Unsupervised multi-tasks learning of structured data

Supervisor's names:

Pierre Beauseroy	(Full professor)	pierre.beauseroy@utt.fr
Alexandre Baussard	(Full professor)	alexandre.baussard@utt.fr

Description of the research project

In recent years, machine learning has made substantial advances in various domains and has gained great attention. Basically, machine learning methods tends to train models that relates input data to desired output using training data. Machine learning has become a wide field of studies with many different types of methods. These methods can be categorized depending on the data available for learning and the objective that is pursued when training a model.

From data point of view, the most studied situation, is the case where one set of observations is available and all observations are characterized by a given set of features.

At least two different situations can be considered depending if training data are labelled (typically the class that should be assigned to each observation) or not. Methods that applies to data without label are said to be non-supervised methods. When dealing with non-supervised situation, the goal of training can be to strengthen the structure of the data in order, for instance, to identify groups in the given population and/or to reduce the number of features that characterizes each observation. The reduction of the feature number helps to reduce the computational load of further processing and above all to improve the robustness of awaited results.

Nowadays, with increase in sensor and data availability, many related databases can be found containing complementary information on given observations. One can figure out this situation as viewing a common object type from different point of view (figure 1).

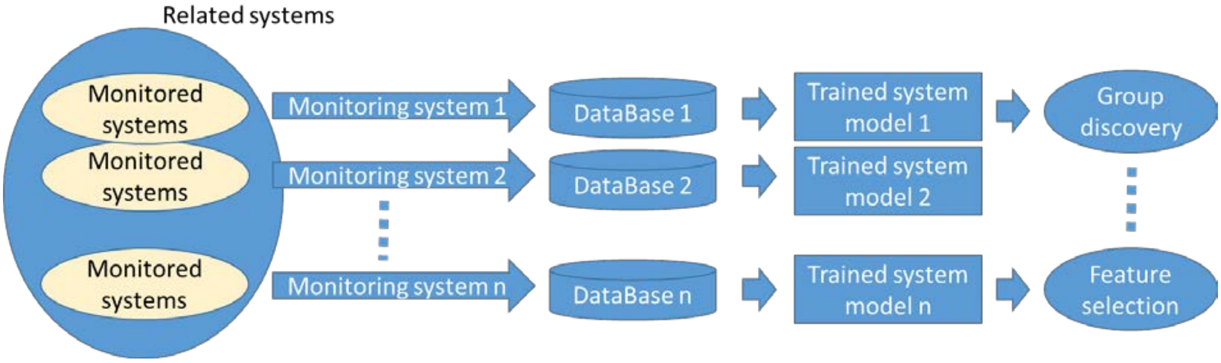

Figure 1 – Classic individual learning scheme – each database is a data source

For instance, simple examples of such case can be found when dealing with medical data. Depending on the location and for many different reasons, for a given cancer type search, patients will not be subject to the exact same series of test (DNA, Gene expression…). To model a given test, one could consider databases coming from different origin that contain information about the targeted test to increase the number of patients and to improve the training of each model. This can be done individually for each test. But one may want to go a step forward and take advantage of the commonalities between the different tests to improve each model. This situation is typical of multi-tasks learning and correspond to the purpose of this research project.

Gathering multisource data leads, most of the time, to databases that are sparse and contain many missing data due to the fact that each source contains only a subset of the all features obtained when combining all sources. A classical way to deal with unknown values is missing data estimation or imputation methods. But, when the number of sources increases, the standard approaches face a combinatorial explosion issue due to the possible combination of available sources and missing sources to be estimated.

The proposed research project is to development new approaches based on neural networks and deep neural networks to performed multi-task unsupervised learning in the case of multisources data. This should enable to solve the combinatorial issue and to offer a common framework to handle missing data in non-controlled context (when only non-controlled part of the sensors is responding in a monitoring system) and the problem of multisource data processing. Simultaneously, it should improve the individual models trained for each sub-data set (figure 2).

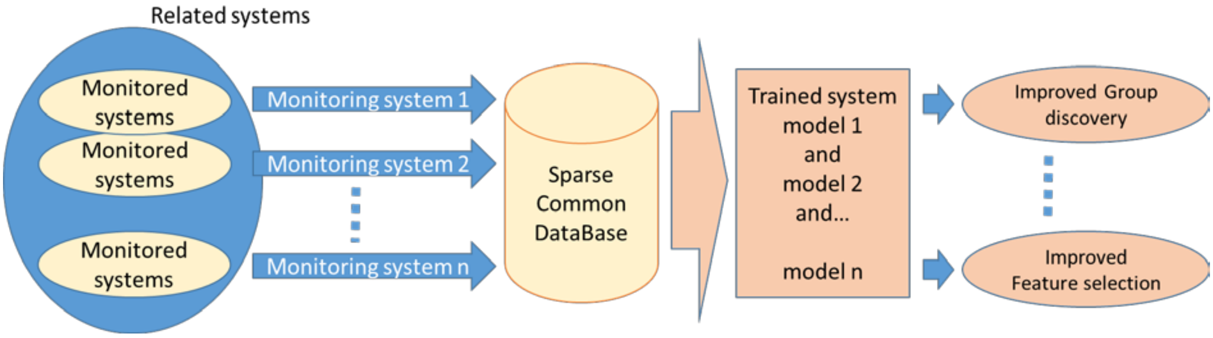

Figure 2 – Diagram of the proposed research project

Two applications may be used to assess the developed new methods. The first one is about multispectral image analysis of agricultural products and the second one about subgroup of cancer identification based on multi-omic data (biological data).

Additional Bibliography

1. Caruana, Rich (1997) Multitask Learning. In Machine Learning 28, 41-75
2. Pedro J. Garcia-Laencina, Jesus Serrano,Anibal R. Figueiras-Vidal, and José-Luis Sancho-Gomez, (2007) Multi-task Neural Networks for Dealing with Missing Inputs. IWINAC, La manga del mar menor, Spain.
3. Yu Zhang and Qiang Yang (2017), A Survey on Multi-Task Learning. In arXiv 1707.08114
4. Antonio de la Vega de León, Beining Chen and Valerie J. Gillet (2018) Effect of missing data on multitask prediction methods. In Journal of Cheminformatics (10:26)
5. Xin J. Hunt, Saba Emrani, Ilknur Kaynar Kabul, Jorge Silva, (2018) Multi-Task Learning with Incomplete Data for Healthcare. ArXiv:1807.02442v1
6. Zheng Zhang a,b, Qi Zhu c, Guo-Sen Xie d, Yi Chen e,f, Zhengming Li g, Shuihua Wangh (2020) Discriminative margin-sensitive autoencoder for collective multi-view disease analysis. In Neural Networks 123, 94-107