

Group-based Meta-classification

Noor A. Samsudin, Andrew P. Bradley
School of Information Technology and Electrical Engineering
The University of Queensland, Australia
{azah, bradley}@itee.uq.edu.au

Abstract

Virtually all existing classification techniques label one sample at a time. In this paper, we highlight the potential benefits of group based classification (GBC), where the classifier labels a group of homogeneous samples. In this way, GBC can take advantage of the additional prior knowledge that all samples belong to the same, unknown, class. We pose GBC in a generic hypothesis testing framework requiring the selection of an appropriate sample and test statistic. We then evaluate one simple example of GBC on both synthetic and real data sets and demonstrate that GBC may be a promising approach in applications where the test data can be arranged into homogenous subsets.

1. Introduction

Classification is a fundamental task in statistical pattern recognition. The existing literature is typically divided into approaches that perform supervised versus unsupervised classification [1-3]. In supervised classification, class labels are known for each sample, while in unsupervised classification class labels are unknown. More recently, semi-supervised techniques have also been proposed for situations where both labeled and unlabeled data is available [4]. In all of these approaches there is first a learning (training) phase, where training samples are presented and attributes of, or a decision boundary between, various subsets of the training data are determined. Next there is a classification (evaluation) phase, where test samples are presented and the performance of the classifier measured.

Conventionally, little distinction is made between whether training samples are presented one at a time (incremental learning) or as a group (batch learning). However, in this paper we will refer to incremental learners, such as the Perceptron [2] and Multiscale Classifier (MSC) [5], as Individual-based Learning (IBL) and batch learners, such as Support Vector

Machines (SVMs) [3] and most Multi-layer Perceptrons (MLPs) [1], as Group-based Learning (GBL). Traditionally, supervised learning has employed either IBL or GBL techniques, whilst unsupervised learning has almost exclusively employed GBL techniques. Semi-supervised learners, including self-training and co-training [4], are also examples of GBL.

However, when it comes to the classification phase virtually all, well known, supervised, un-supervised and semi-supervised classification techniques classify the test data one sample at a time. Therefore, they are all instances of individual-based classification (IBC) and, as is highlighted in Table 1, there are very few techniques that employ group-based classification (GBC). This seems a little surprising, as knowing that a set of test samples comes from the same, but *unknown*, class would appear to be valuable prior knowledge that currently is being under utilized.

Table 1: Pattern classification approaches.

	Learning (training)	Classification (evaluation)
Individual-based	IBL, e.g., incremental learning in Perceptron and MSC.	IBC, e.g., supervised, unsupervised, semi-supervised
Group-based	GBL, e.g., SVMs, batch training, self-training,	Group-based Classification (GBC)

So, the question is; in which applications is GBC appropriate? Clearly, this would appear to be in applications where the test data is either in, or can be arranged into, homogenous subsets. For example, in Pap smear screening for cervical cancer [6] the data is presented as a sample of cells on a microscope slide. Ultimately, the goal is to classify the slide (i.e., the patient) as either normal or abnormal (pre-cancerous). Unfortunately, each slide contains many thousands of cells and, in principle; each cell must be analyzed before the slide can be classified. However, only in

rare circumstances does it make sense to a human expert to label *individual* cells as either normal or abnormal. Therefore, Pap smear screening can be thought of as *meta-classification* problem, where it is only a collective group of cells on a slide that can sensibly be classified. Specifically, Nordin and Bengtsson [6] have identified three approaches to automated Pap smear screening, which are briefly described in Table 2. They are termed Rare Event (RE), fixed proportion, and Malignancy Associated Changes (MACs). The rare event approach can be viewed as IBC. Fixed proportion attempts to achieve GBC via a two-step process, but is prone to the same problems as RE and so does not really satisfy the principle of GBC. Instead, *MACs* is the best example of a practical approach to GBC.

Table 2: Approaches to Pap smear screening.

