UNIVERSITE CLAUDE BERNARD – LYON I

DIPLÔME NATIONAL DE DOCTORAT (Arrêté du 25 mai 2016)

Date de la soutenance : 15 décembre 2016

Nom de famille et prénom de l'auteur : **Raywat MAKKHONGKAEW**

Titre de la thèse : « Co-sélection instances-variables en mode semi-supervisé. Application au diagnostic de transport ferroviaire. »

Résumé de la thèse

We are drowning in massive data but starved for knowledge retrieval. It is well known through the dimensionality tradeoff that more data increase informative but pay a price in computational complexity, which has to be made up in some way. When the labeled sample size is too little to bring sufficient information about the target concept, supervised learning fail with this serious challenge. Unsupervised learning can be an alternative in this problem. However, as these algorithms ignore label information, important hints from labeled data are left out and this will generally downgrades the performance of unsupervised learning algorithms. Using both labeled and unlabeled data is expected to better procedure in semi-supervised learning, which is more adapted for large domain applications when labels are hardly and costly to obtain. In addition, when data are large, feature selection and instance selection are two important dual operations for removing irrelevant information. Both of tasks with semi-supervised learning are different challenges for machine learning and data mining communities for data dimensionality reduction and knowledge retrieval.

In this thesis, we focus on co-selection of instances and feature in the context of semi-supervised learning In this context, co-selection becomes a more challenging problem as the data contains labeled and unlabeled examples sampled from the same population. To do such semi-supervised co-selection, we propose two unified frameworks, which efficiently integrate labeled and unlabeled parts into the co-selection process. The first framework is based on weighting constrained clustering and the second or is based on similarity preserving selection. Both approaches evaluate the usefulness of features and instances in order to select the most relevant ones, simultaneously.

Finally, we present a variety of empirical studies over high-dimensional data sets, which are well-known ir the literature. The results are promising and prove the efficiency and effectiveness of the proposed approaches. In addition, the developed methods are validated on a real world application, over data provided by the State Railway of Thailand (SRT). The purpose is to propose the application models from our methodological contributions to diagnose the performance of rail dry port systems. First, we present the results of some ensemble methods applied on a first data set, which is fully labeled. Second, we show how can our co-selection approaches improve the performance of learning algorithms over partially labeled data provided by SRT.