A Complete Data Mining process to Manage the QoS of ADSL Services

Françoise Fessant and **Vincent Lemaire**[1]

Abstract. In this paper we explore the interest of computational intelligence tools in the management of the Quality of Service (QoS) for ADSL lines. The paper presents the platform and the mechanisms used for the monitoring of the quality of service of the Orange ADSL network in France. The context is the availability of the VoIP services. In particular, this platform allows the detection and the classification of the unstable lines of the network. The interpretation of the classification results allows the discovery of some new knowledge used to improve the ADSL lines labeling and to prevent inefficient supervision of the network.

1 INTRODUCTION

Internet is now the common platform for voice and video services. The services are delivered through ADSL lines. At the end of the line, at customer's home the connection point between the customer and the internet is in most cases a box. The box is also the mean by which an internet access provider can deliver its services (telephony, internet, television, video over demand).

The quality of the multimedia services given on internet can be affected by various factors dependant of what occurs in the network like congested links, latency, data loss, or dependant of the good working of the platforms delivering the services.

Quality of Service (QoS) is crucial to guaranty that the services will be delivered with good quality to the customers. QoS refers to a broad collection of networking technologies and techniques [12]. The monitoring of the QoS is a way to detect for instance when a process or an element of network are working outside of their working area or are not working properly. The data collection and the observation of some end to end QoS measures is a way to determine the origin of the trouble: element of network, physical line or box. The precise knowledge of the source of the trouble is a key point to an appropriate and rapid reaction of the supervision service to assure a good quality of service.

Orange, the French telecommunication company has implemented a complex chain dedicated to the monitoring of the QoS for its ADSL services in order to increase the satisfaction of its customers. This chain is based on the collection of a large number of end to end measures and on the creation of indicators.

This paper focuses on one specific part of this chain and on the detection of one type of problem that is the identification of unstable ADSL lines. The context of the QoS is restricted to the availability of the telephony on IP (VoIP) service from the Orange box (the live box). We show how the detection of unstable ADSL lines helps to improve the QoS and allows the extraction of additional knowledge to reinforce the global supervision chain.

We will successively present the different parts of the chain from the point of view of a data mining process (from objective and data understanding to modeling and results exploitation). Two steps of the process will be more specifically described: the modeling step and the result interpretation step. The detection of the unstable ADSL lines is performed with a dedicated classifier based on naïve Bayes. Then a deep analysis of the classification results has been performed to better understand which explanatory variables explain the classification results and what are the actions that can be applied to improve the ADSL lines stability but also the labeling process.

2 MANAGEMENT OF THE QoS WITH A COMPLETE DATAMINING PROCESS

Data mining can be defined "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [7]. Several industrial partners have proposed to formalize this process using a methodological guide named CRISP-DM, for CRoss Industry Standard Process for Data Mining [5]. The CRISP-DM model provides an overview of the life cycle of a data mining project, which consists in the following phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. The CRISP-DM model is mainly a process guide for data mining project. The presentation of the QoS processing chain is based on this structure. In our discussion we are specifically interested in the modeling and evaluation phases. The whole process in the QoS context is given on Figure 1.

2.1 Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. In our context the objective is clearly: how to increase the customer satisfaction? In this study, the detection of the unavailability of the services delivered by the live box (VoIP or internet) is the mean chosen to contribute to this objective.

The Orange's service in charge of the QoS management has proposed two kinds of answers. The first one consists in the creation and publication of weekly dashboards to follow some QoS metrics and to be able to react quickly when a degradation is detected. The second one is the automatic classification of ADSL lines (in two main classes: stable or unstable).

[1] Orange labs, Lannion, France, email: francoise.fessant@orange.com

2.2 Data understanding

The data understanding phase starts with an initial data collection and proceeds to an explanatory analysis to get familiar with the data and to identify data quality problems.

The availability of the IP telephony service is watching by a functionality included in the live box. The functionality works when the live box is on. It is able to detect the unavailability of the service but also to raise the reason of this unavailability. It can also perform some measures of vocal quality during calls. The main advantage of this system is that it is located as close as possible of the customer and gives an end to end vision of the service as it is perceived by the customer. This paper is mainly concerned by the VoIP service availability.

Each time an unavailability event is detected the mentioned functionality produces a ticket which is send to a platform collecting and centralizing the whole tickets. To be more precise, the availability ticket is produced twice: the first time when the service is lost and the second time when the service comes back. This last ticket gives the nature of the event that has caused the lost of the service. It is also possible to assess to the duration of unavailability. The content of a ticket concerns only anonymous information about the state of the live box according to the services it has to deliver. In no way private information about the network activity of the customer is seen.

