Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 10 (13): 1336-1340, 2010
ISSN 1812-5654

The Decision Tree and Support Vector Machine for the Data Mining

D. Benhaddouche and A. Benyettou
Département D'informatique, Laboratoire Simpa, Faculté Des Sciences,
Université Des Sciences et de la Technologie d'Oran-USTO, BP. 1505, Oran El-Mnaouer 31000, Algérie

Abstract: In this study one used the dated mining to extract from knowledge biomedical, one used in the training course of training two algorithms Decision trees and SVMs. These methods of supervised classification are used to make diagnoses it for the disease hypothyroid and which gives custom the model of the extraction of data. We have studied the advantages of both methods, also by concrete results; we conclude that the method of SVM (Support Vector Machine) is better in our case.

Key words: Medical data, learning, classifiers, algorithms decision trees, support vector machine, extracting knowledge

INTRODUCTION

The quality of the data recorded in the great biomedical data bases is not guaranteed by the strict procedures of dated management, as it is the case for the clinical trials. It thus appears necessary to set up specific methods of pretreatment of the data before carrying out analyses that it is by traditional statistical methods or recent methods of excavation of data. In general, the biomedical complex data collected at the time of the studies of populations are treated by statistical methods, which constitute the reference for the majority of the biologists, epidemiologists or doctors confronted with the analysis of the results. However, the technological projection in medicine implies a volume of data to be treated increasingly large. The statistical methods are robust but they are not alone enough to exploit all the potential richness of the data. The principal problems are to extract from the units of knowledge starting from these data, which are new and potentially useful. In this context, we proposed to promote the study and the use of the techniques symbolic systems in excavation of data. We used the advantages of a method to avoid the drawbacks of the other method.

For the method SVM (Support Vector Machine) time is important but the error is smaller than the method of decision trees (Ayat, 2004; Marée, 2005)

DATA MINING

The term of extraction of the data, more known under the name data mining is often employed to indicate the whole of the tools making it possible the user to reach the data of the company and to analyze them. One could define the excavation of the data like a step having the aim of discovering relations and facts, at the same time new and significant, on great sets of data. The data are without value if they are not interpreted. By interpreting data, one obtains information and it is necessary that information is received, included/understood and classified to obtain knowledge from them (Boulicaut, 2002) There are five great stages which it is necessary to traverse 2.

Processes in a project of excavation of the data (Jouini, 2003):

- To pose the problem
- Seek and selection of data
- The préparation data
  - Reduction
  - Cleaning
  - Transformation
- Development of the model (modeling)
- Application of the model

Tasks: Contrary to the generally accepted ideas, the excavation of the data is not the miracle cure able to solve all the difficulties or needs for the company. However, a multitude of problems of a intellectual, economic or commercial nature can be gathered, in its formalization, one of the following tasks: Classification, estimate, prediction, grouping by similarities, segmentation (or clusterisation), description, optimization (Bouchard, 2005).

Corresponding Author: Djamila Benhaddouche, Département d'Informatique, Laboratoire Simpa, Faculté Des Sciences, Université des Sciences et de la Technologie d'Oran-USTO, BP. 1505, Oran El-Mnaouer 31000, Algérie

Methods: We adopt only certain methods which come to supplement the traditional tools which are requests SQL, the requests of crossed analysis, the tools for visualization, the descriptive statistics and the analysis of the data. The methods are the algorithm for the segmentation, the rules of association, the closest neighbors (reasoning starting from case), the decision trees, the networks of neurons, the genetic algorithms, the networks bayésiens, Support Vector Machine (SVM), the methods of regression, the analysis discriminating.

SUPPORT VECTOR MACHINES (SVM)

Support Vector Machines (SVM) are new discriminating techniques in the theory of the statistical training. For the data processing specialists, the SVM is a linear classifior with broad margin in a space with core. For the statisticians, the SVM is a nonparametric estimator. It is based on a minimization of the empirical risk regularized on a functional space of Hilbert and with a linear function of loss per pieces.

Mathematical principle and general: SVM are algorithms based on the three following mathematical principles (Ayat, 2004).

- Principle of Fermat (1638)
- Principle of Lagrange (1788)
- Principle of KuhnTucker (1951)

Principle general: The principle general is the construction of a classifior with actual values and the division of the problem in two pennies problems:

- Nonlinear transformation of the entries
- Choice of a linear separation optimal (Kharoubi, 2002)

Concepts of bases

Problem of training: One is interested in a phenomenon F (possibly not determinist) which starting from a certain set of entries X product an exit y = f(x) generally, only the case (Y = {-1, 1} interests us in SVMs but one can easily extend to the case |y| = m > 2).

The goal is to find this function F starting from the only observation of a sample:

$$S = \{(X_1, y_1,), ... (x_n, y_n)\}$$

Of n independent copies of (X, Y).

Optimal hyperplane: One calls optimal hyperplane the separating hyperplane which is located at the maximum distance from the vectors X closest among the unit to the examples; one can also say that this hyperplane maximizes the margin.

