World Academy of Science, Engineering and Technology
International Journal of Industrial and Manufacturing Engineering
Vol:8, No:10, 2014

Data Mining in Medicine Domain Using Decision Trees and Vector Support Machine

Djamila Benhaddouche, Abdelkader Benyettou

Open Science Index, Industrial and Manufacturing Engineering Vol:8, No:10, 2014 publications.waset.org/10000925.pdf

***Abstract*—**In this paper, we used data mining to extract biomedical knowledge. In general, complex biomedical data collected in studies of populations are treated by statistical methods, although they are robust, they are not sufficient in themselves to harness the potential wealth of data. For that you used in step two learning algorithms: the Decision Trees and Support Vector Machine (SVM). These supervised classification methods are used to make the diagnosis of thyroid disease. In this context, we propose to promote the study and use of symbolic data mining techniques.

***Keywords*—**A classifier, Algorithms decision tree, knowledge extraction, Support Vector Machine.

I. 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 pre-treatment 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 propose to promote the study and the use of the techniques symbolic systems in excavation of data.

II. Data Mining

The term of excavation 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. [4]

Djamila Benhaddouche is with the SIMPA Laboratory, University of Science and Technology, Mohamed Boudiaf UST Oran, Algeria (Phone: +213- 772-934019; e-mail: djamila.benhaddouche@univ-usto.dz).

Abdelkader Benyettou is Professor at the University of Science and Technology, Mohamed Boudiaf UST Oran, Algeria, SIMPA Laboratory (e-mail: a_Benyettou@yahoo.fr).

A. Processes

There are five great stages which it is necessary to traverse in a project of excavation of the data [9]:

- ✓ To pose the problem
- ✓ Seek and selection of data
- ✓ Data preparation
- ❖ Reduction
- ❖ cleaning
- ❖ Transformation
- ✓ Development of the model (modeling)
- ✓ Application of the model.

B. 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 an 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. [2].

C. 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 neighbor's (reasoning starting from case), the decision trees, the networks of neurons, the genetic algorithms, the networks Bayesians, support vector machine (SVM), the methods of regression, the analysis discriminating..

III. 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 classifier 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.

A. Mathematical & General Principle

SVM are algorithms based on the three following mathematical principles [1]:

- ➢ principle of Fermat (1638)

Open Science Index, Industrial and Manufacturing Engineering Vol:8, No:10, 2014 publications.waset.org/10000925.pdf

- principle of Lagrange (1788)
- principle of KuhnTucker (1951)

General principle:

The construction of a classifier with actual values and the division of the problem in two pennies problems:

- Nonlinear transformation of the entries;
- Choice of a linear separation "optimal**".** [8]

B. Concepts of Bases

1) 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).

2) Optimal Hyper Plane

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

3) Supports Vectors (SV)

The Vectors Supports (term which one could translate by points of support) are the vectors x_i for which equality: $y_i((w^0 x_i) + b^0) = 1$ is checked, concretely, they are the points closest to the optimal hyper plane.

4) The Margin

The margin represents the smallest distance between the various data of the two classes and the hyper plane.

C. Construction of the Optimal Hyper Plane

To describe the technique of building the optimal hyper plane separating data belonging to two different classes in two different cases: The case of linearly separable data and the case of not linearly separable data. We consider the following formalism is driving D, such as:

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

D. Principle of the SVM

Classifiers SVM use the idea of HO (Optimal Hyper plane) 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 $g \neq 0$ with:

$$\int g^2(z)dz \geq 0$$

One calls these functions the cores of Hilbert-Schmidt. Several cores were used by the researchers; here are some (Table I) [1]

IV. 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 [7]. 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

A. Fundamental Principle

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: [3]

- **Node :** each non final 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).

B. 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:

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

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 decision tree). [10]

C. 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:

- pre-pruning
- post-pruning
- Decision

Open Science Index, Industrial and Manufacturing Engineering Vol:8, No:10, 2014 publications.waset.org/10000925.pdf

D. Algorithms

1) Algorithm CART is published by [5].
2) Algorithm CHAID is published by [6].
3) Algorithm ID3 is published by [11].
4) Algorithm C4.5 was worked out [12], this algorithm is in fact only one improvement of ID3.

V. Conception and Realisation

A. Description of the Base of The data

We will use in our application the base of the data we will use in our application database which was published [13].

TABLE I
CORES OF HILBERT-SCHMIDT

Core	Formula
linear	$K(x,y) = x \cdot y$
sigmoid	$K(x,y) = \tanh(ax \cdot y + b)$
polynomial	$K(x,y) = (ax \cdot y + b)^d$
RBF	$K(x,y) = \exp(-\|x-y\|^2/\sigma^2)$
Laplace	$K(x,y) = \exp(-\gamma\|x-y\|)$

The database contains 3163 biomedical recordings connected of the patients' hypothyroid 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 goiter, 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

B. Structure of the Data Used

1) Entries

Age: concerning the age of the patient, oldest are touched by the hypothyroid.

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

2) Results

Class: there are two classes:

- Hypothyroid: the case where the diagnosis is confirmed positive.
- Negative: as the word indicates it: non sick people.

C. 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:

1) The Comprehension of the Problem

For this spot, one tried to control well our problems concerning the disease "Hypothyroid" using the experts in the field.

2) Selection of the Data

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

3) 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.

4) The Algorithm General of SVM

Entry labelled units 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=1 j until L do:
To learn:

$$\max\left[\sum_{i=1}^{L}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{L} y_i y_j \alpha_i \alpha_j K(x_i, x_j)\right]$$ With the constraints

$$0 \le \alpha_i \le C, \quad i = 1,...,L$$

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

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

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

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:

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

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

5) 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 pass to the following node not explored if there are;
Until obtaining a decision tree;
End.

6) Results

TABLE II
COMPARISON BETWEEN DECISION TREE AND SVM ON LEARNING TIME AND ERROR RATE

	SVM	decision trees
The time of training	16h	8h
The error rate	0.054	0.061

VI. CONCLUSION

Both methods gave satis factory results of both: that of Support Vector Machines (SVM) gave a better classification compared to decision trees, this is due to its simplicity and mathematical rigor, so the SVM can better generalization and time despite significant learning as opposed to the method of decision.

REFERENCES

[1] N. E. Ayat ," Automatic selection of model of the machines with vectors of suppo rt application to the recognition of i mages and handwritten figure", These, 2004.
[2] G. Bouchard," Generative models in supervised classification and applications to the categorization of images and industrial reliability", these, 2005.
[3] L. Bougrain, "Decision Trees".
[4] J-F Boulicaut," State of the art on th e extraction of frequent reasons", Paris; 2002.
[5] L.Briemen and associates, "Algorithm CART", 1984.
[6] Hartigan, "Algorithm CHAID", 1975.
[7] W. Jouini," Methods and techniques of Retrieval of Knowledge of Data bases", 2003.
[8] J. Kharroubi "Methods and techniques of Retrieval of Knowledge of Data bases".
[9] R. Marée, "Automatic classification of images by decision trees", these; 2005.
[10] N. Pasquier, "Data Mining: Algorithms of extraction and reduc tion of the rules of association in the data base, these, 2000.
[11] R. Quinlan, "Algorithm ID3", 1986.
[12] Quinlan, "Algorithm C4.5", 1993.
[13] www.axon.cs.byu.edu/~martinez/classes/470/MLDB/thyroid-disease/hypothyroid.data.

Open Science Index, Industrial and Manufacturing Engineering Vol:8, No:10, 2014 publications.waset.org/10000925.pdf