Pairwise Naive Bayes Classifier

Jan-Nikolas Sulzmann
Technische Universität Darmstadt
D-64289, Darmstadt, Germany
sulzmann@ke.informatik.tu-darmstadt.de

Abstract

Class binarizations are effective methods that break multi-class problem down into several 2-class or binary problems to improve weak learners. This paper analyzes which effects these methods have if we choose a Naive Bayes learner for the base classifier. We consider the known unordered and pairwise class binarizations and propose an alternative approach for a pairwise calculation of a modified Naive Bayes classifier.

1 Introduction

The Naive Bayes classifier (NB) is a Bayesian learner which outperforms more sophisticated learning methods like Neural Networks, Nearest Neighbour, or Decision Tree Learning in many fields of application. NB is widely deployed because of its simplicity, versatility, and efficiency. We seek to increase its performance by combining it with class binarization methods which are a way to enhance a weak learner. To this end we consider on the one hand the well known unordered and pairwise classbinarization and on the other hand alternative methods for a pairwise calculation of NB.

We developed a couple of alternative methods which all are based on the same probabilistic approach. This approach uses not only the probabilities of classes but also the probabilities of pairs of classes which can be computed in two different ways. The first one which we call regular estimates the probabilities much like NB, the second one estimates not only the probabilities of classes but also the probability of pairs of classes.

This paper summarizes the results of [Sulzmann, 2006] and is assembled as follows. In the section *"Fundamentals & Notations"* we introduce the required notations and give a short survey of the basics of a Naive Bayes classifier. After this we describe the used class binarizations and several decoding methods.

The second section *"Pairwise Naive Bayes Classifier"* consist of three subsections. In the first two subsections we draw some conclusions about class binarizations with a Naive Bayes classifier. We show that contrary to expectations the unordered class binarizations is not equal to a Naive Bayes classifer. Afterwards we prove that the pairwise class binarization with decoding methods we described in the previous sections is equivalent to a Naive Bayes classifier. In the third subsection we introduce an alternative approach for a pairwise calculation of a Naive Bayes classifier. After this we describe four methods that implement this approach differently. All of them can be computed in two ways which we introduce thereafter.

In the third section *Experiments* we describe the experiments we have made. They consist of a comparison of our own methods, class binarizations and the Naive Bayes classifier which we test with several options (e.g. discretization and computation techniques). After this we analyse the results of our experiments which are summarized in two tables in the appendix.

In the last section *Conclusion* we resume the conclusions we have drawn about class binarizations with a Naive Bayes classifier and about our own methods.

2 Fundamentals & Notations

This section prides a short survey of the fundamentals and notations which are relevant for this paper.

2.1 Notation

A: an attribute, a set of attribute values

A_i: the ith attribute, $i \in \{1, ..., n\}$

a_i: an abitrary value of attribute A_i

v_i: the quantity of different attribute values of A_i

$D = (a_1, ..., a_n)$: an example described by attribute values

$c_i \in \{c_1, c_2, ..., c_m\}$: a class (a probabilistic event)

$c_{ij}, i \neq j$: a class pair, (the probabilistic event $c_{ij} = c_i \cup c_j$)

Pr: a probability

$\widehat{\Pr}$: a estimated probability

$\Pr(c_i|D)$: the probability that the example is of class c_i

$\Pr(c_i|D, c_{ij})$: the probability that the example is of class c_i under observation of class pair c_{ij}

t: the quantity of training examples

t_{c_i}: the quantity of training examples who belong to class c_i

$t_{c_j}^{a_i}$: the quantity of training examples who belong to class c_j and whose j^{th} attribute value matches a_i

2.2 Naive Bayes Classifier

The Naive Bayes classifier is a Bayesian learning method and therefore based on the theorem of Bayes:

$$\Pr(A|B) = \frac{\Pr(B|A) \cdot \Pr(A)}{\Pr(B)}$$

The goal of NB is to predict from a training set of classified examples the class of an example $D = (a_1, , a_n)$, where a_i is the value of the i^{th} attribute. The error can be minimized by selecting $argmax_{c_i} \Pr(c_i|D)$, where

$c_i, ..., c_m$ are the m classes. Therefor we need estimates $\widehat{\Pr}(c_i|D)$ of $\Pr(c_i|D)$,$\forall i$. One can adapt the theorem of Bayes to solve this problem:

$$\Pr(c_i|D) = \frac{\Pr(D|c_i) \cdot \Pr(c_i)}{\Pr(D)},$$

where $\Pr(D) = \sum_j (\Pr(D|c_j) \cdot \Pr(c_j))$

With this approach we gain the following basic version of a Bayesian learner:

$$\begin{aligned} c_B &= \arg\max_{c_i} \frac{\Pr(D|c_i) \cdot \Pr(c_i)}{\Pr(D)} \\ &= \arg\max_{c_i} \Pr(D|c_i) \cdot \Pr(c_i) \\ &= \arg\max_{c_i} \Pr(a_1, a_2, ..., a_n|c_i) \cdot \Pr(c_i) \end{aligned}$$

If we make the naive assumptions that the attributes are independent the classifier is called *naive* and the probability $\Pr(D|c_i)$ can be calculated as follows:

$$\begin{aligned} \Pr(D|c_i) &= \Pr(a_1, a_2, ..., a_n|c_i) \\ &= \prod_{j=1}^{n} \Pr(a_j|c_i) \end{aligned}$$

With this formula we obtain the basic version of NB:

$$c_{NB} = \arg\max_{c_i} \Pr(c_i) \cdot \prod_{j=n}^{n} \Pr(a_j|c_i)$$

The probabilities $\Pr(c_i)$ and $\Pr(a_j|c_i)$ can be estimated with the relative frequencies of training examples of class c_i in the training set and accordingly with the relative frequencies of trainining examples of this class whose attribute value of the corresponding attribute has the value a_j.