Supports Vecteurs (SV): The Vectors Supports (term which one could translate by points of support) are the vectors for which equality:

$$y_i ((w^0 x_i)+ b^0) = 1 \quad (1)$$

is checked, concretely, they are the points closest to the optimal hyperplane.

The margin: The margin represents the smallest distance between the various data of the two classes and the hyperplane.

Construction of the optimal hyperplane: For describing well the technique of construction of the optimal hyperplane separating from the data belonging to two different classes in two different cases: The case of the linearly separable data and the case of the data not linearly separable. We consider the following formalism Is the unit D such as:

$$D = \{(x_i, y_i) \in R^n \times \{-1; 1\} \text{ for } i = 1, .., m\} \quad (2)$$

Principle of the SVM: Classifieurs SVM use the idea of HO (Optimal Hyperplane) to calculate a border between groups of dots. They project the data in space of characteristics by using nonlinear functions. In this space, one builds the HO which separates the transformed data. The principal idea is to build a linear surface of separation in the space of the characteristics which corresponds to a nonlinear surface in the space of entry. For any function g 0 with:

$$\int g^2(z)dx \geq 0 \quad (3)$$

One calls these functions the cores of Hilbert-Schmidt. Several cores were used by the researchers; here are some (Table 1), (Ayat, 2004):

Table 1: Cores of hilbert-schmidt

Noyau	Formula
Linéaire	$k(X, Y) = x \cdot y$
Sigmoïde	$K(x, y) = \tanh(ax \cdot y+b)$
Polynomial	$K(x, y) = (zx \cdot y+b)^d$
RBF	$K(x, y) = \exp(-\|x-y\|^2/\sigma^2)$
Laplace	$K(x, y) = \exp(-\|\gamma\|x-y\|)$

DECISION TREES

The decision trees are most popular of the methods of training, the popularity of the method rests mainly on its simplicity. A decision tree is the chart of a procedure of classification. Indeed, with any complete description only one sheet with the decision tree is associated. This association is defined while starting with the root of the tree and while going down according to answers' to the tests which label the internal nodes (Marée, 2005).

The associated class is then the class by defect associated with the sheet which corresponds to description. The procedure of classification obtained has an immediate translation in term:

Principe fonda mental: One gives oneself a unit X of N examples noted x_i whose P attributes are quantitative or qualitative. Each example X is labelled, i.e., it is associated for him a class or a target attribute which one notes $y \in Y$.

From these examples, one builds a tree such as (Bougrain, 2004):

- **Node:** Each nonfinal node corresponds to a test (IF….THEN…) on the value of one or more attributes
- **Arc:** Each branch on the basis of a node corresponds to one or more values of this test
- **Break into leaf:** With each final node called sheet a value with the target attribute is associated (class)

Construction of a decision tree: The best method is that which consists in testing all the possible trees, but this solution is not possible.

Example: If one has NR attributes which can take on average V values, the number of trees studied:

$$\sum_{i=1}^{N}(n-i+1)^{v^{i-1}} \Rightarrow$$

Get the number of possible trees

Thus 4 attributes with 3 values gives 526 possible trees.

One thus seeks to build the tree by a downward induction (top-down induction of décision tree) (Pasquier, 2000).

Problems: This apparent simplicity should not mask real difficulties which arise during the construction of the tree.

- Choice of the discriminating attribute (choice of the attribute of segmentation)
- Stop of the segmentation (choice depth of the tree)
- There are two various techniques:
  - Pré-pruning
  - Post-pruning
- Décision

Algorithms

- **CART:** Algorithm CART (Breiman, 1984)
- **CHAID:** Algorithm CHAID (Hartigan, 1975)
- **ID3:** Algorithm ID3 (Quinlan, 1986)
- **C4.5:** Algorithm C4.5 was worked out by Quinlan (1993), this algorithm is in fact only one improvement of ID3

CONCEPTION AND REALISATION

Description of the base of the data: We will use in our application the base of the data «hypo py published on the site http://axon.cs.byu.edu/~martinez/classes/470/MLDB/thyroid-disease/hypothyroid.data.

It contains 3163 biomedical recordings connected of the patients hypothyroïdiens with 25 attributes and a result of diagnosis. By using this base of the data, we can show some models related on the age of the patient, the kind, the questions, the pregnancies, the thyroid treatment, the surgery and the drug, as well as their clinical test results, such as the disease, the tumour, lithium, the goitre and measurements of TSH, T3, TT4, T4U, FTI, levels of TBG.

We asked for the opinion of an expert who proposed to us to eliminate some attributes from the original base of the data which do not have impact on the result of the diagnosis and another examination which are not available in our laboratories. Thus one obtained another data base more adapted to the problems.

