Correlation Analysis in Classifiers

Vincent Lemaire
Orange Labs, 2 avenue Pierre Marzin, 22307 Lannion Cedex, France

Carine Hue
GFI Informatique, 11 rue Louis de Broglie, 22300 Lannion, France

Olivier Bernier
Orange Labs, 2 avenue Pierre Marzin, 22307 Lannion Cedex, France

Abstract: This chapter presents a new method to analyze the link between the probabilities produced by a classification model and the variation of its input values. The goal is to increase the predictive probability of a given class by exploring the possible values of the input variables taken independently. The proposed method is presented in a general framework, and then detailed for naive Bayesian classifiers. We also demonstrate the importance of "lever variables", variables which can conceivably be acted upon to obtain specific results as represented by class probabilities, and consequently can be the target of specific policies. The application of the proposed method to several data sets shows that such an approach can lead to useful indicators.

Key-words: Exploration, Correlations, Classifiers, Naïve Bayes.

INTRODUCTION

Given a database, one common task in data analysis is to find the relationships or correlations between a set of *input* or *explanatory* variables and one target variable. This knowledge extraction often goes through the building of a model which represents these relationships (Han & Kamber, 2006). Faced with a classification problem, a probabilist classifier allows, for all the instances of the database and given the values of the explanatory variables, the estimation of the probabilities of occurrence of each class.

These probabilities, or scores, can be used to evaluate existing policies and practices in organizations and governments. They are not always directly usable, however, as they do not give any indication of what action can be decided upon to change this evaluation. Consequently, it seems useful to propose a methodology which would, for every instance in the database, (i) identify the importance of the explanatory variables; (ii) identify the position of the values of these explanatory variables; and (iii) propose an action in order to change the probability of the desired or target class. We propose to deal with the third point by exploring the model relationship between each explanatory variable independently from each other and the target class. The proposed method presented in this chapter is completely automatic.

This article is organized as follows: the first section positions the approach in relation to the state of the art; the second section describes the method at first from a generic point of view and then for the naive Bayes classifier. Through three illustrative examples the third section allows a discussion and a progressive interpretation of the obtained results. In each illustrative example different practical details of the proposed method are explored. Finally we shall conclude.

BACKGROUND

Machine learning abounds with methods for supervised analysis in classification. Generally these methods propose algorithms to build a model, a probabilist classifier, from a training database made up of a finite number of examples. The output vector gives the predicted probability of the occurrence, or score, of each class label. In general, however, this probability of occurrence is not sufficient and an interpretation and analysis of the result in terms of correlations or relationships between input and output variables is needed. The interpretation of the model is often based on the parameters and the structure of the model. One can cite, for example: geometrical interpretations (Brennan & Seiford, 1987), interpretations based on rules (Thrun, 1995) or fuzzy rules (Benitez, Castro, & Requena, 1997), statistical tests on the coefficient's model (Nakache & Confais, 2003). Such interpretations are often based on averages for several instances, for a given model, or for a given task (regression or classification).

Another approach, called sensitivity analysis, consists in analyzing the model as a black box by varying its input variables. In such "what if" simulations, the structure and the parameters of the model are important only as far as they allow accurate computations of dependant variables using explanatory variables. Such an approach works whatever the model. A large survey of "what if" methods, often used for artificial neural networks, are available in (Leray & Gallinari, 1998; Lemaire, Féraud, & Voisine, 2006).

Variable importance

Whatever the method and the model, the goal is often to analyze the behavior of the model in the absence of one explanatory variable, or a set of explanatory variables, and to deduce the importance of the explanatory variables, for all examples. The reader can find a large survey in (Guyon, 2005). The measure of the importance of the explanatory variables allows the selection of a subset of relevant variables for a given problem. This selection increases the robustness of models and simplifies the understanding of the results delivered by the model. The variety of supervised learning methods, coming from the statistical or artificial intelligence communities often implies importance indicators specific to each model (linear regression, artificial neural network ...).

Another possibility is to try to study the importance of a variable for a given instance and not in average for all the examples. Given a variable and an instance, the purpose is to obtain the variable importance only for this instance: for additive classifiers see (Poulin et al., 2006), for Probabilistic RBF Classification Network see (Robnik-Sikonja, Likas, Constantinopoulos, & Kononenko, 2009), and for a general methodology see (Lemaire & Féraud, 2008). If the model is restricted to a naive Bayes Classifier, a state of art is presented in (Možina, Demšar, Kattan, & Zupan, 2004; Robnik-Sikonja & Kononenko, 2008). This importance gives a specific piece of information linked to one example instead of an aggregate piece of information for all examples.

Importance of the value of an explanatory variable

To complete the importance of a variable, the analysis of the value of the considered variable, for a given instance, is interesting. For example Féraud et al. (Féraud & Clérot, 2002) propose to cluster examples and then to characterize each cluster using the variables importance and importance of the values inside every cluster. Framling (Framling, 1996) uses a "what if" simulation to place the value of the variable and the associated output of the model among all the potential values of the model outputs. This method which uses extremums and an assumption of monotonous variations of the output model versus the variations of the input variable has been improved in (Lemaire & Féraud, 2008).

Instance correlation between an explanatory variable and the target class

This chapter proposes to complete the two aspects presented above, namely the importance of a variable and the importance of the value of a variable. We propose[1] to study the correlation, for one instance and one variable, between the input and one output of the probabilist classifier, the score of the target class.