If one of the latter relative frequencies equals zero for one of the classes, then the total predicted probability of this class equals also zero. This problem can be solved by assuming that the relative frequencies have prior distributions. A well known representative of this approach is the Laplace estimate. It presumes that each attribute value occurs one more time than it appears in the training set.

The probabilities $\Pr(a_j|c_i)$ and $\Pr(c_i)$ can be estimated (with the Laplace estimate) as follows:

$$\widehat{\Pr}(a_i|c_j) = \frac{t_{c_j}^{a_i} + 1}{t_{c_j} + v_i} \qquad \widehat{\Pr}(c_i) = \frac{t_{c_i}}{t}$$

2.3 Class Binarization

Class binarization techniques [Fürnkranz, 2002; 2003] solve multi-class problems by turning them into a set of binary problems. This enables machine learning methods which are inherently designed for binary problems (e.g. perceptrons, support vector machines (SVM) etc.) to solve multi-class problems.

Definition 2.1 (class binarization, decoding, base learner) *A* class binarization *is a mapping of a multi-class learning problem to several two-class learning problems in a way that allows a sensible* decoding *of the prediction, i.e., it allows the derivation of a prediction for the multi-class problem from the predictions of the set of two-class classifiers. The learning algorithm used for solving the two-class problems is called the* base learner .

The most popular class binarization technique is the *unordered or one-against-all class binarization* (abbr.: 1vsAll), where one takes each class in turn and learns binary concepts that discriminate this class from all other classes. It has been independently proposed for rule learning, neural networks, and SVM.

Definition 2.2 (unordered/one-against-all class binarization) *The* unordered class binarization *transforms a m-class problem into m binary problems. These are constructed by using the examples of class i as the positive examples and the examples of classes j ($j = 1, ..., c, j \neq i$) as the negative examples.*

A more complex binarization technique is the *pairwise or round robin class binarization*. The basic idea is quite simple, namely to learn one classifier for each pair of classes.

Definition 2.3 (round robin/pairwise class binarization) *The round robin or pairwise class binarization transforms a m-class problem into $m(m-1)/2$ two-class problems $< i, j >$, one for each set of classes $\{i, j\}$,$i = 1, ..., m-1$,$j = i+1, ..., m$ The binary classifier for problem $< i, j >$ is trained with examples of classes i and j, whereas examples of classes $k \neq i, j$ are ignored for this problem.*

A crucial point of this technique is how to decode the predictions of the pairwise classifiers to a final prediction. We use Voting, Weighted Voting and methods which are based on the Bradley-Terry-modell for decoding the predictions [Wu *et al.*, 2004].

Voting (abbr.: V) is a simple technique. When we classify a new example, each of the learned base classifiers determines to which of its two classes the example is more likely to belong to. The winner is assigned a point, and in the end, the algorithm will predict the class that has accumulated the most points.

$$\arg\max_{c_i} \sum_{j=i} [\Pr(c_i|D, c_{ij})], \quad [x] = \begin{cases} 1, \text{ if x } \geq 0.5 \\ 0, \text{ else.} \end{cases}$$

For *Weighted Voting* (abbr.: WV) we need additionally the confidence of the base classifiers in its predictions. In contrast to Voting we do not add up points but weighted votes which correspond to the confidence measures.

$$\arg\max_{c_i} \sum_{j=i} \Pr(c_i|D, c_{ij})$$

The last two methods are based upon the *Bradley-Terry modell* which consists of the assumption that the following holds:

$$\Pr(c_i|D, c_{ij}) = \frac{\Pr(c_i|D)}{\Pr(c_i|D) + \Pr(c_j|D)}$$

The methods that we will introduce try to estimate with this assumption and the estimates of $\Pr(c_i|D, c_{ij})$ the probability $\Pr(c_i|D)$ for each class c_i. They attempt to minimize the distance between the estimation $\widehat{\Pr}(c_i|D, c_{ij})$ and $\overline{\Pr}(c_i|D, c_{ij})$ which can be calculated as follows:

$$\overline{\Pr}(c_i|D, c_{ij}) = \frac{\widehat{\Pr}(c_i|D)}{\widehat{\Pr}(c_i|D) + \widehat{\Pr}(c_j|D)}$$

The methods differ in their approach of minimizing the distance betweens the estimates and calculations. The first

Algorithm 2.1 Method of Hastie & Tibshirani

Input: $\widehat{Pr}(c_i|D,c_{ij})$,$t_{c_{ij}}$, $i,j \in \{1,...,m\}$, $j \neq i$
Output: $\widehat{Pr}(c_i|D)$, $i = 1,...,m$

1: Start with some initial $\widehat{Pr}(c_i|D)$,$\forall i$ and corresponding $\widehat{Pr}(c_i|D,c_{ij})$
2: **repeat** $\{i = 1,...,m,1,...\}$
3: $\quad \alpha = \frac{\sum_{j\neq i} t_{c_{ij}} \cdot \widehat{Pr}(c_i|D,c_{ij})}{\sum_{j\neq i} t_{c_{ij}} \cdot \overline{Pr}(c_i|D,c_{ij})}$
4: $\quad \overline{Pr}(c_i|D,c_{ij}) \leftarrow \frac{\alpha \cdot \overline{Pr}(c_i|D,c_{ij})}{\alpha \cdot \overline{Pr}(c_i|D,c_{ij}) + \widehat{Pr}(c_i|D,c_{ij})}$
5: $\quad \overline{Pr}(c_j|D,c_{ij}) = 1 - \overline{Pr}(c_i|D,c_{ij})$
6: $\quad \widehat{Pr}(c_i|D) = \alpha \cdot \widehat{Pr}(c_i|D)$
7: $\quad \widehat{Pr}(c_i|D) = \frac{\widehat{Pr}(c_i|D)}{\sum_j \widehat{Pr}(c_j|D)}$ (optional)
8: **until** m consecutive α are all close to ones
9: **return** $\widehat{Pr}(c_i|D) = \frac{\widehat{Pr}(c_i|D)}{\sum_j \widehat{Pr}(c_j|D)}$

method which was proposed by [Price *et al.*, 1994] (referred as PKPD) is based on a calculation formula. The second one that was suggested by [Hastie and Tibshirani, 1997] (referred a HT) specifies an algorithm which solves this minimization problem.

