Patients Classification by Risk Using Cluster Analysis and Genetic Algorithms*

Max Chacón and Oreste Luci

Informatic Engineering Department, University of Santiago de Chile,
Av. Ecuador 3659, PO Box 10233, Santiago, Chile.
{mchacon, oluci}@diinf.usach.cl

Abstract. Knowing a patient's risk at the moment of admission to a medical unit is important for both clinical and administrative decision making: it is fundamental to carry out a health technology assessment. In this paper, we propose a non-supervised learning method based on cluster analysis and genetic algorithms to classify patients according to their admission risk. This proposal includes an innovative way to incorporate the information contained in the diagnostic hypotheses into the classification system. To assess this method, we used retrospective data of 294 patients (50 dead) admitted to two Adult Intensive Care Units (ICU) in the city of Santiago, Chile. An area calculation under the ROC curve was used to verify the accuracy of this classification. The results show that, with the proposed methodology, it is possible to obtain an ROC curve with a 0.946 area, whereas with the APACHE II system it is possible to obtain only a 0.786 area.

1 Introduction

In order to determine the admission risk, it is necessary to know the patient's condition at the moment he or she is admitted to a medical unit. This condition represents very valuable information to make clinical decisions such as the admission of the patient to an Intensive Care Unit (ICU) and helps decide the distribution of resources within the unit [1–2]. Besides, it is crucial to carry out a health technology assessment [3]. The problem that exists when determining the admission risk is that, after making the preliminary assessment, the patient is treated, thus modifying his initial condition. A true quantification of this risk could be achieved by assessing the results of the natural evolution of the illness without applying medical technologies, which would be an ethical impossibility.

In medicine, the traditional way to deal with this problem has been the creation of physiological indexes intended to determine the seriousness of the patient's condition at the moment of admission, such as: Simplified Acute Physiology Score (SAPS) [4] or Acute Physiology and Chronic Health Evaluation (APACHE [5]). Among the main disadvantages of these indexes, we may quote the lack of an adequate characterization of the medical information at the moment of admittance, such as a quantification of comorbidities [6], the linear characteristics of the composition of these indexes and the lack

* This study has been supported by FONDECYT (Chile) project No. 1990920 and DICYT-USACH No. 02-0219-01CHP.

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 350–358, 2003.

of precision to determine each individual patient's risk. This last point makes possible that these indexes permit to carry out global comparisons between units (they offer accurate risk values on the average) but they are not of great help to assess technologies within a medical unit.

At present, there exists a number of publications that intend to forecast the outcome of medical procedures. To this effect they have applied different Data Mining methods such as: Logistic Regression, Cluster Analysis, Neural Networks and Bayesian Networks [2,7–9]. Especially, the work carried out by Peña-Rojas and Sipper [8] contains an extensive revision of genetic algorithm applications in various fields of medicine. Of the aforementioned works, the one that is closest to admission risk estimate is the one published by Dubowski et al [7], which analyses the outcome considering the deaths that take place during admittance to the hospital. In spite of the fact that the prediction of the outcome is useful when making some decisions - such as the admission of patients- these predictions do not represent the condition of the patient at the moment of admission and the use of these results to make a technological assessment in health is questionable since it includes the utilization of the same technology it intends to assess [10].

This work's proposal is based on two fundamental points: the first one consists of using a group of variables similar to those used in physiological indexes, such as APACHE [5]. It uses an original way of quantifying diagnostic hypotheses (including co-morbidities) as a way to incorporate the medical knowledge that exists at the moment of admission. The second point consists of using a non-supervised learning method, which takes as a basis the clustering of *k-means*, together with a strong search method like the *genetic algorithms*. This type of learning allows the formation of groups of patients, based exclusively on the information we have about each patient at the moment of admission. Later on we use the information contained in the outcome of the interventions, just to assign risk to each group. It is always possible to do this when accepting the hypothesis that interventions tend to modify the patient's state positively. In this way, the groups that show a greater number of patients who did not survive, in spite of the treatment, represent the groups of higher initial risk. Thus it will be possible, later on, to classify a new patient depending exclusively on the conditions presented at the moment of admission.

In order to assess the proposed methodology, we have used the information taken from 294 medical records of patients who were admitted in two Intensive Care Units (ICU) for adults. These were classified in accordance with afore described methodology. The results were compared using the classification that resulted from applying the APACHE II [11] index by means of the use of ROC [12] curves.

