CLASSIFICATION BINARY MODELS FOR BIOMEDICAL DATA: SIMPLE PROBABILISTIC NETWORKS AND LOGISTIC REGRESSION

Anderson ARA[1]
Francisco LOUZADA[2]
Luis Aparecido MILAN[3]

- ABSTRACT: In the biomedical area a critical factor is whether a classification model is accurate enough in order to provide correct classification whether or not a patient has a certain disease. Several techniques may be used in order to accommodate such situation. In this context, Bayesian networks have emerged as a practical classification technology with successful applications in many fields. At the same time, logistic regression is a widely used statistical classification method and evidenced in the literature. In the current paper we focus on investigating the preditive performance of a probabilistic networks in its simple particular case, the so called naive Bayes network, compared to the logistic regression. A systematic simulation study is performed and the procedures are illustrated in some benchmark biomedical data sets.

- KEYWORDS: Binary classification; simple probabilistic networks; naïve Bayes; logistic regression.

1 Introduction

Currently different biomedical types of data, such as medical diagnosis, sequences, protein structures and families, proteomics data, ontologies, gene expression and other experimental data are often collected in research centers. In this plot, classification is an essential task used to predict group membership for

[1] Universidade Federal da Bahia - UFBA, Departamento de Estatística, CEP: 40170-110, Salvador, BA, Brazil. E-mail: *anderson.ara@ufba.br*

[2] Universidade de São Paulo - USP, Instituto de Ciências Matemáticas e Computação, CEP: 13566-590, São Carlos, SP, Brazil. E-mail: *louzada@icmc.usp.br*

[3] Universidade Federal de São Carlos - UFSCAR, Departamento de Estatística, CEP: 13565-905, São Carlos, SP, Brazil. E-mail: *dlam@ufscar.br*

data instances. Therefore binary classification can be considered one particular case on classification and has been successfully applied to wide range of medical problems.

Thus many techniques can be used in the binary classification but methods with high performance are highly required to minimize risks where diagnosis mistakes can cost the life of the patients. In this context, Bayesian networks have emerged as a practical classification technology with successful applications in many fields and provides some advantages such as the ability to combine expert opinion and experimental data (NIELSEN *et al.*, 2009; HECKERMAN *et al.*, 1995).

Otherwise, logistic regression is a widely used statistical method, as evidenced in the literature (KING and ZENG, 2001). Alternatively other techniques are: probit analysis, mathematical programming, expert systems, neural networks, genetic algorithms and others (HAND and HENTLEY, 1997).

Generally, the best technique for all data sets does not exist, but we can compare a set of methods using some statistical criteria. Therefore, the main thrust of this paper is to investigate the ability of probabilistic networks in a simple particular case of naïve Bayes network, and so called simple probabilistic networks, compared to logistic regression. Then we compute a systematic confrontation through simulation and real data analysis involving both methods. The basic idea consists in applying the models to several replicated artificial datasets and some real datasets. Hence study the behavior of the specific statistical performance measures.

We only considered the naive Bayes network and logistic regression classification strategy because they are consolidated casual classification methods.

This paper is organized as follows. In Section 2 the naive Bayes network and logistic regression procedures like that ROC Curve and some performance measures are presented. In Section 3 we present the simulation results with artificial data and some analysis applied in benchmark biomedical real databases. We finish the paper with some final comments in Section 4.

2 Methodology

In this section we expose shortly the procedures of naïve Bayes network and logistic regression and how the ROC curve is applied in both methods. Also we present some statistical performance measures.

2.1 Naïve Bayes network

The naïve Bayes procedure, described by Good (1965), Duda and Hart (1973) and Flach and Lachiche (2004), is based in computing the posterior probability distribution $P(Y|X)$ where $Y = \{y_1, y_2, ..., y_p\}$ is the class variable and $X = \{X_1, X_2, ..., X_k\}$ is a set of attribute variables that explain the domain. However, this classifier has strong independence assumption and this computation is quite feasible. In other words $P(Y = y_i|X) \propto P(Y = y_i) \prod_{j=1}^{k} P(X_j|Y = y_i)$.

Thus predict to the most plausible category through $\arg\max_Y P(Y|X)$. Besides, the naïve Bayes procedure can be interpreted as a simple probabilistic network. The Figure 1 shows naïve Bayes network and a particular case of probabilistic network.

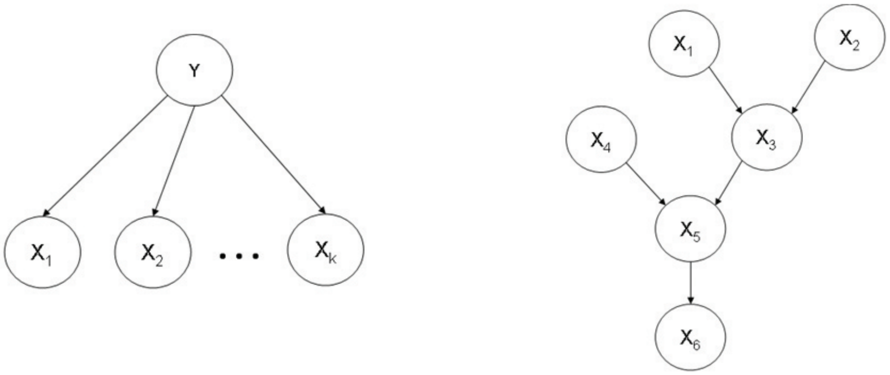

Figure 1 - In the left sub-image we present the traditional structure of naïve Bayes network and in the right sub-image a probabilistic network with six random variables is shown.

2.2 Logistic regression

In a similar way, we consider a set of attribute variables $X = \{X_1, X_2, ..., X_n\}$ and a class variable with binary categories $Y = \{y_1, y_2\}$. Thereby the logistic regression method consists of appointing a linear relation between X and a logit transformation of Y. If we take the y_1 as the category in focus, this model can be represented as $\log\left[\frac{\pi}{1-\pi}\right] = X\beta$ where $\pi = P(Y = y_1)$ and β the vector of coefficients. Hence a possible way to represent this model is $\pi_i = \frac{\exp X_i\beta}{1-\exp X_i\beta}$ where π_i is the probability of the i-th patient belonging to the category of interest. Through specific considerations we can trace a cut-off point used in classification, in other words, set a C point and classify a patient i as a diseased in category y_1 on the study if $\pi_i > C$.

2.3 Some performance measures

A misclassification takes place when the modeling procedure fails to correctly allocate a patient to its true category. Then the modeling procedure misclassification rates can be easily calculated. Thus, to control the misclassification we shall particularly consider the overall correct prediction rate also known as accuracy rate (ACC), but also the sensitivity (SEN) and specificity (SPE). In this context we also consider the Matthews Correlation Coefficient (MCC) a balanced measure which can be used even if the classes are of very different sizes, it returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 an inverse prediction (NIELSEN; RUMI and SALMERÓN, 2009). The performance measures are defined as,

$$ACC = \frac{TP+TN}{TP+TN+FN+FP},\ SEN = \frac{TP}{TP+FP},\ SPE = \frac{TN}{TN+FN}$$

and

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}}$$

where TP is the number of true positive, test positive in actually positive cases, FP is the number of false positive, test positive in actually negative cases, FN is number of false negative, test negative in actually positive cases, TN is the number of true negative, test negative in actually negative cases.

2.4 ROC Curve

The receiver operating characteristic (ROC) curve is an effective method of evaluating the quality or performance of diagnostic tests, and is widely used in several biomedical applications (PARK *et al.*, 2004).

The ROC curve generalizes the notions of sensitivity (SEN) and specificity (SPE) seeking maximum values for sensitivity and specificity measures. The lower misclassification error is guaranteed by higher area of this curve, so the point closer to the left corner can be considered like a cut-off point.

In this context, like often achieved with logistic regression, we can consider the naïve Bayes network for binary classification and apply the ROC curve to define a cut-off point to classify the posterior probability of a patient being assigned to the adequate group.

3 Experimental results

In this section we expose the application results considering the logistic regression and naïve Bayes network in some real and artificial datasets.

As a first set of experiments we consider for Blood Transufion, Breast Cancer, Diabetes and Statlog(Heart) benchmark database.All of these are in the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Table 1 gives a numerical summary of the data sets and their perfomence measures. We can observe a very similar performance.

As a final set of experiments we generated datasets according to a binary random variable indicating presence or absence of a particular disease. Thus, we achieved a comparative evaluate between both methods through a thorough simulation where we consider 399 replications, this number was used by Hall (1986) to construct confidence intervals for the boostrap technique.

Table 1 - Perfomance measures obtained by biomedical real data analysis

						Methods	
DATA	SIZE	ATTRIBUTES	TYPE	RATE	MEASURE	NB	LR
Blood Transufion	748	5	quantitative	24%	ACC	0.69	0.68
					SEN	0.68	0.76
					SPE	0.69	0.65
					MCC	0.32	0.35
Breast Cancer	286	10	qualitative	30%	ACC	0.72	0.74
					SEN	0.66	0.68
					SPE	0.75	0.76
					MCC	0.39	0.42
Diabetes	768	8	quantitative	35%	ACC	0.75	0.77
					SEN	0.73	0.74
					SPE	0.76	0.78
					MCC	0.48	0.51
Statlog(Heart)	270	13	quantitative qualitative	24%	ACC	0.85	0.86
					SEN	0.85	0.86
					SPE	0.85	0.86
					MCC	0.70	0.72

Then we consider a population with one class variable and ten attribute variables, so fixed the samples size at 100, 300, 1000 and 10000 elements. The attribute variables values in X were generated according to Breiman (1998), such distribution of patients without a particular disease has a 10-dimensional normal distribution with mean vector equals to $(0, ..., 0)$ and covariance $4I_{10}$, the distribution of patients with a particular disease has a 10-dimentional normal distribution with mean vector equals to $(\frac{1}{\sqrt{10}}, ..., \frac{1}{\sqrt{10}})$ and covariance I_{10}, where I_{10} is identity matrix of order 10. Also we consider four setups, 50%, 25%,10% and 1% rates of patient with a particular disease.

Overall, four datasets were generated through the rates, hereafter called Setup 1, Setup 2, Setup 3 and Setup 4, respectively. Hence we took four samples with different sizes. For all resamples we fitted the usual logistic regression model and naïve Bayes network. Table 2 shows the performance measures and the 95% confidence intervals based on their resample distributions. The interval results show that in both methods the performance measures are statistically equal, except for MCC measure in the largest sample from Setup 1 where naïve Bayes (NB) appears slightly better than logistic regression (LR), showing a significant improvement on this criterion.

4 Final comments

In this paper we observed a straight approximation between naïve Bayes network and logistic regression in biomedical data results. And statistically we observed equal classification perfomance with slightly naïve Bayes superioty in MCC measure in 50%-50% setup. In general we can say both methods have close performances. Since the principle of parsimony suggests it is better to stick to the simplest model when compared to others with similar performances, in the case of binary classification we can consider the naive Bayes model as a better option than the logistic regression, being easier to implement.

ARA, A.; LOUZADA, F.; MILAN, L. A. Modelos de classificação binária para dados biomédicos: redes probabilísticas simples e regressão logística. *Rev. Bras. Biom.,* Lavras, v.36, n.1, p.48-55, 2018.

- *RESUMO: Na área biomédica, um fator crítico é verificar se um modelo de classificação é preciso o suficiente para fornecer classificação correta se um paciente possui ou não uma determinada doença. Várias técnicas podem ser utilizadas a fim de acomodar tal situação. Neste contexto, as redes probabilísticas, também chamadas de redes Bayesianas, emergiram como uma tecnologia de classificação prática, com aplicações bem sucedidas em muitos campos. Paralelamente, a regressão logística é um método de classificação estatística amplamente utilizado e evidenciado na literatura. No presente trabalho nos concentramos em investigar a capacidade preditiva das redes probabilísticas em seu caso mais simples, a chamada rede de Naïve Bayes, em comparação com a regressão logística. Um estudo de simulação sistemática é realizada, bem como os procedimentos são ilustrados em alguns conjuntos de dados biomédicos de referência.*

- *PALAVRAS-CHAVE: Classificação binária; redes probabilísticas simples; naïve Bayes; regressão logística.*

References

BREIMAN, L. Arcing classifiers. *The Annals of Statistics*, v.26, p.801-849, 1998.

DUDA, R. O.; HART, P. E. *Pattern Classification and Scene Analysis.* New York: JWS, 1973.

FLACH, P. A.; LACHICHE, N. Naive Bayesian Classification of Structured Data. *Machine Learning*, v.57, n.3, p.233-269, 2004.

GOOD, I. J. *The Estimation of Probabilities. An essay on modern Bayesian methods*. Cambridge: The MIT University Press, 1965.

HALL, P. On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics*, v.14, p.1453-1462, 1986.

HAND, D. J.; HENTLEY, W. E. Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of Royal Statistical Society: Series A*, v.160, 523-541, 1997.

HECKERMAN, D.; GEIGER, D.; CHICKERING, D. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, v.20, p.197-243, 1995.

KING, G.; ZENG, L. Logistic Regression in Rare Events Data. *Political analysis*, v.9, n.2, p.137-163, 2001.

NIELSEN, J. D.; RUMI, R. and SALMERÓN, A. Supervised classification using probabilistic decision graphs. *Computational Statistics and Data Analysis*, v.53, p.1299-1311, 2009.

