Integration of Data Mining and Hybrid Expert System

I.G.L. da Silva, B.P. Amorim, P. G. Campos, L.M. Brasil

Informatics Laboratory in Health (LABIS)
Nucleus of Studies and Technology in Biomedical Engineering (NETEB)
Federal University of Paraíba (UFPB), Campus I - Postal Box 5095
ZIP Code: 58051-970, Tel: +55-83-2167067, Fax:+55-83-2167369, João Pessoa - PB, Brazil
(ismenia, bruno, paulemir, lmb)@neteb.ufpb.br

Abstract

In the latest decades a great growth in the capacity of generating and collecting data has happened. The progresses in the data collection and storage, combined with the extensive use of the Database Management System and Data Warehousing technology, have contributed to this growth. However, the traditional methods used to manipulate these data can generate informative reports, but they cannot analyze the content of the data to point out what knowledge interests most. These difficulties contributed to the arising of intelligent tools and techniques to analyze data resultant from the emergent field of the Knowledge Discovery in Databases (KDD). KDD is not a trivial process and it is used to identify valid, potentially useful and comprehensible patterns in the database. The main goals of this work are to apply some of these tools and techniques to a medical database of breast cancer, so that detection and prediction patterns are discovered, and use the database resulting of the mining process in a hybrid expert system that will help in medical diagnosis.

Keywords: Data Mining, Hybrid Expert System, Artificial Neural Networks, Fuzzy Logic, Extraction of Rules.

Knowledge Discovery in Databases

According to Fayyad et al. 1996, the KDD process can be divided into several stages: exploratory analysis of the data, selection of attributes, pre-processing, choosing of mining algorithm, transformation of the data, data mining and interpretation of the results [1]. These stages can be repeated many times until the desired knowledge is obtained. The following sections describe how each stage was accomplished in this work.

Exploratory Analyzing of the Data

The goal of this stage is to analyze thoroughly the database that will be mined, so that it is possible to select the appropriate variables to the knowledge discovery. Thus, it was held a study of the database to comprehend its structure, the meaning and the attribute domain.

The database used contained 699 records [2]. Each record corresponding to a case of breast cancer has 11 attributes. The first attribute corresponds to the number that identifies the case, the nine following attributes are concerned to the necessary symptoms to reach a diagnosis and the last one refers to the type of cancer. Table 1 describes record format.

Attribute	Description	Domain
Code	Sample code number	Id number
S1	Clump Thickness	1 – 10
S2	Uniformity of Cell Size	1 – 10
S3	Uniformity of Cell Shape	1 – 10
S4	Marginal Adhesion	1 – 10
S5	Single Epithelial Cell Size	1 – 10
S6	Bare Nuclei	1 – 10
S7	Bland Chromatin	1 – 10
S8	Normal Nucleoli	1 – 10
S9	Mitoses	1 – 10
Type	Class	2 if benign or 4 if malignant

Table 1. Record format of the database

Selection of Attributes

At this stage, one must select a set of data or point out a subset of variables or data samples that will be used for the knowledge discovery.

Because a domain expert has already provided the database and some of the attributes describe the symptoms and associated diagnosis, the whole database was chosen to be mined.

Pre-processing

At this stage, one must eliminate noises and errors; establish procedures to verify the lack of data, conventions for naming and other long-term steps to the construction of a consistent database.

The following steps were accomplished in this work:

1. Elimination of the attribute *code* because it is not important to the problem;
2. The database had many redundant records. As these records are not going to contribute to a better performance of the expert system, they were eliminated. To accomplish this task, it was developed a program in C^{++} language;

3. In 16 records, the attribute *S6* had null value. Instead of eliminating these records, we decided to substitute it for the most frequent values in the other records.

From the 699 initial records, 457 were selected. Table 2 shows the distribution of the records.

Type of cancer	Before the pre-processing	After the pre-processing
2	458 records (65.52%)	219 records (47.92 %)
4	241 records (34.48%)	238 records (52.08 %)

Table 2. Percentage distribution of the records according to the type of cancer

Choosing of Mining Algorithm

At this stage, one must select the methods that will be used to search for patterns of the data. This includes the choosing of appropriate models and parameters, and the association of a data mining method.

An expert system was used to accomplish the tasks of data mining and interpretation of the results. The following sections will describe this system.

Transformation of the Data

At this stage, the format of the attribute values must be modeled on the formats demanded by the mining algorithm. This transformation is accomplished by the hybrid expert system (HES) in the step where the input values are mapped by fuzzy values [3][4].

