In *Proceedings of the Fourteenth Australian Joint Conference on Artificial Intelligence*, Adelaide, December, 2001, Berlin: Springer, pp 545-556.

Candidate elimination criteria for Lazy Bayesian Rules

Geoffrey I. Webb

School of Computing and Mathematics
Deakin University
Geelong Vic. 3217
webb@deakin.edu.au

Abstract. Lazy Bayesian Rules modifies naive Bayesian classification to undo elements of the harmful attribute independence assumption. It has been shown to provide classification error comparable to boosting decision trees. This paper explores alternatives to the candidate elimination criterion employed within Lazy Bayesian Rules. Improvements over naive Bayes are consistent so long as the candidate elimination criteria ensures there is sufficient data for accurate probability estimation. However, the original candidate elimination criterion is demonstrated to provide better overall error reduction than the use of a minimum data subset size criterion.

Keywords: machine learning

1 Introduction

Naive Bayes [4] is a simple and efficient approach to classification learning that has clear theoretical motivation and support. It has been demonstrated to provide competitive prediction error to more complex learning algorithms [8, 11], especially when training set sizes are small [17].

Lazy Bayesian Rules (LBR) [17, 18] modifies naive Bayes, seeking to retain its simplicity, efficiency, and clear theoretical foundations, while weakening the attribute independence assumption that can reduce naive Bayes' prediction accuracy. LBR has been demonstrated to provide prediction accuracy comparable to boosting decision trees [18].

This paper describes naive Bayes and LBR. It then examines one of the components of LBR, the candidate elimination criterion by which LBR determines whether an attribute should be a candidate for factoring out of the attribute independence assumption. Experiments demonstrate that improvements over naive Bayes are consistent so long as the candidate elimination criterion ensures there is sufficient data for accurate probability estimation. The original candidate elimination criterion is demonstrated to be better at determining when to stop than the use of a minimum data subset size criterion.

2 Naive Bayes

Naive Bayes is motivated as follows. When classifying an instance $X = x_1, x_2, \ldots x_n$, whose class y is unknown, classification error will be minimized by selecting

$$argmax_y(P(y \mid X)) \tag{1}$$

the class that is most probable given X. A problem arises where $P(y \mid X)$ is to be estimated from the frequencies of X and y in a set of data $\mathcal{D} = \langle X_1, y_1 \rangle, \langle X_2, y_2 \rangle, \ldots \langle X_k, y_k \rangle$. In the limit, when the dataset contains the entire domain with respect to which probabilities are to be determined,

$$P(W) = F(W) \tag{2}$$

where $F(W)$ is the frequency with which W occurs in $\mathcal{D}$. As $P(W \mid Z) = P(W \wedge Z)/P(Z)$, $P(y \mid X)$ might be estimated by the approximation

$$P(y \mid X) \approx \frac{F(y \wedge X)}{F(X)} \,. \tag{3}$$

However, in many cases X and $y \wedge X$ will not occur frequently enough in the data for accurate estimation of the probabilities from the frequencies. In fact, unless the set of data is very comprehensive, X and $y \wedge X$ may not occur at all. In this context, Bayes rule

$$P(y \mid X) = \frac{P(y)P(X \mid y)}{P(X)} \tag{4}$$

may be used to derive alternative probabilities, by estimation of which the target probability can be estimated. As P(X) is invariant across different values of y,

$$P(y \mid X) \propto P(y)P(X \mid y) \tag{5}$$

and hence we need not estimate the denominator. However, this still leaves the problem of estimating $P(X \mid y)$ when $y \wedge X$ does not occur frequently in the data. By making the conditional independence assumption

$$P(x_1, x_2, \ldots x_n \mid y) = \prod_{i=1}^{n} P(x_i \mid y) \tag{6}$$

$P(X \mid y)$ can be estimated by estimation of each $P(x_i \mid y)$, latter estimates being more reliable as each conjunct is likely to occur with relatively high frequency.

