Machine Learning Techniques for Data Mining: A Survey

Seema Sharma[1], Jitendra Agrawal[2], Shikha Agarwal[3], Sanjeev Sharma[4]
School of Information Technology,UTD, RGPV, Bhopal, M.P., India.
[1]seemasharmacg@gmail.com, [2]jitendra@rgtu.net, [3]shikha@rgtu.net, [4]sanjeev@rgtu.net

Abstract–**Data mining (DM) is a most popular knowledge acquisition method for knowledge discovery. Classification is one of the data mining (machining learning) technique that maps the data into the predefined class and group's .It is used to predict group membership for data instance. There are many areas that adapt Data Mining techniques such as medical, marketing, telecommunications, and stock, health care and so on. This paper presents the various classification techniques including decision tree, Support vector Machine, Nearest Neighbor etc. This survey provides a comparative Analysis of various classification algorithms.**

Keywords - **Data Mining, Data Classification, Decision Tree, Bayesian network, Nearest Neighbour, Support Vector Machine (SVM).**

I. INTRODUCTION

Data mining is used to extract the required data from large database .Data mining in the database is a new interdisciplinary field of computer science. It is the process of performing automated extraction and generating the predictive information from large database. it is actually the process of finding the hidden information /patterns from the repositories [2]. The data mining process used a variety of analysis tool to determine the relationship between data in large database. Data mining consists of the various technical approaches including machine learning, statistic, and database system. The goal of the data mining process is to extract information from large database and transform into a human understandable format. The DM and knowledge discovery are essential components due to its decision making strategy.

In data mining, classification, regression and clustering are three approaches in which instances are grouped into identified classes [2]. Classification is a popular task in data mining especially in knowledge discovery and future plan, it provides the intelligent decision making, classification is not only used to study and examine the existing sample data but also predicts the future behavior to that sample data. The classification includes two phases, first is learning process phase in which the training data is analyzed, then the rules and patterns are created. The second phase tests the data and archives the accuracy of classification patterns [3].Clustering approach is based on unsupervised learning because there are no predefined classes. In this approach data may be grouped together as a cluster [4] [6]. Regression is used to map data item into a really valuable prediction variable. In Classification technique various algorithms such as decision tree, nearest neighbor, genetic algorithm support vector machine (SVM) etc. [1]. In this paper, we examine the various classification algorithms and compare them. In the rest of this paper, we first give Decision Tree Concepts in Section II, Bayesian Network in Section III, K-Nearest Neighbor in section IV, Support Vector Machine in Section V. Introduce some related work in section VI. Comparative studies on accuracy are presented in Section VI. The paper is concluded in Sections VII

II. DECISION TREE

The Decision tree is one of the classification techniques in which classification is done by the splitting criteria. The decision tree is a flow chart like a tree structure that classifies instances by sorting them based on the attribute (feature) values. Each and every node in a decision tree represents an attribute in an instance to be classified. All branches represent an outcome of the test, each leaf node holds the class label. The instance is classified based on their feature value.

There are numerous methods for finding the feature that best divide the training data such as information gain, gain ratio, Gini index etc. The most common way to build decision trees by using top down greedy method partitioning, starting with the training set and recursively finding a split feature that maximizes some local criterion. Decision tree generates the rule for the classification of the data set. The three basic algorithms are widely used that are ID3, C4 .5 and CART. [2]

A ID3

ID3 is an iterative Dichotomer 3. It is an older decision tree algorithm introduced by Quinlan Ross in 1986 [9]. The basic concept is to make a decision tree by using the top-down greedy approach. ID3 uses the information gain for selecting the best feature.For defining the information gain we have to first calculate the entropy.

$$\text{Entropy (s): } - \Sigma \ [P(I) \log (2) P(I)] \qquad (1)$$

Where – P (I) refers to proportion of S belong to Class I, S are all the records (instance), C refer as Class, Σ is over C i.e. Summation of all the classifier.

978-1-4799-1597-2/13/$31.00 ©2013 IEEE

$$\begin{aligned}\mathrm{InformationGain}(\mathrm{S},\mathrm{A}) &\\ = \ &\mathrm{Entropy}(\mathrm{S})\\ -\ &\Sigma\ ((|\mathrm{Sv}|/|\mathrm{S}|)\ \mathrm{Entropy}(\mathrm{Sv}))\end{aligned} \tag{2}$$

Where A is feature for which gain will be calculated, V is all the Possible of the feature, Sv is the no of element for each V.

