Building Naive Bayesian Classifiers from Literature: a Case Study in Classical Swine Fever

P.L. Geenen[1] L.C. van der Gaag[1] W.L.A. Loeffen[2]
A.R.W. Elbers[2]

[1]Institute of Information and Computing Sciences, Utrecht University,
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
{petrag,linda}@cs.uu.nl
[2]Central Institute For Animal Disease Control, Wageningen UR,
P.O. Box 2004, 8203 AA Lelystad, The Netherlands
{armin.elbers,willie.loeffen}@wur.nl

Abstract

Naive Bayesian classifiers are typically learned from data, yet in this paper we address the construction of full and selective classifiers from information provided in the literature. We use a paper that reports on an epidemiological study in Classical Swine Fever, to build classifiers for discriminating between pig herds with and without the disease. We show that even without the original dataset, full as well as selective classifiers can be constructed. As in our case study we had the original dataset at our disposal, we were able to compare the accuracies of the constructed classifiers to those of the diagnostic rules reported in the paper under study; we found that the accuracies of our classifiers compared favourably to the accuracies of these rules.

1 Introduction

Naive Bayesian classifiers have proven to be powerful tools for solving classification problems in a variety of domains. 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, modelling the possible outcomes or classes for the problem under study, and a set of feature variables, modelling the features that provide for distinguishing between the various classes; the feature variables are assumed to be mutually independent given the class variable [1]. Instances of the classification problem under study are presented to the classifier as a combination of values for the feature variables; the classifier then returns a posterior probability distribution over the class variable, that is, it yields a probabilistic summary for each class. Naive Bayesian classifiers have been successfully applied in for example the medical domain where they are used for solving diagnostic problems.

Naive Bayesian classifiers are typically learned from data. Learning such a classifier amounts to establishing the prior probabilities of the different classes and estimating the conditional probabilities of the various features given each of the classes. Not for every classification problem, however, will a dataset be available that is rich enough to allow for reasonable probability estimates. We argue that in the absence of data, a classifier may be constructed from information provided in the literature. To support our observations, we construct various Naive Bayesian classifiers for the domain of Classical Swine Fever based upon a paper that reports on an epidemiological study of the clinical symptoms observed in pig herds during the 1997/1998 epidemic of the disease in the Netherlands [2]. In the paper, the sensitivity and specificity characteristics of 32 clinical symptoms are provided. We show how to construct, from this information, a full Naive Bayesian classifier containing 32 feature variables modelling the various symptoms.

For a classification problem under study, the more discriminative features are generally distinguished from the less informative ones. The Naive Bayesian classifier then is built over just the selected subset of features; by constructing the classifier over a subset of the features, a less complex model is yielded that tends to have a better generalisation performance [3]. These restricted classifiers are called selective Naive Bayesian classifiers. A well-known approach to singling out the features to be included in a selective classifier, is the filter approach. With this approach, an information-theoretic function capturing the separability between the different classes is used as a criterion to decide upon inclusion of a feature variable. We show that the sensitivity and specificity characteristics reported for the various clinical symptoms of Classical Swine Fever allow feature selection with the filter approach and, hence, provide for the principled construction of a selective Naive Bayesian classifier.

In our case study in the domain of Classical Swine Fever, we had at our disposal the original dataset that was used in the reported epidemiological study. The reported study was aimed at the selection of readily observed symptoms for a classification rule that could be used as a diagnostic test for the disease [2]. A number of deterministic disjunctive rules resulted from the study. For each of the selected subsets of features, we constructed selective classifiers, once again building upon the reported sensitivity and specificity characteristics. Using the original dataset, we could now compare the accuracies of the full and selective Naive Bayesian classifiers constructed from the reported information to those of the diagnostic rules. We found that the accuracies of the various classifiers compared favourably to the accuracies of these rules. We found moreover, that the selective classifier constructed with the filter approach included a feature that proved highly discriminative for Classical Swine Fever, yet was missing from the rules.

