Proceedings of Machine Learning Research vol 72, 499-510, 2018 PGM 2018

Naive Bayesian Classifiers with Extreme Probability Features

Linda C. van der Gaag L.C.VANDERGAAG@UU.NL
Department of Information and Computing Sciences, Utrecht University, The Netherlands

Andrea Capotorti ANDREA.CAPOTORTI@UNIPG.IT
Dipartimento di Matematica e Informatica, Università di Perugia, Italy

Abstract

Despite their popularity, naive Bayesian classifiers are not well suited for real-world applications involving extreme probability features. As will be demonstrated in this paper, methods used to forestall the inclusion of zero probability parameters in naive classifiers have quite counterintuitive effects. An elegant, principled solution for handling extreme probability events is available however, from coherent conditional probability theory. We will show how this theory can be integrated in standard naive Bayesian classifiers, and then present a computational framework that retains the classifiers' efficiency in the presence of a limited number of extreme probability features.

Keywords: Naive Bayesian classifiers; Extreme probabilities; Coherent conditional probability theory; Computational efficiency.

1. Introduction

Nowadays a multitude of methods and associated software are available for learning probabilistic models from data. Among these are methods for constructing Bayesian network classifiers (Friedman et al., 1997). These models include a designated variable of interest, called the class variable, and multiple feature variables, each of which is related directly to this class variable. Especially *naive Bayesian classifiers* (Duda and Hart, 1973) have become quite popular, as is evidenced by their use for a large range of applications, such as spam filtering, remote-sensing classification, and medical risk prediction. This popularity is readily explained by their ease of construction and use. Despite their strong underlying assumptions of independence moreover, these naive Bayesian classifiers tend to outperform more complex models (Domingos and Pazzani, 1997).

Naive Bayesian classifiers are proving less suited for real-world applications in which quite rare features play an important role, such as the medical diagnosis of serious disease with rare (pathognomonic) findings. When for such an application the classifier's parameter probabilities are learned from data, the estimate obtained for the rare feature at hand will most likely be zero. If the feature is not a logical impossibility, generally some smoothing method is employed to forestall the inclusion of zero probability parameters in the model (see for example Flach (2012)); the thereby established parameter value is dependent of the number of data instances available and, hence, is more or less arbitrary. If the classifier's parameter probabilities would be elicited from domain experts, also a rather arbitrary parameter value would be found for the rare feature, as experts are known to be quite uncomfortable and, in fact, rather unreliable in assessing very small probabilities.

The small parameter values included in a naive Bayesian classifier to forestall zero probability parameters, have various unwanted effects. From advances in sensitivity analysis of Bayesian networks for example, it is well known that inaccuracies in small parameter probabilities can strongly influence output probabilities of interest (Chan and Darwiche, 2002). Since the small parameter val-

ues included in a naive Bayesian classifier as a result of smoothing may easily be incorrect by one or more orders of magnitude, such inaccuracies can in fact change classification results. Absence of a rare feature may further have a changing effect on the output when taken into account explicitly, while the feature's absence essentially is uninformative and would not be expected to be influential.

Where common practices with developing naive Bayesian classifiers result in the inclusion of arbitrary small values to forestall zero probability parameters, *coherent conditional probability theory*, developed by Coletti and Scozzafava (2002) based on the pioneering ideas of de de Finetti (1972), offers a more principled approach to handling zero probability features in naive Bayesian classifiers. This theory allows the computation of probabilities conditioned on zero probability events by building on a representation of conditional probability as a sequence of unconditional probability distributions where each subsequent distribution zooms in on the zero probabilities from the previous one while maintaining coherence throughout the sequence.

In this paper, we will show how coherent conditional probability theory can be integrated with standard naive Bayesian classifiers, to arrive at classifiers that are able to handle zero probability features in a principled way. We will demonstrate that these classifiers do not have the unwanted effects of commonly-used smoothing approaches. We will further illustrate that by exploiting coherence theory, the resulting classifiers may not always return a point class probability for a case at hand; in fact, the probabilistic information represented by the classifier may be found to accommodate multiple probability distributions, thereby causing it to return a probability interval. As straightforward application of coherence theory is exponential in the number of variables of a classifier, and hence rapidly becomes computationally infeasible in practice, we will also propose a computational scheme for naive Bayesian classifiers with extreme probability features which retains the inferential efficiency of standard naive classifiers in the presence of a limited number of such features.

The paper is organised as follows. In Section 2, we provide some preliminaries on naive Bayesian classifiers and on coherent conditional probability theory. Section 3 provides some examples to demonstrate the unwanted effects of commonly-used approaches to handling zero probability features in naive Bayesian classifiers. In Section 4 the use of coherence theory within naive Bayesian classifiers in general is detailed, and in Section 5 our computational scheme is proposed. The paper ends with our concluding observations and plans for further research.

2. Preliminaries

We provide some preliminaries on naive Bayesian classifiers and on coherent conditional probability theory, and thereby introduce our notational conventions.