Naive Bayes is classification using (1), estimating $P(y \mid X)$ by (4) and (6). As (1) minimizes prediction error, naive Bayes will minimize prediction error except in so far as the conditional independence assumption is violated and the estimation from data of probabilities $P(y)$ and $P(x_i \mid y)$ is inaccurate.

However, while the conditional independence assumption makes the estimation of $P(X \mid y)$ feasible, and naive Bayes delivers competitive classification performance for small data sets, the independence assumption is likely to be violated

for many real world classification tasks. Notwithstanding Domingos & Pazzani's [3] observation the such violations are harmless so long as they do not affect the relative rank of each estimate of $P(y \mid X)$, research into semi-naive Bayesian learning has demonstrated that such violations are frequent and that explicit actions to alleviate their effect can reduce error [6, 7, 9–14, 16].

3 Lazy Bayesian Rules

LBR utilizes an alternative to Bayes theorem (4),

$$P(y \mid Z_1 \wedge Z_2) = \frac{P(y \mid Z_2)P(Z_1 \mid y \wedge Z_2)}{P(Z_1 \mid Z_2)}. \tag{7}$$

The derivation of this equality is given in Zheng & Webb [17]. Given that $P(Z_1 \mid Z_2)$ is invariant across values of y,

$$P(y \mid Z_1 \wedge Z_2) \propto P(y \mid Z_2)P(Z_1 \mid y \wedge Z_2). \tag{8}$$

Where Z_1 is a conjunction of terms, $Z_1 = z_1 \wedge z_2 \wedge \ldots z_m$, a conditional attribute indpendence assumption

$$P(Z_1 \mid y \wedge Z_2) \approx \prod_{i=1}^{m} P(z_i \mid y \wedge Z - 2) \tag{9}$$

can be used to estimate $P(Z_1 \mid y \wedge Z_2)$.

Like naive Bayes, LBR estimates $P(y \mid X)$ for each y, selecting the y that maximizes the estimate. LBR differs from naive Bayes by segmenting the conjuncts of X into two groups, Z_1 and Z_2, and then using (7) in place of (4) and (9) in place of (6). Like naive Bayes, LBR will minimize classification error except in so far as its independece assumption is violated and the estimation of the required probabilities is incorrect.

A principal advantage of LBR over naive Bayes is that its independence assumption is weaker. Whereas naive Bayes assumes independence between all conjuncts given the class, LBR assumes independence only between the conjuncts in Z_1 given both the class and the conjuncts in Z_2.

The assumption of independence between fewer attributes is an advantage as fewer attribute interdependencies will be assumed incorrectly.

The assumption of independence under stronger conditions is also a major advantage. Consider the conditions $age > 70$, *senile*, and *nocturia*. Each of these three conditions will be highly interdependent with the others, as senility and nocturia are both correlated with age. However, given $age > 70$, *senile* and *nocturia* may be independent, as the interdependence of senility and nocturia may solely result from the respective interdependencies with age. That is, while $P(senile \wedge nocturia) \neq P(senile)P(nocturia)$, $P(senile \wedge nocturia \mid age > 70) = P(senile \mid age > 70)P(nocturia \mid age > 70)$. If this is the case (and conditioning

on y does not produce independence between these attributes),

$$P(y \mid age > 70 \wedge senile \wedge nocturia) \neq \frac{P(y)P(age > 70 \mid y)P(senile \mid y)P(nocturia \mid y)}{P(age > 70 \wedge senile \wedge nocturia)} \quad (10)$$

so naive Bayes will be inaccurate. However, LBR may be accurate because

$$P(y \mid age > 70 \wedge senile \wedge nocturia) = \frac{P(y \mid age > 70)P(senile \mid age > 70 \wedge y)P(nocturia \mid age > 70 \wedge y)}{P(senile \wedge nocturia \mid age > 70)}. \quad (11)$$

