Expert Systems
The Journal of Knowledge Engineering

Article

DOI: 10.1111/j.1468-0394.2009.00498.x

Extracting new patterns for cardiovascular disease prognosis

Luis Mena,[1] Jesus A. Gonzalez[2] and Gladys Maestre[3]
(1) Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico, and Faculty of Engineering, University of Zulia, Maracaibo, Venezuela
Email: lmena@inaoep.mx
(2) Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico
Email: jagonzalez@inaoep.mx
(3) Neurosciences Laboratory of the Institute for Biological Research and Institute for Cardiovascular Diseases, University of Zulia, Maracaibo, Venezuela, and Gertrude H. Sergievsky Center, Columbia University, New York, USA
Email: gem6@columbia.edu

Abstract: *Cardiovascular diseases constitute one of the main causes of mortality in the world, and machine learning has become a powerful tool for analysing medical data in the last few years. In this paper we present an interdisciplinary work based on an ambulatory blood pressure study and the development of a new classification algorithm named REMED. We focused on the discovery of new patterns for abnormal blood pressure variability as a possible cardiovascular risk factor. We compared our results with other classification algorithms based on Bayesian methods, decision trees, and rule induction techniques. In the comparison, REMED showed similar accuracy to these algorithms but it has the advantage of being superior in its capacity to classify sick people correctly. Therefore, our method could represent an innovative approach that might be useful in medical decision support for cardiovascular disease prognosis.*

Keywords: cardiovascular diseases, machine learning, blood pressure variability, classification, medical decision support, prognosis

1. Introduction

Since the 1950s, cardiovascular diseases have taken first place as a cause of mortality in developed countries. Projections carried out by the Department of Public Health of Massachusetts predict that for the year 2020, if this tendency continues, these diseases could be the main cause of death in the world. This would be the first time that this has happened in the history of the modern world.

The main cardiovascular risk factors are related to a high level of cholesterol in the blood and hypertension; however, in the last four decades cardiology has been one of the branches of medicine that has advanced a lot in the development of techniques to discover new causes of risk. Ambulatory blood pressure monitoring (ABPM) (Mancia, 1990) has been one of the most effective techniques; it consists of measuring the blood pressure (BP) with portable and automatic devices that allow the registration of

BP during a programmable period of time, generally 24 hours. The fundamental contribution of this technique with regard to the traditional measure is that BP readings are obtained outside the hospital environment. This helps to obtain more representative samples and also avoids the alert reaction or phenomenon known as *white-coat hypertension* (Pickering *et al.*, 1988). This phenomenon makes reference to those subjects that present a high BP during the medical appointment and normal BP outside this environment.

In addition, continuous monitoring of BP has allowed the BP variability (BPV) to be classified into global and circadian BPV, and abnormal ABPM registrations of both are considered potential cardiovascular risk factors. Global BPV is the variability present in the set of measures registered by the ABPM. Generally the ABPM readings are programmed periodically considering a 15 minute interval between readings; this is done with the objective of getting a representative sample of measurements and to avoid dispersion of the results (Di Rienzo *et al.*, 1983). Recent studies have suggested that an increase in global BPV is associated with an increase in subsequent cardiovascular events/complications (Frattola *et al.*, 1993; Kikuya *et al.*, 2000; Sander *et al.*, 2000; Mena *et al.*, 2005).

The circadian BPV tries to measure the BP changes between day (activity period) and night (rest period). Usually, a BP descent is detected at night. The presence or absence of this night BP descent allows the studied subjects to be classified as *dippers* and *non-dippers*. In an informal way, subjects are considered *non-dippers* when their values of night BP do not decrease by 10% with regard to their values of day BP (Verdecchia *et al.*, 1990). Several studies seem to associate the absence of BP decrement with the onset of some cardiovascular diseases (Palatini *et al.*, 1992; Rizzoni *et al.*, 1992; Shimada *et al.*, 1992).

In recent years the need to automatically extract knowledge from databases has increased. Thus, the use of machine learning techniques (Witten & Frank, 2005) to discover valid, novel, interesting and comprehensible patterns could be the key to success in either the business or scientific environment. We find a clear example of this in the medical diagnosis/prognosis domain (Kukar & Groselj, 2000; Breault *et al.*, 2002; Podgorelec *et al.*, 2002), where identifying patterns that help to predict the incidence of any kind of disease may represent the opportunity to react on time to avoid, delay or diminish the consequences of exposure of predispositions.

Generally, prediction tasks are solved by applying supervised classification techniques; however, we have to consider some additional challenges associated with the application of machine learning to the medical prognosis domain. One of the most important problems is the selection of relevant attributes. These attributes are known in medical prognosis as risk factors and are classified as changeable (e.g. blood pressure, cholesterol etc.) and non-changeable (e.g. age, sex etc.). According to this, if we consider a non-changeable attribute such as age (a good attribute for classification tasks) it might not be useful as a target to modify disease evolution because no medical treatment exists to modify the age of a patient. Therefore, we should focus on changeable attributes, and this could make the classification task harder.

Another important aspect to consider is the need to obtain classifiers with comprehensible patterns to provide the medical staff with a novel point of view about the given problem. Usually this is done using symbolic learning methods (e.g. decision trees and rules), because it is possible to explain the decisions in an easy way for humans to understand. For this reason, connectionist methods such as neural networks could be excluded, since these almost always behave as black boxes. However, the use of a symbolic learning method generally sacrifices *accuracy* in prediction in order to obtain a more understandable model.

A third problem that hinders obtaining high overall performance is that in general medical data sets exhibit an imbalanced class distribution (Chawla *et al.*, 2004), where there exists a majority or negative class of healthy people (normal data) and a minority or positive class of sick people (the important class), which

generally has the highest cost of erroneous classification. Therefore, the performance of standard classifiers (e.g. C4.5, *k*-nearest neighbour and naive Bayes) tend to be overwhelmed by examples of the majority class and ignore the minority class examples; results have an acceptable performance in terms of *accuracy* and *specificity* (healthy subjects diagnosed correctly), but a low performance in terms of *sensitivity* (sick subjects diagnosed correctly).

