

# A Succinct Reflection on Data Classification Methodologies

Divyanka Hooda<sup>1</sup>, Divya Wadhwa<sup>2</sup>, Hardik Singh<sup>3</sup>, Anuradha<sup>4</sup>

<sup>1</sup>Computer Science, ITM University,  
Gurgaon, Haryana-122001, India.

<sup>2</sup>Computer Science, ITM University,  
Gurgaon, Haryana-122001, India.

<sup>3</sup>Computer Science, ITM University,  
Gurgaon, Haryana-122001, India.

<sup>4</sup>Computer Science, ITM University,  
Gurgaon, Haryana-122001, India.

## Abstract

Classification is a data mining (machine learning) technique used to assign group membership to various data instances. Indeed there are many classification techniques available for a scientist wishing to discover a model for his/her data. This diversity can cause trouble as to which method should be applied to which data set to solve a particular domain concentrated problem. This review paper presents several major classification techniques like Decision Tree Induction, Bayesian Classification, Rule-based Classification, classification by Back Propagation, Support Vector Machines, Lazy Learners, Genetic Algorithms, Rough Set Approach, and Fuzzy Set Approach. The goal of this survey is to provide a comprehensive review of different data classification techniques.

**Keywords:** Classification, Decision tree, SVM, Bayesian Classifier, Rule-Based Learning.

## 1. Introduction

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. We locate data into specific categories for its most effective and efficient use, then we call it data classification. In the bank credit data given in table 1, we can classify each customer into two classes (Fraud/No Fraud) depending on age, mortgage and income. In more technical terms, old data is classified and models are made for the prediction of classes of an object on the basis of some specific attributes.

Table 1: Bank Credit Data

S.NO	AGE	MORTGAGE	INCOME	CLASS
1	22	Yes	21,000	Fraud
2	27	Yes	15,000	Fraud
3	26	No	40,000	Fraud
4	29	Yes	27,000	Not fraud
5	18	No	13,000	Fraud

After applying a suitable classification technique, we can predict whether it would be safe for the bank to give loan or not. Every classification varies from the other on the basis of various parameters like classification accuracy, standard error rate, time and space complexity and many more. Decision tree classification is one of the famous classification technique, which gives better visualization of trained model in the form of a tree as given below in Fig 1 if applied to the data given in Table 1. Internal nodes presents various tests to be conducted on the data fields and leaf node tells us about the class label. For example, if age of a customer is above 25 years and income is less than 20,000 than it would not be advisable to give loan to that customer. Because, as we can see from tree model that the tree branch corresponding to that customer is ending at fraud class.

### 1.1 Pseudo Code for Data Classification

Given below is the algorithmic flow to classify data set. Any given data set is classified on the basis of one of its attributes, which are having critical importance over the others. Every time recursively partitioning the given data set into smaller portions, which are easy to classify depending upon whether that portion is having same class label or not. We will stop this process either all attributes are testing for their significance or pure portions are there. Fig. 1 represents application of following algorithm on the assumed data set.

1. Goal is to classify a data set into categories such that  $D(n) \leftarrow C_i$ , where C represents a class and  $i \rightarrow 1$  to m.
2. Input original data set i.e., D (n), where n is the number of records.
3. Find information gain i.e., Info\_Gain of each attribute  $A_j$ , where  $j \rightarrow 1$  to p.
4. Choose  $A_j$  with max (Info\_Gain), select  $A_j$  as root node to start.
5. Partition D (n) on all possible values of  $A_j$  (i.e., k) such that  $D(n) \leftarrow P_k$ , where P is the partition and  $k \leftarrow 1$  to q.
6. If  $D(n) \leftarrow P_k$  consists of pure class, then stop.
7. Else repeat steps 2 to 6, till the time no more attributes left to partition.
8. End.

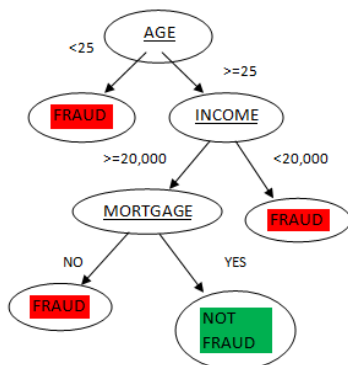


Fig. 1 Tree Generated by decision tree classifier.

### 2. Different Types of Classifiers

In the literature, there are various classifiers, each of which works in their unique way. For example, some generate rules for classification (rule based classification), some use trees (decision trees generating training model in form of a tree), fuzzy set (using truth values between 0.0 and 1.0) and many other.

