Chapter 25

An Enhanced Selective Naïve Bayes Method with Optimal Discretization

Marc Boullé

France Telecom R&D, 2 avenue Pierre Marzin, 22307 Lannion Cedex, France
marc.boulle@francetelecom.com

In this chapter, we present an extension of the wrapper approach applied to the predictor. The originality is to use the area under the training lift curve as a criterion of feature set optimality and to preprocess the numeric variables with a new optimal discretization method. The method is experimented on the NIPS 2003 datasets both as a wrapper and as a filter for multi-layer perceptron.

25.1 Introduction

The Naïve Bayes approach is based on the assumption that the variables are independent within each output label, which can harm the performances when violated. In order to better deal with highly correlated variables, the Selective Naïve Bayes approach (Langley and Sage, 1994) uses a greedy forward search to select the variables. The accuracy is evaluated directly on the training set, and the variables are selected as long as they do not degrade the accuracy. For numeric variables, the probability distribution is evaluated according to a Gaussian distribution whose mean and variance are estimated on the training examples.

Although the approach performs quite well on datasets with a reasonable number of variables, it does not scale on very large datasets with hundred or thousands of variables, such as in marketing applications. We propose to enhance the original method by exploiting a new Bayes optimal discretization method called MODL and by evaluating the predictors with a new criterion more sensitive than the accuracy. In the NIPS Challenge, the method is experimented both as a wrapper approach (ESNB) and as a filter for multi-layer perceptron (ESNB+NN). The method is fast, robust and manages to find a good trade-off between the error rate and the number of selected variables. However, the method needs further improvements in order to reach better error rates.

We detail and comment the ESNB method in section 25.2 and focus on the MODL discretization method in section 25.3. We present the results of the method in the NIPS Challenge in section 25.4.

25.2 The Enhanced Selective Naïve Bayes Method

In this section, we describe three enhancements to the Selective Naïve Bayes method: the use of a new discretization method to pre-process numeric variables, the use of the area under the lift curve to evaluate the performance of predictors and a post-processing correction of the predicted output label probabilities. Lift curves summarize the cumulative percent of targets recovered in the top quantiles of a sample (Witten and Franck, 2000).

The evaluation of the probabilities for numeric variables has already been discussed in the literature (Dougherty et al., 1995, Hsu et al., 2002, Yang and Webb, 2002). Experiments have shown that even a simple Equal Width discretization with 10 bins brings superior performances compared to the assumption using a Gaussian distribution. In a selection process, the risk of overfitting the data raises with the number of variables. Slight losses in the quality of the evaluation of the probability distribution of the variables may have cumulated effects and lead to the selection of irrelevant or redundant variables. We propose to use a new supervized discretization method called MODL, which is Bayes optimal. This method is described in section 25.3.

In the wrapper approach, (Kohavi and John, 1997) propose to evaluate the selection process using accuracy with a five-fold cross validation. However, the accuracy criterion suffers from some limits, even when the predictive performance is the only concern. (Provost et al., 1998) propose to use Receiver Operating Analysis (ROC) analysis rather than the accuracy. In marketing applications for example, the lift curves are often used to evaluate predictors. In the context of variable selection, there are other justifications to replace accuracy by another criterion. In case of a skewed distribution of output labels, the accuracy may never be better than the majority accuracy, so that the selection process ends with an empty set of variables. This problem also arises when several consecutive selected variables are necessary to improve the accuracy. In the method proposed in (Langley and Sage, 1994), the selection process is iterated as long as there is no decay in the accuracy. This solution raises new problems, such as the selection of irrelevant variables with no effect on accuracy, or even the selection of redundant variables with either insignificant effect or no effect on accuracy. We propose to use the area under the lift curve, measured directly on the training set, to evaluate whether a new variable should be selected. If we note TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative) the four possible outcomes of the confusion matrix, the lift curve is obtained by plotting TP (in %) against $\frac{TP+FP}{TP+FP+TN+FN} \times 100\%$ for each confidence value, starting at (0,1) and ending at (1,0). At each step of the algorithm, the variable which brings the best increase of the area under the lift curve is choosen and the selection process stops as soon as this area does not rise anymore. This allows capturing slight enhancements in the learning process and helps avoiding the selection of redundant variables or probes that have no effect on the lift curve.

