NAIVE BAYESIAN CLASSIFIERS FOR THE CLINICAL DIAGNOSIS OF
CLASSICAL SWINE FEVER

P.L. GEENEN1, L.C. VAN DER GAAG1, W.L.A. LOEFFEN2 AND A.R.W. ELBERS2

SUMMARY

Naive Bayesian classifiers have been successfully applied for solving diagnostic problems in the medical domain, but are relatively new to the veterinary field. To demonstrate their potential, we constructed naive Bayesian classifiers for discriminating between Classical Swine Fever (CSF) infected and non-infected herds. To this end, we used data on 490 herds, collected during the 1997/1998 CSF epidemic in the Netherlands. A full naive Bayesian classifier and a selective one were constructed, and their classification accuracies were compared to that of a previously published diagnostic rule. The full classifier had a higher accuracy than the diagnostic rule; the selective classifier proved to be comparable to the rule. In contrast with the diagnostic rule, the two classifiers had the advantage of taking both the presence and the absence of clinical signs into account, which resulted in more discriminative power.

INTRODUCTION

Naive Bayesian classifiers have proven to be powerful tools for solving classification problems in a variety of domains. They have been successfully applied in the medical domain for solving diagnostic problems, such as the diagnosis of heart disease in newborn babies (Spiegelhalter et al., 1993), of dementia severity (Shankle et al., 1998), of ischaemic heart disease (Kukar et al., 1999), and of breast cancer (Butler et al., 2003). Naive Bayesian classifiers have recently also found their way to the veterinary domain which resulted in the cattle disease diagnosis system CaDDIS (McKendrick et al., 2000) and in a system for the diagnosis of scrapie (Kuncheva et al., 2004). A naive Bayesian classifier in essence is a model of a joint probability distribution over a set of stochastic variables. It is composed of a single class variable, modeling the possible outcomes for the problem under study, and a set of feature variables, modeling the domain features that provide for distinguishing between the various outcomes. In the model, it is assumed that the feature variables are mutually independent given the class variable (Friedman et al., 1997); the term ‘naive’ in fact refers to the assumption of mutual independence. Although this assumption is not always valid, naive Bayesian classifiers tend to have a good classification performance and often outperform more sophisticated models (Domingos and Pazzani, 1997).

As an example, Fig. 1 depicts the structure of a simple naive Bayesian classifier for distinguishing between CSF-infected and non-infected pig herds. In the structure, CSF denotes the class variable; the feature variables are cyanosis, ataxia and conjunctivitis. The arrows from

1 Institute for Information and Computing Sciences, Utrecht University, P.O.Box 80.089, 3508 TB Utrecht, The Netherlands. Email: petrag@cs.uu.nl
2 Dept. of Virology, Central Institute for Animal Disease Control, Wageningen UR, P.O. Box 2004, 8203 AA Lelystad, The Netherlands
CSF to the feature variables indicate that whether or not a herd is infected influences the probabilities with which cyanosis, ataxia and conjunctivitis respectively, will be observed. In addition to its graphical structure, the model includes various probabilities capturing the strengths of the diagnostic influences between the variables. These probabilities are typically estimated from data, yet in this paper we will demonstrate that they may also be obtained from the epidemiological literature. The observations of a specific herd are presented to the classifier as values for the modelled feature variables. To classify a herd in which cyanosis and conjunctivitis have been observed and no ataxia, the two positive values and the negative one are entered for the appropriate feature variables. The classifier then returns the posterior probability distribution over the class variable for the herd. For computing this posterior distribution, the classifier builds upon Bayes’ rule; in fact, the term ‘Bayesian’ refers to the prominent role of this rule of probability. From the established posterior distribution, the class with the highest probability determines the classification of the herd. Even if missing values occur, the classifier will return a probability distribution. It will then take the posterior distributions over the missing variables into account upon computing the posterior probabilities for the class variable.