Figure 2 represents some of the techniques for data classification, each incorporating different classifiers.

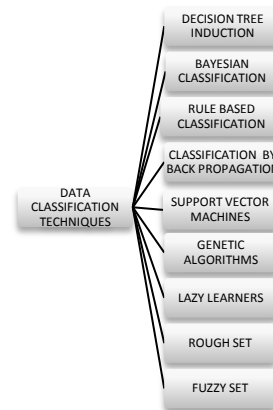


Fig. 2 Different Data Classification Techniques.

#### 2.1 Decision Tree Induction

A tree is a graph without cycles, so a decision tree is a structure where, root node is the parent (topmost) node, having highest information gain, defining the favorable sequence of attributes to investigate a domain centered problem. Internal nodes do testing on an attribute. Branch represents the outcome of the test. Leaf node holds the class label. Table 2 gives famous algorithms for the same.

Table 2: Decision tree induction

ALGORITHM	AUTHOR	DESCRIPTION
ID3 [11][13]	QUINLAN, 1983	Uses Information Gain as splitting criteria.
C4.5 [9]	QUINLAN, 1993	Evolution of ID3. Uses gain ratio for splitting.
CART [10][12]	Breiman et al., 1984	Construction of Binary trees for classification.
CHAID [2]	Kass, 1980	Nominal attributes are handled statistically.
QUEST [2]	Loh and Shih, 1997	Supports linear combinational splits.
CAL5 [2]	Muller, Wyszotzki, 1994	For numerical - valued attributes.
FACT [2]	Loh, Vanichsetakul 1988	Uses statistical and Discriminant analysis.
LMDT [2]	Brodley, Utgoff, 1995	Multivariate tests are used on attributes.
T1 [2]	Holte, 1993	One-level decision tree is used.
PUBLIC [2]	Rastogi, Shim, 2000	Integrates the growing and pruning.
MARS[2]	Friedman, 1991	Multiple regression is approximated.

## 2.2 Bayesian Classification

Naïve Bayesian classification consists of supervised learning algorithms (classifiers) that take a training sample as an input and returns a general classification rule as the output. A Naïve Bayesian classifier is a simple probabilistic classifier which applies Bayesian theorem with strong (or naive) independence assumptions.

### Baye's theorem:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad (1)$$

Here, x and y are two different events. Baye's theorem defines a relationship (as shown) between the probabilities of x and y i.e., P(x) and P(y) respectively, and the conditional probabilities of x given y has already occurred i.e., P(x|y) and y given x has already occurred i.e., P(y|x). Table 3 gives important Bayesian classifiers.

Table 3: Bayesian Classification

ALGORITHM	AUTHOR	DESCRIPTION
Maximum A Posteriori (MAP)[15][16]	Dempster, Laird, et al. - 1977	$c_{map} = \arg \max_{c \in C} p(c e)$ or $\arg \max_{c \in C} p(c) \cdot p(e c)$
Maximum Likelihood (ML)[14]	Redner, Walker - 1984	$c_{mp} = \arg \max_{c \in C} p(e c)$
Naïve Bayes Classifier[17]	Thomas Bayes, 1763	Strong independent assumptions are used.

## 3.3 Rule-Based Classification

The extraction of useful if-then rules from data based on statistical significance or a set of rules (which will be written on the basis of some priority) are used in rule-based classification. Certain association based relationships are applied on a set of objects in a database in association rule algorithms. For example if a rule having an expression of the form X Y, and if a transaction of the database contain X, then it will tend to contain Y. The Apriori association rule based algorithm was developed on this type of relationship for rule mining in large transaction databases by IBM's Quest project team [19]. Table 4 represents rule based classification algorithms.

Table 4: Rule based classification

ALGORITHM	AUTHOR	DESCRIPTION
FOIL[3]	Quinlan and Cameron-Jones, 1993.	Repeatedly searches for the current best rule.
AQ[24]	Ryszard S. Michalski, late 1960s.	Induces a set of rules from relations in Prolog.
PRM[3]	Xiaoxin Yin and Jiawei Han 2003.	Modifies FOIL to achieve higher accuracy.
CPAR[3]	Xiaoxin Yin and Jiawei Han, 2003	Builds rules by adding literals one by one.
RIPPER[4]	Cohen, 1995	Generates "key words spotting rules".
1R[26]	Holte, 1993	Simplest algorithm for discrete attributes

## 3.4 Classification by Back Propagation

A Back propagation object specifies the parameters used by the back propagation learning algorithm. In back propagation, the output error on the training examples is used to adjust the network weights. For better learning of the algorithm, it is divided into two phases: propagation and weight update. First the weights are initialized in the network. Then for each training set, error is computed at its output by the following formula:

$$E = (O_{exp} - O_{actual})^2 \quad (2)$$

Here E is the discrepancy. Lastly, weights are updated until all examples are classified correctly.