2 Data Collection and Pre-processing

The data were obtained from the Intensive Care Units (ICU) of two public hospitals of the city of Santiago, Chile. A total of 294 medical records of adult patients were collected. 50 out of the 294 correspond to patients who died during their stay at ICU. Table 1 shows a list of the variables that were taken at the moment of admission or during the first 12 hours of hospitalization.

The discreet or binary variables, such as infection upon admission, can be represented directly, but the variables used to represent physiological acuteness require a monotonous

severity scale in order to be used efficiently with the clustering method. These variables do not have a monotonous behavior with respect to seriousness. For example, in its original form, temperature has a normal severity range located at the middle of the scale, but both hypothermia and fever produce an increase in the patient's severity condition. To capture the severity increase in only one direction, we resort to the code used by the APACHE system, which uses discrete ranges between zero and four. In this way, severity always increases from zero value, which is considered normal. To encode age we also use the five ranges defined in the APACHE system.

The method used to quantify the information contained in the diagnostic hypotheses was the knowledge of intensivists with the double purpose of enriching the information contained in the diagnoses and carrying out a quantification that may reflect severity. The procedure included four steps: First, the 780 diagnoses (each patient considers the principal diagnosis and co-morbid states up to a maximum of eight) were split into 17 groups that represent physiological systems and morbid groups (Table 2).

Table 1. List of input variables for the grouping method.

Variable	Units
Age *	In years
Sex	Binary
Glasgow Coma Scale	0 - 10
Admittance Infection	Binary
Admittance with Respiratory Cardio arrest	Binary
Pre-Admittance Surgery	Binary
Temperature* (axilar)	°C
Mean arterial pressure *	mmHg
Heart rate *	Breaths per min
Respiratory rate *	Breaths per min
PaO2-AO2 * (Arterial Oxygen Pressure Difference)	mmHg
pH* (Arterial)	pH (Arterial)
Serum Sodium * (Sodium ion in the blood)	Meq/l
Serum Potassium * (Potassium ion in the blood)	Meq/l
Serum Creatinin* (It measures renal function)	mg/dl
Hematocrit * (It measures globular mass)	%
White Blood Count *	Cel /cm^3
Neoplasia Presence	Binary
Multiorganic Failure	Binary

* Variables were codified according to APACHE index.

Table 2. Quantification of the Diagnostic Hypothesis

Physiological System or Morbid Group
Neurological
Respiratory
Cardiovascular
Renal
Metabolic
Digestive
Endocrine
Immunologic
Obstetric-Gynecologist
Osteomyoarticular
Psychiatric
Uro-gynecological
Transplant
Otorhinological
Hematological
Dermatologic

Then, the specialists classified each diagnosis in three different categories: *chronic, acute* and *hyperacute* (except in the cases of trauma, that are all *acute* or *hyperacute*). A third step considered that in each physiological or morbid group there may exist more than one of these categories or there may exist a combination of these, for instance two *chronic* and one *hyperacute* diagnoses. With these combinations, a classification system

was created that had an associate increasing severity order according to the combination of *chronic, acute* and *hyperacute* in each patient. Finally, an algorithm was used to assign severity condition values to each one of the 17 groups for each patient, according to the combination of existing diagnoses.

3 K-Means and Genetic Algorithms

The combination proposed for the clustering rescues the main strengths of the traditional clustering methods, especially those of the *k-means* [13] and *Genetic Algorithms (GA)* [14]. From the *k-means*, we can save the simplicity in the representation for cluster creation, which has an influence in the capacity to look for different alternatives at the same time, thus increasing the solution space and its capacity to avoid being trapped in local minimals. From a general perspective, the first problem to be solved is to determine the optimum number of clusters. We based the solution to this problem on a measure of quality for the clustering. The existing literature on cluster analyses shows that one of the most efficient measurements to determine the quality of the clustering is that of Calinski-Harabasz [15]. We leave as a parameter the cluster range. Clustering is started with the largest number of clusters. After that, the quality of the clustering is measured and then the closest clusters are identified and merged. This process continues until the lower bound of the selected clusters interval is reached. Finally, the number of clusters that present the highest value for the Calinski-Harabasz measurement is selected.

For the particular case of admission risk, the selection of the clusters ranking made by the user becomes easy since a large number of clusters (over 10) does not contribute any significant advantages in the final classification and it is difficult to make a later identification of the risk of each cluster by means of the assignation of the death rate that is obtained from the results of the medical interventions.