Approaches	Descriptions
Rare Event (RE)	Exhaustively look for any abnormal cells throughout the slide. Problems: <ul style="list-style-type: none"> • Unavailable diagnostic cells due to sampling error; • Misclassified cells due to human error; • Time-consuming.
Fixed Proportion	Is a two-stage classification. First, RE is accomplished for all cells. Second, based on the proportion of RE cells the slide is classified. Problems: <ul style="list-style-type: none"> • Still prone to RE problems. • Slides with few abnormal cells typically misclassified.
MACs	Classification is based on statistical properties (e.g., μ and σ^2) of a subset of all cells on the slide. Problems: <ul style="list-style-type: none"> • Assumes all cells undergo cancerous changes (field-affect).

Pap smear screening is one example, typical to many biomedical problems, where GBC may be appropriate as each patient must be classified based on measurements from a group of objects. However, it may also be possible pose many other classification problems so that one can classify a group of samples as one homogenous collective. For example, in the commonly used Iris data set, where the goal is to classify sepal and petal measurements as belonging to one of three species of Iris, it seems feasible to present a group of measurements for classification based on the assumption that they were all taken from the same, *unknown* plant. Such example may be viewed as similar to isogenous pattern classification problem [7] as patterns are obtained from same source. Therefore, in this paper, we present a generic approach to both

group based learning and classification. We then go on to evaluate a number of specific instances of the group based paradigm on both synthetic and real data.

2. The group-based classifier

The essence of GBC can be found in statistical hypothesis testing [8]. It lies in the selection of a sample statistic (the result of applying a function to a set of data, e.g., mean or variance) and a test statistic (a measure of similarity between sample statistics, e.g., a two-sample z , t or F -test) Clearly, the sample statistic summarizes a specific property of the data and the class membership decision is made by the test statistic. In this way, we have flexibility in the distributional assumptions we may make regarding our data. In order to demonstrate our approach to GBC we propose the following generic steps:

i denotes the class label, i.e., $i = \{1, \dots, j\}$,

\mathbf{X}_i is the training set of class i , and \mathbf{x} is the test set, which is a group of samples whose class label is unknown.

\mathbf{T}_i is the union of the \mathbf{x} and \mathbf{X}_i , i.e. $\mathbf{x} \cup \mathbf{X}_i$.

Let $s(\mathbf{X}_i)$ denote a sample statistic function of \mathbf{X}_i , and $s(\mathbf{T}_i)$ denote a sample statistic function of \mathbf{T}_i .

Let $m_i = f(s(\mathbf{X}_i), s(\mathbf{T}_i))$ denote the result of the test statistic $f()$, i.e., the similarity between $s(\mathbf{X}_i)$ and $s(\mathbf{T}_i)$.

Let \mathbf{m} denote the set of similarity results on each class $s(\mathbf{X}_i)$ and $s(\mathbf{T}_i)$, such that, $\mathbf{m} = \{m_1, m_2, \dots, m_j\}$

Step 1 For each class, determine $s(\mathbf{X}_i)$.

Step 2 For each class, determine $s(\mathbf{T}_i)$.

Step 3 For each class, determine $m_i = f(s(\mathbf{X}_i), s(\mathbf{T}_i))$.

Step 4 \mathbf{x} is classified into class i , if $\Pr(\mathbf{x} \in i | m_i)$ is maximum.

Note that we have proposed to consider $\mathbf{x} \cup \mathbf{X}_i$ to determine $s(\mathbf{T}_i)$ instead of \mathbf{x} alone. The rationale for this decision is that the number of samples in \mathbf{x} would typically be considerably smaller than \mathbf{X}_i (so as to maximize the amount of training data). This difference in sample size between \mathbf{X}_i and \mathbf{x} , could lead to a large difference in the standard error of $s(\mathbf{X}_i)$ compared to $s(\mathbf{x})$ and so should be avoided.

