A Survey of Classification Techniques in Data Mining

M. Sujatha
Research Scholar,
Computer Science And Systems Engineering,
Andhra university, Visakhapatnam, Andhra Pradesh,

S. Prabhakar
Research Scholar,
Computer Science Engineering,
JNTU,Hyderabad,Andhra Pradesh,

Dr. G. Lavanya Devi
Assistant Professor,
Computer Science And Systems Engineering,
Andhra university,Visakhapatnam,,Andhra Pradesh,

Abstract- This paper provides a survey of numerous data mining classification techniques for innovative database applications. Classification is a model finding process that is used for assigning the data into different classes according to specific constrains. There are several major kinds of classification algorithms including Genetic algorithm C4.5, Naive Bayes, SVM, KNN, decision tree and CART. We deliberate the description of the algorithm.

Keywords - KNN, C4.5, SVM, CART

I. INTRODUCTION

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity [1]. There are several applications for Machine Learning (ML), the most significant of which is data mining. People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. This makes difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines. Numerous ML applications involve tasks that can be set up as supervised. In the present paper, we have concentrated on the techniques necessary to do this. In particular, this work is concerned with classification problems in which the output of instances admits only discrete, unordered values. Our next section presented various classification methods. Section III described evaluating the performance of classifier. Finally, the last section concludes this work.

II. CLASSIFICATION METHODS

A. Genetic Algorithm-

Association Rules mining technique is used in GA, to find undetermined solution [7]. GA implements on small group of categorical data. Implementation of GA generates high-level prediction rules for better attribute selection. A single prediction rule is given by Michigan approach for each individual of a population; reduce the cost [10]. A set of prediction rules of a population of each individual represented by Pittsburgh approach [5]. In the case of classification , we measure the quality of the rule set as a whole, rather than the quality of single rule. Depending on data, rules are generalized or specialized, similar to the logical OR implementing and the logical AND operators.

B. Rules Sets-

The classification rule is "if-then-"rules, rule is (condition) .A Rule r represents an instance X, if the attributes of the instance satisfied the condition of the rule. Individual's rules are ranked .the related to their quality is called rule-based order. The rules that belong to the similar class appear together is known as class-based ordering. A good rule represents without mistakes and must cover as many examples as possible.

$$\text{Accuracy} = \frac{\text{correctly classified examples}}{\text{examples matched by the rule}} \qquad 1$$

$$\text{Coverage} = \frac{\text{examples matched by the rule}}{\text{examples in the trained set}} \qquad 2$$

These two learn sets are learning rules sequentially; one at a time is called separate-and- conquer (Furnkranz 99) and learning all rules together. In Separate- And- Conquer is after each rule is learned, the covered examples are deleted from training set and the process begins again , process completes when there are no examples to cover. Separate- And- Conquer methods are AQ family of rule induction method (Michal ski 69), CN$_2$(Clask an Boswell 91), RIPPER K(Cohen 95) .Separate- And- Conquer method accept new rule, when minimum accuracy and coverage, accept rules are better than default class, use a validation set of examples, unseen before. CN2 rule is initial pool contains an empty predicate, that covers all examples. CN2 rule algorithm does not generate rules, it generates predicates, and the majority class of the example covered by the predicate will be assigned as the class of the rule [28].

C. C4.5-

C4.5 measures numeric attributes, deals with missing values and pruning noisy data [3]. Pruning is used in C4.5 to avoid over fitting to noise in data [9]. C4.5 revised version is C4.8, implemented in WEKA as J4.8 and C5.0 (Rule quest) for commercial successor. Error estimate for sub tree is weighted sum of error estimates for all its leaves. Error estimate for a node (upper bound):

If c = 25% then z = 0.69 (from normal distribution).f is the error on the training data. N *is* the number of instances covered by the leaf.