If these two advantages were the only consideration, it would be advantageous to factor out all conditional interdependencies by placing all attributes in Z_2. However, placing an attribute in Z_2 carries one disadvantage in addition to its advantages. Each conditional probability $P(Z_1 \mid y \wedge Z_2)$ will be estimated by the approximation $P(z_i \mid y \wedge Z_2) \approx F(z_i \wedge y \wedge Z_2)/F(y \wedge Z_2)$. The more attributes in Z_2 the lower the frequency in $\mathcal{D}$ of both $z_i \wedge y \wedge Z_2$ and $y \wedge Z_2$ and hence the lower the expected accuracy of the approximation. Hence, LBR engages in a process of seeking to balance gains in expected accuracy due to factoring out harmful attribute interdependencies against losses in expected accuracy due to decreased expected accuracy of estimation of the required parameters.

LBR manages this trade-off by performing leave-one-out cross-validation once for each attribute-value using the conditional formula that results from including that value in Z_2. An attribute-value v is only considered as a candidate if the number of examples misclassified by including v in Z_2 but correctly classified by excluding it is significantly lower than the number correctly classified by including it but misclassified by excluding it. A matched-pair binomial sign test with significance level 0.05 is used to assess significance. The candidate with the lowest error is selected and the process repeated until no candidates remain.

LBR uses *lazy learning*. Calculation is performed when an object is to be classified. Only the attribute-values of that object are considered for inclusion in Z_2. The algorithm is presented in Table 1. Note that this algorithm does not explicitly maintain Z_2. Each A_{best} found is added to Z_2. Z_1 is the values of the attributes in Att for E_{test}. Z_2 is the remaining attribute values for E_{test}. The effect of factoring out Z_2 is achieved by selecting for $D_{training}$ the subset of instances that satisfy the conditions in Z_2. When the probability of an attribute value conditional on a class is estimated from a training set, the m-estimate [2] with m = 2 is used. When the probability of a class is estimated, the Laplace estimate [2] is used. When applying naive Bayesian classification, if two or more classes obtain equal highest probability estimates, one is selected at random.

4 Alternative candidate elimination strategies

LBR eliminates from consideration as candidates for A_{best} attribute values that fail to reduce error by a statistically significant amount using leave-one-out cross-

Table 1. The Lazy Bayesian Rule learning algorithm

LBR(Att, $D_{training}$, E_{test})
  $INPUT$: Att: a set of attributes,
    $D_{training}$: a set of training examples described using Att and classes,
    E_{test}: a test example described using Att.
  $OUTPUT$: a predicted class for E_{test}.
$LocalNB$ = a naive Bayesian classifier trained using Att on $D_{training}$
$Errors$ = errors of $LocalNB$ estimated using N-CV on $D_{training}$
$Cond = true$
REPEAT
  $TempErrors_{best}$ = the number of examples in $D_{training}$ + 1
  FOR each attribute A in Att whose value v_A on E_{test} is not missing *DO*
    D_{subset} = examples in $D_{training}$ with $A = v_A$
    $TempNB$ = a naive Bayesian classifier trained using $Att - \{A\}$ on D_{subset}
    $TempErrors$ = errors of $TempNB$ estimated using N-CV on D_{subset} +
      errors from $Errors$ for examples in $D_{training} - D_{subset}$
    IF (($TempErrors < TempErrors_{best}$) *AND*
      (***TempErrors* is significantly lower than *Errors***))
    THEN
      $TempNB_{best} = TempNB$
      $TempErrors_{best} = TempErrors$
      $A_{best} = A$
  IF (an A_{best} is found)
  THEN
    $Cond = Cond \wedge (A_{best} = v_{A_{best}})$
    $LocalNB = TempNB_{best}$
    $D_{training} = D_{subset}$ corresponding to A_{best}
    $Att = Att - \{A_{best}\}$
    $Errors$ = errors of $LocalNB$ estimated using N-CV on $D_{training}$
  ELSE
    EXIT from the *REPEAT* loop
classify E_{test} using $LocalNB$
RETURN the class

validation on the training data. The condition that enforces this strategy is set in bold type in Table 1.