2.1 Naive Bayesian classifiers

Naive Bayesian classifiers are Bayesian networks that are tailored to solving problems where cases described by features are to be assigned one of several distinct classes (Duda and Hart, 1973; Friedman et al., 1997). To this end, a naive Bayesian classifier discerns a *class variable* C and a set $\mathbf{F}$ of *feature variables* F_i, $i = 1, \ldots, n$, $n \geq 1$. For ease of exposition, we assume all variables to be binary. The value $F_i = \mathit{true}$ will be denoted by f_i, while $\bar{f}_i$ indicates $F_i = \mathit{false}$; we will use f_i^* to indicate either value. A value combination $\mathbf{f}$ for (a subset of) $\mathbf{F}$ will be called an (input) *case*; the case is *complete* if it includes a value for each variable in $\mathbf{F}$. A case with an associated class is called an *instance*. A naive Bayesian classifier takes the topology of a directed tree, with the class variable C for its unique root and with each feature variable F_i having C for its only parent. This

tree captures dependence of the class variable on each feature variable separately, and models the independence $F_i \perp\!\!\!\perp F_j \mid C$ for any two feature variables F_i, F_j, $i \neq j$, given the class variable.

To supplement its independence structure, a naive Bayesian classifier specifies a probability distribution $\Pr(C)$ over its class variable and conditional distributions $\Pr(F_i \mid C)$ over the separate feature variables F_i; the specified probabilities are the classifier's *parameters*. Through its independence structure and associated parameter distributions, a naive Bayesian classifier represents a unique probability distribution $\Pr(\mathbf{F}, C)$, which factorises according to

$$\Pr(\mathbf{F}, C) = \Pr(C) \cdot \prod_{i=1}^{n} \Pr(F_i \mid C)$$

and from which essentially any probability over the variables involved can be computed. Classifiers are mostly used however, for establishing distributions $\Pr(C \mid \mathbf{f})$ over the class variable, given input cases $\mathbf{f}$. Associated with the classifier is a *decision rule* for assigning a class to an input case (Domingos and Pazzani, 1997; Friedman et al., 1997), based on the computed distribution and the application's utility considerations. A commonly used rule for this purpose takes a *probability threshold* δ, and assigns the class c to a case $\mathbf{f}$ if $\Pr(c|\mathbf{f}) \geq \delta$; otherwise the class $\bar{c}$ is returned.

Naive Bayesian classifiers are most often learned from data. Since their independence structure is fixed, their construction amounts to estimating the required parameter probabilities. The estimates are typically obtained as proportions over (sub-)sets of instances from the available dataset. Datasets being finite, estimates equal to zero are likely to be found for rare features. When such a feature is known not to be a logical impossibility, *Laplace correction* or another smoothing method is employed to forestall the inclusion of a zero probability parameter in the classifier at hand (Flach, 2012; Witten et al., 2005); the value thus included is dependent of the size of the available dataset.

2.2 Zero probabilities in a coherent setting

Upon reviewing *coherent conditional probability theory*, we re-phrase its basic concepts in terms of our context of random variables and probability distributions over these variables; for further introduction, we refer to Coletti and Scozzafava (2002).

We consider the Boolean algebra $\mathcal{V}$ spanned by a set $\mathbf{V}$ of (binary) variables; in this algebra, universal truth is indicated by $\top$, and logical impossibility by $\bot$; we use $\mathcal{V}^0 = \mathcal{V} \setminus \{\bot\}$ to denote the algebra without the impossibility. The value combinations of the variables $\mathbf{V}$ spanning $\mathcal{V}$, are termed the algebra's *atoms*. A probability distribution $\Pr(\cdot)$ over $\mathbf{V}$ now is a function on the algebra $\mathcal{V}$ that is uniquely defined by the probabilities of its atoms; these probabilities are termed the *constituent probabilities* of the distribution $\Pr$ (van der Gaag, 1990). In our context of naive Bayesian classifiers, the primary Boolean sentences of interest are disjunctions of atoms, that is, value combinations for (sub)sets of variables; the probability of such a sentence is expressed as the sum of the constituent probabilities of its composing atoms.

Following de de Finetti (1972), we assume *conditional probability* over $\mathbf{V}$ to be a function $\Pr(\cdot \mid \cdot)$ on $\mathcal{V} \times \mathcal{V}^0$ and look upon the conditioning part as a hypothesis which is allowed to have zero probability. The basic idea of coherent conditional probability theory now is to represent conditional probability as a sequence of unconditional probability distributions where each subsequent distribution zooms in on the zero probabilities from the previous one, maintaining coherence throughout the sequence. Any conditional probability $\Pr(\cdot \mid \cdot)$ over $\mathbf{V}$ thus has associated a unique, linearly ordered class $\mathcal{P} = [P_0, \ldots, P_k]$ of unconditional probability distributions over $\mathbf{V}$, called

its *complete agreeing class*; the index i of a distribution P_i in $\mathcal{P}$ will be termed the *level* of the representation of $\Pr(\cdot \,|\, \cdot)$. For a given conditional probability $\Pr(\cdot \,|\, \cdot)$, its agreeing class is obtained by setting

