JHI JOURNAL OF HEALTH INFORMATICS

Evaluation of Stacking on Biomedical Data*

Avaliação de *Stacking* em Dados Biomédicos

Evaluación de *Stacking* en Datos Biomédicos

Maria Izabela Ruz Caffé[1], Pedro Santoro Perez[2], José Augusto Baranauskas[2]

Keywords: Artificial Intelligence, Classification, Ensembles

ABSTRACT

Objectives: Stacking is a well-known ensemble technique, but some of its aspects still need to be explored, e.g., there are few recommendations on which and how many algorithms should be used at level-0 or even which algorithm should be used to compose the level-1 meta-classifier. The literature indicates the meta-algorithm at level-1 should be simple, and Naive Bayes has been typically used in these studies. **Methods:** In this work, we have analyzed stacking on biomedical datasets, using three different paradigms of machine learning algorithms to compose the meta-classifier. **Results:** The experiments indicate simple meta-algorithms do not provide good results. **Conclusion:** the meta-classifier must have a degree of complexity to provide a nice performance.

Descritores: Inteligência Artificial, Classificação, Combinação de Classificadores

RESUMO

Objetivos: Stacking é uma técnica bem conhecida de combinação de classificadores, mas alguns de seus aspectos ainda precisam ser explorados, por exemplo, existem poucas recomendações sobre quais e quantos algoritmos devem ser utilizados no nível-0 ou ainda qual algoritmo deve ser usado para compor o meta-classificador do nível-1. A literatura indica que o meta-algoritmo no nível-1 deve ser simples e geralmente Naive Bayes tem sido usado nestes estudos. **Métodos:** Neste trabalho, o algoritmo de stacking foi avaliado em dados biomédicos, usando três algoritmos de aprendizado de máquina de diferentes paradigmas para compor o meta-classificador. **Resultados:** Os experimentos indicam que meta-algoritmos simples não fornecem bons resultados. **Conclusão:** O meta-classificador deve ter um grau de complexidade para oferecer um bom desempenho.

Descriptores: Inteligencia Artificial, Clasificación, Combinación de Clasificadores

RESUMEN

Objetivos: Stacking es una técnica de combinación de clasificadores bien conocida, pero algunos aspectos quedan por explorar, por ejemplo, existen pocas recomendaciones sobre cuales o cuantos algoritmos deben utilizarse en el nível-0 o aun cual algoritmo debe usarse para componer el nível-1. La literatura indica que el meta-algoritmo debe ser simple y, generalmente, Naive Bayes ha sido usado en estos estudios. **Métodos:** En este trabajo, se analiza el algoritmo de stacking con datos biomédicos, utilizando tres algoritmos de aprendizaje automático de distintos paradigmas para componer el meta-clasificador. **Resultados:** Los experimentos indican que meta-algoritmos simples no ofrecen buenos resultados. **Conclusión:** El meta-clasificador debe tener un grado de complejidad para obtener un buen rendimiento.

** Research project with financial support from CNPq/FAPEAM – INCT ADAPTA.*

[1] Department of Computer Science and Mathematics, Faculty of Philosophy, Sciences and Languages at Ribeirão Preto. Faculty of Medicine at Ribeirão Preto, University of Sao Paulo - USP, Ribeirão Preto (SP), Brasil.

[2] Faculty of Medicine at Ribeirão Preto, University of Sao Paulo - USP, Ribeirão Preto (SP), Brasil.

Autor Correspondente: **Maria Izabela Ruz Caffé**
e-mail: **mariairc@aluno.ffclrp.usp.br**

Artigo recebido: **10/11/2011**
Aprovado: **21/08/2012**

INTRODUCTION

Combination of classifiers (ensembles) is a machine learning method in which several models (hypothesis) are combined to generate a single final classifier. Generally, the use of ensembles has a tendency to decrease the error rate, making the final classifier more accurate, since it uses all learning algorithms predictions to generate a final hypothesis[1].

