International Journal of Computer Vision and Image Processing
Volume 7 • Issue 2 • April-June 2017

Feature Selection of Interval Valued Data Through Interval K-Means Clustering

D. S. Guru, Department of Studies in Computer Science, University of Mysore, Mysore, India
N. Vinay Kumar, Department of Studies in Computer Science, University of Mysore, Mysore, India
Mahamad Suhil, Department of Studies in Computer Science, University of Mysore, Mysore, India

ABSTRACT

This paper introduces a novel feature selection model for supervised interval valued data based on interval K-Means clustering. The proposed model explores two kinds of feature selection through feature clustering viz., class independent feature selection and class dependent feature selection. The former one clusters the features spread across all the samples belonging to all the classes, whereas the latter one clusters the features spread across only the samples belonging to the respective classes. Both feature selection models are demonstrated to explore the generosity of clustering in selecting the interval valued features. For clustering, the kernel of the K-means clustering has been altered to operate on interval valued data. For experimentation purpose four standard benchmarking datasets and three symbolic classifiers have been used. To corroborate the effectiveness of the proposed model, a comparative analysis against the state-of-the-art models is given and results show the superiority of the proposed model.

KEYWORDS

Feature Selection, Interval Data, Interval K-Means Clustering, Symbolic Classification, Symbolic Similarity Measure

INTRODUCTION

In the current era of digital technology- pattern recognition plays a vital role in the development of cognition based systems. These systems quite naturally handle a huge amount of data. While handling such vast amount of data, the task of data processing has become curse to process. To overcome curse in data processing, the concept of feature selection is being adopted by researchers. Nowadays, feature selection has become a very demanding topic in the field of machine learning and pattern recognition, as it select the most relevant and non-redundant feature subset from a given set of features using a feature selection technique. Basically, the feature selection techniques are broadly classified into: filter, wrapper, and embedded methods (Artur et. al., 2012).

Generally, the existing conventional feature selection methods (Artur et. al., 2012) fail to perform feature selection on unconventional data like interval, multi-valued, modal, and categorical data. These data are also called in general symbolic data. The notion of symbolic data was emerged in the early 2000, which mainly concentrates in handling very realistic type of data for effective classification, clustering, and even regression for that matter (Lynne and Edwin, 2007). As it is a powerful tool in solving problems in a natural way, we thought of developing a feature selection model for any one of

DOI: 10.4018/IJCVIP.2017040105

the modalities. In this regard, we have chosen with an interval valued data, due its strong nature in preserving the continuous streaming data in discrete form (Lynne and Edwin, 2007). Thus, we built a feature selection model for interval valued data in this work.

Initially, Manabu Ichino (1994) provided the theoretical interpretation of feature selection on interval valued data. The method works based on the pretended simplicity algorithm handled in Cartesian space. Later, there are couple of attempts found on feature selection done on mixed type data (i.e., interval, multi-valued, and qualitative). Bapu et. al., (2007) proposed a two-stage feature selection algorithm which can handle both interval as well multi-valued data using Mutual Similarity Value proximity measure for un-supervised data. Qin et al., (2014) proposed an approach based on information theory which selects an optimal feature subset by computing modified heuristic mutual information. This approach handles both numeric and interval valued features. Lyamine et al., (2015) proposed a feature selection model which handles heterogeneous type data viz., interval, quantitative, and qualitative. The proposed feature selection model makes use of similarity margin and weighting scheme for selecting features. In addition to this the model converts the different types of data into a common type and further a common weighting scheme is employed on it. Lyamine et. al., (2011) present the feature selection of interval valued data based on the concept of similarity margin computed between an interval sample and a class prototype. The similarity margin is computed using a symbolic similarity measure. The authors have constructed basis for the similarity margin and then they worked out at the multi-variate weighting scheme. The weight corresponding to each feature decides the relevancy of that feature. Hence, they considered Lagrange Multiplier for optimizing the weights which results with the optimal set of features. The experimentation is done on three standard benchmarking interval dataset and validated using LAMDA classifier. Chih-Ching et. al., (2014) have come up with the model which make use (Lyamine et. al., 2011) in all aspects in selecting the features but with respect to similarity measure computation, the authors have used robust Gaussian kernel. The authors also have given an experimentation and a comparative analysis on only one interval dataset. Jian et al., (2016) proposed a heuristic approach for attribute reduction. This approach makes use of rough set and information theory for attribute reduction. In the theory of rough sets, if the efficiency of the optimal subset equals to the efficiency of original feature set then such process is termed as attribute reduction (instead of feature selection). Guru and Vinay (2016) proposed a feature selection model based on two novel feature ranking criteria for interval valued data. This model makes use of vertex transformation technique before computing the rank of the features. The ranked features are then sorted based on their relevancy before get selected through experimentation. The limitation of this model is the computation of vertex transformation for higher dimensional data which leads to exponential time complexity.

Apart from the above-mentioned works, no work can be seen on feature selection of interval valued data based on clustering of features. Clustering of features and then selecting the cluster representatives helps in improving the prediction accuracy. In addition, it also eliminates the redundant features, as it selects features which are of most relevant (Qinbao et. al., 2013). With this background, here in this paper, a feature selection model is proposed.

The proposed feature selection model is based on clustering the interval features through interval K-Means clustering algorithm. The conventional K-Means clustering is modified to adopt the interval valued data by altering the conventional kernel with the symbolic interval similarity measures. The clustering of interval features for feature selection is realized in two different ways viz., class independent features clustering and class dependent features clustering. In the former approach, initially the supervised interval feature matrix (Figure 1(a)) is transformed and then the transformed feature matrix (Figure 1(b)) is fed into interval K-means clustering algorithm. Thus, results with the K