Extension of Attribute Information based on Naive Bayes Classifiers for Small Data Set Classification

1Prof.Mr.M.Vazralu M.Tech(Phd), 2V. Radha
radhavmtech@yahoo.in

Abstract- In this paper, the naive Bayes classifier, which is a set-valued counterpart of naive Bayes, is extended to a general and flexible treatment of incomplete data, yielding a new classifier called naïve bayes. The new classifier delivers classifications that are reliable even in the presence of small sample sizes and missing values. Extensive empirical evaluations show that, by issuing set-valued classifications able to isolate and properly deal with instances that are hard to classify and to perform as well as naive Bayes on the other instances. The experiments point to a general problem: they show that with missing values, empirical evaluations may not reliably estimate the accuracy of a traditional classifier, such as naive Bayes.

I. Introduction

A more objective-minded model of prior ignorance has been proposed extension of naive Bayes classifier (NBC) to imprecise probabilities [12]. NBC models prior ignorance by a set of prior densities (also called prior credal set), which is turned into a set of posteriors by element wise application of Bayes’ rule. The classification is eventually issued by returning all the classes that are non-dominated by any other class according to the posterior credal set, where class ci is said to dominate cj if for all the posteriors it holds that the probability of ci is larger than that of cj. This makes NBC naturally issue set-valued classifications (i.e., classifications made by more than one class) when faced with instances that are hard to classify, due to a combination of prior ignorance and poor information about those specific instances in the learning set. The shift of paradigm based on set-valued classifications allows NBC to deliver robust classifications in spite of small learning sets. NBC has indeed shown excellent accuracy in real-world case [7], thus demonstrating the usefulness, for classification purposes, of modeling prior ignorance via a credal set. In the following, set-valued classifications are also called indeterminate. Determinate classifications correspond instead to the set being a singleton, and hence to the case usually considered by more traditional classifiers. Similarly, we say that a classifier is determinate when it outputs a single class and indeterminate otherwise. As for the ignorance arising from missing data, we can think of the missingness process (MP) as a process that takes in input the complete data, which we cannot usually observe, and outputs the incomplete data, which we do observe. If data are our only source of information, we are ignorant about the MP because it is usually not possible to learn how it operates, from the observed, incomplete data. In common practice, missing values are often ignored; this entails the idea that the MP is nonselective in producing them, or, in other words, that it is a random (MAR) [17]. However, if one is ignorant about the MP, assuming MAR cannot be regarded as an objective-minded approach, as is well documented, for instance, by Manski (2003). In its original formulation, NBC introduced also an initial attempt to deal with ignorance about the MP. The idea was to model ignorance about it by using a set of likelihoods: a likelihood per each complete learning set consistent with the incomplete one. A similar avenue was also implemented by robust Bayes classifier . These approaches are indeed valuable, but have two problems: (i) they implicitly still assume MAR for the missing values in the instance to classify, thus creating a peculiar asymmetry between learning and test set that is not of general validity in applications; (ii) they may well be too conservative, because for some feature variables one might know that the missingness is MAR, and they do not allow this information to be incorporated in the model. Furthermore, their treatment of missing values rests on intuitive arguments rather than on a principled derivation. we extend NBC to a very general and flexible treatment of incomplete data, both in learning and testing. We call the resulting classifier naïve credal classifier 2 (NBC2), in order to emphasize the advancement made to deal with incomplete data, while keeping the original benefits of NBC on the front of prior ignorance. By NBC2, it is possible to declare that some (possibly all or none) of the feature variables are subject to a MAR process, and the remaining ones are automatically assumed to be subject to an MP that is unknown...
to us. Remarkably, the set of feature variables subject to a MAR MP can be chosen differently from the learning to the test set. This is a key characteristic of NBC2: in fact, if the MP is unknown, it may well change its behavior from unit to unit for all we know (i.e., it may not be identically distributed), and we should act accordingly. The development of NBC2 is based on a recently derived so-called conservative inference rule (CIR) to compute (imprecise) conditional expectations with incomplete data [22]. After giving some notation and briefly specializing CIR to the case of naive classification. In the end we obtain procedures to learn NBC2 and to do classifications with it that do not involve approximations and are computationally fast. (The software which implements NBC2 is released as open source; more details are provided in Next, we concentrate on empirical evaluations: in Section 4 we analyze the behavior of NBC2 from a number of angles and on a number of publicly available data sets. The analysis turns out to be particularly meaningful when we compare NBC2 with its precise-probability counterpart, that is, naive Bayes. We do this by evaluating the accuracy of NBC on the instances of the test set where NBC2 issues a determinate classification separately from those where it does not. In fact, NBC2 is indeterminate on an instance when it deems that there is not enough knowledge in the learning set to make a determinate classification reliably; NBC, on the other hand, issues a determinate classification on such an instance (as well as on any other). Therefore we expect NBC to have different behaviors on the two kinds of instances isolated by NBC2. And this is indeed the case: the experiments show that NBC undergoes a major drop in accuracy moving from the instances classified in a determinate way by NBC2 to the indeterminate ones. The drop is observed on every data set, with no exception. It is important to realize that such a drop points out a key question: the usual way to measure the performance of a classifier, that is, its predictive accuracy, which is an average over all the instances of the test set, may not help uncover a possible bad performance of the classifier on a subset of the test instances. These instances are precisely those that are hard to classify and that NBC2 isolates by delivering set-valued classifications. But set-valued classifications help NBC2 to do more than just isolating the hard instances, they enable it to cope effectively with them: in fact, we show that set-valued classifications are often informative, as they usually lead to drop some unlikely classes; and that the measured set-based accuracy of NBC2 (i.e., the proportion of times the true class is contained in the output set) is often similar to the accuracy obtained on the instances classified in a determinate way. At this point we should say that the mentioned experiments have been carried out with a variety of settings, obtained considering both MAR processes and non-MAR ones, and the mentioned outcomes are confirmed over all of them (although sometimes this is due more to prior ignorance and some others more to the missing data). We have considered such posterior probabilities again by separating the cases where NBC2 is determinate from the others, and by comparing those probabilities with the measured accuracy that NBC actually achieves on the data. What we show is that the NBC probabilities are (also very) unreliable on the instances that are hard to classify, as isolated by NBC2, and definitely more unreliable than on the remaining instances. In other words, we observe another kind of drop that now is related to the quality of the posterior probabilities computed by NBC. Overall, we show that NBC may well be too optimistic in dealing with small data sets and missing data, thus yielding unreliable predictions. It is useful to recall that NBC is known to be very robust to missing data. Therefore, it is not unlikely that the optimism on the front of missing data is even greater with more complex classifiers. This point appears to be worth of serious consideration on its own. At the same time, and in contrast with naive Bayes, our experiments show that NBC2 may sometimes be too pessimistic (i.e., conservative) especially when dealing with missing data. This happens because by construction NBC2 implicitly considers the worst possible MP to have acted on the non-MAR part of the data, and in some cases this hypothesis may be too far from the MP that has actually produced the missing values.

