Knowledge Discovery Applied to Medical Domains

Jesús González[1], Beatriz Flores[1], and Pedro Sánchez[2]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro #1, Sta. Maria
Tonantzintla, Puebla, México
{jagonzalez, baflores}@ccc.inaoep.mx
[2] Instituto Tecnológico de Apizaco, Av. Tecnológico S/N, Apizaco, Tlaxcala, México
bsanchez@ssa.gob.mx

Abstract. The high production rate and the variety of medical data makes necessary the use of new tools to analyze it and get the best from it. Knowledge discovery in databases (also known as data mining) is a research area that comes from the combination of the machine learning, statistics, and pattern recognition areas. In this paper, we apply data mining techniques to two medical domains. In the first domain, we use decision trees and neural networks to detect calcifications in mammograms and in the second domain we use decision trees to analyze a tuberculosis database. Our results show that data mining techniques can be efficiently used to detect calcifications in mammograms with a predictive accuracy of up to 94 % in the ISSSTEP mammograms database and can also be used to find descriptive patterns that help us to understand the increase in cases of people with tuberculosis in Tlaxcala, Mexico.

1 Introduction

The technological advance in the medical area has contributed not only to the automation of complex processes but also to the production of large amounts of data in hospitals. There are a rich variety of medical databases that go from diagnosis data to radiological images. The large amount of stored data needs to be analyzed to find patterns that cannot be found with standard tools (like statistics or spreadsheets). Here is where knowledge discovery techniques can be applied to medical datasets to find hidden patterns from them. In the rest of this section we describe the knowledge discovery in databases process. In section 2 we present our research in the calcifications detection in mammograms domain and in section 3 we show our work in the tuberculosis domain. Finally, in section 4 we present our conclusions and future work.

1.1 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is a field of computer science concerned with finding patterns or interesting knowledge in large databases where it is not possible to manually identify these patterns due to the large amount of data. KDD has been technically defined as "the non-trivial process of identifying valid, novel, poten-

I. Rudomín, J. Vázquez-Salceda, J. Díaz de León (Eds.). e-Health: Application of Computing Science in medicine and Health Care. © IPN, México 2003.

tially useful, and ultimately understandable patterns in data" [1]. From this definition, we see that KDD is a process that involves several activities: data preparation, search for patterns over the prepared data, evaluation of those patterns, and refinement of the whole process (see figure 1). The definition also states that patterns should be valid, novel, potentially useful and ultimately understandable. A pattern is valid if it applies when new data is added to the database. Novel patterns are those that show facts that we did not know were in the database. The patterns are useful when the corresponding knowledge can be used to improve something in the field that the data was taken from. The patterns must also be understandable so that the user is able to identify and use the new knowledge.

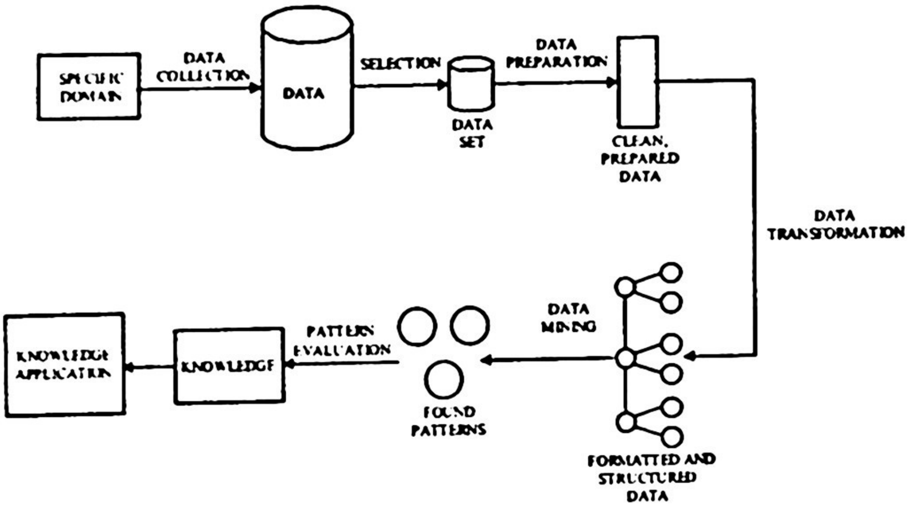

Figure 1. The KDD Process.

