Gestion de la QoS des services ADSL à l'aide d'un processus de data mining

Vincent Lemaire*, Françoise Fessant *

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion

Résumé. Dans cet article l'intérêt d'une approche "fouille de données" est explorée dans le cadre du contrôle de la qualité de service (QoS) des lignes ADSL du réseau de France Télécom. Cet article présente la plateforme et les mécanismes qui ont été mis en place. Ces derniers permettent la détection et la classification des lignes bruitées au sein du réseau. L'utilisation de la classification permet une amélioration de la QoS. L'interprétation de la classification permet la découverte de connaissances actionnables.

1 Introduction

Internet est devenu la plateforme pour les services de voix et de vidéo. Toutefois la qualité des services multimédia offerts sur Internet dépend des congestions, indisponibilités, et autres anomalies survenant dans le réseau. La mesure de la qualité de service (QoS) permet de détecter quand un processus ou un élément du réseau opère en dehors de sa plage de fonctionnement ou ne fonctionne pas correctement. La collecte de mesures de QoS de bout en bout permet, en général, de déterminer la source du problème : problème lié à un élément du réseau (DSLAM (Digital Subscriber Line Access Multiplexor)...), problème lié à la ligne physique, problème lié à la Box (LiveBox^{TM}dans cet article). La connaissance de la source du problème doit permettre de réagir promptement et de manière adéquate.

Orange a mis en place un suivi de la QoS pour ses services ADSL de manière à améliorer la satisfaction de ses clients. Cet article n'a pas l'ambition de décrire l'ensemble de ce processus. Il se concentre sur la détection d'un type particulier de problème : les lignes ADSL "bruitées" et montre en quoi cette classification permet une amélioration de la QoS et l'extraction de nouvelles connaissances. Dans l'application qui est décrite, l'élément de QoS considéré est lié à la disponibilité du service de téléphonie sur IP (VOIP) par le biais des LiveBox (LB). L'ensemble du processus de data mining (Fayyad et al., 1996) qui a été mis en place est présenté Figure 1.

2 Collecte et préparation des données

Provenance des données : La sonde Audiphone est un agent logiciel embarqué dans les Live Box qui, dès que la LB est allumée, surveille la disponibilité du service de téléphonie sur IP et peut également effectuer des mesures de la qualité vocale durant les appels. L'avantage essentiel de l'agent Audiphone est qu'il est positionné au plus près de l'usager. Il est capable de détecter l'indisponibilité du service et de déterminer la cause de cette indisponibilité telle