B C4.5

C4.5 is the decision tree algorithm generated Quinlan [15].It is an extension of ID3 algorithm. The C4.5 can be Refer as the statistic Classifier. This algorithm uses gain radio for feature selection and to construct the decision tree. It handles both continuous and discrete features.C4.5 algorithm is widely used because of its quick classification and high precision. The gain radio "Normalized" the information gain as follows (Quinlan 1993).

$$\mathrm{GainRadio}(\mathrm{A},\mathrm{S}) = \frac{\mathrm{informationGain}(\mathrm{S},\mathrm{A})}{\mathrm{Entropy}(\mathrm{S},\mathrm{A})} \tag{3}$$

C CART

It is stand for Classification Regression Tree introduced by Bremen [8].

TABLE I
COMPARISON OF DECISION TREE ALGORITHM

Algorithms	ID3	C4.5	CART
Measure	Entropy info-gain	Entropy info-gain	Gini diversity index
Procedure	Top-down decision tree construction	Top-down decision tree construction	Constructs binary decision tree
Pruning	Pre-pruning through a single pass algorithm	Pre-pruning through a single pass algorithm	Post pruning based on cost-complexity measure
Advantages	1. Easy to understand. 2. In final decision whole training example is taken.	1. It handles training data with missing feature values. 2. It handles both continuous and discrete features.	1. CART doesn't require Variables to be selected advance. 2. It easily handles outliers.
Disadvantage	1.No backtracking will do a search 2. No global optimization is done	1. This is not suitable small data set.	1. Trees may be unstable. 2. It is a discrete scoring system.

CART uses binary splitting that means the node has exactly two outgoing edges and splitting are done by the Gini index.

$$\mathrm{GiniIndex} = 1 - \sum \mathrm{P}^2(\mathrm{I}). \tag{4}$$

The properties of CART are that it is able to generate the regression tree.

III. BAYESIAN NETWORK

A Bayesian Network (BN) is a graphical model for relationships among a set of various variable features. This graphical model structure S is a directed acyclic graph (DAG) and all the nodes in S are in one-to-one correspondence with the features of a data set. The arcs represent influences among the features while the lack of possible arcs in S encodes conditional independence. Bayesian classifier has exhibited high accuracy and speed when applied to large databases [18] [19] Bayesian networks are used for modeling knowledge Bioinformatics, engineering, medicine, Bio-monitoring, Semantic search image processing. The Naïve Bayes Classifier is based on Bayes Theorem. Theorem is

$$\mathrm{P}(\ \mathrm{H}|\mathrm{X}\ ) = \mathrm{P}(\ \mathrm{X}|\mathrm{H})\ \mathrm{P}(\mathrm{H})\ \mathrm{P}\ (\mathrm{X}) \tag{5}$$

H- Some hypothesis, such that data tuples X belongs to specified class C
X – Some evidence, describe by measure on a set of attributes
P (*H*|*X*) – posterior probability that the hypothesis H holds given the avid
X *P (H)* – prior probability of H, independent of X
P (*X*|*H*) – posterior probability that of X conditioned on H.

A. Advantages

1. Neural networks are able to handle noisy data, classify patterns untrained data on which they are not being trained.
2. Well suited for continuous feature valued inputs and outputs.
3. Real world application of like handwritten character recognition etc..

B. Disadvantage

1. Training time will be large.
2. Poor interpretability.
3. Require number of parameters such as network topology or structure.

C. Challenge in Bayesian Network

To extract the knowledge embedded in trained neural networks and representation of that knowledge symbolically is the challenging issue of neural network.

IV. K-NEAREST NEIGHBOUR

The K-Nearest Neighbor (NN) is the simplest method of machine learning.. In which object classifies based on the closest training example in the feature space. Its role implicitly computes the decision boundary and it is also possible to compute the decision explicitly. So the computational complexity of NN is the function of the boundary complexity [21]. The neighbors are selected from a set of objects for which the correct classification is known. No explicit training step is required this can be thought of as the training set to the algorithm. The k-NN algorithm is sensitive to the local structure of the data set. This is the special case when k = 1 is called the nearest neighbor algorithm. The best choice of k depends upon the data set; higher values of k diminish the effect of noise on the classification [24] but make boundaries between classes less distinct. The various heuristic techniques are used to select the good K. KNN has some strong consistency results. As the infinity approaches to data, the algorithm is guaranteed to yield an error rate less than the Bayes error rate. [24]. If the value of k is small then noisy samples may win the majority votes, which results in misclassification error. That can be solved with larger value of k.