Partition of the Data

To execute the learning and the test stage of the system, it was used the holdout procedure without validation, resulting in the partition of the database in two sets [5]:

1. Training: 2/3 of the database
2. Test: 1/3 of the database

The training and test sets have the same proportion of classes as the whole set of data.

The Hybrid Expert System

Hybrid architectures are a new field of Artificial Intelligence research concerned with the development of the next generation of intelligent systems. Current researches focus on integrating the computational paradigms of Symbolic Manipulation and Artificial Neural Networks (ANN). In this work we developed a Hybrid Expert System (HES) to aide in the medical domain [3][4].

A database represents the knowledge of the domain expert has been used. This helped to overcome the domain expert's difficulty in specifying all rules mainly when imprecision pervades to the problem.

It has been developed a program to extract orthogonal initial rules, which cannot be derived from other rules, from the database to implement a basic structure of a Neural Network Based Expert System (NNES) [3][4].

After, the NNES is refined through a training algorithm, Genetic-Back-propagation Based Learning Algorithm (GENBACK) [7]. This algorithm was inspired in the classical Back-propagation one. The NNES also foresees the possibility of different kinds of variables in its input as quantitative, linguistic, or boolean valued.

A problem related to the trained ANN is its difficulty in explaining how it has gotten to a solution. It was developed a technique to extract rules from a trained fuzzy NNES - Fuzzy Rule Extraction (FUZZYRULEX) [6]. The extracted rules are used to form an equivalent RBES, through which it is possible to obtain the explanations by making use of a backward chaining method. An AND/OR graph is applied as an intermediate process between the NNES and the RBES. So, the visualization and withdrawal of necessary data is easier for the domain expert understanding, as well as for the utilization in the RBES.

It must also be considered that fuzzy logic was used to deal with imprecision. The logic fuzzy theory gives a very good mathematical focus to represent the imprecise knowledge, which is also a goal of this work [3][4].

NNES

The NNES is built based on fuzzy rules obtained from the database during the initial stages of the KDD process. These initial rules are used for the implementation of an ANN initial topology. Two sets of examples composed of the symptoms and diagnosis are also obtained from the database resulting of the mining process. The first set is used to refine the NNES through the proposed learning algorithm. The other one is used to validate the refined NNES (Figure 1).

The initial NNES consists of one input layer, one hidden layer and one output layer. The input layer neurons correspond to the symptoms, in other words, the IF part of the rules. The hidden layer, which will be optimized by the proposed algorithm, represents the intermediate hypotheses that correspond to the initial rule conditions obtained. The output layer represents the diagnosis, which are the rule conclusions or the THEN part.

The AND/OR graph, which represents the concepts and relations, indicates the number of neurons in the ANN input, hidden and output layers. The graph also shows the existence of intermediate concepts and its connections, which are decoded in the hidden layer of the ANN. The hidden layer is defined as the AND nodes, and the output layer as the OR nodes [3][4].

The fuzzy factors in fuzzy rules are treated the following way: the values in fuzzy logic are not abrupt and their membership degree is defined by using a membership function. The method used in this work to match these membership functions was the Max/Min operators. As all input values are numerical, the "fuzzification" process deals with them in the following way: through the interpolation of values, it can have a numerical performance interval that will depend on the problem and data obtained.

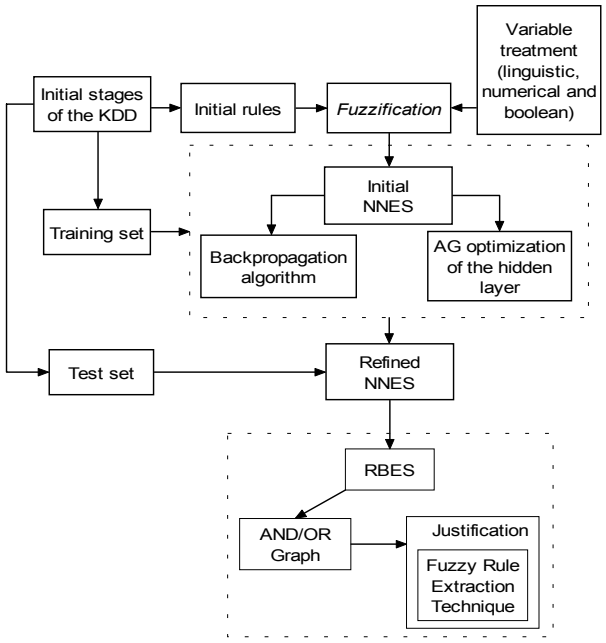

Figure 1. General diagram of the HES

Learning Algorithm. The system optimization and learning, at the refinement stage, is accomplished through the algorithm GENBACK [7]. This algorithm uses Genetic Algorithm (GA) as a tool to help choose the NNES best topology and it is applied to a static feedforward Artificial Neural Network (ANN) of multiple layers and AND/OR fuzzy neurons.