This approach was motivated by the desire to eliminate from consideration attribute values for which factoring out appears to reduce error only by chance. Inevitably different formulae will result in variability in prediction performance, and by chance some will perform better than others. By eliminating candidates for which the difference in performance was not significantly greater than the baseline performance, we reduce the risk of selecting an attribute value that appears to improve performance only by chance. By using leave-one-out cross-validation classification performance as the selection criterion we aimed to measure the effect of both the improvement brought about by weakening the at-

tribute independence assumptions and the decrease in accuracy of estimation brought about by decreased data.

Our previous experiments indicate that this strategy is very effective at managing this trade-off and results in very strong classification performance [17, 18]. However, an alternative argument can be constructed that as the only harm in moving an attribute-value to Z_2 lies in the reduction in accuracy of estimation of the parameters, the candidate elimination strategy should be aimed directly at combating this problem. In other words, an attribute-value should remain a candidate for inclusion in Z_2 so long as there is sufficient data to reliably estimate the required parameters.

This paper tests this proposal by substituting for the LBR candidate elimination test (set in bold type in Table 1) an alternative test that is based solely on the number of examples in $D_{training}$ that have the relevant value. This is predicated on the assumption that if there are sufficient examples of a given value, estimation of the frequency of that value and the probability of each class given that value will be sufficiently accurate for accurate classification. Three values are considered, 30, 100, and 500. The first value, 30, was selected as 30 is commonly held to be the minimum sample from which one should draw statistical inferences. The last value, 500, was selected as a sufficiently large number that accurate estimation of parameters should be possible. 100 was selected as an intermediate value. This new strategy was implemented by substituting the condition $|D_{subset}| \geq MinSize$ for the candidate elimination condition set in bold type in Table 1, where $MinSize$ was set respectively to 30, 100, and 500. This approach will default to naive Bayes when the dataset size is less than $MinSize$ as all candidates will be eliminated.

5 Experiments

For the first experiment, naive Bayes and the four variants of LBR (the original candidate elimination criterion, called hereafter LBR, and candidate elimination using $MinSize$ set to each of 30, 100, and 500, called hereafter $MinSize = 30$, $MinSize = 100$, and $MinSize = 500$, respectively). The 29 datasets from the UCI repository [1] were used that have been used in previous LBR experiments [17, 18] (a selection based on those used in prior semi-naive Bayesian learning research). These datasets are described in Table 2. The experimental method of [18] was replicated, ten repetitions of three-fold cross-validation, with different random selection of folds during each repetition. Numeric attributes were discretized using Fayyad & Irani's [5] MDL discretization algorithm on the training data for a given fold. Each algorithm was evaluated with the same sequence of thirty training and test set pairs formed in this manner.

The average error rates of each algorithm for each data set are presented in Table 3. Also presented for each algorithm is the mean error across all data sets, the geometric mean error ratio compared with naive Bayes, the win/loss record between the algorithm and naive Bayes, and the win/loss record between the algorithm and LBR. The mean error is a very gross measure of performance

Table 2. Description of data sets

Domain	Size	No. of Classes	No. of Attributes	
			Numeric	Nominal
Lung cancer	32	3	0	56
Labor negotiations	57	2	8	8
Postoperative patient	90	3	1	7
Zoology	101	7	0	16
Promoter gene sequences	106	2	0	57
Echocardiogram	131	2	6	1
Lymphography	148	4	0	18
Iris classification	150	3	4	0
Hepatitis prognosis	155	2	6	13
Wine recognition	178	3	13	0
Sonar classification	208	2	60	0
Glass identification	214	6	9	0
Audiology	226	24	0	69
Heart disease (Cleveland)	303	2	13	0
Soybean large	307	19	0	35
Primary tumor	339	22	0	17
Liver disorders	345	2	6	0
Horse colic	368	2	7	15
House votes 84	435	2	0	16
Credit screening (Australia)	690	2	6	9
Breast cancer (Wisconsin)	699	2	9	0
Pima Indians diabetes	768	2	8	0
Annealing processes	898	6	6	32
Tic-Tac-Toe end game	958	2	0	9
LED 24 (noise level = 10%)	1000	10	0	24
Solar flare	1389	2	0	10
Hypothyroid diagnosis	3163	2	7	18
Splice junction gene sequences	3177	3	0	60
Chess (King-rook-vs-king-pawn)	3196	2	0	36