3.1 Representation and Initialisation

One of the crucial aspects in the application of *GA* and meta-heuristic methods in general is the nature of each particular problem. In the case of *GA*, the problem is to determine which elements of the problem will be represented in each individual's chromosome. After assessing different proposals found in the existing literature for the application of *GA* to clustering [16–17], we chose to represent – in each chromosome – the set of centroids calculated in each iteration. In this way, the *GA* will be responsible for running the solution space in a higher hierarchical level and the method component, corresponding to the *k-means*, will be responsible for forming new cluster, assigning the individuals to the nearest centroid.

A random cluster creation was used for the initialization process, making sure that all clusters will have at least one individual. Later on, their centroids, which are used to create chromosomes, were calculated. This initialization process is done for all the individuals of the population.

3.2 Selection, Crossing, and Mutation

The selection was made using the *proportional or roulette method* [17], which is the method that is most commonly used when designing GA. To assign each cluster (or chromosome) its area in the roulette, the measurement of clustering quality called *sum of squares* [17] was used. The proportional method allows to choose, with higher probability, those individuals that have been better evaluated, but it does not discard those who have low qualification. Pairs of chromosomes are gathered for the crossover, the same individual may be selected more than once in subsequent crossovers. In this stage all that matters is to verify that the selected pairs correspond to different chromosomes.

To carry out the crossing, it is necessary to obtain the pair of individuals from the selection stage and to consult the crossing probability, which was previously assigned as an initial parameter of the program. If the indication is no-crossing, the selected chromosomes are copied to the new generation. On the contrary, if the indication is crossing, the crossing operator is applied and the offspring are assigned in the new population. The crossing operator works taking a random number of centroids from both parents in order to transmit them to the offspring. The total number of centroids must correspond to the number of clusters that are being looked for at that stage.

The representation of centroids that have been chosen facilitates this crossing operation because, in order to guarantee the creation of valid individuals it is only necessary to make sure that there are not repeated centroids. In case of repetition, a new centroid is chosen. Besides the aforementioned this representation does not require to manipulate the individual data for the crossing. It only requires to assign data to the nearest centroid and then recalculate the centroids to form the new generation.

The experimental tests showed that the centroid recalculation is a fundamental stage of this method because it improves substantially the results and the speed with which they are obtained. This recalculation may be interpreted as a local optimization that improves the quality of individuals before they are passed on to the next generation. To carry out mutation, we must consider the mutation probability that is established as an initial parameter of the program. In this way the values of a centroid that is chosen at random are aleatorily altered. When modifying some of the centroid's values, this moves through the solutions giving the opportunity of the production of new genetic material.

4 Results

The procedure followed for the cluster formation is centred in the adjustment of parameters to obtain the maximum values in the Calinski-Harabasz quality scale. After trying different parameter combinations, the quality measure reached a maximum of 48.4799 when forming four groups with a population of 20 individuals and crossing and mutation probabilities of 0.75 and 0.05 respectively.

Once the clusters have been formed, the problem is to create an ordinal risk scale for the groups. To this effect, in each group the proportion of dead and survivors was identified (death rate). If we accept that the use of technology avoids death, it is possible to affirm that the groups with the highest death rate correspond to the groups that presented a higher admission risk.

Figure 1 shows the groups in decreasing order according to the risk, with the dead number in each group. It is interesting to mention that when the number of groups increases or diminishes not only does the value of the quality measurement of the clustering decreases but also the difficulty to differentiate the risk based on death proportions increases.

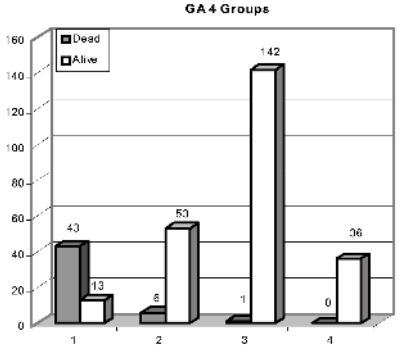

Fig. 1. Grouping obtained by the Genetic Algorithm (GA) for 4 groups. CH=48,4799.

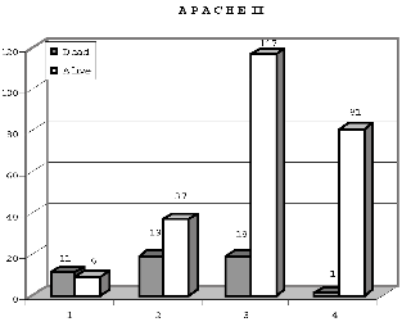

Fig. 2. Distribution of the dead in the groups formed by APACHE II.

