Proposition de recherche doctorale

Machine learning under constrained budget for information extraction and search applications

Mots clés :

- **Directeur de thèse** : Thierry Artières
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'informatique de Paris 6
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Automated tools for machine learning are usually designed to maximize performance criteria like classification accuracy or ranking maximization, regression error minimization, etc . The learning criterion of quality is directly integrated into a training loss to be optimized via some optimization technique like optimization algorithms or gradient descent for example. However, for many modern applications, real learning criteria should also consider external constraints such as the cost of data or feature acquisition, computation time, memory usage, or even very concrete external factors such as power consumption, development time, etc. Just to give a few examples, in medicine, there is a cost associated with each medical procedure ((blood test, x-ray, biological analysis, etc) and acquiring new features or data, deciding new tests may be costly or even dangerous for patients. More recently this problem has started to receive attention in the context of the Big Data framework. Mining very large amounts of data involved in many Big Data applications, for example for information retrieval or information extraction, for the analysis of social networks, for marketing applications on large populations is unfeasible. Therefore, it is important to devise algorithms able to learn specific tasks under different types of constraints that limit the possible exploitation of additional data or more generally additional information. This is a new problem in machine learning, which is emerging in the context, where huge quantities of data have to be processed for different types of applications. This general problem has recently been explored through different directions under the names of Budget Learning or Cost Sensitive Learning.

Résumé du projet de recherche (Langue 2)

Learning under budget constraints raises different challenges. The constraints might be diverse, multiple and their formulation often does not lead to classical learning formulation for which existing algorithms might be used. One particular challenge is allowing the simultaneous consideration of various complex criteria. The thesis will explore a series of problems which fall into this general “constrained budget” framework, like the acquisition of new features, data, and the exploration of external knowledge sources. Different families of techniques will be explored and compared like sampling strategies and sequential exploration methods based on reinforcement learning. Besides this fundamental algorithmic work, case study will be considered in the fields of textual information extraction and search and social recommendation. Typically, the methods used in these domains require very large amounts of data and computing power to process them which represent nowadays critical resources. We will explore new methods for budget learning, i.e. learning under external constraints for discovering automatically search methods able to reach providing a certain search quality, within a limited energy budget.

Informations complémentaires (Langue 2)

REFERENCES Attenberg J, Melville P. Guided feature labeling for budget-sensitive learning under extreme class imbalance. In: ICML-2010 Workshop on Budgeted Learning. Attenberg J, Provost F. Online active inference and learning. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’11. New York, New York, USA: ACM Press; 2011:186. Ji S, Carin L. Cost-sensitive feature acquisition and classification. Pattern Recognition. 2007;40(5):1474-1485. Kanani P, Mccallum A, Hu S. Resource-bounded Information Extraction : Acquiring Missing Feature Values On Demand. In: Proceedings of the 14th PAKDD, Hyderabad, India.; 2010. Kanani P. Prediction-time Active Feature-value Acquisition for Cost-Effective Customer Targeting. In: Proceedings of the Workshop on Cost Sensitive Learning, NIPS 2008.; 2008. Kanani PH, McCallum AK. Selecting actions for resource-bounded information extraction using reinforcement learning. In: Proceedings of the fifth ACM international conference on Web search and data mining - WSDM ’12. New York, New York, USA: ACM Press; 2012:253. Kapoor A, Greiner R. Learning and Classifying Under Hard Budgets. In: Machine Learning: ECML 2005. Springer Berlin / Heidelberg; 2005:170-181. Lizotte DJ, Madani O, Greiner R. Budgeted learning of nailve-bayes classifiers. In: UAI ’03, Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence in Artificial Intelligence. Morgan Kaufmann; 2003:378-385.