L'industrialisation des analyses – Besoins, outils & applications

Françoise Fogelman-Soulié, Erik Marcadé

KXEN, 25 quai Galliéni, 92 158 SURESNES Cedex, France

Francoise@kxen.com, Erik.Marcade@kxen.com

Résumé. Le data mining est aujourd'hui de plus en plus utilisé dans les entreprises les plus compétitives. Ce développement, rendu possible par la disponibilité grandissante de masses de données importantes, pose des contraintes tant théoriques (quels algorithmes utiliser pour produire des modèles d'analyses exploitant des milliers de variables pour des millions d'exemples) qu'opérationnelles (comment mettre en production et contrôler le bon fonctionnement de centaines de modèles). Je présenterai ces contraintes issues des besoins des entreprises ; je montrerai comment exploiter des résultats théoriques (provenant des travaux de Vladimir Vapnik) pour produire des modèles robustes; je donnerai des exemples d'applications réelles en gestion de la relation client. Nous verrons ainsi comment il est possible d'industrialiser le data mining et en faire ainsi un composant facilement exploitable dès qu'on dispose de données.

Abstract. Today data mining is more and more extensively used by very competitive enterprises. This development, brought by the increasing availability of massive datasets, is only possible if challenges, both theoretic and operational, are met : which algorithms should be used to produce models when datasets have thousands of variables and millions of observations; how to run and control the correct execution of hundreds of models. I will present these constraints in industrial contexts; I will show how to exploit theoretical results (coming from Vapnik's work) to produce robust models; I will give examples of real-life applications in customer relationship management. I will thus demonstrate that it is indeed possible to industrialize data mining so as to turn it into an easy-to-use component whenever data is available.

Mots-clés. Data Mining. Robustesse. Passage à l'échelle. Text Mining.

1 Le data mining

1.1 Un peu d'histoire

Le data mining est une discipline qui a émergé progressivement de la convergence de plusieurs domaines : de 1900 à 1990, la *statistique* (Fisher, Cramer, Bayes, Kolmogorov-Smirnoff…) ; de 1940 à 1970, *cybernétique* (Wiener et von Neumann, perceptron de Rosenblatt, Minsky et Papert) ; de 1970 à 1990, *machine learning* : intelligence artificielle, reconnaissance des formes, arbres de décision (Breiman, Friedman) et réseaux de neurones (Hopfield, Kohonen, Rumelhart, LeCun, …), théorie statistique de l'apprentissage (Vapnik) : lors de l'Ecole Modulad de 1996 [5], nous avions montré les liens étroits entre statistique et réseaux de neurones. Depuis, le développement du data mining n'a fait que s'amplifier, grâce aussi à l'informatique qui fournit les moyens matériels indispensables aux traitements de grandes masses de données (ordinateurs rapides, mémoire, disques durs, bases de données).