For a given instance, the distinct values of a given explanatory variable can pull up (higher value) or pull down (lower value) the model output. The proposed idea is to analyze the relationship between the values of an input variable and the probability of occurrence of a given target class. The goal is to increase (or decrease) the score, the target class probability, by exploring the different values taken by the explanatory variable. For instance for medical data one tries to decrease the probability of a disease; in case of cross-selling one tries to increase the appetency to a product; and in government data cases one tries to define a policy to reach specific goals in terms of specific indicators (for example decrease the unemployment rate).

This method does not explore causalities, only correlations, and can be viewed as a method between:

- selective sampling (Roy & McCallum, 2001) or adaptive sampling (Singh, Nowak, & Ramanathan, 2006): the model observes a restricted part of the universe materialized by examples but can "ask" to explore the variation space of the descriptors one by one separately, to find interesting zones.

- and causality exploration (Kramer, Leventhal, Hutchinson, & Feinstein, 1979; Guyon, Constantin Aliferis, & Elisseeff, 2007): as example D. Choudat (Choudat, 2003) propose the imputability approach to specify the probability of the professional origin of a disease. The causality probability is, for an individual, the probability that his disease

[1] The description of the proposed method is done only for classification problems but the method is easily adaptable for regression problems

arose from exposures to professional elements. The increase of the risk has to be computed versus the respective role of each possible type exposures. In medical applications, the models used are often additive models or multiplicative models.

Lever variables

In this chapter we also advocate the definition of a subset of the explanatory variables, the "lever variables". These lever variables are defined as the explanatory variables for which it is conceivable to change their value. In most cases, changing the values of some explanatory variables (such a sex, age...) is indeed impossible. The exploration of instance correlation between the target class and the explanatory variables can be limited in practice to variables which can effectively be changed.

The definition of these lever variables will allow a faster exploration by reducing the number of variable to explore, and will give more intelligible and relevant results. Lever variables are the natural target for policies and actions designed to induce changes of occurrence of the desired class in the real world.

CORRELATION EXPLORATION - METHOD DESCRIPTION

In this section, the proposed method is first described in the general case, for any type of predictive model, and then tested on naive Bayes classifiers.

General case

Let C_z be the target class among T classes. Let f_z be the function which models the predicted probability of the target class (the score) $f_z(X=x) = P(C_z \mid X=x)$, given the equality of the vector X of the J explanatory variables to a given vector x of J values. Let v_{jn} be all the n different possible values of the variable X_j.

The Algorithm 1 describes the proposed method. This algorithm tries to increase the value of $P(C_z \mid X = x_k)$ successively for each of the K examples of the considered sample set using the set of values of all the explanatory variables or lever variables. This method is halfway between selective sampling (Roy & McCallum, 2001) and adaptive sampling (Singh et al., 2006). The model observes a restricted part of the universe materialized by examples but can "ask" to explore the variation space of the descriptors one by one separately, to find interesting zones. The next subsections describe the algorithm in more details.

Exploration of input values

For the instance $\mathrm{x_k}$, $P(C_z \mid x_k)$ is the "natural" value of the model output. We propose to modify the values of the explanatory variables or lever variables in order to study the variation of the probabilist classifier output for this example. In practice, we propose to explore the values independently for each explanatory variable. Let $P_j(C_z \mid x_k, b)$ be the output model f_z given the example x_k but for which the value of its j^{th} component has been replaced with the value b. For example, the third explanatory variable is modified among five variables: $P_3(C_z \mid x_k, b) = f_z(x_k^1,$

x_k^2, b, x_k^4, x_k^5). By scanning all the variables and for each of them all the set of their possible values, an exploration of "potential" values of the model output is computed for the example x_k.