2.3 Data preparation

This phase covers all activities to construct the dataset that will be fed into the modeling tool, from the initial raw or relational data [10, 5].

In the QoS processing chain, the tickets are centralized in an analysis and treatment platform where they are enriched with additional information like the network characteristics of the line and several key point indicators specifying the types of problems and the length of unavailability. The tickets are then fed into a database thanks to a specific application that takes the tickets log and transforms it into a flat table (a table composed of K instances and J variables). A view of the live boxes events for the last 35 days is kept in this database by the mean of all the produced tickets.

2.4 Label of ADSL lines

Orange uses a specific network application to qualify the quality of its ADSL lines. This application is based on the interrogation of the DSLAMs (for Digital Subscriber Line Access Multiplexer), which return the count of the number of resynchronizations for each line and the length of these resynchronizations. This number is strongly correlated to the stability of the line which is, it, directly linked to the availability notion. An inspection period of at least 6 hours is needed before the qualification of the line. The tool is able to label each ADSL line to one of the third following categories: stable, risky or instable.

2.5 Modeling

In this phase, various modeling techniques can be selected and applied, and their parameters have to be calibrated to optimal values according to a success criterion. In the scope of this paper, the detection of the ADSL lines stability is the task to perform. This classification problem will be defined and discussed in-depth in section 3.1.

2.6 Evaluation and result interpretation

Before proceeding to the final deployment of a model, it is important to thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached. The evaluation method and the generalization performance of the classification model are discussed section 3.2.

We propose an interpretation phase in section 4 to better explain the classification results (for example, which explanatory variables explain that an ADSL line has been classified as unstable and how this line can be made less unstable or even become stable). This allows us to extract additional knowledge to reinforce the QoS chain.

2.7 Deployment

The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. The model will be applied to a larger database (the target population) than the training database. In this phase it is important for the user to understand up front what actions will need to be carried out in order to actually make use of the created models. For the QoS application the data collection and preparation phases as well as the modeling phase are currently industrialized. The industrialization of the use of prediction information is still in progress (fine analysis of results, labeling improvement of line state, automatic labeling of tickets for filtering).

3 MODELING: CLASSIFICATION OF ADSL LINES

In this section the modeling step of the data mining process is presented. The model is a classifier designed for the classification of ADSL lines stability. The following notations are introduced: we call T the data table with K instances and J explanatory variables. C is the number of classes. An instance x_k is represented with a vector of J dimensions, f is the probabilistic classifier learned from a modeling table.

3.1 Data and experimental protocol

Five days of tickets have been extracted from the data platform storage for the analysis. These data make the learning database used for the construction of the classification model. The unit we want to analyze is the ADSL line.

