



DATA MINING CLASSIFICATION ALGORITHMS FOR BUSINESS INTELLIGENCE

Dr. Arumugam P, Jose P
Department of Statistics,
Manonmaniam Sundaranar University,
Tirunelveli, Tamil Nadu, INDIA

Abstract: The most useful data mining techniques in Business process is classification. Classification is a supervised machine learning procedure and an important issue in data mining. Decision Trees are considered to be one of the most popular approaches for representing classifiers. In this paper, the classification task is used to predict the market scenario and as there are many approach that is used for data classification. It is an important data mining Technique has found immense uses in medical, market research statistics, educational and business intelligence. Mining to classify the various kinds of data used in nearly every field of our life. Classification is used to classify the item according to the features of the item with respect to the predefined set of classes. This paper based on cost performance evaluation on the correct and incorrect instances of data classification using J48, Random forest, Logistic Model tree classification algorithm. WEKA open source software was used to obtain the classification results of these algorithms. These classification methods are used to classify samples from different datasets. In this paper White industry Business data be used for classification and the result is compared by various datasets.

Keywords: Data mining, Classification, J48 a decision tree algorithm, Random Forest, Logistic Model Tree.

I. INTRODUCTION

Data mining has a great deal of attraction and attention in the Business Industry and in society as a whole in recent years, very huge amount of data availability and the imminent need for turning such type of Big data into useful informative data and retrieve knowledge, rich data. To the gain of information retrieval and knowledge used for many applications like business market analysis, fraud detection and customer retention, to production control and science exploration. Many data mining tools contributing greatly to business strategies, the effective and efficient analysis of data in such different forms becomes a challenging task. Although there are many data mining systems on the market, used for data analysis system performing data mining, interesting knowledge, regularities or high level information can be extracted, viewed from different aspects. The discovered knowledge can be applied to decision making. Rich business databases are holding with hidden information can be used business intelligent decision making. Classification and prediction are important data analysis models. It describing important data classes or to predict future data trends from the business databases. This process widely used for better understanding of large data.

Many classification and prediction methods have been proposed in machine learning, pattern recognition and statistics. Now a day's developing scalable classification and prediction techniques capable of handling large disk resident data. Some of the basic techniques for data classification prove as how to build decision tree classifiers. The types of classifiers are Bayesian, Bayesian belief networks and rule based classifiers. These types of classification and prediction application helps business strategy in the way of target marketing, performance prediction, manufacturing depends upon prediction. A marketing manager at all electronics data analysis to guess whether a customer with a given profile will suitable for product description. They can buy a electronic items. This data analysis task classifier constructs in to predict categorical label "yes or "no" for the marketing data. Suppose that the sales manager like to predict how much a given customer will spend during a sale at all electronics. Which model performed by, and in that particular model which color contributes more that process is analyzed by various classification algorithm. The classification process is structured in fig .1

The business data classification process is a 2 step process.

1. Learning step
2. Classification

1. Learning step : Describing a predetermined set of data classes or concepts. This classifier "learning from a training set made up of sales database tuples and associated with class labels. Classification process also be viewed as the learning of a mapping or function

$$Y=f(x)$$

$x \rightarrow$ input

$y \rightarrow$ prediction the class label.

2. Classification: Test data are used to estimate the accuracy of the classification rules if the accuracy is acceptable, rules can be applied to the classification of new data tuples.

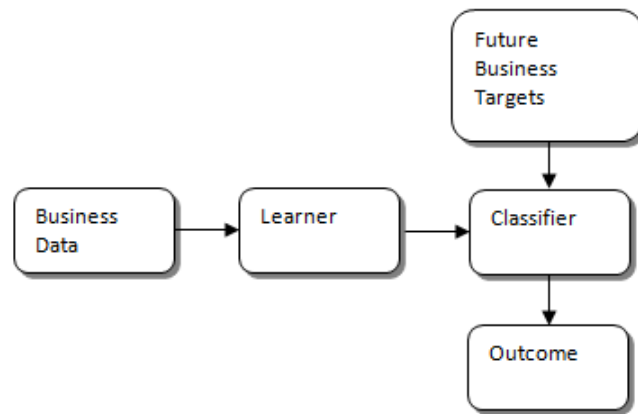


Fig. 1. Classification Process

II. CLASSIFICATION METHODS

J48 algorithm

J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple[1][3]

INPUT

$D \rightarrow$ Training data

OUTPUT

T // Decision tree

ALGORITHM

DTBUILD (*D)

Begin

$T = \emptyset$;

T = Create root node and label with splitting attribute;

T = Add arc to root node for each split predicate and label;

For each arc do

D = Database created by applying splitting predicate to D ;

If stopping point reached for this path, then

T' = create leaf node and label with appropriate class;

Else

$T' = \text{DTBUILD}(D)$;

T = add T' to arc;

End

J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is a program that creates a decision tree based on a set of labeled input data. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

Advantages:

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.
- It uses reduced error pruning

Random Forest:

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and output the class that is the mode of the class output by individual trees. The term came from random decision a forest that was first proposed by Tin Kam Ho of Bell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features. trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. Each tree is grown as described.

