DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 18

11-1-2005

Selection Of Independent Binary Features Using Probabilities: An Example From Veterinary Medicine

Ludmila I. Kuncheva
University of Wales, Bangor, UK, l.i.kuncheva@bangor.ac.uk

Zoë S.J. Hoare
University of Wales, Bangor, UK, z.s.hoare@bangor.ac.uk

Peter D. Cockcroft
University of Cambridge, UK, pdc24@hermes.cam.ac.uk

Recommended Citation

Kuncheva, Ludmila I.; Hoare, Zoë S.J.; and Cockcroft, Peter D. (2005) "Selection Of Independent Binary Features Using Probabilities: An Example From Veterinary Medicine," *Journal of Modern Applied Statistical Methods*: Vol. 4: Iss. 2, Article 18.
Available at: http://digitalcommons.wayne.edu/jmasm/vol4/iss2/18

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 528-537

Selection Of Independent Binary Features Using Probabilities: An Example From Veterinary Medicine

Ludmila I. Kuncheva Zoë S.J. Hoare
School of Informatics
University of Wales, Bangor, UK

Peter D. Cockcroft
Department of Clinical Veterinary Medicine
University of Cambridge, UK

Supervised classification into *c* mutually exclusive classes based on *n* binary features is considered. The only information available is an *n*×*c* table with probabilities. Knowing that the best *d* features are not the *d* best, simulations were run for 4 feature selection methods and an application to diagnosing BSE in cattle and Scrapie in sheep is presented.

Key words: Feature selection, classification, independent features, binary features, veterinary medicine.

Introduction

Consider the differential diagnosis of BSE in cattle based on the probabilistic description of BSE and 56 alternative diseases with similar symptoms. There are many possible disease-related signs that may be observed as present/absent on an animal. For example, over 240 signs related to BSE and the 56 other diagnoses can be listed (Brightling et al., 1996; White, 1984). To build a diagnostic system, a data set is needed with observations for a number of cattle with their verified diagnoses. In the lack of such a data set, one must rely on estimates of the individual class-conditional probabilities that sign x_i is present, given disease ω_j, where $i \in \{1,2,...,n\}$ and $j \in \{1,2,...,c\}$.

The information available in this problem is organized as shown in Table 1.

Table 1. Class-conditional probabilities for the individual features (the only information available)

	ω_1	$\dots\ \omega_i\ \dots$	ω_c
x_1		$\dots$	
...			
x_k	$\dots$	$P(x_k = 1 \mid \omega_i)$	$\dots$
$\dots$			
x_n		$\dots$	

It is unrealistic to expect that a system based on these probabilities will fare well in practice because no relationship between the diagnostic signs (features) has been taken into account. In an ideal scenario, a data set will be collected using all features and the relationships between the features will be estimated from it. In reality, measuring only a small number of relevant features may be feasible.

The goal is to select *d* features (*d* < *n*), which form a subset with the smallest classification error. Denote by **x** the binary vector with the *n* features. The features are assumed to be conditionally independent, that is,

$$P(\mathbf{x} \mid \omega_j) = \prod_{i=1}^{n} P(x_i \mid \omega_j) \qquad (1)$$

Ludmila Kuncheva is a Senior Lecturer. Her interests include pattern recognition and classifier ensembles. Email her at: l.i.kuncheva@bangor.ac.uk. Zoe Hoare studied mathematics at the University of Wales, Bangor, UK. She is a doctoral student in the area of pattern recognition and its application to veterinary medicine. Email her at z.s.hoare@bangor.ac.uk. Peter Cockroft is a Senior Lecturer, member of the Royal College of Veterinary Surgeons (RCVS), and a holder of the RCVS's Diploma in Cattle Health and Production. Email him at pdc24@hermes.cam.ac.uk.

The assumption of independence is enforced upon this study because only (some estimates of) the individual class-conditional probabilities are available. Pattern recognition literature in the 1970s abounds with analyses of the case of independent binary features. Perhaps the most curious result is due to Toussaint (1971). If there are three independent binary features, the best combination of two features may not include the single best feature. Thus, the most desirable selection criterion – the probability of error – will not guarantee the optimal solution if applied in a stepwise manner as in stepwise linear regression.

In this article, four procedures for selecting a subset of features are examined and the results are compared with those obtained with the whole feature set. The feature selection methods are illustrated on two problems taken from veterinary medicine: differential diagnosis of BSE in cattle and Scrapie in sheep.