GENBACK has the following characteristics: it uses GA to help in the optimization of the hidden layer of the ANN, it deals with the network input variables through the fuzzy logic and, in the place of the sum of the weights, it uses AND/OR logical connectives [3][4][7].

GENBACK, at its first stage, considers that the NNES already exists. It uses the set of examples to accomplish changes not only in the connection weights, but also in the network structure. The connections, that may not be in the initial rules, are included or excluded among the neurons, as well as neurons can be included or excluded from the hidden layer by applying GA. So, the algorithm can produce more concepts that were not in these initial rules and, consequently, we can deduce rules that the expert was not able to think of.

GENBACK algorithm has following steps:

1. Creation: individuals of a population. It was applied to the Gauss distribution;
2. Codification: chain of chromosomes. It was used a fixed chain of chromosomes of 8 bits;
3. Training: based on back-propagation algorithm;
4. Evaluation: used the fitness functions, e.g.,

   $F_1 = N_{CH}/ERROR$

   $F_2 = 1/N_{CH} + ERROR$

   where F_1 and F_2 = fitness function, objective function or cost function, N_{CH} = neurons number in the hidden layer of the NNES, *ERROR* = error in the NNES output;
5. Applying the roulette wheel process;
6. Ordering: each individual in a population;
7. Application: selection, crossover and mutation;
8. New generation;
9. Return to step 3: the winner NNES.

RBES

The RBES is composed of two modules. One of them corresponds to the mapping of the data obtained from the output of the AND/OR fuzzy ANN in an AND/OR Graph, so that it is easier to comprehend these data for the implementation of the RBES. The other one is the justification, where the rule extraction technique is applied. The goal of this technique is to give an explanation of the answer obtained from the NNES output. It is used the fuzzy rule extraction algorithm FUZZYRULEXT [6].

The Fuzzy Rule Extraction Technique. The developed technique has two stages: AND/OR Graph and Path Generation by Backtracking. However, it differs from other techniques due to the following characteristics:

1. It uses an AND/OR Graph between the NNES and the RBES;
2. Input data are defined as fuzzy values and represented by membership degrees. Consequently, the properties used to obtain clauses of a rule are simplified [4];
3. About the accumulative weights, the algorithm deals with not only maximum weights between the output and hidden layers, but also minimum weights between the hidden and input layers since some input data are refused;
4. It uses an elimination process of redundant rules, instead of using a stop condition based on the number of created rules. So, the number of rules is reduced, simplifying the RBES construction;
5. Horn clauses are used due to the facility for implementing rules [8].

In the path generation by backtracking stage, the user can ask the system why it has inferred a particular conclusion. The system answers with an IF-THEN rule. It has been noted that these if-then rules are not explicitly represented in the encoded knowledge basis. They are generated by the inference system of the connection weights as it is necessary to get more explanations.

Fuzzyrulext Algorithm. The proposed algorithm has the following steps:

1. Choosing an input-output pattern;
2. Selecting those neurons i in the preceding layer that have a positive impact on the conclusion at output neuron j;
3. Letting the set of m_i neurons of the hidden layer, previously selected, be denoted as $\{a_1, a_2, ..., a_{mi}\}$ and let their connection weights for neuron j in the output layer be given by the set $wet_{(n)ak} = \{w_{(n)ja1}, w_{(n)ja2}, ..., w_{(n)jamj}\}$
4. Determining the set of accumulative link weights $wet_{(n-1)i}$ for neuron i in the hidden layer along the maximum weighted path be as

   $$O_{(n)i} > 0 \quad and \quad w_{(n-1)aki} > 0$$

$$wet_{(n-1)i} = max[wet_{(n)ak} + w_{(n-1)aki}]$$

5. Selecting the set of input neurons: $m_l = \{a_1, a_2, ..., a_{ml}\}$;
6. Making the arrangement in the decreasing order of Net Impact (NI) of the elements of the set of weights obtained in at last step, e.g., $NI_{(n-1)i} = O_{(n-1)i} * wet_{(n-1)i;}$
7. Selecting l_s *input* neurons for the clauses with $w_{(n-2)l} > 0$ and l_p input neurons remaining with $w_{(n-2)l} < 0$, such as $m_l = |l_s| + |l_p|$;
8. Clause generation: For a neuron l_{s1} in the input layer, selected for clause generation, the corresponding input feature u_{sl} for boolean or numeric inputs is obtained as $u_{s1} = (l_{sl} - 1) + 1$;
   The antecedent of the rule is given by the numerical or boolean property being determined as

$$prop = \begin{cases} -0.8 & \text{if } u_{sl} \le -0.8 \\ 1 & \text{if } u_{sl} = 1 \\ -0.8 < x < 1 & \text{otherwise} \end{cases}$$