The KDD process also includes a pattern evaluation phase so that we can identify when the pattern is considered new knowledge or if it is just irrelevant information. We can do that by grading the patterns according to their characteristics like how useful and novel is the pattern and assigning a threshold over that grade. Only if the pattern grade is higher than the threshold is it considered new knowledge. The KDD process steps can be enumerated as:

- Identify and understand the application domain.
- Choose and create the set of information to be used in the process.
- Prepare the data for the process.
- Choose the data mining task.
- Choose the data mining algorithm.
- Execute the data mining algorithm.
- Evaluate the found patterns.
- Apply the discovered knowledge.

Figure 1 shows how the KDD process is related to the information, concepts and tools we just mentioned. In the following sections we will see how we applied the

KDD process to the calcifications detection in mammograms and the tuberculosis domains.

In the following sections we briefly describe the data mining algorithms used for this research. We start with Neural Networks in section 1.2, then we describe decision trees in section 1.3 and finally, we introduce Subdue in section 1.4.

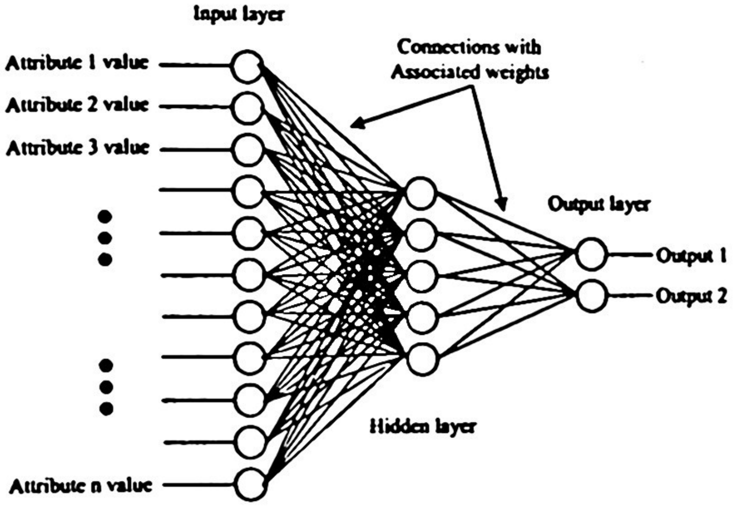

Figure 2. Neural Network Architecture.

1.2 Neural Networks

For a long time researchers have tried to simulate how the human brain works with mathematical models called neural networks [7]. A neural network is composed of a set of cells that are interconnected in a layer fashion. The first layer is called the input layer and its function is to pass the input signals to the next layer. Each cell in the next layer (intermediate cells) receives a signal from each cell in the previous layer modified by a weight factor. The intermediate cells calculate their signal according to a function over its input signals, in the case of the backpropagation algorithm that we use [7], the new signal is calculated with the sigmoid function and then passed to the next layer cells. Once the signals reach the output cells, the result is compared with the real classification of the input feature vector and the error is propagated to the previous layers by adjusting the weights of each connection between cells. In our work with the calcifications detection in mammograms domain, we use a Neural Network (NN) with 11 input nodes, 1 hidden layer with 5 nodes and two output nodes. As we can see in figure 2, the input nodes receive signals from the feature vector and the output nodes show the classification of the NN for the given input vector. In the case of the calcifications detection domain, the possible classifications are positive (the feature vector corresponds to a calcification) or negative (the feature vector does not correspond to a calcification). The NN is trained with the back propagation algorithm comparing the actual output of the network with the real output for each input vector and changing the weights of the network to reduce the output error.

1.3 Decision Trees

Decision trees [8] are a classification method that generates a tree to classify a set of input examples according to their class. Each branch in the tree represents a decision. Each node in the tree refers to a particular attribute. Edges connecting nodes are labeled with attribute values and leave nodes give a classification that applies to the examples that were reached through that branch. At each step of the tree construction, a node is selected according to a statistical measure called information gain that measures how well a node (attribute) distributes the input examples with respect to their class. Figure 3 shows part of an example of a decision tree for the calcifications detection domain. As we can see, the root node for the tree is the area node. If area has a value of less or equal to 13, we verify the value of the diameter attribute. If diameter has a value of less or equal to 2 then the class of the example is positive. If the value for diameter is greater than 2, we verify again the value of the diameter attribute and if it is less or equal to 2.83, we verify the convexity attribute. If the convexity attribute has a value of less or equal to 0.93, the class is positive, otherwise the class is negative.