D. CART-

CART classification and Regression Tree based on accuracy, when data is noisy/ missing values. CART takes a random sample; allow handling missing values by CHAID algorithm. In CART data preprocessing is not needed, It automatically selects relevant attributes. CHAID algorithm [26, 27] considers missing values as distinct categorical value, also this method adopted by C4.5. CART treats a refined method surrogated, as missing primary field. If primary splitter is missing, then we will use a surrogate. CART prunes, the exact order in which each node must be deleted. For small data sets one Standard Error rule is good and generates a optimal tree. For larger data sets zero Standard Error rule generates accurate tree. Both C4.5 and CART are robust tools. Surrogate loss function like Gini index is used when misclassification of Decision tree.

E. Decision Tree Induction-

The problem of constructing optimal binary Decision Tree is an NP-COMPLETE problem and thus theoreticians have searched for efficient heuristic for constructing near optimal Decision Tree. Hunt's algorithm generates a Decision tree by top-down or divides and conquers approach. The sample/row data contains more than one class, use an attribute test to split the data into smaller subsets. Hunt's algorithm maintains optimal split for every stage according to some threshold value as greedy fashion [2].

Hunt's algorithm uses greedy in approach for attribute test, select 'best' split and when to stop splitting on over fit or under fit conditions. Hunt's algorithm measures both numeric and data set. For best split considers misclassification error, Gini index, Entropy. The decrease in entropy is called "Information Gain". Misclassification error (Gini index or Entropy) on the test data set and stop when this begins to increase. Gini Index for a given node t:

$$\mathrm{GINI}(t) = 1 - \sum \left[p\left(\frac{j}{t}\right) \right]^2 \qquad 3$$

where [p(j/t)] is the relative frequency of class j at node t. Gini index is used in CART[13],SLIQ and SPRINT. When a node p is split k partitions (children) the quality of split computed as

$$GINI_{split} = \sum_{i=1}^{n} GINI(i) \text{ , Where } n_i = \text{no of records at child i, n= no of records at node p.} \qquad 4$$

To choose an attribute to partition data, the key building a decision tree which attribute to choose in order to branch. The objective is to reduce impurity or uncertainty in data as much as possible. A subset of data is pure if all instances belong to same class. The heuristic in c4.5 is to choose the attribute with maximum information gain or gain ration based on information theory.

1. Information gain

Given a set of examples D, we first compute its entropy.

$$\mathrm{Entropy}(D) = -\sum_{j=1}^{|c|} p\left(c_j\right) \log_2 p\left(c_j\right) \qquad 5$$

where $p(c_j)$ is the probability of class $c_j$in data set D. we use entropy as a measure of impurity or disorder of data set D.(or a measure of information in a tree)[11]. As the data become purer and purer the entropy value becomes smaller and smaller. Information gained by selecting attribute A_i[4].To branch to partition the data is

$$\mathrm{Gain}(D, A) = \mathrm{Entropy}(D) - \mathrm{Entropy}A_i(D) \qquad 6$$

We choose the attribute with the highest gain to branch/split the current tree. Decision Tree Induction is greedy algorithm constructs Decision Tree in Top-down, recursive Divide-and-Conquer method. There are two methods for improving uncertainty.

- Pre-pruning (Early stopping rule) stop growing a branch when information becomes stop too early. Pre-pruning implements Chi-Squared test.

- Post-pruning takes a fully grown decision tree and discarded unreliable sun tree.

2. Overfitting Model

The classification model, two types of errors are committed. The training sample contains noisy data, and then errors are generated in the form of training errors and generalization errors. The training error is also known as Re- Substitution error or Apparent error, is the number of misclassification errors generated are training records. Whereas the generalization error is an expected error of the classification model on previously unseen records. Once the tree size is too large, its test error rate increases, even though training error rate continues to decrease. This process is known as model over fitting. For a complex tree, training error is zero, test error is large because noisy data present in training sample.

3. Underfitting Model

A training and test error rates of classification model are large and the size of decision tree is small. This situation is known as model underfitting.Decision Tree constructed with new binary features with logical operators such as conjunction, negation, disjunction (Zheng 1998) In addition, Zheng (2000) created at- least- M- of- N features when condition is true otherwise false[4].Gama and brazdil (99) combined a Decision Tree with linear discriminate from constructing multivariate Decision Tree.