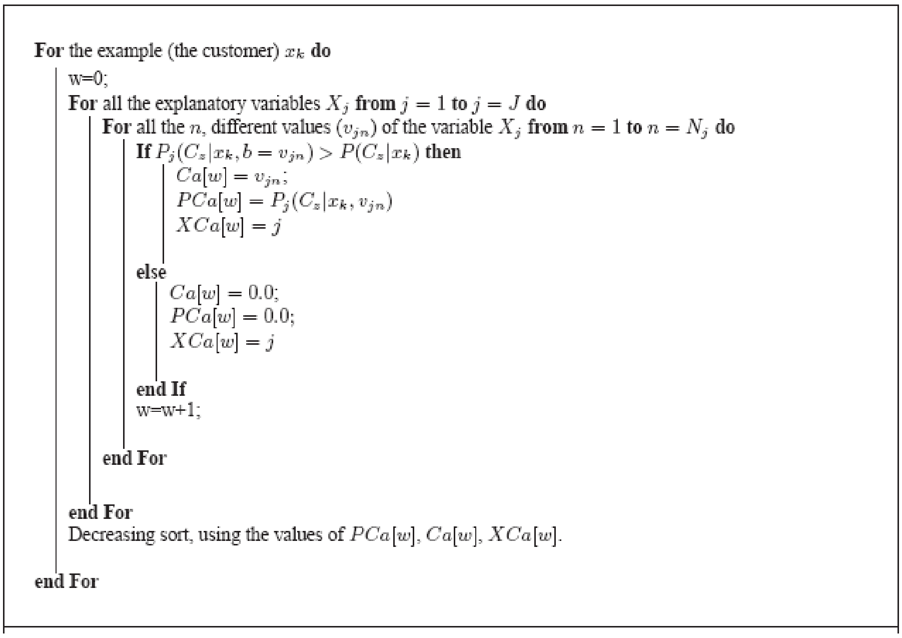
Domain of exploration of each variable

The advantage of choosing the empirical probability distribution of the data as domain of exploration has been showed experimentally in (Breiman, 2001; Lemaire et al., 2006; Lemaire & Féraud, 2008). A theoretical proof is also available for linear regression in (Diagne, 2006) and for naive Bayes classifiers in (Robnik-Sikonja & Kononenko, 2008). Consequently the values used for the *J* explanatory variables will be the values of the *K* examples available in the training database. This set can also be reduced using only the distinct values: let N_j be the number of distinct values of the variable X_j.

Results ranking

The exploration of the explanatory variables or of the lever variables is done by scanning all the possible values taken by the instances in the training set. When the modification of the value of the variable leads to an improvement of the probability predicted by the model, three pieces of data are kept (i) the value which leads to this improvement (*Ca*); (ii) the associated improved probability (*PCa*); and (iii) the variable associated to this improvement (*XCa*). These triplets are then sorted according to the improvement obtained on the predicted probability. Note: if no improvement is found, the tables *CA* and *PCa* only contain null values.

It should also be possible (i) to explore jointly two or more explanatory variables; (ii) or to use the value (*Ca*[0]) which best improves the output of the model (*$P(C_z$ | X =x)*) (this value *Ca*[0] is available at the end of the Algorithm) and then to repeat again the exploration on the example x_k on its others explanatory variables. These other versions are not presented in this chapter but will be the focus of future works.



```
For the example (the customer) $x_k$ do
    w=0;
    For all the explanatory variables $X_j$ from $j = 1$ to $j = J$ do
        For all the $n$, different values ($v_{jn}$) of the variable $X_j$ from $n = 1$ to $n = N_j$ do
            If $P_j(C_z|x_k, b = v_{jn}) > P(C_z|x_k)$ then
                $Ca[w] = v_{jn}$;
                $PCa[w] = P_j(C_z|x_k, v_{jn})$
                $XCa[w] = j$

            else
                $Ca[w] = 0.0$;
                $PCa[w] = 0.0$;
                $XCa[w] = j$

            end If
            w=w+1;

        end For

    end For
    Decreasing sort, using the values of $PCa[w]$, $Ca[w]$, $XCa[w]$.

end For
```

Algorithm 1: Exploration and ranking of the score improvements

Cases with class changes

When using Algorithm 1, the predicted class can change. Indeed it is customary to use the following formulation to designate the predicted class of the example x_k:

$$\arg\max_{Z} P(C_z | x_k)$$

Using Algorithm 1 for x_k belonging to the class t $(t \neq z)$ could produce $P(C_z \mid x_k, b) > P(C_t \mid x_k)$. In this case the corresponding value (Ca) carries important information which can be exploited.

The use of Algorithm 1 can exhibit three types of values (Ca):

- values which do not increase the score (target class probability);
- values which increase the score but without class change (the probability increase is not sufficient);
- values which increase the score with class change (the probability increase is sufficient).

The examples whose predicted class changes from another class to the target class are the primary target for specific actions or policies designed to increase the occurrence of this class in the real world.

Case of a naive Bayesian classifier

A naive Bayes classifier assumes that all the explanatory variables are independent knowing the target class. This assumption drastically reduces the necessary computations. Using the Bayes theorem, the expression of the obtained estimator for the conditional probability of a class C_z is:

$$P(C_z \Big| x_k) = \frac{P(C_z)\prod_{j=1}^{J} P(X_j = v_{jk} \big| C_z)}{\sum_{t=1}^{T} P(C_t)\prod_{j=1}^{J} P(X_j = v_{jk} \big| C_t)} \quad (1)$$

The predicted class is the one which maximizes the conditional probabilities. Despite the independence assumption, this kind of classifier generally shows satisfactory results (Hand & Yu, 2001). Moreover, its formulation allows an exploration of the values of the variables one by one independently.

The probabilities $P(X_j = v_{jk} \mid C_z)$ ($\forall$ j, k, z) are estimated using counts after discretization for numerical variables or grouping for categorical variables (Boullé, 2008). The denominator of the equation above normalizes the result so that $\sum_z P(C_z \mid x_k)$=1.

The use of the Algorithm 1 requires to compute $P(C_z \mid X = x_k)$, and $P_j(C_z \mid X = x, b)$ which can be written in the form of Equations 2 and 3:

$$P(C_z | x_k) = \frac{\overbrace{P(C_z)\prod_{j=1}^{J} P(X_j = v_{jk} | C_z)}^{e^{L_z}}}{\sum_{t=1}^{T} P(C_t)\prod_{j=1}^{J} P(X_j = v_{jk} | C_t)} \quad (2)$$

$$P_j(C_z | x_k, b) = \frac{\overbrace{P(C_z)\prod_{j=1, j\neq q}^{J} P(X_j = v_{jk} | C_z) P(X_q = b | C_z)}^{e^{L_{z'}}}}{\sum_{t=1}^{T} \left[P(C_t)\prod_{j=1}^{J} P(X_j = v_{jk} | C_t)\right] P(X_q = b | C_t)} \quad (3)$$

In Equations 2 and 3 numerators can be written as e^{Lz} and $e^{Lz'}$ with:

$$\mathrm{L_z} = \log(\mathrm{P(C_z)}) + \sum_{\mathrm{j=1}}^{\mathrm{J}} \log(\mathrm{P(X_j = v_{jk} \mid C_z)})$$

and

$$\mathrm{L_{z'}} = \log(\mathrm{P(C_z)}) + \sum_{j=1, j\neq q}^{J} \left[\log(\mathrm{P(X_j = v_{jk} \mid C_{z'})}) + \log(\mathrm{P(X_q = b \mid C_z)})\right]$$

This formulation will be used below.