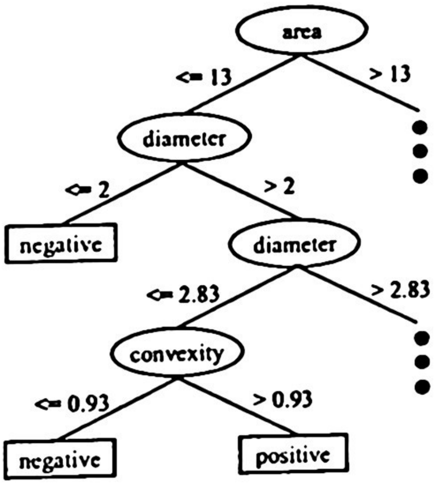

Figure 3. Partial Decision Tree.

1.4 Subdue

Subdue ([4], [5]) is a relational learning system used to find substructures (subgraphs) that appear repetitively in the graph representation of databases. Subdue starts by looking for the substructure that best compresses the graph using the Minimum Description Length (MDL) principle [6], which states that the best description of a data set is the one that minimizes the description length of the entire data set. In relation to Subdue, the best description of the data set is the one that minimizes:

$$I(S) + I(G|S)$$

where S is a substructure found in the input graph G, $I(S)$ is the length (number of bits) required to encode S, and $I(G|S)$ is the length of the encoding of graph G after being compressed using substructure S.

After finding the first substructure, Subdue compresses the graph and can iterate to repeat the same process. Subdue is able to perform an inexact match that allows the discovery of substructures whose instances have slight variations. Another important characteristic of Subdue is that it allows the use of background knowledge in the form of predefined substructures.

The model representation used by Subdue is a labeled graph. Objects are represented by vertices, while relations are represented by edges. Labels are used to describe the meaning of edges and vertices. When we work with relational databases, each row can be considered as an event. Events may also be linked to other events through edges. The event attributes are described by a set of vertices and edges, where the edges identify the specific attributes and the vertices specify the values of those attributes for the event. Figure 4 shows the star graph based representation that we used with Subdue for the tuberculosis domain. As we can see in figure 4, vertices represent either objects as in the case of the "EVENT" node (there is an event vertex for each case in the database) or attribute values as in the case of vertex labeled "2002" that corresponds to the value of attribute "Year" that appears in the label of an edge. This graph-based representation applies for a flat domain but graphs can be used to represent complex structured domains such as chemical compounds.

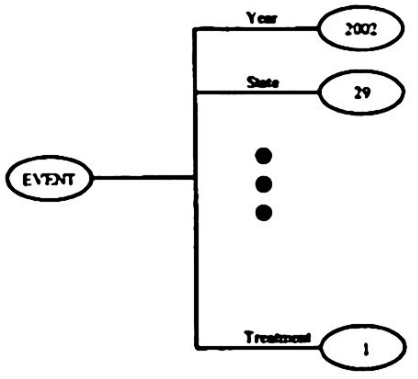

Figure 4. Graph-Based Representation for Subdue

After describing the KDD process and the data mining algorithms used in this research, the following sections show our work in the calcifications detection in mammograms and the tuberculosis domains.

2 Calcifications Detection in the Mammograms Domain

Breast cancer is the second cause of death for women with cancer after cervical uterine cancer and is considered a public health problem. According to statistical data from INEGI, breast cancer was the 12th cause of death for Mexican women with 3,574 deaths in 2001.This is the reason why we decided to attack this problem and found that early detection is one of the best measures to do this. We are working with Dr. Nidia Higuero from the ISSSTEP hospital and created a database of mammograms, which is described in section 2.1. Our domain expert selected the set of cases and gave them to us for scanning. After the images were digitized, Dr. Higuero put marks to those images with calcifications in the places where those calcifications were found; we will refer to these marked images as positive mammograms. We needed these marked images for training purposes, as we will mention in the methodology section. The goal is to find patterns describing calcifications so that we can use them to predict the existence of calcifications in new images and provide the radiologist with a second opinion of his diagnosis. Since our dataset consists of images of mammograms, we need to preprocess the images and create a dataset with the desired characteristics corresponding to those Regions Of Interest (ROI's) known as calcifications. For the preprocessing step of the mammograms we used the machine vision techniques described in the methodology section.

