



Ensemble classifier generation using non-uniform layered clustering and Genetic Algorithm

Ashfaqur Rahman^{a,*}, Brijesh Verma^b

^a Intelligent Sensing and Systems Laboratory, CSIRO, Hobart, TAS, Australia

^b Center for Intelligent and Networked Systems, Central Queensland University, QLD, Australia

ARTICLE INFO

Article history:

Received 4 July 2012

Received in revised form 26 November 2012

Accepted 2 January 2013

Available online 28 January 2013

Keywords:

Ensemble classifier

Genetic Algorithm

Multiple Classifier Systems

Cluster Based Ensemble Classifiers

Diversity in Ensemble Classifiers

ABSTRACT

In this paper, we propose a novel cluster oriented ensemble classifier generation method and a Genetic Algorithm based approach to optimize the parameters. In the proposed method the data set is partitioned into a variable number of clusters at different layers. Base classifiers are trained on the clusters at different layers. Due to the variability of the number of clusters at different layers, the cluster compositions in one layer are different from that in another layer. Due to this difference in cluster contents, the base classifiers trained at different layers are diverse among each other. A test pattern is classified by the base classifier of the nearest cluster at each layer and the decisions from different layers are fused using majority voting. The accuracy of the proposed method depends on the number of layers and the number of clusters at the corresponding layer. A Genetic Algorithm based search is incorporated to obtain the optimal number of layers and clusters. The Genetic Algorithm is evaluated under three different objective functions: optimizing (i) accuracy, (ii) diversity, and (iii) accuracy \times diversity. We have conducted a number of experiments to evaluate the effectiveness of the different objective functions.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

An ensemble classifier [1,2,9] refers to a collection of base classifiers that are trained simultaneously on the data. Their decisions on a pattern are combined to obtain a classification verdict. Ensemble classifiers are also known as multiple classifier systems and committee of classifiers. The training process in an ensemble classifier aims to produce the base classifiers in such a way that they are accurate and also differ from each other in terms of the errors they make on identical patterns. This phenomenon is known as diversity [3–5]. The fusion methods on the other hand explore ways to merge the decisions from the base classifiers into a final verdict. A commonly used approach to generate the base classifiers is by training them on different subsets of the data. This ensures diversified learning of the base classifiers and achieves higher accuracy. The subset selection algorithm varies among the different ensemble generation methods.

Patterns in data set tend to scatter over the Euclidian space and form subgroups. The clustering process identifies these natural subgroups [6]. Divide-and-conquer approach towards ensemble classifier generation [7] produces training subsets for the base classifiers by clustering the data set. Space decomposition process

identifies multiple clusters within the classified data (Fig. 1). Classified data refers to a labelled data set where each pattern is associated with a class label. A cluster produced in this way can be homogeneous or heterogeneous in nature. A homogeneous cluster contains patterns belonging to a single class and only the class label needs to be memorized as the decision is always unique. A heterogeneous cluster contains overlapping patterns from multiple classes that are close in Euclidian space. Each cluster is learned by a base classifier in the divide-and-conquer approach. The learning of the base classifiers is thus focussed and specialized. A pattern is classified by finding the nearest cluster and the corresponding base classifier to provide a verdict.

Space decomposition or divide-and-conquer (i.e. clustering classified data) is in practice for quite a while. As reported in [8] the performance of this approach in some cases is even worse than unique classifiers. This is due to the fact that a pattern can belong to one cluster only and as a result the decision can be obtained from a single classifier. The concept of diversity thus does not apply on to this approach and only one classifier is in fact trained on a pattern leading to poor classification performance.

We aim to address this issue of lack of diversity using overlapping clustering. In this regard we make use of the fact that a data set can be partitioned into different number of clusters. In k -means clustering algorithm k indicates the number of clusters and in hierarchical clustering algorithm [42] the *cutoff* threshold can be used to define the final number of clusters. In order to achieve diversity

* Corresponding author. Tel.: +61 362325536.

E-mail addresses: ashfaqur.rahman@csiro.au (A. Rahman), b.verma@cqu.edu.au (B. Verma).

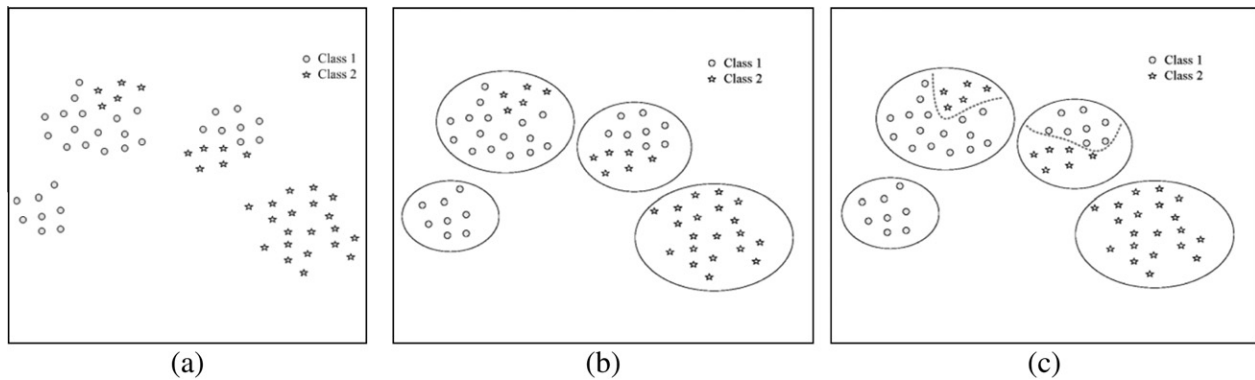


Fig. 1. Impact of clustering on classified data [10]: (a) the data set with two classes, (b) clustered data set with multiple homogeneous and heterogeneous clusters, and (c) class boundaries of heterogeneous clusters.

