Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 1752–1760
Pattern Recognition Letters
www.elsevier.com/locate/patrec

The Naive Bayes Mystery: A classification detective story

Adrien Jamain, David J. Hand *

Imperial College, Department of Mathematics, 180 Queen's Gate, London, SW7 2AZ, United Kingdom

Received 7 January 2005; received in revised form 15 February 2005
Available online 14 April 2005

Abstract

Many studies have been made to compare the many different methods of supervised classification which have been developed. While conducting a large meta-analysis of such studies, we spotted some anomalous results relating to the Naive Bayes method. This paper describes our detailed investigation into these anomalies. We conclude that a very large comparative study probably mislabelled another method as Naive Bayes, and that the Statlog project used the right method, but possibly incorrectly reported its provenance. Such mistakes, while not too harmful in themselves, can become seriously misleading if blindly propagated by citations which do not examine the source material in detail.

Keywords: Supervised classification; Naive Bayes; Statlog; Comparative studies

1. The setting

Many different competing methods have been developed for supervised classification. As a corollary, there have also been many comparative studies of the performance of classification methods. The aims of these studies range from analysing the relative merits of different variations of the same methodology (e.g., Dietterich, 2000) to providing large-scale benchmarks (e.g., Zarndt, 1995). As a by-product of these specific aims, such studies have created the general body of opinion that researchers hold about the relative performance of the different methods. However, the question of whether the view that this literature provides is representative of the 'real' use to which classification methods will be put has been raised by a few (Duin, 1996; Salzberg, 1997). A related but often unrecognised problem is that studies, even if carried out in complete good faith and using sound methodology, are not always accurate enough in their reporting. Indeed, the literature abounds with clear evidence of harmless omissions; for example minor imprecision about which UCI dataset is used or about how data have been pre-processed in a study are commonplace. Other

* Corresponding author. Fax: +44 20 7594 8517.
E-mail addresses: adrien.jamain@imperial.ac.uk (A. Jamain), d.j.hand@imperial.ac.uk (D.J. Hand).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.02.001

more serious kinds of confusions are not straightforward to spot, and can only be established with some uncertainty (after all, the authors always have the benefit of the doubt). While we were carrying out a meta-analysis of classification studies (Jamain, 2004), we discovered what we think is a worrying confusion in two large comparative studies. We present here the evidence that has led us to this belief, and try to unveil what really happened in both comparative studies.

2. The initial clue

One tool which we have routinely used in the course of the meta-analysis described in (Jamain, 2004) is clustering. The basic idea is to consider a primary study as a set of n observations (the methods) in p dimensions (the datasets), and to cluster the methods according to their results. The method of our choice was hierarchical clustering, with complete linkage (that is, the distance between two clusters is taken to be the largest one between two points belonging to each cluster), but any other clustering method could also be used. The resulting dendrograms can be visually inspected to discover interesting patterns, and indeed they often offer considerable insights into the relative performance of the methods.

We were producing such a clustering tree for the Zarndt study (Zarndt, 1995), which is as far as we know the largest comparative study ever made and which is available in the Citeseer online library,[1] when we noticed a very curious pattern. We reproduce this tree in Fig. 1, so that the reader will be able to see if they can spot this anomalous pattern as well. As the figure clearly shows, six different clusters may be identified in the Zarndt study:

- The multi-layer perceptron—`bp`—on its own.
- The noise-tolerant extensions of the nearest-neighbour method, `ib3` and `ib4`.
- The nearest-neighbour and its condensed version, `ibl` and `ib2`.

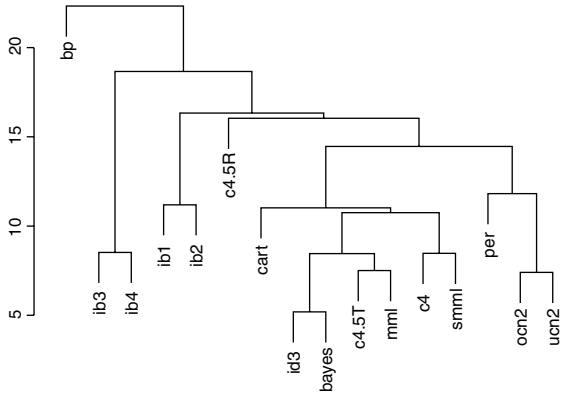

Fig. 1. Hierarchical clustering of the Zarndt study (using complete linkage).