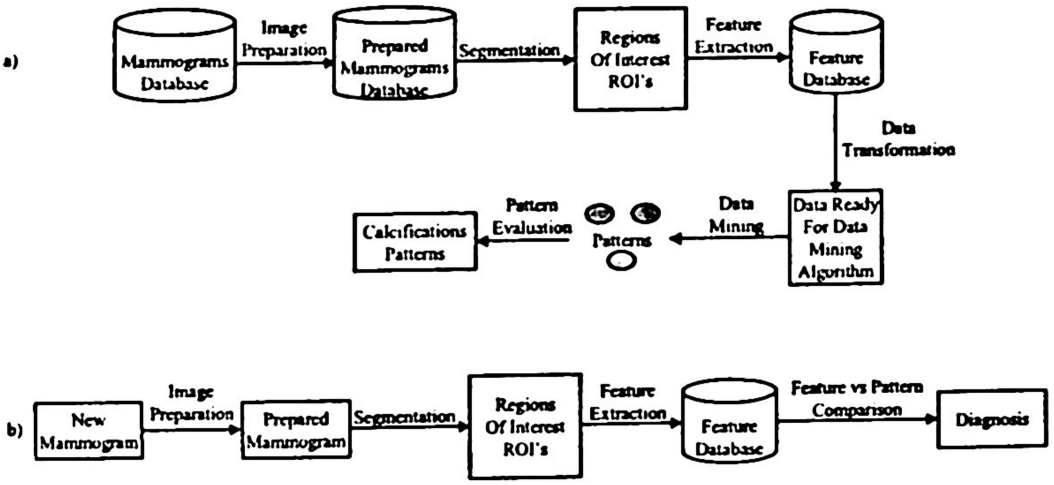

Figure 5. KDD Process Applied to Find Calcifications in Mammograms. a) Training Phase, b) Diagnosis Phase.

2.1 Database

For our experiments we are creating a mammograms database in coordination with our domain expert, Dr. Nidia Higuero from the ISSSTEP hospital. Until now, we have a set of 84 cases of mammograms (one case per patient), each case contains four images, one craniocaudal and one oblique view of each breast. The images were digitized with an Epson Expression 1680 fire-wire scanner at 400 dpi's, with a size of 2,500 x 2,500 pixels in bmp format. From the 84 cases, 54 have calcifications and 30 are normal (with no calcifications). Our domain expert selected the set of cases and gave them to us for scanning.

2.2 Methodology

The combination of machine vision and data mining techniques to find calcifications in mammograms is shown in figure 5. Figure 5a shows the knowledge discovery in databases (KDD) process that we use to find patterns that describe calcifications from known images (those images for which we know if they contain calcifications or not). The process starts with the image database that consists of the original and marked mammograms. As we mentioned before, marked mammograms identify where calcifications are located in a positive mammogram. An image preparation process with a wavelet filter is applied to these images to make calcifications easier to detect. After this, a segmentation algorithm is applied to each positive image in order to get our positive ROI's corresponding to calcifications. A different segmentation algorithm is used to get our negative ROI's, that is; regions of interest of areas that are very similar to calcifications but that are not. After this step we have a ROI's database where each ROI is classified as positive or negative. Next we apply a feature extraction process to each region of interest to create a feature database that will be used to train the data mining algorithms. In our case we use a back-propagation neural network and a decision tree for the data-mining step. After applying the data-mining algorithm, we get patterns to be evaluated in the pattern evaluation step. In the case of the decision tree, our patterns are human understandable as we will show in section 2.3, but the patterns found with the neural network are difficult to interpret because are hidden in the neural network weights and architecture. The patterns found in the training phase will be used in the diagnosis phase shown in figure 5b. The diagnosis starts with the image to be analyzed. We first perform the image preparation filter and then use a segmentation algorithm to find our ROI's (in this case we do not have a marked image) that correspond to possible calcifications. Once we have our ROI's, we execute the feature extraction algorithm to get our feature database, where each ROI is represented by a feature vector. In the next step, we compare each ROI (represented by its feature vector) with the patterns found in the training phase and if the ROI matches any of the positive patterns we diagnose that ROI as positive or a calcification and as negative otherwise.