II. Related Work

classification is to predict the class of the M-th unit, given the previous units (1......N) and the values of the M-th attribute variables. To this extent, a traditional probabilistic classifier outputs what it deems to be the optimal prediction: that is, the class with the highest probability (in the case of 0-1 loss function) on the basis of a uniquely computed posterior density. In the imprecise setting, however, the optimality criterion has to be extended to
manage a set of posterior densities (derived from a set of priors and a set of likelihoods), instead of a single posterior; in particular, according to [21], the optimality criterion in the imprecise setting prescribes to return the non-dominated classes. The definition of dominance is as follows: class $c_i$ dominates $c_j$ if for all the computed posterior densities, the posterior probability of $c_i$ is greater than that of $c_j$; clearly, $c_j$ is non-dominated if no class dominates $c_j$. The second procedure, based on pairwise comparison of classes, identifies the non-dominated classes. Observe that, as a result of the uncertainty arising from both prior specification and non-MAR missing values, there can be several non-dominated classes; in this case, the classifier returns an indeterminate (or set-valued) classification. Classifiers that issue set-valued classifications are called credal classifiers[16]. A key point is that non-dominated classes are incomparable:6 this means that there is no information in the model that allows us to rank them. In other words, credal classifiers are models that allow us to drop the dominated classes, as sub-optimal, and to express our indecision about the optimal class by yielding the remaining set of non-dominated classes. the test of dominance can be re-written as follows: $c_00$ is dominated by $c_0$ if and only if it holds that

