A Survey of Modern Data Classification Techniques

Ravijeet Singh Chauhan[1]
[1]M. Tech Student
[1]SATI, Vidisha, Madhya Pradesh, India

Abstract— In this paper, we present an overview of existing data classification algorithms. All these algorithms are described more or less on their own. Classification is a very popular and computationally expensive task. We also explain the fundamentals of data classification. We describe today's approaches for data classification. From the broad variety of efficient algorithms that have been developed we will compare the most important ones. We will systematize the algorithms and analyze their performance based on both their run time performance and theoretical considerations. Their strengths and weaknesses are also investigated. It turns out that the behavior of the algorithms is much more similar as to be expected.

I. INTRODUCTION

Decision tree learning, used in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. This page deals with decision trees in data mining.

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 variable based on several input variables. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. 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 (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data, but it is not the only strategy. In fact, some approaches have been developed recently allowing tree induction to be performed in a bottom-up fashion.

II. RELATED WORK

Decision tree induction algorithms have been used for classification in many application areas, such as medicine, manufacturing and production, financial analysis, astronomy, and molecular biology. Tree-based learning methods are widely used for machine learning and data mining applications. These methods have a long tradition and are commonly known since the works of [2, 3 and 4]. They are conceptually simple yet powerful. The most common way to build decision trees is by top down partitioning, starting with the full training set and recursively finding a univariate split that maximizes some local criterion (e.g. gain ratio) until the class distributions the leaf partitions are sufficiently pure Pessimistic Error Pruning [4] uses statistically motivated heuristics to determine this utility, while Reduced Error Pruning estimates it by testing the alternatives on separate independent pruning set. In a decision tree learner named NB Tree is introduced that has Naive Bayes classifiers as leaf nodes and uses a split criterion that is based directly on the performance of Naive Bayes classifiers in all first-level child nodes (evaluated by cross-validation) an extremely expensive procedure[8]. In [7, 11] a decision tree learner is described that computes new attributes as linear, quadratic or logistic discriminate functions of attributes at each node; these are then also passed down the tree. The leaf nodes are still basically majority classifiers, although the class probability distributions on the path from the root are taken into account.

A recursive Bayesian classifier is introduced in [7]. Lots of improvement is already done on decision tree induction method for 100 % accuracy and many of them achieved the goal also but main problem on these improved methods is that they required lots of time and complex extracted rules. The main idea is to split the data recursively into partitions where the conditional independence assumption holds. A decision tree is a mapping from observations about an item to conclusions about its target value [9, 10, 11, 12 and 13]. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. A decision tree (or tree diagram) is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [14]. Decision tree Induction Method has been successfully used in expert systems in capturing knowledge. Decision tree induction Method is good for multiple attribute Data sets.

III. CONCLUSION

In this paper, we surveyed the existing data classification techniques. We restricted ourselves to the classic classification problem.

In a forthcoming paper, we pursue the development of a novel classification algorithm that efficiently predicts the value of a target attribute.

REFERENCES

[1] Singh Vijendra. Efficient Clustering For High Dimensional Data: Subspace Based Clustering and Density Based Clustering, Information Technology Journal; 2011, 10(6), pp. 1092-1105.
[2] D Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. "Classification and Regression Trees". Wadsworth International Group. Belmont, CA: The Wadsworth Statistics/Probability Series1984.
[3] Quinlan, J. R. "Induction of Decision Trees". Machine Learning; 1986, pp. 81-106.
[4] Quinlan, J. R. Simplifying "Decision Trees. International Journal of Man-Machine Studies" ;1987, 27:pp. 221-234.
[5] Gama, J. and Brazdil, P. "Linear Tree. Intelligent Data Analysis",1999,.3(1): pp. 1-22.
[6] Langley, P. "Induction of Recursive Bayesian Classifiers". In Brazdil P.B. (ed.), Machine Learning: ECML-93;1993, pp. 153-164. Springer, Berlin/Heidelberg~lew York/Tokyo.
[7] Witten, I. & Frank, E,"Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.ch. 3,4, pp 45-100.
[8] Yang, Y., Webb, G. "On Why Discretization Works for Naive-Bayes Classifiers", Lecture Notes in Computer Science, vol. 2003, pp. 440 – 452.
[9] H. Zantema and H. L. Bodlaender, "Finding Small Equivalent Decision Trees is Hard", International Journal of Foundations of Computer Science; 2000, 11(2):343-354.
[10] Huang Ming, Niu Wenying and Liang Xu , "An improved Decision Tree classification algorithm based on ID3 and the application in score analysis", Software Technol. Inst., Dalian Jiao Tong Univ., Dalian, China, June 2009.
[11] Chai Rui-min and Wang Miao, "A more efficient classification scheme for ID3",Sch. of Electron. & Inf. Eng., Liaoning Tech. Univ., Huludao, China; 2010,Version1, pp. 329-345.
[12] Iu Yuxun and Xie Niuniu "Improved ID3 algorithm",Coll. of Inf. Sci. & Eng., Henan Univ. of Technol., Zhengzhou, China;2010,pp. ;465-573.
[13] Chen Jin, Luo De-lin and Mu Fen-xiang," An im pr oved ID3 decision tree algorithm",Sch. of Inf. Sci. & Technol., Xiamen Univ., Xiamen, China, page; 2009, pp. 127-134.
[14] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann, 2006, ch-3, pp. 102-130.