2.3 Experiments and Results

Data-mining is the task of finding interesting, useful, and novel patterns from databases. In our case we want to find patterns that describe calcifications in mammograms so that we can use them to predict whether a new mammogram has calcifications or not. For this purpose, we use a back-propagation neural network and a decision tree as our data mining algorithms. Neural networks have the property of achieving high accuracies for the classification task but what they learn is not easy to understand. On the other hand, decision trees are known to achieve high accuracies in the classification task and are also easy to understand. For our experiments we find positive and negative regions of interest from 70 mammogram images with calcifications and 60 mammogram images with no calcifications. From these images, we found a total of 649 ROI's, from which 327 were positive (calcifications) and 326 were negative (non calcifications). We performed the feature extraction process to these ROI's and trained our data mining algorithms with them. We used the 10 fold cross validation technique to evaluate the algorithms performance. With the neural networks algorithm we achieved a predictive accuracy of 94.3 % and with decision trees 92.6%. As we can see, we got better results with the neural network algorithm than with the decision tree. The only problem with the neural network results is that they are not easy to understand and we needed to show the learned patterns to our domain expert. This is why we generated rules with the decision tree and asked our domain expert to study them. Dr. Higuero found the rules very interesting and told us that she was relating them to the way she does her diagnosis. Figure 6 shows two of these rules. The first rule says that if the area of the ROI is less or equal to 13, and its diameter is less or equal to 2, then it might not be a calcification. Rule 2 says that if the roundness of the ROI is less than 0.68, and its contour length is greater than 10.49, then it might be a calcification. These rules are obtained from the decision tree by following the path starting at the root node to the leaves.

1) If area <= 13, and diameter <= 2, then calcification = negative
2) If roundness > 0.68, and contour length > 10.49, then calcification = positive

Figure 6. Decision Rules from the Calcification Detection Domain.

Another medical domain that we have been working with is the tuberculosis domain, which is discussed in the following section.

3 Tuberculosis Domain

Since 1993, the world program against Tuberculosis created by the World Health Organization (WHO) took the decision to declare tuberculosis as a "world emergency". A lot of efforts have been focused to get it into control and eradicate it. Although some advances have been achieved with the SSST (Strictly Supervised Shortened Treatment) [2], this year there will be more deaths caused by tuberculosis than any other year in history [3]. According to the WHO with the World Health Report 2002, tuberculosis represents the number eight cause of death in the world, which is even more dangerous than car accidents, breast cancer and bronchial and tracheal diseases. Although we recognize the success in some countries in the fight against tuberculosis, it has not been overcome, even less with its new allied: the AIDS virus. Some of the causes of the tuberculosis problem are due to human factors (as an individual or as a determined group of risk). The Mycrobacterium tuberculosis bacillus, also known as the Kotch bacillus, knows how to protect itself. After thousands of years of evolution it has developed extraordinary surviving strategies. Because of this, it is very difficult to win a battle against such an enemy with a lot of advantages. This is the reason why we need to use new technologies for a good strategy against the disease. We think that we can find important knowledge to fight tuberculosis from treatment databases and here is where data mining techniques can help us to design a fight strategy against tuberculosis in Tlaxcala so that we can definitely eradicate it.

3.1 Database

The tuberculosis database contains data about the evolution of patients with tuberculosis under treatment. The database consists of 232 cases with information about the patient, the diagnosis, the treatment, and the classification of the case. Patient related data identifies the patient and the place where he was under treatment. Diagnosis data describes laboratory results and the diagnosis method that was used among other information. Treatment information describes how the patient evolved to the treatment, which medication he used, etc. The final classification of the patient shows the treatment length, treatment schema and the class of the patient by the time he finished the treatment. The class can be healthy, non-healthy, death, unfinished treatment, moved to another location, in treatment, or failure.

3.2 Methodology

In this section we explain the steps of the KDD process applied to the tuberculosis domain. In this domain we did not need the image preprocessing steps as we did for the calcifications detection domain. In the first step of the process (development and understanding of the application domain), we studied the tuberculosis disease so that we could understand it. We also invited a domain expert (a doctor who has studied the disease for a long time) to help us to understand the domain and to evaluate the patterns found with the data mining techniques in the database. The domain expert wants to use the discovered knowledge to enforce the prevention and control program to eradicate the tuberculosis disease from Tlaxcala. In the second step (creation of the target data set), we selected a subset of the attributes and examples from the database to use in the KDD process. For our first experiment, we selected all the data of all the registered patients with the tuberculosis disease (that is, we used the whole database). In other experiments we used only data for a specific period of time (months of treatment) in order to find patterns that help us to make suggestions about the treatment process, which lasts for 6 months. In the data preparation step we eliminated noise, discretized some attributes and replaced some null values with a constant. The data transformation phase involved the transformation of information into star graphs, one star graph per case. For the data mining task we chose to use the Subdue algorithm that uses a graph-based representation as described before. For the pattern interpretation and evaluation step we had the participation of our domain expert. He carefully analyzed the patterns found and decided which of them were useful and at the same time contained something important about the data. Finally, in the knowledge consolidation step, we documented the patterns that we considered knowledge (patterns that were novel and useful) and we will try to use them to create a program that help us to have a better control in the eradication of the tuberculosis disease. It is important to say that some of the steps in the KDD process are not clearly separated. It is possible to go back to a previous step and change a decision taken before in order to enhance the results (KDD is an iterative process). A light modification in one of the steps might strongly affect the rest of the process. The whole process is crucial in order to succeed.