- $P_0(\cdot) = \Pr(\cdot \,|\, H_0^0)$, with the Boolean sentence $H_0^0 = \top$;
- for each successive level i, $P_i(\cdot) = \Pr(\cdot \,|\, H_0^i)$, with $H_0^i = \bigvee_{H \in H_0^{i-1}, P_{i-1}(H)=0} H$, that is, with H_0^i being the disjunction of all Boolean sentences of zero probability at the previous level;

with the iterative construction halting when $H_0^{k+1} = \bot$. For each Boolean sentence $H \in \mathcal{V}^0$, there now is a minimum index $i \in \{0, \ldots, k\}$ such that $P_i(H) > 0$; the associated level is called the *zero layer* of H with repect to $\mathcal{P}$. For every conditional probability $\Pr(E \,|\, H)$ with $E \in \mathcal{V}$, $H \in \mathcal{V}^0$, and i the index of the zero layer of H, we have that

$$\Pr(E \,|\, H) = \frac{P_i(E, H)}{P_i(H)}$$

Equivalent with the representation of conditional probability by a complete agreeing class of distributions $\mathcal{P}$ is the representation by a sequence $\mathcal{S}$ of compatible systems of linear equations (Coletti and Scozzafava, 2002). The first system S^0 of the sequence has all constituent probabilities x_j of the joint distribution $\Pr(\cdot)$ over the variables $\mathbf{V}$ for its unknowns; these constituents are now denoted as x_j^0, with the superscript indicating the system's level in the sequence. Each subsequent system S^i in the sequence includes for its unknowns just the constituent probabilities x_j^i which have $x_j^{i-1} = 0$ at the previous level $i-1$. The constituent probabilities x_j^i are called the *supporting constituents* for level i; the set of all constituent probabilities at level i will be denoted as X^i. The system of equations S^i for level i in the sequence now is constructed as follows:

$$S^i = \begin{cases} \displaystyle\sum_{x_j^i \in X^i} a_j^i \cdot x_j^i \;=\; p \cdot \sum_{x_j^i \in X^i} b_j^i \cdot x_j^i, \;\; \text{for all available } p = \Pr(E \mid H) \\ \displaystyle\sum_{x_j^i \in X^i} x_j^i = 1 \\ x_j^i \geq 0, \;\; \text{for all } x_j^i \in X^i \end{cases}$$

where $\sum_{x_j^i \in X^i} a_j^i \cdot x_j^i$ expresses the joint probability $\Pr(E, H)$ in its constituent probabilities by means of appropriate indicator coefficients a_j^i, and $\sum_{x_j^i \in X^i} b_j^i \cdot x_j^i$ similarly encodes $\Pr(H)$; the numerical value p of the probability $\Pr(E \mid H)$ is taken from the available probabilistic information.

The representation of conditional probability by a sequence of systems of equations is the operational tool used in coherence theory for checking if probabilistic information, specified either on the whole space $\mathcal{V} \times \mathcal{V}^0$ or on a proper subspace, is coherent and for extending the information to yet unconsidered probabilities of interest.

3. Motivating Examples

Despite their popularity, naive Bayesian classifiers are ill suited for real-world problems involving rare features. By means of two examples, we show that common methods for handling zero-probability features in such classifiers have unwanted effects. The examples further illustrate the

advantages of modelling these features in the principled setting of coherence theory.

Example 1. We consider a classification problem in which a rare feature is of importance; this feature occurs with (almost) zero probability, yet if it occurs the class for the case at hand is known to certainty. We model the problem by a standard naive Bayesian classifier, with the class variable A and the two feature variables B, C; the value b of B represents the rare feature and is assumed to occur only with a. For its parameters, the classifier includes the probability $\Pr(a) = 0.8$ for the class variable A, and the conditional probabilities $\Pr(c \mid a) = 0.4, \Pr(c \mid \bar{a}) = 0.2$ for the variable C. It further includes the zero probability $\Pr(b \mid \bar{a}) = 0$ to express the logical impossibility of b given $\bar{a}$. To capture the information that b is rare yet not impossible to occur, the small probability $\Pr(b \mid a) = 0.001$ is assigned to the event of b given a. The decision rule associated with the classifier uses the probability threshold $\delta = 0.75$ for assigning the class a.

From the constructed classifier, the conditional probability of a given any feature combination involving b is found to be equal to one, as a consequence of the logical impossibility of b given $\bar{a}$. Now, for the case $\bar{b}\bar{c}$ not involving the impossibility, the probability $\Pr(a \mid \bar{b}\bar{c}) = 0.7498$ is computed. With the classifier's decision rule therefore, the class $\bar{a}$ is assigned to the case. For the (incomplete) case with just the feature $\bar{c}$, the probability $\Pr(a \,|\, \bar{c}) = 0.7500$ would be found and the class a would be returned. Explicitly adding the feature $\bar{b}$ to the case would thus change the output, even though the feature in essence is uninformative. We note that this counterintuitive effect would occur, also if we reduced the parameter $\Pr(b \mid a)$ by several orders of magnitude in size.