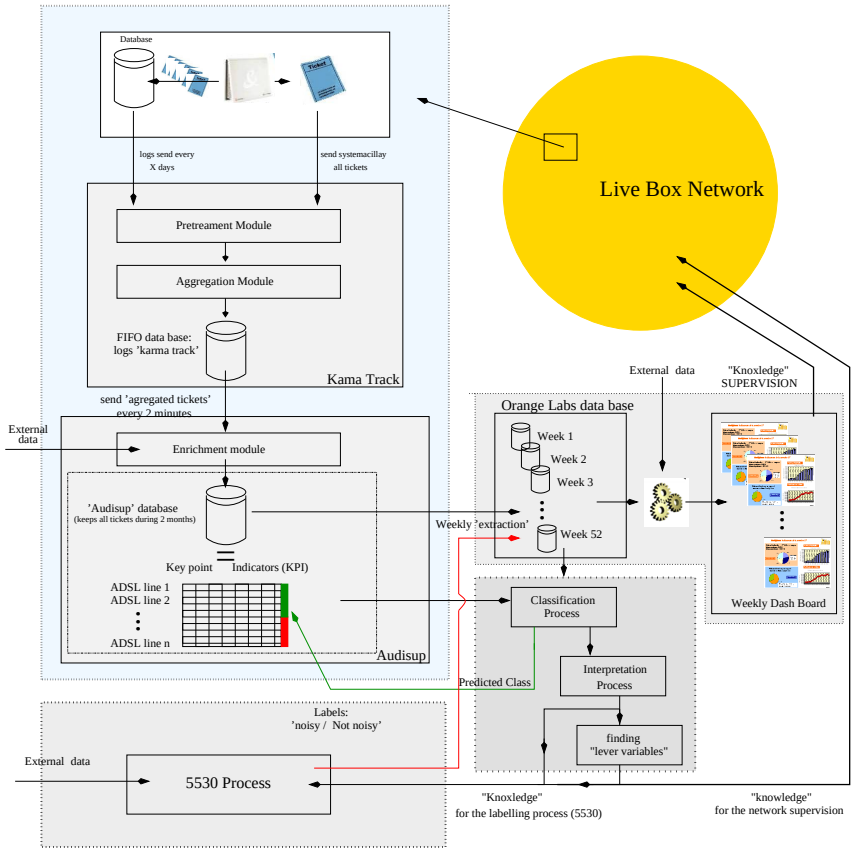

FIG. 1 – *Le processus complet de data mining : des tickets Audiphone aux indicateurs de QoS*

que vue par la livebox en extrémité de la chaîne de transmission. Il permet alors une vision du service tel qu'il est perçu par l'usager. Dans cet article nous nous intéressons plus spécifiquement à la disponibilité du service. A chaque évènement d'indisponibilité, l'agent logiciel génère un ticket qui est envoyé ensuite vers une chaîne de collecte pour stockage et traitement. Il est important de préciser que le contenu des tickets ne viole en rien la vie privée de l'utilisateur de la LB. Le contenu d'un ticket ne comporte que des informations anonymes sur la qualité et la conformité de la LB vis-à-vis des services qu'elle doit rendre.

Mise en base des données : Les tickets sont remontés vers une plateforme d'analyse et de traitement (KarmaTrack-Audisup) qui réalise l'enrichissement des tickets à l'aide d'informations réseau et le calcul de Key Point Indicators (KPI), sur différents axes tels que les types de problèmes et la durée d'indisponibilité. Les tickets sont ensuite mis en base au sein d'une application nommée 'Audisup'. Audisup traite les logs de tickets pour produire une base de données "à plat" (un tableau de N instances représentées par J variables). Cette base de données contient l'ensemble des tickets émis par les LB au cours des 35 derniers jours. Un ticket

contient différents champs. Certains de ces champs sont de type 'variable continue' d'autres sont des champs à valeurs catégorielles.

Etiquetage des lignes ADSL : L'application 5530 (Le Meur et Santos-Ruiz, 2010), développée par Alcatel, permet d'observer la qualité des lignes ADSL, par la collecte de paramètres représentatifs de leur état, au travers de l'interrogation de DSLAM télé surveillés. Le 5530 est un moyen de service après vente (SAV) supplémentaire et de gestion de QoS permettant de fiabiliser à distance le diagnostic de signalisations complexes sur l'ADSL et le multiservice. La notion de stabilité (notion extrêmement corrélée à la notion d'indisponibilité) au sein du 5530 est évaluée à l'aide de deux compteurs : le MTBR (Mean Time Between Retrains) et le MTBE (Mean Time Between Errors) basés essentiellement sur les nombres et durées de resynchronisations de la ligne. En fonction du MTBR et des MTBE, le 5530 classe les lignes en 3 catégories : Stable ; Risquée ; Instable (Le Meur et Santos-Ruiz, 2010). Il est à noter que pour connaître la stabilité de la ligne à l'aide du 5530, il faut que l'inspection ait une durée de plus de six heures de synchronisation entre le modem client et le modem DSLAM.

3 Modélisation : Classification des lignes ADSL

Notations utilisées par la suite : (*) une table de modélisation T contenant K instances et J variables explicatives ; (*) un problème de classification à C classes ; (*) un classifieur probabiliste f entraîné sur la table de modélisation ; (*) une instance x_k représentée sous la forme d'un vecteur à J dimensions.

3.1 Données et protocole expérimental

On dispose pour l'analyse de la classification des lignes ADSL d'une extraction des tickets remontés par la plateforme Audisup sur une période de 5 jours. Tous les tickets associés à une même date journalière sont stockés dans le même fichier. Chaque ligne ADSL est caractérisée par différents types d'indicateurs : les paramètres d'identification de la LB et du réseau (fixes quel que soit le jour considéré) et les paramètres qui concernent l'évènement d'indisponibilité (variables selon le jour de l'évènement). Une ligne ADSL est décrite par 123 indicateurs (variables explicatives). A ces indicateurs on rajoute l'information sur la caractéristique de la ligne "bruitée/non bruitée" provenant du service 5530. La base de modélisation contient 71164 lignes ADSL. Les priors sur les classe 'Stable', 'Risquée et 'Instable' sont respectivement de 0.881, 0.054, 0.064. Les expérimentations de classification ont été réalisées à l'aide du logiciel Khiops (développé par Orange Labs, `www.khiops.com`). Le classifieur utilisé est un classifieur naif de Bayes moyenné (Boullé, 2007).