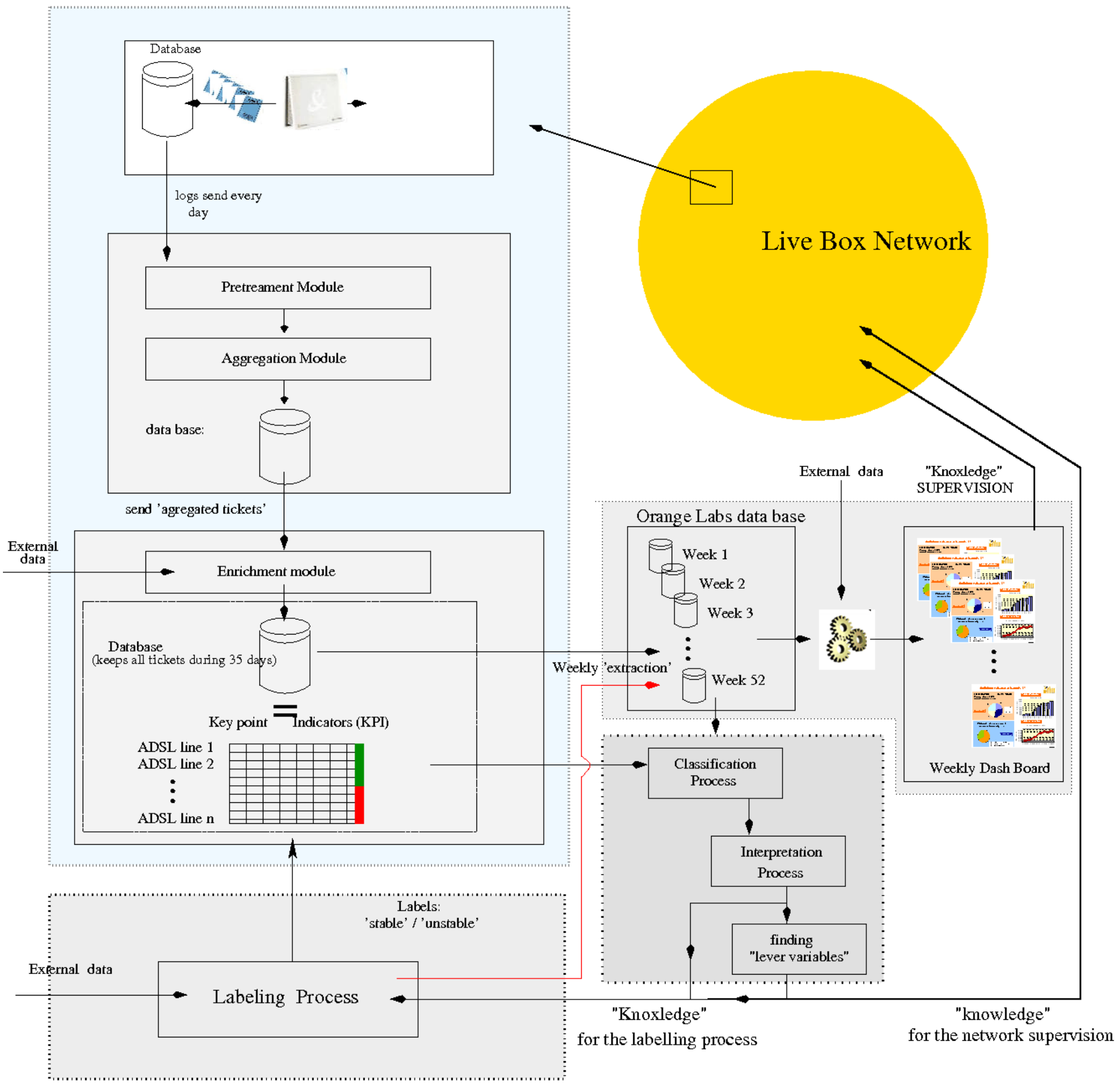

Figure 1. The data mining process: from tickets to QoS indicators

A line is characterized by a set of indicators gathered from the tickets. The indicators are of different sorts. Some of them identify the live box (MAC address of the box, serial number, model and firmware for example). Others concern the line and the network elements the line is connected up (DSLAM model and name, BAS name, etc.). The last group is for the events that have led to the ticket production (total number of tickets, number of tickets of service lost and service return during the period, total duration of events, duration of events of each type, etc.). Five types of unavailability events are detected by the live box: reboots, desynchronizations of the ADSL line, lost of session PPP (Point to Point Protocol), ToIP connection problem (Telephony over IP), ToIP configuration lost at the live box level.

Finally a line is represented by a vector of 39 explanatory variables that are categorical or continuous. We add to these descriptors the information about the state of the line coming from the network application as described above. The modeling database holds 71164 instances. The priors on the classes stable, risky and unstable are respectively of 0.881, 0.054 and 0.064.

Experiments have been made with the Khiops[2] software (developed by Orange Labs). Khiops is a data mining tool allowing to automatically building successful classification models. Khiops offers an optimal preprocessing of the input data, efficient selection of the variables and averaging the models. It is a parameter-free classification method that exploits the naive Bayes assumption. Khiops operates in two steps. In the first step consisting in preparing data it estimates the univariate conditional probabilities using the MODL (Minimum Optimized Description Length) method, with Bayes optimal discretizations and value groupings for numerical and categorical variables [1, 2]. In the second step, the modeling step, it searches for a subset of variables consistent with the naive Bayes assumption, using an evaluation based on a Bayesian model selection approach and efficient add-drop greedy heuristics [3]. Finally, it combines all the evaluated models using a compression-based averaging schema [4].

This approach also quantitatively evaluates the predictive importance of each variable for the target. At the end of the preparation step (preprocessing of the input data) Khiops returns a value (the level) that is directly the predictive importance of the

[2] www.khiops.com

explicative variable for the target. A level of zero means that the variable has been discretized in one single interval (for a numerical variable) or the modes are in one single group (for a symbolic variable). The variable is not informative for the modelisation and thus can be reliably discarded.