The present paper is organised as follows. In Section 2, we briefly introduce the domain of Classical Swine Fever and review the information available from an epidemiological study of the clinical symptoms observed during an outbreak of the disease. In Section 3, we show how full and selective Naive Bayesian classifiers can be constructed from the available information. In Section 4, we compare the accuracies of the constructed classifiers to those of the reported diagnostic rules. The paper ends with our concluding observations in Section 5.

2 A Reported Study in Classical Swine Fever

We review the information available from an epidemiological study in Classical Swine Fever. Before doing so, we give some background knowledge of the disease.

Classical Swine Fever is a highly infectious viral disease of pigs that has a potential for rapid spread. The virus causing the disease is transmitted mainly by direct contact between infected and non-infected susceptible pigs, yet transmission by vehicles, farmers, veterinarians, and artificial insemination may also occur. When a pig is infected, the virus first invades the lymphatic system and subsequently affects the blood vessels which may give rise to various bleedings. The virus will ultimately affect several organs and the pig will die. As a consequence of the infection, a pig will show different disease symptoms, among which are fever, reduced feed intake, inflammation of the eyes, walking disorders, and haemorrhages of the skin. Classical Swine Fever is quite common in parts of Europe and Africa, and in many countries of Asia, Central and South America [4].

Since the occurrence of Classical Swine Fever has a major impact on international trade of animals and animal products, extensive measures have been taken within the European pig husbandry to prevent the introduction and spread of the virus. Unfortunately, however, each year several outbreaks of the disease occur. Such an outbreak has serious socio-economical consequences. In the 1997/1998 epidemic in the Netherlands, for example, 429 herds were infected and 12 million pigs had to be killed. The total costs were estimated at 2.3 billion US dollars.

Clinical symptoms seen by the farmer or by a veterinarian are usually the first indications of the presence of Classical Swine Fever in a herd. When an early suspicion of the disease is reported to the ministry, a veterinary expert team will visit the farm and inspect the pig herd. During the 1997/1998 epidemic in the Netherlands, these teams regularly encountered herds in which the disease was indicated. The body temperature of the diseased pigs was measured and an anamnesis, or disease history, was recorded on an investigation form. In the anamnesis, the presence of disease symptoms within the herd were recorded; if a single pig within the herd was observed to suffer from inflammation of the eyes, for example, then this feature was marked as being present. Pigs with apparent disease symptoms and/or fever were killed and submitted to the Animal Health Service for a post-mortem examination. If one or more pigs from such a submission proved to be infected with the virus, then the herd was diagnosed as positive for Classical Swine Fever. If all pigs from 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.

From the investigation forms that were available from 245 herds that were diagnosed as positive for Classical Swine Fever and 245 herds that were diagnosed as negative for the disease, a dataset was constructed; the 490 forms had been filled in by 185 different veterinary inspectors. On the forms, a total of 32 distinct clinical symptoms were recorded. Upon constructing the dataset, the recorded symptoms were encoded as '1's for the appropriate variables; symptoms that were not recorded explicitly were assumed to be absent and were encoded as '0's. In the positively classified herds the mean number of recorded symptoms was 3.5 which

Table 1: The combinations of clinical signs and the associated sensitivity, specificity and accuracy of the three diagnostic rules.

Rule	*Clinical signs*	*Sens.*	*Spec.*	*Acc.*
Optimally sensitive	Walking disorder, reduced feed intake, not responding to antibiotics, inflammation of the eye, apathy, haemorrhages, abortion, cyanosis of the ears, not drinking water, dead pigs	0.890	0.212	0.55
Optimally specific	Reduced water intake, low milk production, raised hairs, pigs fatigued after small exertion, blue colour of the ears, hard faecal pellets	0.086	0.980	0.53
Optimally efficient	Walking disorder, reduced feed intake, not responding to antibiotics, inflammation of the eye, hard faecal pellets	0.727	0.527	0.63

was significantly higher than the mean number of 3.0 recorded symptoms in the negative herds (Mann Whitney U-test, $p < 0.01$).