Structure of the data used
Entries

- **Age:** Concerning the age of the patient, oldest are touched by the hypothyroïdie
- **Sex:** This disease more frequently assigns the women
- **Enclosure:** If the woman is pregnant or not
- **Patient:** If the patient is already sick or not
- **Test TSH:** If the test were carried out or not
- **TSH (THYROGLOBULINE):** Anti-hypophyseal hormone simulating the thyroid one
- **Test T3:** If the test were carried out or not
- **T3 (TRIIODOTHYRONINE):** Iodized hormone secreted by the thyroid one, also coming from the peripheral desiodation of T4

- **Tt4 test:** If the test were carried out or not
- **Tt4 (THIROXINE):** Iodized hormone secreted by the thyroid one, the free form (T4L or free-T4 or FT4), which presents 0.02 A 0.04% of T4 total and only activates it
- **Test TBG:** If the test were carried out or not
- **TBG (THYROXIN BINDING GLOBULIN):** Protein of transport of the thyroid hormone

RÉSULTS

- **Class:** There are two classes:
- **Hypothyroïdie:** The case where the diagnosis is confirmed positive
- **Négative:** As the word indicates it, they is the nonsick people

Scheme of work: To be able to treat our data base, one was obliged to pass by the following stages which enter the processes of excavation of the data:

The comprehension of the problem: For this spot, one tried to control well our problems concerning the disease Hypothyroïdie using the experts in the field.

Selection of the data: Our original data base contains 25 attributes where according to the expert the majority of them are unutilised for our diagnosis; thus one filtered the original data base by selecting only the useful fields.

Cleaning: Cleaning consists with

- To eliminate the doubled blooms appeared in our data base
- Filling of the missing values
- The replacement of the disturbed data

The algorithm general of SVM

Entry a labelled unit X.

Algorithm

That w, b = 0;
That is to say unit X for the whole of training.
That is to say together of the labels of Y.
Beginning
For i = 1 j until L do:
    For j = 1 j until L do:

To learn:

$$\text{Max}\left[\sum_{i=1}^{L}\alpha i - \frac{1}{2}\sum_{i,j=1}^{L} y_i y_j \alpha_i \alpha_j \; K(x_i, x_j)\right]$$

With the constraints

Table 2: Results of comparison between SVM and decision tree

Parameters comparison	SVM	Decision tree
The time of training (H)	16	8
The error rate	0.054	0.061

$0 \le \alpha_i \le C$, I = 1, ..., L

$$\sum_{i=1}^{L} \alpha_i y_i = 0$$

$$H = \sum_{i,j=1}^{L} y_i y_j \; K(x_i, x_j)$$

To solve the problem of optimization with QP such as the function to be minimized is:

$$\text{Min } -1/2\alpha'^{*} H^{*} \alpha + c^{t*}\alpha$$

C: control level of error in classification

After the resolution of this problem, one obtains which is used for the calculation of the function of decisions as follows:

$$f(x) = \text{sign}\left(\left(\sum_{i=1}^{L} \alpha_i y_i \; K\left(x_i, x_j\right)\right) + b\right) \quad (4)$$

To evaluate the number of support vector.
To evaluate the value of B.
End

The algorithm general of C4.5

Entry: language of description: sample S;
Beginning
To Initialize with the empty tree; the root is the current node;
    To repeat
        To calculate the entropies for each value of each attribute
    To calculate the profit for each attribute
        The choose the maximum profit
        The choose the test for the current node;
        To decide if the node courant is final;
            If the node is final then to affect a class;
            If not to select a test and create the under tree;
            End if
        To passe to the following node not explored if there are;
    Until obtaining a decision tree;
End.

We have compiled a table (Table 2) comparison between the two methods and found that in the case of SVM, the learning time is greater than that of decision trees, cons by the error rate of SVM is inferior to that of decision tree.

CONCLUSION

One arrived at the following conclusion: The two methods gave satisfactory results but the method Support Vector Machines gave one better classification

compared to the method of decision tree; that is due to its simplicity and its mathématical rigour, also SVM allows better a generalisation than the decision tree althougth the time of training was too long contrary to the method of the decision tree

REFERENCES

Ayat, N.E., 2004. Automatic selection of model of the machines with vectors of support application to the recognition of images and handwritten figures. Ph.D. Thesis, University of Montréal Canada.

Bouchard, G., 2005. Generative models in supervised classification and applications to the categorization of images and industrial reliability. Ph.D. Thesis, University of Grenoble France.

Boulicaut, J.F., 2002. State of the Art on the Extraction of Frequent Patterns. National Institute for Applied Sciences, France, Tunis.

Bougrain, L., 2004. Decision Trees: Courses in Data Mining. University of Nancy, France.

Breiman, L., 1984. Algorithm CART. Classification and Regression Trees. California Wadsworth International Group, Belmont, California.

Hartigan, J.A., 1975. Algorithm CHAID. Clustering Algorithms. John Wiley and Sons, New York.

Jouini, W., 2003. The Methods and Techniques for Knowledge Discovery in Databases. Ecole Centrale, Paris, France.

Kharoubi, J., 2002. Methods and techniques of retrieval of knowledge of data bases. Ph.D. Thesis, TELECOM Paris France.

Marée, R., 2005. Automatic classification of images by decision trees. Ph.D. Thesis, University of Liege Belgique.

Pasquier, N., 2000. Data mining: Algorithms of extraction and reduction of the rules of association in the data bases. Ph.D. Thesis, University of Clermont-Ferrand II France.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learn., 1: 81-106.

Quinlan, R., 1993. C4.5.Programs for Machine Learning. Morgan Kaufman Publisher Inc., San Francisco, USA.