![Graphical Structure](image)

**Fig. 1 The graphical structure of an example naive Bayesian classifier**

When a naive Bayesian classifier is constructed to cover all features that have been gathered upon data collection, the resulting model is called a *full* naive Bayesian classifier. Real-life data, however, often record more features than are strictly necessary for classification. Moreover, the collected features may be correlated to some extent, thereby violating the independence assumption that underlies the classifier. Now, from among the recorded features, the more discriminative features can be distinguished from the less informative ones by means of statistical concepts. A classifier built from such a subset of discriminative features, is called a *selective* naive Bayesian classifier. These less complex classifiers tend to have a better performance in general (Langley and Sage, 1994).

In this paper, we will demonstrate how to construct a full and a selective naive Bayesian classifier from an epidemiological study, to discriminate between CSF-infected and non-infected pig herds. Clinical signs seen by the farmer or by a veterinarian are usually the first indications of a CSF-infection in a herd. Unfortunately, the clinical signs of CSF are mainly atypical and may vary from mild to severe (Van Oirschot, 1999), as a consequence of which the disease can remain undetected for weeks. Our classifiers aim to improve upon early detection and are based upon clinical signs only. For the construction of the classifiers, data on 32 clinical signs of 490 herds collected during the 1997/1998 CSF epidemic in the Netherlands were used. The data were analysed before, from which had resulted a number of diagnostic rules for the classification...
of CSF (Elbers et al., 2002). We will compare the classification accuracies of our classifiers with the reported optimally efficient diagnostic rule. We will further compare the selection of feature variables for the rule and for our selective naive Bayesian classifier.

MATERIALS AND METHODS

The dataset

The data that were used for constructing the naive Bayesian classifiers, were collected and analysed before by Elbers et al. (2002). The data record the absence or presence of 32 clinical signs for 490 herds. These herds were visited by veterinary expert teams during the 1997/1998 epidemic in The Netherlands, due to a clinical suspicion of infection. For each herd, the body temperature of the diseased pigs was measured and an anamnesis was taken. On a standardized investigation form, the presence of disease signs within the herd were recorded; if one or more pigs within the herd were observed to suffer from cyanosis for example, then this feature was marked as being present in the herd. Pigs with apparent disease signs and/or fever were euthanised and submitted to the Animal Health Service for a post-mortem examination. The tonsils and samples of the spleen, ileum and kidney were collected and sent to the Institute for Animal Science and Health in Lelystad (currently CIDC-Lelystad) to be tested by means of a CSF-specific immune fluorescence assay (IFA). Later during the epidemic, also blood samples were collected for virus isolation and antibody detection. If one or more pigs from a submission from a single herd proved to be infected with the virus, that is if one or more pigs were positive in the IFA or in the virus isolation, then the herd was diagnosed as positive for CSF. If all pigs of the submission were negative upon examination and the herd remained to be so for at least six months after the submission, then the herd was diagnosed as negative for the disease. Of the 490 herds from which a completed investigation form was available, 245 herds were diagnosed as CSF positive and 245 herds were diagnosed as negative for the disease.

From the available investigation forms, a dataset was constructed for further analysis. Upon constructing the dataset, the recorded clinical signs were encoded as ‘1’ s for the appropriate variables; signs that were not recorded explicitly were assumed to be absent and were encoded as ‘0’ s. The recorded clinical signs were relatively sparse, possibly since during the epidemic, most infected herds were detected rather early in the disease process. In the CSF-positive herds the mean number of recorded clinical signs was 3.5; for the CSF-negative herds this number was 3.0, which is significantly lower (Mann Whitney test, p < 0.01). The number of ‘1’ s in the dataset as a consequence constitutes just 10.2% of the total number of data.

The diagnostic rules

The dataset was analysed by Elbers et al. (2002). The goal of the analysis was to arrive at diagnostic classification rules composed of clinical signs, that could be used as simple diagnostic tests for establishing the presence of CSF in a herd. As the recorded clinical signs were relatively sparse, disjunctive rules were constructed: if at least one sign mentioned in such a rule is present in a herd, then the herd is classified as positive for CSF; if all signs from the rule are absent, then it is classified as CSF negative. To select the clinical signs that serve to explain most variation in the classification, logistic regression with backward selection was applied to the data, with the classification of a herd for the response variable and the clinical signs for the explanatory variables. Subsequently, all possible disjunctive rules that could be
constructed from the selected signs were evaluated using receiver operating characteristic (ROC) analysis. Moreover, three arbitrary diagnostic rules were chosen, that are epidemiological meaningful for disease detection: a rule that combines maximum sensitivity with the highest available specificity (‘optimally sensitive’), a rule with maximum specificity and the highest available sensitivity (‘optimally specific’), and a rule with maximised sensitivity and specificity (‘optimally efficient’).