Another problem refers to the fact that medical data are often obtained from longitudinal studies that consist of observing the incidence of a disease in a group of subjects during a specific period of time. At the end of the study, a binary classification is done and every subject is classified as sick (positive class) or healthy (negative class), depending on whether the studied disease had developed or not. However, the fact that these studies were designed to culminate at a certain time might make the classifiers' task harder, because a subject that presented clear risk factors during the period of study but whose death was not caused by the studied disease (e.g. died in an accident) or who did not present the disease by the end of the study (it could appear just after the end of the study) is classified as healthy (considered as class label noise), and both situations tend to confuse the classifiers.

Finally, we mention some features that a machine learning algorithm should have to satisfactorily solve medical diagnosis/prognosis tasks. Besides creating a classifier that achieves a good overall performance and provides medical staff with comprehensible prognostic knowledge, it is necessary to have the ability to support decisions and to reduce the number of tests necessary to obtain a reliable prognosis (Kononenko, 2001; Bosnić & Kononenko, 2008). What we mean by achieving a good overall performance and the comprehensibility of the prognostic knowledge was described earlier. The ability to support decisions refers to the fact that it is preferable to accompany the predictions with a measure of reliability, e.g. the probability of an example belonging to a class, which could provide medical staff with enough trust to use the new prognostic knowledge in practice. In addition, it is desirable to have a classifier that is able to predict reliably using a small amount of data about the patients, because the collection of the data is often expensive, time consuming and harmful for them (Kononenko, 2001).

In this paper we propose a machine learning method for medical decision support in the prognosis of cardiovascular diseases, which tries to solve the previously exposed disadvantages. This work is based on an ABPM study and the development of a new classification algorithm named REMED. We focused on the discovery of new patterns for the abnormal BPV as possible cardiovascular risk factors. We compared our results with other classification algorithms such as Bayesian methods, decision trees and rule induction techniques. In the comparison, REMED showed similar *accuracy* to these algorithms but with the advantage that it is superior in its capacity to predict sick people correctly.

2. Methods

2.1. Ambulatory blood pressure monitoring

All subjects underwent a 24 hour ABPM with a fully automatic device (SpaceLabs 90207) that met the criteria of the Association for the Advancement of Medical Instrumentation (Imai *et al.*, 1990). Readings were obtained every 15 minutes during the day period (06:00–22:59 h) and every 30 minutes for the night period (23:00–05:59 h). Systolic BP (SBP) values greater than 260 mmHg or lower than 70 mmHg as well as diastolic BP (DBP) readings greater than 150 mmHg or lower than 40 mmHg were automatically discarded by the ABPM device (as non-valid measures).

2.2. Representation of the cardiovascular risk factors

The cardiovascular risk factors studied in this research were obtained from analysis of the subject's ABPM registrations. We analysed abnormal BPV (global and circadian) and hypertension (a well-known cardiovascular risk

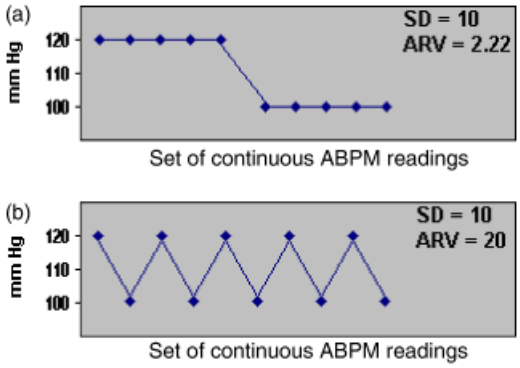

Figure 1: *Variability for two distinct blood pressure signals, but with the same values. ARV, average real variability; SD, standard deviation.*

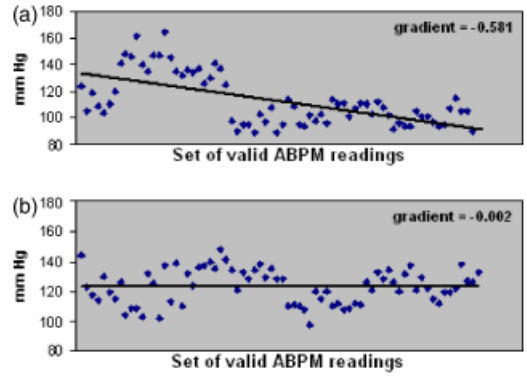

Figure 2: *Linear approach for blood pressure readings of (a) a dipper and (b) a non-dipper subject.*

factor). To evaluate the presence of hypertension we calculated the mean of the SBP and the DBP from the valid readings of the ABPM in a period of 24 hours. In addition, in order to analyse the global and circadian BPV we proposed two new measures.

2.2.1. Global BPV Most of the previous works that have studied the BPV as a cardiovascular risk factor have done it using the *standard deviation* (SD) as a variability index (Frattola *et al.*, 1993; Kikuya *et al.*, 2000; Sander *et al.*, 2000). However, this is a statistical measure that only reflects the dispersion around the central value (the mean); thus it does not account for the order in which the measurements were registered, as can be seen in Figure 1. Therefore, we proposed the *average real variability* (ARV) (Mena *et al.*, 2005) as the variability index. The ARV consists of calculating the arithmetic average of the differences in absolute value of the BP continuous measures.

$$\mathrm{ARV} = \frac{1}{n-1}\sum_{i=1}^{n-1}|x_{i+1} - x_i| \qquad (1)$$

where n denotes the total number of valid readings and x_i denotes the BP reading obtained at time i.

