Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 1384–1393
Expert Systems with Applications
www.elsevier.com/locate/eswa

Classifier design with feature selection and feature extraction using layered genetic programming

Jung-Yi Lin [a,*], Hao-Ren Ke [b], Been-Chian Chien [c], Wei-Pang Yang [a,d]

[a] Department of Computer Science, National Chiao Tung University, 1001 Ta Hsueh Road, HsinChu 300, Taiwan
[b] Library and Institute of Information Management, National Chiao Tung University, Taiwan
[c] Department of Computer Science and Information Engineering, National University of Tainan, Taiwan
[d] Department of Information Management, National Dong Hwa University, Taiwan

Abstract

This paper proposes a novel method called FLGP to construct a classifier device of capability in feature selection and feature extraction. FLGP is developed with layered genetic programming that is a kind of the multiple-population genetic programming. Populations advance to an optimal discriminant function to divide data into two classes. Two methods of feature selection are proposed. New features extracted by certain layer are used to be the training set of next layer's populations. Experiments on several well-known datasets are made to demonstrate performance of FLGP.

Keywords: Feature generation; Feature selection; Pattern classification; Genetic programming; Multi-population genetic programming; Layered genetic programming

1. Introduction and review

1.1. Feature selection

This research concentrate on three research topics: feature selection, feature generation, and classifier design. Feature selection is an important technique of pattern recognition dealing with raw features. It focuses on removing useless, irrelevant, and redundant features. The classification accuracy of data derived by selected features is better than that by no selection. Many research working on feature selection have been proposed (Ahmad & Dey, 2005; Dash & Liu, 1997; Jain & Zongker, 1997; John, Kohavi, & Pfleger, 1994; Kittler et al., 1978; Kohavi & John, 1997; Kudo & Sklansky, 2000; Pernkopf, 2005; Pudil, Novovicova, & Kitter, 1994). John et al. (1994) divide feature selection methods into *filter* group and *wrapper* group. Filter methods applies by measuring the degree of feature relevance and deciding to leave or remove it. By contrast, wrapper methods (Jain & Zongker, 1997; Kohavi & John, 1997) function by cooperating with particular classifiers. Dash and Liu (1997) did a solid survey on many feature selection methods. They categorize feature selection methods into five based on the evaluation of features discrimination ability of: *distance measure*, *information measure*, *dependence measure*, *consistency measure*, and *classifier error rate*. Table 1 shows the five categories (Dash & Liu, 1997). Both Jain and Zongker (1997) proposed taxonomies of feature selection methods as shown in Fig. 1. In Jain and Zongker (1997), experiments with 15 different feature selection algorithms on extracted 18 features of SAR images are made. Kudo and Sklansky also compare excellently the differences among many feature selection methods in Kudo and Sklansky (2000). There are 16 different feature selection methods in comparison with 1-NN classifier, including sequential algorithms and branch-and-bound algorithms. Besides, they also give comments on these methods and

* Corresponding author. Tel.: +886 922 912 530.
E-mail addresses: jylin@cis.nctu.edu.tw (J.-Y. Lin), claven@lib.nctu.edu.tw (H.-R. Ke), bcchien@mail.nutn.edu.tw (B.-C. Chien), wpyang@cis.nctu.edu.tw (W.-P. Yang).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.01.006

Table 1
A comparison of evaluation functions (cited from Dash and Liu (1997))

Evaluation function	Generality	Time complexity	Accuracy
Distance measure	Yes	Low	–
Information measure	Yes	Low	–
Dependence measure	Yes	Low	–
Consistency measure	Yes	Moderate	–
Classifier error rate	No	High	Very high

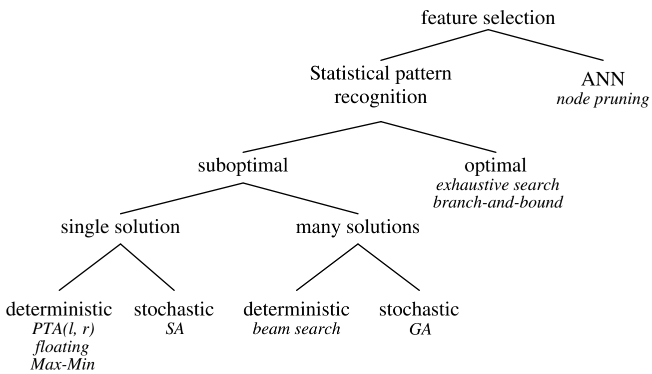

Fig. 1. Taxonomy of feature selection algorithms (cited from Jain and Zongker (1997)).

recommend some useful algorithms, which can provide the number of raw features. Kittler et al. (1978) and Pudil et al. (1994) proposed many feature selection techniques, such as (generalized) sequential forward selection, (generalized) sequential backward selection, (generalized) plus l-take away r selection, sequential forward floating selection, and sequential backward floating selection. Pernkopf compared the accuracy of Bayesian network classifier and k-NN classifiers with different feature selection methods in Pernkopf (2005).