3.2 Résultats

On a utilisé une procédure de validation croisée par k-folds (avec k=10) pour produire les résultats. Les performances en taux de bonne classification (ACC) et d'AUC sont respectivement de 0.8924 +/- 0.0017 et de 0.8185 +/- 0.0078 ; ce qui en font de très bons résultats. Les erreurs de classification proviennent majoritairement de l'affectation à tort de l'étiquette stable à une ligne jugée instable par l'application 5530. Ce résultat est cohérent avec le service 5530 qui a identifié un problème d'étiquetage pour certaines lignes étiquetées instables qui le sont à

tort. En effet certaines cartes implémentées dans les DSLAM ne permettent pas de distinguer les extinctions de modem des resynchronisations liées à la transmission ADSL. Ceci a pour conséquence d'augmenter le compteur de resynchronisations utilisé par le 5530 pour déterminer l'état d'une ligne et donc d'aboutir à l'évaluation d'une ligne instable à tort. La courbe de lift obtenue est présentée Figure 2.

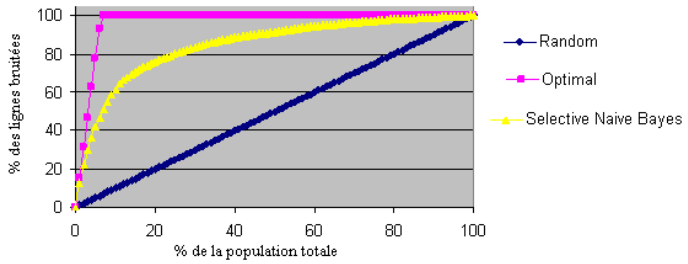

FIG. 2 – *Courbe de lift*

4 Evaluation et interprétation des résultats

4.1 Importances individuelles des variables explicatives

On propose de calculer l'importance individuelle des variables explicatives pour chaque ligne ADSL (et non en "moyenne"). On utilise dans cette section une méthode de calcul, G, permettant, connaissant f (le classifieur utilisé section 3), de calculer l'importance d'une variable en entrée du classifieur. Etant donné l'état de l'art et le type de classifieur utilisé (un classifieur naif de Bayes) on choisit d'utiliser comme mesure d'importance le "Weight of Evidence" (WoE) décrit dans (Robnik-Sikonja et Kononenko, 2008). L'indicateur WoE mesure le log d'un odds ratio. Il est calculé pour l'ensemble des variables explicatives présentes en entrée du classifieur et pour une classe d'intérêt. La classe d'intérêt (q) est en général la classe d'appartenance prédite de l'instance x_k. Une variable qui a une importance (WoE) positive contribue positivement à définir la classe prédite, à l'inverse une variable qui a une importance (WoE) négative contribue négativement à définir la classe prédite (donc contribue positivement à définir une autre classe du problème de classification).

La figure 3 illustre ce point avec 3 lignes ADSL : l'une classée par le classifieur comme 'STABLE', la seconde classée 'RISKY' et la troisième comme 'UNSTABLE'. On remarque que : (i) la ligne classée 'Stable' est caractérisée par une multitude de petites contributions positives (importances positives) et quelques contributions négatives ; (ii) la ligne classée 'Risky' est caractérisée par une multitude de petites contributions positives (importances positives) et presque aucunes contributions négatives ; (iii) la ligne classée 'Unstable' par quelques fortes contributions positives. On a donc une interprétation complètement individuelle pour chaque ligne ADSL, permettant un diagnostic précis. Pour chaque ligne on a : (i) la classe prédite par le classifieur ; (ii) un score de confiance sur cette prédiction ; un ordonnancement des variables explicatives en fonction de leur importance ; (iii) la valeur de l'importance pour chaque variable explicative.