[Price *et al.*, 1994] consider that

$$\left(\sum_{j\neq i} \Pr\left(c_{ij}|D\right)\right) - (m-2)\cdot \Pr\left(c_i|D\right) = \sum_{j=1}^{m} \Pr\left(c_j|D\right)$$

holds. If we adapt this equation and

$$\Pr\left(c_{ij}|D\right) = \Pr\left(c_i|D\right) + \Pr\left(c_j|D\right)$$

to the estimated probabilities we obtain the following calculation formula:

$$\widehat{\Pr}\left(c_i|D\right)_{PKPD} = \frac{1}{\sum_{j\neq i} \frac{1}{\widehat{\Pr}(c_i|D,c_{ij})} - m + 2}$$

The approach of [Hastie and Tibshirani, 1997] tries to minimize the Kullback-Leibler distance $l\left(p\right)$ between $\widehat{\Pr}\left(c_i|D,c_{ij}\right)$ and $\overline{\Pr}\left(c_i|D,c_{ij}\right)$.

$$\begin{aligned} l\left(p\right) &= \sum_{i<j} t_{c_{ij}} \cdot \left(\widehat{\Pr}\left(c_i|D,c_{ij}\right) \cdot \log \frac{\widehat{\Pr}\left(c_i|D,c_{ij}\right)}{\overline{\Pr}\left(c_i|D,c_{ij}\right)} \right. \\ &+ \left. \widehat{\Pr}\left(c_j|D,c_{ij}\right) \cdot \log \frac{\widehat{\Pr}\left(c_i|D,c_{ij}\right)}{\overline{\Pr}\left(c_i|D,c_{ij}\right)} \right), \end{aligned}$$

where $t_{c_{ij}} = t_{c_i} + t_{c_j}$ is the sum of the training examples of the classes c_i + c_j.

To this end [Hastie and Tibshirani, 1997] propose to find estimated probabilities $\widehat{\Pr}(c_i|D)$ for each class which satisfy the following conditions:

$$\sum_{j\neq i} t_{c_{ij}} \cdot \widehat{\Pr}\left(c_i|D,c_{ij}\right) = \sum_{j\neq i} t_{c_{ij}} \cdot \overline{\Pr}\left(c_i|D,c_{ij}\right)$$

$$\sum_{i=1}^{m} \widehat{\Pr}\left(c_i|D\right) = 1$$

$$\widehat{\Pr}\left(c_i|D\right) > 0, i = 1,...,m$$

This problem can be solved by algorithm 2.1

3 Pairwise Naive Bayes Classifier

As aforementioned we seek to improve the performance of NB by combining it with class binarization methods. Therefore we consider the unordered and pairwise class binarization and alternative methods.

3.1 Unordered Class Binarization

The structure of the unordered class binarization with NB as its base classifier is very similar the structure to NB. So one might think they compute the same predictions for a given example. Contrary to expectations these two methods calculate different probabilitiy estimates and if applicable different predictions.

The unordered class binarization splits a m-class problem in m binary problems that consists of discriminating one class from all other. For class c_i the other classes are handled as one class $\overline{c_i}$ and their training example are thrown together. This approach does not change any relative frequencies of class c_i. Therefore the probabilities $\Pr(D|c_i)$ and $\Pr(c_i)$ remain unchanged.

The absolute frequencies of $\overline{c_i}$ have to be calculated:

$$t^a_{\overline{c_i}} = \sum_{j\neq i} t^a_{c_j} \qquad t_{\overline{c_i}} = \sum_{j\neq i} t_{c_j}$$

With the aid of this quantities and the Laplace estimate the required probabilities for $\overline{c_i}$ can be estimated as follows:

$$\widehat{\Pr}(a_k|c_i) = \frac{t^{a_k}_{\overline{c_i}} + 1}{t_{\overline{c_i}} + v_k} = \frac{\left(\sum_{j\neq i} t^{a_k}_{c_j}\right) + 1}{\left(\sum_{j\neq i} t_{c_j}\right) + v_k}$$

and

$$\widehat{\Pr}(\overline{c_i}) = \frac{t_{\overline{c_i}}}{t} = \frac{\sum_{j\neq i} t_{c_j}}{t} = \sum_{j\neq i} \widehat{\Pr}(c_j) = 1 - \widehat{\Pr}(c_i)$$

Now we can estimate $\Pr(D|\overline{c_i})$ as follows:

$$\widehat{\Pr}(D|\overline{c_i})_{UK} = \prod_{k=1}^{n} \widehat{\Pr}(a_k|\overline{c_i})_{UK} = \prod_{k=1}^{n} \frac{\left(\sum_{j\neq i} t^{a_k}_{c_j}\right) + 1}{\left(\sum_{j\neq i} t_{c_j}\right) + v_k}$$

If we consider the abovementioned estimation we can clearly see that the following holds:

$$\widehat{\Pr}(D|\overline{c_i})_{UK} \cdot \widehat{\Pr}(\overline{c_i}) \neq \sum_{j\neq i} \widehat{\Pr}(D|c_j)_{NB} \cdot \widehat{\Pr}(c_j)$$

