JMLR: Workshop and Conference Proceedings, **52**, 2016, 523–534 PGM 2016

On Construction of Hybrid Logistic Regression-Naïve Bayes Model for Classification

Yi Tan & Prakash P. Shenoy VIEIRATY0113@GMAIL.COM, PSHENOY@KU.EDU
University of Kansas School of Business
1654 Naismith Drive
Lawrence, KS 66045

Moses W. Chan & Paul M. Romberg {MOSES.W.CHAN, PAUL.M.ROMBERG}@LMCO.COM
Advanced Technology Center
Lockheed Martin Space Systems
Sunnyvale, CA 94089

Abstract

In recent years, several authors have described a hybrid discriminative-generative model for classification. In this paper we examine construction of such hybrid models from data where we use logistic regression (LR) as a discriminative component, and naïve Bayes (NB) as a generative component. First, we estimate a Markov blanket of the class variable to reduce the set of features. Next, we use a heuristic to partition the set of features in the Markov blanket into those that are assigned to the LR part, and those that are assigned to the NB part of the hybrid model. The heuristic is based on reducing the conditional dependence of the features in NB part of the hybrid model given the class variable. We implement our method on 21 different classification datasets, and we compare the prediction accuracy of hybrid models with those of pure LR and pure NB models.

Keywords: logistic regression, naïve Bayes, hybrid logistic regression-naïve Bayes

1. Introduction

Data classification is a very common task in machine learning and statistics. It is considered an instance of supervised learning. For a standard supervised learning problem, we implement a learning algorithm on a set of training examples, which are pairs consisting of an input object, a vector of explanatory variables $F_1, \ldots, F_n$, called *features*, and a desired output object, a response variable C, called *class* variable. The learning algorithm outputs a *classifier* that constitutes a function that maps a data item into one of several pre-defined classes (Fayyad et al., 1996). Given new feature values, the *classifier* can make a prediction of the values of the corresponding class variable. A classification task is to find a subset of features and a classifier based on data in a training set that maximizes the accuracy of prediction (using the classifier) of data in a test set.

Some classifiers, for example, logistic regression (LR) and naïve Bayes (NB), are much easier to understand and interpret than others (Langley et al., 1992). Because LR and NB models have few parameters, they scale well to high dimensions, and can be trained very efficiently. These two classifiers belong to the family of probabilistic classifiers.

LR is a robust classifier that estimates the probability of occurrence of the class variable directly. It can handle both numeric and categorical features. Also, it is a conditional model (for the class variable given the features) and not a full probability model. Given the observed values of all features, the portion of the model for the features is irrelevant for computing the posterior distribution

of the class variable. It assumes that the log odds of a class is a linear function of the features. A disadvantage of LR is that it cannot predict if some of the features have missing values.

A NB classifier is a full probability model, including the prior distribution of the class variable, that makes an assumption that all features are mutually conditionally independent given the class variable (Friedman et al., 1997). An advantage of NB over LR is that it is capable of using features that have missing data (Liu et al., 2005). While NB can handle numeric features with non-parametric distributions, the number of parameters increases with the number of bins used to discretize the state space of the numeric features, and the simplicity of NB is lost. A numeric feature discretized into 10 bins would result in $2 \cdot (10 - 1) = 18$ parameters for a binary class variable. On the other hand, LR classifiers can include numeric features without discretization using a single parameter per numeric feature.

An outline of the remainder of the paper is as follows. In Section 2, we describe related work on hybrid discriminative/generative classifiers, and state the contributions of our paper. In Section 3, we sketch the LR model. Section 4 describes the NB model. In Section 5, we describe the hybrid model. Section 6 describes our method for construction of a hybrid model. Section 7 shows the empirical results from our experiments using datasets from the UCI Machine Learning Repository. Finally, in Section 8, we summarize and conclude.

2. Related Literature

Ng and Jordan (2001) do an empirical and theoretical study of LR and NB models for classification. They find that a NB model has a lower asymptotic accuracy (as the number of training instances becomes large) compared to LR, but that a NB model approaches its asymptotic error much faster than a LR model. Thus, for large training instances, LR classifiers have higher accuracies, whereas for small training instances, NB classifiers may have higher accuracies than LR. They ignore the issue of missing values.

In recent years, several researchers have explored a hybrid model that combines discriminative and generative models in one model (Rubinstein and Hastie, 1997; Raina et al., 2003; Kang and Tian, 2006; Xue and Titterington, 2010; Fujino et al., 2007; Bishop and Lasserre, 2007).