2.2.2. Circadian BPV In order to identify the *non-dipper* subjects from ABPM studies, we established the time periods of measurement (day/night) and then we contrasted the BP means for every period. Generally, the time interval used goes from 06:00 to 22:59 h during the day and from 23:00 to 05:59 h for night time. However, we consider that it is not practical to establish a strict cut to generalize the activity and rest periods of the subjects and therefore we propose to estimate for each subject the gradient of the straight line that best fits the values obtained with the ABPM during 24 hours (Figure 2). Our hypothesis is that those subjects with a smaller decrement in this gradient could be *non-dippers*.

$$\mathrm{gradient} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2)$$

where n denotes the total number of valid readings from 06:00 to 05:59 h, x_i denotes the BP reading obtained at time i, $y_i = i$ for $i = 1, \ldots, n$, $\bar{x}$ denotes the mean of x_i and $\bar{y}$ denotes the mean of y_i.

2.3. Algorithm

The complete symbolic classification algorithm has been called REMED (rules extraction for medical diagnosis) (Mena & Gonzalez, 2006). The REMED algorithm is composed of three basic procedures: (1) a procedure supported by the simple *logistic regression* model (LRM) (Hosmer & Lemeshow, 2000) for the selection

of the attributes, (2) a second procedure for the selection of initial partitions, and finally (3) the procedure responsible for the construction of classification rules.

2.3.1. Attribute selection procedure The main task of the first part of the algorithm (Figure 3) considers the selection of the best combination of attributes to build precise prediction rules. For this reason we used the simple LRM. This allows us to quantify the risk of suffering the studied disease with respect to the increase or decrease in the value of a specific attribute. We take advantage of the simple LRM in our algorithm because it uses a probabilistic metric called the *odds ratio* (OR) (Zheng *et al*., 2004), which allows us to determine whether there exists any type of association between the considered attribute and the studied disease. Thus, an OR equal to 1 indicates a non-association, an OR greater than 1 indicates a positive association (if the value of the attribute increases then the risk of suffering the disease also increases) and an OR smaller than 1 indicates a negative association (if the value of the attribute decreases then the risk of suffering the disease increases). Therefore, depending on the type of association established (positive or negative) with the OR metric, we can determine the syntax with which each attribute's partition will appear in our rules system.

```
Attributes Selection (examples, attributes)
final_attributes ← ∅
confidence_level ← 1-α  // > 99%
ε ← 1/10^k  // convergence level
for x ∈ attributes do
  e.x [ ] ← {examples of each attribute x}
  // p = p-value, OR = odds ratio
  p,OR ← Logistic_Regression (e.x [ ], ε)
  if p < ( 1 − confidence_level ) then
      final_attributes ← U x , OR
  end-if
end-for
```

Figure 3: *Pseudocode for the attribute selection process.*

However, the establishment of a positive or negative association between the risk of suffering the disease and an attribute is not enough. It is necessary to determine if this association is statistically significant for a certain confidence level. To achieve this, we always use high confidence levels ($>99\%$) to select attributes that are strongly associated with the risk of suffering the disease, and thus we can guarantee the construction of more precise rules. At this time, we only consider continuous attributes. This is because in the clinical environment discrete attributes are usually binary (e.g. smoker and non-smoker) and their association with a certain disease is almost always well known; therefore, continuous attributes have a higher degree of uncertainty than discrete attributes.

2.3.2. Initial partitions selection procedure Partitions are a set of excluding and exhaustive conditions used to build a rule. These conditions classify all the examples (exhaustive) and each example is assigned to only one class (excluding). The second part of the algorithm (Figure 4) finds the initial partitions, trying to maximize the resulting *sensitivity* (the true positive rate) as much

```
Initial Partitions ( examples, final_attributes)
m ← Number (final_attributes)
for i ← 1 ... m do
  e [ ] ← { sorted examples of the attribute i}
  partitions[ i ] ← Average (e [ ])
  pointer ← Position (e [ ], partitions[ i ] )
  k ← pointer
  while e_k.class ≠ 1  // seeking next positive
    if OR[ i ] > 1 then
      k ← k + 1  // positive association
    else
      k ← k − 1  // negative association
    end-if
  end-while
  if pointer ≠ k then
    if OR[ i ] > 1 then
      partitions[ i ] ← ( e_k + e_{k−1} ) / 2
      // positive association
    else
      partitions[ i ] ← ( e_k + e_{k+1} ) / 2
      // negative association
    end-if
  end-if
end-for
```

Figure 4: *Pseudocode to determine the initial partitions.*

as possible, without considerably decreasing *specificity* (the true negative rate) and maintaining an acceptable *accuracy*.

The procedure that REMED uses to select the initial partitions comes from the fact that, if an attribute x has been associated in a statistically significant form with the studied disease, then its mean $\bar{x}$ (mean of the n values of the attribute) is a good candidate for an initial partition of the attribute. This is because a large number of n independent values of an attribute x will tend to be normally distributed (by the central limit theorem); therefore, once a statistically significant association (positive or negative) is established between x and the studied disease, a single threshold above (positive association) or under (negative association) $\bar{x}$ will be a partition that indicates an increase of the risk of suffering the disease.

Then, we sort the examples by the attribute's value and we search for the next positive example in the direction of the established association according to the OR metric from the initial partition of each attribute ($\bar{x}_i$). Later, we calculate a new partition by computing the average between the value of the selected example and the value of its predecessor or successor. This displacement is carried out only once for each attribute, because another displacement to calculate a new partition would include at least a positive example at the opposite side of the threshold, and this could decrease the risk of suffering the disease in the new partition (loss of *sensitivity*).

