Proceedings of the 19th International FLAIRS Conference (FLAIRS-2006), Melbourne Beach, Florida, May 11-13, 2006

Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic

Luis Mena [a,b] and Jesus A. Gonzalez [a]

[a] Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico.
[b] Department of Computer Science, Faculty of Engineering, University of Zulia, Maracaibo, Venezuela.
{lmena,jagonzalez}@.inaoep.mx

Abstract

In this paper, we present a new rule induction algorithm for machine learning in medical diagnosis. Medical datasets, as many other real-world datasets, exhibit an imbalanced class distribution. However, this is not the only problem to solve for this kind of datasets, we must also consider other problems besides the poor classification accuracy caused by the classes distribution. Therefore, we propose a different strategy based on the maximization of the classification accuracy of the minority class as opposed to the usually used sampling and cost techniques. Our experimental results were conducted using an original dataset for cardiovascular diseases diagnostic and three public datasets. The experiments are performed using standard classifiers (*Naïve Bayes*, *C4.5* and *k-Nearest Neighbor*), emergent classifiers *(Neural Networks* and *Support Vector Machines*) and other classifiers used for imbalanced datasets (*Ripper* and *Random Forest*). In all the tests, our algorithm showed competitive results in terms of accuracy and area under the *ROC* curve, but overcomes the other classifiers in terms of comprehensibility and validity.

Key words: machine learning, imbalanced datasets, medical diagnosis, accuracy, validity and comprehensibility.

1. Introduction

Many real-world datasets exhibit an imbalanced class distribution, where there exists a majority class with normal data and a minority class with abnormal or important data. Fraud detection, network intrusion and medical diagnosis are examples of this kind of datasets; however, opposite to other machine learning applications, the medical diagnostic problem does not end once we get a model to classify new instances. That is, if the instance is classified as sick (the most important class) the generated knowledge should be able of provide the medical staff with a novel point of view about the given problem. This could help to apply a medical treatment on time to avoid, delay, or diminish the incidence of the disease. Then, besides the classification accuracy, we should also consider the comprehensibility of diagnostic knowledge. Furthermore, we must consider an additional problem, the selection of relevant attributes (or risk factors). We should be focused over changeable attributes (can be changed with medical treatment) such as blood pressure or cholesterol levels and should not consider non-changeable attributes such as age and sex (usually good attributes for classification). This makes even harder the classification task.

Another important issue is that medical datasets used for machine learning should be representative of the general incidence of the studied disease. This is important to make possible the use of the generated knowledge with other populations. Therefore, the over-sampling and under-sampling techniques (Kubat and Matwin 1997, Chawla et al. 2002) frequently used to balance the classes and to improve the minority class prediction of some classifiers, could generate biased knowledge that might not be applicable to the general population due to the artificial manipulation of the datasets. For this reason, we propose a different strategy that tries to maximize the classification accuracy of the minority class (sick people) without modifying the original dataset. Thus, each of the steps of our algorithm is guided to reach this objective. Since we are dealing with binary classification problems, the majority class accuracy is guaranteed by default.

In section 2 we describe the methodology of our algorithm. In section 3, we present a brief description of the datasets and classifiers used in our experiments. We then compare (section 4) the performance of our algorithm with some standard classifiers, emergent classifiers, and classifiers specifically used for imbalanced datasets. This comparison makes reference to the accuracy, comprehensibility and validity (only for the symbolic classifiers) of the obtained results. In section 5 we show an analysis of the results and finally, in section 6 we present our conclusions and future work.

2. Methodology

In this section we propose a new algorithm called *REMED* (Rule Extraction for MEdical Diagnostic). The *REMED* algorithm includes three main steps: 1) attributes selection, 2) selection of initial partitions, and finally 3) rule construction.