Methodology

Feature selection is one of the oldest topics in pattern recognition and machine learning (Stearns, 1976; Van Campenhout, 1982; Jain and Chandrasekaran, 1982; Patrick, 1972). Surveys on more recent state-of-the-art and comparisons between feature selection procedures can be found in (Dash & Liu, 1997; Blum & Langley, 1997; Jain & Zongker, 1997; Aha & Bankert, 1995).

Evaluation of the Feature Subsets

The most intuitive measure of quality of a feature subset is the error of a classifier built on these features. In theory, one can calculate the error under the assumption that the probabilities are equal to their expert estimates. The optimal classifier for independent features is the Naïve Bayes classifier. Denote by P_j the prior probability for class ω_j. Let $\mathbf{x} = [x_1,\ldots,x_n]^T$ be a binary vector to be labeled into one of the c mutually exclusive classes. A discriminant function is calculated for each class,

$$\begin{aligned}\mu_j(\mathbf{x}) &= P_j P(\mathbf{x} \mid \omega_j) \\ &= P_j \prod_{i=1}^{n} P(x_i \mid \omega_j), \quad j = 1,\ldots,c\end{aligned} \tag{2}$$

$\mathbf{x}$ is labeled in the class with the largest discriminant value. There are 2^d possible binary vectors $\mathbf{x}$ for a candidate subset S with d features. The (probability for the) minimum classification error for the subset can be calculated as

$$\begin{aligned}P_e &= \sum_{\mathbf{x}} P(\mathbf{x}, \text{error}) \\ &= 1 - \sum_{\mathbf{x}} \max_j \left[P_j \prod_{i \in S} P(x_i \mid \omega_j) \right]\end{aligned} \tag{3}$$

Equation (3) shows the difficulty in calculating the error for large d. Every $\mathbf{x}$ must be visited to decide which class label to assign to it. There are indirect criteria related to the error which may be faster to calculate, but direct calculation of the error in some form is preferable (Dash & Liu, 1997). Monte Carlo simulations were chosen for estimating the error of the selected feature subset. The probabilities for each class were available and it was therefore possible to generate randomly a sample from each class with n independent features. Using the selected feature subset, the Naïve Bayes classifier was applied for the objects in this sample.

The Single-Best Method (SB)

It is known that the individually best d features do not necessarily form the best subset of d features (Toussaint, 1971). Nonetheless, the method is quick and sometimes surprisingly efficient. The error for each feature is calculated separately using (3) (note that there are only two possible $\mathbf{x}$'s for each feature: present or absent), the errors are sorted in ascending order and the top d features are retained. In this method, one can pick a desired value for d.

The complexity of a feature selection algorithm is typically measured by the number of calculations of the classification error needed to select d out of n features. Thus the single-best method needs just n evaluations regardless of the number d.

Sequential Forward Selection (SFS)

This is the method traditionally used in stepwise regression. To start, there is an empty set, *S*, of chosen features. Each feature must be evaluated separately as in the single-best method and the best individual feature is placed in *S*. At the next step, all pairs of features which contain the feature selected already and one other feature are evaluated. The pair with the smallest error is retained as *S*. Then, one must check all triples of features, and so on, until the desired cardinality *d* of *S* is reached. This procedure does not guarantee finding the optimal set of *d* features even in this simple case of independent binary features. The reason for this can be explained again with the Toussaint's counter example: the best set of two does not necessarily contain the single best feature.

Below, an example illustrating both the non-optimality of the sequential feature selection (SFS) and the calculation of the error though equation (3) is shown.

Consider three features, x_1, x_2, and x_3, and two classes, $\Omega = \{\omega_1, \omega_2\}$. The non-traditional data considered in this study is given in the form of probability estimates $P(x_i = 1 \mid \omega_j)$, as shown in Table 2.

