Unsupervised multi-task learning from incomplete multisource data

Supervisor's names:
Pierre Beauseroy (Full professor) pierre.beauseroy@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 on whether the training data is labelled (typically the class that should be assigned to each observation) or not. Methods that apply 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).

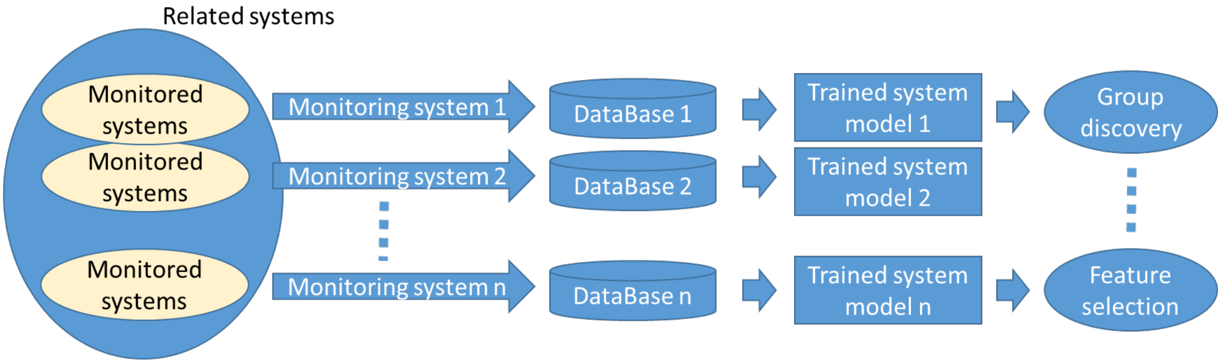

Figure 1 – Classic individual learning scheme – each database is a data source.

Simple examples of this type of case can be found in the field of medical data. Depending on the examination centre and for many different practical reasons, for a given type of cancer, patients will not undergo exactly the same series of tests (DNA, gene expression, MRI, etc.). To build a diagnostic aid model based on a given type of test (MRI, for example), we can consider databases of different origins, different medical centers, each of which contains information about the targeted test. Proceeding in this way increases the number of patients and improves the training of the diagnostic model in question. However, there are two limitations to this approach. On the one hand, it is not easy to compile such homogeneous data sets, and on the other, if you are interested

in several types of examination, you will have to repeat these data consolidation operations between different databases for each examination and each diagnostic aid model to be developed.
To limit the amount of work and get the most out of the data, you can try to use all the data sources together instead. On the one hand, this would make it possible to avoid creating a specific database for each model and, on the other hand, it would make it possible to take advantage of the common points between the different examinations to improve each model. This situation is typical of multi-task learning and corresponds to the objective of this research project.

In most cases, multi-source data collection leads to sparse databases containing a large amount of missing data, as each source contains only a subset of all the characteristics potentially available and each observation is only visible in a very limited number of sources (one patient will not undergo all the possible analysis). For each observation available in at least one database, estimating the missing data using regression methods or imputation methods, for example, is a conventional way of dealing with the unknown values and completing a common unified database (Figure 2). However, when the number of sources increases, these standard approaches become inoperative because they are faced with a combinatorial explosion problem due to the multiplication of possible groups of available sources and therefore of the complementary 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 multisource data. This should enable to solve the combinatorial issue related to the multiplication of sources number and to offer a common framework to handle missing data in non-controlled context (typically when only a subset of the sensors in a monitoring system transmits information about an observed system and the sensors member of this subset are not selected or controlled). Simultaneously, it should improve the individual models trained for each sub-data set (figure 2) taking advantage of the commonalities between the different sensors.

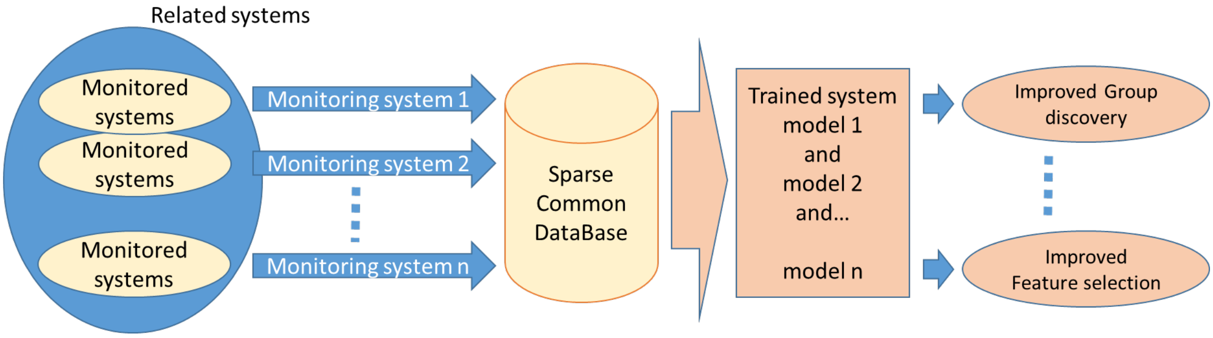

Figure 2 – Diagram of the proposed research project

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, Qi Zhu, Guo-Sen Xie, Yi Chen, Zhengming Li, Shuihua Wangh (2020) Discriminative margin-sensitive autoencoder for collective multi-view disease analysis. In Neural Networks 123, 94-107
7. Xu, Longfei, Xu, Lingyu, Yu, Jie (2023) A multi-task learning-based generative adversarial network for red tide multivariate time series imputation, In Complex & Intelligent Systems 9, 1363-1376