Constructing naive Bayesian classifiers

Constructing a naive Bayesian classifier starts by defining the class variable with its possible values and the feature variables with their values. We created a binary class variable, modelling whether or not a herd is infected with CSF, and 32 feature variables, each modelling whether a specific clinical sign is present or absent in a herd. For the class variable, prior probabilities for the various classes discerned have to be specified; for a binary diagnostic class variable these probabilities usually reflect the prevalence of the disease for the population under study. For our classifiers, the prior probabilities were computed from the numbers of positively and negatively diagnosed herds and were established to be \( p(\text{CSF} = \text{yes}) = p(\text{CSF} = \text{no}) = 0.5 \). Our classifiers as a consequence assume that an arbitrary herd is equally likely to be infected as it is to be non-infected; we will return to this assumption in the discussion. To complete the construction of a naive Bayesian classifier, various conditional probabilities have to be obtained. For each feature variable included in the classifier, more specifically, conditional probability distributions over its values given the different classes have to be defined. For our classifiers, these probabilities were taken to be the estimated sensitivity \( p(\text{clinical sign} = \text{yes} \mid \text{CSF} = \text{yes}) \) and specificity \( p(\text{clinical sign} = \text{no} \mid \text{CSF} = \text{no}) \) for each clinical sign; these sensitivities and specificities were reported by Elbers et al. (2002). For example, for conjunctivitis these probabilities are \( p(\text{Conjunctivitis} = \text{yes} \mid \text{CSF} = \text{yes}) = 0.229 \) (sensitivity) and \( p(\text{Conjunctivitis} = \text{no} \mid \text{CSF} = \text{no}) = 0.861 \) (specificity). We would like to note that, as the reported sensitivities and specificities were established from a relatively small dataset, zero probabilities not necessarily indicate a logical impossibility of the clinical sign occurring. To prevent inconsistencies when entering the data into our classifiers, therefore, we replaced these zero probabilities by 0.0001.

While a full classifier includes all possible feature variables, a selective naive Bayesian classifier includes just a carefully selected subset of the available feature variables. Building a selective naive Bayesian classifier therefore involves singling out the feature variables that best serve to separate the different classes under study. In our study, we used the so-called filter approach for this purpose. With this approach, the selection of appropriate feature variables is performed in a pre-processing step before the classifier is actually constructed. For each feature variable, its ability to separate the various outcome classes is investigated by means of an information-theoretic criterion. Only the variables that show a high discriminative ability then will be included in the classifier under construction. In our study, the concept of mutual information was used to decide upon inclusion of a feature variable. The mutual information \( I(X,Y) \) of two variables \( X \) and \( Y \) is defined as

\[
I(X,Y) = \sum_{x,y} p(x,y) \cdot \ln \frac{p(x,y)}{p(x) \cdot p(y)}
\]

(1)
where the feature variables are taken for the variable $X$ and the class variable is taken for the variable $Y$. To decide upon inclusion of a specific feature variable, we exploited the property that the quantity $2 \cdot N \cdot J(X,Y)$ asymptotically follows a $\chi^2_{(r-1)\cdot(r_0-1)}$ distribution, where $r$ is the number of possible values of $X$ (for our classifiers $r = 2$), $r_0$ is the number of values of $Y$ (for our classifiers $r_0 = 2$), and $N$ is the number of observations (for our classifiers $N = 490$ equals the number of herds). To decide whether or not to include the feature variable under study in the classifier, we used a significance level of $\alpha = 0.01$; only feature variables $X$ for which $2 \cdot N \cdot J(X,Y) > 6.64$ were thus included. We would like to note that since $p(x, y) = p(x | y) \cdot p(y)$ and $p(x) = \sum_y p(x, y)$, all probabilities mentioned in Eq. (1) can be expressed in terms of sensitivity, specificity, and the numbers of positively and negatively diagnosed herds. The mutual information of the various feature variables with the class variable can therefore be calculated directly from the information reported in Elbers et al. (2002).