Implementation details on very large databases

To measure the reliability of our approach, we tested it on marketing campaigns of France Telecom (results not allowed for publication until now). Tests have been performed using the PAC platform (Féraud, Boullé, Clérot, & Fessant, 2008) on different databases coming from decision-making applications. The databases used for testing had more than 1 million of customers, each one represented by a vector including several thousands of explanatory variables. These tests raise several implementation points enumerated below:

- To avoid numerical problems when comparing the "true" output model $P(C_z | x_k)$ and the "explored" output $P_j(C_z | x_k, b)$, $P(C_x | x_k)$ is computed as:

$$P(C_x | x_k) = \frac{1}{\sum_{t=1}^{T} e^{L_t - L_x}}$$

where $L_t = \log(P(C_t)) + \sum_{j=1}^{J} \log(P(X_j = v_{jk} | C_t))$

- To reduce the computation time: the modified output of the classifier can be computed using only several additions or subtractions since the difference between L_z (used in Equation 2) and $L_{z'}$ (used in Equation 3) is:

$$L_{z'} = L_z - log(P(x_q=v_{jk} | C_z)) + log(P(X_q=b | C_z))$$

- Complexity: For a given example x_k, the computation of tables presented in Algorithm 1 is of complexity $O(\sum_{j=1}^{d} N_j)$.

This implementation is "real-time" and can be used by an operator who asks the application what actions to do, for example to keep a customer.

EXPERIMENTATIONS

In this section we describe the application of our proposed method to three illustrative examples. This first example, the Titanic database, illustrates the importance of lever variables. The second example illustrates the results of our method on the dataset used for the PAKDD 2007 challenge. Finally, we present the results obtained by our method on a government data problem, the analysis of the type of contraceptive used by married women in Indonesia.

The Titanic database:

Data and experimental conditions

In this first experiment the Titanic (www.ics.uci.edu/~mlearn/) database is used. This database consists of four explanatory variables on 2201 instances (passengers and crew members). The first attribute represents the class trip (status) of the passenger or if he was a crew member, with values: 1st, 2nd, 3rd, crew. The second (age) gives an age indication: adult, child. The third (sex) indicates the sex of the passenger or crew: female or male. The last attribute (survived) is the target class attribute with values: no or yes. Readers can find for each instance the variable importance and the value importance for a naive Bayes classifier in (Robnik-Sikonja & Kononenko, 2008).

Among the 2201 examples in this database, a training set of 1100 examples randomly chosen has been extracted to train a naive Bayes classifier using the method presented in (Boullé, 2008). The remaining examples constitute a test set. As the interpretation of a model with low performance would not be consistent, a prerequisite is to check if this naive Bayes classifier is correct. The model used here (Guyon, Saffari, Dror, & Bumann, 2007) gives satisfactory results:

- Accuracy on Classification (ACC) on the train set: 77.0%; on the test set: 75.0%;
- Area under the ROC curve (AUC) (Fawcett, 2003) on the train set: 73.0%; on the test set: 72.0%.

The purpose here is to the see another side of the knowledge produced by the classifier: we want to find the characteristics of the instances (people) which would have allowed them to survive.

Input values exploration

Algorithm 1 has been applied on the test set to reinforce the probability to survive (score). Table 1 shows an abstract of the results: (i) it is not possible to increase the probability for only one passenger or crew; (ii) the last column indicates that, for persons predicted as surviving by the model (343 people), the first explanatory variable (status) is the most important to reinforce the probability to survive for 118 cases; then the second explanatory variable (age) for 125 cases; and at last the third one (sex) for 100 cases. (iii) For people predicted as dead by the model (758) the third explanatory variable (sex) is always the variable which is the most important to reinforce the probability to survive.

	Size	Status / Age / Sex
Predicted 'yes'	343	118 / 125 / 100
Predicted 'no'	758	0 / 0 / 758

Table 1: Ranking of explanatory variables

These 758 cases predicted as dead are men and if they were women their probability to survive would increase sufficiently to survive (in the sense that their probability to survive would be greater than their probability to die). Let us examine then, for these cases, additional results obtained by exploring the other~~s~~ variables using Algorithm 1:

- the second best variable to reinforce the probability to survive is (and in this case they survive):
    - for 82 of them (adult + men + 2nd class) the second explanatory variable (age);
    - for 676 of them (adult + men + (crew or 3rd class)) the first explanatory variable (status);
- the third best variable to reinforce the probability to survive is (and in this case nevertheless they are dead):
    - for 82 of them (adult + men + 2nd class) the first explanatory variable (status);
    - for 676 of them (adult + men + (crew or 3rd class)) the second explanatory variable (age).

