Decision Tree for the Weather Forecasting

Rajesh Kumar
Asst. Prof., Dept. of ECS
Dronacharya College of Engineering, Gurgaon, India.

ABSTRACT

Predicting the classification of data in a suitable class is a challenging task. It depends on various factors to predict the dependent variables. Since decision tree evaluation can be quantified and it is simple to use, so a model using decision tree has been proposed by the author to predict the event like fog, rain and thunder by inputting average temperature, humidity and pressure. Which can be used by farmers or by peoples of all walk of life in taking the intelligent decisions. This model can be used in machine learning and further the proposed model has scope for improvement as more and more relevant attributes can be used in predicting the dependent variable. Decision tree(Decision stump) has been implemented in Weka to facilitate the forecasting of weather..

Keywords

Decision tree, Data mining, Classification, Genetic algorithm.

1. INTRODUCTION

Classification is an utmost important task in data mining for the purpose of machine learning. First of all we create a model than model is trained by a sample of data called training set . A trained model is provided with unseen data called test set in predicting the future event classification with an accuracy[10]. Few of the classification tasks has been in use since long, like classifying a tumor cells as benign or malignant, Classifying credit card transactions as legitimate or fraudulent, Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil, Categorizing news stories as finance, entertainment, sports. Following classification methods has been used commonly depending on their forte.

- Decision Tree based Methods
- Rule-based Methods
- Memory based reasoning
- Neural Networks
- Naïve Bayes and Bayesian Belief Networks
- Support Vector Machines

Rest of the paper is organized as follows

- Section 2 covers Data mining.
- Section 3 covers Decision tree.
- Section 4 covers Analysis .
- Section 5 covers Conclusion.
- Section 6 covers References.

2. DATAMINING

The manual extraction of patterns from data has been done by human being since centuries ago and it was like cleaning of Augustean stable. Information technology revolution has dramatically increased data collection, storage and manipulation ability, this lead to increase data set size and complexity. As data sets have grown in size and complexity, the need of special tools like neural networks, cluster analysis, genetic algorithms , decision trees and support vector machines emerged for the analysis of data. Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets[22]. It uses the statistical methods and artificial intelligence algorithms in indexing and storing of data bases so that information retrieved from it can be rationalized with an efficiency. Knowledge discovery in databases field is concerned with the development of methods and techniques for making sense of data.[23] . KDD refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in the KDD process. Data mining is the application of specific algorithms for extracting patterns from the huge data[23]. Following are the goals of data mining.

2.1Classification[25]

In the data analysis , it is essential to put the instances in a desired class. It categorizes the instance in a particular category. The ability of a classifier refers to the ability to correctly classifying the unseen data in a class. The bagging and boosting are the techniques for improving the classification accuracy. Bagging improves generalization performance by reducing variance of base classifier. If a base classifier is unstable ,bagging helps to reduce the error associated with the random fluctuation in training data .In boosting each classifier is dependent on the previous one and focuses on previous error by giving them more weights. Following methods are being used in classification.

2.1.1 Rule based methods[25]

Data mining system learns from examples. It formulates classification rules in order for the prediction of future. For instance, in customer database in a bank, a query is made whether a new customer applying for a loan is a good investment? Typical rule are as follows which may be produced by rule based systems.

if STATUS = married and INCOME > 10000 and HOUSE_OWNER =yes
then INVESTMENT_TYPE = good.

2.1.2 Neural Network

Neural network can be used in the classification purpose. They simulate the human brain . Artificial Neuron can be supervised or unsupervised. They are composed of many units called neuron. Artificial neuron require long training time and are black

box which lacks explanation, but it has high tolerance to noisy data so it can classify untrained data [25].

2.1.3 Bayesian classification

Bayesian classification predicts class membership using Bayes theorem, which further uses probability. Its performance is comparable to selected neural network and decision tree. They can facilitate decision making even on computational intractable problems[25].

2.1.4 Support Vector Machine[25]

Support vector machine can classify both linear and non linear data..Data from two classes are separated by hyper plane, Support vector machine finds the hyper plane by using training data. Its training is slow but accuracy is very high and SVM can model non linear problems also.

2.1.5 Genetic Algorithm[25]

Genetic algorithm has taken a queue from the natural evolution. Initial population is created using randomly generated rules. Each rule is represented by a string of bits. In next generation, survival of the fittest selects the fittest rules. Crossover and mutation are used in production of offspring. In cross over substring of a rule are exchanged with substring of another rule. In mutation randomly selected bits are inverted. It being an iterative purpose, a rule will get position in next generation, if it crosses a threshold. Genetic algorithm can be used in classification besides optimization purpose.

2.1.6 Case Based Reasoning

Case based reasoning stores the old instances in a database to classify the unseen instances as equal to stored instance, if it does not exist than it search for another very similar instance[25].