Feature generation (Jolliffe, 1986; Lee & Landgrebe, 1997; Ma, Theiler, & Perkins, 2004; Mao & Jain, 1995; Park & Park, 2004; Wang & Paliwal, 2003) deals with features in different way. It generates representative features or classification-oriented ones (Ma et al., 2004). The former is mainly applied when the given data are not represented by a distinguishable form for the classifier on hand, e.g. handwriting and images. Mostly common feature extraction methods are principle component analysis (PCA) (Jolliffe, 1986) and linear discriminant analysis (LDA). Ma et al. (2004) proposed a general feature extraction framework and two nonlinear feature extraction algorithms, kernel function and mean-STD-norm, cooperating with a SVM classifier. Mao and Jain (1995) proposed several artificial neural network models such as PCA-networks, LDA-networks, and nonlinear discriminant analysis (NDA) network in Mao and Jain (1995). Wang and Paliwal (2003) investigates performance of PCA, LDA, minimum classification error (MCE), and generalized MCE with SVM classifier on vowel databases.

Traditionally, GP maintains a single population to find optimal solutions (Banzhaf, Nordin, Keller, & Framcone, 1998; Koza, 1992). Multipopulation GP (MGP) (Brameier & Banzhaf, 2001; Fernández, Tomassini, & Vanneschi, 2003) developed in recent years uses multiple populations to extend individual diversity and create different evolving environments. Fernández et al. (2003) performed several experiments about parallel and distributed GP (PADGP), isolated MGP (IMGP) ("isolated" means that there is no migration between populations), and traditional single population GP. Their experiments show that PADGP and IMGP usually perform better than traditional SGP. Moreover, MGP with small populations performs better than traditional GP using single large population. Brameier and Banzhaf (2001) combine linear GP and MGP techniques to design a classifier. Individuals represented as strings migrate between demes, i.e. subpopulations, according to their fitness. Fig. 2 (Brameier & Banzhaf, 2001) shows the circle topology where each circle stands for a population and arrows are migration directions.

Many classifiers based on GP have been proposed recently (Bojarczuk, Lopes, & Freitas, 1999; Chien, Lin, & Yang, 2003; Falco, Cioppa, & Tarantino, 2002; Freitas, 1997; Kishore, Patnaik, Mani, & Agrawal, 2000; Konstam et al., 1998; Lin, Chien, & Hong, 2002; Loveard & Ciesielski, 2001). Kishore et al. (2000) considered a k-class classification problem as a combination of k two-class classification problems and then generated corresponding expressions or discriminant functions for each class, and so did we (Chien et al., 2003; Lin et al., 2002). These methods need k runs for a k-class classification problem. To generate classification rules, Freitas (1997) proposed Tuple-Set-Descriptor (TSD), a logical formula to represent an individual. Muni, Pal, and Das (2004) proposed a method to solve multi-class problem by representing each individual as a multitree. Each tree stands for a candidate solution corresponding to each class. To evolve an individual is equivalent to evolve k trees simultaneously.

Researches on feature selection and feature extraction using evolutionary computation boom rapidly. The use of genetic algorithm (GA) methods can be found in Oh,

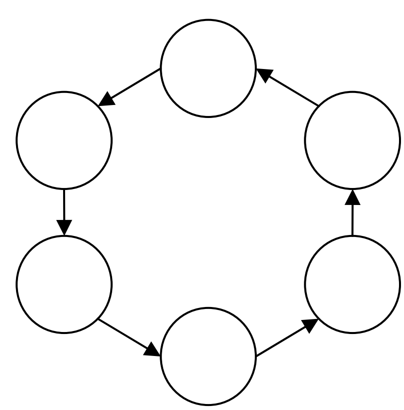

Fig. 2. The circle topology MGP.