Rubinstein and Hastie (1997) are among the earliest to suggest combining discriminative and generative models. They suggest that features that satisfy the assumption of a generative model be retained in the generative part, with the remaining moved to the discriminative part. They compare linear discriminant analysis (LDA), a generative model, with LR, a discriminative model, for three different simulated datasets, and discover that LDA does better than LR when the class densities were Gaussian, and vice-versa. They also compare NB, a generative model, with generalized additive model (GAM), a discriminative model, for a simulated dataset with log-spline density that satisfies the assumptions of the GAM model. The GAM model has a lower error rate than the NB model. However, when the training set is a small subset (25 observations) of the entire dataset, NB models had a lower average error than GAM. While they propose combining the two approaches, they do not describe any experimental results of a hybrid model.

Raina et al. (2003) investigate a hybrid model with LR as the discriminative component and NB as the generative component for text classification problems. They run experiments using pairs of newsgroups from a dataset of USENET new postings, where the documents have two disjoint regions–a subject region, and a message body region. A NB model treats the two regions in exactly the same way (due to the strong assumption of a NB model). A hybrid model treats the 2 regions

differently using a different weight parameter for the two sections. As the subject region has fewer words than the message body region, the words in the subject region are weighted higher than the words in the message body region. Depending on how the weight parameters are estimated from a dataset, the hybrid model reduces to a LR model. Experimental results show that hybrid models can provide lower test error than either pure LR or pure NB. As the size (number of words) of the training set increases the accuracy of the hybrid model is asymptotically equal to the accuracy of the LR model. For small training sets, the hybrid model does much better than either NB or LR.

Kang and Tian (2006) introduce a hybrid discriminative-generative classifier where the discriminative component is LR, and the generative component is NB or tree-augmented NB (TAN). They learn a hybrid model that includes all features by starting with an empty generative component, and then greedily add one feature at a time (that which results in the maximum accuracy gain) to the generative component until the classification accuracy does not improve. They test their algorithm for 20 different datasets, which are pre-processed so that there are no missing values and all features are categorical. They measure classification accuracy using either 10-fold cross-validation (for small datasets) or 3-fold cross-validation (for large datasets). This is done just once, so they get a point estimate of the classification accuracy. The average point estimate of the classification errors for all 20 datasets is lowest for the hybrid LR-NB model.

Xue and Titterington (2010) study hybrid discriminative-generative classifiers where the discriminative component is LR, and the generative component is Fisher's linear discriminant analysis (LDA), which assumes that the features have a joint multivariate normal distribution with a covariance matrix that does not depend on the class variable. They construct a hybrid model that consists of all features in a dataset. They test all features for univariate normality (using the Shapiro-Wilk test with 0.01 level of significance), and those that fail the test are assigned to the LR portion of the hybrid model. They test their algorithm for 6 datasets that have only numeric features, where they measure classification accuracy using a random split of a dataset into training and test set for six different sizes of the training set, which is repeated 200 times. They find that for smaller sizes of the training set, the hybrid model does better than the pure LR and pure LDA models.

Fujino et al. (2007) investigate hybrid discriminative-generative classifiers similar to Raina et al. (2003) for text classification having multiple components (such as titles, hyperlinks, anchor text, images, etc.). They use a generative classifier for each component that are then combined using weights learnt using a maximum entropy principle. They do an empirical evaluation on four text-classification datasets, and find that hybrid classifiers outperform NB and LR models.

Bishop and Lasserre (2007) investigate hybrid models that are a convex combination of generative and discriminative models. They apply their approach to object recognition in static images. Each image has two sets of features—observable features, and latent patch labels—in addition to class. Their training set consists of a total of 50 images in each class with 8 classes (for a total of 400 images), with 45 images in each class being unlabelled (missing values of patch labels). The test set consists of 800 images (100 images in each class) with no missing values. They compare the performance of hybrid models with different combination weights and find that the best performance is obtained with a blend between generative and discriminative extremes.