We reconsider the classification problem in the setting of coherent conditional probability, and model b given a as a zero probability event. For this purpose, the parameter $\Pr(b \mid a) = 0$ is included in the classifier; all other parameter probabilities are as above. We note that in view of the semantics of standard naive Bayesian classifiers, the information $\Pr(b \mid a) = \Pr(b \mid \bar{a})$ would indicate independence of B of the class variable A. In the setting of coherence theory however, the feature variable B is *not* independent of the variable A as, informally speaking, the two zeroes involved have different strengths. For computing probabilities of interest from the thus constructed classifier, its parameters are first expressed in terms of the following constituent probabilities:

$$\begin{array}{lll} x_1^0 = \Pr(abc) & x_3^0 = \Pr(ab\bar{c}) & x_5^0 = \Pr(a\bar{b}\bar{c}) \\ x_2^0 = \Pr(a\bar{b}c) & x_4^0 = \Pr(\bar{a}\bar{b}c) & x_6^0 = \Pr(\bar{a}\bar{b}\bar{c}) \end{array}$$

As the atoms $\bar{a}bc$ and $\bar{a}b\bar{c}$ describe logical impossibilities, they are not considered explicitly. The parameter information from the classifier now gives rise to the system S^0 of equations for level zero of the sequence of compatible systems:

$$S^0 = \left\{ \begin{array}{ll} x_1^0 + x_2^0 + x_3^0 + x_5^0 \;=\; 0.8 \cdot (x_1^0 + x_2^0 + x_3^0 + x_4^0 + x_5^0 + x_6^0) & [\Pr(a \mid \top) = 0.8] \\ x_1^0 + x_3^0 \;=\; 0 \cdot (x_1^0 + x_2^0 + x_3^0 + x_5^0) & [\Pr(b \mid a) = 0] \\ x_1^0 + x_2^0 \;=\; 0.4 \cdot (x_1^0 + x_2^0 + x_3^0 + x_5^0) & [\Pr(c \mid a) = 0.4] \\ x_4^0 \;=\; 0.2 \cdot (x_4^0 + x_6^0) & [\Pr(c \mid \bar{a}) = 0.2] \\ x_1^0 + x_2^0 + x_3^0 + x_4^0 + x_5^0 + x_6^0 \;=\; 1 & [\Pr(\top) = 1] \\ x_i^0 \;\geq\; 0,\; i = 1, \ldots, 6 & \end{array} \right.$$

This system has the following unique solution:

$$\begin{array}{lll} x_1^0 = 0 & x_3^0 = 0 & x_5^0 = 0.48 \\ x_2^0 = 0.32 & x_4^0 = 0.04 & x_6^0 = 0.16 \end{array}$$

We find that the conditional probability $\Pr(a \mid \bar{b}\bar{c})$ can be established at this level and compute it to be $\Pr(a \mid \bar{b}\bar{c}) = x_5^0/(x_5^0 + x_6^0) = 0.7500$; we note that at this level we would also establish that $\Pr(a \mid \bar{c}) = 0.7500$. The case $\bar{b}\bar{c}$ thus is assigned the class a, as is the incomplete case $\bar{c}$. We conclude that in the setting of coherence theory, conditioning on the uninformative value $\bar{b}$ no longer has the counterintuitive effect seen with the standard naive Bayesian classifier. □

Where the above example illustrates the counterintuitive effect of more or less arbitrary small parameter values in a standard naive classifier upon taking the complement of a zero-probability feature into consideration, our next example demonstrates that such arbitrary small values can further have quite strong effects on an established output probability.

Example 2. We consider the same problem as in Example 1, yet now assume that the conditional events of b given a and of b given $\bar{a}$ both are rare yet not logically impossible. To express the information that b is highly unlikely to occur, both in the presence of A and in its absence, in a standard naive Bayesian classifier very small values are assigned to the probabilities involved. Included in our classifier now are $\Pr(b \mid a) = 0.001$, $\Pr(b \mid \bar{a}) = 0.0001$, expressing that, although in both cases quite rare, b is an order of magnitude more likely to be found with a than with $\bar{a}$. For the variables A, C, the classifier includes the same parameter probabilities as before. The decision rule associated with the classifier uses the probability threshold $\delta = 0.90$ for assigning the class a.

From the constructed classifier, the probability of a given the feature combination $b\bar{c}$ is computed to be $\Pr(a \mid b\bar{c}) = 0.9677$; with the classifier's decision rule, the class a is assigned to the case. Now, if the probabilities of the two events of b given a and b given $\bar{a}$ differed, not by a factor ten, but by a factor two instead, that is, if $\Pr(b \mid \bar{a}) = 0.0005$ instead of $\Pr(b \mid \bar{a}) = 0.0001$, then the probability of a given $b\bar{c}$ would have been found to be $\Pr(a \mid b\bar{c}) = 0.8571$ and the class $\bar{a}$ would be assigned to the case. Changes in the small parameter values included in a classifier to forestall zero probability parameters, can thus have considerable impact on the returned output. If the difference between the probabilities of b given a and b given $\bar{a}$ could not have been quantified and equally small parameter values would have been included for both events, in essence an independency of A and B would have been introduced, in which case we would have found that $\Pr(a \mid b\bar{c}) = \Pr(a \mid \bar{b}\bar{c}) = 0.75$, regardless of the value included for $\Pr(b \mid a) = \Pr(b \mid \bar{a})$.