Figure 5 shows an example that illustrates this procedure. We assume that a positive association between the studied disease and a continuous attribute such as the SBP was previously established using the simple LRM. Then, we select SBP $\geq$ 132.71 as our initial partition (the mean of the n SBP examples). After this we move the partition to the next example with class = 1 (example 179 in Figure 5). It is important to mention that, since the number of examples belonging to the negative class is a lot larger than that of the positive class (because of the class imbalance), there is a high probability of finding negative examples between the initial partition and the next positive example to make a displacement (jumping negative examples); therefore, the negative class imbalance is a required input assumption to apply REMED efficiently. Finally, we establish the new partition calculating the average for attribute SBP using the values of examples 178

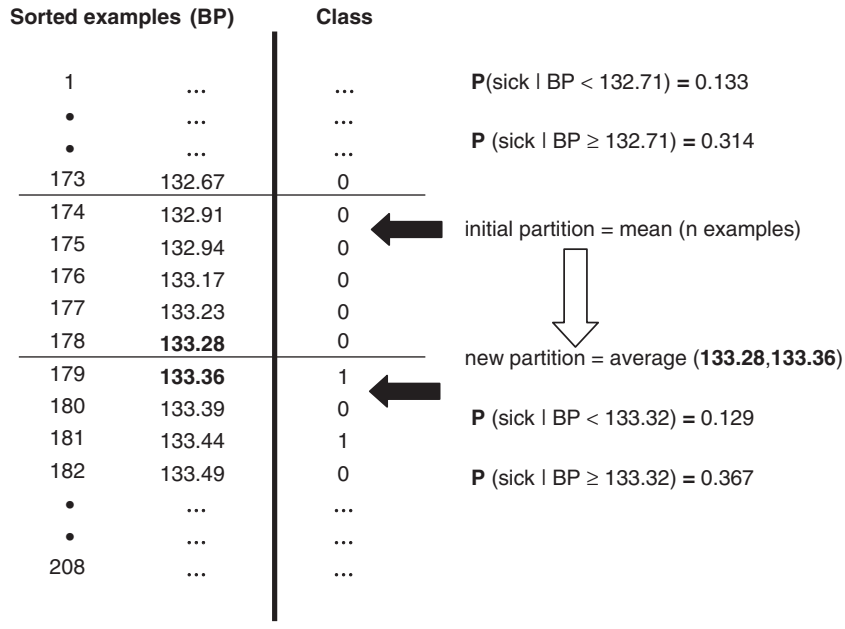

Figure 5: *Procedure to determine the initial partitions.*

and 179 (SBP $\geq$133.32). The goal of this strategy consists of increasing the risk of suffering the disease above this partition. For this reason we do not make a new displacement to search the next positive example, because this possible new partition calculated with the values of examples 180 and 181 (SBP $\geq$133.42) decreases the risk of suffering the disease above the threshold ($p = 0.357$) and increases again this risk under this threshold ($p = 0.133$).

2.3.3. Rules construction procedure Once we obtain the initial partitions for each of the m selected attributes, we build a simple system of rules which contains m conditions (one for each selected attribute j), in the following way:

if 1 $\langle$relation$\rangle$ p_1 **and** j $\langle$relation$\rangle$ p_j **and** ... **and** m $\langle$relation$\rangle$ p_m **then** $class = 1$
else $class = 0$

where $\langle$relation$\rangle$ is either $\geq$ or $\leq$ depending on whether j is positively or negatively associated with the positive class through p_j (partition for attribute j).

We make a first classification with this rules system. Then, we try to improve the performance of our rules system by increasing or decreasing the threshold of each partition as much as possible. For this we apply the *bisection* method (Burden & Faires, 2000) to calculate possible new partitions starting with the current partition of each attribute and the maximum or minimum value for this attribute in the examples. We build a temporal system of rules changing the current partition value for the new partition value and classify the examples again. We only keep a new partition if it decreases the number of false positives (FP) (healthy subjects diagnosed incorrectly) but does not decrease the true positive rate (sick subjects diagnosed correctly). This step is repeated for each attribute until we overcome the established convergence level for the bisection method or the current system of rules is not able to decrease the FP rate. This is done with the procedure shown in Figure 6. In this figure we grouped sets of instructions in sections identified with letters from A to E, which are described below.

(A) We build an initial system of rules from the set of initial partitions. Then we make a first classification and save the results. We also store the number of sick subjects classified correctly in k_1 and the total number of sick subjects predicted by the initial system of rules in k_2.

(B) Then, we begin an iterative process (1, ..., m) to try to improve the predictive value of each of the partitions. We estimate a new partition for attribute i by averaging its initial partition with the maximum or minimum value of the examples for this attribute (depending on the type of established association). With the goal of evaluating the performance of the new partition, we make a copy of the initial partitions in the *copy_partitions* [ ] array.

(C) We build a new system of rules by changing the current partition of attribute i by the new partition and then we classify the examples again. We store the number of sick subjects classified correctly in k_3 and the total of sick subjects predicted by this rules system in k_4.

(D) We then evaluate the results obtained with the new classification. First, we verify if the number of sick subjects classified correctly decreased ($k_3 < k_1$); if this happens we set the current partition as the maximum benchmark to calculate a new partition. Otherwise we verify if the new classification decreased the number of healthy subjects diagnosed incorrectly ($k_4 < k_2$); if this happens we store the total number of sick subjects predicted by the current system of rules in k_5 and establish it as the minimum benchmark for the current partition. We continue estimating new partitions for attribute i with the bisection method until a stop criterion is met. The stop criterion is when the difference in absolute value between the maximum and minimum benchmarks does not exceed the established convergence level for the

