Le choix d'une bonne mesure de qualité, condition du succès d'un processus de fouille de données

Philippe Lenca, Stéphane Lallich

Le choix d'une bonne mesure de qualité, condition du succès de la fouille de données

Philippe Lenca[*,***], Stéphane Lallich[**]

[*]Institut Télécom, Télécom Bretagne
UMR CNRS 3192 Lab-STICC
philippe.lenca@telecom-bretagne.eu
[**]Université Lyon, Lyon 2
Laboratoire ERIC
stephane.lallich@univ-lyon2.fr
[***]Université européenne de Bretagne, France

Notre réflexion se situe dans le domaine de l'apprentissage supervisé ou non supervisé par induction de règles. La fouille de données est couronnée de succès lorsque l'on parvient à extraire des données des connaissances nouvelles, valides, exploitables, etc. (Fayyad et al. (1996) Kodratoff et al. (2001)). L'une des clefs du succès est, bien sûr, le choix d'un algorithme qui soit bien adapté aux caractéristiques des données et au type de connaissances souhaitées : par exemple les règles d'association en non supervisé ; les arbres de décision, les règles d'association de classe et le bayésien naïf, en supervisé. Cependant, le succès dépend d'autres facteurs, notamment la préparation des données (représentation des données, *outliers*, variables redondantes) et le choix d'une bonne mesure d'évaluation de la qualité des connaissances extraites, tant dans le déroulement de l'algorithme que dans l'évaluation finale des résultats obtenus. C'est de ce dernier facteur que nous allons parler.

En introduction, nous évoquerons rapidement le problème de la représentation des données. Puis, après avoir rappelé le principe de la recherche des règles d'association (Agrawal et Srikant (1994)) ou des règles d'association de classe intéressantes (Liu et al. (1998)), nous montrerons, à partir de quelques exemples, la diversité des résultats obtenus suivant la mesure d'intérêt choisie, que ce soit en comparant les pré-ordres obtenus ou en calculant les meilleures règles (Vaillant et al., 2004). Ces exemples illustrent le fait qu'il n'y a pas de mesure qui soit intrinsèquement bonne, mais différentes mesures qui, suivant leurs propriétés, sont plus ou moins bien adaptées au but poursuivi par l'utilisateur. Une mesure favorise tel ou tel type de connaissance, ce qui constitue un biais d'apprentissage que nous illustrerons par la mesure de Jaccard (Plasse et al. (2007)) .

Nous proposerons ensuite une synthèse des travaux concernant les mesures de qualité des règles d'association en présentant les principaux critères d'évaluation des mesures et en montrant concrètement le rôle de chacun de ces critères dans le comportement des mesures (e.g. Lenca et al. (2003), Tan et al. (2004), Geng et Hamilton (2006), Lenca et al. (2008), Suzuki (2008), Guillaume et al. (2010), Lerman et Guillaume (2010), Gras et Couturier (2010) ; nous renvoyons également le lecteur aux ouvrages édités par Guillet et Hamilton (2007) et Zhao et al. (2009)). Nous illustrerons le lien qui existe entre les propriétés des mesures sur les critères retenus et leur comportement sur un certain nombre de bases de règles (Vaillant et al., 2004).

A côté de ces critères qui permettent d'étalonner les propriétés des mesures, nous présenterons d'autres critères de choix très importants. En premier lieu, nous nous intéresserons aux propriétés algorithmiques des mesures afin de pouvoir extraire les motifs intéressants en travaillant directement sur la mesure considérée, sans fixer de seuil de support, ce qui permet d'accéder aux pépites de connaissances (Wang et al. (2001), Xiong et al. (2003), Li (2006), Le Bras et al. (2009), Le Bras et al. (2009), Le Bras et al. (2010)). Nous exhiberons des conditions algébriques sur la formule d'une mesure qui assurent de pouvoir associer un critère d'élagage à la mesure considérée. Nous nous poserons ensuite le problème de l'évaluation de la robustesse des règles suivant la mesure utilisée (Azé et Kodratoff (2002), Cadot (2005), Gras et al. (2007), Le Bras et al. (2010)).

Enfin, nous traiterons le cas des données déséquilibrées (Weiss et Provost (2003)) en apprentissage par arbres (Chawla (2003)) et nous montrerons comment le choix d'une mesure appropriée permet d'apporter une solution algorithmique à ce problème qui améliore de façon significative à la fois le taux d'erreur global, la précision et le rappel (Zighed et al. (2007), Lenca et al. (2008)). Si l'on veut privilégier la classe minoritaire, cette solution peut être encore améliorée en introduisant, dans la procédure d'affectation des étiquettes opérant sur chaque feuille de l'arbre, une mesure d'intérêt adéquate qui se substitue à la règle majoritaire (Ritschard et al. (2007), Pham et al. (2008)). Une discussion sur les mesures de qualité de bases de règles est présentée dans (Holena, 2009).

En définitive, comment aider l'utilisateur à choisir la mesure la plus appropriée à son projet ? Nous proposerons une procédure d'assistance au choix de l'utilisateur qui permet de retourner à celui-ci les mesures les plus appropriées, une fois qu'il a défini les propriétés qu'il attend d'une mesure (Lenca et al. (2008)).

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules. In *VLDB*, pp. 487–499.