3.3 Experiments and Preliminary Results

As we have not concluded our experiments in the tuberculosis domain, we will only present some preliminary results in this section. For our experiments with Subdue in the Tuberculosis domain we used all the cases (a total of 232) stored in the database. In this experiment we eliminated missing values because if we replaced those missing values with any other value we found that the substructures found were biased by them. Figure 7 shows one of the substructures found by Subdue in the tuberculosis domain. This substructure was found in 156 of 232 cases and is telling us that the 67% of the patients that were under treatment (Treatment = 1) in 2002 (Year = 2002) had lung tuberculosis (Localization = 1) that was diagnosed by the bacillus test (DiagnosisMethod = 1), the patient was found to have tuberculosis in an external consult (DetectionService = 1). The fact that these cases belong to state number 29 (Tlaxcala) is not relevant because all the patients are from Tlaxcala.

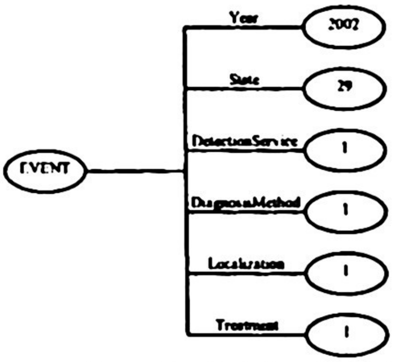

Figure 7. A Substructure Found by Subdue in the Tuberculosis Domain.

We are using the substructures found by Subdue to enhance the methodology used to eradicate the tuberculosis disease. That is, we need to find out what is failing in the process so that we can reduce the number of deaths caused by tuberculosis. This is how the patterns found in this domain are helping the domain expert in this task. We still need to make more experiments with Subdue (including its concept learning version called SubdueCL) and try other algorithms such as decision trees and association rules. We also need to test the predictive accuracy of the patterns found.

4 Conclusions and Future Work

As we could see through this paper, data mining techniques applied to medical domains can really contribute to improve the data analysis capacity to find hidden knowledge from databases. In the case of the calcifications detection in mammograms domain, we were able to find patterns describing calcifications that can be used to detect calcifications in new images with a predictive accuracy of up to 94.3 %. Radiologists can use this method as a second opinion for their diagnosis or to train radiology students. In the case of the tuberculosis domain, we are finding descriptive patterns that are used by the domain expert to study what might be wrong in the current process used to eradicate the tuberculosis disease so that the process can be enhanced. Our future work in the calcifications detection in mammograms domain we want to try different data mining algorithms such as Bayesian networks and also we want to find other characteristics that yield to better results. In the case of the tuberculosis domain, we want to use the association rules and decision trees algorithms to find different patterns and we also want to test the predictive accuracy with the patterns found.

References

1. Gregory Piatetsky-Shapiro and William J. Frawley, "Knowledge Discovery in Databases," AAAI Press/The MIT Press, Menlo Park, California 1991.
2. World Health Organization, ¿What is DOTS/TAES?, Guide to Understand the anti-tuberculosis fight strategy Recommended by the OMS and Known as the DOTS/TAES Strategy, 1999.
3. J. Sauret Valet, "Tuberculosis, Recent Vision", Grupo Aula Medica, 2001.
4. D. J. Cook, and L. B. Holder. Substructure Discovery Using Minimum Description Length and Background Knowledge. Journal of Artificial Intelligence Research. 1:231-55, 1994.
5. D. J. Cook, and L. B. Holder. Graph-Based Data Mining. IEEE Intelligent Systems, 15(2):32-41, 2000.
6. Rissanen, J. Stochastic Complexity in Statistical Inquiry. World Scientific Publishing Company, 1989.
7. David E. Rumelhart, Bernard Widrow, and Michael Alehr, The Basic Ideas in Neural Networks. Communications of the ACM Vol 37, No 3, pp. 87-92, 1994.
8. J. R. Quinlan, Improved Use of Continuous Attributes in C4.5, JAIR vol 4, 77-90, 1996.