Chemometrics and Intelligent Laboratory Systems 173 (2018) 51–64

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Distributed feature selection: A hesitant fuzzy correlation concept for microarray high-dimensional datasets

Mohammad Kazem Ebrahimpour, Mahdi Eftekhari *

Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

ARTICLE INFO

Keywords:
Distributed machine learning
Distributed feature selection
Hesitant fuzzy sets
Microarray high dimensional datasets
Divide and conquer feature selection

ABSTRACT

Feature selection has been the problem of interest for many years. Almost all existing feature selection approaches use all training samples and features at once to select salient features. These approaches are named centralized methods; however, there are other approaches that split the training data on their dimensions in order to run each batch on different clusters (Machine) for the cases which we are dealing with ultra-big data. In this paper, a novel distributed feature selection approach based on hesitant fuzzy sets is proposed. First, datasets are horizontally (by their features) divided into some subsets according to the information energies of hesitant fuzzy sets and shuffling. Then, on each subset our HCPF (Hesitant fuzzy set based feature selection algorithm using Correlation coefficients for Partitioning Features) is applied individually. Finally, a merging procedure is employed that updates the final feature subset according to improvements in the classification accuracy. The effectiveness of the proposed method has been evaluated by twenty two state-of-the-art distributed and centralized algorithms on eight well-known microarray high dimensional datasets. The experimental results reveal that the proposed method has achieved significant results compared to the other approaches due to the statistical non-parametric Wilcoxon signed rank test. Our experiments confirm that the proposed method is effective to tackle feature selection problem in terms of classification accuracy and dimension reduction in ultra-high dimensional datasets.

1. Introduction

In the last two decades, handling DNA microarray high dimensional datasets has created a new line of research in both machine learning [1–5] and bioinformatics [6–9]. These types of datasets suffer from small sample size and huge number of features since they measure gene expression [10]. Therefore, feature selection [11] plays a crucial role in DNA microarray datasets, which removes the irrelevant and redundant features from the dataset. Thus, the learning algorithms concentrate on the important aspects of features that are useful for future predictions.

Typically feature selection approaches are divided into three main groups [12–14]: filters, wrappers and embedded methods [13]. Filter approaches undergo feature selection process by considering nature characteristics of them [8]. Therefore, these approaches are fast and can be used when we are dealing with huge datasets; however, since they do not consider the classifier/regressor in their decision making process, their performance is not as well as other model based approaches [15]. On the other hand, wrapper approaches train a model for evaluating candidate subsets. Thus, they are more accurate than filters; on contrary, since they train a model for each candidate subset, they are computationally expensive. The embedded approaches try to obtain a good subset of features during the training phase. The logic behind these approaches is the more important features get higher weights in the trained model [16]. This idea makes much sense since higher weights means more impact on the outputs. The embedded approaches can be considered as a trade-off between filter and wrapper approaches since they are considerably accurate and fast.

Traditional feature selection algorithms consider we can load the whole data in one computer; therefore, they apply machine learning algorithms on the whole dataset at once. We call this process centralized machine learning [10,12,13,15]. For instance, Hoque et al. [15] considered the feature selection problem as an optimization multi-objective problem. They mentioned that feature selection has two general aims: selecting relevant features to class labels and avoiding redundancy among themselves. Thus, they used the multi-objective NSGA II algorithm in order to deal with it. Moreover, Canul-Reich et al. [17] introduced an iterative embedded feature selection algorithm

* Corresponding author.
E-mail address: m.eftekhari@uk.ac.ir (M. Eftekhari).

https://doi.org/10.1016/j.chemolab.2018.01.001
Received 3 September 2017; Received in revised form 29 November 2017; Accepted 4 January 2018
Available online 6 January 2018

called SVM-RFE. Their method in each iteration removes least important features regarding to SVM. Multi-Filter Multi-Wrapper (MFMW) is one of the hybrid methods which attempts to increase the classification accuracy and robustness of selected features [18,19]. Recently, ensemble feature selection attracts researchers. Ref [19] used different filter methods by various aggregation operators in order to generate the optimal subset of features.

Since data is growing pretty much every day and there is an important question that if we cannot load the whole dataset in a computer, how can we apply machine learning algorithms like feature selection methods on that dataset? The most straight forward answer to this question is using distributed machine learning. Which means dividing the data semantically and let them to run on different clusters (computers) in parallel and finally merging the results.