The main contribution of this paper is to continue the investigations of hybrid models as first suggested by Rubinstein and Hastie (1997). While Xue and Titterington (2010) do the same, they investigate hybrid models with linear discriminant analysis as the generative component for datasets with numeric features. Our focus is on datasets with NB as the generative component for datasets with a mix of numeric and categorical features. Kang and Tian (2006) also investigate hybrid models

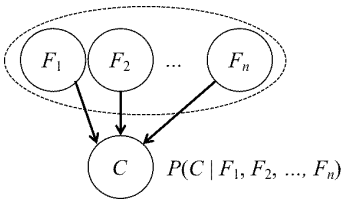

Figure 1: A LR Model as a Bayesian Network

with NB and TAN as generative components, but their heuristic for finding a hybrid model is based on improvement of accuracies. Also, they provide only a point estimate of accuracy of hybrid and other models based on cross-validation. Our strategy for constructing a hybrid model is based on reducing the conditional dependence of features in the NB part of the hybrid model. Also, unlike earlier studies, we do not use all features. Instead, we first estimate a Markov blanket of the class variable to identify relevant features, and then use only these relevant features for constructing a hybrid model. Finally, we find a distribution of the accuracy of the hybrid models (and not just a point estimate) by repeating the experiments many times with different random seeds.

3. Logistic Regression

In this section, we discuss LR as a method for classification. Suppose C is the categorical class variable, and $F_1, \ldots, F_n$ are real-valued features used to predict C. The features can be numeric or Boolean (with values 0 or 1). If we have a categorical feature with k distinct values, then we can represent such a feature with $k-1$ Boolean features.

The LR model assumes that the log odds for a class c_j is a linear function of the features:

$$\ln\left(\frac{P(C = c_j \mid \mathbf{f})}{1 - P(C = c_j \mid \mathbf{f})}\right) = \beta_{0\,j} + \sum_{i=1}^{n} \beta_{i\,j} f_i \tag{1}$$

Notice that for a dataset with a binary class variable (with 2 classes), and n real-valued features, we have $n+1$ parameters. If we have a class variable with c classes, then we have $(c-1)\cdot(n+1)$ parameters. The small number of parameters is one reason for the simplicity and robustness of the LR classifier. Using Eq. 1, we can compute the probability distribution of all classes in C. The predicted class is the one with the highest probability.

Rijmen (2008) has modeled a LR model as a Bayesian network, where Eq. (1) constitutes the conditional probability distribution for C given $\mathbf{F} = (F_1, \ldots F_n)$. LR assumes a parametric form for the distribution $P(C|\mathbf{F})$, and has its model structure as shown in Fig. (1). In this figure, the dotted oval around the features denotes that the Bayesian network structure of the feature variables is not represented, as it is irrelevant to C, assuming that we have observed values of all features.

4. Naïve Bayes

In this section we discuss NB model as a method for classification. NB (Hand and Yu, 2001) is a probabilistic classifier that is based on Bayes rule. It makes an assumption that features are mutually conditionally independent given the class variable. This assumption reduces the complexity of the model (number of parameters), which makes it a robust model. Suppose C is the binary class

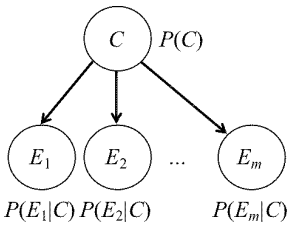

Figure 2: A NB Model as a Bayesian Network

variable, whose value we wish to predict based on observation of a subset of m features $\mathbf{E} = (E_1, \ldots, E_m)$. The features can be numeric or categorical. Fig. 2 is a Bayesian network depiction of a NB classifier. Assuming C, E_1, ..., E_m are all binary, the number of parameters is $2\,m + 1$.

Based on the assumption of a NB model, it can be shown that

$$odds(C = 1 \mid e_1, \ldots, e_m) = odds(C = 1) \prod_{i=1}^{m} lr(e_i, C = 1), \tag{2}$$

where $lr(e_i, C = 1) = \frac{P(e_i|C=1)}{P(e_i|C=0)}$, the likelihood ratio for $C = 1$ based on the observed value $E_i = e_i$. In words, the posterior odds of $C = 1$ is equal to prior odds of $C = 1$ times the likelihood ratios of observed features for $C = 1$. If a feature is not observed, we can regard its likelihood ratio as equal to 1.

Missing data are common in real-life datasets. Depending on the amount of missing data, they may significantly affect the efficiency and accuracy of classifiers. Simple methods for handling missing data are either to delete the instances with missing data, or to impute the missing data using expectation-maximization algorithms (Dempster et al., 1977) assuming that the data are missing at random. Liu et al. (2005) analyze the impact of missing data to classification algorithms and show that NB classifier is the least sensitive to missing data among six classifiers it considers.