A. Advantages

1. Easy to understand and implement classification technique.

B. Disadvantages

1. Computational costs are expensive, when sample is large.

2. The local structure of the data is very sensitive and require large storage

C. Challenge in K-Nearest Neighbor

1. Maintain the classification accuracy of the KNN classifier.
2. Reduce large amount of work on the application of proximity graphs to the KNN problem.

V. SUPPORT VECTOR MACHINE

The support vector machine [SVM] is a training algorithm. It trains the classifier to predict the class of the new sample. SVM is based on the concept of decision planes that defined decision boundary and point that form the decision boundary between the classes called support vector treat as parameter. SVM is based on the machine learning algorithm invented by vapnik in 1960's. It is also based on the structure risk minimization principle to prevent over fitting.

There are 2 key implementations of SVM technique that are mathematical programming and kernel function. It finds an Optimal separates hyper plane between data point of different classes in a high dimensional space. Let's assume two classes for classification. the classes being P and N for Yn= 1,-1, and by which we can extend to K class classification by using K two class classifiers. Support vector classifier (SVC) searching hyper plane. But SVC is outlined so kernel functions are introduced in order to non line on decision surface.

A. Linear SVC

Which data is linearly separable. Let we is weight vector, his base, Xn is the nearest data point. $w^T + b \geq 1$ for $x_n \in$ PAnd $w^T x_n + b \leq -1$ for$x_n \in$ N.

For optimization the problem minimizes the $\frac{1}{2}[w^T w]$ Subject to $y_n(w^T w + b) \geq 1$ for n=1 to N. The Lagrangian formula for this problem formula for this problem

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \alpha_n y_n(w^T x_n + b) - 1 \qquad (6)$$

Where α is the Lagrange multiplier, for the quadratic program is maximized with respect to α≥0 and minimize with respect to w, b

$$\Delta wL = w - \sum_{n=1}^{N} y_n \alpha_n x_n = 0. \qquad (7)$$

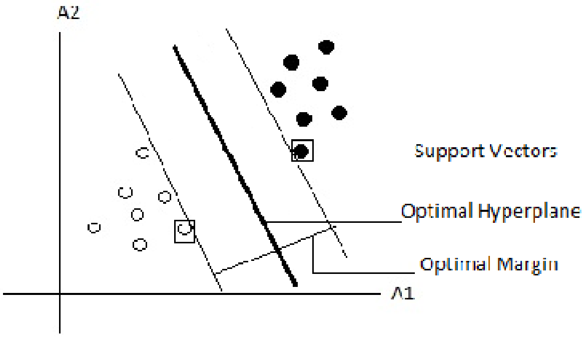

Fig 1 linear SVC

$$\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} y\alpha_n = 0 \ . \qquad (8)$$

Substitute the value in equation 6 we get

$$L(\alpha) = \sum_{n=1}^{N} \alpha_n \frac{1}{2}\sum_{n=1}^{N} \alpha_n\alpha_m y_n y_m w^T w. \qquad (9)$$

B. Non –linear SVC

A linear classifier not suitable for c class hypothesis. It can be used to learn nonlinear decision function space SVM can also be extended for learning non-linear decision function. Mapping function space denoted by

$$X \to H \ \text{ and } \ x \to \varphi \ . \qquad (10)$$

Mapping the data into H is time consuming and difficult. So require a function to give the inner product value, introduce the kernel function.

$$K(x, z) = \varphi(x)^T\varphi(z) \ . \qquad (21)$$

The kernel function allows us to hyper plane without explicitly perform the calculation

C. Non separable case

Noise is present in the training data, some data point may be misclassified .We introduce a vector of slack variable $\varepsilon = (\varepsilon 1 \ldots)^T$

A. Advantages

1. Accurate methods among all machine learning algorithms. It finds the best classification function of training data
2. SVM prevents over fitting than other methods.

B. Disadvantages

1. It is computationally expensive.
2. SVMs are requiring large amount of training time and large amount of storage and poor interpretability of results.

C. Challenge in SVM

1. Implementation of SVM is determined by the kernel so find best kernels for appropriate application is the open challenge.
2. SVM model is very expensive for the space and evaluation.

VI. RELATED WORK

ID3 algorithm are famous due to its high classification speed, easy to construction easy to understand but problem are arisen to chose the attribution which has many value this problem are solved by Liu Yuxu et al [10], they proposed the algorithm that uses attribute-importance to increase the information gain of feature, introduce attribute importance is a vagueconcept, often referring to prior knowledge about the transaction. The proposed algorithm analyzes with ID3 by an example. This indicates that the improved ID3 algorithm can make more effective rules. Another feature of ID3 algorithm is to classify the spatial data set.