3.2 Classification results

A k-fold cross validation procedure (with *k=10*) has been used. The performance of every model is computed on the fold which has not been used to train the classifier. The ten 'test' results are then combined to give an estimation of the generalization error of the model. The folds used to do the training of the initial classifier do not cross the test set. Note that the classification model Khiops is robust to unbalanced classes. There is no need to sample the data to balance the classes.

The predictive model is evaluated using the accuracy on classification (ACC) and the area under the ROC curve (AUC) [6] (the higher the criteria, the better, with 1.00 indicating perfect performance). The numerical results for the two criteria are respectively *0.8924 +/- 0.0017 and 0.8185 +/- 0.0078* that are very good results.

Another way to present the performance of a classification model is the cumulative gain curve. It is a graphical representation of the advantage of using a predictive model to choose which unstable lines to select. The x-axis gives the proportion of the lines with the best probability to correspond to the target (unstable class), according to the model. The y-axis gives the percentage of the targeted lines reached. The curves are plotted on Figure 2. The diagonal represents the performance of a random model. If we target 20% of the lines with the random model, we are able to reach 20% of the unstable lines. With the current model, when 20% of the lines are observed, 77% of the unstable lines are reached.

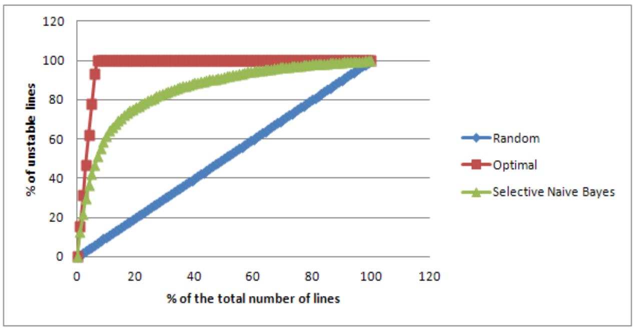

Figure 2. The lift curve

A deep analysis of the classification errors has shown that main classification errors come from the wrongly assignment of the label stable to a line that had been found unstable by the network application. This difference is coherent with the observation made by the network service in charge of the labeling of ADSL lines that has detected a problem in the labeling of some lines (with the label unstable while they are not). This happens in some DSLAMs with specific cards that are unable to distinguish modem extinctions from desynchronizations due to ADSL transmission. The consequence is the increase of the count of desynchronizations used by the network application to determine the state of a line and so to wrongly label a line as unstable.

Khiops allows further interpretation of the numerical results. As said earlier, the tool analyses each variable independently for the target in the preparation step and return a value that is directly its predictive importance (the level) for the classification model. The table 1 contains the 5 most informative variables found by Khiops. 31 variables have been found useful for the model (with a level >0).

Table 1. The variables ranked by level of importance.

rank	Variable name	level
1	nb_tickets_306_1	0.172868
2	nb_tickets_308_1	0.172453
3	nb_tickets_C	0.151913
4	nb_tickets	0.150877
5	nb_tickets_308	0.14135

306_1 is the code for a lost of service for an event of type desynchronization, 308_1 is the code for a return of service for the same type of event.

The observation of the more informative variables confirms that the stability notion is strongly correlated to the information of ADSL line desynchronization. This is perfectly consistent with the lines labeling method used by the network application whereas the view given by the software agent in the live box is only a service view completely independent of the network view.

4 EVALUATION AND RESULT INTERPRETATION

This section is dedicated to the evaluation and interpretation steps. We propose a method to *(i)* identify the importance of the explanatory variables for every ADSL line in the database (and not "in average" for all the examples) and *(ii)* propose an action in order to change the probability of the desired class. The method is completely automatic. It is based on the analysis of the link between the probabilities at the output of the classifier and the values of the explanatory variables at the input.

4.1 Individual importance of explanatory variables

For a classifier as the naive Bayes we choose the method called Weight of Evidence (WoE) described in [9] and [11] as the importance measure. This indicator measures the log of the odds ratio. It is computed for all the explanatory variables at the input of the classifier and for one specific class. The class of interest *(q)* is generally the predicted class for the instance x_k. A variable with a positive importance (WoE) contributes positively to define the predicted class. At the opposite, a variable with a negative importance (WoE) contributes negatively to define the predicted class (and so positively contributes to define another class of the classification problem).