5. A Hybrid LR-NB Model

In this section, we discuss a hybrid LR-NB model (hybrid, in short) as a method for classification. The graphical structure of the hybrid model represented as Bayesian network is shown in Fig. 3. Node C is the class variable, whose value we need to predict based on observation of two subsets of features: $\{F_1, \ldots, F_n\}$, the parents of C in Fig. 3, called the LR part, and $\{E_1, \ldots, E_m\}$, the children of C, called the NB part.

The conditional independence assumptions of a hybrid model are as follows. First, the features in the LR part of the model are conditionally independent of the features in the NB part of the model given the class variable C. Second, the features in the NB part of the model are mutually conditionally independent given C.

One implication of the first conditional independence assumption is that to learn the parameters of the conditional distribution of C given the features in the LR part, the features in the NB part are irrelevant for this task. Thus, one can use standard LR parameter estimation methods to learn these parameters. Similarly, to learn the parameters of the NB part of the hybrid model, the features in the LR part are irrelevant for this task, and thus, we can use standard NB parameter estimation methods for learning these parameters.

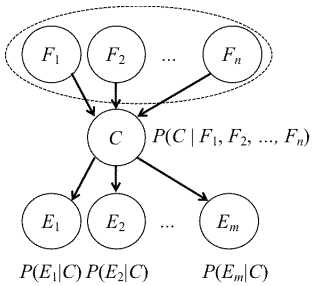

Figure 3: A Hybrid LR-NB Model as a Bayesian Network

Making inferences in a hybrid model is easy. For simplicity of exposition, we assume that C is a Boolean variable. Using variable elimination, after we eliminate the observed features in the LR part, the posterior distribution of the class variable C is given by the LR model:

$$\ln(odds(C = 1|\mathbf{f})) = \beta_0 + \sum_{i=1}^{n} \beta_i f_i \tag{3}$$

This gives the posterior odds of $C = 1$ given $\mathbf{f} = (f_1, \ldots, f_n)$ as:

$$odds(C = 1|\mathbf{f}) = \exp(\beta_0 + \sum_{i=1}^{n} \beta_i f_i) \tag{4}$$

After elimination of the features $\mathbf{F}$ in the LR part, what's left is a NB model such that the posterior distribution of C (given $\mathbf{F} = \mathbf{f}$) is as given in Eq. (4). Thus, we can now compute the posterior distribution of C given $\mathbf{F} = \mathbf{f}$ and $\mathbf{E} = \mathbf{e}$ using the NB model as follows:

$$odds(C = 1|\mathbf{e}, \mathbf{f}) = \exp(\beta_0 + \sum_{i=1}^{n} \beta_i f_i) \prod_{j=1}^{m} lr(e_j, C = 1) \tag{5}$$

As mentioned earlier, if we are missing a value, say, of E_j, then we can assume that the corresponding likelihood ratio $lr(e_j, C = 1)$ is equal to 1. Eq. (5) is the equation for making inferences from a hybrid classifier, which estimates the probability that the binary-valued class variable C will take the value of either 1 or 0 given the observed values of all features. The general hybrid model where C is not binary-valued is a simple extension of binary hybrid model, which combines multinomial LR and NB models. Hosmer and Lemeshow (2004) provide detailed discussions of multinomial LR.

A hybrid model has many other advantages beyond relaxing the conditional independence assumptions of a NB model. A hybrid model inherits the best features from both LR and NB models. First, the hybrid model retains the simplicity of LR and NB models. If C is binary-valued, and if all the features in the NB part of the model are binary-valued, then the number of parameters in the hybrid model is $n + 1 + 2m$, where n is the number of features in the LR part of the model, and m is the number of features in the NB part of the model. Second, if we have features with missing values, we can restrict such features to the NB part as a NB model is able to easily handle missing values. Third, if we have numeric features with non-parametric distributions, we can restrict such features to the LR part, as LR can easily handle numeric features using a single parameter

without discretization. In a NB model, a numeric feature discretized into 10 bins would result in $2 \cdot (10 - 1) = 18$ parameters assuming a binary class variable.

6. A Method for Constructing a Hybrid LR-NB Model

The main focus of this paper is on construction of a good hybrid model that predicts well. We ignore the issues of continuous vs. categorical features, and missing values of features. Instead, we focus on the assumptions of the NB model and try to find a hybrid model that satisfies these assumptions as much as possible. This is the strategy suggested in Rubinstein and Hastie (1997), but has not been implemented previously for the case of hybrid models with LR as the discriminative component, and NB as the generative component.