Stacking[2] is one of the heterogeneous classifiers combining methods. The proposed method is a two-layer structure: at level-0 learning algorithms take as input the training data, thus generating the level-0 classifiers; the next layer (level-1) takes as input the predictions of the previous layer (level-0) and a level-1 meta-algorithm combines them to provide the final meta-classifier h^*, as shown in Figure 1. More precisely, assume L different learning algorithms $A_1, A_2, \ldots, A_L$ and a set of n examples $\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$, where is implicit the fact each element x_i is a multidimensional vector. Each level-0 algorithm $\{A_1, A_2, \ldots, A_L\}$ is applied to the training set, inducing classifiers $\{h_1, h_2, \ldots, h_L\}$. Then, each level-0 classifier is used to label the examples. This implies that, for each example x_i, a tuple is formed, composed by the class predicted by each level-0 classifier as well as its true class label, i.e., $(h_1(x_i), h_2(x_i), \ldots, h_L(x_i), y_i)$. These tuples compose the level-1 training set, where attributes are the classes predicted by each one of the L classifiers. This level 1 training set is taken as input to the level-1 meta-algorithm to learn the final h^* meta-classifier.

One of the motivations to use stacking is that, passing through the levels, the meta-classifier learns the previous classifiers errors. In his article, Wolpert[2] relates that many aspects about stacking still are unknown (*black art*) in the sense that there are no recommendations on which or how many algorithms should be used to compose the meta-classifier at level-1. Generally, studies in literature use a variable number of algorithms at level-0 and Naive Bayes to generate the meta-classifier.

Normally, biomedical data are associated to a large quantity of classes, different numbers of examples, high dimensionality and examples with missing and redundant data, which has motivated the application of machine learning algorithms to this domain[3-4]. This way, Tanwani's group research[5] sought to construct a guideline to use ensembles, with emphasis on biomedical datasets. Nevertheless, the authors used only Naive Bayes as the meta-classifier in their analysis of stacking. With that in mind, the objective of this work consists in evaluating *stacking* on biomedical datasets using different classifiers at both level-0 and level-1, extending Tanwani's work[5] on *stacking*.

The remainder of this paper is organized as follows. Section 2 describes the algorithms used to induce classifiers and meta classifiers, the datasets, and the three experiments performed in this work. In Section 3, the results are exposed and discussed; the conclusions are presented in Section 4.

EXPERIMENTAL METHODOLOGY

As mentioned before, a previous study[5] used only Naive Bayes as meta-classifier. In this paper, the use of *stacking* is extended, comparing the performance under three different paradigms at level-1: statistical-based (Naive Bayes), induction of simple rules (One Rule) and one-