It is important to assess what has been achieved with the clustering. To this effect we can resort to the APACHE method, which is the most commonly used method to determine risk at the moment of admission to the ICU. In the specific case of the units that have been mentioned, all the variables that conform the APACHE II index are registered. Using these variables it is possible to estimate the risk of each patient, using a logistic transformation that assigns a probability to the weighed score obtained from the APACHE II variables.

To make an equivalent comparison we formed four groups, gathering the individual risks in a decreasing order. Figure 2 shows the clustering made according to APACHE II index, indicating the number of dead and survivors in each cluster.

It is also interesting to compare the efficiency of the proposed *GA* method in relation to other traditional clustering methods. To carry out this comparison, a clustering was made using the k-means method assigning risk in the same manner as in the case of *GA*.

The presentation of results based on histograms is useful to identify the group risk, but does not offer an objective comparative parameter between the different methods. Given the fact that the comparison is based on the classification between dead and survivors, it is possible to measure the sensitivity and specificity to detect death as if each method were a binary death rate classifier, with the discrimination threshold taken from each of the generated clusters. Each binary classification generates a point (sensitivity, specificity). Using these, it is possible to draw an ROC curve and, in this way, the area below the ROC curve will be an objective comparative parameter. Figure 3 shows the ROC curves that measure the advantage of the death classification of APACHE II index (area equals

0.786); the k-means method (with a 0.888 area) and the proposed method that uses *GA* (with a 0.946 area).

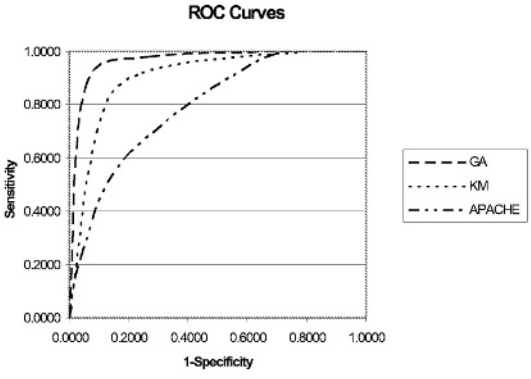

Fig. 3. ROC curves for GA, K-Means and APACHE.

As mentioned in the introduction, an important point in this design is the incorporation of information that contains the diagnosis hypotheses and therefore it would be interesting to assess the contribution of this information. In order to carry out this assessment, the 17 variables that represented the diagnosis hypotheses were withdrawn and the proposed method was applied again. The area below the ROC curve gave a 0.824 value.

5 Discussion

The histogram that appears in Figure 2 (APACHE II) shows the groups in a decreasing order according to their seriousness, However, in spite of this, the dead shows a more homogeneous distribution among the groups than the proposed model (Figure 1). This proves the difficulty that is found with the APACHE system to classify the patients' individual risks correctly (despite the fact that in this case the patients are grouped together). For this reason the APACHE system has been criticized because, although everybody recognizes its effectiveness to determine the number of patients that may die in a unit, it fails in the individual classification [18–19]. This limits its field to comparative studies between units and it does not offer guarantees when it is necessary to know the admission risk of a specific patient. But this is exactly what is needed to carry out an effectiveness analysis of the technologies that are used in a unit.

A more precise quantification of the differences between the *GA* grouping method and the APACHE system can be made comparing the ROC curve areas. Knowing that the perfect discriminator has an area that is equal to that of the unit, the *GA* grouping reaches 0.946, exceeding the ROC curve area reached by the APACHE II system by 0.161.

The influence of incorporating the *GA* to the method can be assessed when comparing the classification obtained through *GA* with the *k-means* classification. In this case we can see that the *GA* classification surpasses the curve obtained by *k-means* by 0.058. This is an indication that the sole idea of using simple grouping method produces an increase of 0.103 in the ROC curve area with respect to the classification obtained by the APACHE II system (36% of the gain is attributed to the *GA* and the remaining 64% corresponds to the grouping in an isolated way).

The experiment of deleting the variables that quantify diagnosis hypotheses shows that the inclusion of this information produces a decrease of 0.122 of the area, which corresponds to 75.8% of the gain obtained with the total of the variables proposed with respect to APACHE II.

From this analysis we can infer that the basic idea of using a non-supervised method such as clustering produces improvement with respect to the standard risk evaluation systems in medicine. But the highest values are achieved when we incorporate adequately the information contained in the diagnosis hypotheses and the use of *GA* to achieve a better clustering of patients by the risk.

6 Conclusions

Starting from the idea of using clustering analysis to build classifiers that may allow to classify patients according to their admission risk, this paper includes the diagnosis hypothesis quantification and the *GA* as the two main pillars that proved fundamental to achieve the results shown here.