- `C4.5R` on its own.
- The decision trees—`cart`, `id3`, `c4.5T`, `mml`, `c4` and `smml`—and Naive Bayes—`bayes`.
- The perceptron—`per`—and the variants of `CN2`—`ocn2`, `ucn2`.

So, where is the anomaly? Well, as the reader may have guessed (aided, perhaps by the title of the paper), it is in the location of the Naive Bayes method, which is deep inside the decision tree cluster. We have noted that in general apparent anomalies can occur in the clusters formed in this study and others. However, this one is stranger because it is not situated in a small cluster or on the edges of a large one. For example, the position of the perceptron within what can be called the rules cluster (next to the two variants of `CN2`) is potentially intriguing as well, but in fact it is situated on the periphery of this very small cluster and hence the pattern is not that intriguing. In contrast, the decision tree cluster is clearly built 'around' what the author calls Naive Bayes.[2] This peculiarity led us to carry out some further investigations.

3. The evidence

We pursued our inquiry by carefully scrutinizing what the author of the study wrote in his report. Perhaps it is worth mentioning here that

[1] http://citeseer.nj.nec.com

[2] See Appendix A for the classical definition of Naive Bayes.

the study was the result of an M.Sc. project, and hence, fortunately for us, the report gives great detail about all the experimental steps. Had the study been reported in a journal, the story would probably have ended inconclusively here.

In the body of his thesis (p. 22), Zarndt is silent about which implementation he used for Naive Bayes. However, in the appendices (p. 72) he reports having used the well-known IND package (Buntine, 1993), and describes which command lines he used for 'Naive Bayes':

```
mktree -e -v -s bytes <file name>
```

The term `mktree` of course reinforced our suspicion. Our doubts were confirmed after further consultation of the IND documentation (Buntine, 1993, p. 22) since it appeared that this particular line creates a Bayesian tree, similar to the MML and SMML trees, but certainly not anything like the classical Naive Bayes method. This would explain the high experimental similarity between the so-called `bayes` method of Zarndt and all the decision trees observed in Fig. 1.

To be fair and honest, this confusion has probably had little consequence due to the relatively limited availability of the study, and would not alone be the object of a communication such as the present one. Still, the Zarndt study has already been referenced four times within the Citeseer library itself, and will probably be more in the future since it has now been discovered and cited. Apart from the obvious similarity in name between Naive Bayes and Bayesian tree, it was still quite puzzling how such a misinterpretation could have happened; after all, Zarndt himself qualifies the IND package as 'decision tree software' on page 22, the same page on which he writes about Naive Bayes. However, further investigation unveiled what may have been the explanation.

4. The plot thickens

Also on page 22, the same page on which Zarndt 'describes' the Naive Bayes method, although he writes little of particular relevance (seven very general lines), he does makes a reference to the famous Statlog study (Michie et al., 1994). This is perhaps not surprising since this study is generally considered to be the golden reference in comparative studies of classification methods. The reasons for Statlog's popularity are manifold, but perhaps the main one is that its interdisciplinary character makes it accessible to researchers in all areas interested in classification methods.[3] Statlog produced results for 23 methods tested on 22 datasets from varied domains, and involved many research teams distributed all over Europe. As we will see shortly in more detail, a 'Naive Bayes' routine from the IND package was also apparently used in this study. Hence, it seems possible that Zarndt read the Statlog study, downloaded the IND package, and found that the most similar command to 'something Bayesian' was the one above. This raises the question of what happened in Statlog?

When the Statlog authors describe the Naive Bayes method (p. 40), it is the usual, statistical one. They however write that the specific routine used came from the IND package. This is confirmed in the appendix (p. 263), where the following methods are allegedly taken from the IND package: `INDCart`, `Bayes Tree` and `Naive Bayes`. The first two probably correspond to (unspecified) sequences of IND instructions according to the respective tree building and pruning principles. But the third one is something of a mystery.