As there are 3^p possible hybrid model structures where p is the number of features, searching the space of all possible hybrid models is computationally intractable for large values of p. To narrow the search, we first select the initial set of features based on an estimated Markov boundary of the class variable C. We used 4 constraint-based methods: (i) grow-shrink (*gs*) (Margaritis, 2003); (ii) incremental association (*iamb*) (Tsamardinos et al., 2003); (iii) fast incremental association (*fast.iamb*) (Yaramakala and Margaritis, 2005); and (iv) interleaved incremental association (*inter.iamb*) (Tsamardinos et al., 2003); with 3 different conditional independence tests (asymptotic chi-square (*mi*), asymptotic chi-square with adjusted degrees of freedom (*mi-adf*), and shrinkage estimator for the mutual information (*mi-sh*)) for categorical features, and with 4 different conditional independence tests (exact Student's t-test (*cor*), Fisher's Z-test (*zf*), asymptotic chi-square test (*mi-g*), and sequential Monte Carlo permutation test (*mi-g-sh*)) for numeric features, with significance level of 0.05, all of which are implemented in an R package called *bnlearn* (Scutari and Denis, 2014). For a dataset with some numeric features and some categorical features, we discretize the numeric features so that all features are discrete, and the discretization of numeric features is done only for the purpose of estimating a Markov boundary of C. Notice that a Markov blanket of C can be learnt directly from a dataset without first estimating a graphical model. We estimate the Markov boundary of C as the union of all the estimated Markov boundaries. The logic here is that all methods unanimously agree that the features not in the estimated Markov boundary are irrelevant to C given the features in the estimated Markov boundary.

Using only features in the estimated Markov boundary, we select the LR and NB parts of the hybrid model using a heuristic that reduces the conditional dependence between the features in the NB part given class variable C. The heuristic is as follows. We start with all features in the estimated Markov boundary in the NB part, and none in the LR part. We compute the normalized conditional mutual information ($norMI$) given class variable C (Strehl and Ghosh, 2002) for each pair of features in the NB part, and select a pair of features with the highest $norMI$. For each of these two features (in the selected pair), we compare their second highest $norMI$ and remove the one with higher value from the NB part and include it in the LR part. We continue with this process until the $norMI$ for each pair of features in the NB part is less than 0.05, or there is only one feature left. When there is only one pair of features in the NB part that has $norMI$ more than 0.05, we select one at random to move to the LR part.

The cutoff value of $norMI = 0.05$ is ad-hoc. Lowering this cutoff value leads to more features in the LR part, and raising this value leads to more features in the NB part of the hybrid model. One alternative is to use a cutoff value that is statistically significant to some level as suggested by a G-test (McDonald, 2014a) or a chi-square test (McDonald, 2014b) for independence.

Algorithm 1 outlines the procedure that selects which features are in the LR part and which are in the NB part.

Algorithm 1 Find structure of a hybrid model

input: A set of labelled instances.
output: A hybrid network structure with identified *LR-part* and *NB-part*.

1: Find the Markov boundary of class variable C using all 4 constraint-based method with all different conditional independence tests.
2: Set *NB-part* $= \cup MB(C)$ and *LR-part* $= \emptyset$
3: do while |*NB-part*| > 2
4: Let (X_i, X_j) denote the pair of features that maximizes $norMI(X_i; X_j \mid C)$ where $X_i, X_j \in$ *NB-part*
5: If $max\, norMI \geq 0.05$ then
6: Let $X_s \in$ *NB-part* $\setminus \{X_j\}$ be the attribute that maximizes $norMI(X_i; X_s \mid C)$
7: Let $X_t \in$ *NB-part* $\setminus \{X_i\}$ be the attribute that maximizes $norMI(X_j; X_t \mid C)$
8: If $norMI(X_i; X_s \mid C) < norMI(X_j; X_t|C)$, then *NB-part* $=$ *NB-part* $\setminus \{X_j\}$,
9: *LR-part* $=$ *LR-part* $\cup \{X_j\}$
10: If $norMI(X_i; X_s \mid C) > norMI(X_j; X_t \mid C)$, then *NB-part* $=$ *NB-part* $\setminus \{X_i\}$,
11: *LR-part* $=$ *LR-part* $\cup \{X_i\}$
12: else end do;
13: If |*NB-part*| $= 2$ and $norMI(X_i; X_j \mid C) \geq 0.05$, then
14: pick one at random and add to *LR-part*.
15: else end algorithm

7. Experimental Analysis

To evaluate the performance of hybrid models, we conducted experiments on 21 different machine learning datasets from the UCI Machine Learning Repository. A summary of these datasets is given in Table 1.