We now reconsider the classification problem in the setting of coherence theory, and model b given a and b given $\bar{a}$ as zero probability events by setting $\Pr(b \mid a) = \Pr(b \mid \bar{a}) = 0$; all other parameters of our classifier are as above. For computing probabilities of interest, the classifier's parameter information is first expressed in terms of the following constituent probabilities:

$$\begin{array}{llll} x_1^0 = \Pr(abc) & x_3^0 = \Pr(a\bar{b}c) & x_5^0 = \Pr(\bar{a}\bar{b}c) & x_7^0 = \Pr(a\bar{b}\bar{c}) \\ x_2^0 = \Pr(\bar{a}bc) & x_4^0 = \Pr(ab\bar{c}) & x_6^0 = \Pr(\bar{a}b\bar{c}) & x_8^0 = \Pr(\bar{a}\bar{b}\bar{c}) \end{array}$$

The parameter information now gives rise to the system S^0 for level zero of the sequence of systems:

$$S^0 = \begin{cases} x_1^0 + x_4^0 \;=\; 0 \cdot (x_1^0 + x_3^0 + x_4^0 + x_7^0) & [\Pr(b \mid a) = 0] \\ x_2^0 + x_6^0 \;=\; 0 \cdot (x_2^0 + x_5^0 + x_6^0 + x_8^0) & [\Pr(b \mid \bar{a}) = 0] \\ x_1^0 + x_3^0 + x_4^0 + x_7^0 \;=\; 0.8 \cdot (x_1^0 + x_2^0 + x_3^0 + x_4^0 + x_5^0 + x_6^0 + x_7^0 + x_8^0) & [\Pr(a \mid \top) = 0.8] \\ x_1^0 + x_3^0 \;=\; 0.4 \cdot (x_1^0 + x_3^0 + x_4^0 + x_7^0) & [\Pr(c \mid a) = 0.4] \\ x_2^0 + x_5^0 \;=\; 0.2 \cdot (x_2^0 + x_5^0 + x_6^0 + x_8^0) & [\Pr(c \mid \bar{a}) = 0.2] \\ x_1^0 + x_2^0 + x_3^0 + x_4^0 + x_5^0 + x_6^0 + x_7^0 + x_8^0 \;=\; 1 & [\Pr(\top) = 1] \\ x_i^0 \;\geq\; 0,\ i = 1, \ldots, 8 \end{cases}$$

and has the following unique solution:

$$\begin{array}{llll} x_1^0 = 0 & x_3^0 = 0.32 & x_5^0 = 0.04 & x_7^0 = 0.48 \\ x_2^0 = 0 & x_4^0 = 0 & x_6^0 = 0 & x_8^0 = 0.16 \end{array}$$

Since from the computed solution the probability of interest $\Pr(a \mid b\bar{c}) = x_4^0/(x_4^0 + x_6^0)$ cannot be established, level one of the sequence of systems is considered. This level pertains to the zero constituent probabilities x_i^0 from level zero, which will be renamed, for all i involved, to x_i^1 to indicate the new level. As will be elaborated upon in Section 4, at level one we have to also include the information that the feature variables B and C are independent given the class variable A; we note that this information was entailed before by the equations of the system S^0, and therefore did not need to be considered explicitly at level zero. The following system of equations S^1 now results:

$$S^1 = \begin{cases} x_1^1 = 0.4 \cdot (x_1^1 + x_4^1) & [\Pr(c \mid ab) = 0.4] \\ x_2^1 = 0.2 \cdot (x_2^1 + x_6^1) & [\Pr(c \mid \bar{a}b) = 0.2] \\ x_1^1 + x_2^1 + x_4^1 + x_6^1 = 1 & [\Pr(\top) = 1] \\ x_i^1 \geq 0,\ i = 1, 2, 4, 6 \end{cases}$$

in which we encoded the standard notion of conditional independence used in naive Bayesian classifiers. The system S^1 has a convex polytope of solutions, with the following extreme points:

$$(x_1^1 = x_4^1 = 0, x_2^1 = 0.2, x_6^1 = 0.8) \qquad (x_2^1 = x_6^1 = 0, x_1^1 = 0.4, x_4^1 = 0.6)$$

From this polytope, we find for the probability of interest that $\Pr(a \,|\, b\bar{c}) = x_4^1/(x_4^1 + x_6^1) \in [0, 1]$. The vacuousness of this conditional probability conveys the information that, without any further knowledge of the comparative relation between the two zero probabilities $\Pr(b \mid a)$, $\Pr(b \mid \bar{a})$, nothing more concrete can be concluded about the probability of interest. □

4. Exploiting Coherent Conditional Probability Theory in Naive Bayesian Classifiers

Having illustrated the advantages of modelling extreme probability features through coherent conditional probability theory, we now extend beyond the simple examples with two feature variables only and describe in Section 4.1 how the theory is applied in naive Bayesian classifiers in general. In Section 4.2 we will illustrate how the framework of coherent conditional probability allows the incorporation of additional information about the extreme probabilities in a classifier.