Of course, in this case, most explanatory variables are not in fact lever variables, as they cannot be changed (age or sex). The only variable that can be changed is status, and even in this case, only for passengers, not for crew members. The change of status for passengers means in fact buying a first class ticket, which would have allowed them a better chance to survive. The other explanatory variables enable us to interpret the obtained survival probability in terms of priority given to women and first class passengers during the evacuation.

Application to sale: results on the PAKDD 2007 challenge

Data and experimental conditions

The data of the PAKDD 2007 challenge are used (http://lamda.nju.edu.cn/conf/pakdd07/dmc07/): The data are not on-line any more but data descriptions and analysis results are still available. Thanks to Mingjun Wei (participant referenced P049) for the data (version 3).

The company, which gave the database, has currently a customer base of credit card customers as well as a customer base of home loan (mortgage) customers. Both of these products have been on the market for many years, although for some reasons the overlap between these two customer bases is currently very small. The company would like to make use of this opportunity to cross-sell home loans to its credit card customers, but the small size of the overlap presents a challenge when trying to develop an effective scoring model to predict potential cross-sell take-ups.

A modeling dataset of 40,700 customers with 40 explanatory variables, plus a target variable, had been provided to the participants (the list of the 40 explanatory variables is available at http://perso.rd.francetelecom.fr/lemaire/data_pakdd.zip). This is a sample of customers who opened a new credit card with the company within a specific 2-year period and who did not have an existing home loan with the company. The target categorical variable "Target_Flag" has a value of 1 if the customer then opened a home loan with the company within 12 months after opening the credit card (700 random samples), and has a value of 0 otherwise (40,000 random samples).

A prediction dataset (8,000 sampled cases) has also been provided to the participants with similar variables but withholding the target variable. The data mining task is to produce a score for each customer in the prediction dataset, indicating a credit card customer's propensity to take up a home loan with the company (the higher the score, the higher the propensity).

The challenge being ended it was not possible to evaluate our classifier on the prediction dataset (the submission site is closed). Therefore we decide to elaborate a model using the 40 000 samples in a 5-fold cross validation process. In this case each 'test' fold contains approximately the same number of samples as the initial prediction dataset. The model used is again a naive Bayes classifier (Boullé, 2008; Guyon, Saffari, et al., 2007). The results obtained on the test sets are:

- Accuracy on Classification (ACC): 98.29% ± 0.01% on the train sets and 98.20% ± 0.06% on the test sets.
- Area under the ROC curve (AUC): 67.98% ± 0.74% on the train sets and 67.79% ± 2.18% on the test sets.
- Best results obtained on one of the folds: Train set AUC=68.82%, Test set AUC=70.11%.

id participant	AUC for test set	Rank	Modeling Technique
P049	70.01%	1	TreeNet + Logistic Regression
P085	69.99%	2	Probit Regression
P212	69.62%	3	MLP + n-Tuple Classifier

Table 2: PAKDD 2007 challenge: the first three best results

Table 2 shows the first three best results and corresponding method of winners of the challenge. Results obtained here by our model are coherent with those of the participants of the challenge.

Input values exploration

The best classifier obtained on the test sets in the previous section is used. This naive Bayes classifier (Boullé, 2007) uses 8 variables out of 40 (the naïve Bayes classifier takes into account only input variables which have been discretized (or grouped) in more than one interval (or group) see (Boullé, 2006)). These 8 variables and their intervals of discretization (or groups) are presented in Table 3. All variable are numerical except for the variable "RENT_BUY_CODE" which is symbolic with possible values of 'O' (Owner), 'P' (Parents), 'M' (Mortgage), 'R' (Rent), 'B' (Board), 'X' (Other).

Explanatory Variables	Interval 1 or Group 1	Interval 2 or Group 2
RENT_BUY_CODE	M,R,B,X	O,P
PREV_RES_MTHS	]-∞,3.5[	[3.5,+∞ [
CURR_RES_MTHS	]-∞,40.5[	[40.5,+∞ [
B_ENQ_L6M_GR3	]-∞,0.5[	[0.5,+∞ [
B_ENQ_L3M	]-∞,0.5[	[3.5,+∞ [
B_ENQ_L12M_GR3	]-∞,1.5[	[1.5,+∞ [
B_ENQ_L12M_GR2	]-∞,0.5[	[0.5,+∞ [
AGE_AT_APPLICATION	]-∞,45.5[	[45.5,+∞ [

Table 3: Selected explanatory variables (there is no reason in (Boullé, 2006) to have two intervals for each variable, it is here blind chance).

The lever variables were chosen using their specification (see http://lamda.nju.edu.cn/conf/pakdd07/dmc07/ or the appendix A). These lever variables are those for which a commercial offer to a customer can change the value. We define another type of variable which we will explore using our algorithm, the observable variables. These variables are susceptible to change during a life of a customer and this change may augment the probability of the target class, the propensity to take up a home loan. In this case, the customers for which this variable has changed can be the target of a specific campaign. For example the variable "RENT_BUY_CODE" can not be changed by any offer but is still observable. The customer can move from the group of values [O,P] ('O' Owner, 'P' Parents) to [M,R,B,X] ('M' Mortgage, 'R' Rent, 'B' Board, 'X' Other). Among the eight variables (see Table 3) chosen by the training method of the naive Bayes classifier, two are not considered as 'lever' variables or observable variables ("AGE_AT_APPLICATION" and "PREV_RES_MTHS") and will not be explored.

