New Approaches for Geometric Decision Tree Performance

Vijendra C. Lad
Department of Computer Engineering,
SKNCOE,Pune.

Deepa Kulkarni
Department of Computer Engineering,
SKNCOE,Pune.

Abstract - **Decision trees have proved to be valuable tools for the description, classification and generalization of data. Work on constructing decision trees from data exists in multiple disciplines such as statistics, pattern recognition, decision theory, signal processing, machine learning and artificial neural networks. This paper proposes a modified geometric decision tree algorithm using boosting process. Decision Trees are considered to be one of the most popular approaches for performing classification. Classification tree produced by decisi on tress are simple understand and interpret. This paper presents modifications for constructing geometric decision tree classifier s using adaptive boosting. In this algorithm tree is constructed using the geometric structure of data. The algorithm evaluates th e best angle bisector (hyper planes) as split rule for decision tree structure. The multidimensional hyper planes help to builds small decision trees and gives better performance. Furthermore to enhance the performance, we proposed use an Adaptive Boosting met hod for boosting decision tree which will create multiple small geometric decision trees in the view of improving the performance of prediction in terms of accuracy, confusion matrix, precision and recall. The proposed modification to geometric decision tree is applied on standard data set from UCI repository.**

1. INTRODUCTION

Advances in data collection methods, storage and processing technology are providing a unique challenge and opportunity for automated data exploration techniques. Enormous amounts of data are being collected daily from major scientific projects (e.g., Human Genome Project, the Hubble Space Telescope, Geographical Information Systems), from stocks trading, from hospital information systems, from computerized sales records and other sources. In addition, researchers and practitioners from more diverse disciplines than ever before are attempting to use automated methods to analyze their data. As the quantity and variety of data available to data exploration methods increases, there is a commensurate need for robust, efficient and versatile data exploration methods. Decision trees are a way to represent rules underlying data with hierarchical, sequential structures that recursively partition the data. A decision tree can be used for data exploration in one or more of the following ways: Description: To reduce a volume of data by transforming it into a more compact form which preserves the essential characteristics and provides an accurate summary. **Classification:** Discovering whether the data contains well-separated classes of objects, such that the classes can be interpreted meaningfully in the context of a substantive theory.

Generalization: Uncovering a mapping from independent to dependent variables that is useful for predicting the value of the dependent variable in the future Automatic construction of rules in the form of decision trees has been attempted virtually in all disciplines in which data exploration methods have been developed. It has been tradi-tionally developed in the fields of statistics, engineering (pattern recognition) and decision theory (decision table programming). Recently renewed interest has been generated by re-search in artificial intelligence (machine learning) and the neurosciences (neural networks). Though the terminology and emphases differ from discipline to discipline, there are many similarities in the methodology.

Decision trees automatically constructed from data have been used successfully in many real-world situations. Their effectiveness has been compared widely to other automated data exploration methods and to human experts. Several advantages of decision tree-based classification have been pointed out. Knowledge acquisition from pre-classified examples circumvents the bottleneck of acquiring knowledge from a domain expert.[2] Tree methods are exploratory as opposed to inferential. They are also non-parametric. As only a few assumptions are made about the model and the data distribution, trees can model a wide range of data distributions. [2] Hierarchical decomposition implies better use of available features and computational efficiency in classification. [2] As opposed to some statistical methods, tree classifiers can treat uni-modal as well as multi-modal data in the same fashion. [2] Trees can be used with the same ease in deterministic as well as incomplete problems. (In deterministic domains, the dependent variable can be determined perfectly from the independent variables, whereas in incomplete problems, it cannot be.)[2] Trees perform classification by a sequence of simple, easy-to-understand tests whose semantics are intuitively clear to domain experts. The decision tree formalism itself is intuitively appealing. For these and other reasons, decision tree methodology can provide an important tool in every data mining researcher/practitioner's tool box. In fact, many existing data mining products are based on constructing decision trees from data.