PARK, S. H.; GOO, J. M.; JO, C. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean Journal of Radiology*, v. 5, n. 1, p. 11-18, 2004.

Received on 15.04.2016.

Approved after revised on 06.06.2017.

Table 2 - Perfomance measures obtained by systematic simulation for all setups

		n=100		n=300		n=1000		n=10000	
DESIGNS		NB	LR	NB	LR	NB	LR	NB	LR
Setup 1	ACC	0.72(0.66; 0.79)	0.73(0.66;0.80)	0.68(0.64;0.72)	0.68(0.64;0.72)	0.66(0.64;0.69)	0.66(0.64;0.69)	0.67(0.66;0.67)	0.65(0.65;0.66)
	SEN	0.72(0.61; 0.84)	0.73(0.62;0.84)	0.68(0.59;0.77)	0.69(0.60;0.78)	0.66(0.60;0.72)	0.66(0.61;0.72)	0.70(0.67;0.71)	0.67(0.64;0.69)
	SPE	0.72(0.60; 0.83)	0.73(0.60;0.84)	0.67(0.59;0.76)	0.68(0.60;0.76)	0.66(0.60;0.72)	0.66(0.60;0.71)	0.63(0.62;0.66)	0.64(0.62;0.67)
	MCC	0.44(0.32; 0.58)	0.46(0.32;0.60)	0.36(0.28;0.45)	0.37(0.29;0.45)	0.32(0.28;0.37)	0.32(0.28;0.37)	0.33(0.33;0.34)	0.31(0.30;0.31)
Setup 2	ACC	0.74(0.66; 0.82)	0.75(0.65;0.83)	0.69(0.63;0.75)	0.69(0.63;0.75)	0.66(0.63;0.70)	0.66(0.63;0.70)	0.65(0.65;0.66)	0.65(0.64;0.66)
	SEN	0.75(0.62; 0.88)	0.76(0.64;0.87)	0.70(0.61;0.79)	0.70(0.61;0.78)	0.67(0.61;0.73)	0.67(0.61;0.73)	0.68(0.66;0.69)	0.68(0.65;0.70)
	SPE	0.74(0.61; 0.85)	0.74(0.62;0.86)	0.68(0.60;0.77)	0.69(0.59;0.78)	0.66(0.60;0.72)	0.66(0.60;0.72)	0.64(0.63;0.67)	0.64(0.62;0.66)
	MCC	0.43(0.30; 0.58)	0.45(0.30;0.59)	0.34(0.25;0.42)	0.34(0.26;0.42)	0.29(0.25;0.34)	0.29(0.25;0.34)	0.28(0.28;0.29)	0.27(0.27;0.28)
Setup 3	ACC	0.78(0.67; 0.90)	0.80(0.68;0.91)	0.72(0.63;0.79)	0.72(0.63;0.81)	0.68(0.61;0.74)	0.68(0.62;0.75)	0.64(0.62;0.66)	0.64(0.62;0.66)
	SEN	0.81(0.67; 1.00)	0.82(0.67;1.00)	0.73(0.62;0.85)	0.74(0.63;0.85)	0.70(0.62;0.77)	0.70(0.62;0.78)	0.68(0.65;0.70)	0.68(0.65;0.70)
	SPE	0.78(0.66; 0.91)	0.80(0.67;0.92)	0.72(0.61;0.80)	0.71(0.62;0.82)	0.68(0.60;0.75)	0.68(0.60;0.75)	0.63(0.61;0.66)	0.64(0.61;0.67)
	MCC	0.41(0.26; 0.57)	0.44(0.28;0.62)	0.29(0.20;0.38)	0.29(0.20;0.38)	0.23(0.19;0.28)	0.23(0.19;0.28)	0.19(0.18;0.20)	0.19(0.19;0.20)
Setup 4	ACC	–	–	0.79(0.63;0.93)	0.80(0.64;0.94)	0.75(0.61;0.88)	0.75(0.62;0.88)	0.68(0.60;0.74)	0.67(0.60;0.74)
	SEN	–	–	0.82(0.66;1.00)	0.81(0.62;1.00)	0.77(0.62;0.90)	0.77(0.62;0.91)	0.68(0.60;0.76)	0.68(0.61;0.76)
	SPE	–	–	0.79(0.62;0.93)	0.80(0.64;0.94)	0.75(0.61;0.88)	0.75(0.62;0.88)	0.68(0.60;0.74)	0.67(0.60;0.74)
	MCC	–	–	0.21(0.11;0.33)	0.21(0.12;0.34)	0.13(0.08;0.19)	0.13(0.08;0.19)	0.08(0.06;0.09)	0.08(0.06;0.09)