We have tried the obvious approach to solving this mystery (i.e. contacted the authors) but, at the time of writing, received no answer. The problem is that, besides Statlog now being a rather old piece of work, it is probably difficult for anyone to know exactly what was done by each research team in a project of such a scale. This means that all that now remains is speculation, based on the evidence at our disposal.

'Naive Bayes' in Statlog could indeed be:

(1) a Bayesian tree confusingly made via IND (similarly to what Zarndt seems to have done);

[3] And the full book is downloadable for free from http://www.amsta.leeds.ac.uk/~charles/statlog/

(2) a specifically programmed routine, or one from another package, implementing the correct statistical concept but erroneously reported as from IND;
(3) a routine of the IND package which has escaped our inspection, or maybe which has been deleted since then (the package went from version 1.0 at the time of Statlog to 2.1 at present).

However, the third hypothesis is very unlikely, given that the package is allegedly for the 'creation and manipulation of decision trees from data',[4] and that the history of the modifications of IND does not make any reference to such a routine. Now, the question is: can we find any evidence supporting either of the other two hypotheses?

5. First witness: Direct comparison

The uncertainty described above risks creating a confusing picture of the performance of Naive Bayes. Can we resolve it by finding evidence of what Naive Bayes is in Statlog by comparing results from Statlog with those from other studies, where we can ascertain—as far as possible—that the correct implementation of Naive Bayes has been used? We know of six other such studies (of course it is likely that there are many other studies in the literature that include some results related to Naive Bayes, but these are the ones we included in our meta-analysis; Jamain, 2004):

- (Asparoukhov and Krzanowski, 2001): 30 results on 5 non-UCI datasets which have not been used elsewhere, to the best of our knowledge.
- (Titterington et al., 1981): 12 results on the non-UCI `head` dataset which has not been used in Zarndt, but has been in Statlog.
- (Kontkanen et al., 1998): 9 results on 9 different datasets, three of which have been used in Statlog (`Cr. Aust`—Australian credit scoring—and `Diab`—Pima indian diabetes—, `Heart`—heart disease, two-class version).
- (Weiss and Kapouleas, 1989): 4 results on 4 different datasets. None of them have been used in Statlog and 3 of them have been by Zarndt (`Ann-thyroid`, `Iris`, `Breast Yugoslavian`).
- (Cestnik et al., 1987): 4 results on 4 different datasets (`Lymphography`, `Primary`, `Breast Yugoslavian`, `Hepatitis`), all used by Zarndt but none in Statlog.
- (Clark and Niblett, 1987): 3 results on 3 datasets (`Breast Yugoslavian`, `Lymphography`, `Primary`), again none of them are in Statlog, but all in Zarndt.

Most of the studies do not share any dataset with Statlog, and hence an investigation of the similarity between Naive Bayes in Statlog and Naive Bayes in other studies can only be based a priori on the unique dataset of Titterington and 3 datasets of Kontkanen. Unfortunately, it turns out that the evidence is even weaker than this suggests: the results from the `Heart` dataset in Kontkanen are not comparable with those of Statlog since the two studies use different misclassification cost matrices (identity in Kontkanen, non-identity in Statlog). The other two datasets of Kontkanen are used with the same cost matrix and the same pre-processing as Statlog. As far as the `head` dataset common to Titterington and Statlog goes, although both studies use the same cost matrix and the same set of variables (set III in Titterington), there are a number of discrepancies between how the dataset has been processed which makes a completely fair comparison impossible. Indeed, the Titterington version has 1000 examples, split half-and-half between training and test samples, and no missing values replaced (methods included in the study were able to deal with missing values directly), whereas the Statlog version has only 900 examples, was processed by ninefold cross-validation, and more importantly had all its missing values replaced by class medians.

