One Dependence Augmented Naive Bayes

Juan Bernabé Moreno
University of Granada, Department of Computer Science and Artificial Intelligence

Abstract— **This work subjects the article about the suggested improvement for the Naïve Bayes Classification Algorithm to a critical review. The algorithm is described and placed within the classifiers taxonomy. The pre-processing techniques and the data sets employed for the experimental part are commented. The goodies and the baddies are discussed and a benchmark presented, where the algorithm is compared with other algorithms intended to improve the Naïve Bayes as well.**

***Index Terms*— ODANB, TAN, SBN**

I. INTRODUCTION

The One Dependency Augmented Naïve Bayes classifier proposed by Liang Xiao Jiang et all [40] aims at improving the performance of the know Naïve Bayes algorithm by relaxing the conditional independency assumption

The ODANB classifier, like other Bayes Network classifiers has two separate components for the BN definition:

- *Structural*: a DAG *g* representing the (in)dependencies among the n+1 one-dimensional variables
- *Parametrical*: set of local probability distributions for *g*

And other two components for the making a classifier out of this BN:

- A *learning algorithm* to build up a BN s out of the *data set* d, $P(C, X_1,\ldots,X_n)$
- An *inference algorithm* to compute the available evidence for the object to be classified $P(C|Ev)$

The subsequent section provides a description of the ODANB algorithm in pseudo-code. Then, a classification of the ODANB according to several criteria is given in order to find the best place for this algorithm among the BN classifiers. The data pre-processing techniques employed in the experiments are briefly described, as well as the data sets used. The next section comments on the benchmarking results and the other *to-beat* algorithms, to end with a conclusion.

II. ALGORITHM

```
Foreach instance to classify
max ← 0
cOutput ← 1
product ← 1
  foreach c in C
   for i ← 1 to n do
    for j ← 1 to n, j <> i do
        Product *= PMutualCondInfo (Ai, Aip, c)
    end for
   end for
   if max < Product then
      max ← Product
      cOutput ← c
   end if
  end foreach
  end Foreach
```

```
Function calcAverage
average ← 0
 for i ← 1 to n do
  for j ← 1 to n, j <> i do
    average += calcIP (Ai;Aj | C)
  end for
 end fot
 average ← average /n(n-1)
 return average
End Function
```

```
Function calcIP
max ← 0
m ← 0
 for i ← 1 to n do
  for j ← 1 to n, j <> i do
   for c ← 1 to n do
      maxN = P(ai,aj,cc) * log P(ai,aj,cc)
             *P(cc)/P(ai,cc)P(aj,cc)
      If maxN > max then
        max ← maxN
        m ← j
      End If
   end for
  end for
 end fot
 return max
End Function
```

```
Function PMutualCondInfo (Ai, Aip, C)
  average ← calcAverage
  ip ← calcIp (Ai, Aip, C)
```

```
  if ip < average then
    return P(Ai|Am,C)
  else
    return P(Ai| C)
  end If
End Function
```

III. Classification algorithm type

All BN classifiers can be viewed as a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items (referred to as traits, variables, characters, etc) and based on a training set of previously labeled items. In other words, the statistical information inherent in the data is employed for the classification and thus, Bayesian Networks classifiers are a kind of *Statistical classifiers* (like k-Neighbors, etc)

According to the way the classifier represents the knowledge, the ODANB algorithm, like all Bayesian classification methods are *probabilistic graphical models*

According to the type of learning, Bayesian Network classifiers can be considered *example-based learners*

According to the way the required probability distribution is provided, BN classifiers can be considered *supervised classifiers*, if one subject matter expert provides the probabilities, or rather *unsupervised*, it probabilities are computed using the information stored in a data base

IV. Bayesian Network type

The kind of Bayesian Network (BN) retrieved by the algorithm is a so called Augmented Naïve BN, characterized mainly by the points bellow:

1) All attributes have certain influence on the class

2) The conditional dependency assumption is relaxed (certain attributes have been added a parent)

V. Pre-Processing Techniques

The author mentions following data pre-processing techniques applied to the data before running the ODANB algorithm:

ReplaceMissingValues: this filter will scan all (or selected) nominal and numerical attributes and replace missing values with the modes and mean.