## 3.5 Support Vector Machines

The basic goal of a SVM classifier is to predict, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Vapnik & Chervonenkis' statistical learning theory in 1960s laid the groundwork for Vapnik to release SVM approach in 1992.[21]

Linear SVM:

D is training data having a set of n points of the form:

$$D = \{(t_i, y_i) | t_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n \quad (3)$$

Where  $y_i$  is either -1 or 1, indicating the class to which the point  $t_i$  belonging. Table 5 is the representation of SVM algorithms.

Table 5: Support vector machines

<i>ALGORITHM</i>	<i>AUTHOR</i>	<i>DESCRIPTION</i>
SVM-RS[5]	Kang and Yoo, 2007	Binarizing available user preference data.
TV PROGRAM RS[6]	Xu and Araki, 2006	Electronic Program Guide as features.
SSVM[7]	Xia et al, 2006	Estimate missing elements in matrix.
CSVM[8]	Oku et al, 2006	For context-aware RS (CSVM).

### 3.6 Lazy Learners

Lazy learning algorithms are an instance-based algorithm that stores only the training data or minor processing, and waits until it is given a test tuple. Instance-based learning algorithms are lazy-learning algorithms (Mitchell, 1997), as they delay the induction or generalization process until classification is performed. These algorithms require less computation time during the training phase (than eager-learning algorithms such as decision trees, neural and Baye’s networks) but more computation time during the prediction process. Some of the lazy learner’s algorithms are represented in Table 6.

Table 6: Lazy learners

<i>ALGORITHM</i>	<i>AUTHOR</i>	<i>DESCRIPTION</i>
<i>K</i> -nearest neighbor[22]	Trevor Hastie, Robert Tibshirani, 1994	Instances represented as points in a Euclidean space.
Racing algorithm[18]	Oded Maron, Andrew Moore, 1997	Provides better allocation of computational resources among candidate configurations.
Locally weighted regression	Cleveland and Devlin, 1988	Constructs local approximation, to minimize the weighted error.

### 3.7 Genetic Algorithms

GA is based on an analogy to biological evolution. Each rule is represented by a string of bits. An initial population is created consisting of randomly generated rules. e.g., IF A1 and Not A2 then C2 can be encoded as 100 based on the notion of survival of the fittest; a new population is

formed to consist of the fittest rules and their off springs. The fitness of a rule is represented by its classification accuracy on a set of training examples.  $t$ . The schema theorem specifies the (expected) number  $X(s, t+1)$  of chromosomes carrying schema  $s$  in the next generation. A simplified version has the following form:

$$X(s, t + 1) = u(s, t)[1 - e]X(s, t) \quad (4)$$

Where  $u(s, t)$  is the average fitness of the chromosomes carrying schema  $s$  at time  $t$  (the observed average fitness), and  $e$  is the overall probability (usually quite small) that the cluster  $s$  will be destroyed (or created) by mutation or crossover [23].

### 3.8 Rough Set

This theory can be used to classify imprecise or noisy data and to discover structural relationships within them. It applies to discrete-valued attributes. Continuous-valued attributes must therefore be discredited before its use. Z. Pawlak [20] first released this approach to data analysis in 1982.

### 3.9 Fuzzy Set

Rather than having a precise cut-off between categories, fuzzy logic uses truth values between 0.0 and 1.0 representing the degree of membership that a certain value has in a given category. Each category represents the fuzzy set. Fuzzy set theory is also known as possibility theory [24]. A fuzzy set is a pair  $(U, m)$  where  $U$  is a set and  $m : U \rightarrow [0,1]$ .

## 4. Conclusion

In this survey, we have tried to give an overview of Decision Tree Induction, Bayesian Classification, rule based Classification, classification by Back Propagation, Support Vector Machines, Lazy Learners, Genetic Algorithms, Rough Set Approach, and Fuzzy Set Approach. Every technique, however, has its own pros and cons like nowadays mostly research is done in Support Vector Machines (SVM). Thus, we have given an insight on the different techniques used in machine learning for data classification.