Therefore NB and the unordered class binarization compute different estimations for $\Pr(c_i|D)$:

$$\begin{aligned} & \widehat{\Pr}(c_i|D)_{UK} \\ = & \frac{\widehat{\Pr}(D|c_i) \cdot \widehat{\Pr}(c_i)}{\widehat{\Pr}(D|c_i) \cdot \widehat{\Pr}(c_i) + \widehat{\Pr}(D|\overline{c_i})_{UK} \cdot \widehat{\Pr}(\overline{c_i})} \\ \neq & \frac{\widehat{\Pr}(D|c_i) \cdot \widehat{\Pr}(c_i)}{\widehat{\Pr}(D|c_i) \cdot \widehat{\Pr}(c_i) + \sum_{j\neq i} \widehat{\Pr}(D|c_j)_{NB} \cdot \widehat{\Pr}(c_j)} \\ = & \widehat{\Pr}(c_i|D)_{NB} \end{aligned}$$

3.2 Pairwise Class Binarization

Contrary to the expectations the pairwise class binarization with NB as its base classifier is equivalent to NB. That means they predict always the same class for an example. Different predictions can be traced back to imprecise implementations of the class binarization shemes.

Before we can give a proof of the aforementioned, we have to show some relations between the probabilities of both classifiers. The base classifier of class pair c_{ij} estimates the probabilities $\Pr(c_i|D, c_{ij})$ and $\Pr(c_j|D, c_{ij})$ that can be calculated as follows:

$$\Pr(c_i|D, c_{ij}) = \frac{\Pr(D|c_i) \cdot \Pr(c_i)}{\Pr(D|c_i) \cdot \Pr(c_i) + \Pr(D|c_j) \cdot \Pr(c_i)}$$
$$\Pr(c_j|D, c_{ij}) = 1 - \Pr(c_i|D, c_{ij})$$

As a reminder the probability $\Pr(c_j|D)$ can be calculated as follows:

$$\Pr(c_i|D) = \frac{\Pr(D|c_i) \cdot \Pr(c_i)}{\sum_j (\Pr(D|c_j) \cdot \Pr(c_j))}$$

If we compare both calculations, we are able to see that the probabilities differ only in their normalization factors. This leads to the following lemma:

Lemma 3.1 *For any two mutual different classes c_i and c_j holds*

$$\frac{\Pr(c_i|D, c_{ij})}{\Pr(c_j|D, c_{ij})} = \frac{\Pr(c_i|D)}{\Pr(c_j|D)}$$

PROOF.

$$\frac{\Pr(c_i|D, c_{ij})}{\Pr(c_j|D, c_{ij})} = \frac{\frac{\Pr(c_i|D)}{\Pr(c_i|D)+\Pr(c_j|D)}}{\frac{\Pr(c_j|D)}{\Pr(c_i|D)+\Pr(c_j|D)}} = \frac{\Pr(c_i|D)}{\Pr(c_j|D)} \quad \square$$

This relation gives a hint to some essential transitive correlation between the probabilities of class pairs.

Lemma 3.2 *For any mutual different classes c_i, c_j and c_k holds:*

(a) The following inequalities are equivalent

$$\begin{aligned} & \Pr(c_i|D) < \Pr(c_j|D) && (1)\\ \Leftrightarrow\quad & \Pr(c_i|D, c_{ij}) < \Pr(c_j|D, c_{ij}) && (2)\\ \Leftrightarrow\quad & \Pr(c_i|D, c_{ik}) < \Pr(c_j|D, c_{jk}) && (3) \end{aligned}$$

(b)

$$\begin{aligned} & \Pr(c_i|D, c_{ik}) < \Pr(c_k|D, c_{ik})\\ \wedge\quad & \Pr(c_k|D, c_{jk}) < \Pr(c_j|D, c_{jk})\\ \Rightarrow\quad & \Pr(c_i|D, c_{ij}) < \Pr(c_j|D, c_{ij}) \end{aligned}$$

PROOF. (a)

(1) $\Leftrightarrow$ (2)

$$\begin{aligned} & \Pr(c_i|D) < \Pr(c_j|D)\\ \Leftrightarrow\quad & \tfrac{\Pr(c_i|D)}{\Pr(c_i|D)+\Pr(c_j|D)} < \tfrac{\Pr(c_i|D)}{\Pr(c_i|D)+\Pr(c_j|D)}\\ \Leftrightarrow\quad & \Pr(c_i|D, c_{ij}) < \Pr(c_j|D, c_{ij}) \end{aligned}$$

(1) $\Leftrightarrow$ (3)

$$\begin{aligned} & \Pr(c_i|D) < \Pr(c_j|D)\\ \Leftrightarrow\quad & \Pr(c_i|D) + \Pr(c_k|D) < \Pr(c_j|D) + \Pr(c_k|D)\\ \Leftrightarrow\quad & \tfrac{\Pr(c_k|D)}{\Pr(c_i|D)+\Pr(c_k|D)} > \tfrac{\Pr(c_k|D)}{\Pr(c_j|D)+\Pr(c_k|D)}\\ \Leftrightarrow\quad & \tfrac{\Pr(c_i|D)}{\Pr(c_i|D)+\Pr(c_k|D)} < \tfrac{\Pr(c_j|D)}{\Pr(c_j|D)+\Pr(c_k|D)}\\ \Leftrightarrow\quad & \Pr(c_i|D, c_{ik}) < \Pr(c_j|D, c_{jk}) \end{aligned}$$

(b)

$$\begin{aligned} & \begin{pmatrix} \Pr(c_i|D, c_{ik}) < \Pr(c_k|D, c_{ik}) \\ \wedge \Pr(c_k|D, c_{jk}) < \Pr(c_j|D, c_{jk}) \end{pmatrix}\\ \Leftrightarrow\quad & \begin{pmatrix} \Pr(c_i|D) < \Pr(c_k|D) \\ \wedge \Pr(c_k) < \Pr(c_j|D) \end{pmatrix}\\ \Leftrightarrow\quad & \Pr(c_i|D) < \Pr(c_j|D) \end{aligned} \quad \square$$

These transitive relations hold analogous for the equality of probabilities.

