CARI’12

Decision Tree Network

Data mining approach

Faiz Maazouzi *, Halima Bahi**

LabGED Laboratory
Computer Science Department
Badji Mokhtar University
BP. 12, (23000) Annaba
Algeria

* mazouzi@labged.net

** bahi@labged.net

RÉSUMÉ. Le but de cette recherche est de proposer une approche pour l'extraction des données basée sur des arbres de décision. Dans cette approche nous construisons un modèle qui contient plusieurs couches, dans chaque couche nous avons plusieurs arbres de décision, nous proposons ensuite, une méthode pour utiliser notre modèle. L'objectif de notre proposition est d'améliorer les techniques actuelles de fouille de données. Pour juger de ces performances, nous comparons notre approche (DTN pour Decision Tree Network) avec un ensemble de méthodes de classification bien connues telles que : Boosting Decision Trees (BDT) et le bagging decision tree, nous comparons également le DTN avec l’algorithme des arbres de décision C4.5. Les résultats montrent des améliorations substantielles par rapport à des techniques similaires..

ABSTRACT. The aim of this research is to propose data mining approach based on decision trees. In this approach we build a model as a network that contains several layers, in each layer we have several decision trees, and then we propose a method for using our model. The objective of DTN is to improve the present existing data mining techniques. So, we compare the Decision Tree Network (DTN) performances with some well-known methods, namely Boosting Decision Trees (BDT) and bagging decision tree; we also compare DTN with C4.5 decision tree algorithm. Results show substantial improvements when compared to similar techniques.

MOTS-CLÉS : arbre de décision, fouille de données, algorithmes des arbres de décision, techniques de fouille de données.

KEYWORDS: decision tree, data mining, decision tree algorithm, data Mining techniques.

1. Introduction

Nowadays, many modern applications (web data, genomics, finance, e-marketing … etc.) require to manipulate and to process very large data. The discipline that develops and explores practical methods to model this type of data is called statistical learning (statistical machine learning). This is, ultimately, to produce predictive tools and decision support dedicated to a specific application.

Many data mining and machine learning algorithms have been proposed in the past years, such as: Bayesian classifier [9], rules -based classifier [3], neural networks [1], support vector machines [11], and decision trees [4]. Decision trees are widely used in data mining and several approaches of decision trees were developed in the last years [6] [7]. A similar methods to the one proposed in this paper is called Boosting Decision Trees (BDT) [2] and Bagging Decision Trees [12]. The common idea in both methods is that the model contains several trees, each tree is constructed on the basis of a different training set. In our study, we propose a new approach of data mining based on decision trees "decision tree network" (see Fig. 1). The purpose of the use of our method is to improve the learning process.

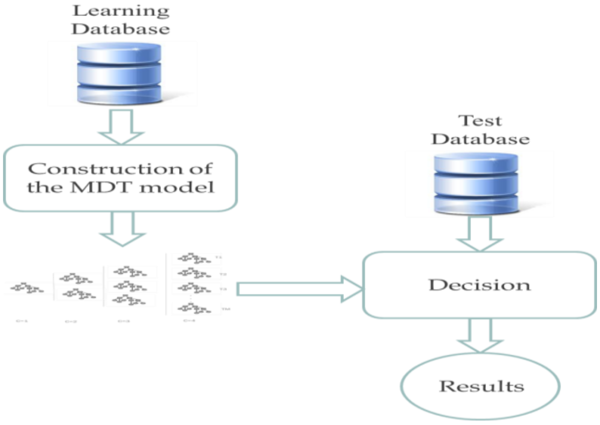

Figure 1. *General system overview*

For decision trees classifier. A small change in the training data can produce a large change in the tree. So, for each tree, any class may depend from different attribute and for each sub base (SB) in our system we get a different tree, and the results of classification for each tree are different.

This paper is organized as follows: next section describes the related works. Section 3 will present DTN model construction. In Section 4, we shall present the decision-making process. Evaluation and experiments are presented in Section 5. Finally, a conclusion is done.

