Computers in Industry 74 (2015) 201–211

Contents lists available at ScienceDirect

Computers in Industry

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

Evolutionary feature and instance selection for traffic sign recognition

Zong-Yao Chen [a], Wei-Chao Lin [b], Shih-Wen Ke [c], Chih-Fong Tsai [a,*]

[a] Department of Information Management, National Central University, Taiwan
[b] Department of Computer Science and Information Engineering, Hwa Hsia University of Technology, Taiwan
[c] Department of Information and Computer Engineering, Chung Yuan Christian University, Taiwan

ARTICLE INFO

Article history:
Received 21 January 2015
Received in revised form 7 August 2015
Accepted 18 August 2015
Available online 11 September 2015

Keywords:
Traffic sign recognition
Genetic algorithm
Feature selection
Instance selection
Data mining

ABSTRACT

The problem of traffic sign recognition is generally approached by first constructing a classifier, which is trained by some relevant image features extracted from traffic signs, to recognize new unknown traffic signs. Feature selection and instance selection are two important data preprocessing steps in data mining, with the former aimed at removing some irrelevant and/or redundant features from a given dataset and the latter at discarding the faulty data. However, there has thus far been no study examining the impact of performing feature and instance selection on traffic sign recognition performance. Given that genetic algorithms (GA) have been widely used for these types of data preprocessing tasks in related studies, we introduce a novel genetic-based biological algorithm (GBA). GBA fits "biological evolution" into the evolutionary process, where the most streamlined process also complies with reasonable rules. In other words, after long-term evolution, organisms find the most efficient way to allocate resources and evolve. Similarly, we closely simulate the natural evolution of an algorithm, to find an option it will be both efficient and effective. Experiments are carried out comparing the performance of the GBA and a GA based on the German Traffic Sign Recognition Benchmark. The results show that the GBA outperforms the GA in terms of the reduction rate, classification accuracy, and computational cost.

1. Introduction

The recognition of traffic signs has become a challenging problem in computer vision to be used in vehicles, where perception and correct interpretation of traffic signs have a critical impact on driver safety [1].

Traffic sign recognition (TSR) can be regarded as an image classification problem. In general, it can be approached by constructing a classifier trained using a given set of traffic sign images represented by relevant image features. Once the classifier is trained, it is able to classify new unknown traffic signs into the classes of the training images.

From the data mining viewpoint, the process generally contains a number of steps, such as dataset selection, data preprocessing, data analysis, and result interpretation and evaluation [2,3]. Data preprocessing is one of the most important steps of data mining. Specifically, the aim of data preprocessing is to make the chosen dataset as 'clean' as possible for eventual analysis and evaluation. In other words, no quality data, no quality mining results should be missing [4,5].

In order to effectively preprocess the chosen dataset, *feature selection* (or dimensionality reduction) and *instance selection* (or record reduction) are two of the more active research problems in data mining. This is because, in most real-world data mining problems, including TSR, the number of features and data samples selected is usually very large. However, there has as yet been no related study focused on examining the effect of performing data preprocessing on TSR performance.

In particular, to automatically recognize a specific traffic sign, the image or video stream is processed by feature extraction for the later recognition where some relevant visual features, such as color and texture, are extracted to represent or describe its visual content. However, feature extraction usually results in a certain high dimensionality; say over a thousand features (c.f. Section 4.1). Therefore, if too many features are used for data analysis (or TSR is this paper), it can cause the curse of the dimensionality problem [6]. Since not all of the pre-chosen features are informative in many cases, especially when the feature dimensionality is very large, the objective of feature selection is to select more representative features which have more discriminative power over a given dataset. This is also called dimensionality reduction [7].

On the other hand, given a training set composed of a certain number of data samples (i.e. images in this paper), it is usually the

* Corresponding author.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

http://dx.doi.org/10.1016/j.compind.2015.08.007
0166-3615/

case that the dataset may contain some proportion of noisy data, which can affect the effectiveness of training a classifier. That is, in TSR, if the chosen training set contains some outliers but outlier removal is not considered, the recognition rate of a classifier could be affected.

According to Aggarwal and Yu [8] and Barnett and Lewis [9], it is often true that data points are not all equally informative and some data points will be further away from the sample mean than is deemed reasonable. This is similar to the process of data reduction aimed at the discarding of faulty data (or outliers), which should be considered as noisy points lying outside a set of defined clusters and could lead to significant performance degradation. The performance in data mining tasks such as classification or prediction performance will very likely be poorer if the instance selection step is not considered [10,11].