Algorithm 1 has been applied on the 40700 instances in the modeling data set. The 'yes' class of the target variable is chosen as target class (C_z = 'yes'). This class is very weakly represented (700 positive instances out of 40700). The AUC values presented in Table 2 or on the challenge website does not show if customers are classified as 'yes' by the classifier. Exploration of lever variables does not allow in this case a modification of the predicted class. Nevertheless Table 4 and Figure 1 show that a large improvement of the 'yes' probability (the probability of cross-selling) is possible.

In Table 4 the second column (C2) presents the best $P_j(C_z \mid x_k, b)$ obtained, the third column (C3) the initial corresponding $P(C_z \mid x_k, b)$, the fourth column (C4) the initial interval used in the naive Bayes formulation (used to compute $P(C_z \mid x_k, b)$) and the last column (C5) the interval which gives the best improvement (used to compute $P_j(C_z \mid x_k, b)$). This table shows that:

- for all lever or observable variables, there exists a value change that increases the posterior probability of occurrences of the target class;
- the variable that leads to the greatest probability improvement is B_ENQ_L3M (The number of Bureau Enquiries in the last 3 months), for a value in [1.5,+∞[rather than in]- ∞,1.5[; This variable is an observable variable, not a lever variable, and means that a marketing campaign should be focused on customers who contacted the bureau more than once in the last three months.
- nevertheless, none of those changes leads to a class change as the obtained probability ($P_j(C_z \mid x_k, b)$) stays smaller than $P(C_z \mid x_k)$.

C1: explored variable	C2	C3	C4	C5
RENT_BUY_CODE	0.6	0.26	[O,P]	[M,R,B,X]
CURR_RES_MTHS	0.36	0.21	[40.5,+∞[	]-∞,40.5[
B_ENQ_L6M_GR3	0.25	0.10	]-∞,0.5[	[0.5,+∞[
B_ENQ_L3M	0.12	0.12	]-∞,1.5[	[1.5,+ ∞[
B_ENQ_L12M_GR3	0.36	0.16	]-∞,0.5[	[1.5,+∞[
B_ENQ_L12M_GR2	0.36	0.24	[0.5,+∞[	]- ∞,0.5[

Table 4: Best $P(C_z)$='yes' obtained

In Figure 1 the six dotted vertical axis represent the six lever or observable variables as indicated on top or bottom axis. On the left hand size of each vertical axis, the distribution of $P(C_z \mid x_k)$ is plotted (□) and on the right hand size the distribution of $P_j(C_z \mid x_k, b)$ is plotted (■). Probability values are indicated on the y-axis. In this Figure only the best $P_j(C_z \mid x_k, b)$ (*PCa*[0] in Algorithm 1) is plotted. This figure illustrates in more details the same conclusions as given above.

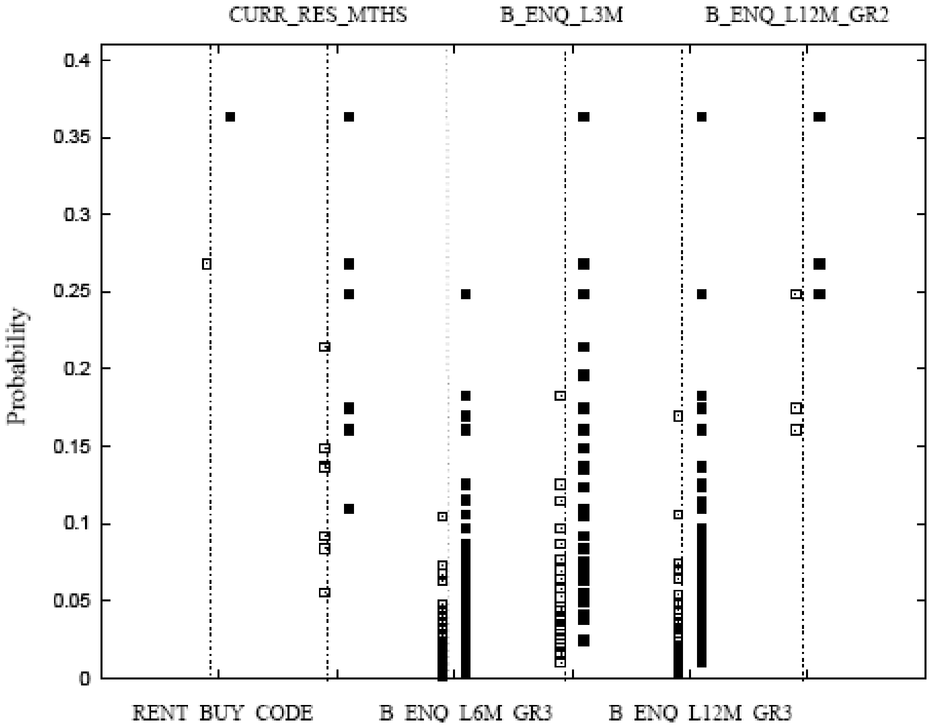


Fig 1: Obtained results on $P_j(C_z \mid x_k, b)$.

