Expert Systems With Applications 46 (2016) 485–493

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Adaptive pairing of classifier and imputation methods based on the characteristics of missing values in data sets

Jaemun Sim [a], Ohbyung Kwon [b,*], Kun Chang Lee [a]

[a] SKKU Business School, Sungkyunkwan University, Seoul 110-745, Republic of Korea
[b] School of Management, Kyung Hee University, Seoul 130-701, Republic of Korea

ARTICLE INFO

Keywords:
Classification algorithms
Imputation methods
Case-based reasoning
Experiments

ABSTRACT

Classifiers and imputation methods have played crucial parts in the field of big data analytics. Especially, when using data sets characterized by horizontal scattering, vertical scattering, level of spread, compound metric, imbalance ratio and missing ratio, how to combine those classifiers and imputation methods will lead to significantly different performance. Therefore, it is essential that the characteristics of data sets must be identified in advance to facilitate selection of the optimal combination of imputation methods and classifiers. However, this is a very costly process. The purpose of this paper is to propose a novel method of automatic, adaptive selection of the optimal combination of classifier and imputation method on the basis of features of a given data set. The proposed method turned out to successfully demonstrate the superiority in performance evaluations with multiple data sets. The decision makers in big data analytics could greatly benefit from the proposed method when it comes to dealing with data set in which the distribution of missing data varies in real time.

1. Introduction

Emerging infrastructures like cloud systems, smart grids, pervasive computing systems and network-related processing are providing managers and practitioners with more flexible utilities for the sake of adopting user-intended applications (Nia, Atani, & Haghi, 2014). Such infrastructures have greatly contributed to producing big data set in a format of massive amounts of streamed information from a wide variety of the network-connected objects (Sowe, Kimata, Dong, & Zettsu, 2014). Correspondingly, data sets in marketing, scheduling, and manufacturing businesses become very large (volume), get rapidly updated by streaming (velocity) (Bifet, 2013), and/or inadvertently tend to be incomplete due to the nature of their sources like IoT (Internet of Things) and social networking services (SNSs) (variety) as well (Chen, Mao, Zhang, & Leung, 2014). It is no surprise that the significant challenges in this type of dataset encompass the unstable data structure and/or characteristics with null value problems caused by either rapidly changing user locations, fault sensors or user's non-responses. The problems like this become more serious when the big data application systems implemented on powerful classifiers tend to repeatedly show poor performances because of the constantly changing patterns of missing data, data volume and data structure embedded in the big data sets.

To cope with these challenges, intelligent applications must be improved in the following ways. First, due to the volume and velocity of data in these data sets, scalable classification is required (Jang, 2014; Liu, Blasch, Chen, Shen, & Chen, 2013). Second, with respect to variety, many null values must be included in order to maintain a satisfactory level of reasoning accuracy (Kim, 2012; Wu, Zhu, Wu, & Ding, 2014). To alleviate these problems, it is necessary to develop a sophisticated method of finding optimal pairs from every possible classifier/imputation method pair in real time.

According to the literature in this area, characteristics of missing data, data sets, and imputation methods may influence the performance of classification algorithms (Sim, Lee, & Kwon, 2015). Research in various data domains has been conducted related to selecting an imputation method that improves the performance of a classifier, and several new imputation methods have been proposed (Farhangfar, Kurgan, & Dy, 2008; Hengpraphrom, Wichian, & Meesad, 2010; Kang, 2013; Liu & Brown, 2013; Luengo, García, & Herrera, 2010; Silva & Hruschka, 2013). Although most imputation methods improve overall classification performance, the magnitude of improvement differs according to the problem domain (Farhangfar et al., 2008; Hengpraphrom et al., 2010; Su, Khoshgoftaar, & Greiner, 2008). The differences in magnitude become clearer as the ratio of missing data increases (Hengpraphrom et al., 2010; Su et al., 2008). To the best of our knowledge, when experimenting with various data sets, no imputation method has always proven superior to other methods in

* Corresponding author. Tel.: +8229612148; fax: +8229610515.
E-mail addresses: deskmoon@gmail.com (J. Sim), obkwon@khu.ac.kr, obyung@gmail.com (O. Kwon), kunchanglee@gmail.com (K.C. Lee).

http://dx.doi.org/10.1016/j.eswa.2015.11.004
0957-4174/

combination with any specific classifiers, because the effect of an imputation method on a classifier differs according to the data set (Farhangfar et al., 2008; Kang, 2013).

If the characteristics of the data set are invariant and fully known beforehand, as prior studies have assumed, identification of an optimal combination of a classifier and imputation method would be possible. However, if the data is collected in real time, the characteristics of the data set will differ depending on the timeline. In this case, the performance of all possible pairs of classifiers and imputation methods for all types of data characteristics must be evaluated in order to select the optimal combination. Moreover, if real-time analysis is needed for an application, an autonomous method of selecting this optimal combination is necessary. However, very few studies have addressed this need for autonomous selection of classifiers and imputation methods based on the characteristics of a data set, especially as regards the structure of null values.