Table 2. An example of a set of probabilities for 3 features and 2 classes

	ω_1	ω_2
x_1	0.1	0.5
x_2	0.6	0.1
x_3	0.8	0.4

Denote $a = P(x_k = 1 \mid \omega_1)$ and $b = P(x_k = 1 \mid \omega_2)$ for some x_k. Assuming equal prior probabilities for the two classes, the probability of correct classification for feature x_k is

$$P(k) = 1/2\{\max(a,b) + \max(1-a, 1-b)\} \quad (4)$$

Using (4), the individual errors for the features are $\varepsilon_1 = 1 - ½\,[\max(.1,.5) + \max(.9,.5)] = 0.30$, $\varepsilon_2 = 0.25$, and $\varepsilon_2 = 0.30$. Consider a pair of features, (x_k, x_j), and denote the probabilities for x_j as $p = P(x_j = 1 \mid \omega_1)$ and $q = P(x_j = 1 \mid \omega_2)$. Substituting again in equation (3), the probability of correct classification for the pair of features is

$$\begin{aligned} P(k,j) = 1/2\{&\max(a\,p, bq) \\ &+ \max[(1-a)p, (1-b)q] \\ &+ \max[a(1-p), b(1-q)] \\ &+ \max[(1-a)(1-p), (1-b)(1-q)]\} \end{aligned} \quad (5)$$

The errors for the three pairs of features for the example in Table 2 are

$$\begin{aligned} \varepsilon_{12} = 1 - ½(&\max(.1\times.6, .5\times.1) \\ &+ \max(.9\times.6, .5\times.1) \\ &+ \max(.1\times.4, .5\times.9) \\ &+ \max(.9\times.4, .5\times.9)) \\ = 0.25&, \end{aligned}$$

$\varepsilon_{13} = 0.24$, and $\varepsilon_{23} = 0.25$.

As ε_{13} is the smallest pair-wise error, and ε_2 is the smallest individual error, the best pair of independent features, (x_1, x_3), does not include the single best feature x_2.

SFS is probably the most widely used procedure because it has both reasonable error and reasonable complexity for "traditional" data sets (Aha & Bankert, 1995; Jain & Zongker, 1997).

At the first step, SFS evaluates all *n* features, at the second step, *n*-1 evaluations are needed as there are *n*-1 possible pairs. For selecting *d* features, SFS needs the following number of evaluations of the error

$$\sum_{i=0}^{d-1} (n-i) \quad (6)$$

However, the complexity calculation is not that simple when the features from probabilistic data as shown in Table 1 are selected. For the calculation of the theoretical error, the algorithm has to visit every **x** in the possible feature space, find out which is the maximum discriminant function, and add the contribution of the error for **x** based on the class label decision. The fact that the features are treated as independent does not make the task any easier. The complexity of SFS will depend heavily on the number of features in the evaluated subsets.

Complexity of feature selection algorithms for probabilistic data can be evaluated by the total number of **x**'s visited in the process of selecting *d* out of the *n* features. The complexity for the single-best method is $C_{SB} = 2n$, and for the SFS, $C_{SFS} = \sum_{i=0}^{d-1}(n-i)2^{i+1}$.

Class-Pairs Feature Selection (CP)

Ji and Bang (2000) proposed the following feature selection method. A single feature is selected for each pair of classes.

Table 3 shows the data structure used by the algorithm, where C_{ij} = class pairs, ($i \neq j$), x_k = *k*-th feature, (k = 1,..,n), $P_{ij}(k)$ = discriminatory power of feature *k* for $C_{ij.}$ Using (4), the values of $P_{ij}(k)$ are calculated as the probability of correct classification between classes ω_i and ω_j for feature x_k.

Table 3. The table for the class-pairs method for feature selection (Ji and Bang, 2000).

		C_{ij}		
		…		
x_k	…	$P_{ij}(k)$	…	T_k
		…		
		E_{ij}		

The following values are then calculated

- $E_{ij} = \sum_k P_{ij}(k)$, the relative ease of classifying the pair C_{ij} , and
- $T_k = \sum_{ij} P_{ij}(k)$, the relative discriminatory power of feature x_k.

The algorithm begins with an empty set of features. The class pair that is the hardest to discriminate (has the smallest E_{ij}) is identified from the table. The feature with the highest discriminatory power for this pair is added to the subset, if not already selected. If more than one feature has the highest $P_{ij}(k)$ in the chosen column, then the feature with the highest value of T_k is selected.. The hardest pair is removed from the table and the process continues with the next hardest pair of classes (Note that the classes are not removed altogether, only the column of the table is removed.). The process stops once all class pairs have been covered.