   The corresponding input feature u_{sl} for linguistic inputs is obtained as $u_{sl} = (l_{sl} - 1) \bmod 3 + 1$;
   The antecedent of the rule is given by the linguistic property being determined as

$$prop = \begin{cases} \text{strong} & \text{if } u_{sl} \ge 0.8 \\ \text{moderate} & \text{if } -0.4 < u_{sl} < 0.7 \\ \text{weak} & \text{if } u_{sl} \le -0.4 \end{cases}$$

   The certainty measurement for each output neuron is defined as

$$bel_j = O_{(n+1)j} - \sum_{i \neq j} O_{(n-1)i}$$

   The consequent of the rule is given by the property being determined as

$$prop = \begin{cases} -0.8 & \text{for } -5 \le bel_j \le -0.2 \\ 1 & \text{for } bel_j > -0.2 \\ \text{Not recognise} & \text{for } bel_j < -5 \end{cases}$$

   The process is repeated until the IF-THEN rules to justify all outputs presented by the input-output patterns of the trained NNES are indicated.
9. Elimination process of redundant rules: determining the total number of rules generated at the last stage; creating one vector for the input-output data; comparing antecedent-consequent of each rule for determining how many rules are similar; keeping those rules that are distinct; eliminating those redundant rules; determining the total number of rules generated without redundancies.

Simulations

The simulations related to the HES were accomplished at first for clinical cases of classification of epileptic crises. The set of data used in the system was obtained at the Hospital of the Federal University of Santa Catarina, Brazil. The initial neural structure has 32 neurons in the input layer (symptoms), 11 neurons in the hidden layer and 4 neurons in the output layer (diagnosis).

Presently, the HES is being tested to a medical database of breast cancer. The set of data used in the system was obtained at the Hospital of the University of Wisconsin, Madison, USA. The initial neural structure has 9 neurons in the input layer (symptoms), 10 or 15 neurons in the hidden layer and 2 neurons in the output layer (diagnosis).

NNES

Table 3 and Table 4 show the results obtained in some simulations for clinical cases of classification of breast cancer. The Gauss distribution has been used in these simulations. The simulations 3 e 4 used the fitness function F1. The others used the fitness function F2.

Data										
S	*NG*	*IP*	*CR*	*MR*	*T*	α	β	*NE*	*Nhlw*	*F*
1	3	6	0.6	0.1	0.1	0.4	0.1	2000	4	0.10
2	3	6	0.6	0.1	0.1	0.4	0.1	2000	4	0.12
3	3	6	0.6	0.1	0.1	0.4	0.1	2000	13	3.09
4	3	6	0.6	0.1	0.1	0.1	0.1	2000	20	2.64

Table 3. Data of the simulations [a]

[a] S - simulations; NG – number of generations; IP - initial population; CR – crossover rate; MR – mutation rate; T – tolerance; α - learning rate; β - momentum; NE – number of epochs; Nhlw - number of neurons in the hidden layer of the winner network; F – fitness

Quantity of test patterns		
S	*NH*	*PH (%)*
1	5	50.0
2	6	60.0
3	7	70.0
4	8	53.3

Table 4. Recognition of patterns [b]

[b] S - simulations; NH – number of hits; PH – percentage of hits.

At the following, see some simulations to for clinical cases of classification of epileptic crises.

Data										
S	*NG*	*IP*	*CR*	*MR*	*T*	α	β	*NE*	*Nhlw*	*F*
1	5	8	0.6	0.1	0.01	0.1	0.05	250	15	1.45
2	5	8	0.6	0.1	0.01	0.1	0.05	1,000	16	1.78
3	5	8	0.6	0.1	0.01	0.3	0.7	1,250	16	1.65
4	5	20	0.6	0.1	0.01	0.1	0.05	1,500	16	1.79

Table 5. Data of the simulations [a]

Quantity of test patterns		
S	*NH*	*PH (%)*
1	9	81.8
2	7	63.6
3	7	63.6
4	8	72.7

Table 6. Recognition of patterns [b]

RBES

Table 7 shows the results of simulations for clinical cases of classification of epileptic crises.