This point is illustrated Figure 3 with 3 ADSL lines. The first line is classified as stable, the second is classified as risky and the third is classified by the model as unstable. The x-axis represents simply the index of the 31 input variables (with level >0). The figure (which can be seen as a nomogram [9]) indicates that:

- the line predicted as stable is characterized by number of little positive contributions (positive importances) and some negative contributions,

- the line predicted as risky is characterized by number of little positive contributions (positive importances) and hardly any negative contributions
- the line predicted as unstable is characterized by some heavy positive contributions.

So we have a completely individual interpretation for each ADSL line allowing a precise diagnosis. For each line we can have:

- the class predicted by the classifier,
- a confidence score on this prediction,
- the importance value for each explanatory variable.

Arguments for the choice of the method and details about the algorithm are given in [8].

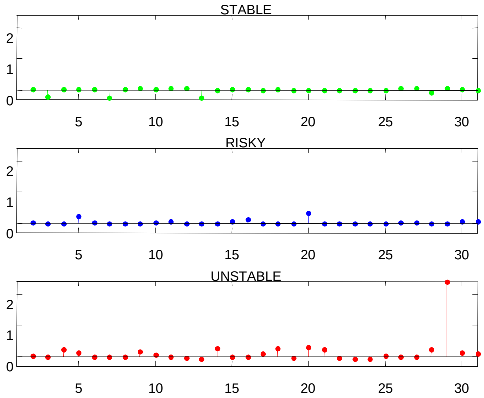

Figure 3. Examples of explanatory variables importances for 3 ADSL lines (from top to bottom) for a line predicted as stable, a line predicted as risky and a line predicted as unstable.

4.2 Indexes for the improvement of the line stability

In this section a method that study the influence of the input values on the output scores of the naïve based classifier is exploited. The goal is to increase the predictive probability of a given class by exploring the possible values of the input variables taken independently. Then it became possible to act on some variables, the lever variables that are defined as the explanatory variables for which it is conceivable to modify their value to induce changes of occurrence of the desired class.

We use the methodology described in [8]. Given C_z the target class among the C target classes (for example here the class unstable line). Given f_z the function which models the predicted probability of the target class $f_z(X = x) = P(C_z|X = x)$. Given the equality of the vector X of the J explanatory variables to a given vector x of J values. The method proposed here tries to increase the value of $P(C_z|X = x_k)$ successively for each of the K examples of the considered data base. $P(C_z|X = x_k)$ is the natural value of the model output. The idea is to modify the values of some explanatory variables (the lever variables) in order to study the variation of the probabilist classifier output for the considered example. The values of the lever variables are modified to see the variations in the posterior probabilities. This is done by covering some interval for continuous variables and by trying the possible values for discrete ones (after the preparation step where variables have been discretized or grouped).

The method is illustrated with the explanatory variable "live box firmware" as lever variable. This symbolic variable can take 4 modalities that will be called *A*, *B*, *C* and *D* for confidential reasons. These four groups are preserved after the first step of pretreatment for the classifier construction.

We focus on the ADSL lines labeled unstable and effectively predicted unstable by the classifier (1611 lines). For 1321 of these lines there exists a value that can increase the probability of the class stable. To do that the variable "live box firmware" has to take the value *D.* For the other 380 ADSL lines the probability of stability cannot be improved (the value of the variable is already *D*).

The improvements of the probability are presented Figure 4. We have, distributed on the x-axis, the 3 modalities of the variable "live box firmware" that can be changed. Then within each modality the values of $(PCa(x_k)\text{-}PCi(x_k))$ are arranged by ascending order (with (Ca) the value which leads to the improvement, (PCa) the associated improved probability and (PCi) the initial probability). To be more precise we have for x_k in [1:407] LBF='A', for x_k in [408:778] LBF='B' and for x_k in [779:1231] LBF='C'. The blue points coincide to an improvement without class change. The red squares coincide to an improvement with class change. We can conclude that the variable "live box firmware" is really a lever variable. When its value is changed to *D*, it allows obtaining more stable lines (from 1231 cases on 1611) or even obtaining stable lines (from 51 lines, Figure 4).