4.1 Expressing classifier information in linear equations

The first step of framing a naive Bayesian classifier in the context of coherent conditional probability is to express the classifier's embedded probabilistic information in terms of the constituent probabilities over the variables involved. This information is composed of:

- the parameter distribution $\Pr(C)$ over the class variable C, and the parameter distributions $\Pr(F_i \,|\, C)$ over the feature variables F_i, $i = 1, \ldots, n$, given the class variable;
- the conditional independence $F_i \perp\!\!\!\perp F_j \mid C$ for all pairs of feature variables F_i, F_j, $i, j = 1, \ldots, n$, $i \neq j$, given the class variable C.

For the system of equations S^0 at level zero of the sequence of compatible systems, the class probabilities $\Pr(c^*)$ of the classifier are expressed in two linear equations of the following form:

$$\sum_{i=1}^{2^{n+1}} a_i^0 \cdot x_i^0 \;\; = \;\; p^* \cdot \Big(\sum_{i=1}^{2^{n+1}} b_i^0 \cdot x_i^0 \Big)$$

where the term $\sum_{i=1}^{2^{n+1}} a_i^0 \cdot x_i^0$ expresses the parameter $\Pr(c^*)$ in its constituent probabilities x_i^0 by means of appropriate indicator coefficients $a_i^0 \in \{0, 1\}$, $i = 1, \ldots, 2^{n+1}$, and the term $\sum_{i=1}^{2^{n+1}} b_i^0 \cdot x_i^0$ encodes the probability of the universal truth by means of the indicator coefficients $b_i^0 \in \{0, 1\}$, $i = 1, \ldots, 2^{n+1}$; we note that the coefficients b_i^0 are equal to 1, with the exception of the coefficients pertaining to atoms involving a logical impossibility. The numerical values $p^* = \Pr(c^*)$ for the two classes c and $\bar{c}$ are taken from the classifier's specification. The four conditional probabilities $\Pr(f_j^* \mid c^*)$ for a feature variable F_j are encoded similarly, yielding linear equations of the form:

$$\sum_{i=1}^{2^{n+1}} a_i^0 \cdot x_i^0 \;\; = \;\; p^{**} \cdot \Big(\sum_{i=1}^{2^{n+1}} b_i^0 \cdot x_i^0 \Big)$$

where $\sum_{i=1}^{2^{n+1}} a_i^0 \cdot x_i^0$ now expresses the joint probability $\Pr(f_j^*, c^*)$ in its constituent probabilities and $\sum_{i=1}^{2^{n+1}} b_i^0 \cdot x_i^0$ encodes $\Pr(c^*)$; the numerical values p^{**} of the four probabilities $\Pr(f_j^* \mid c^*)$ are again taken from the classifier's specification.

As for any Bayesian network in general, the parameter distributions per variable of a naive Bayesian classifier are algebraically independent, in the sense that any such distribution can be numerically specified independently of all other distributions. This property guarantees that the linear equations expressing a classifier's parameter probabilities are jointly coherent and, hence, allow at least one solution. In general, these equations allow multiple solutions for the constituent probabilities involved, since for a unique solution to be guaranteed, information about the conditional probabilities $\Pr(F_i, F_j \mid C)$ for any two feature variables F_i, F_j, $i \neq j$, for example, is lacking. We recall however, that in the simple examples of the previous section, the systems S^0 at level zero all had unique solutions, even though they were built from just parameter probabilities. This uniqueness originated from the only pair of feature variables involving zero probabilities, which effectively served to numerically enforce standard conditional independence of these two variables given C. The independences expressed by the tree structure of the classifiers therefore were entailed by the linear equations derived from the parameter probabilities and the encoded information thus exactly matched the probability distributions of the classifiers under study.

The independence information of a naive Bayesian classifier in general takes the form of a set of conditional independences $F_i \perp\!\!\!\perp F_j \mid C$ for all pairs of feature variables F_i, F_j, $i, j = 1, \ldots, n$, $i \neq j$, given the class variable C. This information is expressed in terms of constituent probabilities in equations of the following form for the system S^0 at level zero of the sequence of systems:

$$\sum_{i=1}^{2^{n+1}} a_i^0 \cdot x_i^0 \;=\; p^{**} \cdot \Big(\sum_{i=1}^{2^{n+1}} b_i^0 \cdot x_i^0 \Big)$$

with $p^{**} = \Pr(f_j^* \mid c^*, \mathbf{h})$, for each feature variable F_j and all value combinations $\mathbf{h}$ for all subsets of feature variables $\mathbf{H} \subseteq \mathbf{F} \setminus \{F_j\}$, $j = 1, \ldots, n$; in the above equation, the term $\sum_{i=1}^{2^{n+1}} a_i^0 \cdot x_i^0$ expresses the probability $\Pr(f_j^*, c^*, \mathbf{h})$ in its constituents by means of appropriate indicator coefficients a_i^0, and $\sum_{i=1}^{2^{n+1}} b_i^0 \cdot x_i^0$ similarly encodes the probability $\Pr(c^*, \mathbf{h})$. For each possible value combination $\mathbf{h}$, the numerical values p^{**} for the four probabilities $\Pr(f_j^* \,|\, c^*, \mathbf{h})$ are taken as the numerical values of $\Pr(f_j^* \,|\, c^*)$ from the classifier. We note that the thus resulting system of equations explicitly expresses all independences represented by the tree structure of the classifier, regardless of whether they are already entailed by the parameter information and of whether they are actually needed for establishing a probability of interest; upon practical application, therefore, not all constructed equations may need to be included in all levels of the sequence of compatible systems. We further note that, while conditional independence in coherence theory not necessarily satisfies the property of symmetry (Vantaggi, 2001), our explicit encoding of the tree structure of a naive Bayesian classifier enforces the standard symmetric notion of conditional independence to hold.

