Proposition de recherche doctorale

Learning representations in structured domains using neural networks and applications to social network analysis

Mots clés :

- **Directeur de thèse** : Patrick Gallinari
- **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)

{{Keywords}} : Machine Learning, Relational Learning, Social Networks , Representation Learning, Latent Models, Deep Neural Networks, Matrix Factorization, Big Data {{Context}} The development of Information technology has led to the production of huge volumes of data. A recent report by IBM [IBM 2013] mentions that every day, quintillions of data are produced, and that the majority of these data has been produced in the last few years only. Managing such large volumes has become a major issue for modern Information Technology. This domain, known as Big Data Analysis, raises several important challenges including the capture, the storage, the search, and the analysis of these data. This thesis proposal focuses on this last aspect: the development of automatic tools for the analysis of large volumes of complex data based on Machine Learning algorithms and their application to the domain of social network analysis. Machine Learning has become a key technology for the analysis of data and is now a major component of the data processing platforms developed by major IT companies. The field has considerably developed over the last ten years together with the growth of data production. Data come now from everywhere: information-sensing mobile devices, wireless networks, cell phone GPS signals, internet, user logs, purchase transaction records, videos, social media sites, biology and many other sources. They come in many different forms and variety as could be inferred from the previous examples. They are often complex, exhibiting different relational structures among the data elements. On social network sites for example, the data may correspond to different sources like text, images, videos; the users or items of these networks may share different relations. To summarize, data analysis and in particular Machine Learning faces important challenges related to the volume, diversity and complexity of the data sources to be processed. {{Subject}} Practical applications in the field of data analysis require intensive pre-processing steps in order to provide a data representation amenable to machine learning. These pre-processing steps usually involve human expertise and knowledge about the task and the data to be handled. A major challenge is then to alleviate this burden by developing automatic techniques for learning good representations. This has become a field of study by itself in the machine learning community, and in 2013, a new international conference has been launched, and is entirely dedicated to this topic [ICLR 2013]. The thesis proposal concerns representation learning for complex relational data, typical of those that can be found on social sites. For many current applications, data are more and more complex. They are often multimodal (image, music, video, comments, etc.) , heterogeneous (different types of entities are present) and dynamic (they evolve with time). Besides, they are often highly relational (e.g. in biology or in social networks) with different types of relations linking the different entities. Classical data analysis techniques, and therefore machine learning algorithms, are not adapted to the processing of these complex data. One possible choice to circumvent this problem is then to learn data representations which can then be handled by classical algorithms. One way towards this direction is to map these complex data onto one or several latent representation spaces where machine learning techniques may then operate. If one is able to learn such representations, it will then be possible to define metrics or similarity functions onto these latent spaces, it will be possible to represent multi-modal and heterogeneous data onto common latent spaces and to consider relational information as constraints on these latent representations. Different families of techniques have been developed for learning such representations. Probabilistic techniques like PLSA or LDA [Hoffman 1999, Blei 2003] allow us learning latent distributions, matrix factorization methods like Non Negative Matrix Factorization [Lee 99] have been successfully used in different domains like recommendation or link prediction. More recently, deep neural networks [Socher 2011, Le 2012, Bordes 2001] have emerged as a powerful family of methods for learning complex transformation of the data.

Résumé du projet de recherche (Langue 2)

The proposed research direction are the following: -* Learning representations for heterogeneous networks and multimodal data Develop a representation learning framework adapted to complex relational data, representative of what can be found on heterogeneous social networks. This framework will exploit earlier developments on manifold learning [Weston 2008]. Both non parametric and parametric method will be developed with a focus on the latter which are more suitable for performing inference in evolving environments. -* Deep Neural Networks An alternative direction is the use of Deep Neural Networks which can be used to implement parametric functions for learning to encode data onto latent representations. Deep Neural Networks allow us to push further the learning of representation by building non linear transformations through the combination of successive mappings learned sequentially. -* Integration of temporal and dynamic information Social data are dynamically evolving. This dynamic or temporal evolution is a major characteristic of these data which should be considered in most applications. The above models have been mainly developed for static unstructured or structured data and the integration of their dynamic nature into the above models is another direction that will be considered in this proposal. -* Applications in the domain of social network analysis These methods will be applied in different applications for the analysis of social data. We forecast two main applicative domains which are respectively social recommendation (i.e. the development of social recommendation systems that take into consideration the social links between the users) and classification and ranking in social networks which is more of a generic problem that can be used for different applications.

Informations complémentaires (Langue 2)

{{References}} -* [IBM 2013] "IBM What is big data? — Bringing big data to the enterprise". 01.ibm.com -* [ICLR 2013] International Conference on Learning Representations 2013, https://sites.google.com/site/representationlearning2013/ -* [Socher 2011] [Socher R., Lin C., Ng A. Y., Manning C.D., Parsing Natural Scenes and Natural Language with Recursive Neural Networks, ICML 2011. -* [Le 2012] Le Q. V. , Ranzato M., Monga R., Devin M., Corrado G., Chen K., Dean J., Ng A., Building high-level features using large scale unsupervised learning, ICML 2012 -* [Lee 1999] Lee D.D., Seung H.S. "Learning the parts of objects by non-negative matrix factorization". Nature 401 (6755), 1999. -* [Poultney 2006] Poultney C. , Chopra S. , Lecun Y., Efficient learning of sparse representations with an energy-based model, NIPS 2006 -* [Bordes 2011] Bordes A., Weston J., Collobert R., and Bengio Y.. Learning structured embeddings of knowledge bases. In AAAI, 2011.