**Data analysis**

Based upon the information available from Elbers et al. (2002), a full and a selective naive Bayesian classifier were constructed as outlined above. Using the constructed classifiers and the previously published diagnostic rule respectively, each herd was classified as either CSF positive or CSF negative. From the results obtained, the sensitivity, the specificity and the overall accuracy of each of the three models were computed, where the accuracy was taken to be the fraction of correctly diagnosed herds. The established sensitivities, specificities and accuracies were compared, and the features selected for the rule and for our selective naive Bayesian classifier were evaluated. Both the classifiers and the diagnostic rule were programmed and run in the software package IDEAL; the significance tests for comparing the results obtained were performed with S-plus.

**RESULTS**

The results of our study are summarised in Table 1, which reports the sensitivities, the specificities and the overall accuracies, with their respective 95%-confidence intervals, for the full and selective naive Bayesian classifiers and for the optimally efficient diagnostic rule. The full naive Bayesian classifier was found to have a significantly higher overall accuracy than the optimally efficient rule ($p<0.05$, proportions test). The accuracy of the selective classifier proved to be comparable to that of the rule and to that of the full naive Bayesian classifier. It was further found that the optimally efficient rule had a significantly higher sensitivity and a significantly lower specificity ($p<0.05$, proportions test) than the two classifiers.

<table>
<thead>
<tr>
<th>Classification model</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
<th>Accuracy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full naive Bayes</td>
<td>0.65 (0.59 - 0.71)</td>
<td>0.73 (0.67 - 0.78)</td>
<td>0.69 (0.65 - 0.73)</td>
</tr>
<tr>
<td>Selective naive Bayes</td>
<td>0.63 (0.57 - 0.69)</td>
<td>0.67 (0.61 - 0.72)</td>
<td>0.65 (0.61 - 0.69)</td>
</tr>
<tr>
<td>Optimally efficient rule</td>
<td>0.73 (0.67 - 0.78)</td>
<td>0.53 (0.46 - 0.59)</td>
<td>0.63 (0.59 - 0.67)</td>
</tr>
</tbody>
</table>

Table 1. The sensitivities, specificities and accuracies, with their 95%-confidence intervals (95% CI) for the two constructed classifiers and the diagnostic rule.
The feature variables that were selected for the optimally efficient rule and for the selective naive Bayesian classifier respectively, are shown in Table 2. It was found that the variables modelling the clinical signs of ataxia, not responding to antibiotics treatment and low feed intake, were selected for both models. The variable coughing/respiratory problems was selected for the classifier only, while hard faecal pellets and conjunctivitis were only selected for the diagnostic rule. We would like to mention that, conjunctivitis would have been selected as the next clinical sign to be included in the selective classifier, based upon its mutual information with the class variable; note that it was not included on account of the mutual information being slightly lower than the threshold value of 6.64. The most striking difference between the two models is the absence, from the diagnostic rule, of the highly discriminative sign of coughing/respiratory problems.

<table>
<thead>
<tr>
<th>Optimally efficient diagnostic rule</th>
<th>Selective naive Bayesian classifier</th>
<th>(I(X,Y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ataxia</td>
<td>Ataxia</td>
<td>21.6</td>
</tr>
<tr>
<td>Low feed intake</td>
<td>Low feed intake</td>
<td>18.9</td>
</tr>
<tr>
<td>Not responding to antibiotics</td>
<td>Not responding to antibiotics</td>
<td>9.0</td>
</tr>
<tr>
<td>-</td>
<td>Coughing/Respiratory problems</td>
<td>16.7</td>
</tr>
<tr>
<td>Conjunctivitis</td>
<td>-</td>
<td>6.5</td>
</tr>
<tr>
<td>Hard faecal pellets</td>
<td>-</td>
<td>4.3</td>
</tr>
</tbody>
</table>