ImasSukaesihSitanggang et al [11] Propose a new spatial decision tree algorithm which is based on the ID3 algorithm for discrete features represented in polygons, points and lines. The proposed algorithm employs the spatial information gain for choosing a best splitting layer from a set of explanatory layers. They used spatial datasets that are composed by a set of layer. The proposed algorithm divides the layer into two groups: explanatory and one reference layer and spatial relationship are applied for construction tuples. Each of the layer has own attribute. In this approach two layers are associated and produce a new layer of spatial relationship. Author shows that this proposed algorithm is 74.72% in real spatial data set.

Khatwani, S. Et al [13] proposed Id3 algorithm to create multiple decision tree each of predicts the performance based on the feature. They use the genetic algorithm for fitness function and apply on the Id3 tree to improve on id3 performance.

NishantMathur et al [14] precede an ID3 tree based on the Havrda and Charvat Entropy. They analyze that traditional ID 3 algorithm is based on Shannon entropy and proposed entropy and show that traditional id3 contain large no of leaf node.

Sathyadevi, G. et al. [20] propose CART derived model along with the extended definition to identify diagnosing hepatitis disease provided an efficient classification accuracy based model.

Mena ChanaTasi et al [5] proposed a novel method for medical problem. They combine the PSO and C4.5, where PSO is used in the feature selection technique and C4.5 adopts PSO fitness function for classification by using five datasets from UCI repository. They compare the proposed algorithm with another algorithm such as logistic regression, Back Propagation Neural network, support vector machine and show that they propose algorithm's accuracy is best among them.

Biao Qin et al [7] proposed novel Bayesian classification technique that is based on the uncertain data. They take 20 data sets from UCI repository and apply uncertain Bayesian classification and prediction technique. Their Implementation proposed algorithm in Weka and show that the result of the proposed approach is better than the Bayesian Classification.

David Tania [5] presents absolute taxonomy of Nearest Neighbor Queries in spatial databases,. The taxonomy comprises four perspectives: space, the result, query-point, and relationship. Lien-Fa Lin et al

[22] highlight on the issue of scalable processing continuous K-nearest neighbor queries over moving objects with uncertainty. They propose a CI-tree to predetermine the candidates for CKNN queries.

XiuboGeng et al [12] proposed a machine learning K nearest neighbor method for query dependent ranking. They first consider the online method and next consider two offline methods which create a ranking model to enhance the efficiency of ranking in advance and approximation are accurate in terms of difference in loss of prediction .

A.Moosavianet at [17] proposed a technique that produces the fault detection of engine journal bearing. The proposed technique is based on the feature selection and machine learning technique. They use SVM and KNN algorithm for classifying the journal bearing fault condition in the system. They show that the proposed algorithm has possibilities and abilities in the fault diagnosis.

Xuemei Zhang et al [25] propose a structure Risk minimization technique in SVM for minimizing the misclassification; they employ a risk decision rule of empirical risk Minimization (ERM) for a non separable sample in MATLAB. Shows computational result is better than the SVM.

Savvaskaratsiolis et al [15] proposed a region based support vector (SVM) algorithm for medical diagnosis. They use new methodology that divides the training set into two subsets, first subset is used to train SVM with RBF kernel,second subset is used to train other SVM with polynomial kernel. This proposed technique used Pima data set that contains 500 normal cases and 265 abnormal cases.

VII. COMPARATIVE STUDY

This comparative study shows that the classification accuracy is based on the data set. Table 4 consist the difference among classification algorithm bases on some factors

TABLE II
CLASSIFICATIONACCURACYOFALGORITHMS
(USINGMATLABCLASSIFICATION TOOLBOX)

Data set	Decision tree	N B	KNN	SVM
Ablone	0.93	0.10	0.99	0.83
Australian	0.84	0.10	0.64	0.99
Bcw	0.53	0.84	0.70	0.99
Bio	0.60	0.99	0.80	0.82
Car	0.711	0.43	0.99	0.72
DNA	0.95	0.92	0.99	0.98
Avg. Accuracy	0.76	0.57	0.86	0.88

TABLE IV
DIFFERENCES AMONG CLASSIFICATION TECHNIQUES

Algorithm	Decisio n tree	NB	K-Nearest Neighbor	Support Vector Machine
Proposed By	Quinlan	Duda and Hurt	Cover and Hart	Vapnik
Avg. Accuracy	76%	57%	86%	88%
Tolerance to noise	Good	Very good	Average	Good
Speed of classification	High	High	Average	High
Generative or Discriminative	Discrim inate	Genera tive	Discriminat e	Discrimi nate

VIII. CONCLUSION

This paper specifies various classification techniques used in many fields, such as Decision Tree, Bayesian network, Nearest Neighbour, Support Vector Machine (SVM). Generally Decision trees and Support vector machines have different operational profiles, where one is very accurate the other is not and vice versa. On the other hand, decision trees and rule classifiers have a similar operational profile. Various algorithms will be combined for classifying the data set. This paper provides compressive overview of various classification techniques used in different fields of data mining. In any field one classification technique is more useful than another. This paper presents various classification techniques. One of the above techniques can be selected based on the required application conditions.