The collected data were analysed in an epidemiological study of the predictive values of the various clinical symptoms of Classical Swine Fever [2]. The aim of the analysis was to arrive at classification rules composed of readily observed symptoms, that could be used as diagnostic tests for establishing the presence of the disease in a herd. As the recorded symptoms are relatively sparse, disjunctive rules were constructed. If at least one symptom mentioned in such a rule is present in a herd, then the herd is diagnosed as positive for the disease; if all symptoms from the rule are absent, then the herd is classified as negative. Logistic regression with backward selection was applied to the data, with the classification of the herd as the response variable and the clinical symptoms as explanatory variables; the clinical symptoms that served to explain the variation in the classification the most were thus selected. Subsequently, all possible rules that could be constructed from the selected symptoms were evaluated using a receiver operating characteristic (ROC) analysis; in the evaluation, the sensitivity and specificity characteristics of the various rules were studied. The analysis resulted in three different diagnostic rules: a rule with maximised sensitivity and specificity ('optimally efficient'), a rule combining maximum sensitivity with the highest possible specificity ('optimally sensitive'), and a rule with maximum specificity and the highest possible sensitivity ('optimally specific'). The combinations of clinical symptoms mentioned in the three rules, and their associated sensitivity, specificity and accuracy, are shown in Table 1. Note that, as the numbers of positively and negatively diagnosed herds are the same, the accuracy of a rule, that is, its proportion of correctly diagnosed herds, simply equals the mean of the rule's sensitivity and specificity.

3 Constructing Classifiers from the Literature

Naive Bayesian classifiers generally are learned from data. If for a classification problem of interest appropriate data are not available, the literature in the domain may provide sufficient information to construct a reliable classifier. We show how both full and selective classifiers can be built from information provided in the literature. We illustrate our observations by constructing various Naive Bayesian classifiers for the domain of Classical Swine Fever based upon the reported results of the epidemiological study reviewed in the previous section.

Constructing a full Naive Bayesian classifier starts by defining the class variable with its possible values and the feature variables with their values. For the class variable, prior probabilities for the various classes discerned have to be established; for each feature variable, moreover, conditional probability distributions over its values given the different classes have to be defined. To build a Naive Bayesian classifier for diagnosing Classical Swine Fever, we created a binary class variable modelling whether or not a herd is infected with the disease, and 32 feature variables; each feature variable served to model the presence or absence of a specific clinical symptom. The prior probabilities for the class variable were computed from the numbers of positively diagnosed and negatively diagnosed herds, that is, the probabilities of the two classes were established to be $p(\text{CSF} = \text{yes}) = p(\text{CSF} = \text{no}) = 0.5$. For each feature variable, the required conditional probabilities were established from the sensitivity and specificity characteristics reported for the corresponding clinical symptom. We recall to this end that the sensitivity of a symptom equals $p(\text{symptom} = \text{yes} \mid \text{CSF} = \text{yes})$; its specificity equals $p(\text{symptom} = \text{no} \mid \text{CSF} = \text{no})$. 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 symptom occurring. To prevent inconsistencies when entering the data, therefore, we replaced these probabilities in the specifications of our classifiers by 0.0001. By including all 32 feature variables, we obtained a full Naive Bayesian classifier for diagnosing Classical Swine Fever. By including just the appropriate subsets of feature variables moreover, we obtained selective classifiers for the three diagnostic rules that resulted from the original epidemiological study.

Building a selective Naive Bayesian classifier involves singling out the feature variables that best serve to separate the different classes under study. The selection of appropriate feature variables generally is based on data. With the well-known filter approach to feature selection, an information-theoretic criterion is used to decide upon inclusion of the various feature variables. The mutual information $I(X, Y)$ of two variables X and Y is often used for this purpose, where

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

The feature variables then are taken for the variable X and the class variable is taken for the variable Y. To decide upon inclusion, the property that $2 \cdot N \cdot I(X, Y)$ asymptotically follows a $\chi^2_{(r-1)(r_0-1)}$ distribution is exploited, where r is the number of possible values of X, r_0 is the number of values of Y, and N is the

size of the dataset used. With a level of significance of $\alpha = 0.01$, for example, only feature variables for which $2 \cdot N \cdot I(X, Y) > 6.64$ are included in the classifier.