as error rates on different domains are incommensurable, but provides an approximate indication of relative performance. The geometric mean error ratio is the geometric mean of the value for each data set of the error of the algorithm divided by the error of naive Bayes. The geometric mean is more appropriate than the mean as an aggregate measure of ratio values [15]. The win/loss records with respect to naive Bayes and LBR list the number of domains for which the error of the algorithm is lower than the error of, respectively, naive Bayes and LBR.

The first point of interest is that LBR has scored slightly fewer wins and slightly more losses with respect to naive Bayes than in previous experiments [17, 18]. However, it is notable that all of LBR's losses to naive Bayes occur with smaller data sets. The largest is credit screening, containing 690 examples, and for which the training set size in three-fold cross-validation will be 430. It is also

Table 3. Error rates

			MinSize		
	NB	LBR	30	100	500
Lung cancer	0.534	0.544	0.534	0.534	0.534
Labor negotiations	0.098	0.098	0.105	0.098	0.098
Postoperative patient	0.378	0.386	0.383	0.378	0.378
Zoology	0.059	0.059	0.063	0.059	0.059
Promoter gene sequences	0.109	0.112	0.170	0.109	0.109
Echocardiogram	0.296	0.297	0.306	0.296	0.296
Lymphography	0.182	0.182	0.196	0.182	0.182
Iris classification	0.066	0.066	0.065	0.066	0.066
Hepatitis prognosis	0.144	0.144	0.175	0.144	0.144
Wine recognition	0.023	0.023	0.030	0.023	0.023
Sonar classification	0.245	0.245	0.240	0.248	0.245
Glass identification	0.238	0.237	0.240	0.246	0.238
Audiology	0.277	0.277	0.290	–	0.278
Heart disease (Cleveland)	0.171	0.171	0.200	0.177	0.171
Soybean large	0.143	0.101	0.149	0.115	0.143
Primary tumor	0.534	0.535	0.568	0.551	0.534
Liver disorders	0.361	0.363	0.359	0.355	0.361
Horse colic	0.208	0.199	0.197	0.192	0.208
House votes 84	0.100	0.067	0.086	0.057	0.100
Credit screening (Australia)	0.146	0.147	0.166	0.154	0.146
Breast cancer (Wisconsin)	0.026	0.026	0.041	0.034	0.026
Pima Indians diabetes	0.252	0.251	0.267	0.253	0.252
Annealing processes	0.030	0.028	0.030	0.026	0.030
Tic-Tac-Toe end game	0.295	0.185	0.145	0.220	0.295
LED 24 (noise level = 10%)	0.261	0.260	0.265	0.263	0.259
Solar flare	0.039	0.015	0.020	0.017	0.031
Hypothyroid diagnosis	0.018	0.015	0.020	0.017	0.018
Splice junction gene sequences	0.046	0.044	0.077	0.057	0.043
Chess (King-rook-vs-king-pawn)	0.124	0.028	0.021	0.021	0.032
Mean	0.185	0.174	0.186	0.178	0.183
Geo mean vs NB		0.930	1.081	0.960	0.975
W/L vs NB		12/7	8/19	10/9	4/1
W/L vs LBR			8/21	10/13	9/11

notable that of the seven losses to naive Bayes, only three are by more than 0.002, a very small margin. While the win loss record is not significant at the 0.05 level using a one-tailed binomial sign test (p=0.1796), the mean across all data sets is substantially lower, and, more significantly, the geometric mean error ratio strongly favours LBR. It is notable that for the largest data sets LBR is consistently winning, halving naive Bayes' error with respect to solar flare and quartering it with respect to chess.