2.2 Association[25]

Rules that associate one attribute of a relation to another attribute approaches are the most efficient means of discovering such rules like in supermarket database. If a certain percentage of all the records that contain items A and B also contain item C .the specific percentage of occurrences is the confidence factor of the rule . Association rule mining is useful in mining single dimensional Boolean association rule from the transactional databases, it can be further extended for mining multilevel rule from the transactional databases[25].

2.3 Sequence/Temporal

Sequential pattern functions identifies the collections of related records and detects frequently occurring pattern over a period of time under study .Difference between sequence rules and other rules is the temporal factor. For example - Retailers database can be used to discover the set of purchases that frequently precedes the purchase of a microwave oven or harvesting season.

3. DECISION TREE

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target parameter based on several input parameter. A tree can be made to learn by splitting the source data set into subsets based on an attribute value test[16]. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data in data mining. In TDIDT systems machine learning can be classified on the basis of following[2]

- Learning strategy used.
- .Representation of the knowledge acquired by the system.
- The application domain of the system.

Decision trees can be described as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data to facilitate the machine learning. In decision tree dependent variable is predicted from the independent variable. Decision trees used in data mining are generally of following types[16].

- Classification tree analysis is used in prediction of data in a class.
- Regression tree analysis is required in prediction of independent variable as a unit of number (e.g. the price of a house, or a patient's length of stay in a hospital).

The Classification And Regression Tree (CART) analysis is a term commonly used to refer to both of the above procedures [4]. Trees used for regression and classification resembles in procedure but differs at procedures of splitting a node. Decision tree learning is the construction of a decision tree from class-labeled training data. A decision tree is a flow-chart, where each internal node denotes a test on an attribute and each branch represents the outcome of a test, and each leaf node holds a class label or the prediction. The topmost node in a tree is the root node as in tree.

There are many specific decision-tree algorithms. Few of them are enumerated as follows[16].

- ID3 (Iterative dichotomiser3)
- C4.5 (successor of ID3)
- CART (Classification And Regression Tree)
- CHAID (CHI-squared Automatic Interaction Detector).
- MARS(extends decision trees to better handle numerical data.)

Decision tree splits the attributes by using greedy search that optimizes on certain criterion. Test conditions are specified depending on the attributes types whether it is nominal, ordinal or continuous. Determining the best split remains an issue. Greedy method advocates that nodes with homogeneous class distribution are preferred so there is a need to measure a method of node impurity. Node impurity is measured by the following.

- Gini index.
- Entropy.
- Misclassification error.

3.1 Gini index[19]

Used by the CART as acronym of classification and regression tree algorithm. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum value zero when all cases in the node fall into a single target category.

$$Gini(t) = 1 - \sum [P(j \mid t)]^2 \qquad (1)$$

For example one decision classifies 1 instance in class c1 and rest 5 instances are classified in c2 then probability p(c1)= 1/6 and p (c2)=5/6. Then

$$Gini = 1 - [(1/6)^2 - (5/6)^2] \quad (2)$$

After finding the above value, sort the values of attributes and find the Gini index of each values of the attributes and choose the position of split at the least Gini index.

3.2 Information gain[19]

Entropy at a given node t is denoted by formula

$$Entropy(t) = -\sum p(j|t) log\, p(j|t) \quad (3)$$

Where as p(j | t) is the relative frequency of class j at node t. It Measures homogeneity of a node. Where as Maximum (log n) occurs when records are equally distributed among all classes implying least information. and Minimum as 0.0 when all records belong to one class, implying most information. For example a decision is to be made and it splits one instance in class c1 and 5 instances in class c2.then entropy is calculated as follows P(C1) = 1/6 P(C2) = 5/6.
Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65bits

Then Information Gain in the form of gain split is calculated as follows

$$Entropy(p) - (\sum (ni/n) Entropy(i)) \quad (4)$$

where as Parent Node, p is split into k partitions
Where ;n_i is number of records in partition i .
It is used in ID3 and c4.5.it measures reduction in entropy as a result of split. Split with maximum reduction is selected.
Various advantages of using decision tree has been identified [16]

- It is simple to understand and interpret.
- It requires only training set and test set of data.
- It can handle both numerical and categorical data.
- It uses a white box model, So model is capable of explanation of results.
- Results of the decision tree can be validated by statistical tests That makes it possible to account for the reliability of the model.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated, so it provides fault tolerance to some extent.
- Huge amount of data can be analyzed using standard computing resources in reasonable time.

Various disadvantages of using decision tree are enumerated as follows[25].

- Decision tree suffers the over fitting problem.
- Anomalies are observed due to noise or outlier.
- Poor accuracy is also observed in testing set of data due to over fitting problem.