Corollary 3.3 *For any mutual different classes c_i, c_j and c_k holds:*

(a) The following inequalities are equivalent

$$\begin{aligned} & \Pr(c_i|D) = \Pr(c_j|D) && (1)\\ \Leftrightarrow\quad & \Pr(c_i|D, c_{ij}) = \Pr(c_j|D, c_{ij}) && (2)\\ \Leftrightarrow\quad & \Pr(c_i|D, c_{ik}) = \Pr(c_j|D, c_{jk}) && (3) \end{aligned}$$

(b)

$$\begin{aligned} & \Pr(c_i|D, c_{ik}) = \Pr(c_k|D, c_{ik})\\ \wedge\quad & \Pr(c_k|D, c_{jk}) = \Pr(c_j|D, c_{jk})\\ \Rightarrow\quad & \Pr(c_i|D, c_{ij}) = \Pr(c_j|D, c_{ij}) \end{aligned}$$

PROOF. Follows directly from Lemma (3.2) or can be analogous proven. $\square$

Now we have all utilities that we need for our latter equivalency proofs. Let us have a look at the basic structure of the Round Robin class binarization whose class pairs are evaluated by voting methods.

$$\arg\max_{c_i} \sum_{j \neq i} vote(c_i, c_j), \qquad (1)$$

where vote is a function which determines depending on the voting method how class c_i is rated under the class pair c_{ij}. The voting methods Voting and Weighted Voting can be written as follows:

$$vote_V(c_i, c_j) = [\Pr(c_i|D, c_{ij})]$$
$$vote_{WV}(c_i, c_j) = \Pr(c_i|D, c_{ij})$$

Comparing these functions with each other and NB we draw two conclusions. First both functions are equivalent in respect to the voting result. That means the predictions of the Round Robin class binarization with one of these functions are the same. Second the Round Robin class binarization with these voting methods is equivalent to NB. Before we proof this conclusion we have to introduce some minor facts.

Lemma 3.4 *For any mutual different classes c_i and c_k holds:*

(a)

$$\begin{aligned} & \Pr(c_i|D, c_{ij}) < \Pr(c_j|D, c_{ij})\\ \Leftrightarrow\quad & [\Pr(c_i|D, c_{ij})] < [\Pr(c_j|D, c_{ij})] \end{aligned}$$

(b)

$$\begin{aligned} & \Pr(c_i|D, c_{ij}) < \Pr(c_j|D, c_{ij})\\ \Rightarrow\quad & [\Pr(c_i|D, c_{ik})] \leq [\Pr(c_j|D, c_{jk})] \end{aligned}$$

PROOF. (a)

$$\begin{aligned} \Leftrightarrow\quad & \Pr(c_i|D, c_{ij}) < \frac{1}{2} < \Pr(c_j|D, c_{ij})\\ \Leftrightarrow\quad & [\Pr(c_i|D, c_{ij})] < [\Pr(c_j|D, c_{ij})] \end{aligned}$$

(b)

$$\begin{aligned} & \Pr(c_i|D, c_{ij}) < \Pr(c_j|D, c_{ij})\\ \Leftrightarrow\quad & \Pr(c_i|D, c_{ik}) < \Pr(c_j|D, c_{jk})\\ \Rightarrow\quad & [\Pr(c_i|D, c_{ik})] < [\Pr(c_j|D, c_{jk})] \end{aligned} \quad \square$$

This leads directly to the following corollary:

Corollary 3.5 *For any mutual different classes c_i and c_j holds:*

$$\begin{aligned} & \Pr(c_i|D,c_{ij}) = \Pr(c_j|D,c_{ij}) \\ \Leftrightarrow \quad & [\Pr(c_i|D,c_{ij})] = [\Pr(c_j|D,c_{ij})] \\ \Leftrightarrow \quad & [\Pr(c_i|D,c_{ik})] = [\Pr(c_j|D,c_{jk})] \end{aligned}$$

Hence we draw the conclusion that the rankings of NB and of the Round Robin class binarization with voting methods are related as follows:

Lemma 3.6 *For any mutual different classes c_i and c_j holds:*

(a)

$$\begin{aligned} & \Pr(c_i|D) \leq \Pr(c_j|D) \\ \Leftrightarrow \quad & \sum_{k\neq i} \Pr(c_i|D,c_{ik}) \leq \sum_{k\neq j} \Pr(c_j|D,c_{jk}) \end{aligned}$$

(b)

$$\begin{aligned} & \Pr(c_i|D) \leq \Pr(c_j|D) \\ \Leftrightarrow \quad & \sum_{k\neq i} [\Pr(c_i|D,c_{ik})] \leq \sum_{k\neq j} [\Pr(c_j|D,c_{jk})] \end{aligned}$$

PROOF.

(a) "⇒":

For any class c_k with $c_i \neq c_k \neq c_j$ holds

$$\begin{aligned} & \Pr(c_i|D,c_{ij}) < \Pr(c_j|D,c_{ij}) \\ \Rightarrow \quad & \Pr(c_i|D,c_{ik}) < \Pr(c_j|D,c_{jk}) \end{aligned}$$

due to Lemma 3.2 and

$$\begin{aligned} & \Pr(c_i|D,c_{ij}) = \Pr(c_j|D,c_{ij}) \\ \Rightarrow \quad & \Pr(c_i|D,c_{ik}) = \Pr(c_j|D,c_{jk}). \end{aligned}$$

due to Lemma Korollar 3.3.