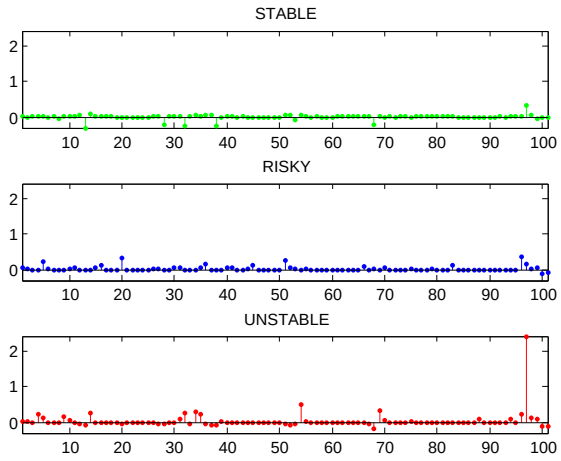

FIG. 3 – *Exemple d'importance des variables explicatives pour 3 lignes ADSL (de haut en bas) pour une ligne classée stable, une ligne classée risky et une ligne classée unstable.*

4.2 Indices d'amélioration de la stabilité des lignes

On utilise ici la méthodologie décrite dans (Lemaire et Hue, 2010). Soit C_z la classe cible d'intérêt parmi les C classes cible, par exemple ici, la classe "ligne bruitée". Soit f_z la fonction qui modélise la probabilité d'occurrence de cette classe cible $f_z(X = x) = P(C_z|X = x)$ étant donné l'égalité du vecteur X des J variables explicatives à un vecteur donné x de J valeurs. La méthode proposée ici recherche à augmenter la valeur de $P(C_z|X = x_k)$ pour chacun des K exemples de la base de données. Parmi les variables disponibles en entrée du classifieur, on exclura de l'exploration celles pour lesquelles les valeurs ne peuvent pas être modifiées. On conserve les variables dites levier, c'est à dire celles sur lesquelles on pense pouvoir agir.

On utilise, à titre d'exemple pour l'article, la variable 'LB_firmware' (LBF) comme variable levier. D'après la table de modélisation (la table ayant servi à créer le classifieur) cette variable peut prendre 4 modalités différentes que l'ont nommera A, B, C, D pour des raisons de confidentialité. L'étape de prétraitement des variables catégorielles (groupage de modalités (Boullé, 2005)) qui est la première étape lors la construction du classifieur naif de Bayes a déterminé que ces 4 groupes étaient optimaux (pas de création d'un groupe contenant plusieurs de ces modalités).

On s'intéresse ensuite plus précisément aux lignes ADSL 'Unstable' et effectivement prédites comme 'Unstable' par le classifieur (soit 1611 lignes ADSL). 1231 d'entre elles peuvent voir leur probabilité d'être stable augmenter. Pour celà la variable LBF doit prendre comme valeur 'D'. La probabilité de stabilité des 380 autres lignes ADSL ne peut pas être améliorée (la variable LBF est déjà à 'D'). On présente Figure 4 l'amélioration de la probabilité. Dans cette figure l'axe des abscisses coincide avec une modalité de LBF, puis au sein d'une modalité de LBF les valeurs de $PCa(x_k) - PCi(x_k)$ ont été ordonnées de manière croissante. On a (i) pour $x_k \in [1 : 407]$ LBF='A' ; (ii) pour $x_k \in [408 : 778]$ LBF='B' ; (iii) pour $x_k \in [779 : 1231]$ LBF='C'. Les points bleus montrent une amélioration sans changement de classe. Les points

rouges montrent une amélioration avec changement de classe. On en conclut que la variable 'L_firmware' est effectivement une variable levier. Elle permet lorsque l'on force sa valeur à 'D' d'obtenir des lignes plus stables (1231 cas sur 1611) voir d'obtenir des lignes stables (les 51 carrés rouges dans la figure 4).

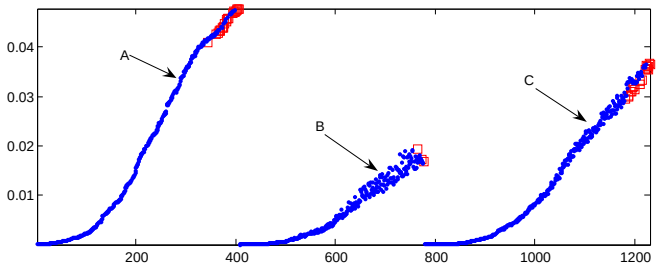

FIG. 4 – *Amélioration possible* $(PCa(x_k) - PCi(x_k))$ *de la stabilité pour les 1280 lignes ADSL* (x_k)*.*

On a donc une exploration des corrélations ligne ADSL par ligne ADSL permettant de rechercher des moyens d'améliorer leur stabilité. Pour chaque ligne classée instable on a : (i) la probabilité initiale d'instabilité $(PCi(.))$; (ii) la probabilité améliorée (diminuée) d'instabilité $(PCa(.))$; (iii) la variable explicative qui permet le gain ; (iv) la valeur que doit prendre cette variable explicative pour obtenir le gain.

5 Discussion

De bonnes performances de classification ont été obtenues, validant le processus d'extraction et de création des données. L'analyse des variables les plus importantes montre que l'information de ligne bruitée est très corrélée à l'information de désynchronisation de la ligne ADSL (les deux variables les plus informatives pour la cible sont les tickets de perte de service de type 306_1 et de retour de service 308_1 qui comptabilisent le nombre de désynchronisations de la ligne). Ce qui est complètement cohérent avec la manière dont l'application 5530 étiquette l'état d'une ligne.

Deux manières d'utiliser le modèle de classification se dégagent de l'étude.

Une première piste pour l'utilisation du modèle en mode opérationnel est le filtrage des lignes instables pour ne garder que les lignes stables. De cette manière on filtre un grand nombre de tickets qui sont produits parce que la ligne est instable.

Une autre utilisation possible serait de renforcer la connaissance du 5530 pour améliorer l'étiquetage des lignes. En effet, un certain nombre de lignes étiquetées instables le sont à tort. Certains types de DSLAM remontent mal les compteurs des nombres de resynchronisations journaliers sur lesquels se base l'application 5530 pour en déduire l'état de stabilité d'une ligne. Ces DSLAM ne font pas la différence entre les resynchronisations et les extinctions/allumages électriques. Par exemple, il suffit que le client laisse son modem allumé moins de 4 heures dans une journée pour être déclaré instable à coup sur. Toutes les lignes pour lesquelles le client éteint son modem la nuit et pendant la journée de travail sont déclarées instables à tort. La sonde Audiphone remonte les désynchronisations de la ligne ADSL de manière indépendante des

DSLAM et le compte des tickets de ce type de remonté pourrait aider à préciser les compteurs alimentant la décision de stabilité.

Dans le cadre de l'étude qui a été menée la 'classification' des lignes ADSL les trois quarts du processus a été entièrement industrialisé : de la collecte des données à la prédiction des lignes ADSL bruitées. L'information de prédiction est elle en cours d'industrialisation : analyse précise des résultats et utilisation dans la compréhension du phénomène, exploitation des résultats dans l'amélioration du service 5530 et enfin étiquetage automatique des tickets à des fins de filtrage.

La phase d'interprétation des résultats de classification permet quant à elle d'obtenir une interprétation complètement individuelle d'une ligne ADSL autorisant un diagnostic précis. La phase d'exploration des corrélations existantes au sens du classifieur permet de rechercher des moyens d'améliorer la stabilité des lignes et ainsi d'avoir des propositions de plan d'intervention.

Références

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

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

Fayyad, U. M., G. Piatetsky-Shapiro, et P. Smyth (1996). *Advances in Knowledge Discovery and Data Mining*, Chapter From data mining to knowledge discovery : An overview., pp. 1–34. AAAI/MIT Press.

Le Meur, R. et L. Santos-Ruiz (2010). Evaluation de la stabilité d'une ligne ADSL par le 5530. Technical report, France Telecom Research and Development.

Lemaire, V. et C. Hue (2010). *Correlation Analysis in Classifiers*, Chapter From data mining to knowledge discovery : An overview., pp. 1–34.

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

Summary

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 mechanism used to monitoring the quality of service of the Orange ADSL network in France. This platform allows the detection and the classification of the noisy lines in the network. The interpretation of results given by the classification process allows the discovery of a knowledge used to improve the process which labels the lines (noisy / not noisy) and to prevent inefficient supervision of the network.