Expert Systems with Applications 42 (2015) 8259–8273

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Diversity-driven generation of link-based cluster ensemble and application to data classification

Natthakan Iam-On [a,*], Tossapon Boongoen [b]

[a] School of Information Technology, Mae Fah Luang University, Tasud, Muang District, Chiang Rai 57100, Thailand
[b] Department of Mathematics and Computer Science, Royal Thai Air Force Academy, Bangkok 10220, Thailand

ARTICLE INFO

Article history:
Available online 18 July 2015

Keywords:
Ensemble clustering
Data classification
Optimization
Feature transformation

ABSTRACT

Over decades, a large number of research studies have concentrated on improving the accuracy of classification model. This is the case as several types of classifiers prove to be useful in real-life problems, including the prediction of system failure risk and microarray-based cancer diagnosis. Despite this, the accuracy of existing classifiers has been constrained by uninformative variables typically observed in modern data. In addition to feature selection, one may transform the original data to another variation, where only key feature components are included. Unlike conventional transformation-based techniques found in the literature, this paper presents a novel method that makes use of cluster ensembles, specifically the summarized information matrix, as the transformed data for the following classification step. Among different state-of-the-art methods, the link-based cluster ensemble approach (LCE) provides a highly accurate clustering, and thus particularly employed here. This is uniquely coupled with a diversity-driven generation of ensemble, which provides informative and diverse sets of clusterings. The performance of this transformation model is evaluated on published synthetic, standard and gene expression datasets; using C4.5, Naive Bayes, KNN, Neural Network and Random Forest classifiers; in comparison with benchmark techniques. The findings suggest that the new model can improve the classification accuracy of original data and performs better than the other transformation methods investigated in the empirical study.

1. Introduction

Classification that is the task of assigning objects of interest to one of several predefined categories or classes. It is a pervasive problem encompassing many diverse applications. These include detecting spam email messages based upon the message header and content, categorizing cells as malignant or benign based on the results of MRI (Magnetic Resonance Imaging) scans, and classifying observed objects based upon their shapes. In recent decades, the development of high-throughput technologies has resulted in the exponential growth of harvested data with respect to dimensionality and sample size. As such, effective management and analysis of these data becomes increasingly challenging. For instance, in the biomedical domain, microarray technology has been regarded as an essential tool, where expression levels of a large number of genes can be simultaneously monitored and studied (Kim & Lee, 2007). It inspires several down-streaming applications, including the identification of differentially expressed genes for molecular studies or drug therapy response (Ramaswamy, Ross, Lander, & Golub, 2003; Tusher, Tibshirani, & Chu, 2001; Wallqvist, Rabow, Shoemaker, Sausville, & Covell, 2002) and the creation of classification systems for improved cancer diagnosis (Cleator & Ashworth, 2004; Spang, 2003; Wang et al., 2004). The significance of data from DNA microarrays in cancer diagnosis, and its advantage over the traditional methods of morphological appearance-based classification, have been widely recognized in bioinformatics and medical communities (de Souto, Costa, de Araujo, Ludermir, & Schliep, 2008; Golub et al., 1999; Khan et al., 2001).

This sort of complex data is typically of high dimensionality with a large variable (or feature) per sample ratio. However, it has been observed that although there are hundreds or thousands of variables for each sample, a few may account for much of the data variation (West et al., 2001). With this characteristic, an effective prediction may not be achieved due to the fact that any proximity or similarity metric exploited by a classification model is increasingly inaccurate as dimensionality increases (Boongoen, Shang, Iam-On, & Shen, 2011). To deal with the problem, one

* Corresponding author.
E-mail addresses: nt.iamon@gmail.com (N. Iam-On), tossapon_b@rtaf.mi.th (T. Boongoen).

http://dx.doi.org/10.1016/j.eswa.2015.06.051
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

may reduce dimension of the data by selecting a subset of interesting features; or generating feature components/super features (i.e., combinations of features), which are widely known as feature selection and transformation, respectively. Following these approaches to dimensionality reduction, the conventional classification procedure has been re-shaped and comprises two steps of: data preprocessing by dimensionality reduction; and the classification process, in which samples are classified into known categories by applying standard statistical or machine learning models (Nguyen & Rocke, 2002). Recently, feature transformation techniques have been effective for electromyography (EMG) signal classification (Phinyomark, Phukpattaranont, & Limsakul, 2012) and text categorization (Bharti & Singh, 2015).

Dimensionality reduction techniques transform data for sake of easier computation, modeling and inference by analysis of variable interdependence and interobject similarity (Carroll, Green, & Chaturvedi, 1997). On one hand, feature selection algorithms try to search for the most valuable feature subset heuristically under certain predefined feature subset evaluation criterion (Dettling & Buhlmann, 2003). On the other, feature transformation methods like principal component analysis or PCA (Domeniconi & Gunopulos, 2008) transform original features into some new features, which are probably difficult to interpret for human beings (Duda, Hart, & Stork, 2012). Specific to the latter category, PCA has shown to be efficient for finding the underlying feature components (Bicciato, Luchini, & Bello, 2003). As such, this linear transformation has been extensively used in gene expression data analysis (Yeung & Ruzzo, 2001). In addition, kernel principal component analysis (KPCA) is put forward as a nonlinear variation of PCA, with many successful applications in the field of machine learning (Ng, Jordan, & Weiss, 2001). These initial algorithms have motivated the development of several advanced techniques such as Isometric Projection (Cai, He, & Han, 2007), Neighborhood Preserving Embedding (He, Cai, Yan, & Zhang, 2005a), and Locality Preserving Projection (He, Yan, Hu, Niyogi, & Zhang, 2005b).