REFERENCES

[1]. Bakar, A. A., Othman, Z. A., & Shuib, N. L. M. (2009, October). Building a new taxonomy for data discretization techniques. In *Data Mining and Optimization, 2009. DMO'09. 2nd Conference on* (pp. 132-140). IEEE.

[2]. Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

[3]. Balagatabi, Z. N., & Balagatabi, H. N. (2013). Comparison of Decision Tree and SVM Methods in Classification of Researcher's Cognitive Styles

in Academic Environment. *Indian Journal of Automation and Artificial Intelligence*, *1*(1), 31-43.

[4]. Merceron, A., & Yacef, K. (2005, May). Educational Data Mining: a Case Study. In *AIED* (pp. 467-474).

[5]. Taniar, D., & Rahayu, W. (2013). A taxonomy for nearest neighbour queries in spatial databases. *Journal of Computer and System Sciences*.

[6]. Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. *China Machine Press*, *8*, 3-6.

[7]. Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009, December). Naive bayes classification of uncertain data. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on* (pp. 944-949). IEEE.

[8]. Burrows, W. R., Benjamin, M., Beauchamp, S., Lord, E. R., McCollor, D., & Thomson, B. (1995). CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver, Montreal, and Atlantic regions of Canada. *Journal of applied meteorology*, *34*(8), 1848-1862.

[9]. Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

[10]. Jin, C., De-lin, L., & Fen-xiang, M. (2009, July). An improved ID3 decision tree algorithm. In *Computer Science & Education, 2009. ICCSE'09. 4th International Conference on* (pp. 127-130). IEEE.

[11]. Sitanggang, I. S., Yaakob, R., Mustapha, N., & Nuruddin, A. A. B. (2011, June). An extended ID3 decision tree algorithm for spatial data. In *Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on* (pp. 48-53). IEEE.

[12]. Geng, X., Liu, T. Y., Qin, T., Arnold, A., Li, H., & Shum, H. Y. (2008, July). Query dependent ranking using k-nearest neighbor. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 115-122). ACM.

[13]. Khatwani, S., & Arya, A. (2013, January). A novel framework for envisaging a learner's performance using decision trees and genetic algorithm. In *Computer Communication and Informatics (ICCCI), 2013 International Conference on* (pp. 1-8). IEEE.

[14]. Mathur, N., Kumar, S., Kumar, S., & Jindal, R. The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree.

[15]. Karatsiolis, S., & Schizas, C. N. (2012, November). Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset. In *Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on* (pp. 139-144). IEEE.

[16]. Tsai, M. C., Chen, K. H., Su, C. T., & Lin, H. C. An Application of PSO Algorithm and Decision Tree for Medical Problem.

[17]. Moosavian, A., Ahmadi, H., & Tabatabaeefar, A. (2012). Journal-bearing fault detection based on vibration analysis using feature selection and classification techniques.

[18]. Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*(2-3), 131-163.

[19]. Jensen, F. V. (1996). *An introduction to Bayesian networks* (Vol. 210). London: UCL press.

[20]. Sathyadevi, G. (2011, June). Application of CART algorithm in hepatitis disease diagnosis. In *Recent Trends in Information Technology (ICRTIT), 2011 International Conference on* (pp. 1283-1287). IEEE.

[21]. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. Miscellaneous Clustering Methods. *Cluster Analysis, 5th Edition*, 215-255.

[22]. Li, G., Li, Y., Shu, L., & Fan, P. (2011). Cknn query processing over moving objects with uncertain speeds in road networks. In *Web Technologies and Applications* (pp. 65-76). Springer Berlin Heidelberg.

[23]. Khan, R. Z., & Allamy, H. (2013). Training Algorithms for Supervised Machine Learning: Comparative Study. *International Journal of Management & Information Technology*, *4*(3), 354-360.

[24]. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, *13*(1), 21-27.

[25]. Zhang, X., & Yang, L. (2012). Improving SVM through a risk decision rule running on MATLAB. *Journal of Software*, *7*(10), 2252-2257