S	*NIL*	*NHL*	*NOL*	*DST*	*NRR*	*NRW*	*NH*	*PH (%)*
1	6	4	3	6	8	7	5	83.3
2	6	10	3	6	23	12	5	83.3
3	6	10	3	6	30	13	5	83.3
4	32	13	4	9	43	27	7	77.8

Table 7. Classification of epileptic crises [a]

[a] S- simulations; NIL – number of neurons in the input layer; NHL – number of neurons in the hidden layer; NOL – number of neurons in the output layer; DST – data set of test; NRR – number of generated rules with redundancies; NRW – number of generated rules without redundancies; NH – number of hits; PH – percentage of hits

The RBES is still being adapted to the breast cancer domain. Based on the simulations for clinical cases of classification of epileptic crises, whose results were among 77.8 up to 83% of accuracy, we believe that the performance of the RBES will be similar to first medical example.

Conclusions

The database obtained by the KDD process helped us to overcome the domain expert's difficulty in specifying all rules mainly when imprecision pervades to the problem and to discover detection and prediction patterns.

The training of the fuzzy ANN was inspired in the classic backpropagation algorithm, with some alterations: the optimisation of the hidden layer was supported by GA, incorporation of logical operators AND/OR in place of the weighted sum, and formation of the ANN by fuzzy logic. Nevertheless, it was observed that at the backward step the error propagation between the output and the hidden layer has reached expected values.

Another reached goal was related to the optimisation of the topology to be adopted by the fuzzy ANN. The optimisation of the hidden layer was supported by GA. Meantime when it is applied for this goal, we must take care to respect a maximum and minimum number of neurons of this layer.

The fuzzy rule extraction technique has achieved satisfactory values for clinical cases of classification of epileptic crises. Therefore, the system was able to translate the encoded knowledge among the connection weights of the NNES into rules. Besides, all examples applied demonstrated that the procedure for reducing the number of redundant rules produced after the learning and refinement stages of the ANN were appropriate. The RBES prototype developed for this medical application achieved a satisfactory result.

At moment, the HES is being tested to a medical database of breast cancer. The first results are being obtained.

Acknowledgements

We would like to thank The National Advice of Scientific and Technological Development (CNPq), the Institutional Program of Scientific Initiation Scholarships (PIBIC) for the research grant, and The Hospital of the University of Wisconsin (Madison-USA) and Dr. William H. Wolberg for giving us the database on breast cancer.

References

[1] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. eds. 1996. *Advanced in Knowledge Discovery and Data Mining*. Cambridge, MA: AAAI/MIT Press.

[2] Campos, P. G., Brasil, L.M., Filho, M. T. B. 2001. BRECADES – Breast Cancer Aided by Diagnosis Expert System. In Proceedings of the First Vietnam-Japan Bilateral Symposium on Biomedical Imaging/Medical Informatics and Applications (VJMEDIMAG'2001), 84-91. Hanoi, Vietnam.

[3] L.M. Brasil, F.M. de Azevedo and, J.M. Barreto. "A hybrid expert system for the diagnosis of epileptic crisis". Special Issue in Artificial Intelligence in Medicine (AIM), ISSN 0933-3657/00,v.1-3, no. 21, p. 227-233, Amsterdam, Holland, 2000.

[4] Brasil, L.M. Brasil. 1999. *A proposal of architecture for hybrid expert system and a corresponding elicitation/representantion methodology of knowledge*. Ph.D, diss., Dept. of Electric Engineering, Federal University of Santa Catarina (UFSC), Brazil (in Portuguese).

[5] Witten, Ian H., and Frank, E. eds. 2000. *Data Mining: Tratical Machine Learning Tools and Techiniques with Java Implementations.* Sun Francisco, CA: Morgan Kaufmann.

[6] Amorim, B.P., Brasil, L.M., Rojas, J.C.C., Silva, I.G.L., Filho, M.T.B., Azevedo, F.M., Almeida, A.E.M. 2001. Extraction of fuzzy rules for and/or fuzzy artificial neural networks. In Proceedings of the First Vietnam-Japan Bilateral Symposium on Biomedical Imaging/Medical Informatics and Applications (VJMEDIMAG'2001), 108-114. Hanoi, Vietnam.

[7] Silva, I.G.L., Brasil, L.M., Amorim, B.P., Rojas, J.C.C., Filho, M.T.B., Azevedo, F.M., and Almeida, A.E.M. 2001. Optimization and learning algorithm for neuro-fuzzy-ga expert system. In Proceedings of the First Vietnam-Japan Bilateral Symposium on Biomedical Imaging/Medical Informatics and Applications (VJMEDIMAG'2001), 70-77. Hanoi, Vietnam.

[8] Kowalski, R. eds. 1979. *Logic for Problem Solving*. New York: Computer Science Library.