With the encoding of the probabilistic information of a naive Bayesian classifier as described above, the resulting system of equations S^0 at level zero of the sequence of compatible systems $\mathcal{S}$ is guaranteed to be coherent and to have a unique solution. The latter property derives from the observation that the unknowns x_i^0 are the constituent probabilities of the distribution defined by the classifier with its independence structure. Since any system S^i at a deeper level $i > 0$ is constructed to be compatible with the system at level $i - 1$, coherence is guaranteed throughout $\mathcal{S}$.

4.2 Including additional information about extreme probabilities

Upon computing conditional probabilities over the class variable of a naive Bayesian classifier for cases involving zero probability features, the solution space of agreeing distributions may be too large to warrant an informative decision. The setting of coherence theory now readily allows the inclusion of additional information to further constrain this solution space; such information may be available from domain experts or from literature. We illustrate the basic idea by means of an example, and refer to (Capotorti, 2005) for a more in-depth discussion of the types of information that can be dealt with in the setting of coherent conditional probability in general.

Example 3. We consider again the problem from Example 2 in Section 3, in the setting of coherence theory. We suppose that we know, for example from literature, that in case the zero probability feature b is observed, the presence of A is at least as likely as its absence, that is, we know that $\Pr(a \mid b) \geq \Pr(\bar{a} \mid b)$. To incorporate this additional information, the linear equation

$$x_1^0 + x_4^0 \;\geq\; x_2^0 + x_6^0$$

is inserted into the system of equations S^0 at level zero of the sequence of compatible systems. The thus extended system S_+^0 is still coherent and its unique solution of constituent probabilities is the

same as that of the original system. Because the extra equation pertains to probabilities conditioned on a zero probability event, it is inherited by the system S^1_+ at level one:

$$S^1_+ = \begin{cases} x^1_1 + x^1_4 \;\geq\; x^1_2 + x^1_6 & [\Pr(a \mid b) \geq \Pr(\bar{a} \mid b)] \\ x^1_1 \;=\; 0.4 \cdot (x^1_1 + x^1_4) & [\Pr(c \mid ab) = 0.4] \\ x^1_2 \;=\; 0.2 \cdot (x^1_2 + x^1_6) & [\Pr(c \mid \bar{a}b) = 0.2] \\ x^1_1 + x^1_2 + x^1_4 + x^1_6 \;=\; 1 & [\Pr(\top) = 1] \\ x^1_i \;\geq\; 0,\; i = 1, 2, 4, 6 & \end{cases}$$

With the extra information, S^1_+ has a polytope of solutions with the following extreme points:

$$\begin{array}{ll} (x^1_1 = 1, x^1_2 = x^1_4 = x^1_6 = 0) & (x^1_1 = x^1_2 = 0.5, x^1_4 = x^1_6 = 0) \\ (x^1_4 = 1, x^1_1 = x^1_2 = x^1_6 = 0) & (x^1_1 = x^1_6 = 0.5, x^1_2 = x^1_4 = 0) \\ & (x^1_2 = x^1_4 = 0.5, x^1_1 = x^1_6 = 0) \\ & (x^1_4 = x^1_6 = 0.5, x^1_1 = x^1_2 = 0) \end{array}$$

from which the probability $\Pr(a \mid b)$ is found to lie within the interval $[0.5, 1]$. If the decision rule associated with the classifier would use a probability threshold δ smaller than 0.5 for assigning the class a therefore, the case b would be readily classified as belonging to a. The probability $\Pr(a \mid b\bar{c})$ is still found to be vacuous, however, as the two features b and $\bar{c}$ reside on different levels of unexpectedness and, hence, cannot be compared given the available information. □

5. Practicable Coherent Inference

Straightforward application of coherence theory for handling zero probability features in naive Bayesian classifiers, would involve the constituent probabilities of the joint distribution over a classifier's variables. As the number of constituent probabilities typically is exponential in the number of variables involved, classification would rapidly become infeasible, even with the recent MIP approach of Cozman and di Ianni (2015). Building upon the notion of *locally strong coherence*, introduced by Capotorti and Vantaggi (2002), we now propose a simple computational scheme for practicable coherent inference in naive Bayesian classifiers with zero probability features.