Random forests are an ensemble method for Classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler,¹ and "Random Forests" is their trademark. In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows: Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the k^{th} tree.

Each tree is grown as follows:

1. If the number of cases in the training set is N , sample N cases at random - but *with replacement*, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number $m < M$ is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning

INPUT

$N \rightarrow$ Training cases N ,

$M \rightarrow$ number of variables in the classifier

ALGORITHM

Begin

For each tree choose training dataset D

To determine the decision at a node of the tree;

$m < M$

consider a bootstrap sample

Build the next split

Compute Gini Index

Calculate prediction error

Calculate the best split

Advantages

- It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
- It runs efficiently on large databases.
- It can handle thousands of input variables without variable deletion.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

Logistic Model Tree:

A logistic model tree basically consists of a standard decision tree structure with logistic regression functions at the leaves, much like a model tree is a regression tree with regression functions at the leaves. As in ordinary decision trees, a test on one of the attributes is associated with every inner node. For a nominal (enumerated) attribute with k values, the node has k child nodes, and instances are sorted down

one of the k branches depending on their value of the attribute. For numeric attributes, the node has two child nodes and the test consists of comparing the attribute value to a threshold an instance is sorted down the left branch if its value for that attribute is smaller than the threshold and sorted down the right branch otherwise.

More formally, a logistic model tree consists of a tree structure that is made up of a set of inner or non-terminal nodes N and a set of leaves or terminal nodes T . Let S denote the whole instance space, spanned by all attributes that are present in the data. Then the tree structure gives a disjoint subdivision of S into regions S_t , and every region is represented by a leaf in the tree

$$S = \bigcup_{t \in T} S_t, \\ t \cap S_t = \emptyset \text{ for } (t \neq t')$$

And the regression is a better way to use regression for classification tasks is to use a Logistic regression model that models the posterior class probabilities $\Pr(G=j|X=x)$ for the j classes. Given estimates for the class probabilities, can classify unseen instances by

$$j^* = \operatorname{argmax} \operatorname{pr}(G=j|X=x)$$

LMT uses the CART cost complexity pruning method. This is a powerful (but computationally expensive) pruning method that produces very compact trees. If LMT hasn't produced a tree structure then there is no non-linear structure that can be exploited to improve performance over a single linear regression. LMT is extremely slow, so we want to make sure that we don't try to run it on large files.

INPUT

$T \rightarrow$ Training Data set

ALGORITHM

Begin

Training dataset T

Growing the tree depends splitting criterion of attribute

Stop the splitting criterion. If the node is not split

Building the logistic model tree with attribute selection

Prune with minimize error

End

Advantages

- LMT produces a single tree containing binary splits on numeric attributes, multiway splits on nominal ones, and logistic regression models at the leaves.
- The algorithm ensures that only relevant attributes are included.
- LMT uses cost complexity pruning.

The following dataset selected for tests based on the performance comparison of three classification algorithms In this paper, we have used weka (Waikato environment for knowledge analysis) tool for comparison of J48, Random Forest, LMT algorithm and calculating efficiency based on accuracy regarding correctly classified instances and incorrectly classified instances. And

III. PERFORMANCE ANALYSIS

Table 1. Dataset for the Test 1.

Data set	Instances	Attributes	Classes
Letter	20000	16	26
Nursery	12960	8	5
Optical digits	5620	64	0
Wave form	5000	40	3
spam	4601	58	2
German	1000	20	2
Segment	2310	19	7
Sick	3772	30	2

Table 2. Dataset for the Test 2.

