Applying Feature Selection Techniques to Neonatal Risk Indexes Analysis

Anália Lourenço[1], Ana Cristina Braga[2] and Orlando Belo[1]

[1] Departamento de Informática, Universidade do Minho, 4700-320 Braga, Portugal
{analia,obelo}@di.uminho.pt
[2] Departamento de Produção e Sistemas, Universidade do Minho, 4700-320 Braga, Portugal
acb@dps.uminho.pt

Abstract. Scoring systems that quantify neonatal mortality have an important role in health services research, planning and clinical auditing. They provide means to monitoring, in a more accurate and reliable way, the quality of care among and within hospitals. The classical analyses based on a simple comparison of mortality or the dealing with the newborns birthweight have proved to be insufficient. There are a large number of variables that influence the survival of newborns that must to be taken into account. From strictly physiological information through more subjective data, concerning medical care, there are many variables to attend to. Scoring systems try to embrace such elements, providing more reliable comparisons of the outcome. Notwithstanding, if a clinical score intends to gain widespread between clinicians, it must be simple and accurate and use routine data. In this paper, it is presented a neonatal mortality risk evaluation case study, pointing out data specificities and how different data preparation approaches (namely, feature selection) will affect the overall outcome.

Area: Machine Learning, Knowledge Discovery and Data Mining.
Keywords: Data Mining, Classification Models, Data Preparation, Feature Selection, Neonatal Death Risk Evaluation.

1. Introduction

The idea of acquiring knowledge through data analysis is not a new one. Data manipulation has always been a major priority within organizations, allowing the creation and spread of a wide range of information systems. However, these efforts have been clearly insufficient, incapable of coping with the increasing organizational needs. Usually information systems assist daily activities, supplying means to solve operational problems and tasks, but frequently they cannot ensure data quality. In fact, data quality is perhaps the most forgotten factor of all those that may be involved in data analysis. The increasing data volumes that are generated nowadays have highlighted its real value and strongly suggest a bound between data analysis approaches and decision support systems.

The main idea is to establish a path from raw operational data, through decision support systems processing and storage, to data preparation for data mining, analyzing its leading processes and consequently, their implications. A data volume is meaningless if there is no conviction on its contents or its volatility. Moreover, data should be structured according to analytical interests, selecting the best observation perspectives and measures in order to provide better decision support. Decision support systems databases offer the best elements for Knowledge Discovery, preparing most of the needed data and ensuring its quality, non-volatility and availability. This is a major step towards adequate and correct data mining, as well as to achieve good performance improvements in terms of accuracy and comprehensibility. Most of the preparation efforts that were normally taken are now prevented, allowing a redirection of efforts and resources to specific Data Mining preparation issues.

Before performing a data mining process, it is important to give the problem a closer look, analyzing its major variables and constraints. Decision-makers must realize that although an automatic knowledge extraction process is desirable, it is rarely achievable. Real-life problems usually bring along many elements that data mining algorithms are not ready to deal with (at least not in the manner users would like). It is too risky to perform certain steps without recurring to user interaction. It is true that decision support systems take care of a part of the problem, ensuring some mechanisms to the gathering, transformation and loading of operational data into pre-defined and rigid analytical structures. But, it is also true that the knowledge discovery process will eventually demand a more specific processing of the features, to select the most relevant transformations. Thus, it is crucial to understand the problem and to know the available data in order to apply the best techniques and to ensure that data is not misconstrued.

Over the next sections, there is a very generic description of the two main feature selection approaches – wrapper and filter – and an explanation of a case study concerning neonatal mortality risk evaluation for very low birthweight babies. The main implications of the problem are highlighted and then, feature selection and transformation techniques were used. Data mining was applied constructing a mortality prediction model based in scoring systems and there were constructed *Receiver Operating Curves* (ROCs) in order to analyze the predictive capacity of the indexes.

2. Feature Selection Approaches

Some of the issues raised during the definition of a new dataset are "subject independent" and, thus, may be modeled and treated in a systematic manner. There is always a main repository (or a certain number of collections) from which the data will be extracted. Although the selected variables are domain specific and play a very precise role in the problem, the way the process flows is more or less straightforward, according to the user's task directives. The real challenge lays in the manipulation of these elements, providing a consistent and meaningful set. From the aggregation of

values to the establishment of thresholds and discriminative classes, there are a large number of techniques that may come in hand.