2. Related work

In the last few years, decision trees has been used widely in the data mining domain, in text mining [16], web mining [17], image mining [20]. If the decision trees have met very interesting in the 90s with a very large number of publications designed to enhance their performance, it is clear that no breakthrough has been produced in recognition rate compared to the reference algorithms that are ID3, CHAID, CART and C4.5 [15].The good thing is that with many studies we can make more control over the properties of trees. It is possible to characterize the variations and the context in which they work best.

Another group of researchers proposed to build a new classifier which contains a set of decision trees. For example the work of [13] on "bagging" Freund and Schapiro [18] on the 'boosting' and their use in the trees [19] have shown that it is possible to significantly improve the performance of the classification model.

3. Construction of DTN model

The aim of machine learning methods is to construct a set of predictive models and combine their outputs into a single prediction. In the last few years, several methods were developed, such as, bagging [13], random forests [14] and boosting [15]. The two most famous algorithms are CART [8] and C5 (the most recent version of ID3 and C4.5 [8]). These algorithms are used for their performance.

Our model represents a set of decision trees combined in the form of a network, where decision tree represent a graphical representation of a classification procedure.

The model is built in three stages:

1) The first step is to choose the number of layers n.

2) The second step, we built the set $LiSBj$ for each layer as:

$$LiSBj = \frac{TB}{\mathrm{L_N}} \tag{1}$$

Where:

$i,j = 1 \dots n$

TB: The Training set size.

$\mathrm{L_N}$: Layer number

3) The last step is to apply the decision tree algorithm (C4.5) [8] for each data set $LiSBj$.

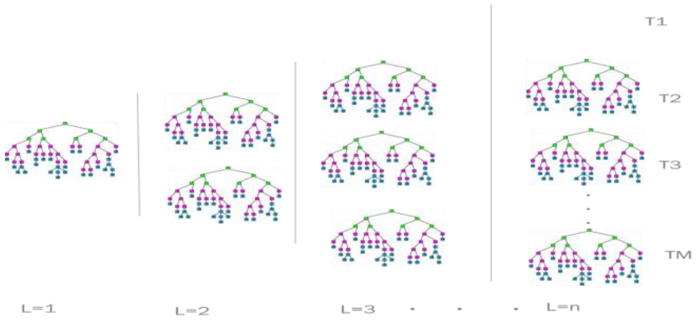

Figure 2.*Structure of Decision Tree Network (DTN) model*

Base	Base size	Training set size	Number of layer	SB size for each tree	Number of trees	Number of classes
Iris	150	120	2	L1SB1 :120 L2SB1 :60 L2SB2 :60	3	3
Pima	306	245	4	L1SB1 :245 L2SB1 :122 L2SB2 :123 L3SB1 :81 L3SB2 :81 L3SB3 :82 L4SB1 :61 L4SB2 :61 L4SB3 :61 L4SB4 :62	10	2
Glass	214	165	3	L1SB1 :165 L2SB1 :82 L2SB2 :82 L3SB1 :55 L3SB2 :55 L3SB3 :55	6	7
wine	178	142	2	L1SB1 :142 L2SB1 :71 L2SB2 :71	3	3
Iono-sphere	351	280	4	L1SB1 :280 L2SB1 :140 L2SB2 :140 L3SB1 :90 L3SB2 :90 L3SB3 :91 L4SB1 :70 L4SB2 :70 L4SB3 :70 L4SB4 :70	10	2

Table 1.*Information on models built from bases “iris, Pima, glass, wine, Ionosphere”.*

In the construction stage, the number of layers depends always from the size of database, if the size of the database is large the number of layers is large and vice versa.

As an illustration, Information about models built from the Datasets: iris, Pima, glass, wine and Ionosphere [6] are reported in Table 1.

The Table 1 content the different information extracted from different database and this information is very important for the construction of our model (such as: the number of layer, $LiSBj$ for each layer and the number of tree in DTN). The number of trees in the DTN model always depends from the number of layers.

4. The decision-making process with DTN method

In this section we present the decision-making process used in our technique of data mining. Let us consider a problem *Pr* with *M* classes such as C_M represents the class *M*.

The purpose of decision-making process is to calculate the percentage that a P_C (Probability of class) test data belongs to the class C_M.