The maximum number of features this method will select is max{(c(c-1)/2, n}. However, Ji and Bang (2000) claim that the number selected will be much less than either of these. This method may also be restricted at any point to pick only *d* features. The complexity of the class-pair method (measured again by the total number of **x**'s visited) is $C_{CP} = c\ (c\text{-}1)\ n$. This calculation reflects only the preparation phase (setting up Table 3), and does not take into account the actual procedure which constructs the feature subset.

Feature-Pairs Feature Selection (FP)

The selection methods considered above are either overly simplistic but scale well with *n*, *c*, and *d* (single-best) or they are computationally demanding but more accurate (SFS). Optimality of the selected feature subset is not guaranteed in any case. The class-pairs method is one possible method that scales well and may be accurate. Here, another method is proposed for feature selection from probabilities, called feature-pairs method.

The process is started with an empty set of features. All pairs of features are evaluated and the best pair is added to the set. While the desired number of features is not reached, add the features from the next best pair which are not already among the selected features. Suppose that *d*-1 features are already in the set, and there is a pair of features such that neither of the two members of the pair is in the set. One may either take both features and exit with d+1 features or randomly select one member of the pair to make up the total of d features in the set. The complexity of the feature-pairs method (using the number of visited **x**'s) is $C_{FP} = n\,(n\text{-}1)$.

All four methods are based on a true calculation of the classification error plus some heuristic about how one forms the feature subset. The experimental results in the next section help to evaluate the assets and drawbacks of the four methods.

Results

A Small-Scale Simulation Study

To include SFS in the comparisons, a relatively small example with n = 20 features was chosen and the number classes, c, was varied from 3 to 10. The number of selected features, d, was varied from 2 to 10.

For each c, 50 random matrices of size 20×c were generated from uniform random distribution. Each matrix represented the probabilities for the features and classes as shown in Table 1. For each such matrix and each d, the four feature selection algorithms were applied and the best subset of size d was found.

To evaluate the selected subsets, a traditional data set was generated randomly for every pair (c,d). One hundred data points were generated from the distribution of each class and the Naïve Bayes classifier was used to label these points. The error was estimated as the percent mismatch with the true class label.

An example of the simulation algorithm is given below. Consider the problem presented in Table 2. Suppose that Method X picked features (x_1, x_3). Set a misclassification counter to 0. The steps below are repeated 100 times for each class.

(Step 1) Generate a data point from class ω_1. To do this, pick a vector of 3 random numbers, one for each feature, e.g. $[0.2736, 0.9241, 0.7102]^T$. Compare this vector with the first column of Table 2 (corresponding to ω_1). If the generated number for x_i is smaller than the corresponding probability in the table, set x_i to 1; else set x_i to 0. For this example, the generated data point is **x** = [0, 0, 1].

(Step 2) Classify the data point using Naïve Bayes and only the chosen features. For this example (x_$_1$=0, x_$_3$=1), the two discriminant functions for **x** are

$$\mu_1(\mathbf{x}) = 1/2\,(0.9\times 0.8) = 0.36$$
$$\mu_2(\mathbf{x}) = 1/2\,(0.5\times 0.4) = 0.10$$

(Step 3) Choose a class label by the maximum discriminant function and note whether there is a mismatch with the class label whose distribution is currently being used. In the example, label ω_1 is chosen so the misclassification counter remains unchanged.

Figure 1 shows the probability of error versus the number of selected features, d, for c = 10 classes. Each point on the figure is the average error over the 50 random matrices.

As expected, SFS gives the lowest error. The single-best and the feature-pairs methods are approximately the same with a slight preference to feature-pairs, and the class-pairs method is the worst. For d=2 selected features, SFS is the second best method because feature pairs selects the true best pair features.

Figure 1. Probability of error versus the number of selected features (n=20, c=10).

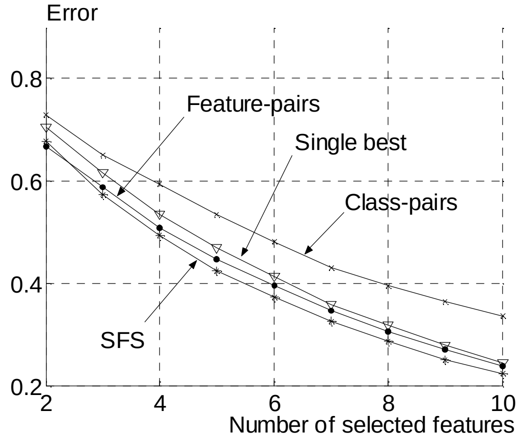


Table 4 gives the classification error averaged across the 50 random matrices of probabilities for 2 and 10 selected features (out of 20), for c = 3,…, 10 classes.

