SOUTENANCE DE THESE

Gen YANG

Unité de Recherche : **UMR 7253 Laboratoire Heudiasyc**

soutiendra sa thèse de **Doctorat**

sur le sujet :

Modèles prudents en apprentissage statistique supervisé

A l'Université de technologie de Compiègne
Le mardi 22 mars 2016 à 10h
Bâtiment Blaise Pascal – GI042

Résumé de thèse

Modèles prudents en apprentissage statistique supervisé

Dans certains champs d'apprentissage supervisé (e.g. diagnostic médical, vision artificielle), les modèles prédictifs sont non seulement évalués sur leur précision mais également sur la capacité à l'obtention d'une représentation plus fiable des données et des connaissances qu'elles induisent, afin d'assister la prise de décisions de manière prudente. C'est la problématique étudiée dans le cadre de cette thèse. Plus spécifiquement, nous avons examiné deux approches existantes de la littérature de l'apprentissage statistique pour rendre les modèles et les prédictions plus prudents et plus fiables: le cadre des probabilités imprécises et l'apprentissage sensible aux coûts. Ces deux domaines visent tous les deux à rendre les modèles d'apprentissage et les inférences plus fiables et plus prudents. Pourtant peu de travaux existants ont tenté de les relier, en raison de problèmes à la fois théorique et pratique. Nos contributions consistent à clarifier et à résoudre ces problèmes.

Sur le plan théorique, peu de travaux existants ont abordé la manière de quantifier les différentes erreurs de classification quand des prédictions sous forme d'ensembles sont

produites et quand ces erreurs ne se valent pas (en termes de conséquences). Notre première contribution a donc été d'établir des propriétés générales et des lignes directrices permettant la quantification des coûts d'erreurs de classification pour les prédictions sous forme d'ensembles. Ces propriétés nous ont permis de dériver une formule générale, le *coût affaiblie généralisé* (CAG), qui rend possible la comparaison des classifieurs quelle que soit la forme de leurs prédictions (singleton ou ensemble) en tenant compte d'un paramètre d'aversion à la prudence. Sur le plan pratique, la plupart des classifieurs utilisant les probabilités imprécises ne permettent pas d'intégrer des coûts d'erreurs de classification génériques de manière simple, car la complexité du calcul augmente de magnitude lorsque des coûts non unitaires sont utilisés. Ce problème a mené à notre deuxième contribution, la mise en place d'un classifieur qui permet de gérer les intervalles de probabilités produits par les probabilités imprécises et les coûts d'erreurs génériques avec le même ordre de complexité que dans le cas où les probabilités standards et les coûts unitaires sont utilisés. Il s'agit d'utiliser une technique de décomposition binaire, les dichotomies emboîtées. Les propriétés et les pré-requis de ce classifieur ont été étudiés en détail. Nous avons notamment pu voir que les dichotomies emboîtées sont applicables à tout modèle probabiliste imprécis et permettent de réduire le niveau d'indétermination du modèle imprécis sans perte de pouvoir prédictif.

Des expériences variées ont été menées tout au long de la thèse pour appuyer nos contributions. Nous avons caractérisé le comportement du CAG à l'aide des jeux de données ordinales. Ces expériences ont mis en évidence les différences entre un modèle basé sur les probabilités standards pour produire des prédictions indéterminées et un modèle utilisant les probabilités imprécises. Ce dernier est en général plus compétent car il permet de distinguer deux sources d'indétermination (l'ambiguïté et le manque d'informations), même si l'utilisation conjointe de ces deux types de modèles présente également un intérêt particulier dans l'optique d'assister le décideur à améliorer les données ou les classifieurs. De plus, des expériences sur une grande variété de jeux de données ont montré que l'utilisation des dichotomies emboîtées permet d'améliorer significativement le pouvoir prédictif d'un modèle imprécis avec des coûts génériques.

Thesis summary

Cautious models in supervised machine learning

In some areas of supervised machine learning (e.g. medical diagnostics, computer vision), predictive models are not only evaluated on their accuracy but also on their ability to obtain more reliable representation of the data and the induced knowledge, in order to allow for cautious decision making. This is the problem we studied in this thesis. Specifically, we examined two existing approaches of the literature to make models and predictions more cautious and more reliable: the framework of imprecise probabilities and the one of cost-sensitive learning. These two areas are both used to make models and inferences more reliable and cautious. Yet few existing studies have attempted to bridge these two frameworks due to both theoretical and practical problems. Our contributions are to clarify and to resolve these problems.

Theoretically, few existing studies have addressed how to quantify the different classification errors when set-valued predictions are produced and when the costs of mistakes are not equal

(in terms of consequences). Our first contribution has been to establish general properties and guidelines for quantifying the misclassification costs for set-valued predictions. These properties have led us to derive a general formula, that we call the *generalized discounted cost* (GDC), which allow the comparison of classifiers whatever the form of their predictions (singleton or set-valued) in the light of a risk aversion parameter.

Practically, most classifiers basing on imprecise probabilities fail to integrate generic misclassification costs efficiently because the computational complexity increases by an order (or more) of magnitude when non unitary costs are used. This problem has led to our second contribution, the implementation of a classifier that can manage the probability intervals produced by imprecise probabilities and the generic error costs with the same order of complexity as in the case where standard probabilities and unitary costs are used. This is to use a binary decomposition technique, the nested dichotomies. The properties and prerequisites of this technique have been studied in detail. In particular, we saw that the nested dichotomies are applicable to all imprecise probabilistic models and they reduce the imprecision level of imprecise models without loss of predictive power.

Various experiments were conducted throughout the thesis to illustrate and support our contributions. We characterized the behavior of the GDC using ordinal data sets. These experiences have highlighted the differences between a model based on standard probability framework to produce indeterminate predictions and a model based on imprecise probabilities. The latter is generally more competent because it distinguishes two sources of uncertainty (ambiguity and the lack of information), even if the combined use of these two types of models is also of particular interest as it can assist the decision-maker to improve the data quality or the classifiers. In addition, experiments conducted on a wide variety of data sets showed that the use of nested dichotomies significantly improves the predictive power of an indeterminate model with generic costs.