[4] From the IND website http://ic.arc.nasa.gov/projects/bayes-group/ind/IND-program.html

Table 1
Direct comparison of error rates between Naive Bayes in Statlog and Naive Bayes in other studies

Dataset	Statlog	Other study	Statlog MAD
Cr. Aust	0.151	0.150 (Kontkanen et al., 1998)	0.019
Pima	0.262	0.243 (Kontkanen et al., 1998)	0.021
Head	23.950	21.900 (Titterington et al., 1981)	9.990

Misclassification costs are equal for Cr. Aust and Pima, and non-equal for Head (for the exact form of the cost matrix see Titterington et al., 1981, p. 154; or Michie et al., 1994, p. 150).

Anyway, completely fair or not, the direct evidence of the comparative performance of Naive Bayes amounts to a set of three pairwise comparisons, which we reproduce in Table 1. To try to make some sense of these numbers, we also show an estimate of the dispersion of observed results for each dataset. For a given dataset we took this estimate to be the MAD (the median absolute deviation about the median) of all the results related to the dataset in the Statlog study. Results by dataset are typically right-skewed; hence, the choice of the MAD estimator over the classical standard deviation—but using standard deviation would not change our conclusion anyway. Since all other results are within one MAD of Statlog there seems to be very little discrepancy of performance between the different applications of Naive Bayes. This tends to support our second hypothesis, which is that the mistake in Statlog was only in the reporting. However, this is only a very small sample of datasets on which to base any assertion, and besides, had we found some discrepancy, we could not have concluded anything due to the tendency of the data to be dispersed anyway. Indeed, to take one example, error rates for the real Naive Bayes on the Breast Yugoslavian dataset show more variation between themselves than with the presumably false Naive Bayes of Zarndt: 0.28 for Weiss and Kontkanen, 0.38 for Clark, 0.22 for Cestnik, and in the middle of all 0.31 for Zarndt. Hence, we have now to look for further evidence, which will have to be indirect.

6. Second witness: Overall accuracy

One question which we may try to answer to shed some (indirect) light on the problem is the following: is there a difference in *overall* accuracy of Naive Bayes between Statlog and the other studies? In other words, we may consider the global performance of Naive Bayes within each study instead of looking at results for individual datasets. A major difficulty with this approach is how to create some measure of this global performance. Bearing in mind that it has to remain simple (due to the small size of some studies), here is our proposal:

- Within each study, scale each 'set of comparable results' (see below) between 0 and 1, with 0 being the best and 1 the worst.
- Still within each study, average all these scaled results by method.

This gives a simple estimate of the overall performance of each method within each study, and we call this estimate 'overall error rate'. What we call a 'set of comparable results' is, broadly speaking, a set of results which are on the same scale. Such a set usually consists of all the results related to a given dataset in a study, but not always. For example, when within a study a dataset is tested with different sets of variables, or different accuracy measures (e.g. different cost matrices), then the corresponding sets of results are not comparable in the strict sense because they are not on the same scale. Hence these sets need to be distinguished before scaling to $[0,1]$. For example, Titterington et al. (1981) use four different accuracy measures (error rate with equal misclassification costs, error rate with non-equal misclassification costs, average logarithmic score, average quadratic score) and four different sets of variables, all with the same dataset head. This amounts to 16 different 'sets of comparable results'. There are even more subtleties when one considers a study such as (Asparoukhov and

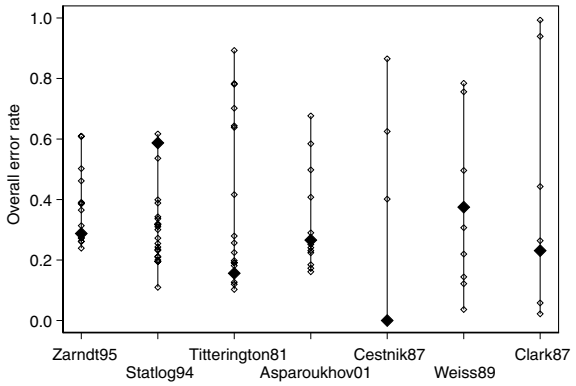

Fig. 2. Overall error rate of Naive Bayes and other methods in the literature (results for Naive Bayes are in large plain diamonds).