Table 4. Classification error (in %) with 2 and 10 features for c = 3,…, 10 classes. CP stands for class-pairs method, SB for the single-best method and FP for the feature-pairs method.

(a)

d = 2 selected features

c	CP	SFS	SB	FP
3	21.2	17.9	22.7	16.8
4	40.1	31.7	36.1	30.3
5	49.6	42.9	47.2	41.1
6	57.9	51.0	54.2	49.4
7	62.6	56.2	60.3	54.3
8	67.5	61.3	64.3	59.4
9	70.2	65.1	67.8	63.8
10	72.8	67.8	70.6	66.8

(b)

d = 10 selected features

c	CP	SFS	SB	FP
3	14.4	4.2	4.4	4.5
4	16.8	7.3	7.9	8.0
5	16.1	9.8	10.8	11.2
6	21.2	13.7	15.0	15.1
7	25.0	15.5	17.2	17.3
8	29.1	18.4	20.4	19.8
9	31.2	20.8	23.0	22.8
10	33.6	22.3	24.3	23.9

The results in Table 4 confirm the superiority of SFS for more than 2 features and it also shows that the class-pairs method gives the largest error. There is an interesting turn about the single-best and feature-pairs methods. For small number of classes (3 to 7) SB was slightly better whereas for larger number of classes (8 to 10) FP was the better of the two methods. This behavior is an indication that for larger scale problems FP may be the more accurate method.

A Larger-Scale Simulation Study

SFS was excluded from this experiment because of its large computational time. The same experiments, as in the previous section, were run with a total number of features $n = 100$ and number of classes $c = 50$. The number of selected features was $d \in \{5, 10, 15,…, 50\}$. Figure 2 shows the error versus the number of selected features for SB, CP and FP. The curves are close together but the errors for all d are related as $E_{FP} < E_{SB} < E_{CP}$. The differences between E_{FP} and E_{SB} are not statistically significant.

Figure 2. Probability of error versus the number of selected features (n = 100, c = 50).

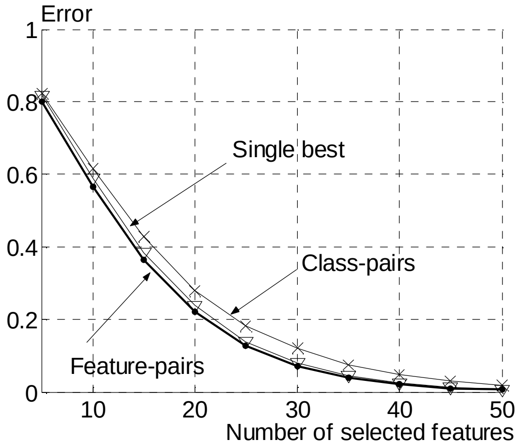
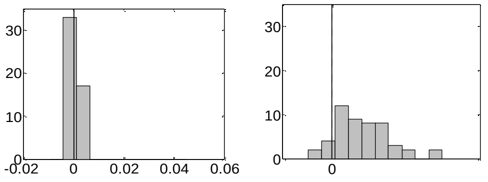


Figure 3 shows the histogram of the 50 differences $E_{SB} - E_{FP}$ for 50 and 25 selected features. For 50 features, $E_{SB} - E_{FP}$ was positive in 64% of the runs, the same in 6% of the runs and negative in 30% of the runs. For 25 selected features, $E_{SB} - E_{FP}$ was positive in 94% of the runs and negative in 6% of the runs. This suggests that there may be optimal ratios c:d:n for which FP is distinctly better than SB.

Figure 3. Histograms of the 50 differences $E_{SB} - E_{FP}$ for d = 50 selected features (a) and d = 25 selected features (b).



(a) (b)

The computational time ratio for the three methods was approximately C_{SB}:C_{CP}:C_{FP} = 1:8:23.

The above simulations do not assume any relationship between the classes. The matrices are generated uniformly which means that the correlations between the columns will be close to 0, as will be the correlations between the rows. In real problems, the class profiles will rarely be uncorrelated. Below, the four methods are explored on two real diagnostic problems where only probabilistic data is available.

An Application to Diagnosis of BSE in cattle and Scrapie in Sheep

The above feature selection methods were applied for selecting diagnostic signs in two problems coming from veterinary medicine.