The purpose of this paper is to propose an adaptive method of selecting the optimal classification algorithm/imputation method pair. An autonomous, adaptive selection method should be able to recognize the features of a data set and, if necessary, make changes automatically. To develop this method, we amended case-based reasoning as follows: the original case base is preprocessed to derive a compound metric of a null data structure. Then a candidate set is formed by identifying multiple pairs, and a pair is selected from among the candidate pairs. To demonstrate the feasibility and superiority of the proposed method, we conducted experiments with multiple benchmark data sets and several classifiers and imputation methods that have been deemed suitable in previous studies for reasoning with incomplete data sets.

The paper is organized as follows: Section 2 describes the related works on imputation methods and classifiers. The proposed method and corresponding experiment, which shows the performance of the method, are delineated in Sections 3 and 4, respectively. Finally, in Section 5, we conclude with the implications of the study results to researchers and practitioners.

2. Related works

2.1. Selection of imputation methods

Researchers using supervised learning algorithms, such as those used for classification, have generally assumed that training data sets are complete and that all occurrences contain a value. Missing values are filled in using many imputation methods. Imputation techniques are based on the idea that missing data for a variable are replaced by an estimated value that is drawn from the distribution of existing values. In most cases, attributes of data sets are interdependent; thus, through identification of relationships among attributes, missing values can be determined (Batista & Monard, 2003; Kang, 2013; Li, Li, & Li, 2014).

There is no single superior imputation algorithm for replacing all missing data in a set, because all imputation methods are affected by the characteristics of the data set and the missing values (Kwon & Sim, 2013; Loh & H'ng, 2014). Thus, if the characteristics of a data set and its missing values are changed by some event, then the performance of the selected imputation method may be altered. For example, for sensor-based traffic data, which vary periodically under certain expected conditions such as changed load capacity or altered timeline, robust imputation algorithms using historical information may be prepared (Tan, Wu, Cheng, Wang, & Ran, 2014). However, various data sets, such as those from SNSs, may be changed by uncertain and complex events (Wrzus, 2013); therefore, the characteristics of missing values may also change. Due to this uncertainty, it is impossible to prepare a robust imputation method using data from prior experiments. Moreover, most sensor-based data require real-time decisions. The need for swift execution makes it difficult to select a suitable imputation method fast enough using the techniques outlined in existing studies in which comparative experiments among candidate imputation methods were performed. Considering the two factors of missing data variability and execution time, we assert that only metadata that influence the performance imputation method should be used to select a suitable imputation method. In addition, the following factors with respect to meta-data must be considered.

Missing ratios: When the ratio of missing to present data increases, the error of the imputation also increases and the difference in performance of the imputation method compared to other methods becomes larger. Each imputation method has a different pattern of performance for a given missing ratio (Hengpraphrom et al., 2010; Su et al., 2008).

Missing value distribution: For any given missing ratio, each imputation method has a different performance pattern according to the distribution of missing cells. For example, even if the same imputation method is used repeatedly, its performance may change according to the probability of missing cells in each feature (Wasito & Mirkin, 2006). Various patterns of missing data, such as missing completely at random (MCAR) and missing at random (MAR), can cause differences in the performance of the imputation method (Ghannad-Rezaie, Soltanian-Zadeh, Ying, & Dong, 2010). Here, MCAR refers to a missing data process that does not depend on either observed or missing values, whilst MAR is defined as a situation in which missingness depends on observed values, not on unobserved values (Wang, Xie, & Fisher, 2011).

Data set characteristics: The characteristics of a data set, such as the degree of imbalance, the size of the sample, and the number of features, influence imputation performance (Sim et al., 2015) because an imputation method is a form of machine learning. The performance of a machine learning algorithm depends on the characteristics of the data set (Kwon & Sim, 2013).

2.2. Selection of classifiers

The classification algorithm is one of the most important functions in the analysis of large data sets. Classification algorithms are the most widely used data mining models to extract valuable knowledge from huge amounts of data (Dogan & Zuhal, 2013). Classification is a data mining process that assigns items in a collection to target categories or classes. The goal of classification is to predict a target class for each case in the data set accurately (Akhila, Madhu, Madhu, & Pooja, 2014). Many comparative analyses are used to determine which algorithm is best suited for a particular data set. Classification capability depends on the types of algorithms and the characteristics of the data, such as the degree of imbalance, number of features, number of instances, and number of class types (Kwon & Sim, 2013; Liu & Zhou, 2006; Okamoto, 1963; Raudys & Pikelis, 1980). There is no superior classification algorithm for all types of data sets, because each classification algorithm is affected by the characteristics of the data set (Kwon & Sim, 2013). Moreover, when missing values are treated by a certain imputation method, the classification algorithm is also affected by the imputation method. Thus, each different imputation method/classifier pair results in a different performance, even if they treat the same data with the same missing values (Bastista & Monard, 2003; Farhangfar et al., 2008; Silva and Hruschka, 2013).

The descriptions of this causal relationship in the literature are insufficient. Intuitively, it seems that the cause may be related to the choice of the method of estimation and model of the machine learning algorithm, as both imputation methods and classifiers are forms of machine learning. The machine learning algorithm builds a model via its own method. For example, J48 divides classes with a split point (Safavian, 1991), whereas SVM divides classes with outer boundary points, such as marginal vectors (Suykens, 1999), and k-NN divides classes with similar instances (Zhang, 2007). This means that the chosen imputation method must estimate the determinant point