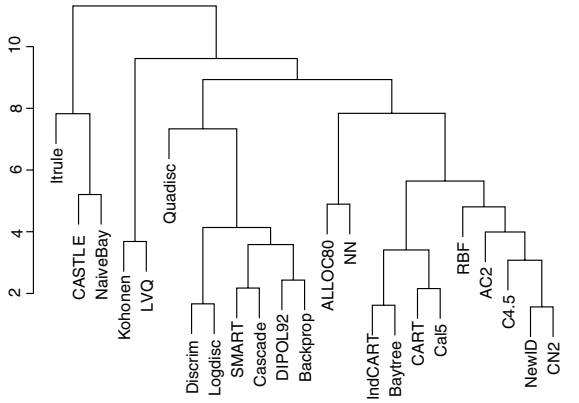

Fig. 3. Hierarchical clustering of the Statlog study (using complete linkage).

Krzanowski, 2001): in this particular study, methods are used with different class priors. Using different priors being equivalent to using different cost matrices, each set of results related to a given set of priors constitutes a different 'set of comparable results'.

We show overall error rates for all methods in Fig. 2, including those of Zarndt for comparison. The striking feature is that Naive Bayes seems to perform much worse in Statlog than in other studies. Of course, the overall error rate is very variable in the case of small studies (and there are quite a few of them), and it is also relative to the particular choice of datasets and methods within each study as we will see in the next section. For example, Naive Bayes could appear bad when compared with certain methods and good with others. At first sight, anyway, this would be an argument for our first hypothesis, which is a confusion between methods in Statlog; however, before drawing a perhaps hasty conclusion we should look more closely at the results of Statlog to see why Naive Bayes appears to perform so badly there.

7. Third witness: Inside Statlog

Using an approach similar to that which initially led to our suspicions about Naive Bayes in the Zarndt study, we may look at the clusters of methods in Statlog (Fig. 3). These are quite interesting on their own; one can see clearly differentiated clusters for decision trees and rules (`IndCART`, `Baytree`, `CART`, `Cal5`, `AC2`, `C4.5`, `NewID`, `CN2`), statistical and neural methods (`Discrim`, `Logdisc`, `SMART`, `Cascade`, `DIPOL92`, `Backprop`), and other more minor patterns. Only the quadratic discriminant `Quadisc`, the rule-induction algorithm `Itrule`, and radial-basis functions `RBF` seem quite out of their logical place. Concerning Naive Bayes, it is located closely to the Bayes causal network `CASTLE`, and away from all the decision trees and rules (except the aforementioned `Itrule`). Notably far away are the other methods taken from IND, namely `IndCART` and `Baytree`. This is of course strong evidence against our first hypothesis, and suggest that it was the correct statistical method which was used in Statlog.

Now, to return to the question above, what can explain the difference between the overall accuracy of Naive Bayes in Statlog and in other studies? In fact, a possible explanation is simply in the choice of datasets that the different studies have made. For example, one noticeable fact is that all the smaller studies used medical diagnosis datasets, which usually incorporate a considerable amount of prior knowledge in variable selection. In particular, it is common in such situations that the variables have relatively little correlation (on the principle that highly correlated variables are likely

to contribute less unique information about class separability; see Hand and Yu, 2001). In contrast, Statlog has a wider range of datasets, including for example image datasets—and Statlog's Naive Bayes does badly on these. In Fig. 4 we have shown the 0–1 scaled error rates, where for each dataset the best method of Statlog is given 0 and the worst 1, as before in our measure of overall error rate. Naive Bayes is the worst method for five datasets: `Vehicle` (vehicle silhouette recognition), `SatIm` (satellite imaging), `Dig44` (optical digit recognition), `Cut50` (character segmentation from handwritten word images), and `BelgII` (industrial dataset of unknown domain). Among these, the first four are image-related, and the origins of the last one are unknown. Besides, Naive Bayes is not far from the worst method on `KL` (a processed version of `Dig44`), `Letter` (letter recognition), and `Cut20` (a processed version of `Cut50`). In contrast, Naive Bayes is best or close to the best for `Heart` and `Head`, two medical diagnosis datasets, and does quite well on the credit scoring datasets also (`Cr.Aust`, `Cr.Ger`, `Cred.Man`).