Decision Tree is complex because of replicated data set, Decision Tree construction is complex representation to avoid replicated data in Decision Tree, implement FICUS construction algorithm, which receives standard input or feature representation to produce a set of generated features by Markovitch and Rosenstein (2002). The study shows that C4.5 has a very good combination of error rate and speed. Based on this analytical evolution implemented a more efficient version of algorithm is known as EC4.5

F. Bayesian network -

Bayesian network is based on DAG (Directed Acyclic Graph) and one to one corresponding feature [12]. Bayesian network is divided into two tasks learning DAG and structure of network [8].

1. The network structure is fixed, learning the parameter in the conditional probability tables (CPT).
2. If structure is unknown, one approach is introduce scoring function that evaluates the "fitness" of the network with respect to training data and then to search for best network according to this score [16].

Compare Bayesian network to Decision Tree or Neural Network takes prior information about a given problem. Large data set with many features not suitable for Bayesian network [6].

G. Instance Based Learning -

Instance based learning's is lazy-learning algorithm, as a delay the induction or generalization process until a classification is performed. Last learning algorithm require less computation time during the training phase than eager-learning algorithm (such as Decision tree, neural & Bayesian network) but more computation time during the classification process. In KNN, Irighton & Mellish 2002) found that their ICF algorithm and RT3 algorithm (Wilson & Martinez 2003).The disadvantage of Instance Based Learning takes more computation time for classifications. The available input features should be used in modeling via feature selection (Yu & Liu 2004) which improves classification accuracy and slow down classification time. If suitable distance metric is chosed, then accuracy of instance based classifier.

H. Support Vector Machine -

Support Vector Machine is a new classification method for both linear and non-linear data. It uses a non-linear mapping to transform the original training data into a higher dimension. With the new dimension, it searches for the linear optimal separating hyper plane (i.e. "decision boundary"). With an appropriate non-linear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyper plane. SVM finds this hyper plane using support vectors ("essential training tuples") and margins (defined by the support vectors)[25]. Features: Training can be but accuracy is high owing to their ability to model complex non-linear decision boundaries (margins – maximization). SVM used for both classification and prediction.

I. K-Nearest-Neighbour (KNN) -

KNN is a non- Parametric classification method which is simple but effective in many cases [19]. The KNN has major drawbacks 1. Its low efficiency- being a lazy learning method prohibits it in many such applications such as dynamic web mining for a large repository and 2. Its dependency is based on selection of a good for K. There are [15, 30] that are used to significantly reduce the computation required at query time such as indexing training examples. As the KNN classifier requires storing the whole training set, when this is not at the redundancy of the training set to alleviate this problem [29, 31, 24, and 21].

Condensed Nearest Neighbor [CNN] by minimizing the number of stored patters and storing only a sub-set of the training set for classification. The basic idea is that patterns the training set may be very similar and some do not add extra information and thus may be discarded. Gate [6] proposed the Reduced Nearest Neighbor (RNN) rule that aims to further reduce the stored subset after having applied CNN. It simply removes those elements from the subset which will not cause an error.

III. EVALUATING THE PERFORMANCE OF CLASSIFIER

A. *Hold – Out method-*

The original data with labeled examples is classified into two sets, called training set and test set [34]. The set should not be used in testing and the test set should not be used in learning. Unseen test set provides unbiased estimate of accuracy. This method is mainly used when the data set is large.

B. *n-fold Cross-validation-*

The available data is partitioned into n equal-size disjoint subsets. Use each subset as the training set to learn a classifier. The procedure is run n times, which given accuracies average of the n accuracies.10-fold and 5-fold cross- validations are commonly used. This method is used when the available data is not large.

C. *Leave-one-out cross validation-*

This method is used when the data set is very small. It is a special case of cross-validation. Each fold of cross validation has only a single test example and all the test of the data is used in training [17]. If the original data has m examples, this is m-fold cross-validation.

D. *Validation set -*

The available data is divided into three subsets, 1. Training set 2. Validation set and 3. Test set. A validation set is used frequently for estimation parameters in learning algorithm. In such cases, the values that give the best accuracy on the validation set are used as the final parameter values. Cross validation can be used for parameter estimating as well.