Azé, J. et Y. Kodratoff (2002). Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'association. In *EGC*, pp. 143–154.

Cadot, M. (2005). A simulation technique for extracting robust association rules. In *CSDA*, pp. 143–154.

Chawla, N. (2003). C4.5 and imbalanced datasets : Investigating the effect of sampling method, probalistic estimate, and decision tree structure. In *ICM Workshop on Learning from Imbalanced Data Sets*.

Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, et R. Uthurusamy (Eds.) (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

Geng, L. et H. J. Hamilton (2006). Interestingness measures for data mining: A survey. *ACM* (3, Article 9).

Gras, R. et R. Couturier (2010). Spécificité de l'A.S.I. par rapport à d'autres mesures de qualité de règles d'association. In *Analyse Statistique implicative*, pp. 175–198.

Gras, R., J. David, F. Guillet, et H. Briand (2007). Stabilité en A.S.I. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles d'association. In

Qualité des Données et des Connaissances, pp. 35–43.

Guillaume, S., D. Grissa, et E. M. Nguifo (2010). Propriété des mesures d'intérêt pour l'extraction des règles. In *Qualité des Données et des Connaissances*, pp. 15–28.

Guillet, F. et H. J. Hamilton (Eds.) (2007). *Quality Measures in Data Mining*. Springer.

Holena, M. (2009). Measures of ruleset quality for general rules extraction methods. *Int. J. Approx. Reasoning* (6), 867–879.

Kodratoff, Y., A. Napoli, et D. Zighed (2001). Bulletin de l'Association Française d'Intelligence Artificielle, Extraction de connaissances dans des bases de données.

Le Bras, Y., P. Lenca, et S. Lallich (2009). On optimal rules mining: a framework and a necessary and sufficient condition for optimality. In *PAKDD*, Volume 5476 of *Lecture Notes in Computer Science*, pp. 705–712. Springer-Verlag Berlin Heidelberg.

Le Bras, Y., P. Lenca, et S. Lallich (2010). Mining interesting rules without support requirement: A general universal existential upward closure property. *Annals of Information Systems 8*, 75–98.

Le Bras, Y., P. Lenca, S. Moga, et S. Lallich (2009). On the generalization of the all-confidence property. In *ICMLA*, pp. 759–764. IEEE Press.

Le Bras, Y., P. Meyer, P. Lenca, et S. Lallich (2010). A robustness measure of association rules. In *ECML/PKDD*, Volume 6322 of *Lecture Notes in Computer Science*, pp. 227–242. Springer-Verlag Berlin Heidelberg.

Lenca, P., S. Lallich, T.-N. Do, et N.-K. Pham (2008). A comparison of different off-centered entropies to deal with class imbalance for decision trees. In *PAKDD*, Volume 5012 of *Lecture Notes in Computer Science*, pp. 634–643.

Lenca, P., P. Meyer, P. Picouet, B. Vaillant, et S. Lallich (2003). Critères d'évaluation des mesures de qualité en ECD. *Revue des Nouvelles Technologies de l'Information* (RNTI-1), 123–134.

Lenca, P., P. Meyer, B. Vaillant, et S. Lallich (2008). On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. *EJOR* (2), 610–626.

Lerman, I.-C. et S. Guillaume (2010). Analyse comparative d'indices d'implication discriminants fondés sur une échelle de probabilité. Technical Report IRISA 1942/INRIA 7187.

Li, J. (2006). On optimal rule discovery. *TKDE 18*(4), 460–471.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining*, pp. 80–86.

Pham, N.-K., T.-N. Do, P. Lenca, et S. Lallich (2008). Using local node information in decision trees: Coupling a local decision rule with an off-centered entropy. In *DMIN*, Volume 1, pp. 117–123.

Plasse, M., N. Niang, G. Saporta, A. Villeminot, et L. Leblond (2007). Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics & Data Analysis 52*(1), 596–613.

Ritschard, G., D. A. Zighed, et S. Marcellin (2007). Données déséquilibrées, entropie décentrée et indice d'implication. In *Analyse Statistique implicative*, pp. 315–327.

Suzuki, E. (2008). Pitfalls for categorizations of objective interestingness measures for rule discovery. In *Statistical Implicative Analysis, Theory and Applications*, pp. 383–395.

Tan, P.-N., V. Kumar, et J. Srivastava (2004). Selecting the right objective measure for association analysis. *IS* (29), 293–313.

Vaillant, B., P. Lenca, et S. Lallich (2004). A clustering of interestingness measures. In *DS*, pp. 290–297.

Wang, K., Y. He, et D. W. Cheung (2001). Mining confident rules without support requirement. In *CIKM*, pp. 89–96. ACM.

Weiss, G. M. et F. Provost (2003). Learning when training data are costly: The effect of class distribution on tree induction. *J. of Art. Int. Research 19*, 315–354.

Xiong, H., P.-N. Tan, et V. Kumar (2003). Mining strong affinity association patterns in data sets with skewed support distribution. In *ICDM*, pp. 387–394.

Zhao, Y., C. Zhang, et L. Cao (Eds.) (2009). *Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction*. IGI Global.

Zighed, D. A., S. Marcellin, et G. Ritschard (2007). Mesure d'entropie asymétrique et consistante. In *EGC*, pp. 81–86.