Unlike these transformation methods, the use of clustering information in addition to the original feature has been recently reported to improve the accuracy of intrusion detection problem (Nguyen, Harbi, & Darmont, 2011). In particular, a fuzzy clustering technique is applied to the data under examination, and the resulting memberships of each sample to k clusters (where k is a user-defined number of clusters) are used to form additional k variables or features for the following classification step. Similar classification approaches of Sang-Woon (2010) and Nasierding, Tsoumakas, and Kouzani (2009) have combined cluster labels and conventional supervised algorithms for face recognition and image annotation, respectively. Inspired by this integrative approach to feature transformation and the success of link-based ensemble clustering or LCE (Iam-On, Boongoen, & Garrett, 2010; Iam-On, Boongoen, Garrett, & Price, 2011; Iam-On, Boongoen, Garrett, & Price, 2012) in recent years, the paper presents a novel use of ensemble-information matrix in the context of classification problem. With respect to LCE, the refined sample-cluster association matrix can be considered as the representation of samples in the transformed space, which is discovered from multiple clusterings in the setting of original features. Having accomplished this, the initial data dimensions are reduced to a set of cluster labels, with which each sample associates to a certain degree.

The proposed approach is largely different from the aforementioned clustering-oriented methods, which make use of the result from a single run of clustering technique. In contrary, this method concentrates on transforming the original data to another information matrix that represents associations within a cluster ensemble. As LCE has proven effective for the clustering problem, the richness of its underlying link-based matrix may well boost the accuracy of a classification model. Note that the present study of ensemble clusterings for the task of feature transformation follows the initial investigation with microarray data analysis (Iam-On & Boongoen, 2013). With the aim to extend the previous framework, the quality of ensemble, hence that of the resulting data matrix, is optimized through the diversity-driven selection of ensemble members. This is motivated by previous studies of Fern and Lin (2008) and Kuncheva and Vetrov (2006), which reported the improvement of ensemble clustering with a set of medium to highly diverse base partitions. Building on existing works, the contribution of this research is a new collective framework where aggregation of multiple clusterings and ensemble generation methods are brought together for the classification problem. In addition, this can be generalized to many models developed in the field of cluster ensemble. The research also provides a new and original insight to the integrative application of link-based ensemble clustering and intelligent optimization to classification problem, which has been one of the major fields of data analysis. Also, the empirical study would reveal facts regarding the relation between level of ensemble diversity and the accuracy of down-steaming classifier. The corresponding findings will be useful as a guideline to applying the aforementioned model to any particular classification problem.

The rest of this paper is organized as follows. Section 2 provides background knowledge of ensemble clustering as to set the scene for the concepts and notations used throughout the paper. Following that, Section 3 introduces the ensemble-based data transformation framework, including the generation of cluster ensemble with diversity optimization and the summarization of selected data partitions into ensemble-information matrix. The link-based matrix refinement is also emphasized here. Section 4 includes the evaluation of the proposed model as compared to several well-known unsupervised transformation methods, using conventional classification techniques like C4.5, k-Nearest Neighbors (KNN), Naive Bayes, Neural Network and Random Forest; on synthetic and benchmark datasets. The paper is concluded in Section 5 with the perspective of future research.

2. Basis of ensemble clustering

Ensemble clustering, sometimes referred to as consensus clustering or cluster ensemble, has recently become an attractive alternative for analyzing complex data, especially those obtained from microarray experiments (Iam-On et al., 2010; Kim, Kim, Ashlock, & Nam, 2009; Monti, Tamayo, Mesirov, & Golub, 2003; Yu, Wong, & Wang, 2007). This meta-learning methodology is motivated by the fact that the performance of most clustering techniques are highly data dependent. One method may produce an acceptable result on one dataset, but possibly become less successful with others (Duda, Hart, & Stork, 2000; Fred & Jain, 2005; Xue, Chen, & Yang, 2009). To tackle this, many researchers working in the field of data mining or machine learning have attempted to combine multiple clusterings into a single consensus clustering, with higher quality than those initial decisions. This process can provide more robust and stable solutions across different problem domains and datasets (Fred & Jain, 2005; Iam-On et al., 2010; Topchy, Jain, & Punch, 2005).

Formally, the problem of ensemble clustering can be defined as follows. Let $X = \{x_1, \dots, x_N\}$ be a set of N data points or samples, where each $x_i \in X$ is represented by a vector of D feature or attribute values, i.e., $x_i = (x_{i,1}, \dots, x_{i,D})$. Also, let $\Pi = \{\pi_1, \dots, \pi_M\}$ be a cluster ensemble with M base clusterings, each of which is referred to as an ensemble member. Each base clustering returns a set of clusters $\pi_g = \{C_1^g, C_2^g, \dots, C_{k_g}^g\}$, such that $\bigcup_{t=1}^{k_g} C_t^g = X$ and $\bigcap_{t=1}^{k_g} C_t^g = \emptyset$, where k_g is the number of clusters in the g^{th} clustering. For each $x_i \in X, C^g(x_i)$ denotes the cluster label in the g^{th} base clus-