Over fitting can be avoided by pre pruning and post pruning. In pre pruning, a threshold is identified and goodness below a threshold is not splitted [25]. In post pruning, branches are removed by using different test set from the training set to get the best pruned tree[25].The bagging and boosting are the techniques for improving the classification accuracy. Bagging improves generalization performance by reducing variance of base classifier. If a base classifier is unstable ,bagging helps to reduce the error associated with the random fluctuation in training data .In boosting each classifier is dependent on the previous one and focuses on previous error by giving them more weights[25].

4. ANALYSIS

In this study data is taken for one year from the http:///www.wundergrounds.com.
A training set for the 64 instances was prepared from the dataset .Only three parameters has been taken into account(average temp, average humidity and sea level). Then 72 test instances are prepared by taking data randomly. Weka data mining tools has been downloaded from http:///www.kaz.dl.sourceforge.net/project/weka/weka3-6-windows-jre/3.6.9/weka-3-6-9jre.exe for the analysis purpose of predicting an event by decision tree on the basis of above mentioned parameters. Data cleaning was done so that classes of events are not complex. Result shows that out of 72 test instances, 46 tests were classified properly which gave an kappa statistics of .0584. Results can be further improved by taking more attributes in the model and increasing the training set data.

Table 1. Confusion Matrix.

A	B	C	D	E	F	Classified As
46	0	0	0	1	0	A=Fog
0	0	0	0	0	0	B=Null
22	0	0	0	3	0	C=Rain
0	0	0	0	0	0	D=Thunderstorm
0	0	0	0	0	0	E=Rain with other event
0	0	0	0	0		F=Fog with other event

5. CONCLUSION

To tap the potential of huge amount of data, decision tree can be used in predicting the dependent variable like fog and rain. Software equipped with decision tree can provide artificial intelligence to the machine. Such software can be used by trekkers, mountaineers and drivers, which can facilitate them in decision making, because many location exists where telemetric data does not exist and if it exists, it is too much location sensitive, so generalizing to a vast geographical are still remains doubtful. In this information age it is quite cost effective to equip a machine with the sensors and artificial intelligence software so that machine exhibits intelligence.

6. REFERENCES

[1] Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

[2] Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers

[3]Varun Chandola et al,(2009) Anamoly detection survey , ACM Computing Surveys, Vol. 41(3), Article 15.

[4] Breiman, Leo; Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8.

[5] Mahesh V. Joshi, Vipin Kumar, and Ramesh C. Agrawal, Evaluating boosting algorithms to classify rare classes, First IEEE International Conference on Data Mining.

[6] Friedman, J. H. (1999). Stochastic gradient boosting. Stanford University.

[7] Hastie, T., Tibshirani, R., Friedman, J. H. (2001). The elements of statistical learning : Data mining, inference, and prediction. New York: Springer Verlag.

[8] Rodriguez et al. (2006), Rotation forest : A new classifier ensemble method, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(10).

[9] Pages downloaded from http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[10] Pages downloaded from http://www-users.cs.umn.edu/~kumar/papers/papers.html

[11] Hyafil, Laurent; Rivest, RL (1976). "Constructing Optimal Binary Decision Trees is NP-complete". Information Processing Letters 5 (1): 15–17. doi:10.1016/0020-0190(76)90095-8.

[12] Shyam Boriah, Varun Chandola and Vipin Kumar,(2008),Similarity measure of categorical datas, In Proceedings of SIAM Data Mining Conference, Atlanta, GA.

[13] Principles of Data Mining.(2007). doi:10.1007/978-1-84628-766-4. ISBN 978-1-84628-765-7.

[14] Horváth, Tamás; Yamamoto, Akihiro, eds. (2003). Inductive Logic Programming. Lecture Notes in Computer Science 2835.doi:10.1007/b13700. ISBN 978-3-540-20144-1.

[15 Anurag Srivastava et al. Parallel formulation of decision trees and classification algorithms,Kluwer academic publishers.

[16] pages downloaded from http://en.wikipedia.org/wiki/Decision_tree_learning

[17] Varun Chandola, Shyam Boriah, and Vipin Kumar,(2009) In Proceedings of SIAM Data Mining Conference, Sparks.

[18] Breiman, L. (1996). Bagging Predictors. "Machine Learning, 24": pp. 123-140.

[19] Pages downloaded from http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap4_basic_classification.pdf

[20] Anurag Srivastava,E hong Han,Vipin Kumar,Vineet singh Kluwer academic publishers, Boston,'parallel formulation of decision tree- classification algorithms.

[21] pages downloaded from http://citeseer.ist.psu.edu/oliver93decision.html

[22] Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

[23 Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,(1996) ,"To Knowledge Discovery in Databases" 6, American Association for Artificial Intelligence. AI Magazine Volume 17 Number 3.

[24] pages downloaded from http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-1.html

[25] Jiawei Han and Michelin Kamber ,"(2006) Data Mining Concepts and Techniques",2nd edition, Morgan Kaufman.