These results suggest that the LBR's candidate elimination strategy might be suboptimal for small numbers of examples. In other words, it is credible that the candidate elimination strategy does not take adequate account of whether

there is sufficient data for reliable estimation of the required parameters. It was this supposition, derived from previous experiments, that motivated the current study.

Of the three minimum example settings, it seems clear that $MinSize = 30$ provides the worst performance. On all metrics it performs worse than naive Bayes. The geometric mean error ratio strongly favours naive Bayes as does the win/loss record (significantly at the 0.05 level, one-tailed binomial sign test p=0.0261). The win/loss record against LBR strongly and significantly favours LBR (p=0.0120).

The situation with respect to $MinSize = 100$ is less clear cut. It wins as often as it loses against naive Bayes. The mean, and more significantly, the geometric mean error ratio, both favour $MinSize = 100$ over naive Bayes, indicating that the magnitude of its wins tends to be greater than the magnitude of its losses. The win/loss record with respect to LBR favours the latter, but not significantly so (p=0.3388).

The results with respect to $MinSize = 500$ appear much more straightforward, however. First, it is necessary to consider the outcome for audiology. It might initially appear anomalous that $MinSize = 500$ achieves a different outcome to naive Bayes for a dataset with fewer than 500 examples. The explanation, however, is straightforward. For this dataset there is one classification during the ten sets of three-fold cross-validation for which naive Bayes scores two classes as equi-probable and for which the random resolution of this draw selected different classes for naive Bayes and $MinSize = 500$. In this case the random outcome favoured naive Bayes. Of the larger datasets, for which $MinSize = 500$ had the opportunity to move attribute-values to Z_2, $MinSize = 500$ consistently wins over naive Bayes. Restricting the analysis to datasets for which $MinSize = 500$ modifies the behaviour of naive Bayes, the win/loss record is 4/0, which approaches significance at the 0.05 level (p=0.0625).

Table 4 presents the average size of Z_2 ($|Z_2|$) and the average number of examples from which the probabilities are estimated ($|\mathcal{D}|$) for each dataset for LBR and its three variants. It is striking that when there is sufficient data for the constraint on minimum numbers of examples to be satisfied, this alternative approach tends to add many more values to Z_2. Consider, for example, $MinSize = 500$ on the King-rook-vs-king-pawn data. More than three times the number of attribute values are added to Z_2 even though there is not a large difference in the average number of examples selected by each Z_2. This is because $MinSize = 500$ can keep selecting additional attribute values so long as they cover sufficient cases while LBR requires that the selection results in a significant reduction in error.

Of the six datasets for which $MinSize = 500$ is able to select attribute values for Z_2, LBR obtains lower error for four and higher for two. However, for the two for which LBR obtains higher error, the magnitude of the difference is very small whereas the magnitude is relatively high for those datasets for which LBR achieves lower error. These results suggest that the significance test in LBR's candidate elimination strategy does confer an advantage. Further support