Also in the absence of data can a principled selection of the feature variables to be included in a selective classifier be made. To build a filter-based selective Naive Bayesian classifier for Classical Swine Fever, we applied the mutual-information criterion with the variable modelling the presence or absence of the disease for the variable Y and each of the feature variables modelling a clinical symptom for the variable X. We now observe that since $p(x, y) = p(x \mid y) \cdot p(y)$ and $p(x) = \sum_y p(x, y)$, all probabilities mentioned in the formula for $I(X, Y)$ can be rewritten in terms of the numbers of positively and negatively diagnosed herds and the sensitivity and specificity characteristics of the various symptoms. For each feature variable, therefore, its mutual information with the class variable could be readily established from the information that we had available from the epidemiological study. To decide upon whether or not to include a feature variable in the classifier, we used a significance level of $\alpha = 0.01$ as indicated above. Four features variables, or clinical symptoms, were thus selected for our classifier: 'walking disorder', 'low feed intake', 'respiratory problems', and 'not responding to antibiotics'.

4 An Experimental Comparison

For our case study, we had at our disposal not only the reported results from the epidemiological study in Classical Swine Fever, but also the original dataset from which the various diagnostic rules were constructed. We were able, therefore, to compare the accuracies of the different Naive Bayesian classifiers that we built using the reported information, to the accuracies of the rules.

Using the available data, we calculated the sensitivity, specificity and accuracy of the full Naive Bayesian classifier, of the selective classifiers constructed from the three diagnostic rules, and of the selective Naive Bayesian classifier constructed with the filter approach; we also calculated the sensitivity, specificity and accuracy of the three rules. The results are summarised in Table 2. The accuracies of the 'optimally efficient', 'optimally sensitive' and 'optimally specific' classifiers were equally good, or even better, than the accuracies of the corresponding diagnostic rules. We found a large difference in characteristics between the optimally sensitive rule and its corresponding classifier. This difference originates from the way in which the two models interpret the available data. If a herd shows just a single symptom from among the mentioned symptoms, the rule serves to classify the herd as positive, thereby accounting for its high sensitivity and low specificity. The classifier uses information of both the presence and the absence of the mentioned symptoms and tends to classify herds with just a single symptom as negative, which results in a much lower sensitivity and a higher specificity. No such differences were found for the optimally specific rule and its corresponding classifier; as both models include symptoms that are hardly ever observed, they both tend to classify most herds as negative for the disease. We would like to mention that, while the three rules were designed to be optimal in a particular sense, the corresponding classifiers were not constructed with such an aim in mind, which renders a detailed

Table 2: The sensitivity, specificity and accuracy, with their 95%-confidence intervals, for the various classifiers and diagnostic rules, calculated from the original dataset ($n = 490$).

Classifier	*Sensitivity*	*Specificity*	*Accuracy*
'Optimally efficient'	$0.66(0.61 - 0.72)$	$0.59(0.53 - 0.65)$	$0.63(0.59 - 0.67)$
'Optimally sensitive'	$0.65(0.59 - 0.71)$	$0.65(0.59 - 0.71)$	$0.65(0.61 - 0.69)$
'Optimally specific'	$0.09(0.05 - 0.13)$	$0.98(0.96 - 1.00)$	$0.53(0.49 - 0.58)$
Full Naive Bayes	$0.65(0.59 - 0.71)$	$0.73(0.67 - 0.78)$	$0.69(0.65 - 0.73)$
Selective Naive Bayes	$0.63(0.57 - 0.69)$	$0.67(0.61 - 0.72)$	$0.65(0.61 - 0.69)$
Rule	*Sensitivity*	*Specificity*	*Accuracy*
Optimally efficient	$0.73(0.67 - 0.78)$	$0.53(0.46 - 0.59)$	$0.63(0.59 - 0.67)$
Optimally sensitive	$0.89(0.85 - 0.93)$	$0.21(0.16 - 0.26)$	$0.55(0.51 - 0.59)$
Optimally specific	$0.09(0.05 - 0.12)$	$0.98(0.96 - 1.00)$	$0.53(0.49 - 0.57)$