2.1 Attributes Selection

For the first step we consider that in medical practice the collection of datasets is often expensive and time consuming. Then, it is desirable to have a classifier that is able to reliably diagnose with a small amount of data about the patients. In the first part of *REMED* we use *simple logistic regression* to quantify the risk of suffering the disease with respect to the increase or decrement of an

attribute. We always use high confidence levels (>99%) to select attributes that are really significant and to guarantee the construction of more precise rules. Other important aspect to mention is that depending on the kind of association established (positive or negative) through the *odds ratio* metric, we build the syntax with which each attribute's partition will appear in the rules system. This part of the algorithm is shown in the top of figure 1.

2.2 Partitions Selection

The second part of *REMED* comes from the fact that if an attribute x has been statistically significant in the prediction of a disease, then its mean $\overline{x}$ (mean of the values of the attribute) is a good candidate as initial partition of the attribute. We sort the examples by the attribute's value and from the initial partition of each attribute, we search the next positive example (class = 1) in the direction of the established association. Then, we calculate a new partition through the average between the value of the found example and the value of its predecessor or successor. This displacement is carried out only once for each attribute. This can be seen in the middle part of figure 1.

2.3 Rules Construction

In the last part of the algorithm, we build a simple rule system of the following way:

if $(e_{i,1} \geq p_1)$ **and** $(e_{i,j} \leq p_j)$ **and** … **and** $(e_{i,m} \geq p_m)$ **then** *class* = 1
else *class* = 0

where $e_{i,j}$ denotes the value of attribute j for example i, p_j denotes the partition for attribute j and the relation $\geq$ or $\leq$ depends on the association attribute-disease.

With this rule system we make a first classification. We then try to improve the accuracy of our system by increasing or decreasing the value of each partition as much as possible. For this we apply the bisection method and calculate possible new partitions starting with the current partition of each attribute and the maximum or minimum value of the examples for this attribute. We build a temporal rule system changing the current partition by each new partition and classify the examples again. We only consider a new partition if it diminishes the number of false positives (FP) but does not diminish the number of true positives (TP). This step is repeated for each attribute until we overcome the established convergence level for the bisection method or the current rule system is not able to decrease the number of FP (healthy persons diagnosed incorrectly). This part of the algorithm is exemplified at the bottom of figure 1.

We can appreciate that the goal of *REMED* is to maximize the minority class accuracy at each step, first selecting the attributes that are strongly associated with the positive class. Then stopping the search of the partition that better discriminates both classes in the first positive example, and finally trying to improve the accuracy of the rule system but without diminishing the number of TP (sick persons diagnosed correctly).