There are two types of partitioning data. vertically, i.e. by samples and horizontally, i.e. by features [20]. Since microarray high dimensional datasets are suffering from small sample size and extremely large feature size it is desirable to partition them vertically in order to reduce the complexity of feature selection algorithms. Bolon-Canedo et al. proposed distributed feature selection [21] and they applied their proposals on well-known filters feature selection algorithms in the literature. According to their results [20], none of their distributed approaches have any significant superiority compared to others. Furthermore, according to their paper we conclude that semantically partitioning the data is vital and it has a direct impact on the final subset of features. Moreover, the authors of this paper previously proposed an efficient feature selection algorithm called Hesitant fuzzy set based feature selection algorithm using Correlation coefficients for Partitioning Features (HCPF) [6] that suffers from computational complexity and it couldn't be applied on big data. Thus, the aforementioned reasons motivate us to develop a distributed version of HCPF. The main contributions of this paper are listed as follows:

- Proposing a semantically splitting approach by taking advantages of sophisticated ranking feature selection algorithms.
- Proposing a hybrid partitioning method by combination of information energies of hesitant fuzzy sets and shuffling as partitioning approach for reinforcing the search strategy to find more promising features.
- Proposing highly parallel HCPF algorithm [6] in order to deal with microarray high dimensional datasets.
- A comprehensive comparison with twenty seven state of the art algorithms is done and results are presented.
- For confirmation the results statistical non-parametric tests were applied on the average outcomes and the excellence of the proposed method is approved.

The rest of the paper is organized as follows. In Section 2, the fundamental concepts of Hesitant Fuzzy Sets (HFSs), ranking algorithms, similarity measures are explained as preliminaries. Then in section 3 the proposed method is presented. After that in section 4 the experimental results are given. Finally the paper is concluded in section 5.

2. Preliminaries

Fuzzy sets are introduced by Zadeh in 1965 [22]. He argued that in the logic, we don't have just zero or one (false or true). There are a lot of gray areas in between those bounds. After that there were a lot of extensions to the pure fuzzy sets called intuitionistic [23], type-2 [24] and hesitant fuzzy sets [25]. In the hesitant fuzzy sets which is proposed by Torra, he argued that the membership degree in a fuzzy set can be a vector instead of just one value and it opens a new line of research. In this section, the reader notices about some vital information that are used in the proposed method. Therefore, the hesitant fuzzy sets [26] and its necessary aspects are given in section 2.1. After that a brief explanation on filter algorithms and similarity measures are discussed.

2.1. Hesitant fuzzy sets

Definition1: Let X be a universe of discourse. A hesitant fuzzy set (HFS) A on X is defined in terms of function $h_A(x)$, and when is applied to X, will return a finite subset of [0, 1] [25].

$$A = \{\langle x, h_A(x)\rangle | x \in X\} \tag{1}$$

where $h_A(x)$ is the set of all possible values in the interval [0,1]. Indeed, $h_A(x)$ is named the Hesitant Fuzzy Elements (HFE) [27].

Definition2: Given a HFE h, and the lower and upper bounds of the element are defined as follows: Lower Bound: $h^-(x) = min\, h(x)$ and Upper Bound: $h^+(x) = max\ \ h(x)$ [25].

Definition3: For an HFS $A = \{\langle x, h_A(x)\rangle | x_i \in X, i = 1, 2, ..., n\}$, the information energy of A is defined as follows [28]:

$$E_{HFS}(A) = \sum_{i=1}^{n}\left(\frac{1}{l_i}\sum_{j=1}^{l_i} h^2_{A\sigma(j)}(x_i)\right) \tag{2}$$

where n is the cardinality of the universe of discourse, l_i is the number of experts and $h_{A\sigma(j)}(x_i)$ is the j^{th} element of i^{th} universe of discourse members in the respective HFS. Therefore, if Eq. (2) applies on each elements of a HFS the equation will become:

$$E_{HFS}A(i) = \left(\frac{1}{l_i}\sum_{j=1}^{l_i} h^2_{A\sigma(j)}(x_i)\right) \quad i = 1, 2, ..., n \tag{3}$$

Definition 4: For the two typical HFSs, the correlation between two HFSs is defined as [27]:

$$C_{HFS}(A, B) = \sum_{i=1}^{n}\left(\frac{1}{l_i}\sum_{j=1}^{l_i} h_{A\sigma(j)}(x_i)h_{B\sigma(j)}(x_i)\right) \tag{4}$$

where the parameters are the same as Eq. (2).

Definition 5: For the two typical HFSs A and B, the correlation coefficient between them is defined as follows [27].

$$\rho_{HFS}(A, B) = \frac{C_{HFS}(A, B)}{[C_{HFS}(A, A)]^{\frac{1}{2}}.[C_{HFS}(B, B)]^{\frac{1}{2}}} = \frac{\sum_{i=1}^{n}\left(\frac{1}{l_i}\sum_{j=1}^{l_i} h_{A\sigma(j)}(x_i)h_{B\sigma(j)}(x_i)\right)}{\left[\sum_{i=1}^{n}\left(\frac{1}{l_i}\sum_{j=1}^{l_i} h^2_{A\sigma(j)}(x_i)\right)\right]^{\frac{1}{2}}\left[\sum_{i=1}^{n}\left(\frac{1}{l_i}\sum_{j=1}^{l_i} h^2_{B\sigma(j)}(x_i)\right)\right]^{\frac{1}{2}}} \tag{5}$$

where it must satisfy the following circumstances:

1) $\rho_{HFS}(A, B) = \rho_{HFS}(B, A)$
2) $0 \le \rho_{HFS}(A, B) \le 1$
3) $\rho_{HFS}(A, B) = 1 \quad if\ A = B$

The first situation indicates that the correlation coefficient matrix must be symmetrical. The second circumstance implies that all