Preparing data to be mined implies a good understanding of the problem and a certain care while choosing the features. There are many definitions of feature selection, but they all agree in one thing: it is imperative to reduce dataset dimensionality, according to some criteria, in order to improve data mining results in terms of performance, representativeness, accuracy and comprehensibility. The criteria might prevent accuracy significant decline, or ensure that the resulting class distribution is as close as possible to the original one [4]. In this framework, we will focus on accuracy variation analysis.

Two different approaches have emerged towards this search for the optimal subset of features of a given problem: the wrapper and the filter models. On one hand, if the focus is put in the minimization of the classifier error rate, i. e., aiming to achieve the highest predictive accuracy possible, then a wrapper model is used, implying that feature selection will target a particular mining algorithm appliance [10]. On the other hand, when an algorithm-independent approach is desired, filter models become the best candidates. They analyze the feature set, evaluating each variable "weight" over the sample, i. e., its discriminative power over the goal classes and pointing out the most relevant features. Therefore, decision-makers might study the dimensionality proposals, performing any data mining algorithm over them and comparing the outcomes.

Choosing the best feature selection approach to a problem will imply a previous dialog with the decision-maker. First, the goal of the analysis has to be cleared up along with the meaning of the different features and, if applicable, the way they are created (it is important to have an idea about the degree of subjectivity of each variable). Then, it has to be discussed possible feature discretization, in order to detect and establish only the strongest and most meaningful classes, which will allow algorithms to focus and improve the results comprehensibility. And finally, feature selection approaches can be debated having all the previous aspects in mind and also a notion of the sample's dimensionality and size.

3. Medical Mortality Risk Evaluation

During the last years, information systems have been imposing their influence over organizations. The need to satisfy today's expectations and technological developments has transformed these systems into the core of many organizational structures. The medicine systems do not escape from this reality. Highly specialized medical departments are classical candidates for new technologies and, within these, the critical care medicine stands out. Patients of intensive care units have severe or acute life-threatening diseases. These circumstances imply a great demand for special abilities of medical staff and doctors, as well as create great challenges for appropriate pharmacology and technology [18]. Obviously, today's amounts of data make manual approaches to the problem almost impossible.

Therefore, there has been an increasing implantation of information systems within these clinical departments, having in mind the gathering and processing of operational

data. Every element that might be relevant for decision-making must be preserved and "arranged" in order to have real value. Operational information systems ensure daily data gathering, but they are not intended to focus in analytical needs, leaving this job to decision support systems. Physicians specify the decision-making issues they are interested in covering and a specialized storage structure is assembled and populated accordingly. Then, it is finally possible to extract the maximum gain from the data, applying different processing and arranging the elements, following decision-makers' directives. In particular, intensive care units have large amounts of data generated by different tools, as well as, gathered while physicians and medical staff are inspecting each patient. Decision support systems not only ensure their right arranging and storing, but also allow certain data aggregations or pre-defined calculus, in order to generate valuable indexes.

The application of scoring systems is an example of a methodology to be taken. In the case of neonatal intensive care units, scoring systems quantify neonatal mortality which plays an important role in health services research, planning and clinical auditing. They provide means to monitor, in a more accurate and reliable way, the quality of care among and within hospitals. The classical analyses based on a simple comparison of mortality or dealing with the newborns birthweight solely have proved to be insufficient. There are a large number of variables that influence the survival of newborns that must to be taken into account. Scoring systems integrate many of those variables, providing more reliable comparisons of the outcome. [16] [17].

In particular, this study involves four scoring systems – the *Clinical Risk Index for Babies* (CRIB) [3], the *Neonatal Therapeutical Intervention Score System* (NTISS) [8], the *Score for Neonatal Acute Physiology* (SNAP) [5] [14] and the *Score for Neonatal Acute Physiology – Perinatal Extension* (SNAP-PE) – and analyses their predictive potential in a particular neonatal data sample. Table 1 presents the number of variables requested by each one of these indexes, as well as the "stand-by" time necessary to collect the correspondent sample.

The CRIB score involves fewer variables and may be calculated in half a day, while the other scores require a period of 24 hours to collect their wider set of data. Moreover, this index is based only in physiologic information, while the other three indexes include qualitative information (for instance, relating the medical staff care ministered to the newborn), more subordinated to incorrect or misconstrued impressions.