The last problem is related to the Naïve Bayes algorithm itself, which is a good rank estimator, but a weak probability estimator (Hand and Yu, 2001). We propose to add a correction to the estimation of the output labels probabilities at the end of the learning process, instead of using the standard 50% probability threshold to predict the output label. For a given probability threshold, we compute the resulting confusion matrix on the training set and score it owing to the chi-square criterion. The higher the chi-square criterion is, the more correlated are the predicted output labels and the true output labels. The best probability threshold is found by evaluating all possible confusion matrices, once the training examples have been sorted by decreasing probability of output label. This corresponds to finding the best point on the lift curve, owing to the maximization of the chi-square criterion of the related confusion matrix.

Altogether, the algorithm can be optimized in $O(n^2 m \log(m))$ time, where n is the number of input variables and m the number of training examples. The preprocessing step needs $O(nm \log(m))$ to discretize all the variables. The forward selection process requires $O(n^2 m \log(m))$ time, owing to the decomposability of the Naïve Bayes formula on the variables. The $O(m \log(m))$ term in the complexity is due to the evaluation of the area under the lift curve, based on the sort of the training examples. The post-processing correction needs $O(m \log(m))$ time by sorting the training examples and evaluating all possible probability thresholds. However, the irrelevant variables can be detected just after the discretization step: they are discretized in a single interval. If n_r is the number of relevant variable and n_s is the number of selected variables at the end of the learning process, the practical complexity of the algorithm is $O(n_r n_s m \log(m))$ time, which is often far below the theoretical complexity when the number of input variables is very high.

Enhanced Selective Naïve Bayes algorithm:

```
• Initialization
  – Discretize each variable with the MODL discretization method
  – Create an initial empty selected variable set and a set of relevant variables
• Selection process
  Repeat the following steps
  – For each unselected relevant variable
      · Compute the Naïve Bayes predictor with the additional variable, on the
        basis of the previous best predictor
      · Evaluate the resulting predictor with the area under the lift curve
  – If the evaluation is strictly improved
      · Add the best variable to the selected variable set
      · Update the best predictor
• Post-processing
  – Find the best decision threshold by maximizing the chi-square criterion of
    the contingency table
```

25.3 The MODL Discretization Method

In this section, we present the MODL approach which results in a Bayes optimal evaluation criterion of discretizations and the greedy heuristic used to find near-optimal discretizations.

The objective of the process is to induce a list of intervals that splits the value domain of a numeric input variable. The training examples are described by pairs of values: the numeric input value and the output label. If we sort the training examples according to their input values, we obtain a string S of output labels. We now propose the following formal definition of a discretization model.

Definition: A *standard* discretization model is defined by the following properties:

- the discretization model relies only on the order of the output labels in the string S, without using the values of the input variable;
- the discretization model allows to split the string S into a list of substrings (the intervals);
- in each interval, the distribution of the output labels is defined by the frequencies of the output labels in the interval.

Such a discretization model is called a SDM model.

Notations:
m : number of training examples
J : number of output labels
I : number of intervals
m_i : number of training examples in the interval i
m_{ij} : number of examples with output label j in the interval i
A SDM model is completely defined by the set of parameters $\{I, m_i(1 \leq i \leq I), m_{ij}(1 \leq i \leq I, 1 \leq j \leq J)\}$.

This definition is very general, and most discretization methods rely on SDM models. They first sort the samples according to the variable to discretize (property 1) and try to define a set of intervals by partitioning the string of output labels (property 2). The evaluation criterion is always based on the frequencies of output labels (property 3).

In the Bayesian approach, the best model is found by maximizing the probability $P(model/data)$ of the model given the data. Using the Bayes rule and since the probability $P(data)$ is constant under varying the model, this is equivalent to maximize $P(Model)P(Data/Model)$. We define below a prior which is essentially a uniform prior at each stage of the hierarchy of the SDM model parameters. We also introduce a strong hypothesis of independence of the distributions of the class values. This hypothesis is often assumed (at least implicitely) by many discretization methods, that try to merge similar intervals and separate intervals with significantly different distributions of class values.

Definition: The following distribution prior on SDM models is called the three-stage prior:

- the number of intervals I is uniformly distributed between 1 and m;
- for a given number of intervals I, every division of the string to discretize into I intervals is equiprobable;
- for a given interval, every distribution of output labels in the interval is equiprobable;
- the distributions of the output labels in each interval are independent from each other.

Theorem 1. *(Boullé, 2004b) A SDM model distributed according to the three-stage prior is Bayes optimal for a given set of training examples to discretize if the following criterion is minimal on the set of all SDM models:*

$$\log(m) + \log\binom{m+I-1}{I-1} + \sum_{i=1}^{I} \log\binom{m_i+J-1}{J-1} + \sum_{i=1}^{I} \log\left(\frac{m_i!}{m_{i,1}!\ldots m_{i,J}!}\right). \tag{25.1}$$

The first term of the criterion stands for the choice of the number of intervals, the second term for the choice of the bounds of the intervals and the third term for the choice of the output labels distribution in each interval. The last term encodes the probability of the data given the model.

