KAWAB
Un outil pour explorer les corrélations existantes dans un classifieur naïf de Bayes

Vincent LEMAIRE
Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
vincent.lemaire@orange-ftgroup.com, http://perso.rd.francetelecom.fr/lemaire/

Résumé : Cette démonstration présente un outil qui analyse le lien qui existe entre les probabilités produites par un classifieur et les valeurs des variables explicatives positionnées en entrées de ce classifieur. Le but de cette analyse est d'accroitre la probabilité d'apparition d'une classe d'intérêt en explorant l'ensemble des valeurs possibles des variables explicatives; variables explicatives étudiées indépendamment les unes des autres.

Mots Clefs : Exploration, Corrélation, Classifier

1 Introduction

Etant donné une base de données, une tache habituelle, de fouille de données, est de trouver la relation, la corrélation, existante entre d'une part un ensemble de variables explicatives et d'autre part une variable à expliquer, la variable cible. Ce processus d'extraction de connaissance débouche souvent sur la construction d'un modèle qui "représentera" cette relation [1]. Par exemple, face à un problème de classification, un classifieur probabiliste permet, pour toutes les instances de la base de données, d'estimer la probabilité d'apparition d'une classe, étant donné les valeurs des variables explicatives propre à chaque instance.

Ces probabilités peuvent être utilisées pour évaluer une politique existante au sein d'entreprises ou d'organismes gouvernementaux. Mais ces probabilités ne sont pas toujours exploitables car elles ne donnent aucune indication sur les actions à entreprendre pour essayer de modifier leur évaluation. Comment changer la probabilité de décéder suite à la contraction d'une hépatite ?

L'outil présenté dans cette démonstration propose d'explorer la relation existante au sein du modèle de classification entre les variables explicatives, prise une à une et indépendamment, et la sortie du modèle : une classe d'intérêt. Cette exploration permet de trouver des actions qui permettent de changer la probabilité d'apparition de la classe d'intérêt.

2 Exploration des corrélations – Cas général

- Soit x une instance représentée sous la forme d'un vecteur de J variables explicatives (un vecteur de J valeurs).
- Soit C_z la classe d'intérêt parmi T classes.
- Soit f_z la fonction (le classifieur) qui associe à chaque instance (X) la probabilité d'apparition de la classe d'intérêt, tel que $f_z(X=x) = P(C_z \mid X=x)$.
- Soit v_{jn} le nombre n de valeurs différentes d'une variable X_j.
- Soit ξ_j l'ensemble des valeurs différentes de la variable j, ensemble de valeurs observé sur l'ensemble d'apprentissage utilisé lors de la construction du modèle.
- Soit $P(C_z \mid X=x_k)$ la valeur "naturelle" en sortie du modèle pour une instance x_k et une classe d'intérêt z;
- Soit $P_j(C_z \mid X=x_k, b)$ la sortie du modèle étant donné l'instance x_k mais pour lequel la valeur de variable j a été remplacée par la valeur b. Par exemple, si c'est la valeur de la troisième variable explicative qui est modifiée: $P_3(C_z \mid x_k, b) = f_z(x_k^1, x_k^2, b, x_k^4, x_k^5)$.

L'algorithme implémenté dans Kawab essaye d'accroitre la valeur de $P(C_z \mid X = x_k)$ en utilisant l'ensemble des valeurs possibles de chacune des J variables explicatives (ξ_j). Cette démarche est réalisée successivement pour chacune des K instances de la base de données considérée.

En scrutant chaque variable explicative, présente en entrée du modèle, une exploration des valeurs "potentielles" de la sortie du modèle est réalisée pour chaque instance.

Cet exploration peut permettre d'observer 3 types de valeurs au sein des valeurs des variables explicatives : des valeurs qui ne permettent pas d'améliorer la probabilité d'apparition de la classe d'intérêt ($P_j(C_z \mid X=x_k) < P(C_z \mid X=x_k)$); des valeurs qui permettent d'améliorer la probabilité d'apparition de la classe d'intérêt ($P_j(C_z \mid X=x_k) > P(C_z \mid X=x_k)$) mais sans que l'instance x_k change de classe; des valeurs qui permettent d'améliorer la probabilité d'apparition de la classe d'intérêt et de plus x_k change de classe.

Lorsque la modification de la valeur "naturelle" d'une variable explicative produit une amélioration de la probabilité d'apparition de la classe d'intérêt, 3 informations sont alors mémorisées : (i) la valeur de cette variable qui produit l'amélioration (Ca); (ii) la valeur de la

probabilité associée (*PCa*); and (iii) la variable explicative associée à cette amélioration.

Cet algorithme est présenté en détails et appliqué à différents problèmes dans [2][3].