Table 4. Mean $|Z_2|$ and examples available for estimation of parameters

	LBR		MinSize=30		MinSize=100		MinSize=500	
	$\|Z_2\|$	$\|\mathcal{D}\|$	$\|Z_2\|$	$\|\mathcal{D}\|$	$\|Z_2\|$	$\|\mathcal{D}\|$	$\|Z_2\|$	$\|\mathcal{D}\|$
Lung cancer	0.07	20.7	0.00	21.3	0.00	21.3	0.00	21.3
Labor negotiations	0.00	38.0	0.24	36.4	0.00	38.0	0.00	38.0
Postoperative patient	0.05	58.9	1.25	40.9	0.00	60.0	0.00	60.0
Zoology	0.00	67.3	4.13	35.0	0.00	67.3	0.00	67.3
Promoter gene sequences	0.01	70.2	0.47	53.5	0.00	70.7	0.00	70.7
Echocardiogram	0.02	87.1	1.85	49.7	0.00	88.0	0.00	88.0
Lymphography	0.05	97.8	4.31	43.0	0.00	98.7	0.00	98.7
Iris classification	0.00	100.0	0.84	48.9	0.00	100.0	0.00	100.0
Hepatitis prognosis	0.02	102.2	4.28	36.4	0.00	103.3	0.00	103.3
Wine recognition	0.00	118.7	0.74	86.2	0.00	118.7	0.00	118.7
Sonar classification	0.27	126.3	12.39	40.0	5.91	102.4	0.00	138.7
Glass identification	0.12	135.3	3.41	58.1	1.01	118.8	0.00	142.7
Audiology	0.18	145.6	43.33	48.5	26.24	103.0	0.00	150.7
Heart disease (Cleveland)	0.05	175.5	3.31	47.2	1.66	128.1	0.00	180.0
Soybean large	0.99	161.0	13.38	47.6	8.37	109.9	0.00	204.7
Primary tumor	0.10	221.3	3.30	136.8	2.51	161.9	0.00	226.0
Liver disorders	0.28	217.5	4.60	61.3	2.97	138.8	0.00	230.0
Horse colic	0.47	192.4	3.59	54.1	2.01	130.5	0.00	245.3
House votes 84	0.67	188.5	5.44	54.8	2.43	115.7	0.00	290.0
Credit screening (Australia)	0.20	425.2	4.51	84.9	3.06	160.6	0.00	460.0
Breast cancer (Wisconsin)	0.00	466.0	2.38	150.6	1.82	269.9	0.00	466.0
Pima Indians diabetes	0.23	455.3	2.83	100.0	1.76	187.2	0.00	512.0
Annealing processes	0.09	570.0	5.05	121.4	4.76	208.1	2.52	545.0
Tic-Tac-Toe end game	1.65	165.1	2.86	45.3	1.85	121.0	0.00	638.7
LED 24 (noise level = 10%)	0.50	571.1	5.11	129.8	3.54	197.8	0.50	603.9
Solar flare	0.80	534.6	4.71	235.1	4.35	267.4	3.01	695.0
Hypothyroid diagnosis	0.28	1923.7	14.92	532.5	14.61	616.7	14.04	832.6
Splice junction gene sequences	0.39	1686.8	1.98	413.3	1.75	448.4	1.14	878.1
Chess (King-rook-vs-king-pawn)	3.67	572.5	15.62	136.2	15.30	169.2	11.28	551.7

for this conclusion is provided by a second study that compared naive Bayes, LBR, and $MinSize = 500$ in five larger datasets: phoneme (5438 examples), mush (8124), pendigits (10992), adult (48842), and shuttle (58000). As ten runs of three-fold cross-validation was infeasible for such large data sets, leave-one-out cross-validation was performed for 1000 randomly selected examples from each data set. For each of these examples, each algorithm was presented all the remaining examples in the dataset as a training set and the withheld example was then classified. The resulting error rates are presented in Table 5. As can be seen, both LBR and $MinSize = 500$ consistently achieve lower error than naive Bayes for these larger datasets. The win loss records of 5/0 are in both cases statistically significant at the 0.05 level using a one-tailed sign test (p=0.0313). While $MinSize = 500$ obtains marginally lower error than LBR on one dataset, LBR obtains substantially lower error on one and slightly lower on two.

Table 5. Error for large datasets

Dataset	NB	LBR	MinSize=500
phoneme	0.265	0.215	0.244
mush	0.014	0.000	0.000
pendigits	0.123	0.028	0.025
adult	0.163	0.132	0.137
shuttle	0.002	0.000	0.001

6 Conclusions

This paper makes two contributions to the literature on lazy Bayesian rules. First, it presents empirical results on much larger datasets than previously explored, providing statistically significant support for the hypothesis previously advanced [17] that LBR provides consistent advantage over naive Bayes for large datasets.

