Journal of Visual Languages and Computing 33 (2016) 24–36

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

View points

Classifying high dimensional data by interactive visual analysis

Ke-Bing Zhang [a,*], Mehmet A. Orgun [a], Rajan Shankaran [a], Du Zhang [b]

[a] Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
[b] Faculty of Information Technology, Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau

ARTICLE INFO

Article history:
Received 19 November 2015
Accepted 24 November 2015
Available online 2 December 2015

Keywords:
Interactive Visual Analysis (IVA)
Classification
Visual classifer
Data projection

ABSTRACT

Data mining techniques such as classification algorithms are applied to data which are usually high dimensional and very large. In order to assist the user to perform a classification task, visual techniques can be employed to represent high dimensional data in a more comprehensible *2D* or *3D* space. However, such representation of high dimensional data in the *2D* or *3D* space may unavoidably cause overlapping data and information loss. This issue can be addressed by interactive visualization. With expert domain knowledge, the user can build classifiers that are as competitive as automated ones using a *2D* or *3D* visual interface interactively. Several visual techniques have been proposed for classifying high dimensional data. However, the user's interaction with those techniques is highly dependent on the experience of the user in the visual identification of classifying data, and as a result, the classification results of those techniques may vary and may not be repeatable. To address this deficiency, this article presents an interactive visual approach to the classification of high dimensional data. Our approach employs *the enhanced separation feature* of a visual technique called HOV^3 by which the user plots the training dataset by applying statistical measurements on a *2D* space in order to separate data points into groups with the same class labels. A data group with its corresponding statistical measurement which separated it from the others is taken as a visual classifier. Then the user mixes the data points in a classifier with the unlabeled dataset and plots them in HOV^3 by the measurement of the classifier. The data points which overlap the labeled ones in the *2D* space are assigned the corresponding label. Our approach avoids the randomness in the existing interactive visual classification techniques, as the visual classifier in this approach only depends on the training dataset and its statistical measurement. As a result, this work provides an intuitive and effective approach to classify high dimensional data by interactive visualization.

1. Introduction

Classification refers to the task of predicting class labels of items in an unlabeled dataset. The process of classification has two steps, namely, the *learning step* and the *classification step*. In the *learning step*, the user builds up a model for classification using the training dataset by a particular technique. The classification model built up by the user is called a classifier. Then the user applies the classifier and predicts class labels from an unknown dataset in the *classification step* [7]. The increasing computation capability of modern computers has allowed the development of many automated and/or algorithmic

* Corresponding author.
E-mail addresses: kebing.zhang@mq.edu.au (K.-B. Zhang), mehmet.orgun@mq.edu.au (M.A. Orgun), rajan.shankaran@mq.edu.au (R. Shankaran), duzhang@must.edu.mo (D. Zhang).

http://dx.doi.org/10.1016/j.jvlc.2015.11.003
1045-926X/

methods to be used effectively in classification applications [24].

Usually, designing a classification algorithm is an iterative process, in which several decisions and choices are involved, such as, picking a satisfied classifier, adjusting the parameters of the classifier, modifying the classifier for specific cases and feature selection [1]. To optimize the performance of the classifier, this process can be iterated while considering measuring criteria such as the error rate. Therefore, algorithmic or automated classification approaches need the involvement of the user so that the user can interactively tune parameters for the classification process and test his/her modeling assumptions. It is in general impossible to achieve the above adjustment by a fully automated approach.

Thus, for each of the above-mentioned stages in the classification process, the user needs to understand the studied data intuitively in order to have a better classification result. Clearly, without any intuitive insight of the data, the user may lack a full understanding of the underlying classification model. To solve the issue caused by automated approaches, Ware et al. [23] state that the user can build classifiers that can compete with automated approaches by using a simple two-dimensional (2D) visual interface, because he/she can leverage the domain knowledge in the process of building up a classifier.

A visual interface could assist the user by providing an intuitive visual insight in the process of discovering knowledge [12]. In particular, interactive visualization may be able to assist the user to build up a better classification model [2], as Interactive Visual Analysis (IVA) is a suitable technique for analyzing high-dimensional and very large data, and for providing an understanding of the data provided by simple graph manipulation and non-interactive techniques [11]. As a result, IVA could assist the user in building up, optimizing and comparing classifiers [1].

Several visual techniques and tools have been developed to support the user on classifying data in data mining [2,21,9,23]. However, the process of classification by those visual techniques depends overly on the experience of the individual user during data classification, which renders those techniques ineffective for classifying unlabeled data interactively.

To remedy the problem of randomness of user's interaction in the aforementioned techniques, we employ an IVA technique, called *Hypothesis-Oriented Verification and Validation by Visualization* (HOV^3) to visualize high dimensional data as a collection of points scattered in a *2D* space [27]. The main advantage of this approach is that point-based visualization techniques are able to display more data items on the limited space of a monitor screen, and also provide a better comprehension of the relationships among the data items. In this research, we leverage the enhanced separation feature of HOV^3 to determine how dissimilar the data points are from one another in the HOV^3 *2D* space [28] so that they can then be separated.

In practice, in the learning step of our approach, the labeled training data points are first plotted by the user in a *2D* space in HOV^3 using the enhanced separation feature of HOV^3. Then the user separates each predefined group in the training dataset by applying a measure vector, once or as many times as required to the training dataset. The data items of a well-separated predefined group together with its measure vector make up a visual classifier which is applied to the unlabeled dataset in the classification step of our approach. In practice, the user first mixes the data items of a visual classifier and the unlabeled dataset together. Then the user plots the mixed dataset by applying the measure vector of the visual classifier in HOV^3. Intuitively, the class label is assigned to the unlabeled data points which overlap the labeled data points in the well-separated predefined group of the classifier.

Note that this paper is an extended and revised version of [26]. It provides a complete description of our approach as well as a detailed discussion of revised experiments. In the rest of the paper, Section 2 reviews several visual classification techniques that are most relevant to our work. Section 3 discusses the HOV^3 technique and the enhanced separation feature of HOV^3. Section 4 provides a detailed explanation of the classification technique by HOV^3. Section 5 presents the performance analysis of our approach, and also compares it with several existing classification techniques. The contributions of this paper are summarized in Section 6. Finally, we discuss future research directions in Section 7.

2. Classification by visualization

2.1. The role of visualization in classification

One of the key goals of data mining is prediction. As a widely applied data mining technique, classification is a typical predictive process to employ a classifier as the prediction to determine the membership of unknown objects.

Many algorithmic and automatic approaches have been proposed to address the requirements of a broad and diverse range of applications of classification [22]. Given the ever increasing amount of complex and huge data produced in the real world, the user can hardly rely on fully automatic techniques to analyze such data. In fact, there is a need for an iterative and interactive process in-between automated techniques and human intervention while analyzing such data. For example, algorithmic classification methods need the user to be involved in the classification process to tune parameters so that he/she can intuitively and interactively refine his/her predictions. However, the process of revision and refinement cannot be achieved by using a fully automated approach exclusively.

Graphic displays can help human users to obtain most of the information of an object, because, from the user's point of view, visual thinking can assist the user to observe and understand complex data more effectively. Therefore, it has been observed that interactive visual techniques could play an important role in the process of classification [8]. Following this observation, several visual techniques have been employed in the applications of classification. We discuss some of the techniques that are most relevant to this research below.