1.1Terminology

Structures similar to decision trees have been called classification trees, branched testing sequences, discriminant trees and identification keys. Training sets consist of objects, also known as samples, observations, examples or instances. Attributes have been referred to as features, predictors or independent variables. In an ordered attribute space, a decision tree imposes a partitioning that can be geometrically represented as a collection of hyper-surfaces and regions. Much of the work on decision trees uses only a specific type of surface, namely hyper-planes. (For exceptions, see the Neural Trees and Other Methods paragraphs in Section 3.2.) For this reason, splits are often referred to as hyper-planes, attributes as dimensions and objects as points. category or dependent variable is the same as class label. Ordered domains are equiva-lent to or comprise continuous, integer, real-valued and monotonous domains. Unordered domains have categorical, discrete or free variables. Internal nodes are the same as non-terminals or test nodes. Leaf nodes are referred to as the terminal nodes or decision nodes. Goodness measures are also known as feature evaluation criteria, feature selection criteria, impurity measures or splitting rules.

2. RELATED WORK

2.1 Geometric Decision Tree :

The performance of any top-down call tree algorithmic rule depends on the live accustomed rate totally different hyper planes at every node. The problem of getting an appropriate algorithmic rule to search out the hyper plane that optimizes the chosen rating perform is additionally necessary. As an example, for all impurity measures, the optimization is tough as a result of finding the gradient of the impurity perform with relation to the parameters of the hyper plane isn't doable. Driven by these concerns, here, we have a tendency to propose a replacement criterion perform to assess the suitableness of a hyper plane at a node which will capture the geometric structure of the category regions. For our criterion perform, the optimization downside also can be solved additional simply. We have a tendency to initial make a case for our technique by considering a two-class downside. Given the set of coaching patterns at a node, we have a tendency to initial realize 2 hyper planes, i.e., one for every category. Every hyper plane is specified it's nearest to all or any patterns of 1 category and is farthest from all patterns of the opposite category. We decision these hyper planes because the clump hyper planes (for the 2 classes). Thanks to the manner they're outlined, these clump hyper planes capture the dominant linear tendencies within the samples of every category that are helpful for discriminating between the categories. Hence, a hyper plane that passes in between them may be sensible for rending the feature house. Thus, we have a tendency to take the hyper plane that bisects the angle between the clump hyper planes because the split rule at this node. Since, in general, there would be 2 angle bisectors, we elect the bisector that's higher, supported associate impurity live, i.e., the Gina index. If the 2 clump hyper planes happen to be parallel to every different, then we have a tendency to take a hyper plane midway between the 2 because the split rule. As are often seen, though this hyper plane promotes the (average) purity of kid nodes, it doesn't very alter the classification downside; it doesn't capture the isobilateral distribution of sophistication regions during this problem. {The 2|the 2} clump hyper planes for the 2 categories and also the two angle bisectors, obtained through our algorithmic rule, at the foundation node on this downside. As are often seen, selecting any of the angle bisectors because the hyper plane at the foundation node to separate the info.

2.2 Mathematical Model

1] Identify the input data set as

S= {s1, s2, s3, s4…..}

Where S - Input data set

2] Identify the data classes

CL = {cl1, cl2, cl3, cl4….}

Where CL - The data class

Process:

Given a set of data points at a node, we find hyper-planes that have most of the data points of same class. Find the

best hyper-plane to split the data.

Let C1 be the set of points contains points of the majority class

C2 be the set of points contains points of the remaining classes

Let

A be the matrix containing points of C1

B be the matrix containing points of C2

Let

W1 be the first hyper-plane

W2 be the second hyper-plane

Let

W3 be the first bisecting angle

W4 be the second bisecting angle

Such that

W3 = W1 + W2

W4 = W1 – W2

Choose the angle bisectors having lesser Gini index G

$$Gini(\widetilde{w_t}) = \frac{m^{tl}}{m^t}[\,1-({}_1^{tl}m/m^{tl})^2\ -({}_{-1}^{tl}m/m^{tl})^2 + \frac{m^{tr}}{m^t}[\,1-({}_1^{tr}m/m^{tr})^2 - ({}_{-1}^{tr}m/m^{tr})^2\,]$$

3. PROPOSED WORK

3.1 Adaboost Algorithm:

1. Build distribution D_1 , assuming all samples equally important

2. For **t= 1,….,T** (rounds of boosting)

- Select weak classifier with the lowest error from a group

- Check if error larger than ½

(Yes:terminate ; NO: go on)

-Calculate confidence parameter , weight of sub-classfier:

$$a_{t=\frac{1}{2}In(\frac{1-\in t}{\in t})>0}$$

- Re-weight data samples to give poorly classified samples an increased weight

$$D_{t+1=\frac{Dt}{Zt}\begin{cases} e-at\ if\ yi=h_{t(xi)} \\ e^{at\ if\ yi\neq h_{t(x_i)}} \end{cases}}$$

Where is the normalization factor 3 .