Application to government data: results for the Contraceptive Method Choice Data Set

Data and experimental conditions

The Contraceptive Method Choice Data Set is a freely available data set in the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice). This data set is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. It consists of 1473 instances, corresponding to married women either not pregnant or who did not know if they were at the time of the survey. The problem is to predict, from 9 explanatory variables (age, education, husband's education, number of children ever born, religion, working or not, husband's occupation, standard of living index, good media exposure or not) the type of contraceptive method used (no contraceptive method, short-term contraceptive method or long-term contraceptive method). Three explanatory variables are binary (religion either Islam or not, working or not, and good media exposure or not), two are numerical (age and number of children ever born) and the others are categorical.

The model used is a selective naive Bayes classifier (Boullé, 2007), trained on 75 percent of the dataset (1108 instances), the rest of the dataset being used for testing purposes. On the training subset, we obtained an AUC (Area Under ROC Curve) of 0.74, and an AUC of 0.73 for the test subset.

Input values exploration

The selective naive Bayes classifier (Boullé, 2007) uses 8 of the 9 explanatory variables, discarding the binary variable working or not. Among these variables, only two are chosen as lever variables, education and good media exposure or not. The other variables are not considered as possible targets for policies. Education is a categorical variable with four values from 1 (low education) to 4 (high education), partitioned into three groups by the classification algorithm: low education (value 1), middle education (values 2 or 3) and high education (value 4). Algorithm 1 has been applied on the 1473 instances. The target variable is in this case a three class variable (no contraceptive, short-term contraceptive, and long-term contraceptive). As the proposed algorithm can only try to increase the probability of one class, it was applied twice, once to try to increase the probability of using a short-term contraceptive (first target class), once to try to increase the probability of using a long-term contraceptive (second target class).

Applying our method to increase the probability of using a long-term contraceptive showed that the most significant lever variable is the education level. Table 5 indicates the number of instances for each predicted class and each level of education.

	No contraceptive	Short term	Long term
Low education	137	0	15
Middle education	387	19	338
High education	67	211	299
Total	591	230	652

Table 5: Number of instances for each predicted class and level of education.

Out of 1473 instances, 577 instances are already at a high education level. Out of the remaining 895 instances, 99 were predicted to switch from no contraceptive to a long term contraceptive if the education level was changed from whatever value (low or middle) to a high value, and 30 instances were predicted to switch from short term contraceptive to long term contraceptive with the same change in education level. Media exposure do not seem to have any significant impact (only 2 instances of 'class changes' to long term contraceptive, by changing the media exposure to good media exposure). Applying our method to increase the probability of using a short term contraceptive, 157 instances were predicted to switch from no contraceptive to short term contraceptive with a higher education, and 18 with change to good media exposure. This example illustrates the great importance of education level for the choice of contraceptive in developing countries.

CONCLUSION AND FUTURE TRENDS

In this chapter we proposed a method to study the influence of the input values on the output scores of a probabilistic classifier. The method has first been defined in a general case valid for any model, and then been detailed for naive Bayes classifier. We also demonstrate the importance of "lever variables", explanatory variables which can conceivably be changed. Our method has first been illustrated on the simple Titanic database in order to show the need to define lever variables. Then, on the PAKDD 2007 challenge databases, a difficult problem of cross-selling, the results obtained show that it is possible to create efficient indicators that could increase sells. Finally we demonstrated the applicability of our method to a government data case, the choice of contraceptive for Indonesian women.

The case study presented on the Titanic dataset illustrates the point of applying the proposed method to accident research. It could be used for example to analyze road accidents or air accidents. In the case of the air accidents any new plane crash is thoroughly analyzed to improve the security of air flights. Despite the increasing number of plane crashes, the relative frequency of those in relation to the volume of traffic is decreasing and air security is globally improving. Analyzing the correlations between the occurrence of a crash and several explanatory variables could lead to a new approach to the prevention of plane crashes.

This type of relationship analysis method has also great potential for medicine applications, in particular to analyze the link between vaccination and mortality. The estimated 50% reduced overall mortality currently associated with influenza vaccination among the elderly is based on studies neither fully taking into account systematic differences between individuals who accept or decline vaccination nor encompassing the entire general population. The proposed method in this paper could find interesting data for infectious diseases research units. Another potential area of application is the analysis of the factors causing a disease, by investigating the link between the occurrence of the disease and the potential factors.

The proposed method is very simple but efficient. It is now implemented in an add-on of the Khiops software™ (see http://www.khiops.com), and its user guide (including how to obtain the software) is available at:
http://perso.rd.francetelecom.fr/lemaire/understanding/Guide.pdf

This tool could be useful for companies or research centers who want to analyze classification results with input values exploration.

REFERENCES

Benitez, J. M., Castro, J. L., & Requena, I. (1997). Are artificial neural networks black boxes? *IEEE Transactions on Neural Networks*, *8*(5), 1156-1164. (September)

Boullé, M. (2006). MODL: a Bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1):131-165.

Boullé, M. (2007). Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research (JMLR)*, *8*, 1659-1685.

Boullé, M. (2008). Khiops: outil de préparation et modélisation des données pour la fouille des grandes bases de données. In *Extraction et gestion des connaissances (EGC)* (p. 229-230).

Breiman, L. (2001). Random forest. *Machine Learning*, *45*. (`stat-www.berkeley.edu/users/breiman/Breiman`)

Brennan, J. J., & Seiford, L. M. (1987). Linear programming and l1 regression: A geometric interpretation. *Computational Statistics & Data Analysis*.

