Journal of Intelligent & Fuzzy Systems 28 (2015) 2329–2330
DOI:10.3233/IFS-141516
IOS Press

Learning from big data with uncertainty – editorial

Xizhao Wang*
Big Data Institute, College of Computer Science and Software Engineering, ShenZhen University, China

Abstract. Focusing on learning from big data with uncertainty, this special issue includes 5 papers; this editorial presents a background of the special issue and a brief introduction to the 5 papers.

Keywords: Big data, machine learning, uncertainty, 3V characteristics

Uncertainty is a natural phenomenon in machine learning, which can be embedded in the entire process of data preprocessing, learning and reasoning. For example, the training samples are usually imprecise, incomplete or noisy, the classification boundaries of samples may be fuzzy, and the knowledge used for learning the target concept may be rough. Uncertainty can be used for selecting extended attributes and informative samples in decision tree inductive learning and active learning respectively. If the uncertainty can be effectively modeled and handled during the process of processing and implementation, machine learning algorithms will be more flexible and more efficient. The representation, measure, and handling of uncertainty have a significant impact on the performance of learning algorithms.

Usually the modeling/handling of uncertainty is associated with the feature-type and volume of dada. Recent research shows that making clear the change/adaptation of uncertainty with feature-type and volume of data is a very difficult issue. This difficulty is significantly increasing if we deal with the big data, a really hot topic in recent years, which refers to a data set with all or some of the following 3V characteristics: (1) Volume, i.e., the data size is becoming so large that the traditional learning algorithms are insufficient to work well; (2) Variety, i.e. features of the data can be multimodal unstructured ones such as image, video, or text, rather than the traditionally symbolic or real-valued; and (3) Velocity, i.e. the data acquisition is so quick that we must consider the incremental streaming data.

Focusing on learning from big data with uncertainty, we propose to organize a special issue section in this journal. After a strict peer review we selected 5 papers for this issue from 20 online/offline submissions. The following is a brief introduction to the 5 selected papers.

The paper "Interval extreme learning machine for big data based on uncertainty reduction" authored by Yingjie Li, Ran Wang and Simon C. K. Shiu, mainly investigates the problem of interval value learning with big data. In this paper, based on uncertainty reduction, an interval extreme learning machine model is developed for large-scale binary classification problems. This model is built up with two techniques, i.e., discretization of conditional attributes and fuzzification of class labels. The experimental results reported in this paper demonstrate that the proposed method is not only able to improve generalization capability, but also effective to compress large volume data.

The paper "A hierarchical fuzzy cluster ensemble approach and its application to big data clustering"

*Corresponding author. X. Wang, Big Data Institute, College of Computer Science and Software Engineering, ShenZhen University, China. Tel./Fax: +86 0312 5079638; E-mail: xizhaowang@ieee.org.

authored by Pan Su, Changjing Shang and Qiang Shen, discusses the problem of cluster ensembles, which organically integrate individual component methods which may utilize different parameter settings and features, and which may themselves be generated on the basis of different representations and learning mechanisms. This paper extends the cluster ensemble approach to fuzzy clustering, with an aim to be applied for clustering of big data. The proposed algorithm first generates fuzzy base clusters with respect to each data feature and then, employs a fuzzy hierarchical graph to represent the relationships between the resulting base clusters. The work employs fuzzy c-means and hierarchical clustering in generating base cluster and implementing consensus function respectively. When applied to large datasets it has lower time complexity than the original fuzzy c-means and hierarchical clustering.

In the paper "Combination of OSELM classifiers with fuzzy integral for large scale classification" by Junhai Zhai, Jinggeng Wang and Wenxiang Hu, an approach is proposed for large scale classification, which combines OSELM (Online Sequential Extreme Learning Machine) classifiers with fuzzy integral. Firstly, the proposed method trains component classifiers with sequentially strategy on subsets of a large data set, the instances previously used will be excluded from training the following component classifiers. Secondly, the trained component classifiers are combined according to a fuzzy integral model. Thirdly, the aggregation learning system is used for classifying the unseen samples. Experimental results compared with two state-of-the-art approaches verified the effectiveness of the proposed method.

The paper "Division-based large point set registration using coherent point drift with automatic parameter tuning" authored by Junfen Chen, Iman Yi Liao, Bahari Belaton and Munir Zaman studies the problem of large point cloud registration. Point cloud, a collection of coordinates of the spatial points, has the characteristics of big data. Coherent Point Drift (CPD) is able to handle large point cloud registration. However, its registration performance degrades rapidly with the increasing of data points. To overcome this problem, this paper presents a strategy that divides a large point set into several smaller subsets which are extended on the margins. These extended subsets are independently registered using CPD and then merged for final registration. In addition, an approach to tuning the width parameter of Gaussian kernel in CPD for improving the registration accuracy is also proposed in this paper. Experimental results show that the proposed method is able to register large scale datasets with both quicker speed and higher registration accuracy.

In the paper "Enhanced soft subspace clustering through hybrid dissimilarity" authored by Lijuan Wang, Zhifeng Hao, Ruichu Cai and Wen Wen, they discuss a problem of high-dimensional clustering. Soft subspace clustering is one of effective methods to deal with this problem, but most existing soft subspace clustering algorithms assign weights to features relying on Euclidean distance, which only allows soft subspace clustering algorithms to extend or shrink feature space with respect to feature weights. In this work, the authors aim to break this limitation and facilitate feature space transformations directed by an enhanced soft subspace clustering algorithm through a hybrid dissimilarity measure. The proposed hybrid dissimilarity measure assigns weights to features based on Euclidean distance and Cosine dissimilarity, which can extend, shrink and rotate feature space so that the uncertainty in high dimensionality is reduced. It results in an improvement of performance for soft subspace clustering.

I would like to thank all the authors and reviewers for their contributions to this special issue section. Also I would like to thank Professor Reza Langari, the Editor-in-Chief of Journal of Intelligent and Fuzzy Systems, for his assistance in organizing this category.