Genetic algorithms (GA), one of the most widely used techniques for feature and instance selection, can improve the performance of data mining algorithms [12,13]. In particular, Cano et al. [14] have shown that better results can be obtained with a GA than with many traditional and non-evolutionary instance selection methods in terms of better instance selection rates and higher classification accuracy. Moreover, GAs have proven to be suitable for large-scale feature selection problems [15]. However, GAs only pursue the simplest evolutionary process.

Therefore, in this study a genetic-based biological algorithm (GBA) is proposed for both feature and instance selection. The GBA simulates the biological evolution process and rules of nature where, after long-term evolution, organisms find the most efficient way to allocate resources and evolve [16]. Inspired by nature, the GBA is constructed as an efficient and effective means of problem solving for instance selection.

The rest of this paper is organized as follows. Section 2 describes the concepts of feature selection and instance selection. In addition, an overview of traffic sign recognition is provided. The proposed GBA method for feature and instance selection is introduced in Section 3. Section 4 presents the research design and experimental results. Finally, some conclusions are given in Section 5.

2. Literature review

In this literature review section, two important data preprocessing steps are overviewed, which are feature selection and instance selection. Then, brief discussions of GAs and related optimization methods are provided. Next, related literatures of traffic sign recognition are described.

2.1. Feature and instance selection

2.1.1. Feature selection

It is usually the case that the number of features (or variables) collected in a dataset is relative large (i.e., the curse of dimensionality), and not all of these features are informative or can provide high discriminative power [17]. The aim of feature selection is to remove the irrelevant and/or redundant features from the chosen dataset, improving the performance of the classification and/or clustering algorithms. In addition, given a dataset, feature selection can help analysts understand which features are important as well as how the features are related.

Feature selection can be defined as the process of choosing a minimum subset of M features from the original dataset of N features ($M < N$), so that the feature space (i.e., the dimensionality) is optimally reduced according to the following evaluation criteria [18]:

- the classification accuracy does not significantly decrease; and
- the resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features.

A feature selection algorithm usually consists of four steps: subset generation, subset evaluation, stopping criterion, and result validation [18]. Subset generation is a search procedure which generates subsets of features for evaluation. Each subset generated is evaluated by a specific evaluation criterion and compared with the previous best one with respect to this criterion. If a new subset is found to be better, then the previous best subset is replaced by the new subset.

The reader can refer to Kudo and Sklansky [15] and Guyon and Elisseeff [7] for more information.

2.1.2. Instance selection

Regarding Wilson and Martinez [19], one problem with using the original data points is that there may not be any data points located at the precise points that would make for the most accurate and concise concept description. Therefore, the aim of instance selection, or record reduction, is to reduce a dataset while still maintaining the integrity of the original dataset. In some cases, generalization accuracy can increase when noisy instances are removed and when decision boundaries are smoothed to more closely match the true underlying function.

These instances can also be regarded as outliers (or bad data). Specifically, outliers are the data points which are highly unlikely to occur given a model of the data. One approach to performing this task is to derive the distances to neighboring data points by implementing a clustering algorithm [20].

Instance selection can be defined as follows. Let X_i be an instance where $X_i = (X_{i1}, X_{i2}, \ldots, X_{im}, X_{ic})$ meaning that X_i is represented by m-dimensional features and X_i belongs to class c given by X_{ic}. Now, assume that there is a training set TR which consists of M instances and a testing set TS composed of N instances. If $S \subseteq TR$ is the subset of selected samples that are produced by an instance selection algorithm, then we can classify a new pattern T from TS over the instances of S.

2.1.3. Discussion

Many related studies have shown the superiority of GAs over other feature and instance selection methods. For example, Kudo and Sklansky [15] conducted a comparative study of algorithms for large-scale feature selection (where the number of features is over 50). They show that GAs are suitable for large-scale problems. Moreover, Nanni and Lumini [21] compared several reduction methods, such as fuzzy clustering, particle swarm optimization (PSO), and GA. They found that the GA approach performs best.

There are some other related optimization algorithms proposed in literature. For example, Valdez et al. [22] develops a fuzzy logic system to combine the result of PSO and GA, which is a kind of multi-population method. Although it is a good idea to find out a better solution, the time cost of searching for the best solution is larger than using PSO or GA alone since more individuals are used and they need to be evaluated. In Precup et al. [23], the gravitational search algorithm (GSA) is introduced to reduce the parametric sensitivity of the fuzzy control systems. The solution quality of GSA is impeccable. However, according to the no free lunch theory [24], it is difficult to focus on both speed and quality simultaneously. In GSA, its computational cost is still higher than general GAs. On the other hand, Zăvoianu et al. [25] and El-Hefnawy [26] propose some relevant methods for optimizing the performance of electrical drives and bi-level problems. However, they do not consider the time cost and the evaluation time (i.e. the number of visits to the fitness function). Although there are many related algorithms, there is no a comprehensive study to compare