the data set can be independently partitioned n times into variable number of clusters and identical patterns will belong to n alternate clusters. We use the terminology n layers to refer to n alternative clustering of the data set in this paper. The decision provided by the base classifiers trained on the n non-uniform clusters at n layers can be fused to obtain the final verdict on the pattern. The clusters at different layers are non-uniform as the number of clusters vary at different layers. With clustering we can generate the base classifiers and with layers we can achieve the diversity. Note that the number of layers at which maximum diversity is achieved depends on the characteristics of the data set and needs to be optimized. We have adopted a Genetic Algorithm based search approach to identify the optimal number of layers and clusters. The optimal results in GA depend on the objective function. We have used three different objective functions: optimizing (i) accuracy, (ii) diversity, and (iii) accuracy \times diversity.

The research presented in this paper is based on the above philosophy and aims to: (i) develop a novel method for generating ensemble of classifiers using non-uniform cluster layers and optimizing the number of layers, (ii) investigate the impact of number of clusters and number of layers on classification accuracy, (iii) obtain a comparative analysis of the different objective functions in GA, and (iv) obtain a comparative analysis on how well the proposed approach performs compared to the commonly used approaches for ensemble classifier generation.

The paper is organized as follows. Section 2 reviews existing approaches for ensemble classifier generation and decision fusion, some commonly used base classifiers and Genetic Algorithm. Section 3 presents the proposed approach to generate ensemble classifiers. The experimental platform is presented in Section 4. Section 5 presents the experimental results and discussion. Finally, Section 6 concludes the paper.

2. Related works

2.1. Ensemble classifiers

Research towards ensemble classifiers has two major streams: (a) cooperative training methods of the base classifiers and (b) fusion methods for combining the decisions of the base classifiers. The fusion methods combine the discrete class decisions or continuous class confidence values produced by the base classifiers into a class verdict. The proposed ensemble classifier use majority voting fusion method to combine the discrete valued class decisions produced by the base classifiers. We thus confine the following discussions on generation methods only.

Ensemble classifier generation methods using homogeneous base classifiers can be broadly classified into five groups that are

based on (i) manipulation of the training parameters, (ii) manipulation of the error function, (iii) manipulation of the feature space, (iv) manipulation of the output labels, and (v) manipulation of the training patterns. All these methods aim to achieve diversity among the base classifiers.

Diversity can be achieved by *manipulating the training parameters* of the base classifiers in an ensemble. Different network weights are used to train the base neural network learning process in [11,12]. These methods achieve better generalization. A group of ensemble classifier construction methods address this issue by *augmenting the error function* of the base classifiers. An error is imposed if base classifiers make identical errors on similar patterns. Negative correlation learning [13,43] is one such ensemble where all the individual networks in the ensemble are trained simultaneously and interactively through the correlation penalty terms in their error functions.

In another group of ensemble classifiers diversity among the base classifiers is achieved by *manipulating the input feature space*. Different feature subsets are used to train the base classifiers [14–16]. The random subspace ensemble classifiers perform relatively inferior to other ensemble classifiers. Ensemble classifiers can be constructed by *manipulation of the output targets* [17,18]. In class switching ensemble [17], each base classifier is generated by switching the class labels of a fraction of training patterns that are selected at random from the original training set.

The largest set of ensembles generates ensemble classifiers by *manipulating the training patterns* where the base classifiers are trained on different subsets of the training patterns. Our proposed ensemble classifier also belongs to this group. The methods differ in generation of the subsets. Training subsets are generated by partitioning the training set into non-overlapping clusters in *clustered ensembles* [7,19–22,40,41]. The patterns that tend to stay close in Euclidean space naturally are identified by this process. A pattern can belong to one cluster only thus a selection approach is followed for obtaining the ensemble class decision. These methods aim to reduce the learning complexity of large data sets [7]. The clustered ensembles [19–22] do not provide any mechanism for obtaining the optimal number of clusters. Some researchers [8,23,41] provide a mechanism to obtain soft partitioning of the data set that leads to better classification performance [8]. The optimality issues are however not investigated in these works. Cluster based ensembles are normally used to manage the learning complexity of large data sets.

In *bagging* [24] the training subsets are randomly drawn (with replacement) from the training set. Homogeneous base classifiers are trained on the subsets. The class chosen by most base classifiers is the considered to be the final verdict of the ensemble classifier. There are a number of variants of bagging and aggregation

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات