Using a Hierarchical Bayesian Model to Handle High Cardinality Attributes with Relevant Interactions in a Classification Problem *

Jorge Jambeiro Filho
Secretaria da Receita Federal
Alfndega do Aeroporto de Viracopos
Rodovia Santos Dummont, Km 66
Campinas-SP, Brazil, CEP 13055-900
jorge.filho@jambeiro.com.br

Jacques Wainer
Instituto de Computação
Universidade Estadual de Campinas
Caixa Postal 6176
Campinas - SP, Brazil, CEP 13083-970
wainer@ic.unicamp.br

Abstract

We employed a multilevel hierarchical Bayesian model in the task of exploiting relevant interactions among high cardinality attributes in a classification problem without overfitting. With this model, we calculate posterior class probabilities for a pattern W combining the observations of W in the training set with prior class probabilities that are obtained recursively from the observations of patterns that are strictly more generic than W. The model achieved performance improvements over standard Bayesian network methods like Naive Bayes and Tree Augmented Naive Bayes, over Bayesian Networks where traditional conditional probability tables were substituted by Noisy-or gates, Default Tables, Decision Trees and Decision Graphs, and over Bayesian Networks constructed after a cardinality reduction preprocessing phase using the Agglomerative Information Bottleneck method.

1 Introduction

In most countries, imported goods must be declared by the importer to belong to one of large set of classes (customs codes). It is important that each good is correctly classified, because each of the customs codes mean not only different customs duties but also different administrative, sanitary, and safety requirements. The goal of this project is to develop a tool that, considering four attributes: *declared custom code* (DCC), *importer* (IMP), *country of production* (CP) and *entry point in the receiving country* (EPR), will estimate, for each new example, the probability that it involves a misclassification. Such estimates will be used latter by a larger system that allocates human resources for different types of anti-fraud operations.

Our data set has 47826 examples of correct classification (which we will call negative examples) and 590 examples of misclassification (positive examples). In this dataset, the first attribute has 3826 distinct values, the second, 1991 values, the third, 101 values, and the fourth 52 values.

*This work is part of the HARPIA project and is supported by Brazil's Federal Revenue

With only 1.2% of positive examples, dataset is imbalanced what is usually handled with different resampling strategies [Chawla *et al.*, 2002]. However, resampling requires retraining the classifiers for each different assignment of costs for *false positives* and *false negatives*. In our context, such costs are not known in advance (priorities changes acording to other anti-fraud demands) and they may vary from example to example (not all false negatives cost the same). Thus we cannot train the classifiers for all possible cost assignments in advance.

On the other hand, if we can produce reliable probability estimates directly from the original dataset the work of the human resource allocation system becomes much easier. It can, for example, at any time, define a selection rate SR that matches the available human resources for the specific task of detecting wrong customs codes considering all other anti-fraud demands at the moment. The examples to be verified will naturally be the SR examples that are most likely to involve a misclassification. The allocation system may also combine the probability estimates with costs that may vary from to example to example without any retraining. It becomes also unnecessary that customs administration discuss their cost criteria with us. Thus we decided to concentrate on Bayesian techniques and not to use resampling or any other technique that requires retraining when costs change.

Domain specialists claim that there are combinations of attributes values (some involving all of them) that make the probability of an instance being positive significantly higher then it could be expected looking at each value separately. They call such combinations *critical patterns*. To benefit from critical patterns we would like to use a Bayesian Network (BN)[Pearl, 1988] where all attribute nodes are parents of the class node. We call such structure the *Direct BN Structure*.

In a BN, considering that x_{ji} is a possible value for node X_j and π_{jk} is a complete combination of values for Π_j, the set of parents of node X_j, the vector, θ_{jk}, such that $\theta_{jki} = P(x_{ji}|\pi_{jk})$, contained in the CPT of a node X_j, is assessed from the frequencies of the values of X_j among the training instances where $\Pi_j = \pi_{jk}$. The distributions of X_j given any two different combinations of values for the parents of X_j are assumed to be independent and a Dirichlet prior probability distribution for θ_{jk} is usually adopted. Applying Bayes rule

and integrating over all possible values for θ_{jk} it is found that:

$$E(\theta_{jki}) = P(x_{ji}|\pi_{jk}) = \frac{N_{jki} + \alpha_{jki}}{N_{jk} + \alpha_{jk}} \quad (1)$$

where N_{jki} is the number of simultaneous observations of x_{ji} and π_{jk} in the training set, $N_{jk} = \sum_{\forall i} N_{jki}$, α_{jki} is the value of one of the parameters of the Dirichlet prior probability distribution and $\alpha_{jk} = \sum_{\forall i} \alpha_{jki}$, the equivalent sample size of the prior probability distribution.

The Dirichlet prior probability distribution is usually assumed to be noninformative, what yields to:

$$P(x_{ji}|\pi_{jk}) = \frac{N_{jki} + \lambda}{N_{jk} + \lambda M_j} \quad (2)$$

where all parameters of Dirichlet distribution are equal to a small smoothing constant λ, and M_j is the number of possible values for node X_j. We call this *Direct Estimation* (DE).

In the *Direct BN Structure* the node whose CPT is to be estimated is the class node and all other attribute nodes are its parents. The Conditional Probability Table (CPT) of the class node in such a structure contains more than 40×10^9 parameters. It is clear that for rarely seen combinations of attributes the choice of the structure in the *Direct BN Structure* and equation 2 tends to produce unreliable probabilities whose calculation is dominated by the noninformative prior probability distribution. This suggests that using the *Direct BN Structure* and traditional CPTs we will have overfitting problems.

Instead of the *Direct BN Structure*, we can choose a network structure that does not lead to too large tables. This can be achieved limiting the number of parents for a network node. Naive Bayes [Duda and Hart, 1973] is an extreme example where the maximum number of parents is limited to one (the class node is the only parent of any other node). Tree augmented Naive Bayes (TAN) [Friedman *et al.*, 1997] adds a tree to the structure of Naives Bayes connecting the non-class attributes, and thus limits the maximum number of parent nodes to two. However, limiting the maximum number of parents also limits our ability to capture interactions among attributes and benefit from critical patterns. Thus, we would prefer not to do it.

Since the high cardinality of our attributes is creating trouble, it is a reasonable idea to preprocess the data, reducing the cardinality of the attributes. We can use, for example, the Agglomerative Information Bottleneck (AIBN) method [Slonim and Tishby, 1999] in this task. However, the process of reducing the cardinality of one attribute is blind in respect to the others (except to the class attribute), and thus it is unlikely that cardinality reduction will result in any significant improvement in the ability to capture critical patterns, which always depend on more than one attribute.

When the number of probabilities to be estimated is too large when compared to the size of the training set and we cannot fill the traditional conditional probability tables (CPTs) satisfactorily and [Pearl, 1988] recommends the adoption of a model that resorts to causal independence assumptions like the Noisy-Or gate. Using a Noisy-Or the number of parameters required to represent the conditional probability distribution (CPD) of a node given its parents, instead of being proportional to the product of the cardinality of all parents attributes, becomes proportional to the sum of their cardinality. However, causal independence assumptions are incompatible with our goal of capturing critical patterns.

It is possible to use more flexible representations for the conditional probability distributions of a node given its parents, like Default Tables (DFs) [Friedman and Goldszmidt, 1996], Decision Trees (DTs) [Friedman and Goldszmidt, 1996] and Decision Graphs (DGs) [Chickering *et al.*, 1997]. According to [Friedman and Goldszmidt, 1996], using such representations together with adequate learning procedures induces models that better emulate the real complexity of the interactions present in the data and the resulting network structures tend to be more complex (in terms of arcs) but require fewer parameters. Fewer parameter may result in smaller overfitting problems. On the other hand, using traditional CPTs, we assume that the probability distributions for a node given any two combinations of values for the parents are independent. If some of these distribution are actually identical, DTs, DFs and DGs, can reflect it and represent the CPD using a variable number of parameters that is only proportional to the number of actually different distributions.

Using DTs, DFs or DGs to represent the conditional distribution of a node given its parents, we assume that the probability distribution of the node given two different combinations of values for the parents may be either identical or completely independent. It is possible that neither of the two assumptions hold.

In [Gelman *et al.*, 2003] it is asserted that modeling hierarchical data nonhierarchically leads to poor results. With few parameters nonhierarchical models cannot fit the data accurately. With many parameters they fit the existing data well but lead to inferior predictions for new data. In other words they overfit. In contrast hierarchical models can fit the data well without overfitting. They can reflect similarities among distributions without assuming equality.

Observing an slight modification in equation 2 used in [Friedman *et al.*, 1997] in the definition of a smoothing schema for TAN we can see that the data that is used to estimate the CPT of any node that has at least one parent is hierarchical:

$$P(x_{ji}|\pi_{jk}) = \frac{N_{jki} + S \cdot P(x_{ji})}{N_{jk} + S} \quad (3)$$

where S is a constant that defines the equivalent sample size of the prior probability distribution. We call this *Almost Direct Estimation* (ADE). ADE uses the probability distribution assessed in a wider population to build an informative prior probability distribution for a narrower population and so it has a hierarchical nature. Such approach was also used, for example, in [Cestnik, 1990]. ADE is the consequence of instead of a noninformative Dirichlet prior probability distribution, adopting a Dirichlet prior probability Distribution where $\alpha_{jki} \propto P(x_{ji})$.

ADE gets closer to the true probability distribution, but its discrimination power is not significantly better than DE. It is a linear combination of two factors N_{jki}/N_{jk} and $P(x_{ji})$. The second factor is closer to the true probability distribution than its constant counterpart in *Direct Estimation* but it is still

equal for any combination of values of Π_j and thus has no discrimination power.

ADE jumps from a very specific population (the set of training examples where $\Pi_j = \pi_{jk}$) to a very general population (the whole training set). In contrast, we present a model, that we call Hierarchical Pattern Bayes (HPB), which moves slowly from smaller populations to larger ones benefiting from the discrimination power available at each level.

2 The Hierarchical Pattern Bayes Classifier

HPB is a generalization of ADE that employs a hierarchy of patterns. It combines the influence of different level patterns in a way that the most specific patterns always dominate if they are well represented and combines patterns in the same level making strong independence assumptions and a calibration mechanism.

HPB works for classification problems where all attributes are nominal. Given a pattern W and a training set of pairs (X, C), where C is a class label and X is a pattern, HPB calculates $P(C_r|W)$ for any class C_r where a pattern is as defined below:

Definition 1 *A pattern is a set of pairs of the form* $(Attribute = Value)$*, where any attribute can appear at most once. An attribute that is not in the set is said to be undefined or missing.*

Definition 2 *A pattern Y is more generic than a pattern W if and only if $Y \subseteq W$. If Y is more generic than W, we say that W satisfies Y.*

Definition 3 *A pattern Y is strictly more generic than W if and only if $Y \subset W$.*

Definition 4 *The level of a pattern W,* $level(W)$*, is the number of attributes defined in W.*

Definition 5 $G(W)$ *is the set of all patterns strictly more generic than a pattern W*

2.1 The Hierarchical Model

HPB calculates the posterior probability $P(C_r|W)$, using a strategy that is similar to *Almost Direct Estimation*, but the prior probabilities are considered to be given by $P(C_r|G(W))$.

The parameters of the Dirichlet prior probability distribution used by HPB are given by: $\alpha_r = S \cdot P(C_r|G(W))$, where S is a smoothing coefficient. Consequently:

$$P(C_r|W) = \frac{N_{wr} + S \cdot P(C_r|G(W))}{N_w + S} \tag{4}$$

where N_w is the number of patterns in the training set satisfying the pattern W and N_{wr} is the number of instances in the training set satisfying the pattern W whose class label is C_r.

Given equation 4, the problem becomes to calculate $P(C_r|G(W))$. Our basic idea is to write $P(C_r|G(W))$ as a function of the various $P(C_r|W_j)$ where the W_j are patterns belonging to $G(W)$ and calculate each $P(C_r|W_j)$ recursively using equation 4.

Definition 6 $g(W)$ *is the subset of $G(W)$ whose elements have level equal to* $level(W) - 1$.

For example, if W is $\{A = a, B = b, C = c\}$, $g(W)$ is:

$$\{\ \{B = b, C = c\}, \{A = a, C = c\}, \{A = a, B = b\}\ \}$$

We consider that only $g(W)$ influences $P(C_r|G(W))$ directly, so that $P(C_r|G(W)) = P(C_r|g(W))$. The influence of the other patterns in $G(W)$ are captured by the recursive process. The first step for the decomposition of $P(C_r|g(W))$ in an expression that can be evaluated recursively is to apply Bayes theorem:

$$\begin{aligned} P(C_r|g(W)) &= \frac{P(g(W)|C_r)P(C_r)}{P(g(W))} \\ &\propto P(W_1, W_2, \ldots, W_L|C_r)P(C_r) \end{aligned}$$

where W_1,W_2,...,W_L are the elements of $g(W)$.

Then we approximate the joint probability $P(W_1, W_2, \ldots, W_L|C_r)$ by the product of the marginal probabilities:

$$P'(C_r|g(W)) \propto P(C_r) \prod_{j=1}^{L} P(W_j|C_r) \tag{5}$$

but apply a calibration mechanism:

$$P(C_r|g(W)) \propto P'(C_r|g(W)) + B.P(C_r) \tag{6}$$

where B is a calibration coefficient.

Given equations 5 and 6 we need to calculate $P(W_j|C_r)$. Applying Bayes theorem:

$$P(W_j|C_r) = \frac{P(C_r|W_j)P(W_j)}{P(C_r)} \tag{7}$$

We estimate $P(C_r)$ using the maximum likelihood approach: $P(C_r) = N_r/N$, where N_r is the number of examples in the training set belonging to class C_r, and N is the total number of examples in the training set. If it happens that N_r is zero we cannot use equation 7. In this case we just define that $P(C_r|W)$ is zero for any pattern W.

We know that when we substitute $P(W_j|C_r)$ by the right side of equation 7 into equation 5 we are able to clear out the factor $P(W_j)$ because it is identical for all classes, so we do not need to worry about it.

Since W_j is a pattern, the estimation of $P(C_r|W_j)$ can be done recursively using equation 4. The recursion ends when $g(W)$ contains only the empty pattern. In this case $P(C_r|g(W))$ becomes $P(C_r|\{\{\}\}) = P(C_r)$.

2.2 Calibration Mechanism

In spite of its strong independence assumptions, Naive Bayes is know to perform well in many domains when only misclassification rate is considered [Domingos and Pazzani, 1997]. However, Naive Bayes is also know to produce unbalanced probability estimates that are typically too "extreme" in the sense that they are too close to zero or too close to one. In the aim of obtaining better posterior probability distributions, calibration mechanisms which try to compensate the overly confident predictions of Naive Bayes have been proposed [Bennett, 2000; Zadrozny, 2001].

Using equation 5 we are making stronger independence assumptions than Naive Bayes. Naive Bayes assumes that

attributes are independent given the class, what is at least possible. Equation 5 assumes that some aggregations of attributes are independent given the class. Since many of these aggregations have attributes in common we know that such assumption is false. The main consequences of our stronger and unrealistic assumption are even more extreme probability estimates than Naive Bayes' ones. This is compensated by the calibration mechanism in equation 6. This calibration mechanism is analogous to the one used in [Zadrozny, 2001] in the calibration of decision tree probability estimates.

2.3 Selecting HPB Coefficients

Equations 4 and 6 require respectively the specifications of coefficients S and B. In the classification of a single instance, these equations are applied by HPB in the calculation of $P(C_r|W)$ for several different patterns, W. The optimal values of S and B can be different for each pattern.

In the case of the B coefficients, we use an heuristic motivated by the fact that the level of any pattern in $g(W)$ is $level(W) - 1$. The higher such level is, the more attributes in common the aggregations have, the more extreme probability estimates are and the stronger must be the effect of the calibration mechanism. Thus, we made the coefficient B in equation 6 equal to $b(level(W) - 1)$ where b is an experimental constant.

In the case of the S coefficients, we employ a greed optimization approach that starts from the most general pattern family and move toward the more specific ones, where a pattern family is the set containing all patterns that define exactly the same attributes (possibly with different values).

Assuming that the S coefficients have already been fixed for all pattern families that are more generic than a family F, there is a single S coefficient that needs to be specified to allow the use of equation 4 to calculate $P(C_r|W)$ where W is any pattern belonging to F . We select this coefficient, using leave one out cross validation, in order to maximize the area under the hit curve that is induced when we calculate $P(C_r|W)$ for all training patterns, W, in F.

3 Experimental results

All classification methods were tested by the Weka Experimenter tool [Witten and Frank, 1999] using 5 fold cross validation. We compared classifiers built using the following methods:

- *HPB*: HPB as described in this paper;
- *NB*: Naive Bayes;
- *Noisy-Or*: BN with the *Direct BN Structure* using Noisy-Or instead of a CPT;
- *TAN*: TAN with traditional CPTs;
- *ADE*: *Almost Direct Estimation*. BN with the *Direct BN Structure* and the smoothing schema described in [Friedman *et al.*, 1997];
- *DE*: *Direct Estimation*. BN with the *Direct BN Structure* and traditional CPTs;
- *AIBN/TAN*: TAN with traditional CPTs trained over a dataset where cardinality reduction using AIBN was previously applied;
- *DG CBM*: BN with DGs. Complete splits, binary splits and merges enabled;
- *DG CB*: BN with DGs. Complete splits and binary splits enabled;
- *DG C*: BN with DGs. Only complete splits enabled;
- *DG CM*: BN with DGs. Complete splits and merges enabled;
- *HC DT*: BN with Decision Trees learned using Hill Climbing (HC) and MDL as the scoring metric;
- *HC DF*: BN with Default Tables learned using HC and MDL.
- *PRIOR*: Trivial classifier that assigns the prior probability to every instance.

All models involving DGs were constructed following [Chickering *et al.*, 1997]. The models involving DFs and DTs were constructed following [Friedman and Goldszmidt, 1996] using the MDL scoring metric.

We tried different parameterizations for each method and sticked with the parameter set that provided the best results, where best results mean best area under the hit curve [1] up to 20% of selection rate (AUC20) [2]. In the y axis, we chose to represent the $Recall = N_{TruePositives}/N_{Positives}$, instead of the absolute number of hits, because this does not change the form of the curve and makes interpretation easier. We represented the selection rate in log scale to emphasize the beginning of the curves. Besides using the hit curve, we compared the probability distributions estimated by the models with the distribution actually found in the test set using two measures: Root Mean Squared Error (RMSE) and Mean Cross Entropy (MCE). Figure 1, table 1 and table 2 show our results.

Selection of method parameters is explained below:

The smoothing coefficients employed by HPB are all automatically optimized. Such optimization involves a leave one out cross validation that takes place absolutely within the current training set (the 5 fold cross validation varies the current training set) eliminating the possibility of fitting the test set. The B coefficients are defined by the heuristic described in section 2.3 and by the constant b. We varied b over the enumeration $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ and sticked with 2.0, which was the constant that produced the best results in a 5 fold cross validation process. To avoid the effects of fine tuning, we reshuffled the data set before starting another 5 fold cross validation process. The HPB results that we present here came from the second process.

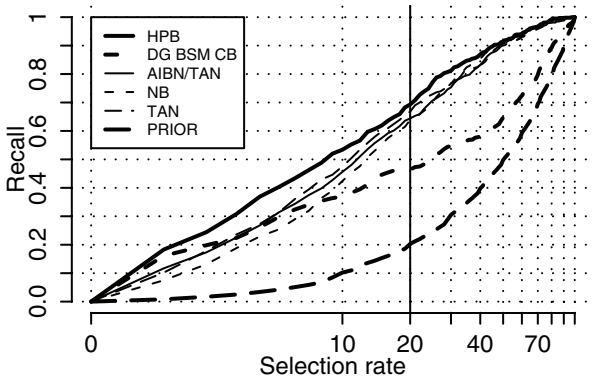

Figure 1: Selection Rate X Recall (to avoid pollution we only present curves related to a subset of the tested methods)

The optimization of AIBN/TAN involves 3 parameters: The TAN smoothing constant (S_{TAN}), the AIBN smooth-

[1] We employed hit curves, instead of the more popular ROC curves, because they match the interests of the customs administrations directly, i.e, the human resource allocation system defines a selection rate and needs an estimate for the number of positives instances that will be detected.

[2] All selection rates of interest are below 20%

ing constant S_{AIBN}, and the minimum mutual information constant MMI. Varying S_{TAN} over the enumeration $\{0.01, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$, S_{AIBN} over the enumeration $[0.1, 0.2, 0.5, 1.0]$ and MMI over the enumeration $\{0.99, 0.999, 0.9999\}$ we found that the best results are obtained with the triple $BestTriple = (S_{tan} = 0.5, S_{AIBN} = 0.5, MMI = 0.999)$. We did not cover the whole grid but since we did not observe any abrupt variations in hit curves and since we covered all immediate neighbors of the $BestTriple$ we believe that such exhaustive covering was not necessary.

Method	1%	2%	5%	10%	20%
HPB	17.8±3.2	25.4±1.3	43.2±2.3	56.4±2.1	72.3±5.0
TAN	8.8±2.6	17.4±1.3	34.0±2.0	48.6±3.3	67.1±3.4
AIBN/TAN	10.8±2.3	17.2±2.7	32.2±2.7	47.6±3.2	68.2±2.7
NB	8.8±0.9	14.2±3.1	28.9±3.9	47.4±4.6	65.4±5.8
Noisy-Or	8.6±2.1	15.2±2.1	30.3±1.5	45.9±3.3	61.5±1.5
BNDG CB	17.6±3.4	21.8±4.5	28.6±3.3	40.8±4.8	53.7±6.5
BNDG CBJ	14.7±2.6	20.6±3.0	32.2±3.5	40.6±2.9	51.5±2.6
BNDG C	9.1±2.2	12.2±2.5	21.1±2.9	30.1±2.2	46.4±1.6
ADE	14.0±2.8	15.4±2.8	20.5±2.4	28.1±3.3	43.4±4.0
DE	10.2±2.0	12.0±2.5	18.1±2.2	24.7±3.1	41.1±3.6
BNDG CJ	3.7±0.7	7.8±0.7	15.5±3.7	24.0±2.4	39.3±3.4
HC DF	4.0±1.8	5.0±3.0	10.8±3.9	22.2±3.1	35.4±1.8
HC DT	1.0±1.0	1.3±1.1	6.7±2.4	11.5±2.5	24.7±3.8
PRIOR	0.8±0.0	1.7±0.0	4.2±0.0	10.2±0.0	20.3±0.0

Table 1: Recall for different selection rates with std. dev.

Method	AUC	AUC20	RMSE	MCE	Parameterization
HPB	84.5±1.6	53.5±2.3	10.6±0.0	4.0±0.1	$b = 2.0$
TAN	82.2±0.9	46.2±2.5	14.1±0.8	7.1±0.4	$s = 0.025$
AIBN/TAN	82.1±1.3	46.0±1.2	12.5±0.5	5.3±0.3	$BestTriple$
NB	81.0±2.1	43.6±3.4	14.5±0.1	6.6±0.2	$s = 0.025$
Noisy-Or	78.2±0.9	43.0±1.6	11.9±0.0	$infinity$	
BNDG CB	68.4±3.7	40.1±3.7	11.4±0.1	6.2±0.2	$s = 0.5$
BNDG CBJ	70.9±2.7	39.2±2.5	11.7±0.1	7.1±1.2	$s = 0.1$
BNDG C	70.4±2.0	30.1±1.7	11.0±0.1	4.6±0.1	$s = 0.01$
ADE	73.7±1.1	28.7±3.0	11.1±0.2	4.9±0.1	$s = 0.025$
DE	72.7±1.3	25.9±2.7	37.5±0.0	31.6±0.0	$s = 2.0$
BNDG CJ	68.5±1.4	24.2±2.2	11.1±0.1	4.7±0.1	$s = 0.01$
HC DF	62.9±3.4	21.3±1.8	10.9±0.0	4.6±0.0	$s = 0.05$
HC DT	51.8±0.5	12.9±1.8	10.9±0.0	4.7±0.0	$s = 0.25$
PRIOR	50.0±0.0	10.2±0.0	10.9±0.0	4.7±0.0	

Table 2: Area Under Curve, Accuracy of Probability Estimates and Optimal parametrization

Noisy-or and PRIOR have no parameters. The optimization of all other methods involves only the smoothing constant, which, in all cases, was exhaustively varied over the enumeration $\{0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 1.0, 2.0\}$. This enumeration covers different magnitudes for the smoothing constant and at the same time avoids fine tuning.

Naive Bayes and Noisy-Or are both unable to capture interactions among the attributes and cannot explore critical patterns. This explains their performance in the very beginning of the hit curve. Tree Augmented Naive Bayes explores interactions among some attributes and performed better than Naive Bayes or Noisy-Or. However, it cannot benefit from some critical patterns involving many attributes that are decisive at the selection rate of 1% and 2%.

Applying cardinality reduction to the attributes before constructing the TAN model did not lead to any significant improvements in the hit curve.

Substituting the traditional CPTs of Bayesian Network by Decision Trees, Default Tables and Decision Graphs with binary splits disabled only made the hit curves worse. The learned Default Tables included very few rows. The learned Decision Trees and Decision Graphs involved very few splits. As a consequence, in all cases, the resulting models had little discrimination power.

Using Decision Graphs with binary splits enabled, the most critical patterns were separated from the others, what resulted in a significant improvement in the beginning of the hit curves. The other patterns were, almost all, left together what resulted in loss of discrimination power for selection rates above 5%.

Hypothesis tests, show that HPB is significantly better than all other classifiers in what regards to AUC, AUC20, RMSE and MCE. We also performed Hypothesis tests for every selection rate from 1% to 100% in steps of 1%. HPB is significantly better than all other classifiers for every selection rate below 30% with the exceptions that it is not significantly better than: TAN in $[17\%, 19\%]$ and $[27\%, 28\%]$; AIBN TAN in $[18\%, 21\%]$ and $[24\%, 28\%]$; BNDG CBJ at 1%; BNDG CB at 1% and 2%.

HPB benefits from critical patterns involving many or even all attributes but also considers the influence of less specific patterns. As a consequence, it performs well for any selection rate.

4 Conclusions

In the domain of preselection of imported goods for verification some combinations of attribute values can constitute critical patterns whose influence over the probability of finding a positive instance is very significant. Due to the high cardinality of the attributes in this domain, exploiting such patterns without overfitting is challenging. We addressed the problem using HPB a novel classification method based on a multilevel hierarchical Bayesian model.

HPB was shown capable of capturing the influence of critical patterns without overfitting and without loosing discrimination power after the exhaustion of critical patterns in the test set. HPB resulted in a hit curve that is almost unambiguously better than any of the other methods. HPB also produced better probability estimates according to two accuracy measures (table 2).

HPB was only validated in a specialized domain, however, its equations are too simple to reflect any particularities of the domain, except that it is characterized by few high cardinality attributes and relevant interactions among them. Thus the use of HPB may be a good option for domains with the same characteristics or, at least, provide some light on how to develop good models for such domains.

HPB training time is exponential in the number of attributes, linear in the number of training instances and independent of the attributes cardinality. Thus HPB is only applicable to domains where there are few attributes, but in such domains it is much faster than methods whose training time depends on the cardinality of the attributes.

HPB is not a full Bayesian model in the sense of [Gelman *et al.*, 2003], where the parameters associated with a sub-population are assumed to be drawn from a general distribution and the calculation of all involved probability distributions is done at once considering all available evidence.

Instead, HPB estimates the probability distributions for the more general populations first and use the results in the estimation of the probability distributions related to the more specific populations. HPB is a generalization of the smoothing techniques used in [Cestnik, 1990] and [Friedman *et al.*, 1997]. In the sense of [Gelman *et al.*, 2003], it is an empirical model.

Hierarchical Bayesian models have been widely used in the marketing community under the name of Hierarchical Bayes [Allenby *et al.*, 1999; Lenk *et al.*, 1996]. These models have also been used in medical domains [Andreassen *et al.*, 2003] and robotics [Stewart *et al.*, 2003]. However, we are not aware of any hierarchical Bayesian model that can be employed to handle high cardinality attributes with relevant interactions in a classification problem. This makes HPB relevant. Moreover, HPB differs from other models by dealing with a multi level hierarchy recursively and also handling the fact that one sub-population is contained by several overlapping super-populations and not only by one super-population.

Based on the literature [Friedman and Goldszmidt, 1996], one can expect that Bayesian Networks with Default Tables, Decision Trees or Decision Graphs can emulate the real complexity of the interactions present in the data with without serious overfitting problems. Another contribution of this paper is to show that BNs with DFs, DTs or DGs, in spite of their theoretical motivations, actually do not result in better overall performance than simpler methods like NB or TAN, in a practical domain where their abilities are truly necessary.

Preprocessing the data reducing the cardinality of the attributes using the Agglomerative Information Bottleneck method did not result in any significant improvements in the hit curves.

Our present mechanism for selecting the smoothing and the calibration coefficients in HPB equations is too simplistic. We leave its improvement as future work.

References

[Allenby *et al.*, 1999] Greg M. Allenby, Robert P. Leone, and Lichung Jen. A dynamic model of purchase timing with application to direct marketing. *Journal of the American Statistical Association*, 94(446):365–374, 1999.

[Andreassen *et al.*, 2003] Steen Andreassen, Brian Kristensen, Alina Zalounina, Leonard Leibovici, Uwe Frank, and Henrik C. Schonheyder. Hierarchical dirichlet learning - filling in the thin spots in a database. In Michel Dojat, Elpida T. Keravnou, and Pedro Barahona, editors, *Proceedings of the 9th Conference on Artificial Intelligence in Medicine (AIME)*, volume 2780 of *Lecture Notes in Computer Science*, pages 204–283. Springer, 2003.

[Bennett, 2000] Paul N. Bennett. Assessing the calibration of naive bayes' posterior estimates. Technical Report CMU-CS-00-155, School of Computer Science, Carnegie Mellon University, 2000.

[Cestnik, 1990] B. Cestnik. Estimating probabilities: a crucial task in machine learning. In *Proceedings of the European Conference on Artificial Intelligence*, pages 147–149, 1990.

[Chawla *et al.*, 2002] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence and Research*, 16:321–357, 2002.

[Chickering *et al.*, 1997] David Maxwell Chickering, David Heckerman, and Christopher Meek. A bayesian approach to learning bayesian networks with local structure. Technical Report MSR-TR-97-07, Microsoft Research, Redmond, WA 98052, 1997.

[Domingos and Pazzani, 1997] Pedro Domingos and Michael J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

[Duda and Hart, 1973] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[Friedman and Goldszmidt, 1996] Nir Friedman and Moises Goldszmidt. Learning bayesian networks with local structure. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Inteligence (UAI)*, pages 252–262, San Francisco, CA, 1996. Morgan Kaufmann Publishers.

[Friedman *et al.*, 1997] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

[Gelman *et al.*, 2003] Andrew B. Gelman, John S. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, second edition, 2003.

[Lenk *et al.*, 1996] P. Lenk, W. DeSarbo, P. Green, and M. Young. Hierarchical bayes conjoint analysis: recovery of part worth heterogeneity from reduced experimental designs. *Marketing Science*, 15:173–191, 1996.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

[Slonim and Tishby, 1999] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In *Advances in Neural Information Processing Systems 12 (NIPS)*, pages 617–623, Denver, Colorado, USA, 1999. The MIT Press.

[Stewart *et al.*, 2003] Benjamin Stewart, Jonathan Ko, Dieter Fox, and Kurt Konolige. The revisiting problem in mobile robot map building: A hierarchical bayesian approach. In Christopher Meek and Uffe Kjærulff, editors, *Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 551–558, Acapulco, Mexico, 2003. Morgan Kaufmann.

[Witten and Frank, 1999] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers Inc., 1999.

[Zadrozny, 2001] Bianca Zadrozny. Reducing multiclass to binary by coupling probability estimates. In *Proceedings of the Advances in Neural Information Processing Systems 14 (NIPS)*, Cambridge, MA, 2001. MIT Press.