In fact, it is possible to go a bit further, thanks to the availability of some dataset characteristics in Statlog (pp. 171–172). If one looks at the averaged absolute correlation between variables (Fig. 5), one can see a fairly clear relationship between the performance of Statlog's Naive Bayes and this particular dataset characteristic. More precisely, it

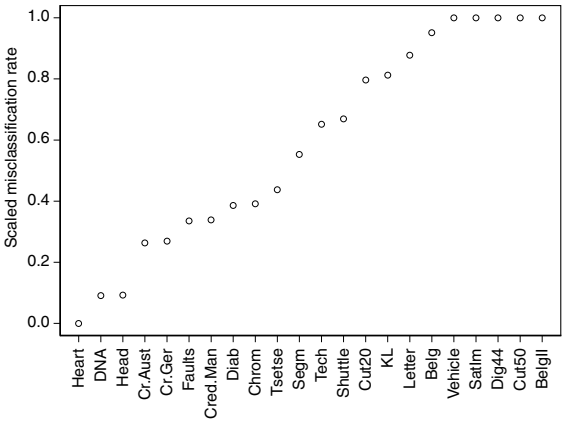

Fig. 4. Scaled misclassification rate of Naive Bayes for each dataset in Statlog (datasets ordered by misclassification rate).

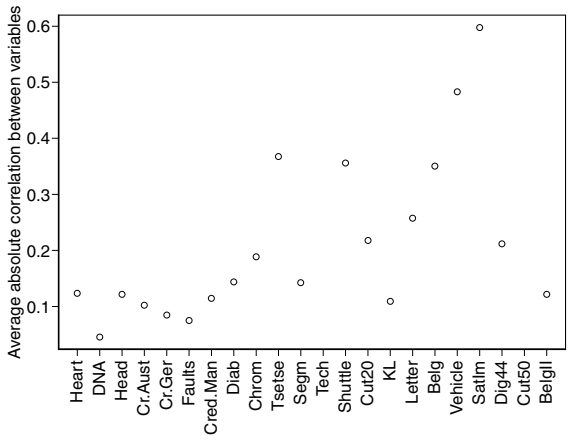

Fig. 5. Averaged between-variable correlation for each Statlog dataset (same dataset order as Fig. 4—hence Naive Bayes has poor performance on the datasets on the right).

seems that the datasets on which it does relatively well are those with a very low averaged absolute correlation. And those on which it does badly tend to have high values for this characteristic. One may remark that some of these do not (`BelgII` is an example), but having a low value for this averaged absolute correlation does not mean *necessarily* that all pairs of variables are weakly correlated—and some other factor may play a part as well. Anyway, it seems that the method tends to do badly when variables are correlated, and do well when they are not, providing a further piece of evidence that the correct Naive Bayes method was used in Statlog.

8. The verdict

In the present communication we have presented all the evidence that has led us to doubts about what 'Naive Bayes' exactly is in two large comparative studies, Zarndt and Statlog. Now that we have studied the available evidence, perhaps we should tentatively draw a conclusion. Let us first recall the main points of this evidence.

For Zarndt's Naive Bayes:

- In hierarchical clustering, presence in a decision tree cluster.

- Explicit statement about the use of a computer routine from the IND decision tree package.

For Statlog's Naive Bayes:

- Vague statement about the use of a computer routine from IND.
- Similarity of performance with other reported results from 'true' Naive Bayes on three datasets.
- In hierarchical clustering, presence in a small cluster, next to a Bayes causal network method.
- Poor overall accuracy, but explained by the presence of datasets with high between-variable correlation.

Perhaps the reader will have drawn their own conclusion by now, but we believe that these points suggest that:

(1) Zarndt's Naive Bayes *is not* the correct method.
(2) Statlog's Naive Bayes *is* the right method, but reported with a wrong provenance.