$$1 < \min_{x_0 \in \mathcal{M}} \min_{d_0} \inf_{p(0) \in \mathcal{P}(0)} \frac{p(c_i|d_i \in \mathcal{D}, x^c \subset \hat{\mathcal{O}})}{p(c_j|d_j \in \mathcal{D}, x^c \subset \hat{\mathcal{O}})}$$

the general form of the test of dominance for any classifier based on the conservative inference rule presented [19]; CIR is a conditioning rule (i.e., a rule for computing conditional expected values) that generalizes the traditional conditioning; it assumes that prior beliefs are dealt with via a credal set $\mathcal{P}(\cdot)$ and it accounts for data sets in which the missingness process is MAR for some variables (the term "x+ 2"o+ refers indeed to the missing data of MAR feature variables in the training set), and unknown for some others. Moreover, CIR is able to manage variables whose MP is MAR in learning and unknown in testing, or vice versa, which we obtain considering any precise prior in the prior credal set, as well as any completion of the non-MAR missing values both in the sample and in the unit to classify. CIR can be regarded as unifying two rules [13] a conservative learning rule, which prescribes how to learn the classifier from an incomplete training set, and a conservative updating rule, which prescribes how to classify a novel instance that contains missing values. Such a distinction is made clear by two distinct optimization loops of the middle optimization loop (mind2o) realizes the conservative learning rule, by prescribing to loop on the completions of the non-MAR part of the learning set, that is, d 2 o, while the outer minimum implements the conservative updating rule, prescribing to loop on the replacements for the non-MAR missing values of the unit to classify. The inner loop, which minimizes over the prior credal set, is common to both learning and updating rules. NBC2 specializes the test of the case of naive classification. In the following, we will move from the precise setting (corresponding in fact to naive Bayes) to NBC2. we describe the precise setting (assuming hence that there is a single prior, and that there is a single likelihood as non-MAR data are complete) we finally relax the assumptions of completeness about non-MAR data in learning and testing respectively, thus managing a set of likelihoods and instances to classify.

III. Naive Bayes Classifiers(NBC)

we introduce the NBC2 framework. In particular, we derive an expression that specializes the test of Equation (1) to the case of naive classification; such an expression realizes both the minimizations over possible priors and over possible unobserved values by distributing some pseudo-counts in a way that minimizes the ratio of the posterior probabilities. we assume that a single prior is specified and that only the observations of the features affected by the MAR MP contain some missing data; the observations of the feature affected by the unknown MP, instead, do not contain any missing data. In practice, this corresponds to the naive Bayes setting, with the difference that we explicitly separate variables affected by the MAR MP and the unknown MP. In our setting, the naive hypothesis (i.e., the assumption of mutual independence of the latent attribute variables $A_1$, ..., $A_k$, $A_{k+1}$, ..., $A_r$ on the class variable $C_i$) can be formalized as follows:

$$\theta_{(d, x)} = \theta_{c_i} \prod_{j=1}^{k} \theta_{a_j|c_i} \prod_{l=1}^{r} \theta_{d_l|c_i} \forall (d, x) \in \mathcal{D} \times \mathcal{X},$$

The underlying idea of such a procedure is to split the outer minimum in each one related to a different feature variable, and to distribute them
at different places into the objective function. This makes it clear that the function to optimize is the lower envelope of a set of convex functions. We show that it is easy to compute the points where the function that determines the envelope changes, thus in fact obtaining a partition of the envelope function's domain with the property that in each of its elements the envelope function is convex. At that point, the function can be locally optimized efficiently on every element of the partition; and the global optimum is then just the minimum of the local optima so computed. The fact that the overall procedure is polynomial follows because the size of the partition is bounded by a polynomial.