comparison of especially the optimally sensitive and optimally specific rules against their corresponding classifiers less meaningful. For the optimally efficient rule, we note that the full Naive Bayesian classifier has a significantly higher accuracy ($p = 0.03$, proportions test) and that the filter-based selective classifier does not show a significantly higher proportion of correctly diagnosed herds.

The three diagnostic rules that resulted from the epidemiological study, represent selections of clinical symptoms that are considered optimal in a particular sense for distinguishing between infected and non-infected herds. With the filter approach, also a subset of the symptoms was selected. The symptoms that were selected for the optimally efficient rule and for the filter-based selective classifier are shown in Table 3. We observe that the symptoms 'walking disorder', 'not responding to antibiotics' and 'low feed intake' were selected for both models. The symptom 'respiratory problems' was selected for the classifier only, while 'hard faecal pellets' and 'inflammation of the eyes' were only selected for the rule. We would like to note that, with the filter approach, 'inflammation of the eyes' would have been selected as the next symptom to be included in the selective classifier, based upon its mutual information with the class variable. It is the absence, from the diagnostic rule, of the highly discriminative symptom 'respiratory problems' that constitutes the more striking difference between the two models. The data

Table 3: The selected clinical symptoms.

Optimally efficient rule	*Filter-based selective classifier*
Walking disorder	Walking disorder
Not responding to antibiotics	Not responding to antibiotics
Low feed intake	Low feed intake
Hard faecal pellets	Respiratory problems
Inflammation of the eyes	

reveals that respiratory problems were encountered more often in herds that were negative for Classical Swine Fever than in positive herds. The symptom therefore is an indication against rather than for the disease. Because of the construction of the optimally efficient rule, however, the presence of a symptom can only be used as an indicator for the disease. The 'respiratory problems' symptom could therefore not show up in the rule. Note that taking the absence of respiratory problems as an indicator for the disease would serve to classify the majority of the herds as positive, which would significantly decrease the rule's specificity.

5 Conclusions

We illustrated, by means of a case study in Classical Swine Fever, that information from the domain literature can provide for the principled construction of full and selective Naive Bayesian classifiers. The classifiers that we constructed for our domain of application from published information, were shown to exhibit good performance on the data that we had available. We note that feature selection based upon published information is necessarily restricted to the filter approach. An alternative approach to feature selection is the wrapper approach in which the accuracy of the resulting classifier is used to guide the search for a suitable subset of features. As for this purpose the availability of a dataset is imperative, wrapper-based feature selection cannot be performed based upon information provided in the literature only. We conclude by observing that there are many classification problems in domains where datasets are not readily available. We feel that for these problems it might be profitable to screen the literature for information that provides for the construction of full and selective classifiers.

Acknowledgements

This research has been partly supported by the Netherlands Organisation for Scientific Research (NWO).

References

[1] N. Friedman, D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning*, vol. 29, pp. 131 – 163.

[2] A.R.W. Elbers, A. Bouma, and J.A. Stegeman (2002). Quantitative assessment of clinical signs for the detection of classical swine fever outbreaks during an epidemic. *Veterinary Microbiology*, vol. 85, pp. 323 – 332.

[3] P. Langley and S. Sage (1994). Induction of selective Bayesian classifiers. *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*. pp. 399 – 406.

[4] S. Edwards, A. Fukusho, P.C. Lefevre *et al.* (2000). Classical Swine Fever: the global situation. *Veterinary Microbiology*, vol. 73, pp. 103 – 119.