Recapitulating we obtain the following implication.

$$\begin{aligned} & \Pr(c_i|D) \leq \Pr(c_j|D) \\ \Rightarrow \quad & \sum_{k\neq i} \Pr(c_i|D,c_{ik}) \leq \sum_{k\neq j} \Pr(c_j|D,c_{jk}) \end{aligned}$$

"⇐": Analogous to "⇒" holds

$$\begin{aligned} & \Pr(c_i|D,c_{ij}) > \Pr(c_j|D,c_{ij}) \\ \Rightarrow \quad & \sum_{k\neq i} \Pr(c_i|D,c_{ik}) > \sum_{k\neq j} \Pr(c_j|D,c_{jk}) \end{aligned}$$

due to Lemma 3.2. Via contraposition we obtain the following:

$$\begin{aligned} & \sum_{k\neq i} \Pr(c_i|D,c_{ik}) \leq \sum_{k\neq j} \Pr(c_j|D,c_{jk}) \\ \Leftrightarrow \quad & \neg \left(\sum_{k\neq i} \Pr(c_i|D,c_{ik}) > \sum_{k\neq j} \Pr(c_j|D,c_{jk}) \right) \\ \Rightarrow \quad & \neg \left(\Pr(c_i|D,c_{ij}) > \Pr(c_j|D,c_{ij}) \right) \\ \Leftrightarrow \quad & \Pr(c_i|D,c_{ij}) \leq \Pr(c_j|D,c_{ij}) \end{aligned}$$

□

(b) Can analagous be proven with the aid of Lemma 3.4 and Korollar 3.5.

According to Lemma 3.6 the rankings of NB and the Round Robin class binarization with Voting or Weighted Voting are equivalent. Hence we can draw the following conclusion:

Theorem 3.7 *The Round Robin class binarization with NB as its base classifier and the voting methods Voting or Weighted Voting predicts the same ranking and classification as a Naive Bayes classifier:*

$$\begin{aligned} & \arg\max_{c_i} \Pr(c_i|D) \\ = \quad & \arg\max_{c_i} \sum_{j\neq i} \Pr(c_i|D,c_{ij}) \\ = \quad & \arg\max_{c_i} \sum_{j\neq i} [\Pr(c_i|D,c_{ij})] \end{aligned}$$

PROOF. According to Lemma 3.6 the rankings of these methods are equivalent. Therefore their predictions are also equal. □

This theoretical equivalency exists also in practice if the estimated probability $\widehat{\Pr}(c_i|D)$ is equal for any class c_i and all of its class pairs. This is the case if the estimations $\widehat{\Pr}(a|c_i)$ and $\widehat{\Pr}(c_i)$ have the same value for all class pairs. These estimated probabilities depend only on the relative frequency of training examples which belong to class c_i and if applicable whose attribute values match a. These frequencies are not affected by class binarization techniques. Therefore the estimated probabilities $\widehat{\Pr}(D|c_i)$ and $\widehat{\Pr}(c_i)$ are also the same for every class and all of its class pairs. Thus the same holds for $\widehat{\Pr}(c_i|D)$ and the theoretical equivalency of the pairwise class binarization and NB occurs also in practice.

Differences between the pairwise class binarization and NB can be traced back to imprecise implementations of these methods. For example if the continuous attributes should be discretized the discretization can be applied on all training example (referred as global, abbr.: G) or on the training examples of each class pair (referred as binary, abbr.: B). The former does not change the abovementioned relative frequencies but the latter clearly does.

Considering the methods that are based on the Bradley-Terry modell we draw the conclusion that the pairwise class binarization with these methods is also equivalent to NB. These methods try to minimize the distance between $\widehat{\Pr}(c_i|D,c_{ij})$ and $\overline{\Pr}(c_i|D,c_{ij})$. As we have seen the estimated probability $\widehat{\Pr}(c_i|D)$ is the same for each class c_i and any of its class pairs. Therefore the following holds:

$$\widehat{\Pr}(c_i|D,c_{ij}) = \frac{\widehat{\Pr}(c_i|D)}{\widehat{\Pr}(c_i|D) + \widehat{\Pr}(c_i|D)} = \overline{\Pr}(c_i|D,c_{ij})$$

Thus the distance of the estimated probabilities is zero and the results of methods HT and PKPD equal those of NB.

3.3 Alternative Method

In the previous subsection we have seen that the pairwise class binarization does not improve the performance of NB. Thus we considered an alternative probabilistic approach for a pairwise estimation of NB. Its basic idea is to estimate $\Pr(c_i|D)$ by the pairwise probabilities $\Pr(D|c_{ij})$ which can be computed in two different ways. Before we

explain these ways we want to introduce our approach.

$$\begin{aligned}(m-1)\Pr(c_i|D) &= \sum_{j\neq i}\Pr(c_i|D)\\ &= \sum_{j\neq i}\Pr(c_i|D,c_{ij})\cdot\Pr(c_{ij}|D)\end{aligned}$$

$$\Leftrightarrow \Pr(c_i|D) = \frac{1}{m-1}\sum_{j\neq i}\Pr(c_i|D,c_{ij})\cdot\Pr(c_{ij}|D)$$

This approach leads to the following basic classifier:

$$\begin{aligned}&\arg\max_{c_i}\frac{1}{m-1}\sum_{j\neq i}\Pr(c_i|D,c_{ij})\cdot\Pr(c_{ij}|D)\\ =\ &\arg\max_{c_i}\sum_{j\neq i}\Pr(c_i|D,c_{ij})\cdot\Pr(c_{ij}|D)\end{aligned}$$