In this section, we will describe the experimental setup and experimental results from our method of construction of a hybrid model, and compare the prediction accuracies of the hybrid model with pure LR and pure NB models using only features in the estimated Markov blanket.

7.1 Experimental Setup

Our goal is to evaluate the performance of hybrid classifier in terms of out-of-sample prediction accuracy. We randomly divided each datasets into two parts, a training set with about 90% of the instances, and a testing set with the remaining 10% of the instances. For datasets that have missing values of features, instances with missing values were restricted to the training set. Using the training set, we implemented the algorithm described in Section 6 to identify the model structure, and trained the corresponding hybrid model. We compare the hybrid model with pure LR model and pure NB model by their prediction accuracies (percent of instances in the test set correctly classified).

To estimate the parameters of a NB model, we discretize the numeric features using an entropy-based method (MDL method), proposed by Fayyad and Irani (1993). We carried out the discretiza-

Dataset	*# Features*	*# Numeric*	*# Categorical*	*# Instances*	*# Classes*	*Missing Values?*
Pima Indians Diabetes	8	8	0	768	2	yes
Adult Census Income	14	6	8	48,842	2	yes
Credit Approval	15	6	9	690	2	yes
Glass Identification	10	10	0	214	6	no
Hypothyroid	19	7	12	3,163	2	yes
Statlog Vehicle Silhouettes	18	18	0	846	4	no
Wine	13	13	0	178	3	no
Bank Marketing	19	9	10	41,188	2	yes
Banknote Authentication	4	4	0	1,372	2	no
Car Evaluation	6	0	6	1,728	4	no
Chronic Kidney Disease	24	11	13	400	2	yes
Blogger	5	0	5	100	2	no
Breast Tissue	9	9	0	106	6	no
Congressional Voting Records	16	0	16	435	2	yes
Connectionist Bench	60	60	0	208	2	no
Default of Credit Card Clients	23	14	9	30,000	2	no
Ecoli	7	5	2	336	8	no
Mushroom	22	0	22	8,124	2	yes
Nursery	8	0	8	12,960	3	no
Qualitative Bankruptcy	6	0	6	250	2	no
EEG Eye State	14	14	0	14,980	2	no

Table 1: A Summary of 21 Bench-Mark Datasets

tion procedure using a filter in WEKA. Also, when we learn the parameters of the LR part of the hybrid model, and the parameters of the pure LR model, using the training set, we disregard instances with missing values.

We repeated the entire procedure (division of dataset, estimation of Markov blanket, identification of model structure, and estimation of model parameters using the training set, and computation of prediction accuracies using the test set) 100 times. For six datasets (Glass Identification, Breast Tissue, Wine, Chronic Kidney Disease, Blogger, and Connectionist Bench) we repeated the procedure 500 times (as the standard errors of accuracy were too large from 100 repetitions). In Table 2, we report the mean results for each datasets, in terms of number of features in Markov blanket, number of features in the LR part of the hybrid model, number of features in NB part of the hybrid model, and the prediction accuracies of all three classifiers (hybrid, LR, and NB) with their standard errors in units of %.

7.2 Experimental Results

Table 2 shows the results from our experiment regarding the estimated Markov blanket, and the average structure of the hybrid model found using our conditional independence based heuristic. Notice that we are guaranteed to get at least one variable in the NB part of the hybrid model. On an average for the 21 datasets, the Markov boundary consists of 62% of the features in the dataset. Also, on an average for the 21 datasets, the hybrid model consists of 52% of the features in the Markov blanket in the LR part, with the remaining 48% in the NB part.

Table 2 also describes the mean accuracies of hybrid, pure LR, and pure NB, models (with standard errors) from 100 repetitions of our procedure (500 for six datasets). Using a paired t-test with 0.05 significance level, the hybrid model has the highest accuracy or tied for highest for 6 of 21 datasets. The pure LR model has the highest accuracies or tied for highest for 15 of 21 datasets. The pure NB model has the highest or tied for highest for 7 of 21 datasets.