The first step is to calculate x

$$x = \sum_{k=1}^{n} \left(\frac{1}{k}\right)^{-1} \qquad (2)$$

n is the number of layers in the model.

The second stage of decision-making process is to calculate the weight (P_i) for each layer.

$$Pi = \frac{1}{i} * x \qquad (3)$$

In the next step we calculated the probability (PL_i) of the C_M class for each layer i (i = 1 ... n)

$$PL_i = \frac{\sum_{j=1}^{i} PT_{i,j}}{i} \qquad (4)$$

Where:

$$PT_{i,j} = \begin{cases} 1 & \text{if the classification result } = C_M \\ 0 & \text{if the classification result } \neq C_M \end{cases}$$

The last step is to calculate P_C (Probability of class) such as:

$$P_C = \frac{\sum_{i=1}^{n} PC_i * P_i}{n} * 100 \qquad (5)$$

The following figure shows the variable used in the process of decision.

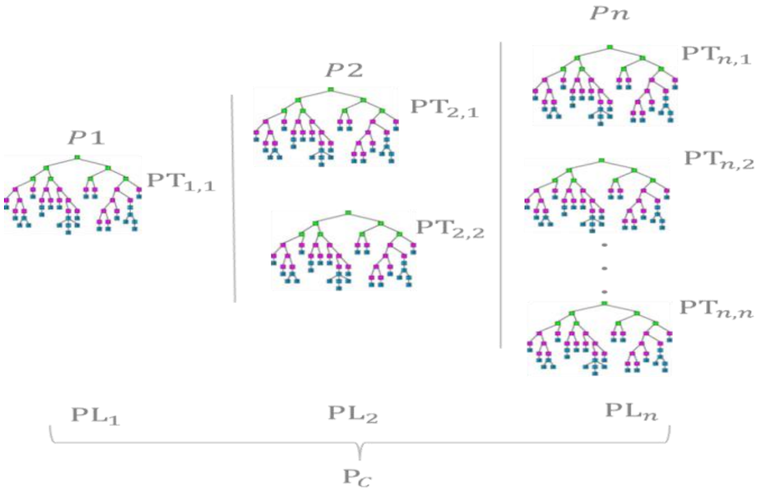

Figure 3. *Variable (P_n, $PT_{n,n}$, PL_n, P_C) in DTN*

Finally, we got a percentage for each class.

If we had a problem of class M, the result of classification is the class that has the highest percentage.

Result of classification = max ($P_{C1,}P_{C2,}\ldots P_{CM}$).

5. Experimental Results

The system described above was applied to some data sets taken from the Machine Learning Repository [6] in order to compare the capabilities of the system with other known decision tree generators.

In this experiment we present a comparative study of Decision Tree Network classification technique and C4.5 algorithm of decision tree and AdaBoost decision tree algorithm. We tested our system with databases: iris, Pima, glass, wine and Ionosphere. Each data set was split randomly into two sets, the training set which comprised 80% of the data and the test set, which comprised 20% of the data. The test database contains attributes with missing values.

We compared the results of Decision Tree Network (DTN) system with the results of algorithms C4.5 (implemented in WEKA as J48 [11]) and the result of AdaBoost decision tree algorithm and Bagging decision tree, to do this we compute the recall and precision of both techniques. The results are shown in the Table 2.

Bases	Techniques							
	C4.5 Algorithm		AdaBoost decision tree		Bagging decision tree		Decision Tree Network (MTN)	
	recall	Precision	recall	Precision	recall	Precision	recall	Precision
Iris	76.8 %	81.2 %	88.6 %	90.4 %	89.9 %	90.1 %	92.2 %	93.1 %
Pima	78.3 %	79.5 %	88.3 %	89.6 %	89.3 %	90.9 %	89.7 %	95.8 %
Glass	81.8 %	81.7 %	81.1 %	80.1 %	83 %	83 %	84.7 %	85.1 %
Wine	82.2 %	79,7 %	77.7 %	80.1 %	81.8 %	80.8 %	82.1 %	80.8 %
Ionosphere	90.5 %	89.8 %	90.1 %	88.9 %	90.5 %	90.7 %	90.8 %	90.9 %

Table 2.*Shows a comparison of the results of MDT and C4.5 AND AdaBoost for "Iris, Pima, Glass, Wine, Ionosphere "data sets*

In this section we have compared the performance of our system with different systems (AdaBoost decision tree, Bagging decision tree, C4.5 Algorithm). According to the experiments and result analysis presented in this paper, good results are obtained from oursystem (DTN).