A final remark is that, to make notation simpler, we refer to the possible realizations of the non-MAR variables in the instance to be classified as oM; such a notation implies however that the non-MAR variables of the test set are the same of the training set. If instead the non-MAR variables of the test set are different from those of the training set, oM should contain all the possible realizations of the variables which are non-MAR in the test set. The NBC2 procedures are summarized [10]. Learning has linear complexity with respect to the number of attributes, while testing has roughly quadratic complexity, if the procedure carried out in Appendix A is adopted. Please refer to Appendix A also for the exact expression for the complexity.

IV. Performance Evolution

Here Taken 18 data sets from the UCI repository, and available in the ARFF format from the WEKA data sets page. All the data sets are complete, that is, they do not contain any missing data. We discretized the numerical features via MDL-based discretization (Fayyad and Irani, 1993), and then we split each data set into a training and a test set. Feature discretization is necessary, as NBC2 is designed to work with categorical variables. Discretizing the features on the entire data set introduces a slight optimistic bias in the evaluation of the classifiers accuracy (in principle, the discretization intervals should be computed on the training set, and then applied unchanged in the test set); yet, this is not a problem as our goal is to compare NBC and NBC2 in identical conditions, rather than to compare our findings with previous results obtained on the same data sets. Also, since our goal is to fairly compare NBC and NBC2, and not to finely tune them for maximum performance, we did not perform feature selection (although, in fact, we remove numerical feature variables discretized into a single bin); yet, as both NBC and NBC2 are based on the naive hypothesis, redundant or mutually dependent feature variables might significantly bias the learning process; hence, in order to achieve maximum performance, one should consider selecting feature variables. In the experiments presented in the following we generate artificial missingness on the original, complete data sets by using different MPs and then we compare NBC andNBC2 accuracy on the incomplete data sets. We consider two different artificial MPs: (i) a MAR one, and (ii) a non-MAR, non-identically distributed one (nonMAR). The MPs we consider do not affect the class variable, as the class variable has been assumed to be always observed. Note that we do not test mixed cases of MAR and non-MAR feature variables, which NBC2 is actually designed to treat. In fact, such settings would simply lead to results intermediate between those obtained under the MAR and the non-MAR settings. Yet, mixed settings are valuable and should be considered whenever possible when operating the classifier, as they allow for finely tuning the treatment of missing data to the characteristics of the MP. The MAR MP turns into missing, with 5% probability, all the features of both the training and test sets. Such a missingness process actually meets not only the definition of MAR, but also the more restrictive definition called MCAR, that is, missing completely at random, see Little and Rubin (1987). We have taken into consideration also a non-MCAR, MAR MP; however, the results do not differ significantly from those obtained with the MAR MP (which satisfies also MCAR) described above. The non-MAR MP works as follows: (i) it splits the categorical values of each feature variable into two halves; (ii) for each feature variable, it turns into missing, with probability 5%, the observations falling in the first half of values, on the training set; (ii) for each feature variable, it turns into missing, with probability 5%, the observations falling in the second half of values, in the test set. Such MP is not identically distributed, as it follows a different pattern from training to test set. We split each data set into equally sized training and test subsets. Using this training/test split, for each data set and for each MP we generate artificial missingness 100 times, producing hence 100 different training and test sets. The results we present are obtained as an average over 100 runs for each data set-MP pair. In this section...
we consider all the usual 18 data sets with the same training/test splits. That is, 50%-50%.