The two term of this classifier will be referred as

$$v_{ij} = \Pr(c_i|D,c_{ij})$$

and

$$w_{ij} = w_{ji} = \Pr(c_{ij}|D).$$

We compute these terms as follows:

$$v_{ij} = \frac{\Pr(D|c_i)\cdot\Pr(c_i)}{\Pr(D|c_i)\cdot\Pr(c_i)+\Pr(D|c_j)\cdot\Pr(c_j)}$$

$$w_{ij} = \frac{\Pr(D|c_{ij})\cdot\Pr(c_{ij})}{\Pr(D)}$$

As abovementioned $\Pr(D|c_{ij})$ can be computed in two different ways. The first one (referred as R) calculates it with the estimations of $\Pr(D|c_i)$ and $\Pr(c_i)$ a regular NB.

$$\Pr(D|c_{ij})_R = \frac{\Pr(D|c_i)\Pr(c_i)+\Pr(D|c_j)\Pr(c_j)}{\Pr(c_i)+\Pr(c_j)}$$

The second one (referred as P) calculates it by merging the training examples of class pair c_{ij}. Hence not only the quantities of classes are used for the prediction but also those of class pairs. We get:

$$t_{c_{ij}} = t_{c_i} + t_{c_j} \qquad t^{a_k}_{c_{ij}} = t^{a_k}_{c_i} + t^{a_k}_{c_j}$$

and

$$\Pr(c_{ij}) = \frac{t_{c_i}+t_{c_j}}{t} \quad \Pr(a_k|D,c_{ij}) = \frac{t^{a_k}_{c_i}+t^{a_k}_{c_j}+1}{t_{c_i}+t_{c_j}+v_k}$$

$\Pr(D|c_{ij})$ can be computed as follows:

$$\Pr(D|c_{ij})_{PW} = \prod_{k=1}^{n}\Pr(a_k|c_{ij})$$

These methods have different computational complexities on training time. The first one has the same as NB, $O(tn)$. The second one considers each training example of a given class c_i not only once but one time for each class pair c_{ij}. This increases the training time by the factor m-1. Hence the second one has a computational complexity of $O(tnm)$.

Consequentially we cannot both use this basic classifier and calculate w_{ij} regularly because this results in the same prediction as NB. Therefore we have to consider modifications of the basic classifier. We will introduce two pairs of method which are related by their modifications. The first pair uses v_{ij} for voting and w_{ij} as the weight of the votes. The second pair uses the basic classifier but estimates v_{ij} by $\Pr(c_i|c_{ij})$.

The classifier of the first pair will be referred as PNB1 and PNB2 and the classifier of the second pair accordingly PNB3 and PNB4.

$$c_{\text{PNB1}} = \arg\max_{c_i}\sum_{\substack{j\neq i\\ v_{ij}\geq v_{ji}}} w_{ij}$$

$$c_{\text{PNB2}} = \arg\max_{c_i}\sum_{j\neq i} v_{ij}\cdot w_{ij}$$

$$c_{\text{PNB3}} = \arg\max_{c_i}\sum_{\substack{j\neq i\\ \Pr(c_i)\geq\Pr(c_j)}} w_{ij}$$

$$c_{\text{PNB4}} = \arg\max_{c_i}\sum_{j\neq i}\Pr(c_i|c_{ij})\cdot w_{ij}$$

As abovementioned PNB2 is equal to the basic classifier if we calculate it regularly. This holds also for PNB1. Therefore both will be calculated only pairwise. PNB3 and PNB4 can be computed regularly and pairwise.

4 Experiments

4.1 Experimental Setup

In this section we compare the methods which we introduced in the former section with NB. To this end we used the learning environment WEKA (short for The Waikato Environment for Knowledge Analysis) of the university of Waikato, New Zealand [Witten and Frank, 2005]. We extended WEKA by two new classifier.

The first one adds new utilities to the MulticlassClassifier which deals with multiclass problems by class binarization or ECOCs. The MulticlassClassifier has been augmented by the decoding method Weighted Voting, HT and PKPD and the option of choosing the discretization type. Now it is possible to discretize on all training examples or on the trainining examples of each class pair.

The second one implements our own methods PNB1 to PNB4 which can be regularly or pairwise computed.

Our experiments consist of two test series. We use data sets of the UCI repository that represent multiclass problems. Before we apply the classifiers the data set will be the one way or another discretized. We use the discretization method of [Fayyad and Irani, 1993] which is already implemented in WEKA.

The first test series compares our own PNB methods with NB. PNB1 and PNB2 will be calculated pairwise. The other methods use both computation techniques. In any case the discretization will be applied on all training examples.

The second test series compares unordered and pairwise class binarizations with NB. The discretization will be applied on the class pairs. In the unordered class binarizations it will also be applied on the whole data.

We compare the methods with a sign test. If the null hypothesis is discarded with niveau 0.95 or 0.99 we call the methods significant (abbr.: S) or accordingly highly significant (abbr.: HS) different. Else we call them equivalent (abbr.: E).

The results of the experiments are summarized in the table in the appendix. The tables contain the error rates of each method on the data sets, the quantities (how many times the method won, lost or was equal to NB) Win, Loss and Tie and the result of the sign tests. The error rates were obtained by stratified 10x10 cross validation and rounded

to fit in the table. The quantities Win, Loss and Tie were computed on the error rates before rounding was applied.

4.2 Results

In the first test series the PNB methods showed higher error rates than NB. Though PNB1, PNB2 and the regularly computed PNB 4 are equivalent to NB. PNB3 and the pairwise computed PNB4 are significant worse than NB. Comparing the regular and the pairwise computation the pairwise one is slightly worse than the regular. The application of PNB method is not advisable because of their bad results and if applicable their higher computational costs.

In the second test series we draw two conclusions. First the unordered class binarization is equivalent to NB, irrespective of choosing the binary or global discretization. Second the pairwise class binarization with the binary discretization is equivalent to NB for each decoding method we introduced but produces lower error rates in most of the cases.