3. Experiment methodology

The purpose of our experiments was to initially investigate the efficacy of GBC on both synthetic and real data sets. As all of our experiments involve approximately equally class priors, error rate was thought an appropriate measure of classification performance. This was estimated using 10-fold cross-validation [1, 2]. For our experiments, all samples in the data sets were used and each cross validation partition (fold) was randomly selected so as to preserve prior class probability. For every class i one

partition was used as the test data, \mathbf{x} , and the remaining partitions as the training data, \mathbf{X}_i . However, in order to plot the classification error rate as a function of group size, we evaluated all possible sub-samples of \mathbf{x} of size one to six. For every combination size, the error rate was estimated by dividing the number of misclassified samples by the total samples. The standard error of the error rate for GBC was also calculated and is shown by the error bars in the Figures 1 to 4. In this paper, we investigate just one instance of GBC in which mean squared Euclidean distance from the class centroid, σ^2 [10] is used as the sample statistic and the F -test ratio as the test statistic. Specifically, our initial GBC was implemented as follows:

Step 1: For every \mathbf{X}_i , determine $s(\mathbf{X}_i) = \sigma^2$ of \mathbf{X}_i

Step 2: For every \mathbf{T}_i , determine $s(\mathbf{T}_i) = \sigma^2$ of \mathbf{T}_i .

Step 3: For every $s(\mathbf{X}_i)$ and $s(\mathbf{T}_i)$, perform F -test, such that $m_i = s(\mathbf{T}_i)/s(\mathbf{X}_i)$.

Step 4 \mathbf{x} is classified into class i , if $\Pr(\mathbf{x} \in i | m_i)$ is maximum.

The mean squared Euclidean distance, σ^2 , is one way to represent dissimilarity of patterns in a cluster [10]. It was chosen here as a relatively simple one-dimensional estimate of intra-class variance. Assuming that the data is drawn from a Normal population, the estimated variance will have a chi-squared sampling distribution and the ratio of two such variances will follow an F -distribution. Thus, the F -test is an appropriate test statistic to compare the variability of the samples in \mathbf{X}_i and \mathbf{T}_i .

In all experiments, the performance of this simple GBC was compared against two commonly used techniques for IBC, namely naive Bayes and nearest neighbor (NN). The Bayes classifier was chosen because the synthetic data sets were indeed Normally distributed and so it should perform well. The NN classifier was chosen because it makes no distributional assumptions about the data and it is known that its performance will never be worse than twice the Bayes error [2]. Note, that for both of the IBC approaches only one test sample is presented to the classifier at a time.

3.1. Synthetic data

The advantage of using synthetic data is that we can control the distributions of the generated samples. Therefore, we generated three, two-dimensional, data sets all with Normally distributed features and with 1000 samples per class. Table 3 summarizes the synthetic data sets used: in Case 1 both classes have the same covariance; in Case 2 the classes have differing covariance; and Case 3 is similar to Case 1,

but with correlated features. As we are using 10-fold cross-validation each test partition, \mathbf{x} , consists of 100 samples per class. For every test partition all possible subsets of size one to six were evaluated. In this way, the proposed GBC technique determined the class labels for variously sized sub-groups of the test data.

Table 3: Summary of the synthetic data sets.

Case	Mean	Covariance	
		Class 1	Class 2
1	Mean for class 1: [5.5 5.5]	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$
2	Mean for class 2: [2.5 2.5]	$\begin{bmatrix} 1.0 & 0 \\ 0 & 1.0 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$
3		$\begin{bmatrix} 1.5 & 0.75 \\ 0.75 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$

3.2. Iris data

The Iris data set contains total of 150 samples with 50 samples in each of the three classes. Each sample consists of four continuously valued features of: sepal length and width plus petal length and width. With 10-fold cross validation each test partition, \mathbf{x} , consists of five (homogenous) samples of unknown class. Therefore, we classified every combination of the test set of size one to five samples.

4. Experimental results

4.1. Synthetic data

Figures 1 to 3 show the plots of error rate as a function of group size for each case of synthetic data. In all cases, the error rate for this simple GBC technique approaches zero as the group size approaches six. The poorest performance is observed in case 3, which has correlated features, and so mean squared Euclidean distance is a poor estimate of the variance in any one direction. However, it should be noted that the performance in this case could be improved by the application of a pre-whitening transform to remove the correlation and normalize the covariance of the data [2].

For all three synthetic cases, the proposed classifier outperforms both the Bayes and NN classifiers when the combination size is two or more. Indeed, it is interesting to note that an error rate of zero can be achieved with this data which has, by definition, overlapping class probability distributions and so has a non-zero Bayes error. This indicates the potential benefits of utilizing the additional prior knowledge implicit to GBC, i.e., that a group of test samples has the same, but unknown, class membership.