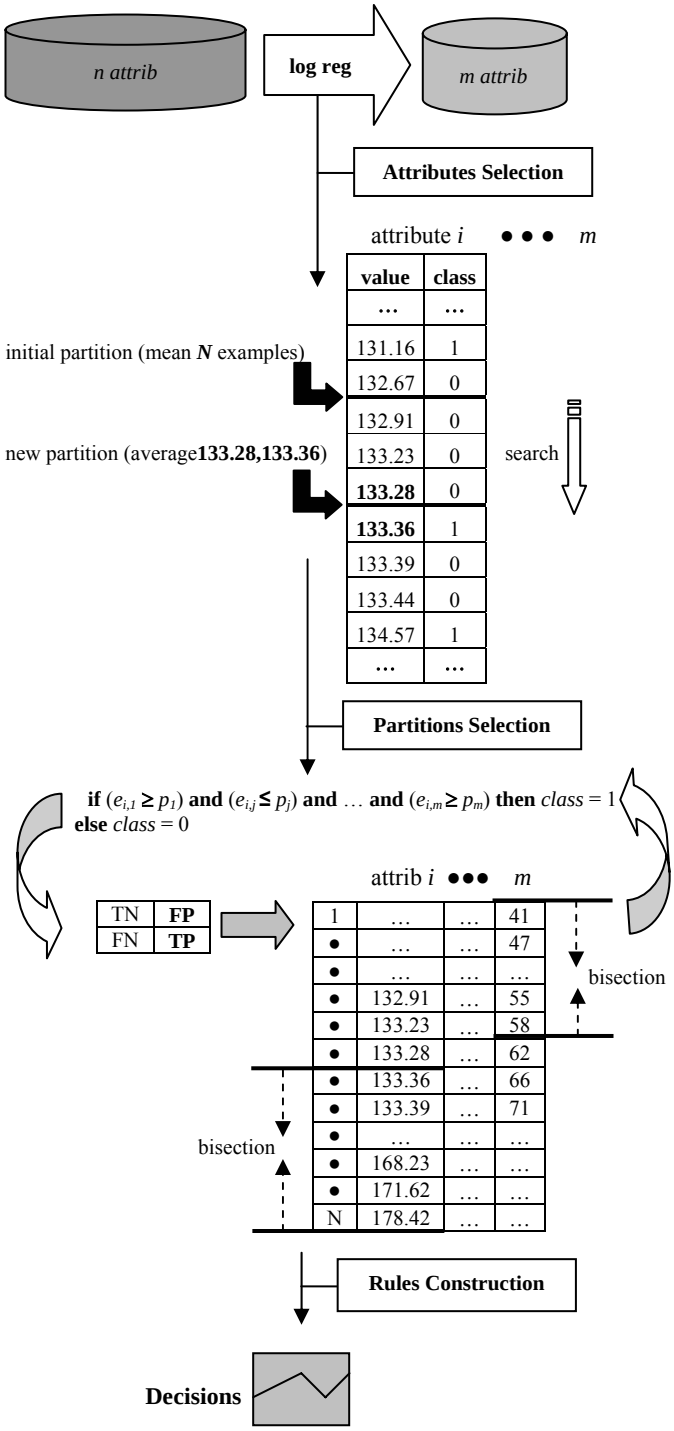

Figure 1. The outline of the *REMED* rule induction method.

3. Datasets and Classifiers

All the datasets have two classes and with the exception of the cardiovascular disease dataset, all were obtained from the UCI repository (Murphy and Aha 1994). In all the cases we only considered changeable (as discussed before) and continuous attributes (with more uncertainty than discrete attributes).

3.1 Cardiovascular Disease

This dataset was obtained from an Ambulatory Blood Pressure Monitoring (ABPM) (Mancia 1990) study named "The Maracaibo Aging Study" (Maestre et al. 2002) conducted by the Institute for Cardiovascular Diseases of the University of the Zulia, in Maracaibo, Venezuela. The final dataset was conformed by 312 observations and at the end of the study 55 individuals registered a cardiovascular disease (one of the world's most important causes of mortality).

The attributes considered were the mean of the readings of systolic blood pressure (SBP) and diastolic blood pressure, systolic global variability (SGV) and diastolic global variability measures with the average real variability (Mena et al. 2005) and systolic circadian variability (SCV) (Frattola et al. 1993), represented with the gradient of the linear approximation of the readings of SBP. All the attributes were calculated from the ABPM valid readings during the period of 24 hours and the dataset did not present missing values.

3.2 Hepatitis

This is a viral disease that affects the liver. The attributes considered were the levels of albumin (AL), bilirrubin (BL), alkaline phosphatase and serum glutamic oxaloacetic transaminase in the blood. The final dataset was conformed by 152 samples, with 30 positive examples and a rate of missing values of 23.03%.

3.3 Hyperthyroid

This is an extremely imbalanced dataset with 3693 negative samples and only 79 positive samples. The attributes considered to evaluate this disease of the thyroid glands were: thyroid-stimulating hormone (TSH), triiodothyronin (T3), total thyroxine (TT4), T4 uptake (T4U) and free thyroxine index (FTI). The dataset presented 27.07% of missing values.

