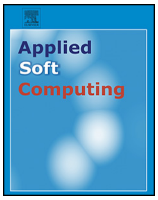
Contents lists available at ScienceDirect

Applied Soft Computing

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



Mutually-exclusive-and-collectively-exhaustive feature selection scheme

Chia-Yen Lee*, Bo-Syun Chen

Institute of Manufacturing Information Systems, National Cheng Kung University, Tainan City, 701, Taiwan

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 24 March 2017
Accepted 26 April 2017
Available online xxx

Keywords:
Feature selection
Mutually-exclusive-and-collectively-exhaustive
Data mining
Semiconductor manufacturing
Bioinformatics

ABSTRACT

In the fields of machine learning and data mining, feature selection methods are used to identify the most cost-effective predictors and to give a deeper understanding of pattern recognition and extraction. This study proposes a novel mutually-exclusive-and-collectively-exhaustive (MECE) feature selection scheme. Based on the MECE principle in decision science, the scheme, which has three stages including evaluation of independence, evaluation of importance and evaluation of completeness, aims to identify the independent and important variables with complete information. A case study of fault classification in semiconductor manufacturing and a study of breast cancer relapse identification in bioinformatics are used to validate the proposed scheme. The results demonstrate that the proposed MECE scheme selects fewer variables, avoids the multicollinearity problem, and improves fault classification accuracy in the two case studies.

1. Introduction

Feature selection is a technique of selecting a subset of features which characterizes the concept of target variables for the purpose of model construction [1]. Improvements in commonly used feature selection techniques are needed to reduce the computational burdens resulting from the expansion of modern high-throughput technologies accompanied by the exponential growth of the collected data. A “good” technique should reduce the actual costs of feature collection and pre-processing, and even improve classifier accuracy [17,19]. Motivated by the need to filter out more irrelevant, redundant, and biased variables in order to support model construction, we present a mutually-exclusive-and-collectively-exhaustive (MECE) feature selection scheme. The scheme has three stages including measuring the independence, measuring the importance, and measuring the completeness. We use the following two practical case studies involving the collection of large amounts of data to test and compare the scheme to commonly used selection methods.

The semiconductor manufacturing process is normally under a constant, real-time surveillance via the monitoring of signals collected from sensors or process points. As wafers moves through a factory, the process data from one manufacturing step is transferred to the next tool located in the next step. At the end of production, nearly 750 tigabytes (TB) of data can be associated with each wafer [26]. While these datasets provide unprecedented opportunities for effective controlling and optimizing the production process, the amount of information collected overwhelms the ability to identify key variables and detect faults in a timely manner.

Similarly, the field of bioinformatics has become increasingly dependent on microarray or high throughput sequencing technologies for identifying genes that are implicated in biological processes. The identified sequences or genes are being used, for instance, to classify future observations and to develop personalized treatments [11]. The number of variables in the raw data, however, ranges from 6000 to 60,000, and while some initial filtering can usually bring the number down to a few thousand, gene sequencing experiments are costly, and it is difficult to extract useful knowledge and patterns from the large and noisy datasets [12].

In both cases, feature selection can help to manage the resulting hundreds of thousands of variables and avoid the “curse of dimensionality” [13]. In particular, dimensionality reduction is a popular technique to remove noisy and redundant features. The techniques can be categorized into feature extraction and feature selection. Feature extraction projects the existing variables into a new feature space with lower dimensions, whereas feature selection selects a subset of the existing variables without a transformation [2]. Lots of previous studies introduce feature selection or feature extraction techniques [2,3,12]. The purpose of feature selection attempts to select the minimally sized subset of features for (1) improving prediction accuracy, and (2) approximating original class distribution

* Corresponding author.
E-mail addresses: cylee@mail.ncku.edu.tw (C.-Y. Lee), p98981057@mail.ncku.edu.tw (B.-S. Chen).

http://dx.doi.org/10.1016/j.asoc.2017.04.055
1568-4946/© 2017 Elsevier B.V. All rights reserved.

given only the values for the selected variables [1]. Indeed, feature selection technique benefits the classification performance of the predictors, selects cost-effective predictors, and provides a better understanding of pattern recognition and extraction. The latter, however, can generate selected variables with a multicollinearity problem or can select insufficient information to support decision-making.

Therefore, we develop a novel mutually-exclusive-and-collectively-exhaustive (MECE) feature selection scheme having the evaluation of independence, the evaluation of importance, and the evaluation of completeness [15].[1] This study contributes to the literature and provides the following benefits: (1) selecting independent and important variables with sufficient information for decision-making, 2) addressing the curse of dimensionality, 3) avoiding the multicollinearity problem, and 4) selecting variables retaining the original physical characteristics.

The remainder of the paper is organized as follows. Section 2 reviews the existing feature selection techniques and discusses their advantages and disadvantages. Section 3 presents the MECE feature selection scheme. Sections 4 and 5 describe the two case studies and the results of testing the proposed scheme. Also compares the proposed scheme to the other popular techniques. Section 6 concludes and suggests future work.