BSE and Scrapie are fatal neurodegenerative diseases. Both are notifiable diseases which have no known cure. There is currently no ante-mortem test for the diseases that can be used routinely in the field. Notifiable diseases have a major impact on human health, welfare and economics. There was a BSE epidemic in Britain in the 90’s and with the first ever BSE case diagnosed in the USA at the end of 2003, the problem of these diseases is global. Therefore, the recognition of the clinical presentations of the two diseases and the need to differentiate them from other diseases is important. In veterinary medicine, prevalence of disease, the conditional dependencies of clinical signs, and the sign frequencies within diseases are rarely, if ever available; demonstrating the need to work with probability data.

Table 5 shows the results from the feature selection experiments with the BSE data. SFS was applied to select 10 of the 242 features and simulated data from the distributions of the 57 classes. The three selection methods SB, CP, and FP, which have lower capacity than SFS were run for $d = 10$ features too. The first 4 rows in Table 4 show the classification error for $d = 10$.

Next, the class-pairs method was run letting it stop when all class pairs have been accounted for. CP selected a total of 58 features. Leaving SFS aside, the other two low-complexity methods were run for 58 features. The classification error is displayed in rows 5-8 in Table 5. Finally, the error with using all features was estimated as a tight lower bound on the classification error.

Table 5. Results from feature selection on the BSE probabilities.

Method (*d*)	Error
SFS (10)	0.4258
SB (10)	0.6432
CP (10)	0.5865
FP (10)	0.5482
CP (58)	0.0172
SB (58)	0.0309
FP (58)	0.0256
ALL (242)	0.0049

The results show that the closest rival to SFS for small number of features is the FP method proposed here. Contrary to the results in the previous section though, CP is better than SB. This shows that in real-life problems when there is dependency between the classes, CP may be a better solution than SB. When run all the way, CP provides the smallest classification error of the three low complexity methods followed by FP and then SB.

Note the large differences between the error probabilities for small number of features. These differences strongly suggest that SFS should be applied as long as the computation time is acceptable. To illustrate the differences between the selected sets of features, Table 6 shows the signs selected by SFS (a) and SB (b) in the order they entered the set.

The same pattern of experiments was repeated for the data containing the probabilities for Scrapie and 62 alternative diseases. Twelve features were selected by SFS. The 3 lower-complexity methods were run for $d = 12$. The errors are shown in Table 7. The class-pairs method (CP) was run again until all class pairs were covered. The number of selected features was 77. SB and FP were then run for the same number of features. Table 7 ranks the feature selection methods exactly in the same way as Table 5. Again, the discrepancies with the simulation study in the previous sub-section can be attributed to the fact that the classes here are not independent. The CP method manages to capture some dependency between the classes and, if run all the way, it selects better subsets of features than SB and FP. Table 8 mirrors table 6 by showing the signs selected for diagnosing Scrapie and the 63 alternative diseases.

Table 6. Signs selected by SFS and SB for diagnosing BSE and 56 other diseases in cattle

(a) Signs selected by SFS
Gait abnormal, unspecified
Circling in one direction
Hypo-responsive to external stimuli
Milk yield less than normal (individual)
Rumen rate nil, (0 per 2min)
Eye menace response absent
Hyper-responsive to external stimuli
Dyspoena, unspecified
Posture recumbency
Temperature >39.5 degrees Celsius

(b) Signs selected by SB
Gait abnormal, unspecified
Dyspoena, unspecified
Dyspoena, rate increased shallow
Diarrhoea, unspecified
Gait uncoordinated\exaggerated
Rumen rate slow (1 per 2min)
Diarrhoea, acute, profuse
Circling in one direction
Gait stiff
Head rotated, tilted or deviated

Table 7. Results from feature selection on the Scrapie probabilities.