## 5. References

- [1.] U. Fayyad and K. Irani, On the handling of continuous-valued Attributes in decision tree generation. Machine Learning, Kluwer Academic Publishers, Boston , vol 8, page 87-102 (1992)
- [2.] Almuallim H., An Efficient Algorithm for Optimal Pruning of Decision Trees. Artificial Intelligence vol 83(2): page 347-362, 1996

- [3.] Prafulla Gupta & Durga Toshniwal Performance Comparison of Rule Based Classification Algorithms, International Journal of Computer Science & Informatics, Volume-I, Issue-II, Page 37-42, 2011
- [4.] Cohen, W. W. (1995). Fast effective rule induction. In Machine Learning: Proceedings of the Twelfth International Conference, 1995
- [5.] Kang, H., and Yoo, S., Svm and collaborative filtering-based prediction of user preference for digital fashion recommendation systems. IEEE Transactions on Inf & Syst, 2007.
- [6.] Xu, J., and Araki, K., A svm-based personal recommendation system for TV programs. In Multi-Media Modeling Conference Proceedings, 2006.
- [7.] Xia, Z., Dong, Y., and Xing, G., 2006 Support vector machines for collaborative filtering. In ACMSE 44: Proceedings of the 44th annual Southeast regional conference, pages 169–174, New York, NY, USA. ACM.
- [8.] K. O. et al., Context-aware SVM for context-dependent information recommendation. In International Conference on Mobile Data Management, 2006.
- [9.] Radaideh, Q., August 1989. The Impact of Classification Evaluation Methods on, Rough Sets Based Classifiers, Proceedings of the 2008 International.
- [10.] Crawford S. L., August 1989 Extensions to the CART algorithm. Int. J. of Man Machine Studies, vol 31(2): pages 197-217.
- [11.] Mehmed Kantardzic (2003): Review of Data Mining: Concepts, Models, Methods, and Algorithms. Technometrics vol. 45, no. 3, p. 277-277
- [12.] Breiman L., Friedman J., Olshen R., and Stone C., 1984. Classification and Regression Trees. Wadsworth Int. Group. Chapman & Hall, New York, NY, (1984)
- [13.] Ilim H., An Efficient Algorithm for Optimal Pruning of Decision Trees. Artificial Intelligence 83(2): 347-362, 1996.
- [14.] Dempster, A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm" J. Roy. Statist. Soc. B, 39: pp 1–38, 1977.
- [15.] Cheeseman et al., 1988a] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, 1988. Autoclass: a Bayesian classification system. In Proceedings of the Fifth International Conference on Machine Learning.
- [16.] [Cheeseman et al., 1988b] P. Cheeseman, M. Self, J. Kelly, J. Stutz, W. Taylor, and D. Freeman, 1988. Bayesian classification. In Seventh National Conference on Artificial Intelligence, pages 607–611, Saint Paul, Minnesota.
- [17.] Tan P-N, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison-Wesley, 2nd Edition 2006 by Pearson Education Inc.
- [18.] D.W. Aha, A Study of Instance-Based Algorithms for Supervised Learning Tasks: Mathematical, Empirical and Psychological Evaluations. PhD Thesis; Technical Report No 90-42, University of California, Irvine, November 1990.
- [19.] Agrawal, R. Srikant: "Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept. 1994.
- [20.] Z. Pawlak, "Rough sets", International Journal of Computer and Information Sciences, page 341-356, 1982.
- [21.] V. N. Vapnik, Statistical Learning Theory, Wiley New York, 1998.
- [22.] Kubat, Miroslav, Cooperson Martin, A reduction technique for nearest-neighbor classification: Small groups of examples. Intell. Data Anal vol 5 edition 6 page 463-476, 2001.
- [23.] D.E. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley, New York, 1989.
- [24.] J.R. Castro, O. Castillo and L.G. Martínez, "Interval Type-2 Fuzzy Logic Toolbox", *Engineering Letters Vol 15(1)*, 2007, page 89-98.
- [25.] R. S. Michalski, I. Mozetic, J. Hong, and N. Lavrac, The AQ15 inductive learning system: An overview and experiments, Report 1260, Department of Computer Science, University of Illinois at Urbana-Champaign, 1986.
- [26.] Craig G. Nevill-Mainning Artificial Neural Network and Expert Systems, Second New Zealand International Two-Stream Conference on 20-23 Nov 1995 Pages 239 – 242.