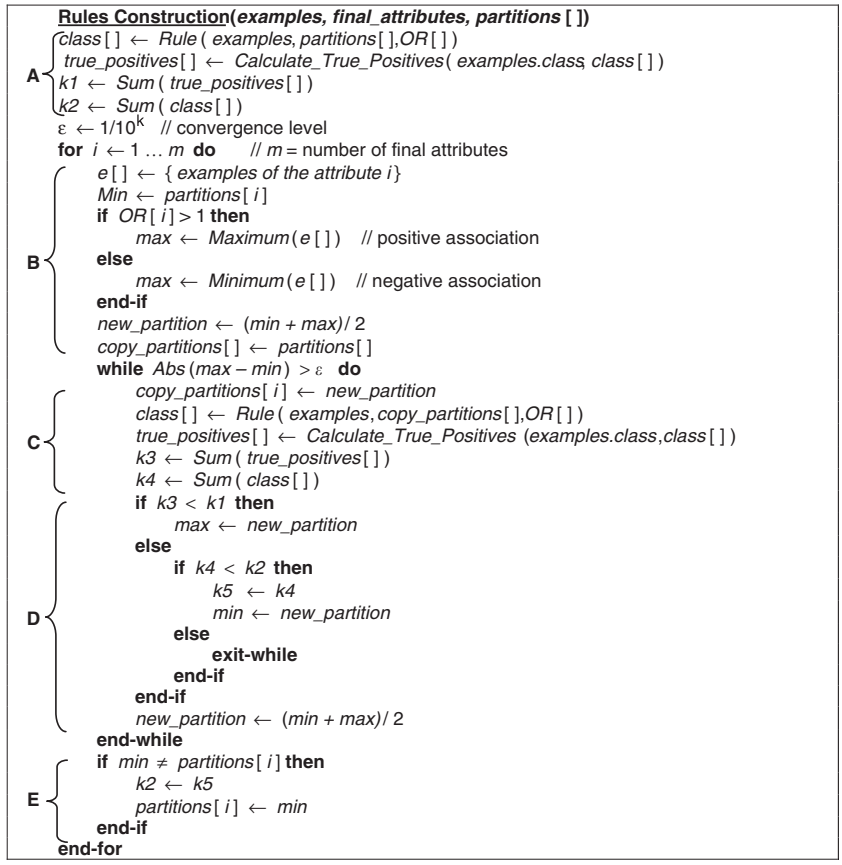

Figure 6: *Pseudocode for the construction of rules.*

bisection method, or the current system of rules is not able to decrease the number of healthy subjects classified incorrectly.

(E) If the new partition for attribute i improves the predictive values, it is included in the set of final partitions. Then, the total number of sick subjects predicted by the current rule is upgraded ($k_2 \leftarrow k_5$); this process is repeated for the m attributes.

2.4. Cardiovascular events under consideration

The cardiovascular events considered were coronary artery diseases, stroke and congestive heart failure. Coronary artery disease can be defined by any of the following: myocardial infarction diagnosed on the basis of at least two of three standard criteria (typical chest pain, electrocardiographic QRS changes and positive ischaemia serum markers) or angina pectoris defined by chest pain, cardiac catheterism showing haemodynamic significant obstructions or revascularization procedures. Stroke was diagnosed on the basis of rapid onset of localizing neurological deficit lasting $\geq$24 hours in the absence of any other disease (or lasting $<$24 hours for transient ischaemic attack). Congestive heart failure was diagnosed using the McKee criteria (McKee *et al.*, 1971; Ritchie *et al.*, 1999).

2.5. Data set

The data set was obtained from a longitudinal study named the Maracaibo Aging Study (Maestre *et al.*, 2002) conducted from October 1998 to June 2001 (2.83 years) by the Institute

for Cardiovascular Diseases of the University of Zulia, in Maracaibo, Venezuela. The study included all subjects with 70% or higher valid ABPM measurements, and without important concomitant diseases. The final data set consisted of 312 subjects ≥ 55 years, mean age 66.9 years, 63% women. The changeable continuous attributes considered were SBP, DBP (both calculated with the mean of ABPM readings), systolic global variability (SGV), diastolic global variability (DGV) (both calculated with the ARV index), and systolic circadian variability (SCV) (calculated with the gradient of the straight line). All these attributes were obtained from the valid readings of the ABPM in a period of 24 hours. The follow-up period for each individual had a mean value of 1.86 years and ended with a non-fatal cardiovascular event or with the arrival of the termination date of the study. At the end of the ABPM study the subjects registered 55 cardiovascular events (17.63%). Informed consent was obtained from every participant and the study protocol was approved by the ethics committee of the Institute.

2.6. *Performance comparison*

In order to compare the predictive capacity of the REMED algorithm we made a performance comparison with a set of 23 classification algorithms implemented in the WEKA framework (Witten & Frank, 2000). The algorithms considered were four Bayesian techniques with different numeric estimator precision values, seven rule learners and 10 decision trees with different methods to calculate information gain, and different pruning techniques. The final selection of methods to use in the comparison was done according to the best *accuracy* shown for our data in terms of the area under the receiver operating characteristics curve (AUC). These chosen algorithms were naive Bayes (Bayesian techniques) (John & Langley, 1995), OneR (rule learner) (Holte, 1993), ADtree (decision trees) (Freund & Mason, 1999) and REMED. In order to improve the performance of all the classifiers we used the simple LRM to select only attributes associated with the positive class with a confidence level $>99\%$ ($p < 0.01$). In all cases we used the 10-fold cross-validation technique to avoid overfitting (Witten & Frank, 2005). Finally, we evaluated the performance of each classification algorithm in terms of *accuracy*, *sensitivity*, *specificity*, AUC calculated with the binormal model (Hanley, 1996), and positive and negative predictive values. We followed the methodology presented by Mitchell (1997) to determine the level of significance by which REMED outperformed the rest of the classification algorithms; we used the popular two-tailed paired t test method with a confidence level of 95%.

3. Results

3.1. *Final selection of attributes*

Using the simple LRM we found statistical significance ($p < 0.01$) for all the attributes corresponding to SBP, but none of the attributes related to the DBP showed statistical significance for the selected confidence level (99%). Therefore, only the systolic attributes were selected for the performance comparison. Table 1 displays the results obtained.