Choudat, D. (2003). Risque, fraction étiologique et probabilité de causalité en cas d'expositions multiples, i : l'approche théorique. *Archives des Maladies Professionnelles et de l'Environnement*, *64*(3), 129-140.

Diagne, G. (2006). *Sélection de variables et méthodes d'interprétation des résultats obtenus par un modèle boite noire*. Unpublished master's thesis, UVSQ-TRIED.

Fawcett, T. (2003). *Roc graphs: Notes and practical considerations for data mining researchers.* Technical Report HPL-2003-4, HP Labs, 2003. Available from `citeseer.ist.psu.edu/fawcett03roc.html`

Féraud, R., Boullé, M., Clérot, F., & Fessant, F. (2008). Vers l'exploitation de grandes masses de données. In *Extraction et gestion des connaissances (EGC)* (p. 241-252).

Féraud, R., & Clérot, F. (2002). A methodology to explain neural network classification. *Neural Networks*, *15*(2), 237-246.

Fern, X. Z., & Brodley, C. (2003). Boosting lazy decision trees. In *International conference on machine learning (ICML)* (p. 178-185).

Framling, K. (1996). *Modélisation et apprentissage des préférences par réseaux de neurones pour l'aide à la décision multicritère*. Unpublished doctoral dissertation, Institut National des Sciences Appliquées de Lyon.

Guyon, I. (2005). *Feature extraction, foundations and applications*. Elsevier.

Guyon, I., Constantin Aliferis, C., & Elisseeff, A. (2007). Computational methods of feature selection. In H. Liu & H. Motoda (Eds.), (p. 63-86). Chapman and Hall/CRC Data Mining and Knowledge Discovery Series. Guyon, I., Saffari, A., Dror, G., & Bumann, J. (2007).

Report on preliminary experiments with data grid models in the agnostic learning vs. prior knowledge challenge. In *International Joint Conference on Neural Networks (IJCNN).*

Han, J. & Kamber M. (2006). *Data mining: concepts and techniques.* Morgan Kaufmann.

Hand, D., & Yu, K. (2001). Idiot's Bayes - not so stupid after all? *International Statistical Review*, *69*(3), 385-399.

Kramer, M. S., Leventhal, J. M., Hutchinson, T. A., & Feinstein, A. R. (1979). An algorithm for the operational assessment of adverse drug reactions. i. background, description, and instructions for use. *Journal of the American Medical Association*, *242*(7), 623-632.

Lemaire, V., & Féraud, R. (2008). Driven forward features selection: a comparative study on neural networks. In *International Joint Conference on Neural Network (IJCNN).*

Lemaire, V., Féraud, R., & Voisine, N. (2006, October). Contact personalization using a score understanding method. In *International Conference On Neural Information Processing (ICONIP).* Hong-Kong.

Leray, P., & Gallinari, P. (1998). *Variable selection* (Tech. Rep. No. ENV4-CT96-0314). University Paris 6.

Lichtsteiner, S., & Schibler, U. (1989). A glycosylated liver-specific transcription factor stimulates transcription of the albumin gene. *CELL.*

Možina, M., Demšar, J., Kattan, M., & Zupan, B. (2004). Nomograms for visualization of naive Bayesian classifier. In *Proceedings of the 8th european conference on principles and practice of knowledge discovery in databases (PAKDD).* (pp. 337–348). New York, USA: Springer-Verlag New York, Inc.

Nakache, J., & Confais, J. (2003). *Statistique explicative appliquée.* TECHNIP.

Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D. S., et al. (2006). Visual explanation of evidence with additive classifiers. In *IAAI.*

Raymer, M. L., Doom, T. E., A., K. L., & Punch, W. L. (2003). Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B.*

Robnik-Sikonja, M., & Kononenko, I. (2008). *Explaining classifications for individual instances.* (to appear in IEEE TKDE)

Robnik-Sikonja, M., Likas, A., Constantinopoulos, C., & Kononenko, I. (2009). *An efficient method for explaining the decisions of the probabilistic RBF classification network.* (currently under review, partially available as TR, `http://lkm.fri.uni-lj.si/rmarko`)

Roy, N., & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th international conf. on machine learning* (p. 441-448). Morgan Kaufmann, San Francisco, CA.

Singh, A., Nowak, R., & Ramanathan, P. (2006). Active learning for adaptive mobile sensing networks. In *Proceedings of the fifth international conference on information processing in sensor networks* (IPSN) (p. 60-68). New York, NY, USA: ACM Press.

Thrun, S. (1995). Extracting rules from artificial neural networks with distributed representations. InM. Press (Ed.), *Advances in neural information processing systems* (NIPS) (Vol. 7).
Cambridge, MA: G. Tesauro, D. Touretzky, T. Leen.

KEY TERMS:

Classifier: a mapping from a (discrete or continuous) feature space X to a discrete set of labels Y.

Probabilist classifier: a classifier with the probability of each label (class) as output.

Exploration: attempt to develop an initial, rough understanding of some phenomenon.

Correlation: the strength and direction of a linear relationship between two variables.

Supervised learning: Supervised learning is a technique for learning a function (a mapping) from training data.

Variable importance: measure of the importance of a variable for the output of a classifier.

Sensibility analysis: analysis of the influence of a change in input variable on the output of the classifier.