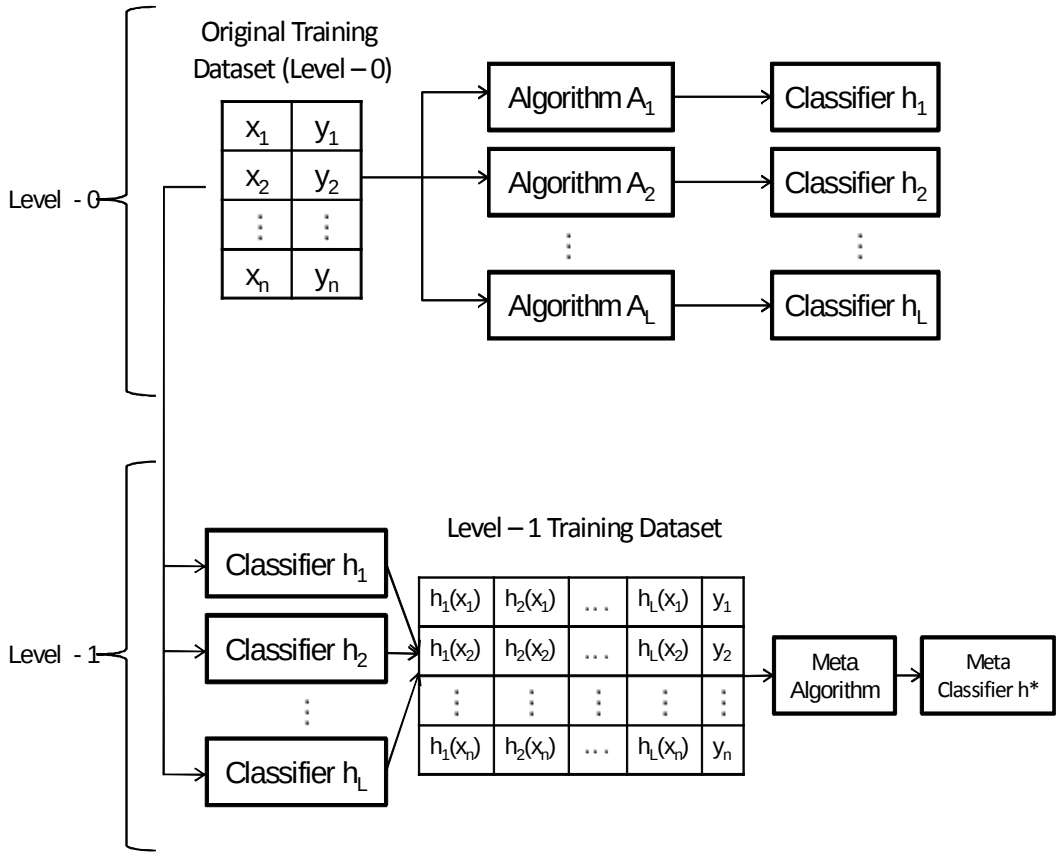

Figure 1 - *Stacking*

level decision tree (Decision Stump). The objective is to try to verify whether algorithms with different paradigms result in a better combination in *stacking*.

In summary, this work is composed of three experiments involving stacking performed in eighteen biomedical datasets using different algorithms, which are described in subsequent sections.

Datasets

The datasets used in this work refer to the biomedical domain selected from the *UCI Machine Learning Repository*[6]. Table 1 contains information about the datasets: name, date of publication, number of examples, number of attributes, number and distribution of classes and whether missing data is present. Datasets have not been pre-processing in anyway. In what follows, a brief biological description about the datasets is given.

Breast Cancer: from attributes with clinical and laboratory information about patients, the task is to distinguish recurrent events from non-recurrent events associated to breast cancer (Oncology Institute of Yugoslavia).

Haberman: from attributes about breast cancer surgeries, the objective is to predict whether the patient survived after the surgery during the study.

Heart-statlog: also based on clinical and laboratory information, the classification problem is to predict the absence or presence of heart disease.

Hepatitis: considering clinical and laboratory data, the task is to predict whether a patient with hepatitis died or survived during the study period.

Liver Disorders: prediction of a particular binary class based on blood tests related to liver problems. The documentation about this dataset does not make clear what the class means. It is known that it can assume two values.

Pima Indians: this problem contains attributes about clinical and laboratory data from female descendants of Pima Indian, whose classification task is to distinguish patients who tested positive for diabetes and patients who tested negative on this disease.

Promoters: based on E. coli DNA sequences, the task consists in discriminating whether or not a particular sequence is a gene promoter.

Sick: the task is to predict whether patients have a certain thyroid disease based on clinical and laboratory data.

Contraceptive Method: based on socio-economic attributes of women in Indonesia, the task is to predict which contraceptive method is used by them.

Lung Cancer: the authors of this dataset do not give information about its attributes. The task is to differentiate among three types of lung cancer.

Postoperative: the problem is to determinate where postoperative recovery patients should be referred to, based on attributes related to clinical measurements about patients (e.g., body temperature, blood pressure).

Ann-thyroid: the task is to distinguish among three situations involving hypothyroidism. Attributes bring data about laboratory measurements related to thyroid problems.

Lymphography: based on clinical and laboratory findings related to lymphography, the task is to discriminate among lymph tumor conditions (Oncology Institute of Yugoslavia).