Discretization: this filter is designed to convert numerical attributes into nominal ones; however the unsupervised version does not take class information into account when grouping instances together. There is always a risk that distinctions between the different instances in relation to the class can be wiped out when using such a filter.

VI. Databases employed

The set of databases employed to benchmark the ODANB algorithm with other known BN algorithms is the one recommended by Weka out of the 171 Data sets maintained by the UCI Machine Learning Repository [1]

The concrete UCI data sets are listed bellow:

- Steel annealing data [2]
- Nominal audiology dataset from Baylor [3]
- Automobile(From 1985 Ward's Automotive Yearbook) [4]
- Balance scale weight & distance database [5]
- Breast Cancer Data (Restricted Access) [6]
- Diagnostic Wisconsin Breast Cancer Database [7]
- Horse Colic [8]
- Credit Approval [9]
- Statlog (German Credit Data) [10]
- Diabetes [11]
- Glass Identification (This data consists of 640 black and white face images of people taken with varying pose (straight, left, right, up), expression (neutral, happy, sad, angry), eyes (wearing sunglasses or not), and size) [12]
- Heart Disease (Data for classifying if patients will survive for at least one year after a heart attack) [13]
- SPECT Heart [14]
- Statlog (Heart) [15]
- Hepatitis [16]
- Thyroid Disease [17]
- Ionosphere (Classification of radar returns from the ionosphere) [18]
- Iris [19]
- King Rook versus King Pawn on a7 (usually abbreviated KRKPA7) [20]
- Labor Relations [21]
- Letter Recognition (Goal: Predict which letter-name was spoken) [22]
- Lymphography [23]
- Mushroom (mushrooms described in terms of physical characteristics) [24]
- Primary Tumor [25]
- Image Segmentation (The problem consists of classifying all the blocks of the page layout of a document that has been detected by a segmentation process) [26]
- Connectionist Bench (Sonar, Mines vs. Rocks)(The task is to train a network to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock) [27]
- Molecular Biology (Splice-junction Gene Sequences) [28]
- Vehicle Silhouettes [29]
- Congressional Voting Records 1984 (Classify as Republican or Democrat) [30]
- Speaker independent recognition of the eleven steady state vowels of British English using a specified training set of lpc derived log area ratios [31]
- CART book's waveform domains [32]
- Zoo [33]

VII. BENCHMARKING

The ODANB has been compared with other existing methods that improves the Naïve Bayes and with the Naïve Bayes itself: TAN and SBN.

In recent years, a lot of effort has focused on improving Naïve-Bayesian classifiers, following two general approaches: selecting feature subset and relaxing independence assumptions [37]

The author picked for the benchmarks one algorithm that relies on relaxing the independence assumption and one that selects a subset of features:

Friedman *et al.* (1997) [34] studied the Tree Augmented Naive Bayes, which allows tree-like structures to be used to represent dependencies among attributes. TAN or Tree Augmented Naive Bayes (TAN) outperforms naive Bayes, yet at the same time maintains the computational simplicity (no search involved) and robustness that characterize naive Bayes. The TAN approximates the interactions between attributes by using a tree structure imposed on the naive Bayesian structure.

Langley and Sage (1994) use forward selection to find a good subset of attributes, then use this subset to construct a *selective Bayesian classifier* (ie, a Naïve-Bayesian classifier over only these variables). [36]

The results of the comparison prove that the ODANB outperforms the other methods.

The comparison criteria that have been introduced are

- Ranking performance. The intuitive idea of ranking for an instance is how far-off this instance is from the class. AUC is introduced to measure it (the area under the Receiver Operating Characteristics curve)
- Accuracy of prediction (measures defined from the confusion matrix outputs)

The table bellow recaps the benchmarked algorithms accuracy for each dataset consider. In each row in bold the best performing algorithm:

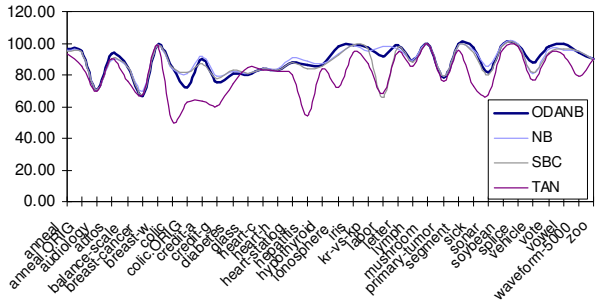
Datasets	ODANB	NB	SBC	TAN
anneal	96.55	94.32	96.88	**96.66**
anneal.ORIG	**90.31**	87.53	88.75	87.98
audiology	62.27	71.23	**76.01**	75.16
autos	**78.55**	64.83	67.71	76.07
balance-scale	**91.36**	91.36	91.36	86.08
breast-cancer	69.61	72.06	**73.45**	66.82
breast-w	96.99	**97.28**	96.42	96.71
colic	81.25	78.81	**81.77**	77.18
colic.ORIG	68.76	75.26	75.53	**75.51**
credit-a	82.90	84.78	**85.51**	84.64
credit-g	73.40	76.30	74.10	73.40
diabetes	73.84	75.40	**75.53**	75.13
glass	60.28	**60.32**	57.99	55.71
heart-c	80.46	**84.14**	82.47	77.53
heart-h	79.66	**84.05**	79.00	79.97
heart-statlog	80.00	**83.70**	79.26	81.11
hepatitis	**85.13**	83.79	80.63	83.83
hypothyroid	92.63	92.79	**93.53**	92.79
ionosphere	90.90	90.89	**91.17**	90.60
iris	94.67	94.67	**97.33**	90.67
kr-vs-kp	90.52	87.89	**94.34**	93.18
labor	**90.00**	93.33	77.00	88.00
letter	77.89	70.00	70.57	**80.45**
lymph	82.43	**85.67**	79.00	84.38
mushroom	**99.94**	95.57	99.67	99.77
primary-tumor	44.26	46.89	46.02	**48.37**
segment	**94.20**	88.92	90.43	86.36
sick	**97.59**	96.74	97.59	97.00
sonar	77.02	**77.50**	70.71	71.62
soybean	91.51	92.08	91.79	**93.41**
splice	93.07	95.36	94.76	**95.39**
vehicle	**71.04**	61.82	60.65	69.86
vote	94.04	90.14	**95.18**	93.12
vowel	**91.82**	67.07	68.69	83.43
waveform-5000	81.26	79.96	81.32	**81.52**
zoo	95.18	94.18	93.18	**97.09**

Bellow same table but for the AUC

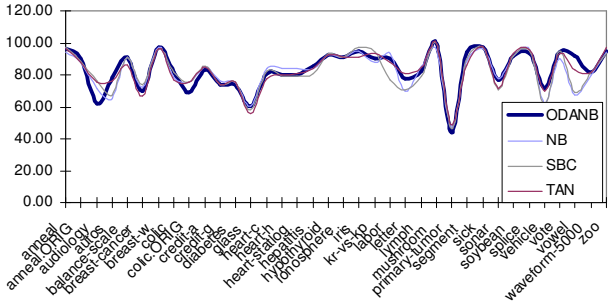
Datasets	ODANB	NB	SBC	TAN
anneal	**96.53**	95.90	94.70	92.97
anneal.ORIG	**95.17**	94.49	94.35	85.42
audiology	70.84	70.96	**70.98**	70.16
autos	**93.07**	89.18	90.43	90.28
balance-scale	**84.46**	84.46	84.46	76.47
breast-cancer	66.57	**69.71**	67.67	67.40
breast-w	99.04	99.19	99.16	**98.74**
colic	84.48	83.71	**84.86**	50.60
colic.ORIG	72.53	80.67	**81.82**	62.89
credit-a	90.19	**92.09**	87.00	63.30
credit-g	75.65	**79.27**	77.41	60.18
diabetes	80.88	82.31	**82.79**	74.18
glass	79.94	80.50	80.97	**84.79**
heart-c	83.85	**84.10**	83.87	82.96
heart-h	83.23	**83.80**	82.83	82.69
heart-statlog	88.18	**91.30**	87.98	80.12
hepatitis	86.04	**88.99**	83.62	53.83
hypothyroid	86.50	**87.37**	85.25	84.03
ionosphere	**97.67**	93.61	92.26	72.05
iris	98.58	98.58	**99.00**	94.17
kr-vs-kp	**97.13**	95.17	96.41	87.21
labor	91.67	**98.33**	65.83	68.33
letter	**98.45**	96.86	97.03	94.50
lymph	89.02	**89.69**	88.14	85.56
mushroom	**100.00**	99.79	99.98	99.87
primary-	78.18	78.85	**78.88**	76.39

tumor				
segment	**99.55**	98.51	98.93	95.35
sick	**97.48**	95.91	94.50	73.25
sonar	81.64	**85.48**	79.89	67.40
soybean	99.46	**99.53**	99.08	96.73
splice	99.05	**99.41**	99.14	97.72
vehicle	**87.97**	80.81	81.31	76.86
vote	**98.16**	96.56	94.26	93.49
vowel	**99.49**	95.81	96.12	92.33
waveform-5000	94.38	**95.27**	95.12	78.90
zoo	**89.88**	89.88	89.06	89.88