3 Démonstration

3.1 Contexte

Kawab est un logiciel (voir Figure 1) qui implémente l'algorithme d'exploration décrit ci-dessus (plus d'autres fonctionnalités non décrites ici). Kawab a besoin d'un classifieur pour l'application de cet algorithme. Ce classifieur est à ce jour exclusivement un classifieur naïf de Bayes généré par le logiciel Khiops.

Khiops est un outil de préparation et de modélisation de données qui produit un modèle de classification supervisé, un classifieur naïf de Bayes. Ce classifieur est notamment construit à l'aide de techniques de sélection de variables et de moyennage de modèles, techniques qui lui confèrent de très bonnes performances. Ce logiciel peut être téléchargé sur http://www.khiops.com.

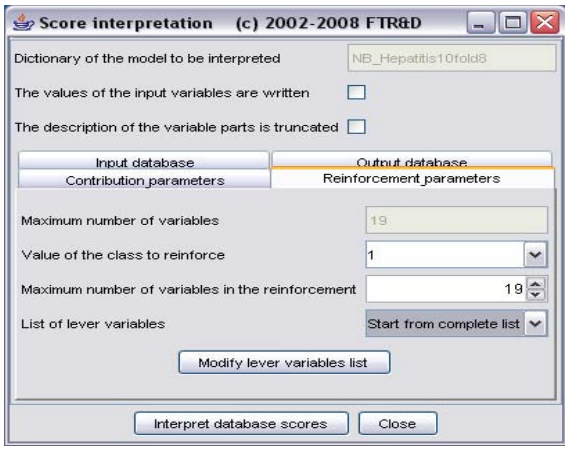

Figure 1 : IHM de Kawab

3.2 Sur un problème jouet - Hépatite

Nous proposons dans cette démonstration de détailler les résultats obtenus sur un problème de médicine. Un classifieur naïf de Bayes est tout d'abord élaboré en utilisant le logiciel Khiops. Puis Kawab est utilisé pour réaliser l'exploration. L'utilisation de Kawab sur les 155 instances que contient cette base de données permet d'améliorer la probabilité de survivre pour 17 'patients'. Quatre des trente patients prédits comme allant décéder voient leur probabilité de survivre augmenter suffisamment pour être classé alors comme 'survivant'.

Parmi les patients prédits comme 'vivant' par le classifieur, mais proche du seuil du décès ($P(C_z = \text{'vivre'} \mid x_k) \approx 0.5$)), 4 d'entre eux sont éloignés du seuil critique de manière significative. L'exploration complète de la base de données montre que la variable explicative 'ALBUMINE' joue un rôle important dans l'hépatite. Cette exploration réalisée de manière automatique montre qu'un accroissement de l'albumine est directement corrélé avec la probabilité de vivre. Cette corrélation trouvée n'est pas une causalité mais apporte néanmoins une connaissance intéressante. De plus, Lichtsteiner dans [4] montre qu'il existe des vecteurs permettant d'accroitre le taux d'albumine …

3.2 Sur votre problème

Nous proposons dans cette démonstration d'appliquer Khiops puis Kawab sur VOTRE problème de classification. Tout d'abord un classifieur sera construit puis Kawab interprètera le résultat de la classification obtenu :

- Exploration des corrélations tel que décrit dans ces deux pages;
- Importance des variables explicatives (non décrit dans cet article mais implémenté dans Kawab).

Un temps sera consacré pour l'installation des versions d'évaluation de Khiops et Kawab (versions limités dans le temps mais pas en fonctionnalités)

4 Pour qui?

L'algorithme d'exploration décrit peut sembler très 'simple' mais il permet de fait de trouver des actions effectives. Il est intégré dans un logiciel qui peut être appliqué sur n'importe quelle base de données, base de données pouvant comporter des milliers de variables et centaines de milliers d'instances.

L'outil est utile aux compagnies et/ou organisations et/ou chercheurs qui veulent comprendre le résultat d'une classification, de manière (i) à accroitre la compréhension qu'ils ont de leur problème et (ii) à potentiellement changer les résultats des classifications obtenues.

Bibliographie

[1] Han, J., Kamber M. Data mining: concepts and techniques. Morgan Kaufmann (2006)

[2] Vincent Lemaire, Carine Hue, Olivier Bernier, "Correlation Explorations in a Classification Model" Workshop "Data Mining Case Studies and Practice Prize" of SIGKDD 2009

[3] Vincent Lemaire, Carine Hue, Olivier Bernier to appear in the Chapter "Correlation Analysis in Classifiers" in "Handbook of Research on Data Mining in Public and Private Sectors: Organizational and Government Applications" A book edited by Professor Antti Syväjärvi and Professor Jari Stenvall.

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