3.2. *Rules system*

Table 2 shows the rules systems generated by the symbolic algorithms: OneR, ADtree and REMED. As we mentioned before (Section 1), in order to satisfactorily solve medical prognosis tasks, it is necessary to provide the medical staff with more comprehensible models. This is why it is important to further analyse and compare the

Table 1: *LRM results for the considered attributes*

Attribute	p
Systolic blood pressure (SBP)	0.0023^{**}
Systolic global variability (SGV)	0.0006^{**}
Systolic circadian variability (SCV)	0.0038^{**}
Diastolic blood pressure (DBP)	0.0417^{*}
Diastolic global variability (DGV)	0.3106

$^{**}p < 0.01$; $^{*}p < 0.05$.

Table 2: *Classification rules generated by OneR, ADtree and REMED*

Algorithm	*Rules*
OneR	If SBP $\geq$ 149.594 and SBP $<$ 153.07 then sick, else healthy If SBP $\geq$ 189.789 then sick, else healthy
ADtree	If SBP $\geq$ 149.594 and SBP $<$ 153.07 then sick, else healthy If SBP $\geq$ 123.596 and SBP $<$ 125.094 and SGV $\geq$ 8.826 and SCV ≥ -0.438 then sick, else healthy If SBP $\geq$ 123.596 and SBP $<$ 125.094 and SGV $\geq$ 8.826 and SCV < -0.705 then sick, else healthy
REMED	If SBP $\geq$ 142.18 and SGV $\geq$ 9.26 and SCV ≥ -0.40 then sick, else healthy

Table 3: *Performance comparison between the classification algorithms*

					Predictive values	
Algorithm	*Accuracy*	*Sensitivity*	*Specificity*	*AUC*	*Positive*	*Negative*
REMED	81.12	31.67	91.43	62.12	45.00	86.4
Naive Bayes	79.52	19.33	92.23	57.55	35.48	84.34
ADtree	77.95	18	90.77	56.93	29.41	83.81
OneR	81.42	16	95.37	56.28	42.86	84.19

performance (comprehensibility and medical validity) of the selected symbolic algorithms.

3.3. Discussion

The results in Table 3 show the average of accuracy, sensitivity, specificity and AUC for each algorithm over 10 runs. We also calculated the predictive positive (*precision*) and negative values from the final confusion matrix. These results indicate that OneR and REMED were the algorithms that reached the best performance in terms of classification accuracy, sensitivity, specificity and AUC. We can also appreciate that REMED maintains similar levels of *accuracy* ($>80\%$) and *specificity* ($>90\%$), but it is clearly superior to the rest of the algorithms in terms of *sensitivity*, diagnosing correctly more than 30% of the sick subjects, while the second best performance in terms of sensitivity (naive Bayes) does not reach 20%. We focused on achieving better *sensitivity* results over *specificity* because in the medical diagnosis/prognosis domain the misclassification cost of false negatives (FN) (sick subjects diagnosed incorrectly) is higher than that of FP, since in the case of a FP more specific medical tests could discover the error but a FN could cause a life-threatening condition that, depending on the kind of disease, could lead to death (Weiss, 2004).

On the other hand, REMED showed the best predictive values (positive and negative). That is, even when REMED predicted rather more sick subjects, it kept the best positive predictive value (45%); therefore, REMED was more precise in its positive predictions. In addition, REMED obtained the best performance in terms of AUC, which can better represent the overall performance of a classifier for imbalanced data sets than when only the accuracy measure is used (Chawla *et al*., 2004).

Tables 4 and 5 show the results of the two-tailed t test with a confidence level of 95% for the accuracy and AUC comparisons respectively. We can appreciate from Table 4 that there were no statistically significant differences ($p < 0.05$) in the accuracy comparison; however, Table 5 shows that in terms of AUC REMED significantly outperformed ($p < 0.05$) OneR (the best performance in terms of *accuracy* and *specificity*) and naive Bayes (the second best

Table 4: *Results of the two-tailed t test (95%) for the accuracy comparison*

Comparison	*Difference (mean x – y)*	*Two-tailed p value*	*Statistical significance*
REMED – naive Bayes	1.604	0.2749	Not significant
REMED – OneR	–0.302	0.8495	Not significant
REMED – ADtree	3.176	0.1198	Not significant

Table 5: *Results of the two-tailed t test (95%) for the AUC comparison*

Comparison	*Difference (mean x – y)*	*Two-tailed p value*	*Statistical significance*
REMED – naive Bayes	4.567	0.0483	Significant
REMED – OneR	5.845	0.0191	Significant
REMED – ADtree	5.191	0.0759	Not significant

performance in terms of *sensitivity* and AUC), and although it showed superior results to ADtree, the difference was not statistically significant ($p = 0.0759$) at the 95% level of confidence. These differences can be better appreciated in Figure 7, which shows the smooth receiver operating characteristics curves calculated with the binormal model; according to other research work in the area (Hanley, 1996), this is a method that behaves well empirically in a wide variety of situations.