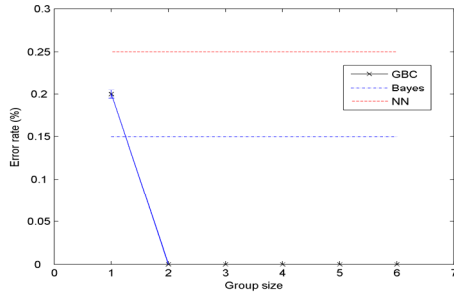


Figure 1: Error rate versus group size for Case 1.

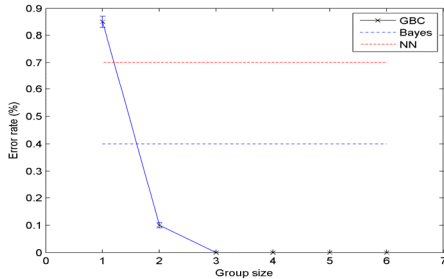


Figure 2: Error rate versus group size for Case 2.

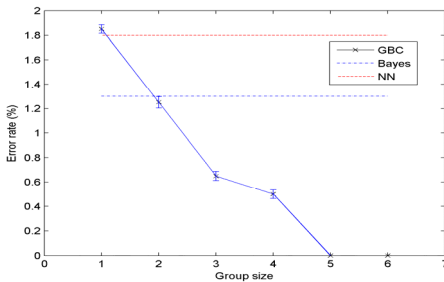


Figure 3: Error rate versus group size for Case 3.

4.2. Iris data

The Iris data set is used as one potential practical application of GBC by arranging the test data into homogenous sub-groups. The results in Figure 4 suggest that GBC has equivalent performance to the Bayes classifier when the group size is four and eventually outperforms both IBC approaches when the group size is five. The use of GBC on the Iris data improves classification performance once the group size is large enough account for any outliers in the test set. Clearly, GBC is benefitting from the prior knowledge that all test samples in the sub-group are homogeneous and should be given the same class label.

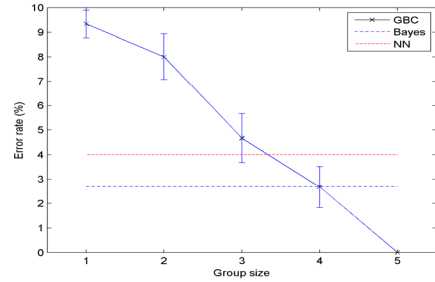


Figure 4: Error rate versus group size for Iris data.

5. Conclusions

In this paper, we have presented the underlying concepts and rationale behind group-based classification (GBC). We have evaluated a simple example of GBC, comparing it to two common examples of individual based classification (IBC). Our empirical results show the benefit of GBC in applications where the test data can be arranged into homogenous subsets.

References

- [1] E. Alpaydin, *Introduction to machine learning*. London: The MIT Press, 2004.
- [2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2 ed. New York: John Wiley & Sons, Inc., 2001.
- [3] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition," *IEEE Transactions on PAMI*, vol. 22, pp. 4-37, 2000.
- [4] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-supervised learning*. London: The MIT Press, 2006.
- [5] B. C. Lovell and A. P. Bradley, "The Multiscale Classifier," *IEEE Transactions on PAMI*, vol. 18, pp. 124-137, 1996.
- [6] B. Nordin and E. Bengtsson, "Specimen analysis by rare event, cell population, and/or contextual evaluation," in *Automated Cervical Cancer Screening*, H. K. Grohs and O. A. N. Husain, Eds. New York: IGAKU-SHOIN Medical Publishers, 1994, pp. 44-51.
- [7] P. Sarkar and G. Nagy, "Style Consistent Classification of Isogenous Patterns," *IEEE Transactions on PAMI*, vol. 27, pp. 88-98, 2005.
- [8] R. J. Wonnacott and T. H. Wonnacott, *Statistics: Discovering its power*. New York: John Wiley & Sons, 1982.
- [10] B. S. Everitt, *Cluster Analysis*, 4 ed. London: Arnold, 2003.