Simple Bayesian Classifier Applied to Learning

Byron Oviedo[1(✉)], Cristian Zambrano-Vega[1], Joffre León-Acurio[2], and Alina Martinez[1]

[1] Faculty of Engineering Sciences, Quevedo State Technical University,
Quevedo, Ecuador
`{boviedo,czambrano,amartinez}@uteq.edu.ec`
[2] Babahoyo Technical University, Babahoyo, Ecuador
`http://www.uteq.edu.ec, https://www.utb.edu.ec/`

Abstract. In this article, we propose the use of a new simple Bayesian classifier (SBND) that quickly learns a Markov boundary of the class variable and a network structure relating class variables and the said boundary. This model is compared with other Bayesian classifiers, then experimental tests are carried out for which 31 well-known ICU databases and two bases of artificial variables have been used. With these databases we compare the results obtained by such algorithms studied in the state of the art such as Naive Bayes, TAN, BAN, RPDag, CRPDag, SBND and combinations with different metrics such as K2, BIC, Akaike, BDEu. The experimental work was done in Elvira software.

Keywords: Bayesian networks educational analysis
Bayesian classifier · Educational analysis

1 Introduction

Learning can be defined as 'any process, through which a system improves its efficiency' [6]. The ability to learn is considered a central feature of intelligent systems [9], and this is why a lot of effort and dedication have been put on the research and development of this topic. The development of knowledge based systems has motivated research in the area of learning with the aim of automatizing the knowledge acquisition process, what is considered one of the main problems in the building of these systems. For some time, algorithms for learning without Bayesian networks restrictions, especially those based on the metric + search paradigm have been considered inadequate for competitive construction of classifiers based on Bayesian networks [1]. This perception is being changed due to the development of generic networks learning methods, which are very competitive [1]. Bayesian networks (without structural restrictions of any kind) can also be used for classifying. In this case, classifiers are referred to as non-restricted Bayesian net-works. These will be used in this paper too. Any Bayesian network can be used in supervised classification, for which it is enough to use

M. Botto-Tobar et al. (Eds.): CITT 2018, CCIS 895, pp. 399–409, 2019.
https://doi.org/10.1007/978-3-030-05532-5_29

the Markov blanket of the case variable. It is necessary to consider that a non-restricted Bayesian classifier has a higher expressive power than a structurally restricted model [8].

In this paper we present a new classifier, which we have called Simple Bayesian classifier, consisting in a generic Bayesian network, but learned from a voracious techniques.

2 Simple Bayesian Classifier

SBND is a new simple Bayesian classifier designed to simplify this activity. To start using this classifier, we will need a PARENTS function, which, given a variable X_i and a set of candidates, can calculate the best parents set in X_i among that set of candidates. The parents set is returned in Π_i and at the same time gives back a numerical value constituting the Score of this variable, given that set of parents meas-ured by a Bayesian Score.

This PARENTS function is the one making a heuristic search for the best set of parents among a set of candidates and this occurs by adding and removing parents as long as the score improves. The idea is to start introducing C as a root node in the Bayesian network $\mathscr{B}$ and keep a set of nodes **X'** of the attributes already introduced in the net (initially empty) [13].

Different Score metrics $Score(X_i, \mathbf{A}|\mathscr{D})$ measuring the suitability of **A** as X_i parents set can be used (these metrics can be BDEu, BIC, K2 o Akaike).

Since we assume to have a procedure PARENTS(X_i, CANDIDATOS, Π_i), which calculates the best set of parents Π_i from X_i using the selected metric, and returns the value of this optimum metric, when implementing this function we will have a voracious algorithm, which starts with an empty Π_i and keeps adding and removing form Π_i the variable producing the highest metric increase, until there is not any possible improvement. In these conditions the values for each $X_i \in \mathbf{X} \setminus \mathbf{X'}$ variable is calculated:

$$Infor(X_i, C) = \text{PADRES}(X_i, \mathbf{X}' \cup \{C\}, \Pi_i) - \text{PADRES}(X_i, \mathbf{X}', \Pi_i')$$

Infor $Infor(X_i, C)$ calculates the differences among the best X_i metrics with parents set chosen between $\mathbf{X}'$ including C and without including C in the candidates. Intuitively, it is a measure of X_i and C conditional dependency, given the already included variables. This value is always theoretically higher or equal to zero, but it could be negative since the best parents set is calculated approximately [13].

Once this value has been calculated for each $X_i \in \mathbf{X} \setminus \mathbf{X}'$ variable, $X_{max} = \arg \text{máx}_{X_i \in \mathbf{X} \setminus \mathbf{X}'} Infor(X_i, C)$ is selected. This would be the variable providing most information about the C class according to the already introduced variables. If $Infor(X_{max}, C) > 0$, then this variable provides additional information about C and is inserted in the network and in $\mathbf{X}'$. Its parents set is calculated with PARENTS($X_{max}, \mathbf{X}' \cup \{C\}, \Pi_i$). In theory, $C \in \Pi_i$, always, since otherwise $Infor(X_{max}, C) = 0$, although due to the voracious nature of the procedure, $C \notin \Pi_i$, what is a remote possibility.

In other words the variable giving most information is added to the network provided that the information is positive, and is considered, as well as its parents set, the best parents set provided by this function. Since the information is positive, the class variable should be supposed to be included in the parents set. The algorithm ends if the function $Infor(X_{max}, C) \leq 0$ [13].

The main characteristics of this classifier are:

- Learns an arbitrary Bayesian network with a subset of initial variables directly influencing this variable. In this sense, it can be considered an algorithm, which calculates a Markov boundary, because it intends to obtain a set of such variables that, once obtained, the rest of the variables are independent.
- The class variable is always a root node and there are links from this node to the rest of the attributes (except for very few occasions due to the approximate nature of the parents calculation). It is, in this sense, similar to other Bayesian classifiers, where there is always a link from the class to each of the attributes.
- The arrangement of the attributes in based on selecting in a voracious way those providing most information on the class, given the selected attributes. In this way the most relevant attributes are introduced first. Obtaining the network with the best metrics is not based on the space of the attributes order, but in obtaining the maximum information for the class. It can even be some network quality lose in this sense, but the algorithm gains speed [13].

3 Experimentation

In this section experimental tests are carried out through using 31 well - known to ICU databases [11] and two bases including artificial variables. The databases can be seen in Table 1. With these two databases. The results obtained by the algorithms mentioned in the state of art, namely Naive Bayes [15], TAN [7], BAN [3], SBND [12], RPDAG and C-RPDAG [1,9] are compared to other combinations with different metrics, namely K2 [4], BIC [14], Akaike [2], BDEu. These methods build up classifiers constituting generic Bayesian networks equivalent in independence and equivalent in classification. The experimental work was made at Elvira [5].

Table 1 provides a brief description of each database characteristics, including the number of instances, attributes, and the states for the class variable. These data sets have been preprocessed as follows: continuous variables have been discretized using the procedure proposed by [10], and the instances having non definite or missing values were removed. For this pre-processing stage, the results obtained by [1] have been used.

4 Results

The results obtained by each classifier and its combinations with the studied metrics can be observed in Tables 2, 3, 4 and 5 (due to their size, they have been divided in 4).

Table 1. Description of the databases

Database	Instances	Attributes	Classes
adult-d-nm	45222	14	2
australian-d	690	14	2
breast-no-missing	682	10	2
car	1728	6	4
chess	3196	36	2
cleve-no-missing-d	296	13	2
corral-d	128	6	2
crx-no-missing-d	653	15	2
diabetes-d-nm	768	8	2
DNA-nominal	3186	60	3
flare-d	1066	10	2
german-d	1000	20	2
glass2-d	163	9	2
glass-d	214	9	7
heart-d	270	13	2
hepatitis-no-missing-d	80	19	2
iris-d	150	4	3
letter	20000	16	26
lymphography	148	18	4
mofn-3-7-10-d	1324	10	2
nursery	12960	8	5
mushroom	8124	22	2
pima-d	768	8	2
satimage-d	6435	36	6
segment-d	2310	19	7
shuttle-small-d	5800	9	7
soybean-large-no-missing-d	562	35	19
splice.dbc	3190	60	3
vehicle-d-nm	846	18	4
vote	435	16	2
waveform-21-d	5000	21	3

Below some non-parametric tests are made of the differences among the different methods in determining the best classifying algorithm, It is important to indicate that the means value for each of the algorithms has been included, as well. The best means is obtained by CRPDAG-BDEu with an 88.354 value,

Table 2. Results with ICU database

Database	SBND BDE	SBND BIC	SBND Ak	SBND K2	TAN	NBayes
adult-d-nm	85.255	85.213	85.405	85.963	85.295	83.090
australian-d	86.667	85.797	86.232	86.667	85.362	85.652
breast-no-missing	97.370	97.662	96.053	97.515	96.196	97.662
car	94.097	85.067	93.635	94.097	94.214	85.299
chess	97.496	96.214	97.653	97.121	91.989	87.765
cleve-no-missing-d	82.115	82.126	80.759	82.103	79.724	82.414
corral-d	99.231	100.0	99.167	99.231	99.231	85.962
crx-no-missing-d	85.916	86.375	86.671	87.140	86.974	86.678
diabetes-d-nm	78.780	79.040	78.908	79.429	77.997	77.341
DNA-nominal	96.171	96.203	93.943	95.983	94.822	95.418
flare-d	82.268	82.268	82.738	82.268	83.018	80.395
german-d	74.0	73.6	71.8	72.3	73.6	75.5
glass2-d	85.882	81.066	84.007	85.882	85.257	83.493
glass-d	73.355	64.545	71.494	69.134	73.852	73.853
heart-d	81.111	81.111	82.593	82.963	82.963	83.333
hepatitis-no-missing-d	90.0	90.0	90.0	86.25	86.25	85.0
iris-d	94.0	95.333	95.333	94.667	94.0	94.667
letter	81.565	74.015	85.655	85.36	86.320	73.6
letter-d	84.320	74.365	84.81	84.615	85.775	73.935
lymphography	77.571	81.620	80.381	80.381	79.048	81.762
mofn-3-7-10-d	92.522	90.790	100.0	93.501	91.237	85.425
nursery	91.890	91.705	97.469	94.537	92.261	90.332
mushroom	100.0	98.523	99.274	100.0	99.963	95.495
pima-d	79.166	79.560	78.259	79.429	79.038	77.994
satimage-d	85.144	82.316	86.807	85.812	88.252	82.440
segment-d	94.459	93.550	95.022	94.372	95.151	92.208
shuttle-small-d	99.741	99.552	99.534	99.759	99.069	99.052
soybean-large-no-missing-d	91.278	84.496	90.918	93.409	94.298	91.269
splice.dbc	96.238	96.270	91.944	96.364	94.796	95.454
vehicle-d-nm	65.849	65.013	70.451	64.892	69.986	61.829
vote	94.715	94.952	95.640	95.174	94.498	90.338
vote-no-missing	95.396	95.375	94.704	95.169	94.244	90.095
waveform-21-d	82.84	82.34	81.9	83.5	83.1	81.84
media	87.770	86.244	88.156	88.030	87.810	85.048

followed by SBND Akaike with 88.156. The basic non parametric test used is Friedman, since it has more than 2 associated samples.

The null hypothesis (H_0) being contrasted is that the answers associated to each of the treatments have the same probability distribution or distributions

Table 3. Results with ICU database

Database	BAN Learning BDEu	BAN Learning BIC	BAN Learning K2
adult-d-nm	85.534	85.472	84.906
australian-d	84.638	86.812	84.203
breast-no-missing	97.662	97.662	97.664
car	93.517	85.414	94.213
chess	96.151	95.745	96.995
cleve-no-missing-d	81.069	79.713	78.701
corral-d	100.0	100.0	98.462
crx-no-missing-d	86.063	86.986	84.681
diabetes-d-nm	78.387	78.389	78.127
DNA-nominal	95.292	95.418	93.158
flare-d	82.830	82.831	83.298
german-d	75.3	75.3	74.1
glass2-d	85.882	85.221	85.846
glass-d	72.013	76.190	74.762
heart-d	82.963	82.222	82.963
hepatitis-no-missing-d	88.75	87.5	88.75
iris-d	94.0	94.0	94.0
letter	84.715	74.880	87.965
letter-d	85.56	77.34	86.945
lymphography	85.0	82.333	74.857
mofn-3-7-10-d	87.617	90.865	93.804
nursery	91.860	91.883	94.892
mushroom	100.0	100.0	100.0
pima-d	78.775	78.780	79.040
satimage-d	88.361	85.175	87.506
segment-d	95.281	92.900	95.195
shuttle-small-d	99.052	99.776	99.741
soybean-large-no-missing-d	93.424	93.418	89.860
splice	95.329	95.705	94.107
vehicle-d-nm	70.689	70.102	69.384
vote	94.493	93.811	92.659
vote-no-missing	93.092	93.552	92.875
waveform-21-d	82.94	82.62	83.5
media	88.090	87.211	87.791

with the same means against the alternative hypothesis stating that at least the distribution of one of the means differs from the others.

The values that will be used in these tests can be seen in Table 6, where the average order of the algorithms are presented. The best performance is shown by SBND K2 algorithm.

Table 4. Results with ICU database

Database	RPDag-BDEu	RPDag-BIC	RPDag-K2
adult-d-nm	85.748	85.576	85.339
australian-d	85.797	85.362	85.797
breast-no-missing	97.662	97.662	97.662
car	93.228	85.878	94.040
chess	97.152	94.931	96.871
cleve-no-missing-d	82.115	81.770	80.425
corral-d	100.0	100.0	100.0
crx-no-missing-d	86.371	86.068	84.387
diabetes-d-nm	79.429	79.040	79.170
DNA-nominal	95.857	96.360	95.450
flare-d	82.268	82.268	82.268
german-d	74.4	74.2	73.8
glass2-d	84.632	84.044	82.169
glass-d	67.294	65.823	73.788
heart-d	80.370	81.481	82.963
hepatitis-no-missing-d	87.5	90.0	86.25
iris-d	96.0	95.333	94.667
letter	83.185	74.835	86.65
letter-d	86.085	74.87	86.325
lymphography	76.905	75.524	74.952
mofn-3-7-10-d	100.0	93.808	96.829
mushroom	100.0	100.0	100.0
nursery	93.465	91.312	94.792
pima-d	79.299	79.299	79.301
satimage-d	84.911	79.285	84.911
segment-d	94.199	94.589	95.325
shuttle-small-d	99.690	94.862	99.534
soybean-large-no-missing-d	89.148	86.096	93.064
splice	95.956	96.238	96.363
vehicle-d-nm	64.902	61.584	64.066
vote	94.720	94.947	95.185
vote-no-missing	94.709	95.153	93.092
waveform-21-d	79.98	81.06	83.320
media	87.666	86.038	87.841

Friedman test's results are shown in Table 7, where a value lower than 0.05 is seen, thus rejecting the null hypothesis and it is determined that the differences measure the statistically significant distributions of the different methods.

When the differences detected are significant, Holm's test is applied for comparing the control algorithm (the best classified) to the rest. Holm's is a multiple

Table 5. Results with ICU database

Database	CRPDag-BDEu	CRPDag-BIC	CRPDag-K2
adult-d-nm	85.257	85.463	85.372
australian-d	86.667	86.232	84.348
breast-no-missing	97.662	97.662	97.664
car	93.228	85.878	94.040
chess	96.621	95.713	96.277
cleve-no-missing-d	81.747	81.057	78.724
corral-d	100.0	100.0	100.0
crx-no-missing-d	86.981	87.135	83.308
diabetes-d-nm	78.387	77.996	78.127
DNA-nominal	96.422	96.202	83.519
flare-d	83.020	82.830	82.738
german-d	73.4	74.1	74.0
glass2-d	86.471	85.221	85.882
glass-d	74.329	70.519	73.377
heart-d	82.593	82.222	82.963
hepatitis-no-missing-d	90.0	83.75	86.25
iris-d	94.0	94.0	94.0
letter	83.845	75.135	87.01
letter-d	86.55	77.34	87.195
lymphography	76.905	80.333	76.333
mofn-3-7-10-d	100.0	93.808	96.829
mushroom	100.0	100.0	100.0
nursery	93.465	91.312	94.792
pima-d	78.775	78.910	78.910
satimage-d	87.553	85.175	86.667
segment-d	95.325	92.900	94.978
shuttle-small-d	99.741	99.707	99.707
soybean-large-no-missing-d	89.859	90.025	93.590
splice	96.332	96.332	90.157
vehicle-d-nm	70.929	72.228	71.517
vote	93.811	93.346	93.351
vote-no-missing	92.854	92.860	93.330
waveform-21-d	82.94	82.74	83.46
media	88.354	86.913	87.528

Table 6. Average score of the algorithms

Algorithm	Ranking
SBND BDE	7.676470588235294
SBND BIC	9.264705882352944
SBND Akaike	7.73529411764706
SBND K2	5.955882352941175
TAN	8.249999999999998
NaiveBayes	11.382352941176473
BAN Learning BDEu	7.058823529411763
BAN Learning BIC	7.955882352941175
BAN Learning K2	7.823529411764705
RPDag Learning BDEu	7.661764705882354
RPDag Learning BIC	9.20588235294118
RPDag Learning K2	7.191176470588233
CRPDag Learning BDEu	6.2647058823529385
CRPDag Learning BIC	8.823529411764708
CRPDag Learning K2	7.749999999999998

Table 7. Friedman test's results

Test	P Value	Hypothesis
Friedman	1,542E−4	Rejected

comparison test, by means of which we confront SBND with K2, the best classifying value, with the rest of the algorithms.

Table 8 shows Holm test's results for 0.05 significance level and Table 9 for 0.10 significance level.

In the first place it was considered $\alpha = 0.05$. P values in Holm test is $P \leq 0.0045$. This value is compared to the rest of the algorithms based on the right column of Table 8. It can be observed that this algorithm is significantly better than Naive Bayes, SBND BIC, RPDag Learning BIC and there are no significant differences with the rest of the algorithms.

In the second place, $\alpha = 0.10$ significance level is considered, P value in Holms test is $P \leq 0.01$. With this value multiple comparisons with the values in the right column of Table 9 are made. It can be determined that our control algorithm SBND with K2 classifies better than Naive Bayes, SBND BIC, RPDag Learning BIC, CRPDag Learning BIC algorithms and there are no significant differences with the rest of the algorithms.

Table 8. Holm Table for $\alpha = 0.05$

i	Algorithm	$z = (R_0 - R_i)/SE$	p	Holm
14	NaiveBayes	5.0029586834427615	5.645704442401309E-7	0.0035714285714285718
13	SBND BIC	3.0505845630748563	0.0022839635380862903	0.0038461538461538464
12	RPDag Learning BIC	2.9963519486201924	0.0027323088004595057	0.004166666666666667
11	CRPDag Learning BIC	2.643839954664876	0.00819714047525547	0.004545454545454546
10	TAN	2.1150719637318978	0.03442381432538883	0.005
9	BAN Learning BIC	1.8439088914585775	0.06519641907813004	0.005555555555555556
8	BAN Learning K2	1.7218855089355838	0.08509026052283775	0.00625
7	CRPDag Learning K2	1.6540947408672533	0.09810826450210172	0.0071428571428571435
6	SBND Akaike	1.6405365872535898	0.10089364763218327	0.008333333333333333
5	SBND BDE	1.586303972798925	0.1126703715408176	0.01
4	RPDag Learning BDEu	1.57274581918526	0.11577768575893127	0.0125
3	RPDag Learning K2	1.1388849035479442	0.2547511629904382	0.016666666666666666
2	BAN Learning BDEu	1.0168615210249505	0.3092193106086886	0.025
1	CRPDag Learning BDEu	0.28472122588698523	0.7758577275237244	0.05