2. Feature selection techniques

Several classical feature selection techniques are commonly used to enhance the data quality and support pattern recognition in the fields of machine learning and data mining. In this study, based on the literature, we include but not limited to principal component analysis (PCA), clustering analysis, stepwise selection (SS), LASSO (least absolute shrinkage and selection operator) and random forest (RF). [18,21]. In particular, the former three are classical methods in the multivariate analysis while the other two are the neoclassical algorithms (named in this study) to address the high-dimensional issue and improve the robustness. Then, the proposed MECE scheme compares with these existing techniques via two types of datasets in Section 4. This section reviews these popular techniques.

2.1. Principal component analysis

Principal component analysis (PCA) is mathematically defined as an orthogonal linear combination of original variables that transforms the data into a new coordinate system for maximizing variance [10]. Observations project to the new space and the value presented on the first axis is called the first principal component (PC) which generally presents the greatest variance (or the greatest eigenvalue). The second PC with the second greatest variance is characterized by the second axis, and so on. In the new PC space, PCs are orthogonal and uncorrelated to each other. Thus, PCA can eliminate the multicollinearity problem and support dimension reduction by choosing fewer PCs instead of original variables. However, PCs generated by converting the original dimension into reduced dimension may partially lose the original physical characteristics and thus are difficult to interpret since they are a linear combination of multiple original variables. Note that, in this study, we don't apply PCA to our case studies since we focus on the original variable selection in semiconductor manufacturing process and gene analysis but PCA conducting variable transformation to PC may lose the original interpretation.

2.2. Clustering analysis

Clustering is based on the concept of replacing a group of similar variables (defined within the same cluster) by a cluster centroid, which becomes a selected feature. In general, there are two categories of clustering methods: the hierarchical method and the nonhierarchical method.

Ward's clustering method, also called the minimum variance method, is an example of the hierarchical method [16]. Given n observations, it starts with n clusters with size 1 (i.e., each cluster has only one observations individually) and continues grouping the observations at each step based on minimum variance criterion (i.e., error sum of square) until all observations are merged into one cluster. Let x_{ik} denote the vector of variables regarding to observation i in cluster k, $\bar{x}_{\cdot k}$ be the vector with the average of variables regarding to observations within cluster k, and $\bar{x}_{\cdot\cdot}$ be the vector with the average of variables regarding all observations. Ward's method defines three metrics according to the squared Euclidean distance between points. The error sum of square (ESS), total sum of square (TSS) and R-square (R^2 or coefficient of determination) are defined as:

$$ESS = \sum_i \sum_j \|x_{ik} - \bar{x}_{\cdot k}\|^2 \tag{1}$$

$$TSS = \sum_i \sum_j \|x_{ik} - \bar{x}_{\cdot\cdot}\|^2 \tag{2}$$

$$R^2 = \frac{TSS - ESS}{TSS} \tag{3}$$

Ward's method generates the best combination with the smallest ESS (or the greatest R-square alternatively), then goes to the next step for forming n-2 clusters, and so on. The algorithm terminates when all observations are merged into one single cluster with size n. The disadvantage is that once the variable is assigned to a cluster at an early stage, the variable cannot be reallocated again. That is, the technique usually generates "local optimum".

K-means clustering is an example of the non-hierarchical clustering method [22]. Given a desired number of cluster, K-means reassigns each observation to clusters iteratively. The objection function is to minimize a squared error function as:

$$\sum_i \sum_j \|x_i^{(k)} - c_k\|^2 \tag{4}$$

where $\|x_i^{(k)} - c_k\|^2$ is a distance measure between the observation i located in cluster k (i.e., $x_i^{(k)}$) and the cluster centroid c_j. This squared error function is an indicator measuring the distance between n observations and its corresponding cluster centroids. The drawback is that K-means is sensitive to the outlier and, therefore may "lose" the physical characteristics of the original variables. To address the problem, the K-medoids method selects the original variable which is most centrally located in a cluster as a reference point instead of taking the mean value [14]. The current study adopts two-phase clustering selection (TPS) including Ward's method (i.e., hierarchical clustering) in the first phase to determine the number of clusters which used in K-means clustering (i.e., non-hierarchical clustering) in second phase. Even though clustering analysis can identify the independent feature and reduce the dimension, the methods cannot ensure the importance of the selected features.

2.3. Stepwise selection

Stepwise selection (SS), or stepwise regression, allows moves in forward and backward directions, dropping or adding variables step by step [9]. Forward selection begins with a regression model including the most significant variable via statistical test. It adds one variable at a time and continues adding variables until none

[1] In general, we use without distinction the terms "variable" and "feature" when there is no impact on the selection algorithms. In fact, "variable" refers to the raw input factors and "features" is constructed by these raw input factors [3].

Please cite this article in press as: C.-Y. Lee, B.-S. Chen, Mutually-exclusive-and-collectively-exhaustive feature selection scheme, Appl. Soft Comput. J. (2017), http://dx.doi.org/10.1016/j.asoc.2017.04.055