Opening new perpectives for classifying and mining textual and web data

Jean-Charles LAMIREL, (LORIA/INRIA – Université de Strasbourg, France)
Mail: Jean-Charles.Lamirel@loria.fr

Guest Speaker—

Office Address: INIST-CNRS
2, allée du Parc de Brabois CS10310
54519 Vandoeuvre-lès-Nancy CEDEX. France
Tel: +33-383504670
Fax: +33-383504646

Jean-Charles LAMIREL is currently teaching Information Science and Computer Science at the Universities of Strasbourg and Nancy and achieving his research at the INRIA laboratory of Nancy (LORIA). He is a research member of the INRIA-CORTEX project whose scope is Neural Networks and Biological Systems.

His main domain of research is Data Mining, Scientometrics and Webometrics based on Neural Networks. He has interests both in theoretical Neural Network models for Data Mining, Scientometrics, and Webometrics, and on applications in these areas. He is more specifically specialized in unsupervised learning methods. He is the creator of the concept of Data Analysis based on Multiple Viewpoints which has been fruitfully implemented in the MultiSOM and MultiGAS models. These models for which it has been theoretically proven that they outperform classical models of Data Analysis begin to be used in many challenging Data Mining, Scientometrics and Webometrics applications.

ABSTRACT

Neural clustering algorithms show high performance in the general context of the analysis of homogeneous textual dataset. This is especially true for the recent adaptive versions of these algorithms, like the incremental growing neural gas algorithm (IGNG) and the labeling maximization based incremental growing neural gas algorithm (IGNG-F). In this paper we highlight that there is a drastic decrease of performance of these algorithms, as well as the one of more classical algorithms, when a heterogeneous textual dataset is considered as an input. Specific quality measures and cluster labeling techniques that are independent of the clustering method are used for the precise performance evaluation. We provide new variations to incremental growing neural gas algorithm exploiting in an incremental way knowledge from clusters about their current labeling along with cluster distance measure data. This solution leads to significant gain in performance for all types of datasets, especially for the clustering of complex heterogeneous textual data.

Keywords:

Neural clustering algorithms ; heterogeneous textual dataset ; performance evaluation ; complex clustering.

Title (FR) : Nouvelles approches pour la classification et la fouille des données textuelles et des données Web.