Table 9. Holm tabla para $\alpha = 0.10$

i	Algorithm	$z = (R_0 - R_i)/SE$	p	Holm
14	NaiveBayes	5.0029586834427615	$5.645704442401309E$-7	0.0071428571428571435
13	SBND BIC	3.0505845630748563	0.0022839635380862903	0.007692307692307693
12	RPDag Learning BIC	2.9963519486201924	0.0027323088004595057	0.008333333333333333
11	CRPDag Learning BIC	2.643839954664876	0.00819714047525547	0.009090909090909092
10	TAN	2.1150719637318978	0.03442381432538883	0.01
9	BAN Learning BIC	1.8439088914585775	0.06519641907813004	0.011111111111111112
8	BAN Learning K2	1.7218855089355838	0.08509026052283775	0.0125
7	CRPDag Learning K2	1.6540947408672533	0.09810826450210172	0.014285714285714287
6	SBND Akaike	1.6405365872535898	0.10089364763218327	0.016666666666666666
5	SBND BDE	1.586303972798925	0.1126703715408176	0.02
4	RPDag Learning BDEu	1.57274581918526	0.11577768575893127	0.025
3	RPDag Learning K2	1.1388849035479442	0.2547511629904382	0.03333333333333333
2	BAN Learning BDEu	1.0168615210249505	0.3092193106086886	0.05
1	CRPDag Learning BDEu	0.28472122588698523	0.7758577275237244	0.1

5 Conclusions

In this article we have introduced a Bayesian classifier known as SBND which is based in quickly obtaining an easy to learn and very competitive Markov's boundary. This classifiers is fast to learn and very competitive as compared to other classifiers of the state of art. Various experiments were made using 31 well known in the ICU databases and two bases of artificial variables.

SBND classifier's performance in some examples is dependent on the metric being used. With BIC the result is not good, Akaike gives good results in reference to the means, and K2 shows good results in non-parametric tests.

For future research, it is important to include the costs of wrong classifications in the problem, since a false positive is not the same as a false negative. If the cost of a false negative was considered better than that of a false positive, it could be detected that more students would drop out, although the number of students at risk of abandoning would increase.

References

1. Acid, S., De Campos, L., Castellano, J.: Learning Bayesian network classifiers: searching in a space of partially directed acyclic graphs. Mach. Learn. **59**(3), 213–235 (2005)
2. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**, 716–723 (1974)
3. Cheng, J., Russell, G.: Comparing Bayesian network classifiers. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, pp. 101–108. Morgan Kaufmann Publishers Inc. (1999)
4. Cooper, G., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**(4), 309–347 (1992)
5. Elvira, C.: Elvira: an environment for probabilistic graphical models. In: Gámez, J., Salmerón, A. (eds.) Proceedings of the 1st European Workshop on Probabilistic Graphical Models, pp. 222–230 (2002)
6. Felgaer, P., Britos, P., Sicre, J., Servetto, A., García-Martínez, R., Perichinsky, G.: Optimización de redes bayesianas basada en técnicas de aprendizaje por instrucción. In: Proceedings del VIII Congreso Argentino de Ciencias de la Computación, vol. 1687 (2003)
7. Friedman, N., Michal, L., Iftach, N., Dana, P.: Using Bayesian networks to analyze expression data. Comput. Biol. **7**(3–4), 601–620 (2000)
8. García, F.: Modelos Bayesianos para la clasificación supervisada: aplicaciones al análisis de datos de expresión genética, Tesis Doctoral, Universidad de Granada (2009)
9. García-Martínez, R., Borrajo, D.: An integrated approach of learning, planning, and execution. J. Intell. Robot. Syst. **29**(1), 47–78 (2000)
10. Irani, K., Jie, C., Usama, F., Zhaogang, Q.: Applying machine learning to semiconductor manufacturing. iEEE Expert **8**(1), 41–47 (1993)
11. Lichman, M.: UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences (2013). http://archive.ics.uci.edu/ml
12. Oviedo, B., Moral, S., Puris, A.: A hierarchical clustering method: applications to educational data. Intell. Data Anal. **20**(4), 933–951 (2016)
13. Oviedo Bayas, B.W.: Modelos gráficos probabilisticos aplicados a la predicción del rendimiento en educación (2016)
14. Schwarz, G.: Estimating the dimension of a model. Ann. Stat. **6**(2), 461–464 (1978)
15. Webb, G.I., Pazzani, M.J.: Adjusted probability Naive Bayesian induction. In: Antoniou, G., Slaney, J. (eds.) AI 1998. LNCS, vol. 1502, pp. 285–295. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0095060