E. *Minimum Description Length (MDL)-*

MDL handles missing values naturally as missing at random. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors. Missing values in nested columns are interpreted as sparse. Missing in columns with simple data types are interpreted as missing at random tm. MDL takes into consideration the size of the model as well as the reduction in uncertainty due to using the model [20]. Both model size and entropy are measured in bits.

MDL considers each attribute as simple predictive model of the target class .model selection refers to the process o comparing and ranking the single-predictor model. Automatic data preparation performs supervised binning for MDL [32]. Supervised binning uses decision trees to create the optimal bin boundaries. It is both categorical and numerical attributes.

F. *Bagging -*

Bagging (Breiman, 1996), a name derived from "bootstrap aggregation", was the first effective method of ensemble learning and is one of the simplest methods of arching [1]. The meta-algorithm, which is a special case of the model averaging, was originally designed for classification and is usually applied to decision tree models, but it can be used with any type of model for classification or regression. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement. Each of these data sets is used to train a different model. The outputs of the models are combined by averaging (in case of regression) or voting (in case of classification) to create a single output. Bagging is only effective when using unstable (i.e. a small change in the training set can cause a significant change in the model) nonlinear models.

Create classifiers using training sets that are boot strapped (drawn with replacement). Sampling with replacement is according to a uniform probability distribution. Each bootstrapped sample D has a same size as original data. Some instance could appear several times in the training set, while others may be omitted. Bagging improves generalization performance by reducing variance of the base classifiers. The performance of the bagging depends upon the base classifier. If the base classifier is unstable, bagging helps to reduce the error associated with random fluctuations in the training data. If a base classifier is stable bagging may not be able to improve, rather than it could degrade the performance.

G. *Boosting-*

Boosting is sequential production of classifier. Each classifier is dependent on previous one, and focuses on the previous ones errors. Examples that are incorrectly predicted in previously classifiers are chosen more often are weighted more evenly. Records that are only classified will have their weights increased. Records that are classified correctly will have their weights decreased.

H. *Ada-Boosting-*

Ada-boosting measures complex hypothesis tend to over fitting. Simple hypothesis may not explain clearly. So it is combined many simple hypotheses into a complex one. There are two approaches

1. Select examples according to error in previous classifier (more representatives of misclassified cases are selected) are more common.
2. Weight error of misclassified cases higher (all cases are in corporate, but weights are different) does not Work for some algorithms.

I. *Occam's Razor-*

Occam's razor refers the simplest hypothesis that fits the data. Occam's razor is most likely to identify unknown objects correctly. Occam's razor given two models of similar generalization error, one should prefer simpler model over more complex model [33]. For complex models, there is a greater chance that is fitted accidentally by errors in data.

J. *Random Forest -*

General purpose tool for classification and regression is WEKA. Random forest is very high accuracy for gradient boosting and support vector machine. There are two categories to construct Random Forest.

1. Classification and regression trees.

2. Bootstrap sample is a sample of the same size as the original dataset drawn from the original dataset with replacement.

IV. CONCLUSION

This paper covers with various classification techniques used in data mining. Each technique has got its own pros and cons as given this paper. 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. There have been a large no of data mining algorithms embedded in these fields to perform different data analysis tasks.

REFERENCES

[1] Jiawei Han and MichelineKamberData Mining: Concepts and Techniques,2ndedition.

[2] Baik, S. Bala, J. (2004), A Decision Tree Algorithm For Distributed Data Mining.

[3] Witten, I. & Frank, E. (2005), "Data Mining: Practical Machine Learning Tools And Techniques", 2nd Edition, Morgan Francisco, 2005.

[4] Zheng, Z. (2000). Constructing *X*-Of-*N* Attributes For Decision Tree Learning. *Machine Learning* 40: 35–75.

[5] Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., "A Data Mining Engine Based On Internet, Emerging Technologies And Factory Automation".