3.4 Breast Cancer

The Wisconsin prognostic breast cancer dataset consists of 10 continuous-valued features computed from a digitized image of a fine needle aspirate of a breast mass. The characteristics of the cell nucleus present in the image were: radius (R), texture (T), perimeter (P), area (A), smoothness (SM), compactness (CM), concavity (C), symmetry (S), concave points and fractal dimension . The mean (me), standard error (se), and "worst" (w) or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. They also considered the tumour size (TS) and the number of positive axillary lymph nodes observed (LN). The dataset was conformed of 151 negative samples and 47 positive samples. Only 2.02% of the data presented missing values.

3.5 Classifiers

To compare the experimental results of *REMED* we used the standard classifiers: *Naïve Bayes* , *C4.5* and *1-Nearest Neighbor* (*1-NN*). We also used emergent classifiers such as the *Multi-Layer Perceptron* (*MLP*), *Radial Basis Function Neural Networks* (*RBF-NN*), and *Support Vector Machines* (*SVM*) and other two classifiers used for imbalanced datasets: *Ripper* (Cohen 1995) and *Random Forest* (Chen, Liaw, and Breiman 2004).

4. Experimental Results

With the exception of *REMED*, all the classifiers were obtained from the *Weka* framework (Witten and Frank 1999). We used *SMO* (Platt 1998) as the *SVM* classifier. In all the cases we applied the *10-fold cross validation* technique to avoid overfitting. The performance of the classifiers is presented in terms of accuracy (Acc), sensitivity (Sens), specificity (Spec), and area under the *RO*C curve (AURC). The results are summarized in tables 1 trough 4. For those classifiers that do not include an attribute selection step we provided them with *logistic regression** or *gain ratio*** in case that this enhanced their result in terms of AURC. The confidence level established for *REMED* to select the statistically significant attributes appears in parenthesis.

Table 1: Results for the Cardiovascular Disease Domain

Classifier	Acc	Sens	Spec	AURC
REMED (99%)	81.09	32.73	91.44	62.78
1-NN **	73.40	27.27	83.27	60.60
Naïve Bayes	80.13	21.82	92.61	58.61
MLP	79.49	12.73	93.77	55.06
Random Forest	78.53	12.73	92.61	55.05
RBF-NN	81.73	3.64	95.33	51.45
C4.5	81.09	1.82	98.05	50.73
SMO	81.73	0	100	50
Ripper	81.09	0	98.44	50

Table 2: Results for the Hepatitis Domain

Classifier	Acc	Sens	Spec	AURC
REMED (99.99%)	78.29	66.67	81.15	74.38
Naïve Bayes	83.55	50	91.80	69.09
1-NN *	78.95	50	84.43	68.94
MLP	86.84	46.67	96.72	67.96
RBF-NN *	82.24	46.67	91.80	67.91
C4.5	84.21	40	95.08	65.52
Ripper	80.26	33.33	91.80	63.01
Random Forest	79.61	30	91.80	61.75
SMO	83.55	13.33	100	55.30

Table 3: Results for the Hyperthyroid Domain

Classifier	Acc	Sens	Spec	AURC
Naïve Bayes	96.79	96.20	96.80	83.19
REMED (99.99%)	98.30	73.42	98.84	76.86
Ripper	98.28	69.62	98.89	75.68
RBF-NN	98.28	54.43	99.21	70.64
MLP	98.41	51.90	99.40	69.81
1-NN *	91.70	50.63	92.58	69.22
C4.5	98.38	48.10	99.46	68.47
Random Forest	98.46	44.30	99.62	67.11
SMO	98.20	17.72	99.92	57.03

Table 4: Results for the Breast Cancer Domain

Classifier	Acc	Sens	Spec	AURC
REMED (99%)	62.63	46.81	67.55	67.88
Naïve Bayes	64.65	44.68	70.86	66.60
1-NN	68.69	36.17	78.81	63.83
MLP	62.63	27.66	78.81	60.66
Random Forest	62.63	14.89	92.72	55.90
C4.5	68.18	14.89	84.77	55.85
Ripper	73.74	6.38	94.70	52.54
RBF-NN	75.25	2.13	98.01	50.85
SMO	76.26	0	100	50