5 Conclusion

In this paper we have drawn several conclusion about class binarizations with a Naive Bayes classifier. First we have shown that the unordered class binarization is not equal to a regular Naive Bayes classifier. Second we have proven that the pairwise class binarization is equivalent to NB for common decoding methods like Voting, Weighted Voting and the proposals of [Hastie and Tibshirani, 1997; Price *et al.*, 1994].

We suggested some alternative methods for a pairwise calculation of a modified Naive Bayes classifier. Our experiments showed that these methods did not improve the performance of a Naive Bayes classifier. Additionally the experiments exhibited that class binarizations can increase the performance of a Naive Bayes classifier if we apply the discretization on the training examples of each binary problem.

For further readings we suggest [Sulzmann, 2006] that gives a more detailed description of the proofs and experiments and considers some additional pairwise approaches.

References

[Fayyad and Irani, 1993] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1022–1029, 1993.

[Fürnkranz, 2002] Johannes Fürnkranz. Round robin classification. *Journal of Machine Learning Research (JMLR)*, 2:721–747, 2002.

[Fürnkranz, 2003] Johannes Fürnkranz. Round robin ensembles. *Intelligent Data Analysis*, 7(5):385–403, 2003.

[Hastie and Tibshirani, 1997] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems (NIPS)*. The MIT Press, 1997.

[Price *et al.*, 1994] David Price, Stefan Knerr, Léon Personnaz, and Gérard Dreyfus. Pairwise neural network classifiers with probabilistic outputs. In Gerald Tesauro, David S. Touretzky, and Todd K. Leen, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1109–1116. MIT Press, 1994.

[Sulzmann, 2006] Jan-Nikolas Sulzmann. Paarweiser Naive Bayes Klassifizierer. Diplomarbeit, Technische Universität Darmstadt, D-64289, Darmstadt, Germany, July 2006.

[Witten and Frank, 2005] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2 edition, 2005.

[Wu *et al.*, 2004] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research (JMLR)*, 5:975–1005, 2004.

data set	NB	$PNB1_P$	$PNB2_P$	$PNB3_P$	$PNB3_R$	$PNB4_P$	$PNB4_R$
anneal	4,00	4,17	3,91	3,51	3,49	3,44	3,42
audiology	26,95	26,72	26,89	31,13	27,79	29,13	27,80
autos	34,97	35,00	34,34	33,78	34,05	33,61	34,79
balancescale	27,93	27,93	27,97	35,59	38,25	27,82	27,88
glass	28,70	29,04	29,06	31,18	31,05	31,41	29,48
hypothyroid	1,74	1,70	1,84	3,19	4,77	3,11	2,86
iris	6,69	6,69	6,69	6,65	6,69	6,65	6,69
letter	25,95	26,80	27,20	47,38	31,27	28,99	25,95
lymph	14,83	14,83	14,70	19,06	17,76	16,12	17,73
primary-tumor	49,66	49,54	49,95	50,97	50,49	50,03	49,95
segment	8,93	9,27	9,36	14,63	8,93	14,63	8,93
soybean	7,14	7,86	8,05	15,05	11,54	10,39	7,28
splice	4,63	4,78	4,93	33,88	48,12	20,00	6,45
vehicle	39,32	39,33	39,04	57,68	57,83	41,65	39,11
vowel	41,27	41,51	41,44	42,77	41,27	42,77	41,27
waveform-5000	20,03	20,03	20,03	54,68	66,16	26,34	19,91
yeast	42,68	42,72	42,74	46,19	43,81	43,79	42,94
zoo	7,22	7,22	6,18	11,18	11,01	10,25	5,95
Mittel	21,81	21,95	21,91	29,92	29,68	24,45	22,13
compared to NB							
Win		4	6	3	2	4	6
Loss		10	11	15	13	14	9
Tie		4	1	0	3	0	3
method statistically equal to NB?							
E/S/HS		E	E	HS	HS	S	E

Table 1: Results of the first test series: PNB1-PNB4: error rates in percent

	Binary discretization		Global discretization				
data set	NB	$1vsAll_B$	$1vsAll_G$	V	WV	HT	PKPD
anneal	4,00	2,28	2,40	3,10	3,07	3,31	3,09
audiology	26,95	27,75	27,75	26,95	26,95	26,82	26,95
autos	34,97	32,80	32,05	32,70	31,89	31,84	32,78
balancescale	27,93	27,97	26,87	26,20	26,20	26,20	26,20
glass	28,70	28,35	31,90	32,08	30,21	30,76	31,38
hypothyroid	1,74	2,13	2,25	1,73	1,68	1,77	1,69
iris	6,69	6,69	6,97	6,56	6,56	6,56	6,56
letter	25,95	26,10	27,38	26,48	26,31	26,28	26,37
lymph	14,83	14,44	14,72	14,80	14,97	14,72	14,73
primary-tumor	49,66	48,62	48,62	49,66	49,66	49,42	49,66
segment	8,93	10,88	9,07	8,54	8,51	22,25	8,53
soybean	7,14	7,63	7,63	7,14	7,14	7,13	7,14
splice	4,63	5,11	5,11	4,63	4,63	4,63	4,63
vehicle	39,32	38,96	37,79	37,66	37,49	38,26	37,55
vowel	41,27	41,03	35,17	36,15	33,41	33,34	34,16
waveform-5000	20,03	21,22	21,32	20,08	20,08	23,47	20,08
yeast	42,68	42,63	41,51	42,80	41,57	41,83	42,26
zoo	7,22	5,35	3,96	7,49	6,12	6,42	6,64
Mittel	21,81	21,25	21,66	21,37	20,91	21,94	21,13
compared to NB							
Win		9	9	9	10	12	11
Loss		8	9	5	4	6	3
Tie		1	0	4	4	0	4
method statistically equal to NB?							
E/S/HS		E	E	E	E	E	E

Table 2: Results of the second test series: unordered and pairwise class binarizations with Voting, Weighted Voting, HT and PKPD: error rates in percent