Once the optimality of the evaluation criterion is established, the problem is to design an efficient minimization algorithm. The MODL method uses a greedy bottom-up merge algorithm to perform this optimization, that can be optimized in $O(m \log(m))$ time. This algorithm exploits the additivity of the MODL criterion, memorizes the variations $\Delta value$ of the criterion related to the merges, and keeps these merge evaluations in a maintained sorted list (such as an AVL binary search tree for example).

The method is fully described and experimented in (Boullé, 2004a,b). Compared to other discretization methods, the MODL method obtains better classification performances with fewer intervals. Random variables are discretized with a single interval since this is the most probable discretization model of such variables. The MODL method is thus efficient at detecting probes.

25.4 Results on the NIPS challenge

In this section, we report the results obtained by the ESNB method on the NIPS 2003 Challenge datasets (Guyon, 2003). Each dataset is divided in 3 sets: training, validation and test. In the first period, only the training sets could be used for training for submission of the *original* challenge entries. In the second period, the training and validation sets were used together for the *validation* entries. We submitted one original challenge entry using the ESNB method. Since the results looked promissing, we decided to submit two validation entries, one using the ESNB method and the other using the method as a filter for multi-layer perceptron (ESNB+NN). The ESNB method is fully automatic and does not require any parameter. In the filter approach, we use a non-linear multi-layer perceptron applied on the variables selected by the ESNB method. The multi-layer perceptron is trained using back-propagation with sigmoid activation function, regularized with orthogonal weight decay. The training set of the challenge is used to train and the validation set to stop training. This predictor is trained with a hidden layer containing 1, 5, 10, 25 or 50 neurons, and the best neural architecture is chosen based on the validation set: 50 neurons for the Madelon dataset and 1 neuron for the other datasets. We report in Table 25.1 the results of our ESNB entry by Dec. 1^{st} and in Table 25.2 the results of our ESNB+NN entry by Dec. 8^{th}.

The ESNB method has a low computation time, with on average 5 mn per dataset on a PC 1.7 Mhz. The MODL discretization methods is very efficient at detecting probes and the use of the area under the lift curve as a feature subset selection criterion helps removing redondant variables. This results in small numbers of selected features, on average 1% of the input variables. Compared to the Dec. 1^{st} original entry, the ESNB method is able to exploit the increased number of training examples available in the Dec. 8^{th} experiments. It selects more features (a total of 321 versus 252) while keeping less probes (1 versus 3) and improves the balanced error rate from 19.85% down to 18.25%. The ESNB+NN method largely improves the results of the ESNB method, especially when the bias of the Naïve Bayes approach is too limiting. Using the ESNB method as a filter approach is thus relevant. In

Table 25.1. Challenge results for the ESNB method (Dec. 1^{st}).

Dec. 1^{st}	ESNB challenge entry					The winning challenge entry					
Dataset	Score	BER	AUC	Feat	Probe	Score	BER	AUC	Feat	Probe	Test
Overall	-57.82	19.85	85.96	1.02	10.6	88.00	6.84	97.22	80.3	47.77	1
Arcene	-78.18	31.25	75.93	0.05	40	98.18	13.30	93.48	100.0	30.0	1
Dexter	-45.45	9.80	96.42	0.17	0	96.36	3.90	99.01	1.52	12.87	1
Dorothea	-45.45	21.03	89.43	0.05	1.89	98.18	8.54	95.92	100.0	50.0	1
Gisette	-56.36	3.12	99.49	3.02	0	98.18	1.37	98.63	18.26	0.0	1
Madelon	-63.64	34.06	68.51	1.8	11.11	100.00	7.17	96.95	1.6	0.0	1

Table 25.2. Challenge results for the (ESNB+NN) method (Dec. 8^{th}).

Dec. 8^{th}	ESNB+NN challenge entry					The winning challenge entry					
Dataset	Score	BER	AUC	Feat	Probe	Score	BER	AUC	Feat	Probe	Test
Overall	-28	12.42	93.12	1.04	1.43	71.43	6.48	97.20	80.3	47.77	1
Arcene	-60	22.92	83.78	0.14	7.14	94.29	11.86	95.47	10.7	1.03	1
Dexter	-25.71	7.20	97.49	0.33	0	100	3.30	96.70	18.57	42.14	1
Dorothea	54.29	14.59	91.50	0.07	0	97.14	8.61	95.92	100	50	1
Gisette	-42.86	2.46	99.64	3.26	0	97.14	1.35	98.71	18.32	0	1
Madelon	-65.71	14.94	93.22	1.4	0	94.29	7.11	96.95	1.6	0	1

order to evaluate the method both on the balanced error rate and the number of the selected variables, we report the method results and the entry results of all the other participants on a bi-criteria plan displayed in figure 25.1.