At the end (**tth round**), the final strong classifier results

$$H_{final}(x) = sgn(\sum_t a_t h_t(x))$$

3.2 Experimental setup

In this section, we present experiential results toshow the effectiveness of our decision tree algorithm. We test the performance of our algorithm on several data sets. We compare our results with GDT & GDT with Adaptive boosting .The proposed enhancement in geometric decision tree is evaluated on a Real world data set & Synthetic data set from UCI repository.

4. RESULT AND CONCLUSION

In this paper, we have offered a new algorithm for acquiring oblique decision trees. The originality is in learning hyperplanes that captures the geometric structure of the class regions. At each node, we have established the two clustering hyperplanes and selected one of a angle bisectors as the split rule.

We have offered some analysis to gain the optimization problem for which the angle bisectors are the resolution. Based on information gain, weargued that our method of choosing the hyperplane at every node is sound. Through wide experiential studies, we mentioned that the method performs enhanced than the other decision tree approaches in terms of accuracy, space of the tree, and time.

We have also exposed that the classifier obtained with GDT is as good quality as that with SVM, whereas it is quicker than SVM. Thus, overall, the algorithm presented here is an excellent and narrative classification method. In future by considering new factors it will also work in very excellent manner.

Data Set	classes	Boosting trial	nodes	Training Data before	Training Data after
2X2 Checkerboard data set	2	9	52	82.875	82.875+0.625
4X4 checkerboard data set	2	12	62	73.625	73.625+2.875
Bank full balanced	2	7	62	62.86	62.86+0
Cancer	2	10	62	76.62	76.62+2.38

table : Improvement in data set percentage before & after

5. REFERENCE

[1] Naresh Manwani and P. S. Sastry, "Geometric Decision Tree" IEEE Trans vol. 42, no. 1, February 2012.

[2] L.Rokach and O. Maimon, "Top-down induction of decision trees classifiers—A survey," IEEETrans.Syst.,Man,Cybern.C,Appl.Rev., vol. 35, no. 4, pp. 476–487, Nov. 2005.

[3] O. L. Mangasarian, "Multisurface method of pattern separation," IEEETrans.Inf.Theory, vol. IT-14, no. 6, pp. 801– 807, Nov. 1968.

[4] T. Lee and J. A. Richards, "Piecewise linear classification using seniority logic committee methods, with application to remote sensing," PatternRecognit., vol. 17, no. 4, pp. 453–464, 1984.

[5] M. Lu, C. L. P. Chen, J. Huo, and X. Wang, "Multi-stage decision tree based on inter-class and inner -class margin of SVM," in Proc.IEEEInt.Conf.Syst.,Man,Cybern., San Antonio, TX, 2009, pp. 1875–1880.

[6] R. Tibshirani and T. Hastie, "Margin trees for high dimensional classification," J.Mach.Learn.Res., vol. 8, pp. 637– 652, Mar. 2007.

[7] M. F. Amasyali and O. Ersoy, "Cline: A new decision-tree family," IEEETrans.NeuralNetw., vol. 19, no. 2, pp. 356–363, Feb. 2008.

[8] A. Kołakowskaand W. Malina, "Fisher sequential classifiers," IEEETrans.Syst.,Man,Cybern.B,Cybern., vol. 35, no. 5, pp. 988–998, Oct. 2005.

[9] D. Dancey, Z. A. Bandar, and D. McLean, "Logistic model tree extraction from artificial neural networks," IEEETrans.Syst.,Man,Cybern.B,Cybern., vol. 37, no. 4, pp. 794–802, Aug. 2007.

[10] W.Pedrycz and Z. A. Sosnowski, "Genetically optimized fuzzy decision trees," IEEETrans.Syst.,Man,Cybern.B,Cybern., vol. 35, no. 3, pp. 633–641, Jun. 2005.

[11] B. Chandra and P. P. Varghese, "Fuzzy SLIQ decision tree algorithm,"

[12] IEEETrans.Syst.,Man,Cybern.B,Cybern., vol. 38, no. 5, pp. 1294– 1301, Oct. 2008. no. 1, pp. 1–32, Aug. 1994.

6. ACKNOWLEDGMENT

The proposed system is based on IEEE Transaction paper under the title “Geometric Decision Tree” published in “IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 42, NO. 1, FEBRUARY 2012”.