Dataset	*# Features*	*# MB*	*# LR-part*	*# NB-part*	*Acc. Hybrid*	*Acc. LR*	*Acc. NB*
Pima Indians Diabetes	8	4.47	0.89	3.58	**80.80** (0.45)	77.95 (0.43)	**80.89** (0.46)
Adult Census Income	14	9.87	3.14	6.73	82.12 (0.06)	**84.68** (0.06)	80.73 (0.08)
Credit Approval	15	8.62	3.07	5.55	**85.76** (0.49)	**85.62** (0.49)	**85.18** (0.47)
Glass Identication	10	5.47	3.49	1.98	**66.49** (0.42)	62.36 (0.39)	63.14 (0.60)
Hypothyroid	18	4.46	1.30	3.16	98.24 (0.08)	97.85 (0.07)	**98.59** (0.07)
Statlog Vehicle Silhouettes	18	17.15	14.48	2.67	78.74 (0.49)	**80.35** (0.40)	65.15 (0.49)
Wine	13	9.82	5.84	3.98	96.52 (0.21)	95.03 (0.22)	**97.90** (0.26)
Bank Marketing	19	12.92	7.91	5.01	88.24 (0.05)	**88.84** (0.04)	83.57 (0.07)
Banknote Authentication	4	3.00	2.00	1.00	96.43 (0.20)	**98.92** (0.08)	92.79 (0.19)
Car Evaluation	6	5.00	0.99	4.01	85.79 (0.30)	**92.52** (0.20)	85.79 (0.30)
Chronic Kidney Disease	24	10.29	6.08	4.21	97.26 (0.36)	**98.80** (0.10)	93.16 (0.69)
Blogger	5	2.04	1.04	1.00	68.24 (0.67)	**69.64** (0.57)	66.48 (0.73)
Breast Tissue	9	7.09	5.42	1.67	66.90 (0.64)	**67.80** (0.64)	66.00 (0.68)
Congressional Voting Records	16	7.39	4.58	2.81	94.06 (0.30)	**95.24** (0.30)	92.76 (0.30)
Connectionist Bench	60	7.48	3.61	3.88	71.48 (0.44)	68.65 (0.45)	**72.36** (0.44)
Default of Credit Card Clients	23	6.93	4.30	2.63	81.72 (0.06)	**82.07** (0.06)	80.50 (0.06)
Ecoli	7	6.00	3.79	2.21	**85.57** (0.57)	**85.77** (0.53)	83.73 (0.66)
Mushroom	22	12.53	10.60	1.93	99.98 (0.009)	**99.99** (0.003)	92.96 (0.099)
Nursery	8	8.00	1.00	7.00	90.29 (0.09)	**92.45** (0.07)	90.29 (0.09)
Qualitative Bankruptcy	6	4.54	1.16	3.38	**99.64** (0.10)	**99.56** (0.10)	**99.64** (0.10)
EEG Eye State	14	13.71	12.18	1.53	66.07 (0.30)	64.03 (0.20)	**68.70** (0.20)

Table 2: Summary of Results: Average Est. Markov Blanket Size, Average Structure of Hybrid Models, and Average Accuracies of Models, in units of % (SE in parenthesis). Highest accuracies are in boldface.

In a pairwise comparison between hybrid and pure LR models for the 21 datasets (using a paired t-test with 0.05 significance level), hybrid models outperform LR for 6 datasets, are tied with LR for 3 datasets, and do worse than LR for 12 datasets. This is consistent with the findings of Ng and Jordan (2001), who find that LR outperforms NB asymptotically. Thus, in very large datasets relative to number of features (such as, e.g., Adult Census Income, Bank Marketing, Default of Credit Card Clients), we would expect LR to predict better than NB and hybrid models.

In a pairwise comparison between hybrid and pure NB models for the 21 datasets, hybrid models outperform NB for 12 datasets, are tied with NB for 5 datasets, and do worse than NB for 4 datasets. This result may be the consequence of our heuristic in constructing hybrid models where we try to reduce the conditional dependence between features in the NB part.

8. Summary and Conclusions

In this paper, we discuss construction of a hybrid LR-NB model from data. First, we restrict features to the estimated Markov blanket of the class variable. Next, we use a heuristic that minimizes the conditional dependence among features in the NB-part of the hybrid model. We test our method on 21 datasets with wide diversity in number of features, number of instances, having a mix of categorical and numeric features, and missing/non-missing values of features. We repeat our method 100 times (500 times for six datasets), and describe statistics on the structure of the hybrid models, and the accuracies of the hybrid, pure LR, and pure NB models.