Table 1. The clinical risk indexes characterization**.**

	CRIB	NTISS	SNAP	SNAPPE
Number of variables	6	48	26	29
Time after childbirth taken to collect the sample (hours)	12	24	24	24

Therefore, it is important to study the "predictive weight" of each one of these scores, ensuring quick, reliable decision-making. As time goes by, complications, such as intracranial haemorrhage grades III and V, retinopathy of prematurity grades 3 and 4, and periventricular leucomalacia, tend to aggravate the baby's sequels and

put in risk his normal development and even his life [1]. In this sense, the first attempt is to use birthweight score to classify the newborns in risk, because it is clearly the "easiest" way out. Nevertheless, the information brought by these four scoring systems supplies "missing" reliability and must to be taken into account. There just has to be established a compromise between time and confidence (amount of gathered and processed data). In this study, each one of the scores is analyzed, figuring out its real predictive value and therefore, being able to select the optimal feature set that may comply with neonatal intensive care needs.

4. Neonatal Classification Models

The physicians' ability to process large amounts of data in an adequate manner in real-time is limited, forcing them often to assess the benefit of a therapy on the basis of rough estimates of the patient's complex medical condition. Intensive care units are an excellent example of this lack of capacity. When patients are severely ill the emergency on the time and quality of response becomes crucial. Doctors have to "absorb" all the available data as quick as possible, deciding the best therapeutical approach, minimizing eventual sequels.

In this scenario, decision support systems emerge having in mind the processing and analysis of vast volumes of data towards the selection of the best therapeutical approaches to a given problem. Knowledge Discovery in Databases, along with new s4tatistics approaches to data analysis, like ROCs, seem to provide an acceptable response to the problem, since they produce accurate and comprehensible knowledge having routine data as basis. The main idea is to use daily data, kept in medical operational information systems, building classification models that express the unit's performance, as well as detecting, representing and using the relevance of certain mortality risk indexes in the performance analysis (within or among hospitals). Here, the emphasis is put in the analysis of the different indexes predictive relevance towards classification models construction.

4.1 The case study

The case study is related to neonatal death risk for very low birthweight babies. In particular, the data concerns 162 newborns with birthweight under 1500g admitted in the Neonatology Unit of the Hospital Garcia da Orta in Portugal, between January 1992 and July 1995 [2] [1]. The various elements were collected retrospectively on the same set of newborns in order to perform the comparison of different risk indexes.

The sample contains many demographic items, as well as, references to the mother condition. Unfortunately, there is not available the elements used in the "calculus" of each risk index, i. e., it is only possible to study the factors as individuals. If those elements were available not only would be possible to analyze the relevance of the different indexes for the problem, but it would also be feasible the study of the prevalence or weight of each one of these variables (and eventual correlation) on the index and on the classes of the problem.

4.2 A ROC Approach

A ROC is created by plotting the true positive rate (sensitivity) on the vertical axis against the false positive rate (1-specificity) on the horizontal axis. A ROC is ordinarily plotted over a range of values, or *cutoffs*, for a given diagnostic test. For diagnostic tests that are expressed as continuous values these cutoffs represent threshold values at which a patient would be classified as positive [7].

The graphical representation of the ROC is of interest in the determination of appropriate diagnostic cutoffs for a given test. The area under the ROC is relevant for demonstrating the ability of the test to classify both true positives and true negatives, simultaneously, as a single measure.

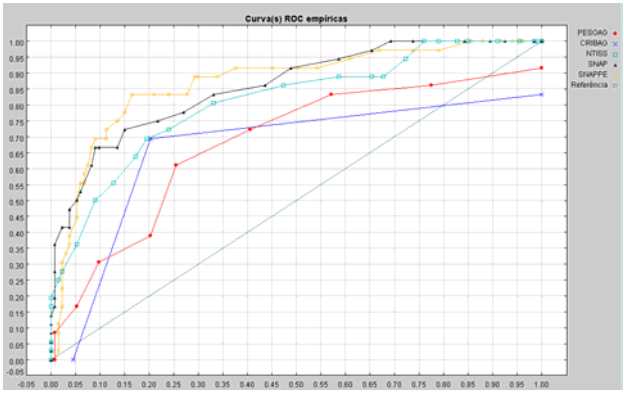

Fig. 1. Risk indexes ROCs.