As we previously mentioned, besides of the classification accuracy, comprehensibility is an important issue for machine learning in medical diagnosis. Without any doubt the symbolic learning classifiers (decision trees and rules) offer better comprehensibility than the rest of the machine learning classifiers. In tables 5, 6, 7, and 8 we show the rule systems produced by each symbolic classifier. In some cases, where the classifier performance in terms of AURC was very low (< 55%), *Weka* did not show the respective rule system (that is why in table 5 only *REMED* appears and in table 8 *Ripper* does not appear). In these rules we can see how many attributes were chosen by *REMED* in the attribute selection phase (i.e. for the Cardiovascular Disease domain, *REMED* chose the SBP, SGV, and SCV attributes). We also analyze the validity of the obtained rules, comparing them with some well-known risk factors for each disease.

Table 5: Rule System for the Cardiovascular Disease Domain

REMED
if SBP ≥ 142.1784 **and** SGV ≥ 9.2575 **and** SCV ≥ -0.4025 **then** *sick* **else** *no sick*

Table 6: Rule Systems for the Hepatitis Domain

REMED
if BL ≥ 1.4 **and** Al ≤ 3.4 **then** *sick* **else** *no sick*
Ripper
if BL > 1.4 **and** Al ≤ 3.6 **then** *sick* **else** *no sick*
C4.5
if BL ≤ 3.5 **and** Al ≤ 2.6 **then** *sick* **if** BL > 3.5 **and** Al ≤ 3.8 **then** *sick* **else** *no sick*

Table 7: Rule Systems for the Hyperthyroid Domain

REMED
if FTI ≥ 156 **and** TT4 ≥ 144 **and** TSH ≤ 0.25 **and** T3 ≥ 1.7 **then** *sick* **else** *no sick*
Ripper
if FTI ≥ 159 **and** T3 ≥ 3.5 **then** *sick* **if** FTI ≥ 171 **and** TT4 ≥ 157 **and** TT4 ≤ 200 **and** TSH ≤ 0.25 **and** T3 ≥ 1.5 **and** T4U ≤ 0.91 **then** *sick* **else** *no sick*
C4.5
if FTI > 155 **and** TT4 > 149 **and** TSH ≤ 0.01 **and** T3 ≤ 4 **and** T4U ≤ 0.91 **then** *sick* **if** FTI > 155 **and** TT4 > 156 **and** TT4 ≤ 167 **and** TSH > 0.01 **and** TSH ≤ 0.26 **and** T3 ≤ 4 **and** T4U ≤ 0.91 **then** *sick* **if** FTI > 155 **and** TT4 > 167 **and** TSH > 0.01 **and** TSH ≤ 0.26 **and** T3 ≤ 4 **and** T4U ≤ 0.85 **then** *sick* **if** FTI > 155 **and** TT4 > 149 **and** TSH ≤ 0.26 **and** T3 > 4 **then** *sick* **else** *no sick*

Table 8: Rule Systems for the Breast Cancer Domain

REMED
If Ame ≥ 981.05 **and** Rw ≥ 21.0218 **and** Pw ≥ 143.4 **and** Aw ≥ 1419 **then** *sick* **else** *no sick*
C4.5
if LN ≤ 3 **and** TS ≤ 2.1 **and** SMw > 0.1482 **and** SMme ≤ 0.111 **and** Tw ≤ 21.43 **then** *sick* **if** LN ≤ 3 **and** TS ≤ 2.1 **and** SMw > 0.1482 **and** SMme > 0.111 **and** SMme ≤ 0.115 **then** *sick* **if** LN ≤ 0 **and** TS > 2.1 **and** TS ≤ 2.8 **and** Tme ≤ 26.29 **and** Rw ≤ 26.51 **and** Cse > 0.04497 **then** *sick* **if** LN > 0 **and** LN ≤ 3 **and** TS > 2.1 **and** Tme ≤ 20.66 **and** Rw ≤ 26.51 **and** CMw ≤ 0.429 **and** CMe ≤ 0.07789 **then** *sick* **if** LN > 0 **and** LN ≤ 3 **and** TS > 2.1 **and** Tme > 20.66 **and** Rw ≤ 26.51 **and** CMw ≤ 0.429 **then** *sick* **if** LN ≤ 3 **and** TS > 2.1 **and** Tme ≤ 26.29 **and** Rw > 26.51 **then** *sick* **if** LN > 3 **and** Aw ≤ 2089 **and** Sw ≤ 0.3277 **and** Tse ≤ 1.198 **and** Pse ≤ 3.283 **and** Tme > 19.22 **then** *sick* **if** LN > 3 **and** Aw ≤ 2089 **and** Sw ≤ 0.3277 **and** Tse > 1.198 **and** Tse ≤ 1.481 **then** *sick* **if** LN > 3 **and** Aw ≤ 2089 **and** Sw > 0.3277 **and** Tw > 34.12 **and** Rme > 13.48 **then** *sick* **if** LN > 3 **and** Aw > 2089 **then** *sick* **else** *no sick*

Finally, in order to evaluate the validity of the obtained rules we show in table 9 some abnormal values of certain attributes that according to the medical literature could represent a risk factor of the corresponding disease.

Table 9: Well-known Risk Factors

disease	abnormal values
Cardiovascular	SBP > 140 mmhg
Hepatitis	BL > 1.2 mg/dl AL < 3.4 g/dl
Hyperthyroid	FTI > 155 nmol/l TT4 > 140 nmol/l TSH < 0.4 mlU/l T3 >1.8 nmol/l

5. Results Analysis

First, we should analyze why it is so difficult to apply machine learning to medical diagnosis. One example of this is the low performance showed by the used classifiers in terms of AURC, since in some cases the best performance did not reach 70% (Cardiovascular Disease and Breast Cancer domains). One reason of this is that most of the datasets are built from longitudinal medical studies that consist on observing the apparition of a disease in a group of individuals during a specific period of time. At the end of the study a binary classification is done, and every subject is classified as either healthy or sick. However, an individual that presented clear risk factors during the period of study, but that his death was not caused by the studied disease (i.e. an accident), or at the end of the study he did not present the disease (being very probable that he developed it just after the end of the study), is classified as healthy, and this situation tends to confuse the classifiers. In spite of this inconvenient, *REMED* showed a regular performance in all the domains, ranked in the first places in terms of AURC and sensitivity. Other classifiers with a constant regular performance were *Naïve Bayes* and *1-NN*, but these have the disadvantage of not being symbolic classifiers and the results are not rich in comprehensibility.

With respect to comprehensibility, *REMED* always produces very simple rule systems, conformed only by two rules. *Ripper* also produced simple rule systems in all the studied cases, but it was not as precise as *REME*D. A clear example can be seen in the hepatitis domain, where apparently both classifiers produced similar rule systems, but *REMED* thoroughly overcomed *Ripper* in terms of the AURC and sensitivity. *C4.5* always produced larger rule systems than *REMED* and *Ripper*. Other advantage of *REMED* for medical domains is that it does not produce rules with enclosed intervals (i.e. $a \leq x \leq b$). This is important because it could represent an inconvenient in medical diagnosis, because the risk of developing a disease is directly proportional to the increase or decrease of a risk factor. Furthermore, the increment or decrement of a risk factor could be related to two different diseases (i.e. hypothyroid and hyperthyroid).

Other aspect important to analyze is the validity of the rule systems. We can appreciate from table 9 that in all the cases the rules proposed by *REMED* are closer to the well-known risk factors for each disease. In the specific case of the cardiovascular disease domain, the rule antecedents related with the BP variability could represent new knowledge to be used for the diagnostic of this important kind of disease. Moreover, the fact that the rule systems of *REMED* are always supported by a selection of attributes with high confidence levels, could provide the medical staff enough trust to use these rules in the practice.

6. Conclusions and Future Work

As we could see from the results, *REMED* is a very competitive algorithm that can be used in the medical diagnostic area. However, we should mention that *REMED* does not pretend to be the panacea of machine learning in medical diagnostic, but a good approach with the desired features to solve medical diagnostic tasks, good performance, the comprehensibility of diagnostic knowledge, the ability to explain decisions, and the ability of the algorithm to reduce the number of tests necessary to obtain reliable diagnosis (Kononenko 2001). It is also important to mention that the *REMED* algorithm can be scaled to work with larger databases than those used in our experiments. This is because the complexity of *REMED* is $O(n^2)$ and independently of the number n of examples and m initial attributes, *REMED* always produces simple rule systems only composed of 2 rules (including the default rule: **else** *class* = 0) and with a maximum of m conditions per rule. However, we still need to work to improve the performance of *REMED*, a possible way to do it could be the combination of *REMED* with *Boosting* techniques (Freund and Schapire 1996) or Cost-Sensitive strategies. We also want to increase the versatility of *REMED*, including modifications that allow it to consider discrete attributes, to work with multi-class problems and inclusive in some cases to generate rule systems with enclosed intervals. This will be done to be able to use *REMED* in other domains with imbalanced datasets.

References

Chawla, N.; Bowyer, K.; Hall, L.; and Kegelmeyer, P. 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research* vol. 16, 321-357.

Chen, C.; Liaw, A.; and Breiman, L. 2004. Using Random Forest to Learn Imbalanced Data. Technical report 666, Statistics Department, University of California at Berkeley.

Cohen, W. 1995. Fast Effective Rule Induction. *Proceedings of the Twelfth International Conference on Machine Learning*, Lake Tahoe, CA, 115-123.

Frattola, A.; Parati, G.; Cuspidi, C.; Albini, F.; and Mancia,G. 1993. Prognostic Value of 24-hour Blood Pressure Variability. *Journal of Hypertension* 11:1133-1137.

Freund, Y.; and Schapire, R. E. 1996. Experiments with a New Boosting Algorithm. *Proceedings of the Thirteenth International Conference on Machine Learning*, 325-332.

Kubat, M; and Matwin, S. 1997. Addressing the Curse of Imbalanced Training Sets: One-Sided Sampling. *Proceedings of the Fourteenth International Conference on Machine Learning*, Morgan Kaufmann pp.179-186.

Kononenko, I. 2001. Machine Learning for Medical Diagnosis: History, State of the Art and Perspective, Invited paper, *Artificial Intelligence in Medicine* – ISSN 0933-3657 23(1):89–109.

Maestre, G.; Pino, G.; Molero, A.; Silva, E.; Zambrano, R.; Falque, L.; et al. 2002. The Maracaibo Aging Study: Population and Methodological Issues. *Neuropidemiology* 21:194–201.

Mancia, G. 1990. Ambulatory Blood Pressure Monitoring: Research and Clinical Applications. *Journal of Hypertension* (Suppl 7): S1-S13.

Mena, L.; Pintos, S.; Queipo, N.; Aizpurua, J.; et al. 2005. A Reliable Index for the Prognostic Significance of Blood Pressure Variability. *Journal of Hypertension* 23:505–512.

Murphy, P.; and Aha, D. 1994. UCI Repository of Machine Learning Databases [machine-readable data repository]. Technical Report, University of California, Irvine.

Platt, J.C. 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Technical Report MSR-TR-98-14, Microsoft Research.

Witten, I.H.; and Frank, E. 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.