With respect to the performance comparison among the symbolic classifiers, although OneR was slightly better in *accuracy* and *specificity* than REMED, our algorithm presented a set of more complete and comprehensible rules, because in the rules system of OneR two of the selected attributes ($p < 0.01$) were excluded (SGV and SCV), while the rules system inferred from the decision tree obtained by ADtree was more complex (with a larger number of rules). In addition, REMED does not produce rules with

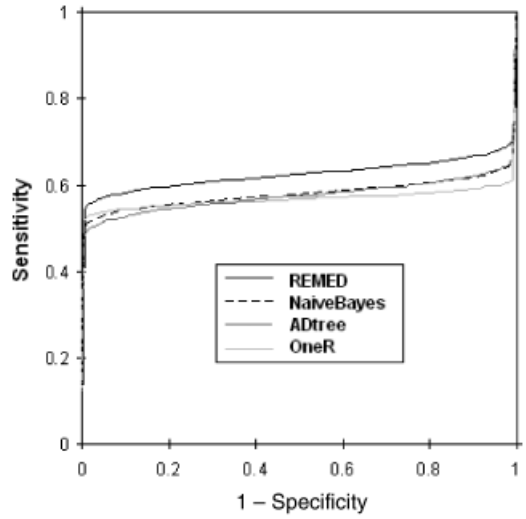

Figure 7: *Smooth receiver operating characteristics curves calculated with the binormal model.*

enclosed intervals (e.g. $a \leq x \leq b$), in contrast to the rest of the symbolic classifiers. This is important because it could represent a disadvantage for medical prognosis, since in general the risk of developing a disease is directly proportional to the increase or decrease of a risk factor (i.e. hyperthyroid and hypothyroid disease).

Finally, we can appreciate in the rules system of REMED that one of the rule antecedents ($SBP \geq 142.18$) is the closest to the clinical standard ($SBP \geq 140$) stated by the American Heart Association to diagnose systolic hypertension (a well-known cardiovascular risk factor). Therefore, for this specific attribute, REMED showed a rules system with greater medical validity than the rest of the symbolic classifiers.

4. Conclusions and future work

The main conclusion is that new patterns found ($SGV \geq 9.26$ and $SCV \geq -0.40$) for the systolic BPV (global and circadian) through REMED and the proposed measures (ARV index and gradient of the straight line) could represent potential risk factors to support medical decisions in cardiovascular disease prognosis.

REMED could be a competitive approach to solve part of the problems associated with the

application of machine learning to medical prognostic domains, because it possesses the desired features to solve medical prognosis tasks: (1) good overall performance, because REMED reached a good overall performance in terms of accuracy, sensitivity, specificity, AUC, and predictive values (positives and negatives), (2) the comprehensibility of the prognostic knowledge, because REMED always generated rules systems with a larger degree of abstraction than the rest of the symbolic classifiers, (3) the ability to support decisions, because the fact that the rules systems of REMED are always supported by an attribute selection with high confidence levels ($>99\%$) could provide the medical staff with enough trust to use these predictive rules in practice, and (4) the ability of the algorithm to reduce the number of medical tests necessary to obtain a reliable prognosis, because REMED uses the simple LRM to only select attributes strongly associated with the studied disease.

In future work, we will work on modifications that improve REMED's predictive capacity in terms of sensitivity ($\geq 50\%$) without significantly degrading its specificity. At the moment the REMED algorithm can only work with continuous attributes; we will include modifications that allow it to deal with discrete attributes. Finally, the rules obtained should be evaluated with data sets from other ABPM clinical studies.

Acknowledgements

The Maracaibo Aging Study is funded by the Venezuelan grant FONACIT G-97000876. MSc Salvador Pintos collaborated in the application of the statistical methods. Dr Jose Aizpurua collaborated in the application of the medical diagnostic tests and the clinical follow-up of the patients. Mallira Rodriguez transcribed the results of the ABPM study.

References

Bosnić, Z. and I. Kononenko (2008) Estimation of individual prediction reliability using the local sensitivity analysis, *Applied Intelligence*, **29**, 187–203.

Breault, J.L., C.R. Goodall and P.J. Fos (2002) Data mining a diabetic data warehouse, *Journal of Artificial Intelligence in Medicine*, **1–2**, 37–54.

Burden, R.L. and J.D. Faires (2000) *Numerical Analysis*, Pacific Grove, CA: Brooks/Cole.

Chawla, N.V., N. Japkowicz and A. Kolcz (2004) Editorial: special issue on learning from imbalanced data sets, *SIGKDD Explorations*, **6**, 1–6.

Di Rienzo, M., G. Grassi, A. Pedotti and G. Mancia (1983) Continuous vs intermittent blood pressure measurements in estimating 24 hour average blood pressure, *Hypertension*, **5**, 264–269.

Frattola, A., G. Parati, C. Cuspidi, F. Albini and G. Mancia (1993) Prognostic value of 24-hour blood pressure variability, *Journal of Hypertension*, **11**, 1133–1137.

Freund, Y. and L. Mason (1999) The alternating decision tree learning algorithm, *Proceedings of the 16th International Conference on Machine Learning*, San Francisco, CA: Morgan Kaufmann, 124–133.

Hanley, J.A. (1996) The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests, *Statistics in Medicine*, **15**, 1575–1585.

Holte, R.C. (1993) Very simple classification rules perform well on most commonly used datasets, *Machine Learning*, **11**, 63–91.

Hosmer, D.W. and S. Lemeshow (2000) *Applied Logistic Regression*, New York: Wiley.

Imai, Y., K. Abe, S. Sasaki, N. Minami, M. Munakata, H. Sekino, M. Nihei and K. Yoshinaga (1990) Determination of clinical accuracy and nocturnal blood pressure pattern by new portable device for monitoring indirect ambulatory blood pressure, *American Journal of Hypertension*, **3**, 293–301.

John, G.H. and P. Langley (1995) Estimating continuous distributions in Bayesian classifiers, in *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, San Mateo, CA: Morgan Kaufmann, 338–345.

Kikuya, M., A. Hozawa, T. Ohokubo, I. Tsuji, M. Michimata, M. Matsubara, M. Ota, K. Nagai, T. Araki, H. Satoh, S. Ito, S. Hisamichi and Y. Imai (2000) Prognostic significance of blood pressure and heart rate variabilities, *Hypertension*, **36**, 901–906.

Kononenko, I. (2001) Machine learning for medical diagnosis: history, state of the art and perspective, *Artificial Intelligence in Medicine*, **23**, 89–109.