In Figure 1, it is presented the ROCs concerning the five indexes. It is possible to observe that the weight index performs badly and the SNAP-PE index is the one that performs better. However, it cannot be forgotten that the CRIB index curve is constructed considering only three cutoffs. As each curve is constructed recurring to the trapezoid formula, in the case of the CRIB curve, a straight line will "connect" the points. This fact makes it appear that the CRIB index does not perform well as a neonatal mortality classifier, because the area beneath the correspondent curve (based in the trapezoid formula) is smaller than the areas concerning other indexes, underestimates its actual predictive capacity. However, there are other formulas that allow a more accurate calculation of the area beneath the ROC as their computation is independent of the number of cutoffs. In this study, the Wilcoxon-Mann-Whitney statistic has been used to compute the areas beneath the ROCs, being the standard errors associated to the areas obtained from the variance of the Wilcoxon statistic [6]. The results obtained with this statistic are presented in Table 2 and they support the previous observation. If the area beneath the ROCs is computed with the Wilcoxon statistic, paying no importance to the number of supporting points, the real predictive value of the indexes emerges. Thus, it is possible to observe that the CRIB index actually is quite a good classifier, presenting only a slight decrease in terms of area, comparing to the SNAP and SNAP-PE indexes.

Table 2. ROC curves and standard errors data.

Index	Area under ROC curve	Standard Error
CRIB	0.867	0.03
WEIGHT	0.768	0.05
SNAP	0.882	0.03
SNAP-PE	0.883	0.03
NTISS	0.845	0.04

4.3 A Data Mining Approach

The problem to be studied and treated can be stated as follows: what are the most relevant risk indexes towards the construction of a neonatal mortality prediction model. Having information concerning very low birthweight newborns, namely, the risk indexes data, the aim is to analyze their discriminative effect over the sample and to infer which are the ones that might induce a more accurate and comprehensible model. Moreover, it cannot be forgotten the nature of the problem. It is not just a matter of optimizing a model, but most of all it is run against time. Newborns are in stake and any time gain will make a difference in their survival options. So, it is important to study how well the CRIB index works. This index contains much more information than the weight, absorbing several physiological data, and spares half day waiting, comparing to the other indexes.

Before proceeding with feature selection activities, there were performed two discretization processes following decision-makers directives. In particular, the weight feature was discretized into equal width intervals and the CRIB values were grouped into three classes (1-low, 2-moderated and 3-high). This does not prejudice the subsequent feature selection, as these transformations are common practice among physicians and, most of the times, they base their studies in these classes rather than in the original intervals of values.

In order to study the effectiveness of different feature selection algorithms it was used the C4.5 as data mining algorithm. This choice was based in two main reasons: this is a well-known mining algorithm, which does not require further description [13]; it selects relevant features by itself in tree branching, so it can be used as a benchmark, as in [9] [11] [15], to verify the effects of the feature selection attributes.

In this study were used only filter algorithms towards feature selection. There was the intention of evaluating the risk indexes without having a special concern about the mining method. The focus was put on the spare of time, i. e., it was important to evaluate how well CRIB index would work by itself (sparing half a day) and which are the most relevant indexes for the problem. If the 24 hours period is indispensable it is important to at least restrict the number of calculations as NTISS, SNAP and SNAP-PE involve large numbers of variables. Therefore, the wrapper approach did not seemed appealing as it bound feature selection to a particular data mining algorithm performance. There were applied four basic ranking filter algorithms - the chi-square, the relief, the information gain and the information gain ratio algorithms -, whose output ranks are presented in Table 3 (order by descendent relevance).

Table 3. Feature ranking based in filter algorithms.

Algorithm	Attribute rank
Chi-square [12] [15]	{CRIB, SNAP-PE, SNAP, NTISS; WEIGHT}
Relief [9]	{WEIGHT, CRIB, SNAP-PE, SNAP, NTISS}
Information gain	{CRIB, SNAP-PE, SNAP, NTISS, WEIGHT}
Gain ratio	{SNAP-PE, CRIB, NTISS, SNAP, WEIGHT}

After analyzing these results and following the proposed aim, several attempts were made, in order to highlight the prediction capacity of each one of the indexes over the data sample. The true and false positives rates of each one (which are the "bricks" of the ROCs) and the correctly classified instances rate are given in Table 4.

Table 4. Accuracy of different prediction models.

Model	True Positives	False Positives	Correctly classified (%)
All features	0.694	0.016	91.9753
CRIB	0.694	0.048	89.5062
SNAP	0.667	0.0887	85.8025
SNAP-PE	0.833	0.143	85.1852
NTISS	0.25	0	83.3333
weight	0.167	0.008	80.8642

The CRIB index presents the same sensitivity of a classifier embracing all the features (indexes) and a decrease in the specificity. The SNAP has a similar sensitivity, but a higher specificity while its extension (SNAP-PE) presents an increase both in sensitivity and in specificity. The NTISS index is clearly too specific and has a poor sensitivity. These facts imply that CRIB index approximates fairly the situation, encouraging its use as a stand-alone neonatal mortality classifier and, therefore, sparing a half-day waiting, establishing a good base for decision-making.