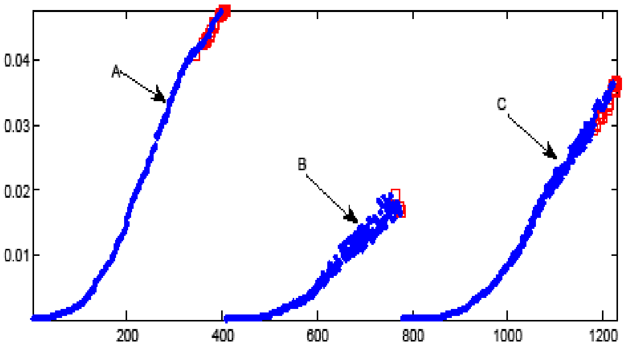

Figure 4. Possible improvement of $(PCa(x_k)\text{-}PCi(x_k))$ of the stability for the 1280 ADSL lines x_k .

The exploration of correlations has been made ADSL line by ADSL line. This exploration offers some means to improve their stability. The knowledge extracted at the end of the exploration step can be listed with the four following points for each ADSL line labeled as instable:

- the initial probability of instability *PCi(.)*
- the improved probability of instability (reduced) *PCa(.)*
- the explanatory variable that allows the gain
- the value that has to take this explorative variable to reach the gain.

5 DISCUSSION

We have presented a complete process for the supervision of VoIP services. Into this process, we have made a focus on a specific point that is the detection of the stability of the ADSL lines. A dedicated classifier based on naïve Bayes has been defined. The model used to classify the ADSL lines according to their stability led to very good classification performances validating the data extraction and data creation steps. The analysis of the most important variables has shown a strong correlation between the stability state of the ADSL line and the information about the desynchronization of the line (the two most informative variables for the target are the number of tickets of lost of service and of return of service for the event of type desynchronization). This result is totally coherent with the network process that is currently used to label the state of ADSL lines.

It appears from this study two ways of using the classification model in an operational way:

It can be used as a filter of the unstable lines, to keep only the stables lines. In this way a large number of tickets, produced mainly because of the line instability are discarded.

Another use could be the reinforcement of the knowledge used by the network tool to label the ADSL lines and in the end the improvement of this labeling. Indeed, we know that the labeling process is imperfect for some model of DSLAM that are unable to distinguish electric start/stop from desynchronizations. As the stability of the ADSL lines is directly correlated to the number of desynchronizations, this involves wrongly unstable labels for some ADSL lines. On the other side, the live box counts the number of desynchronizations in a way that is totally independent of those of the DSLAM. So, the exploitation of the number of tickets of this type could help to make more precise the counters from which is based the stability decision.

The various steps of the data mining process are for the most part industrialized (around 75%): from the data acquisition to the stability ADSL lines prediction. The exploitation of the prediction information is under industrialization.

The interpretation step of the classification results produces an interpretation totally individual of each ADSL line allowing a very precise diagnostic. The exploration of correlation step can give some means to improve the stability of the lines and so intervention plans can be designed.

REFERENCES

[1] M. Boullé, 'A bayes optimal approach for partitioning the values of categorical attributes', *Journal of Machine Learning Research*, **6**, 1431–1452, (2005).

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

[3] M. Boullé, 'Compression based averaging of selective naïve bayes classifiers', *Journal of Machine Learning Research*, **8**, 1659–1685, (2007).

[4] M. Boullé. 'Regularization and Averaging of the Selective Naive Bayes Classifier'. In *IJCNN*, 2989-2997, (2006).

[5] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C Shearer and R. Wirth. *CRISP-DM 1.0: step-by-step data mining guide*, 2000**.**

[6] T. Fawcett, *ROC graphs: Notes and practical considerations for researchers*, Technical Report HPL-2003-4, HP Laboratories, 2003.

[7] U.M. Fayad, G. Piatetsky-Shapiro and P. Smyth, *Advances in Knowledge Discovery and data Mining*, 1–34, Chapter From Data Mining to Knowledge Discovery: An Overview, AAAI/MIT Press, 1996.

[8] V. Lemaire, C. Hue and O. Bernier, *Correlation Analysis in Classifiers*, 204–218, Chapter Data Mining in Public and Private Sectors: Organizational and Government Applications, IGI Global, 2010.

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

[10] D. Pyle, *Data Preparation for Data Mining*, Morgan Kaufman, 1999.

[11] M. Robnik-Sikonja and I. Kononenko, 'Explaining classifications for individual instances'. *IEEE TKDE*, **20** (5), 589–600, (2008).

[12] Z. Wang, *Internet QoS: Architectures and Mechanisms for Quality of Service*, The Morgan Kaufman Series in Networking, San Francisco, 2001.