The results of the ESNB methods used as a filter approach are on the Pareto curve in figure 25.1. Many methods obtain a better error rate, up to twice better than that of the ESNB+NN method, but at the expense of a significantly higher number of selected variables.

25.5 Conclusion

The ESNB method is a feature selection method derived from the Naïve Bayes method enclosed in a wrapper approach. It benefits from the use of the Bayes optimal MODL discretization method and from the evaluation of predictor using the area under the lift curve instead of the accuracy. It can be exploited either directly or as a filter approach with a powerful predictor applied on the selected features.

Experiments on the NIPS 2003 datasets show that this fully automatic method is fast, robust, and exhibits results with a good trade-off between error rate and number of selected variables. However, the method suffers from several weaknesses partly related to the bias of the Naïve Bayes approach. In future work, we plan to improve

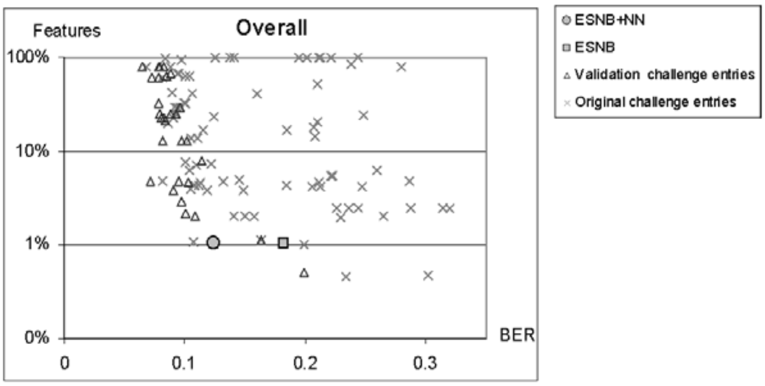

Fig. 25.1. Bi-criteria analysis of the challenge results with the balanced error rate on the x-coordinate and the number of selected variables on the y-coordinate

the error rate of the method, with the smallest possible decay in computation time and selected variable number. The wrapper method can be improved by replacing the forward selection algorithm by a more sophisticated search heuristic. Exploiting ensemble methods is a promising direction for selecting a more comprehensive set of selected variables. Although the univariate evaluation of variables is robust owing to the MODL discretization method, the overfitting behaviour resulting from the selection of a set of variables could be reduced by using regularization techniques. A last direction is to decrease the bias of the Naïve Bayes approach by detecting interactions between variables or building new features combining multiple input variables.

Acknowledgements

I am grateful to Vincent Lemaire (author of the ROBELON submission) for providing the multi-layer perceptron software and enabling the ESNB+NN joined submission to the NIPS Challenge. The MODL method is under French patent N 04 00179. Contact the author for conditions of use.

References

M. Boullé. A bayesian approach for supervized discretization. In *Proceedings of the Data Mining 2004 Conference*. WIT Press, 2004a.

M. Boullé. MODL : une méthode quasi-optimale de discrétisation supervisée. Technical Report, NT/FTR&D/84444, France Telecom R&D, Lannion, France, 2004b.

J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *Proceedings 12th International Conference on Machine Learning*, pages 194–202. Morgan Kaufmann, 1995.

I. Guyon. Design of experiments of the NIPS 2003 variable selection benchmark. http://www.nipsfsc.ecs.soton.ac.uk/papers/datasets.pdf, 2003.

D.F. Hand and K.S. Yu. Idiot's bayes-not so stupid after all? *International Statistical Review*, 69(3):385–398, 2001.

C.N. Hsu, H.J. Huang, and T.T. Wong. Implications of the dirichlet assumption for discretization of continuous variables in naïve bayesian classifiers. *Machine Learning*, 53(3):235–263, 2002.

R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 1997.

P. Langley and S. Sage. Induction of selective bayesian classifiers. In *Proceedings 10th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufman, 1994.

F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings Fifteenth International Conference on Machine Learning*. Morgan Kaufmann, 1998.

I.H. Witten and E. Franck. *Data Mining*. Morgan Kaufmann, 2000.

Y. Yang and G.I. Webb. A comparative study of discretization methods for naïve-bayes classifiers. In *Proceedings of the Pacific Rim Knowledge Acquisition Workshop*, 2002.