This is a rather reassuring conclusion for the Naive Bayes method, and also for the prestige of the Statlog study. However, we may wonder where the reporting mistake came from. We suggest that it probably has something to do with the organization of the Statlog project: apparently, all datasets were sent to each research group, and each had some relative liberty in choosing which method they were going to use. There could thus have been some omission or confusion in the reporting of one particular group (the one which used the other methods from the IND package).

On a slightly broader note, the moral of the story we have told is perhaps twofold: (a) mistakes do happen, even in well designed studies, and (b) if they do happen they might be propagated by further studies, in the same form as they first occur or in another form. The last point is perhaps the crucial one: one could argue that a single mistake is in itself not really harmful, but that the tendency of researchers to refer relatively blindly to previous work is. The present story illustrates that accurate reporting is a very important part of comparative studies. It perhaps also illustrates Murphy's Law: 'if something can go wrong, it will'.

Acknowledgements

The work described in this paper was sponsored by the MOD Corporate Research Programme, CISP. We would like to express our appreciation to Andrew Webb for his support and encouragement for this work.

Appendix A. The real Naive Bayes

The Naive Bayes method originally tackles problems where variables are categorical, although it has natural extensions to other types of variables. It assumes that variables are independent within each class, and simply estimates the probability of observing a certain value in a given class by the ratio of its frequency in the class of interest over the prior frequency of that class. That is, for any class c and vector $\boldsymbol{X} = (X_j)_{j=1,\ldots,k}$ of categorical variables,

$$P(\boldsymbol{X}|c) = \prod_{j=1}^{k} P(X_j|c)$$

and

$$P(X_j = x|c) = \frac{\#\text{ training examples of class } c \text{ where } x_j = x}{\#\text{ training examples of class } c}.$$

Continuous variables are generally discretised, or a certain parametric form is assumed (e.g. normal). One can also use non-parametric density estimators like Kernel functions. Then similar frequency ratios are derived.

References

Asparoukhov, O.K., Krzanowski, W.J., 2001. A comparison of discriminant procedures for binary variables. Comput. Statist. Data Anal. 38, 139–160.

Buntine, W., 1993. IND documentation, version 2.1. NASA Ames Research Center. Available from: <http://ic.arc.nasa.gov/projects/bayes-group/ind/IND-program.html>.

Cestnik, B., Kononenko, I., Bratko, I., 1987. ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In: Progress in Machine Learning: Proc. EWSL-87. Sigma Press, Bled, Yugoslavia, pp. 31–45.

Clark, P., Niblett, T., 1987. Induction in noisy domains. In: Proc. 2nd Eur. Work Session on Learning. Sigma Press, Glasgow, Scotland, pp. 11–30.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decisions trees: Bagging, boosting, and randomization. Machine Learn. 40, 139–157.

Duin, R.P.W., 1996. A note on comparing classifiers. Pattern Recognition Lett. 17, 529–536.

Hand, D., Yu, K., 2001. Idiot's Bayes—not so stupid after all? Internat. Statist. Rev. 69, 385–398.

Jamain, A., 2004. Meta-analysis of classification methods. Unpublished Ph.D. thesis, Department of Mathematics, Imperial College, London.

Kontkanen, P., Myllymaki, P., Silander, T., Tirri, H., 1998. Bayes optimal instance-based learning. In: 11th Eur. Conf. on Machine Learning. Springer-Verlag, Berlin, pp. 77–88.

Michie, D., Spiegelhalter, D.J., Taylor, C.C., 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood.

Salzberg, S.L., 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining Knowledge Discovery 1, 317–328.

Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F., Gelpke, G.J., 1981. Comparison of discrimination techniques applied to a complex data set of head injured patients. J. Roy. Statist. Soc. Ser. A 144, 144–175.

Weiss, S.M., Kapouleas, I., 1989. An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. In: IJCAI89: Proc. 11th Internat. Joint Conf. on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 781–787.

Zarndt, F., 1995. A comprehensive case study: An examination of machine learning and connectionnist algorithms. Available from: <http://citeseer.nj.nec.com/481595.html>.