[6] Friedman, N., Geiger, D. &Goldszmidt M. (1997). Bayesian Network Classifiers.*Machine Learning* 29: 131-163.

[7] Fayyad, U., Piatetsky-Shapiro, G., And Smyth P., "From Data Mining To Nowledge Discovery In Databases," Ai Magazine, American Association For Artificial Intelligence, 1996.

[8] Friedman, N. &Koller, D. (2003). Being Bayesian About Network Structure: A Bayesian Approach To Structure Discovery In Bayesian Networks. *Machine Learning* 50(1): 95-125.

[9] Quinlan, J.R., C4.5 -- Programs For Machine Learning.Morgan Kaufmann Publishers, San Francisco, Ca, 1993.

[10] Bianca V. D.,PhilippeBoula De Mareüil And Martine Adda-Decker, "Identification Of Foreign-Accented French Using Data Mining Techniques, Computer Sciences Laboratory For Mechanics And Engineering Sciences (Limsi)".

[11] Breslow, L. A. & Aha, D. W. (1997). Simplifying Decision Trees:A Survey. *Knowledge Engineering Review 12:* 1–40.

[12] Jensen, F. (1996). An Introduction To Bayesian Networks. Springer.

[13] Introdution To Data Mining By Tan,Steinbach,Kumar.

[14] Collins, M., Schapire, R.E. And Singer, Y. (2000). Logistic Regression, AdaboostAndBregman Distances. Proc. Thirteenth Annual Conference Computational Learning Theory.

[15] T. Mitchell.: Machine Learning. MitpressAndMcgraw-Hill (1997).

[16] Madden, M. (2003), The Performance Of Bayesian Network Classifiers Constructed Using Different Techniques, Proceedings Of European Conference On Machine Learning, Workshopon Probabilistic Graphical Models ForClassification, Pp. 59-70.

[17] Avirm Michael Kearns And Dana Ron,"Algorithmic Stability And Sanity-Check Bounds For Leave-One-Out Cross Validation".

[18] Freund, Y. (1995). Boosting A Weak Learning Algorithm By Majority. Information And Computation 121, 256{285}.

[19] D. Hand, H. Mannila, P. Smyth.: Principles Of Data Mining. The MIT Press. (2001).

[20] Peter D. Grunwald "The Minimum Description Length Principle.

[21] M. Kubat, M. Jr.: Voting Nearest-NeighbourSubclassifiers. Proceedings Of The 17th International Conference Onmachine Learning, ICML-2000, Pp.503-510, Stanford, CA, June 29-July 2, (2000).

[22] Jiang, W. (2000). Process Consistency ForAdaboost. Tech. Report, Dept. Of Statistics,Northwestern University.

[23] Breiman, Leo, 1996. Bagging Predictors, Machine Learning.

[24] E. Alpaydin.: Voting Over Multiple Condensed Nearest Neoghbors. Artificial Intelligence Review 11:115-132, (1997) Kluwer Academic Publishers.

[25] Cristianini, N., Shawe-Taylor, 1.: An Introduction To Support Vector Machines. Cambridge University Press, Cambridge, 2000.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, And C. J. Stone.Classification And Regression Trees. Wadsworth, Belmont, 1984.

[27] D. G. Denison, B. K. Mallick, and A. F. M. Smith. A Bayesian Cart Algorithm.

[28] Clark, P., Niblett, T. (1989), The Cn2 Induction Algorithm. Machine Learning, 3(4):261-283.

[29] P. Hart.: The Condensed Nearest Neighbour Rule, IEEE Transactions OnInformation Theory, 14, 515-516, (1968).

[30] C.M.Bishop.: Neural Networks For Pattern Recognition. Oxford University Press, UK (1995).

[31] G. Gates.: The Reduced Nearest Neighbour Rule. IEEE Transactions OnInformation Theory, 18, 431-433, (1972).

[32] Barron,A,Rissanen,J,AndYu,B.(1998),"The Minimum Description Length Principle In Coding And Modeling IEEE Transaction On Information Theory.

[33] A.Blumer ,A.Ehrefeucht, D.Haussier, M.Warmuth. "Occam's Razor".