Data set	Tree size		
	J48	Random Forest	LMT
Letter	1423	No mem	No mem
Nursery	369	2401	155
Optical digits	247	2809	
Wave form	187	3005	1
spam	161	1681	39
German	64	1279	1
Segment	59	543	5
Sick	45	427	24

Table3.Experimental data

Variable	Description	Possible Values
B No	Identification Number of the bill	1234PHB
B Date	The date which bill is coated	MM.DD.YYYY
Cus Name	Name of the customer	Peters & Co
Mod des	Description of each model	WM7865
Pdt	Name of the product	Washing Machine
Pdt Cod	Code description of the product	Silvergreen123
Total	The total number of count	Numeric
LOB	Line of business	

These classification algorithms can be implemented on different types of data sets like data of Letter, Nursery, Optical digits, Wave form, spam, German credit, Segment, Sick data according to performances. On the basis of the performance of these algorithms, these algorithms can also be used to detect the Sales of particular product in future, prediction of customer behavior etc. Based on the features present in the class labeled training data, develop a description or model for each class. It is used for better understanding of each class, and prediction of certain properties of unseen data

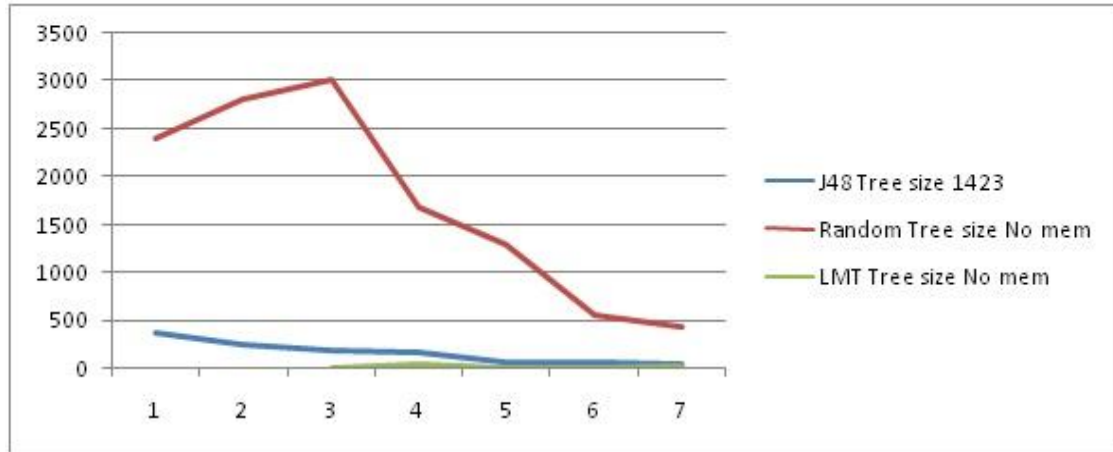


Fig. 2: Comparison analysis

Tested each data with three different classification tree algorithm J48, Random forest, Logistic model trees. For each algorithm both of the test options Percentage split and cross validation used. For the testing of percentage split 66% training set of the instances and remaining part considered for test set. Cross validation reserving for test part the standard 10 fold cross validation is used. Finally the error rate estimates on the holdout set. For j48, Random forest, LMT all the test were run with 10 random seeds..for the LMT restricts the random seed as 5.because it consumes lot of time for execution. J48, Random forest, LMT executes within a minute. But for the LMT algorithm one run could take several times. And Random forests have been observed to over fit for datasets with noisy classification/regression tasks. Random forest is not reliable for some type of data and it builds the largest trees with lowest overall performance. Finally LMT builds smallest classification tree. This process indicates the cost complexity prune. And it produces better performance of more numerical data sets. The result shows form various implementation of classification algorithm cost complexity pruning has a better performance.

IV. Conclusion

Data Mining could be used to improve business intelligence process. This paper shows the implementation result of J48, Random forest, LMT classification algorithm with sales data and the test is comparative with set of other domains. The experiments focus the cost complexity prune of the classification tree and size of the tree. It enhances the efficiency and overall efficiency by optimally utilizing the resources available. The performance, success of sales person in the target review as well as their overall business development could be exponentially accelerated by thoroughly utilizing Data Mining technique to evaluate their product quality and finally the achievement percentage. This is the obvious conclusion of the present investigation. It's that is achieved by the comparative analysis of Classification algorithm. As an experimental result shows LMT produced better performance on data sets with many numerical attributes than other classification algorithms described in this work.

V. REFERENCES

1. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
2. Jiawei han, Micheline Kamber, Data Mining concepts and Techniques, 2nd edition, 2006
3. Insight into Data Mining Theory and practice K.P Soman, shyam Diwakar, and v.Ajay, 2006
4. J.Quinlan, c4.5 Programs for Machine Learning, Morgan Kaufmann publishers, 1993
5. L.Breiman, J.H.Friedman, R.Aolshen and C.J Stone, Classification and Regression Trees, California 1984.
6. Ian H.witten and Eibe Frank. Data Mining : practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., san Francisco, CA, USA, 2nd Edition, 2005.