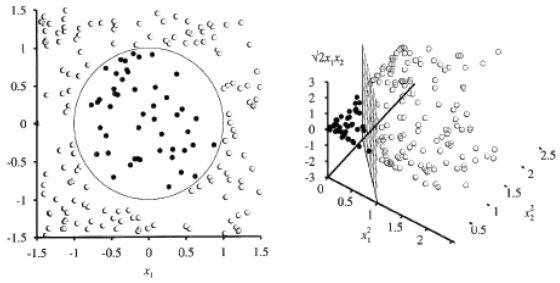

Figure 8: Feature Space Representation [7], [12].

In case of nonlinear data the Kernel trick enables SVM's to form nonlinear boundaries. Steps involved in kernel trick are given below [10] [13].

a) The algorithm is always expressed using only the inner products of data sets. This is also called as dual problem.

b) Pass original data through nonlinear maps to form new data with new dimensions by adding a pair wise product of some original data dimension to each data vector.

c) Instead of an inner product on this new, larger vectors, store in tables and later perform a table lookup, we can now represent a dot product of the data after doing nonlinear mapping on it. This is the kernel function.

3.1.4. Kernel Function

We have various kernel functions that actually help us in classifying the nonlinear data, and the basic idea behind it is operations to be performed in the input space instead of the high dimensional feature space. The most important one is the RBF kernel.

Gaussian Radial Basis Function: Radial basis functions most commonly with a Gaussian form [7], [12].

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right)$$

Exponential Radial Basis Function: A radial basis function generates a piecewise linear solution which can be attractive when discontinuities are acceptable [7], [12].

$$K(x, x') = \exp\left(-\frac{||x - x'||}{2\sigma^2}\right)$$

Selection of parameters for RBF kernel is also a difficult job because many a times the value is close to infinity. Many researchers are working on selecting the automatic parameters for this kernel.

Although neural networks are considered to be easy as compared to Support Vector Machine, they are a bit complex when number of nodes in the network is increased.

Applications of Support Vector Machine:

- In a handwriting recognition task
- Hand writing analysis, face analysis
- Pattern classification and regression

VI. CONCLUSION

The paper presents an overview on machine learning, classification, regression, and machine learning techniques. We have majorly focused on decision tree and support vector machine.ID_3 algorithm is the easiest algorithm and capable of classifying huge datasets. A neural network is the competitor for support vector machine, but it fails in some cases with respect to nonlinear classification. We have seen that SVM is best suited for it. Selection of proper hyperplane and proper parameters for RBF kernel gives more accurate results as compared to neural networks.

VII. REFERENCES

[1] Vikramaditya Jakkula , "Tutorial on Support Vector Machine" ,2013

[2] Lan Li, Shaobin Ma, Yun Zhang, "Optimization Algorithm based on Support Vector Machine"in *Seventh International Symposium on Computational Intelligence and Design*, 2014

[3] Shahrukh Teli M-Tech Student, Prashasti Kanikar Assistant Professor, MPSTME SVKM'SNMIMS University, Mumbai, India."A Survey on Decision Tree Based Approaches in Data Mining", 2015

[4] Lior Rokach and Oded Maimon,IEEE Transaction On System, Man and Cybernetics Part C, Vol 1, No. 11, November Top Down Induction Of Decision Tree Classifier-A Survey,2002

[5] Wikipedia Http://en.wikipedia.org/wiki

[6] David M Skapura, Building Neural Networks, ACM press, 1996

[7] Tom Mitchell, Machine Learning, McGraw-Hill Computer science series, 1997

[8] Tutorial slides by Andrew Moore. Http://www.cs.cmu.edu/~awm

[9] Burges C, "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998

[10] J.P.Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.

[11] http://www.enm.bris.ac.uk/teaching/projects/ 200 4_05/dm1654/kernel.html

[12] Image found on the web search for learning and generalization in svm following links given in the bookabove.

[13] Burges B.~Scholkopf, editor, "Advances in Kernel Methods--Support Vector Learning". MIT press, 1998.

[14] https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=0ahUKEwiV8fPY17_MAhXQI44KHTwcA94QFggqMAM&url=https%3A%2F%2Fwww.ijedr.org%2Fpapers%2FIJEDR1401001.pdf&usg=AFQjCNHiPFlVRDa7Q5HTgJMu8FY6GH7h8A&bvm=bv.121099550,d.c2E

[15] 1Mr. Brijain R Patel, 2Mr. Kushik K Rana Department of computer engineering, GEC Modasa, India Assistant Professor, Department of computer engineering, GEC Modasa, India patelbrijain808@gmail.com , Presentation Slides "A Survey on Decision Tree Algorithm For Classification".

[16] U.V Kulkarni, S.V Shinde, "Neuro –fuzzy classifier based on the Gaussian membership function",4th ICCCNT 2013,July 4-6,2013,Tiruchengode,India.

[17] S.V Shinde, U.V Kulkarni, "Mining Classification Rules from Fuzzy Min-Max Neural Network ",5th ICCCNT 2014,July 11-13,2014,Hefei,China