Cleveland, Switzerland and Hungarian: the problem consists in discriminating among conditions related to heart diseases. The attributes bring information about clinical and laboratory findings.

Dermatology: the problem is to determine eryhemato-squamous diseases from clinical and

Table 1 - Datasets ordered by the number of classes

Dataset	Year	Samples	Attributes	Classes	Distribution of Classes	Missing
Breast cancer	1988	286	9	2	(70.28, 29.72)	Yes
Haberman	1999	306	4	2	(73.53, 26.47)	No
Heart statlog	N/D	270	13	2	(55.56, 44.44)	No
Hepatitis	1988	155	20	2	(20.65, 79.35)	Yes
Liver disorders	1990	345	7	2	(42.03, 57.97)	No
Pima indians	1990	768	9	2	(65.10, 34.90)	No
Promoters	1990	106	58	2	(50.00, 50.00)	No
Sick	1987	3772	30	2	(93.88, 6.12)	Yes
Contraceptive	1997	1473	9	3	(42.70, 22.61, 34.69)	No
Lung cancer	1992	32	56	3	(28.13, 40.62, 31.25)	Yes
Postoperative	1993	90	9	3	(71.11, 2.22, 26.67)	Yes
Ann-thyroid	1992	7200	22	3	(2.31, 5.11, 95.58)	No
Lymphography	1988	148	19	4	(1.35, 54.73, 41.22, 2.70)	No
Cleveland	1988	303	13	5	(54.46, 45.54, 0.00, 0.00, 0.00)	Yes
Hungarian	1988	294	14	5	(63.95, 34.98, 0.00, 0.00, 0.00)	Yes
Switzerland	1988	123	14	5	(6.50, 39.02, 26.02, 24.39, 4.07)	Yes
Dermatology	1998	366	33	6	(30.60, 16.67, 19.57, 13.39, 14.20, 5.46)	Yes
Ecoli	1996	336	8	8	(42.56, 22.92, 15.48, 10.42, 5.95, 1.49, 0.60, 0.60)	No

Table 2 - Algorithms and Meta Algorithms in each Experiment

Experiment	Meta Algorithm	Algorithms
Meta 1	Naive Bayes	Bayes Net, IBK, One Rule, JRip, J48, Decision Stump, MLP, SMO
Meta 2	One Rule	Bayes Net, Naive Bayes, IBK, JRip, J48, Decision Stump, MLP, SMO
Meta 3	Decision Stump	Bayes Net, Naive Bayes, IBK, One Rule, JRip, J48, MLP, SMO

histopathological findings.

Ecoli: the task is to predict the sites of cellular localization of proteins from measures and scores related to protein sequences.

Algorithms and Meta-Algorithms

In the experiments we have used nine machine learning algorithms from the Weka machine learning library[7], with their default parameters, unless otherwise specified. A brief description about each of them is given below.

Naive Bayes uses a probabilistic method to classify[8], assuming independence among attributes.

Bayes Network uses various search algorithms and quality measures, providing structures about the data[9].

IBK (*Instanced Based Learner*) classifies according to the *K*-nearest neighbors[10]. In this study we have used $K = 3$.

JRip induces rules and it is a reimplementation of the Ripper algorithm[11].

One Rule is a very simple classifier that selects one attribute to build the rule with the lowest error rate[7].

J48 generates decision trees and it is a reimplementation of the C4.5 algorithm[12].

Decision Stump constructs a decision tree with one single level using entropy[13].

MLP (*Multi-Layer Perceptron*) is a multilayer neural network using back propagation to optimize the weights during training[14].

SMO (*Support Vector Machine*) John Platt's minimal optimization sequential algorithm to train support vectors[15].

The experiments involving *stacking* are identified as Meta 1, Meta 2 and Meta 3, as shown on Table 2. Each of the nine algorithms also had their performance assessed individually. The performance was estimated using 10-fold stratified cross-validation.

RESULTS AND DISCUSSION

Table 3 presents the AUC values (area under the ROC curve)[16] for experiments Meta 1, Meta 2 and Meta 3 and all other algorithms used in this work, which produced the following ranking order: (i) Meta 1, (ii) Bayes Net, (iii) Naive Bayes, (iv) MLP, (v) J48, (vi) IBK, (vii) SMO, (viii) Meta 3, (ix) Jrip, (x) Meta 2, (xi) Decision Stump, and (xii) One Rule. It is possible to observe that the classifier obtained by Meta 1 had a better rank than Meta 2 and Meta 3. On the other hand, Meta 2 and Meta 3 classifiers, in general, had a worse performance than the other classifiers alone, *i.e.*, Bayes Net, Naive Bayes, MLP, J48, IBK and SMO.

To validate the results under a statistical point of view, the Friedman test[17] was performed with a significance level of $\alpha=0.05$, in order to determine whether there were significant differences among AUC values. As the null hypothesis was rejected (*i.e.*, there exist significant differences among classifiers), a *post-hoc* test[18] was performed, in order to search for pairs with a significant difference, considering all pairwise comparisons. The *post-hoc* test confirmed that Meta 1 did significantly better than all other algorithms, including Meta 2 and Meta 3, except for Naive Bayes, Bayes Net and MLP, against which Meta 1 did better, but not significantly better.

Table 4 shows these results, where Ä (▲) indicates that the algorithm from the line was better (significantly) than the algorithm from the respective column, and □(▼)

Table 3 - AUC (±Standard Deviation) Measures from Experiments

Dataset	Naive Bayes	BayesNet	IBK	JRip	One Rule	Decision Stump	J48	MLP	SMO	Meta 1	Meta 2	Meta 3
Breast câncer	0.72±0.14	0.71±0.14	0.66±0.13	0.61±0.10	0.54±0.07	0.65±0.13	0.63±0.10	0.62±0.13	0.59±0.08	0.69±0.15	0.57±0.09	0.67±0.11
Haberman	0.67±0.11	0.69±0.09	0.63±0.11	0.62±0.10	0.59±0.08	0.64±0.08	0.58±0.08	0.66±0.13	0.51±0.02	0.72±0.05	0.51±0.07	0.55±0.04
Heart statlog	0.90±0.06	0.91±0.04	0.83±0.07	0.80±0.08	0.71±0.06	0.72±0.07	0.76±0.10	0.85±0.06	0.84±0.06	0.89±0.06	0.79±0.08	0.83±0.08
Hepatitis	0.86±0.11	0.89±0.08	0.79±0.15	0.60±0.15	0.65±0.13	0.67±0.13	0.70±0.20	0.82±0.15	0.75±0.13	0.83±0.12	0.73±0.16	0.73±0.16
Liver disorders	0.65±0.12	0.52±0.03	0.64±0.06	0.64±0.09	0.54±0.07	0.54±0.06	0.67±0.08	0.74±0.07	0.50±0.01	0.75±0.08	0.58±0.06	0.66±0.09
Pima indians	0.82±0.05	0.81±0.05	0.74±0.05	0.72±0.06	0.67±0.06	0.69±0.06	0.75±0.08	0.80±0.04	0.72±0.06	0.83±0.05	0.67±0.04	0.73±0.05
Promoters	0.96±0.11	0.96±0.11	0.92±0.08	0.81±0.13	0.70±0.11	0.71±0.12	0.83±0.14	0.98±0.08	0.93±0.09	0.97±0.08	0.58±0.11	0.90±0.09
Sick	0.93±0.05	0.96±0.02	0.88±0.05	0.94±0.05	0.89±0.04	0.94±0.03	0.95±0.04	0.95±0.03	0.50±0.00	0.99±0.01	0.93±0.03	0.94±0.03
Contraceptive	0.69±0.04	0.70±0.04	0.62±0.04	0.62±0.04	0.58±0.02	0.56±0.02	0.66±0.04	0.70±0.02	0.63±0.03	0.73±0.04	0.59±0.04	0.60±0.02

continue...

...continuation

Lung cancer	0.71±0.32	0.71±0.29	0.68±0.31	0.55±0.15	0.54±0.15	0.55±0.11	0.68±0.23	0.56±0.22	0.63±0.20	0.80±0.19	0.70±0.21	0.57±0.15
Postoperative	0.39±0.22	0.40±0.21	0.31±0.13	0.50±0.00	0.48±0.03	0.46±0.13	0.49±0.02	0.41±0.19	0.47±0.04	0,57±0.18	0.46±0.11	0.43±0.12
Ann-thyroid	0.93±0.02	1.00±0.00	0.74±0.04	0.99±0.01	0.95±0.02	0.99±0.00	0.99±0.00	0.97±0.03	0.59±0.02	1.00±0.00	0.97±0.03	0.99±0.00
Lymph	0.91±0.07	0.91±0.07	0.89±0.10	0.78±0.14	0.77±0.11	0.78±0.11	0.79±0.14	0.91±0.07	0.87±0.08	0.87±0.11	0.80±0.06	0.81±0.07
Cleveland	0.90±0.06	0.91±0.04	0.85±0.07	0.84±0.07	0.72±0.07	0.71±0.07	0.80±0.09	0.89±0.06	0.84±0.08	0.91±0.05	0.79±0.08	0.81±0.07
Hungarian	0.90±0.06	0.91±0.07	0,87±0.06	0.76±0.10	0.75±0.11	0.77±0.11	0.77±0.15	0.89±0.04	0.80±0.11	0.89±0.07	0.83±0.06	0.80±0.08
Switzerland	0.53±0.12	0.54±0.04	0.49±0.09	0.56±0.07	0.53±0.10	0.47±0.06	0.55±0.09	0.54±0.12	0.57±0.09	0.52±0.13	0.50±0.07	0.51±0.04
Dermatology	1.00±0.00	1.00±0.00	0.99±0.01	0.95±0.02	0.67±0.03	0.79±0.02	0.97±0.02	1.00±0.00	0.98±0.01	1.00±0.00	0.82±0.03	0.78±0.01
Ecoli	0.97±0.02	0.97±0.02	0.94±0.03	0.91±0.04	0.75±0.04	0.78±0.02	0.92±0.05	0.96±0.02	0.95±0.02	0.93±0.04	0.81±0.04	0.83±0.02
Average *Ranking* AUC	4.00	3.06	6.83	7.64	10.42	9.28	6.56	4.28	7.19	2.58	8.61	7.50
Average *Ranking* Standard Deviation	7.14	5.47	7.31	7.83	6.64	6.08	8.94	5.54	5.14	5.83	6.47	5.33

Table 4 - Friedman Test Results for the AUC measure

Algorithm	Naive Bayes	Bayes Net	IBK	JRip	One Rule	Decison Stump	J48	MLP	SMO	Meta 1	Meta 2	Meta 3
NaiveBayes	-	□	▲	▲	▲	▲	△	△	▲	□	▲	▲
Bayes Net		-	▲	▲	▲	▲	▲	△	▲	□	▲	▲
IBK			-	△	▲	△	□	□	△	▼	△	△
Jrip				-	▲	△	□	▼	□	▼	△	□
One Rule					-	□	▼	▼	▼	▼	□	▼
Decision Stump						-	▼	▼	□	▼	□	□
J48							-	□	△	▼	△	△
MLP								-	▲	□	▲	▲
SMO									-	▼	△	△
Meta 1										-	▲	▲
Meta 2											-	□
Meta 3												-

indicates that the algorithm from the line was worse (significantly) than the algorithm from the respective column. By symmetry, the lower triangle in this table has opposite results to the upper triangle, and has been omitted for clarity.

As the standard deviation can be seen as a measure of algorithm stability to small variations in the training dataset, the Friedman test has been also performed to evaluate significant differences among standard deviations for all algorithms, including *stacking*. The Friedman test found significant differences among standard deviations; however, the *post-hoc* test[18] could not detect those differences using α=0.05. Under these considerations, the results permit us to affirm that, for the assessed datasets, the use of *stacking* neither improves nor worsens algorithm stability.