Following graphic plots the benchmarked algorithms accuracy for each dataset consider.



The AUC comparison for the benchmarked algorithms is plotted in the figure bellow:



The experimental results measured in terms of accuracy and AUC prove that ODANB has better performance than the other algorithms used to compare.

On the other hand, the computational effectiveness is even higher than the TAN's one (only $o(n^2N+n^2logn)$), being n the number of attributes and N the number of training instances.

VIII. Conclusion

The authors' approach relaxing the attribute independence assumption strives for outperforming the Naïve Bayes algorithm in terms of accuracy, but also on ranking measured by AUC.

The ODANB simply adds a parent to some attribute and certainly outperforms NB, SBC and TAN measured by accuracy and AUC, and this without incurring in a complexity increase.

References

[1] Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[2] http://archive.ics.uci.edu/ml/datasets/Annealing

[3] http://archive.ics.uci.edu/ml/datasets/Audiology+%28Original%29

[4] http://archive.ics.uci.edu/ml/datasets/Automobile

[5] http://archive.ics.uci.edu/ml/datasets/Balance+Scale

[6] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer

[7] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[8] http://archive.ics.uci.edu/ml/datasets/Horse+Colic

[9] http://archive.ics.uci.edu/ml/datasets/Credit+Approval

[10] http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29

[11] http://archive.ics.uci.edu/ml/datasets/Diabetes

[12] http://archive.ics.uci.edu/ml/datasets/Glass+Identification

[13] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[14] http://archive.ics.uci.edu/ml/datasets/SPECT+Heart

[15] http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

[16] http://archive.ics.uci.edu/ml/datasets/Hepatitis

[17] http://archive.ics.uci.edu/ml/datasets/Thyroid+Disease

[18] http://archive.ics.uci.edu/ml/datasets/Ionosphere

[19] http://archive.ics.uci.edu/ml/datasets/Iris

[20] http://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King%29

[21] http://archive.ics.uci.edu/ml/datasets/Labor+Relations

[22] http://archive.ics.uci.edu/ml/datasets/Letter+Recognition

[23] http://archive.ics.uci.edu/ml/datasets/Lymphography

[24] http://archive.ics.uci.edu/ml/datasets/Mushroom

[25] http://archive.ics.uci.edu/ml/datasets/Primary+Tumor

[26] http://archive.ics.uci.edu/ml/datasets/Image+Segmentation

[27] http://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+%28Sonar%2C+Mines+vs.+Rocks%29

[28] http://archive.ics.uci.edu/ml/datasets/Molecular+Biology+%28Splice-junction+Gene+Sequences%29

[29] http://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29

[30] http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

[31] http://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+%28Vowel+Recognition+-+Deterding+Data%29

[32] http://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+1%29

[33] http://archive.ics.uci.edu/ml/datasets/Zoo

[34] Friedman, Geiger, and Goldszmidt. "Bayesian Network Classfiers", Machine. Learning, Vol. 29, 131-163, 1997.

[35] D. Lewis. *Naive Bayes at forty: The independence assumption in information retrieval*. In *Conference proceedings of European Conference on Machine Learning*, pages 4-15, 1998

[36] Langley, P., Sage, S. Induction of selective Bayesian classifiers. in Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, 1994, pp. 339-406.

[37] Jie Cheng and Russell Greiner. Comparing bayesian network classifiers. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99), pages 101--107. Morgan Kaufmann Publishers, August 1999.

[38] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition,30,1145-1159.

[39] Provost, F., Fawcett, T.: Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. AAAI Press(1997) 43-48

[40] Liang Xiao Jiang, Harry Zhang, Zhihua Cai and Jiang Su. One Dependence Augmented Naive Bayes, ADMA, 2005, pages 186-194