In a pairwise comparison (using a paired t-test with 0.05 significance level), hybrid models do better than LR for 6 datasets, are tied with LR for 3, and do worse than LR for 12. In a similar

comparison to NB, hybrid models do better than NB for 12 datasets, are tied with NB for 6, and do worse than NB for 3.

The LR model assumes that the log odds of a class is a linear function of the features. If there is an easy way to test this assumption, then we could develop heuristics that move features that violate this assumption from the LR part to the NB part. This may result in a hybrid model that does better than pure LR.

As discussed in Section 5, the hybrid model has several advantages over pure LR and pure NB models, such as ability to use numeric and categorical features, and missing values of features. These issues are not investigated in this paper. We are currently in the process of conducting systematic experiments with datasets that have these characteristics.

Acknowledgments

This project was funded in part by a sponsored grant from Lockheed Martin Space Systems, Sunnyvale, CA, to the University of Kansas. Thanks to Suzanna Emelio for proofreading the paper. Thanks also to three anonymous reviewers of PGM-16 for their constructive comments.

References

C. M. Bishop and J. Lasserre. Generative or discriminative? Getting the best of both worlds. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 8*, pages 3–24. Oxford Univ. Press, 2007.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-1993)*, pages 1022–1027. Morgan Kaufmann, 1993.

U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29 (2–3):131–163, 1997.

A. Fujino, N. Ueda, and K. Saito. A hybrid generative/discriminative approach to text classification with additional information. *Information Processing and Management*, 43:379–392, 2007.

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

D. W. Hosmer, Jr. and S. Lemeshow. *Applied Logistic Regression*. Wiley, 2004.

C. Kang and J. Tian. A hybrid generative/discriminative Bayesian classifier. In *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference (FLAIRS-2006)*, pages 562–567, 2006.

P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence (AAAI-92)*, pages 223–228, Menlo Park, CA, 1992. AAAI Press.

P. Liu, L. Lei, and N. Wu. A quantitative study of the effect of missing data in classifiers. In *IEEE 5th International Conference on Computers and Information Technology (CIT-2005)*, pages 28–33, 2005.

D. Margaritis. *Learning Bayesian network model structure from data*. PhD thesis, Carnegie-Mellon University School of Computer Science, Pittsburg, PA, 2003.

J. H. McDonald. G-test of goodness-of-fit. In *Handbook of Biological Statistics*, pages 53–58. Sparky House Publishing, Baltimore, MD, 3rd edition, 2014a.

J. H. McDonald. Chi-square test of goodness-of-fit. In *Handbook of Biological Statistics*, pages 45–52. Sparky House Publishing, Baltimore, MD, 3rd edition, 2014b.

A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 841–848, 2001.

R. Raina, Y. Shen, A. Y. Ng, and A. McCallum. Classification with hybrid generative/discriminative models. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 545–552, 2003.

F. Rijmen. Bayesian networks with a logistic regression model for the conditional probabilities. *International Journal of Approximate Reasoning*, 48(2):659–666, 2008.

Y. Rubinstein and T. Hastie. Discriminative vs. informative learning. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pages 49–53. AAAI Press, 1997.

M. Scutari and J.-B. Denis. *Bayesian Networks with Examples in R*. Texts in Statistical Science. Chapman & Hall/CRC Press, 2014.

A. Strehl and J. Ghosh. Cluster ensembles – A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

I. Tsamardinos, C. F. Aliferis, and A. R. Statnikov. Algorithms for large scale Markov blanket discovery. In I. Russell and S. Haller, editors, *Proceedings of the 6th International Florida Artificial Intelligence Research Society Conference (FLAIRS-2003)*, pages 376–380, 2003.

J.-E. Xue and D. M. Titterington. Joint discriminative-generative modelling based on statistical tests for classification. *Pattern Recognition Letters*, 31(9):1048–1055, 2010.

S. Yaramakala and D. Margaritis. Speculative Markov blanket discovery for optimal feature selection. In J. Han, B. W. Wah, V. Raghavan, X. Wu, and R. Rastogi, editors, *Proceedings of the 5th IEEE International Conference on Data Mining*, pages 809–812, 2005.