DISCUSSION

In this study, we illustrated the potential of naive Bayesian classifiers for the domain of veterinary medicine by means of a full and a selective classifier for discriminating between CSF-infected and non-infected herds. These classifiers were constructed from the sensitivities and specificities of individual clinical signs and the numbers of positively and negatively diagnosed herds reported by Elbers et al. (2002). While the classifiers could be built from information provided in the literature, to establish their sensitivities, specificities and accuracies, a dataset was needed. Note that for such evaluation purposes, information on the absence or presence of clinical signs in both infected and non-infected herds is needed. For the evaluation of our classifiers, we had available the original data that were analysed before by Elbers et al. These data indeed included observations from both infected and non-infected herds. With the available data, the constructed classifiers exhibited good performance. In fact, they had a comparable (for the selective classifier) or even better (for the full classifier) accuracy than the diagnostic rule that had been established from the same data by Elbers et al. (2002).

Upon singling out the most discriminative clinical signs for our selective classifiers, by means of the concept of mutual information, a large overlap was found with the clinical signs selected by Elbers et al. (2002) for their diagnostic rule. One of the highly discriminative
features, that is the feature of coughing/respiratory problems, that was included in the selective naive Bayesian classifier, however, was not present in the rule. The data reveals that, in the 1997/1998 epidemic in The Netherlands, this sign was encountered more often in herds that were non-infected than in infected herds. If coughing/respiratory problems were present in a herd, therefore, these problems were an indication against CSF rather than for the disease. In Bayesian classification in general, clinical signs that are indicative of the absence of a disease play an equally important role in discriminating between the various outcomes as signs that point to the presence of the disease. In the diagnostic rule, in contrast, the presence of a symptom can only be used as an indicator for the disease, because of its disjunctive character. Since the occurrence of coughing/respiratory problems may vary strongly between CSF epidemics and in literature respiratory problems are said to be a possible indicator for CSF infection, the practical implications of selecting this clinical sign against CSF for early detection of CSF may be limited.

The data that were used in our study, suffered from variance in the way clinical signs were observed and interpreted (Elbers et al., 2002). This variance was caused by the acuteness of the epidemic. In addition, there is some uncertainty as to whether clinical signs that were not recorded, were really absent or were simply not noticed by the veterinarians. In a previous study, we investigated the effect of variation in the absence or presence of non-recorded clinical signs, on the selection of discriminative feature variables using the filter approach taken in this paper (Geenen et al., 2004). It was shown that both the selection of feature variables and the accuracy of the resulting classifier were quite robust against a 10% variation.

Upon using our naive Bayesian classifiers in practice, the observed clinical signs are entered into the computer programme that implements the classifier. This programme subsequently calculates the posterior probability of CSF in the herd. If the calculated probability exceeds a given threshold, a veterinary practitioner should have the possibility to exclude CSF as possible cause for the disease problems by sending samples to the National Reference Laboratory. If there is a high probability for CSF given the observed clinical signs, a serious suspicion of CSF should be notified to the veterinary authorities and an expert team from the veterinary authorities should be warned to visit the farm and possibly take additional measures. Now, the current classifiers are based on data that were collected during an epidemic. For the early detection of CSF, however, an inventory of the clinical signs seen in a situation without outbreaks is needed. In addition, the current classifiers are constructed to classify herds with maximised accuracy, resulting in maximised sensitivity and specificity. For early CSF detection however, one may wish to have a classifier with a somewhat higher sensitivity (and hence, a lower specificity). We are currently developing new methods for the selection of discriminative features that allow for weighing the sensitivity and specificity of the resulting classifiers. For the purpose of our study it was assumed that herds had equal probabilities of being infected and non-infected to allow for comparing the classification performance of the classifiers with that of the diagnostic rule. In practice, however, the prevalence of CSF will be much lower than 0.5. Moreover, the prevalence changes drastically as of the onset of an epidemic. This variation over time has not been included within our classifiers as yet, and will pose an interesting subject for future research.

Judging from the accuracy we attained from these relatively simple models, we feel that naive Bayesian classifiers are promising tools for solving diagnostic problems in the veterinary field. Building upon our experience with the classifiers discussed in the present paper, we are currently developing a more complex model for the early detection of CSF with the help of domain experts, which will hopefully further improve the accuracy of distinguishing between infected and non-infected herds based upon clinical signs only.
REFERENCES