The primary motivation for the paper, however, was to investigate alternatives to the candidate elimination criteria employed in LBR, exploring the hypothesis that it will never be harmful to select candidate attribute values for inclusion in Z_2 that retain sufficient examples for reliable estimation of the required parameters. While some support for this hypothesis was obtained by the consistent capacity of $MinSize = 500$ to reduce error relative to naive Bayes, the error reduction capacity of LBR remains higher. This suggests that the significance test serves a useful function in implicitly assessing the relative gains from factoring out a harmful attribute interdependence against the losses from reducing the amount of data from which parameters are estimated.

Nonetheless, the $MinSize = 500$ strategy may offer computational advantages in some applications. This is because the overheads of assessing how many training cases are selected by a candidate attribute value are very low in comparison to the computational overheads associated with performing a matched-pair binomial sign test. For the extremely large datasets employed in some online datamining applications these computational considerations may outweigh the error reduction capacity of the significance test strategy.

Acknowledgements

I am grateful to Zijian Zheng for developing the lazy Bayesian rules software that was used in these experiments.

References

1. C. Blake and C. J. Merz. UCI repository of machine learning databases. [Machine-readable data repository]. University of California, Department of Information and Computer Science, Irvine, CA., 2001.

2. B. Cestnik, I. Kononenko, and I. Bratko. ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In I. Bratko and N. Lavrač, editors, *Progress in Machine Learning*, pp. 31–45. Sigma Press, Wilmslow, 1987.
3. P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. In *Proc. Thirteenth International Conference on Machine Learning*, pp. 105–112, Bari, Italy, 1996. Morgan Kaufmann.
4. R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York, 1973.
5. U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI-93: Proc. 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1027, Chambery, France, 1993. Morgan Kaufmann.
6. N. Friedman and M. Goldszmidt. Building classifiers using Bayesian networks. In *AAAI-96*, pp. 1277–1284, 1996.
7. R. Kohavi. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *KDD-96*, Portland, Or, 1996.
8. I. Kononenko. Comparison of inductive and naive Bayesian learning approaches to AUtomatic knowledge acquisition. In B. Wielinga, J. Boose, B. Gaines, G. Schreiber, and M. van Someren, editors, *Current Trends in Knowledge Acquisition*. IOS Press, Amsterdam, 1990.
9. I. Kononenko. Semi-naive Bayesian classifier. In *ECAI-91*, pp. 206–219, 1991.
10. P. Langley. Induction of recursive Bayesian classifiers. In *Proc. 1993 European Conference on Machine Leanring*, pp. 153–164, Vienna, 1993. Springer-Verlag.
11. P. Langley and S. Sage. Induction of selective Bayesian classifiers. In *Proc. Tenth Conference on Uncertainty in Artificial Intelligence*, pp. 399–406, Seattle, WA, 1994. Morgan Kaufmann.
12. M. J. Pazzani. Constructive induction of Cartesian product attributes. In *ISIS: Information, Statistics and Induction in Science*, pp. 66–77, Melbourne, Aust., August 1996. World Scientific.
13. M. Sahami. Learning limited dependence Bayesian classifiers. In *Proc. 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 334–338. AAAI Press, 1996.
14. M. Singh and G. M. Provan. Efficient learning of selective Bayesian network classifiers. In *Proc. 13th International Conference on Machine Learning*, pp. 453–461, Bari, 1996. Morgan Kaufmann.
15. G. I. Webb. Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, 40(2):159–196, 2000.
16. G. I. Webb and M. J. Pazzani. Adjusted probability naive Bayesian induction. In *Proc. Eleventh Australian Joint Conference on Artificial Intelligence*, pp. 285–295, Brisbane, Australia, 1998. Springer.
17. Z. Zheng and G. I. Webb. Lazy learning of Bayesian Rules. *Machine Learning*, 41(1):53–84, 2000.
18. Z. Zheng, G. I. Webb, and K. M. Ting. Lazy Bayesian Rules: A lazy semi-naive Bayesian learning technique competitive to boosting decision trees. In *Proc. Sixteenth International Conference on Machine Learning (ICML-99)*, pp. 493–502, Bled, Slovenia, 1999. Morgan Kaufmann.