We consider a naive Bayesian classifier with the class variable C and the set $\mathbf{F}$ of feature variables, as before. We indicate by $\mathbf{F}_0$ the subset of variables modelling zero probability features, and use $\mathbf{F}_n = \mathbf{F} \setminus \mathbf{F}_0$ to denote the remaining set of (standard) feature variables. For a given complete case $\mathbf{f}$, we further partition the set $\mathbf{F}_0$ into the two subsets $\mathbf{F}^r_0(\mathbf{f})$, including the variables from $\mathbf{F}_0$ with rare values in $\mathbf{f}$, and $\mathbf{F}^e_0(\mathbf{f})$, including the remaining variables from $\mathbf{F}_0$. The case $\mathbf{f}$ is now written as $\mathbf{f} = \mathbf{f}_n\, \mathbf{f}^r_0\, \mathbf{f}^e_0$, that is, as composed of three subcases with value combinations for the sets $\mathbf{F}_n$, $\mathbf{F}^r_0(\mathbf{f})$ and $\mathbf{F}^e_0(\mathbf{f})$, respectively. Our computational scheme for coherent probabilistic inference with the classifier given a complete case $\mathbf{f}$ now involves the following two rules:

1. If $\mathbf{F}^r_0(\mathbf{f}) = \varnothing$, then the probability distribution $\Pr(C \mid \mathbf{f})$ of interest is computed using standard probabilistic inference.

2. If $\mathbf{F}^r_0(\mathbf{f}) \neq \varnothing$, then the conditional probability distribution $\Pr(C \mid \mathbf{f})$ of interest is computed by means of the operational tool of coherence theory, from the level in the sequence of systems where the subcase $\mathbf{f}^r_0$ resides; for this purpose, the appropriate system of equations is

constructed in terms of atoms in which the context $\mathbf{f}_n\, \mathbf{f}_0^e$ is fixed and only the class values and the values of the feature variables from $\mathbf{F}_0^e(\mathbf{f})$ differ between atoms.

We note that, by the semantics of a feature being rare, the first rule will apply to the majority of cases offered to the classifier; for all these cases therefore, the computations involved are linear in the classifier's number of variables. The second rule assumes locally strong coherence of the distribution defined by the classifier with respect to the context of non-rare features of a case. This assumption allows the computations involved to be local to the appropriate level in the sequence of systems, by fixing this context of non-rare features; all previous levels of the sequence can then be neglected. By building upon the assumption of locally strong coherence, therefore, the computations involved are exponential only in the number of rare features involved in a case at hand.

6. Concluding Observations

Despite their general popularity, naive Bayesian classifiers are not well suited for real-world applications involving extreme probability features. In this paper, we have demonstrated, by means of simple examples, some unwanted effects of the commonly used approaches to handling zero probability features. These effects are the strong influence on output probabilities of the more or less arbitrary small values used to forestall the inclusion of zero probability parameters in a classifier on the one hand, and the counterintuitive consequences of taking absence of a zero probability feature into consideration on the other hand. We have shown that the theory of coherent conditional probability offers a principled approach to handling extreme probability features in naive Bayesian classifiers and is not associated with these unwanted effects. For practicable application of this theory, we have further proposed a computational scheme for naive Bayesian classifiers with zero probability parameters, which retains the inferential efficiency of standard classifiers for input cases with a limited number of extreme probability features.

While standard naive Bayesian classifiers always establish a single output probability per class for an input case, the examples throughout the current paper have illustrated that the naive Bayesian classifier building upon coherence theory for handling zero probability features, does not. Upon computing conditional probabilities of interest, the classifier may find that the available probabilistic information does not suffice for establishing a single output probability for a case at hand and then returns an interval probability. By employing coherence theory therefore, the naive Bayesian classifier with zero probability features introduces in a natural way the notion of imprecision. From the field of imprecise probability, related approaches to classification have originated, among which is the naive credal classifier initiated by Zaffalon (2002). In our further investigations, we will study the similarities and complementarities of this credal classifier and the approach proposed in the current paper. We will further investigate the effects of relinquishing the assumption of symmetric conditional independence for zero probability features and adopting the notion of conditional independence from coherence theory (Vantaggi, 2001). Last but not least, we plan to study the practicability of our approach in real-world applications and thereby to also further investigate the strengths of different types of additional knowledge about the extreme probability features involved.

Acknowledgement The research reported in this paper was conducted during a sabbatical stay of the first author at Università di Perugia, Italy.

References

A. Capotorti. Benefits of embedding structureal constraints in coherent diagnostic processes. *International Journal of Approximate Reasoning*, 39:211–233, 2005.

A. Capotorti and B. Vantaggi. Locally strong coherence in inference processes. *Annals of Mathematics and Artificial Intelligence*, 39:125–149, 2002.

H. Chan and A. Darwiche. When do numbers really matter? *Journal of Artificial Intelligence Research*, 17:265–287, 2002.

G. Coletti and R. Scozzafava. *Probabilistic Logic in a Coherent Setting*. Kluwer Academic Publishers, Dordrecht, 2002.

F. G. Cozman and L. F. di Ianni. Probabilistic satisfiability and coherence checking through integer programming. *International Journal of Approximate Reasoning*, 58:57 – 70, 2015.

B. de Finetti. *Probability, Induction and Statistics: The Art of Guessing*. John Wiley & Sons, London, 1972.

P. Domingos and M. J. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

P. Flach. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29: 131–163, 1997.

L. C. van der Gaag. Computing probability intervals under independency constraints. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pages 491 – 497, Cambridge, Massachusetts, 1990.

B. Vantaggi. Conditional independence in a coherent finite setting. *Annals of Mathematics and Artificial Intelligence*, 32(1):287–313, 2001.

I. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105:5–21, 2002.