Kukar, M. and C. Groselj (2000) Reliable diagnostics for coronary artery disease, in *Proceedings of the 15th IEEE Symposium on Computer Based Medical Systems*, New York: IEEE Press, 7–12.

Maestre, G., G. Pino, A. Molero, E. Silva, R. Zambrano, L. Falque, M. Gamero and T. Sulbaran (2002) The Maracaibo aging study: population and methodological issues, *Neuroepidemiology*, **21**, 194–201.

MANCIA, G. (1990) Ambulatory blood pressure monitoring: research and clinical applications, *Journal of Hypertension*, **7**, 1–13.
MCKEE, P., W. CASTELLI, T. MCNAMARA and W. KANNEL (1971) The natural history of congestive heart failure, *New England Journal of Medicine*, **285**, 1444–1446.
MENA, L. and J.A. GONZALEZ (2006) Machine learning for imbalanced datasets: application in medical diagnostics, in *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference*, Menlo Park, CA: AAAI Press, 574–579.
MENA, L., S. PINTOS, N. QUEIPO, J. AIZPURUA, G. MAESTRE and T. SULBARAN (2005) A reliable index for the prognostic significance of blood pressure variability, *Journal of Hypertension*, **23**, 505–512.
MITCHELL, T. (1997) *Machine Learning*, New York: McGraw-Hill.
PALATINI, P., M. PENZO, A. RACIOPPA, E. ZUGNO, G. GUZZARDI, M. ANACLERIO and A. PESSINA (1992) Clinical relevance of nighttime blood pressure and of daytime blood pressure variability, *Archives of Internal Medicine*, **152**, 1855–1860.
PICKERING, T., G. JAMES, C. BODDIE, G. HARSHFIELD, S. BLANKS and J. LARAGH (1988) How common is white-coat hypertension?, *Journal of the American Medical Association*, **259**, 225–228.
PODGORELEC, V., P. KOKOL and M.M. STIGLIC (2002) Searching for new patterns in cardiovascular data, in *Proceedings of the 15th IEEE Symposium on Computer Based Medical Systems*, New York: IEEE Press, 111–116.
RITCHIE, J., R. GIBBONS and M. CHEITLIN (1999) ACC/AHA/ACP-ASIM guidelines for the management of patients with chronic stable angina, A report of the ACC/AHA Task Force on practice guidelines (Committee on management of patients with chronic stable angina), *Journal of the American College of Cardiology*, **33**, 2092–2197.
RIZZONI, D., M.L. MUIESAN, G. MONTANI, R. ZULLI, S. CALEBICH and E. AGABITI-ROSEI (1992) Relationship between initial cardiovascular structural changes and daytime and nighttime blood pressure monitoring, *American Journal of Hypertension*, **5**, 180–186.
SANDER, D., C. KUKLA, J. KLINGELHOFER, W. KERSTIN and B. CONRAD (2000) Relationship between circadian blood pressure patterns and progression of early carotid atherosclerosis, *Circulation*, **102**, 1536–1541.
SHIMADA, K., A. KAWAMOTO, K. MATSUYABASHI, M. NISHINAGA, S. KIMURA and T. OZAWA (1992) Diurnal blood pressure variations and silent cerebrovascular damage in elderly patients with hypertension, *Journal of Hypertension*, **10**, 875–878.
VERDECCHIA, P., C. GOTTESCHI, G. BENEMIO, F. BOLDRINI, M. GUERRERI and C. PORCELLATI (1990) Circadian blood pressure changes and left ventricular hypertrophy in essential hypertension, *Circulation*, **81**, 528–536.
WEISS, G.M. (2004) Mining with rarity a unifying framework, *SIGKDD Explorations*, **6** (1), 7–19.
WITTEN, I.H. and E. FRANK (2000) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco, CA: Morgan Kaufmann.
WITTEN, I.H. and E. FRANK (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, San Francisco, CA: Morgan Kaufmann.
ZHENG, Z., X. WU and R. SRIHARI (2004) Feature selection for text categorization on imbalanced data, *SIGKDD Explorations*, **6**, 80–89.

The authors

Luis Mena

Luis Mena received his first degree and an MSc in applied computing from the University of Zulia, Venezuela. Currently he is concluding his PhD in computer science at the National Institute of Astrophysics, Optics and Electronics, Mexico. He is an associate professor in the Department of Computer Science at the Faculty of Engineering of the University of Zulia, Venezuela, and researcher at the National Observatory of Science, Technology and Innovation, Venezuela. His interest areas include data mining, pattern recognition and artificial intelligence. He has written four articles on artificial intelligence in medicine.

Jesus A. Gonzalez

Jesus A. Gonzalez obtained his bachelor degree in computer science and engineering from the University of the Americas, Mexico. He then obtained a Masters degree and PhD in computer science and engineering from the University of Texas at Arlington, USA. He is currently enrolled in the Computer Science Department at the National Institute of Astrophysics, Optics and Electronics, Mexico, as a professor and researcher. He is also the academic coordinator of the campus Mexico of the Regional Center for Space Science

and Technology Education for Latin America and the Caribbean. His interest areas include machine learning, data mining, remote sensing and geographic information systems.

Gladys Maestre

Gladys Maestre graduated from medical school at the University of Zulia, Venezuela, and, following a postdoctoral fellowship in the Department of Psychiatry at Massachusetts General Hospital, USA, completed her PhD in the Department of Pathology at Columbia University. She currently holds the positions of professor of neuroscience at the University of Zulia, Venezuela, and research associate scientist at Columbia University Sergievsky Center, USA. Her research projects focus on gene–environment interactions affecting cognitive and cardiovascular health, particularly related to aging. She leads the Maracaibo Aging Study and coordinates diverse activities related to capacity building for research in Venezuela and the Caribbean region.