6. Conclusion

Decision trees are often used in the field of data mining in this paper we present a new data mining technique: Decision Trees Network.

In the DTN system, the classification results of each instance are as a percentage for each class. We tested our approach with other approaches based on decision trees (C.45 and Boosting decision tree), AdaBoost decision tree algorithm and Bagging decision tree; the results show that it is possible to improve the system of classification when using DTN.

7. References

[1] Freund, Y., Schapire, R.: Experiments with a NewBoosting Algorithm. In Proceedings of the International Machine Learning Conference, pp. 148–156.Morgan Kaufmann, 1996.

[2] Bigus, J. P.: Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. McGraw-Hill, Inc., Hightstown, NJ, USA, 1996.

[3] Hu, Y. C., Chen, R., Tzeng, G. H.: Finding fuzzy classification rules using data mining techniques. Pattern Recogn.Lett. Vol. 24, pp. 509–519, January, 2003.

[4] Wasniowski, R. A.: Using support vector machines in data mining. In Proceedings of the 4th WSEAS International Conference on Systems Theory and Scientific Computation, 2004.

[5] Rokach, L., Maimon, O.: Data Mining with Decision Trees, World Scientific Publishing Co., Singapore 2008.

[6] Machine learning repository, http://archive.ics.uci.edu/ml/datasets.html, last visit June 2011.

[7] Potgieter, G., Engelbrecht, A. P. Evolving model trees for mining data sets with continuous-valued classes. Expert Syst. Vol. 35, pp. 1513–1532, 2008.

[8] Smith, J. F.: Evolving fuzzy decision tree structure that adapts in real-time. In Proceedings of the conference on Genetic and evolutionary computation (GECCO '05), Hans-Georg Beyer (Ed.).ACM, pp.1737–1744, New York, NY, USA, 2005.

[9] Quinlan, J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Vol. 16, pp. 236–240, 1993.

[10] Makki, S., Mustapha, A., Kassim, J. M., Gharayebeh, E. H., Alhazmi, M.: Employing c NeuralNetwork and Naive Bayesian Classifier in Mining Data for Car Evaluation. ICGST AIML-11 Conference,pp.113–119, Dubai, UAE, April, 2011.

[11] Weka 3- Data Mining with open source machine learning software available from: - http://www.cs.waikato. ac.nz/ml/ weka/.

[12] Diepena, M. V., Franses, P. H.: Evaluating chi-squared automatic interaction detection , Information Systems, Vol. 31, pp. 814–831, 2006.

[13] Loh, W. Y.: Classification and regression tree methods. In Encyclopedia of Statistics in Quality and Reliability, pp. 315–323, 2008.

[14]Loh, W. Y., Shih, Y. S.: Split selection methods for classification trees, StatisticaSinica, vol. 7, 815–840, 1997.

[15] Nicolas, V., Marc, B., Carine H.A Bayes Evaluation Criterion for Decision Trees. EGC (best of volume), pp. 21-38, 2009.

[16]Apte, C., Damerau, F., Weiss, S. Text mining with decision rules and decision trees. In Workshop on Learning from text and the Web, Conference on Automated Learning and Discovery, 1998.

[17] Pabarskaite, Z. Decision trees for web log mining. In Journal of Intelligent Data Analysis , pp. 141 - 154, 2003.

[18]Freund Y., Schapire R., A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, 55, 1, 119-139, 1997.

[19] Quinlan R., Bagging, Boosting and C4.5, in Proceedings of the Thirteenth National Conference on Artificial Intelligence, pp. 725-730, 1996.

[20] Rajendran, P., Madheswaran, M. Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm, JOURNAL OF COMPUTING, Vol. 2,pp. 127-136, 2010.