Method (*d*)	Error
SFS (12)	0.5975
SB (12)	0.7635
CP (12)	0.6930
FP (12)	0.6610
CP (77)	0.0625
SB (77)	0.0992
FP (77)	0.0649
ALL (285)	0.0252

Table 8. Signs selected by SFS and SB for diagnosing Scrapie and 63 other diseases in sheep

(a) Signs selected by SFS
Foul odour skin
Mastitis
Exercise intolerance
Paraparesis
Weight Loss
Generalized weakness
Anorexia
Generalized lameness or stiffness
Ataxia
Underweight, thin etc
Dullness
Reluctant to move

(b) Signs selected by SB
Foul odour skin
Mastitis
Matted \ dirty wool \ hair
Moist skin\wool \hair
Skin necrosis
Exercise intolerance
Hyperkeratosis
Lymphadenopathy
Alopecia
Pruritus
Weight loss
Dullness

Conclusion

The problem of selecting a subset of *n* binary features to discriminate between *c* mutually exclusive classes was explored. The information available here is in the form of an $n\times c$ table with class-conditional probabilities for the *n* binary features, i.e., $P(x_i=1|\omega_j)$, $i = 1,\ldots,n$, $j = 1,\ldots,c$. Selecting the best subset of features seems easy because all the probabilistic information is available and the features are assumed to be independent. The difficulty comes from the complexity of the evaluation of the theoretical classification error for a subset of features.

An easy way out would be to generate a sample and run it through the Naïve Bayes classifier using only the features in the subset. Three methods were applied from the literature (SFS, SB and CP) and a method was proposed based on features pairs (FP) for feature selection using probabilities. It was found that SFS was the most accurate but also the most computationally demanding of the four methods. The simulation experiments with generated random distributions suggested that CP was inferior to SB and FP, but did not favor strongly any of SB or FP. The experiments with two real data matrices from veterinary medicine demonstrated that CP is also a valuable method when larger subsets of features are acceptable. FP was found to be the best alternative to SFS for small and medium subsets.

There are at least two caveats that need to be mentioned. First, features are rarely independent in real life problems. By assuming independence, one runs the risk of missing an important feature which does not have a reasonable predictive value on its own, but is highly important in combination with others. However, in the absence of any further information, the independence assumption is the only option. Second, the estimates of the probabilities given as the information to work upon (Table 1) might not be very close to the true probabilities. A sensitivity study can be run by perturbing the probability estimates and observing how the selected feature subset changes.

The acid test for the quality of the selected subset of features would be the error on real data. However, the aim of this study is a preliminary feature selection so that a real data set can be collected using these features. Therefore, at this stage, a reasonably large feature set should be provided. The hope is that highly discriminative combinations of features will be discovered within using systematically collected data.

References

Aha, D.W. & Bankert, R. L. (1995). A comparative evaluation of sequential feature selection algorithms. *In Proc. 5th International Workshop on AI and Statistics*, pages 1-7, Ft Lauderdale, FL.

Blum, A. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*, 245-271.

Brightling, P., Larcombe, M. T., Blood, D. C., & Kennedy, P. C. (1996). Development and the use of Bovid-3, an expert system for veterinarians involved in diagnosis, treatment and prevention of diseases of cattle. In *XIX World Buitrics Congress Proceedings, 2*, pages 528-532.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis, 1*, 131-156.

Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103-130.

Jain, A. K., & Zongker, D. (1997). Feature selection: evaluation, application and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 153-158.

Jain, A. K., & Chandrasekaran, B. (1982). Dimensionality and sample size considerations in pattern recognition practice. In Krishnaiah, P.R. & Kanal, L. N., editors, *Handbook of Statistics*, pages 835--855. North Holland.

Ji, H. & Bang, S. Y. (2000). Feature selection for multiclass classification using pairwise discriminatory measure and covering concept. *Electronics Letters, 36(6)*,524-525

MAFF (2000). Animal Health 2000. MAFF, London.

Patrick, E. (1972). *Fundamentals of Pattern Recognition*. Prentice-Hall, Inc., Englewood Cliffs, N.J.

Stearns, S. (1976). On selecting features for pattern classifiers. *In Proc 3-d International Conference on Pattern Recognition*, pages 71-75, Coronado, CA.

Tax. D., & Duin. R. (2002). Using two-class classifiers for multi-class classification. In *Proc. 16th International Conference on Pattern Recognition*, *2*, 124-127.

Toussaint, G. T. (1971). Note on optimal selection of independent binary-valued features for pattern recognition. *IEEE Transactions on Information Theory*, *17*, 618.

Van Campenhout, J. M. (1982). Topics in measurement selection. In Krishnaiah, P. R. & Kanal, L.N., editors, *Handbook of Statistics*, pages 793-803. North Holland.

White, M.E. (1984). Consultant: computer-assisted differential diagnosis, *Veterinary Computing*, *2*, 9-12.