We consider two settings: the first compares NBC and NBC2-MAR when the MP is MAR, and the second compares NBC and NBC2-nonMAR when the MP is non-MAR. We analyze jointly the predictions issued on all the data sets. One point to consider is that in data sets with two classes, one can spot the instances doubtful for NBC looking at those with probability around 50% for the returned class. However, for data sets with more than two classes, the instances that are doubtful for NBC are not as easy to recognize as before; for instance, with four classes both the posterior mass functions [40%, 40.5%, 15%, 4.5%] and [24.5%, 25.5%, 25%, 25%], lead to doubtful classifications for NBC. For this reason, we only focus now on the case when NBC is confident, that is, on the instances with probability for the returned class greater than or equal to 55%. Looking at the figures we see that the determinacy is much higher in the MAR setup than in the non-MAR; in fact, we have already seen that the non-MAR case leads in general to a much larger amount of instances that are isolated as difficult ones. Nevertheless, to clear drop of accuracy of NBC, for the same level of posterior probability of the predicted class, from the instances classified in a determinate way by NBC2 to the others. The drop is especially striking on the instances classified confidently by NBC, when the computed probability is for instance larger than 70%. This is more evident in the other hand, the drop observed to a much larger set of instances. In fact, NBC2-nonMAR suspends the judgment frequently also on the instances classified by NBC with probability greater than 80%; the reason is that a replacement for missing data especially unfavorable for the class predicted by NBC, can well change the outcome of the classification with respect to the output of NBC, which instead marginalizes out the missing feature variable. The instances for which NBC returns a probability higher than 90% are of particular interest, also because they constitute 62% of the total instances. On this area, NBC2-MAR returns 1.5% of indeterminate classifications on which it achieves the following remarkable performance: NBC2- MAR D: 94%; NBC2-MAR I: 54%. The performance of NBC2-nonMAR in this area is as follows: determinacy 66%, NBC2-NonMAR D: 96%; NBC2-NonMAR I: 89%. The drop is in this case significant, yet smaller than under MAR; in fact, as already pointed out NBC2-nonMAR is more conservative than necessary under this setting. Summing up, hence, uncertainty related to the choice of the prior manifest itself more evidently, but not only, on the instances in which NBC is less confident; yet, only part of such doubtful instances lead to indeterminate classifications. On the other hand, missing data treated as non-MAR can lead to many indeterminate classifications even if the probability computed by NBC for the returned class is high.

V. Conclusion

In this paper the different amounts of information in the data that a classifier can exploit to classify different units. This depends in part of the fact that a learning set can be more informative about some units than some others, and in part on the type and amount of missing values in the units to classify. As a consequence, a classifier’s predictions may be more uncertain on some units than on some others. In this paper we have tried to pursue such a goal just by weakening the assumptions that are traditionally made by classifiers. We have modified the naive Bayes classifier so as to model prior ignorance in an objective-minded way by using a set of prior densities, and by giving the opportunity to treat some of the missing values as originated by a missingness process that we do not know (and the others by a MAR one). The resulting model, called naive credal classifier 2, departs from the more traditional classifiers in a number of ways, the more substantial one being the fact that...
NBC2 makes set-valued classifications in general: it issues a determinate classification (i.e., a singleton) only when it deems that there it has enough information to do so. Extensive empirical evaluations have shown that NBC2 has high accuracy when it issues determinate classifications, and that when it is more cautious, it is very often justified: NBC is clearly shown to considerably decrease its classification accuracy on the instances classified in a set-valued way by NBC2, as well as its ability to compute predictive posterior probabilities. This indicates that set-valued classifications do indeed isolate the instances of the test set that are hard to classify. And they actually do more than just isolating them: they are in fact still informative, as unlikely classes are dropped anyway, and invite the domain expert (for instance, a doctor that has to issue the ultimate diagnosis) to avoid over-confident statements. We have also pointed out that in some extreme cases of missingness processes, the empirical evaluations made for naive Bayes can be completely biased, so that not even its average predictive accuracy can be evaluated reliably. In these cases, NBC2 was instead still reliably evaluated.

References


V. Radha is studying M.Tech. Prepared with the steamed Guidance of Prof. Mr. M. Vazralu.