We must point out the simplicity and capacity to be reproduced offered by diagnosis quantification, provided that we can count on the specialists to carry out the diagnosis classification. The main problem with this classification is the splitting of diagnoses into two degrees of acuteness: acute and hyperacute. This problem can be overcome through consensus meetings with groups of specialists.

From the application of *GA* to the clustering analysis, it is worth pointing out the simplicity of the representation and the crossing stage since the traditional applications for this problem present complex crossing operators [16–17], with high probability to create crippled descendants and loss of genetic material along the generations. The use of centroids in the chromosome formation allows to preserve more adequately the ancestors' best genetic material, circumscribing the random component exclusively to mutation.

A second stage in this research would be to increase the number of patients, trying to include all the existing pathologies in one ICU; to implement a classification system to carry out a prospective study that classifies the patients at the moment of admission and, later on, to assess the results achieved in the unit operation.

The extension of this method to other complex treatment systems such as Neonatal Intensive Care Units, Coronary Units or Multiple Trauma Units offers great possibilities since in these cases the pathologies are sometimes more restricted and it is possible to state the quantification of the diagnosis hypotheses even more accurately. But certainly the richest field for its application are the studies on Technology Assessment in

Health Care, where the admission risk assessment is fundamental to achieve technology effectiveness.

References

1. L.I. Iezzoni: An introduction to risk adjustment. Am J Med Qual 11 (1996) p. 8–11.
2. N. Marvin, M. Bower, J.E. Rowe, and AI Group, De Montfort University, Milton Keynes, UK,: An evolutionary approach to constructing prognostic models. Artif Intell Med, 15 (1999) 155–65.
3. R.W. Evans,: Health care technology and the inevitability of resource allocation and rationing decisions. JAMA 249 (1983) 2047–2053.
4. J.R. Le Gall, S. Lemeshow and F. Saulnier,: new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. JAMA 270 (1993) 2957–63.
5. W.A. Knaus, J.E. Zimmerman, D.P. Wagner, E.A. Draper and D.E. Lawrence,: APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 9 (1981) 591–7.
6. M. Shwartz, L.I. Iezzoni, M.A. Moskowitz, A.S. Ash and E. Sawitz,: The importance of comorbidities in explaining differences in patient costs. Med Care 34 (1996) 767–82.
7. R. Dybowski, P. Weller, R. Chang and V. Gant: Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet 347 (1996) 1146–50.
8. C.A. Peña-Reyes and M. Sipper,: Evolutionary computation in medicine: an overview. Artif Intell Med 19 (2000) 1–23.
9. B. Sierra and P. Larranaga,: Predicting survival in malignant skin melanoma using bayesian networks automatically induced by genetic algorithms. An empirical comparison between different approaches. Artif Intell Med 14 (1998) 215–30.
10. J.D. Horbar, L. Onstad and E. Wright,: Predicting mortality risk for infants weighting 501-1500 grams at birth: A National Institute of Health Neonatal Research Network report. Crit Care Med 21 (1993) 12–8.
11. W.A. Knaus, E.A. Draper, D.P. Wagner and J.E. Zimmerman,: APACHE II: a severity of disease classification system. Crit Care Med 13 (1985) 818–29.
12. M.C. Weinstein and H.V. Fineberg, Clinical Decision Analysis, W.B. Saunders Company, Philadelphia, (1980).
13. M. Aldenderfer and R. Blashfield, Cluster Analysis, Sage University Paper, California (1984).
14. D.E. Goldberg, Genetic Algorithms in Search, Optimization Machine Learning, Addison-Wesley, Reading (1989).
15. T. Calinski and J. Harabasz, A Dendrite Method For Cluster Analysis. Communications in Statistics 3 (1974) 1–27.
16. H. Ding, A.A. El-Keib and R.E. Smith, Optimal Clustering of Power Networks Using Genetic Algorithms, TVGA Report No. 92001, University of Alabama, Tuscaloosa, AL (1992).
17. E. Falkenauer, Genetic Algorithms and Grouping Problems, John Wiley & Sons (1999).
18. R.B. Becker and J.E. Zimmerman,: ICU scoring systems allow prediction of patient outcomes and comparison of ICU performance. Crit Care Clin 12 (1996) 503–14.
19. G. Thibault, Prognosis and clinical predictive models for critically ill patients, Brockton-West Roxbury Veterans Affairs Medical Center, West Roxbury, Massachusetts, (2000).