5. Conclusions and Future Work

In this paper, a neonatal mortality risk case study was considered, aiming the selection of the most discriminative risk indexes and constructing a precise and comprehensible classification model. A ROC based analysis was performed over the five risk indexes chosen, as well as, several feature selection filter algorithms. Merging the various elements and confronting the precision of the constructed model for the different cases (sets of features), it was possible to come up with the conclusion that CRIB index is quite accurate predicting neonatal mortality. This fact

not only implies a significant reduction in the set dimensionality, but, most important of all, it reduces the waiting time. Thus, decision-makers, i. e., physicians and medical staff can judge the most suitable pharmacology and care approach to each specific case faster, having the newborn a better chance of survival.

Acknowledgements

The authors gratefully acknowledge to Professor Paulo Azevedo for its remarks and suggestions.

References

[1] A.C. Braga, P. Oliveira and A. Gomes, "Comparação entre Unidades de Cuidados Intensivos Neonatais", V Congresso Anual da Sociedade Portuguesa de Estatística, 1997. Curia, Portugal. (in portuguese).

[2] A.C. Braga, P. Oliveira and A. Gomes, "A Avaliação do Risco de Morte em Recém-Nascidos de Muito Baixo Peso: uma Comparação de Índices de Risco Baseada em Curvas ROC", IV Congresso Anual da Sociedade Portuguesa de Estatística, 1996. Funchal, Portugal. (in portuguese).

[3] F. Cockburn, R.W.I. Cooke, H.R: Gamsu et al., "The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units", *The Lancet, the International Neonatal Network*, 342: 193-198, 1993.

[4] M. Dash and H. Liu, "Feature Selection for Classification", Intelligent Data Analysis - An International Journal, Elsevier, Vol. 1, No. 3, 1997.

[5] J.E. Gray, D.K. Richardson, M.C. McCormick et al., "Neonatal acute physiology (SNAP) and risk of IVH". Pediatr Res, 31:249^{A}, 1992.

[6] J.A. Hanley and B.J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve", Radiology, 143:29--36, 1982.

[7] J.H. Holmes, "Discovering risk of disease with a learning classifier system", in T. Baeck, ed., Proceedings of the Seventh International Conference on Genetic Algorithms, 1997.

[8] R. Keene and D. Cullen, "Therapeutical Intervention Score System": Update 1983. Critical Care Medicine, bf 11(1), 1-3, 1985.

[9] K. Kira and L.A. Rendell, "The feature selection problem: Traditional methods and a new algorithm", in AAAI-92, Proceedings Ninth National Conference on Artificial Intelligence, p. 129-134. AAAI Press/The MIT Press, 1992.

[10] R. Kohavi and G. John, "Wrappers for Feature Subset Selection" (late draft), in Artificial Intelligence journal, special issue on relevance, Vol. 97, Nos 1-2, pp. 273-324.

[11] H. Liu and R. Setiono, "Feature Selection via Discretization", IEEE Trans Knowledge and Data Engineering 9:4, 642-645, 1997.

[12] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes", the 7th IEEE International Conference on Tools with Artificial Intelligence (TAI'95), Nov.1995, pp. 388-391. Washington D.C., USA.

[13] J.R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann, 1993.

[14] D.K. Richardson, J.E. Gray, M.C. McCormick et al., “Score for neonatal acute physiology: a physiology severity index for neonatal intensive care”, Pediatrics, 91- 617-23, 1993.
[15] R. Setiono and H. Liu, “Chi2: Feature selection and discretization of numeric attributes”, in Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence, 1995.
[16] W. Tarnow-Mordi, G. Parry, C. Gould and P. Fowlie, “CRIB and performance indicators for neonatal intensive care units”, Arch Dis Child Fetal Neonatal Ed 1996 Jan;74(1):F79-80.
[17] W. Tarnow-Mordi and G. Parry, “The CRIB score”, Lancet 1993 Nov 27;342(8883):1365.
[18] I.J. Timm, “Automatic Generation of Risk Classification for Decision Support in Critical Care”, in Bellazzi, R.; Zupan, Blaz (eds.): Workshop Notes on WS21 at the 13th Biennial European Conference on Artificial Intelligence (ECAI '98): Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP 98), p. 38-41, 1998. Brighton, United Kingdom.