CONCLUSION

In this study, an evaluation of *stacking* has been performed using three different paradigms to generate the meta-classifier: statistical-based, induction of simple rules and one-level decision tree, extending a previous work[5], in which *stacking* used only a statistical algorithm as its meta-classifier. The results here permit us to conclude that stacking using a statistical algorithm as meta-classifier has a significantly better performance than induction of simple rules as well as simple trees.

The associated literature recommends the use of simple algorithms as meta-classifiers[1]. However, this study indicates the algorithm used as the meta-classifier must have a certain sophistication degree in order to allow it to adequately represent the concepts from level-0, something rules containing one single attribute or one-level trees clearly do not. Therefore, it is possible to indicate that future studies should use an algorithm of controlled complexity as the meta-classifier – for example, a decision tree with decreasing levels of pruning or a neural network with increasing number of neurons, synapses or training cycles – in order to identify, if possible, the complexity level required to compose a good meta-classifier.

REFERENCES

1. Dzeroski S, Zenko B. Is combining classifiers better than selecting the best one? In: Proceedings of the 19th International Conference on Machine Learning. 2002 jul 8-12; Sydney, Australia.
2. Wolpert DH. Stacked Generalization. Neural Netw. 1992;5(2):241-60.
3. Pereira M, Schmitz A. Inteligência artificial e geotecnologias emergentes aplicadas em estudos ecoepidemiológicos de malária no Município de Bragança-Pará, Brasil, no Período de 2006 a 2008. In: Anais do X Workshop de Informática Médica. Congresso da Sociedade Brasileira de Computação. 2010 jul 20-3; Belo Horizonte (MG). p. 1630-40.
4. Pollettini JT, Tinos R, Panico S, Daneluzzi JC, Macedo AA. Vigilância em atenção básica à saúde a partir do uso de relevance feedback para classificação de pacientes em diferentes níveis de cuidado em saúde. In: Anais do IX Workshop de Informática Médica. Congresso da Sociedade Brasileira de Computação. 2009 jul 21-4; Bento Gonçalves (RS). p. 1945-54.
5. Tanwani AK, Afridi J, Shafiq MZ, Farroq M. Guidelines to select machine learning scheme for classification of biomedical datasets. EvoBIO. 2009;1:128-39.
6. Frank A, Asuncion A. UCI Machine Learning Repository. School of information and computer science. Available from: http://archive.ics.uci.edu/ml
7. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques with java implementations. 2a ed. USA: Morgan Kaufmann; 2005.
8. Rish I. An empirical study of the Naive Bayes classifier. In: Proceedings to IJCAI Workshop on empirical methods in artificial intelligence. 2001 aug 4; Settle (USA). p. 41-6.
9. Chickering DM, Heckerman D, Meek C. Large-sample Learning of Bayesian Networks is NP – Hard. J Mach Learn Res. 2004;5:1287-330.
10. Aha DW, Kibler D, Albert MK. Instance based learning algorithms. Machine Learning. Boston: Kluwer Academic Publishers; 1991.
11. Cohen WW. Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning. 1995 jul 9-12; Tahoe City, California (USA). p.115-23.
12. Quinlan JR. C4.5: programs for machine learning. San Francisco: Morgan Kaufmann; 1993.
13. Iba W, Langley P. Induction of One – Level Decision Trees. In Proceedings of the Ninth International Conference on Machine Learning. 1992 jul 1-3; Aberdeen (Scotland).
14. Haykin S. Neural networks: a comprehensive foundation. 2a ed. London: Pearson Education; 1998.
15. Vapnik VN. Statistical learning theory. USA: Wiley Interscience; 1998.
16. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145-59.
17. Friedman M. A comparison of alternative tests of significance for the problem of m rankings. Ann Appl Stat. 1940;11(1):86–92.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B. 1995;57(1): 289–300.