Pattern Recognition 43 (2010) 3458–3467

Contents lists available at ScienceDirect

Pattern Recognition

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

Nearest neighbour group-based classification

Noor A. Samsudin *, Andrew P. Bradley

School of Information Technology and Electrical Engineering, The University of Queensland, St. Lucia, QLD 4072, Australia

ARTICLE INFO

Article history:
Received 21 October 2009
Received in revised form
10 February 2010
Accepted 4 May 2010

Keywords:
Group-based classification
Nearest neighbour
Compound classification

ABSTRACT

The purpose of group-based classification (GBC) is to determine the class label for a set of test samples, utilising the prior knowledge that the samples belong to same, but unknown class. This can be seen as a simplification of the well studied, but computationally complex, non-sequential compound classification problem. In this paper, we extend three variants of the nearest neighbour algorithm to develop a number of non-parametric group-based classification techniques. The performances of the proposed techniques are then evaluated on both synthetic and real-world data sets and their performance compared with techniques that label test samples individually. The results show that, while no one algorithm clearly outperforms all others on all data sets, the proposed group-based classification techniques have the potential to outperform the individual-based techniques, especially as the (group) size of the test set increases. In addition, it is shown that algorithms that pool information from the whole test set perform better than two-stage approaches that undertake a vote based on the class labels of individual test samples.

1. Introduction

Traditional approaches to classification aim to determine the class label of individual test samples [1] and so have been called individual-based classification (IBC) [2]. Alternatively, group-based classification (GBC) is the task of assigning a single class label to a *group* of test samples [2]. Group-based classification is an example of the use of context to aid decision making [3] which has been found beneficial in such diverse fields as speech recognition, optical character recognition, document classification and remote sensing [1]. For example, the compound classification of collections of cells in flow cytometry, based around the classification and regression tree (CART) algorithm, is presented in [9].

Group-based classification is a simplification of the more commonly considered (non-sequential) compound decision problem ([4], Section 2.12), where we make N decisions jointly for an L-class problem, $\mathbf{c}=\{c_1, \ldots, c_L\}^t$ and an N sample data set, $\mathbf{X}=\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$, i.e.,

$$P(\mathbf{c}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{c})P(\mathbf{c})}{p(\mathbf{X})}$$

However, in group-based classification we assume that the application domain allows us to utilise *a priori* knowledge that the whole group of unlabeled samples belong to the same, but unknown, class. In this way, we only need consider the L possible class labels for the class vector, $\mathbf{c}$, rather than the computationally prohibitive L^N. Clearly, if the group assumption can be made, then we can potentially gain the benefits of making compound decisions without the extreme computational burden of calculating $P(\mathbf{c}|\mathbf{X})$ [4]. Likewise, another solution to the compound decision problem is to make decisions *sequentially* based on previous decisions and/or states. This approach leads to the well-known first-order and hidden Markov models [3,4]. However, the focus of this paper is on non-sequential compound decisions and group-based classification in particular.

Group-based classification is also related to the *bag-of-words* model commonly applied as a simplifying assumption in document classification. Here the occurrence of individual words provides evidence of the document class, but the entire document is assigned a single class label [5]. However, in this paper we concentrate on quantitative feature vectors common to pattern recognition, where there is a natural measure of distance between vectors, as opposed to categorical variables that typically require methods such as contingency tables to be built from word counts in documents.

1.1. Group-based classification

The Bayes formula for a group-based classifier is

$$P(c_l|\mathbf{X}) = \frac{p(\mathbf{X}|c_l)P(c_l)}{p(\mathbf{X})}$$

This clearly shows that the posterior probability, $P(c_l|\mathbf{X})$, is dependent not just on a single sample (feature vector) but a group, or set, of test samples. This is further illustrated in Fig. 1 which

* Corresponding author. Tel.: +61 7 3365 4510; fax: +61 7 3365 4999.
E-mail addresses: azah@itee.uq.edu.au, azasam73@yahoo.com (N.A. Samsudin), bradley@itee.uq.edu.au (A.P. Bradley).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.05.010

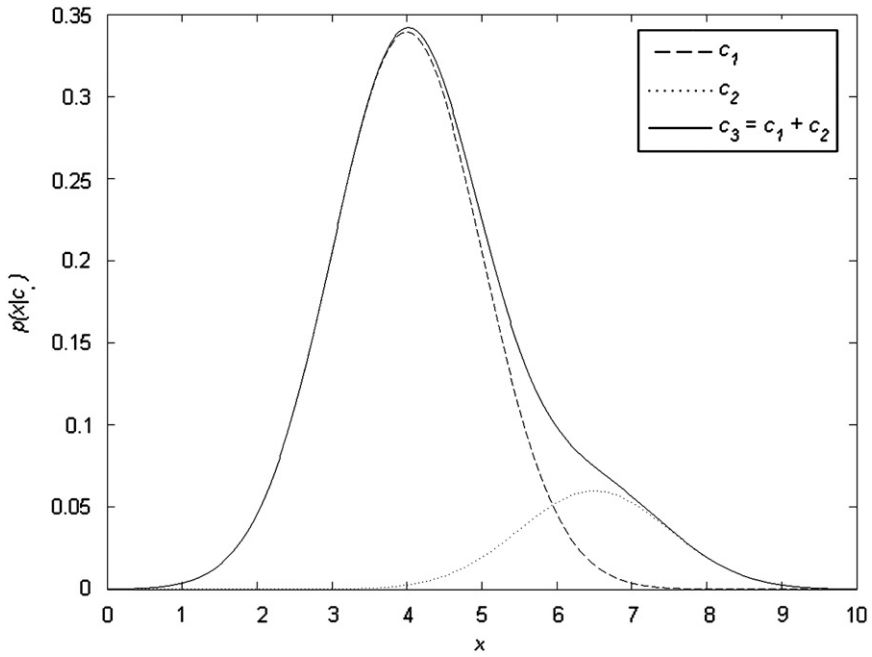

Fig. 1. Three hypothetical class-conditional probability density functions.

shows three hypothetical class-conditional probability density functions. Consider the case where we are interested in discriminating class c_1 from c_2. For an individual-based classifier the optimal decision boundary, assuming equal misclassification costs, is to decide c_1 if $x<6$ or c_2 if $x>6$. However, for a group-based classifier no such simple decision boundary on an individual value of x exists. Rather, a group-based classifier more generally decides c_1 if $P(c_1|\mathbf{X})>P(c_2|\mathbf{X})$ or c_2 otherwise. That is, it selects the class that $\mathbf{X}$, as a whole, was most likely to have come from (taking into account the prior probability of each class). Clearly, one of the key challenges to GBC is to reliably estimate $P(c_l|\mathbf{X})$ for each class from the multi-dimensional feature vectors that make up $\mathbf{X}$. Alternatively, as in [2], one could utilise an appropriate sample statistic, $s(\mathbf{X})$, as a surrogate measure. For example, for the simple case of distinguishing c_1 from c_2 the sample mean, μ_x, would suffice and leads to a reduced Bayes error, provided that the group size is large enough to produce a lower standard error around the class means, than the standard deviation of the original data.

Next, consider the more complex case of discriminating class c_1 from c_3 which is formed by mixing, or contaminating, class c_1 with c_2. This situation occurs frequently in the microscopic analysis of cells, where c_1 could represent the distribution of normal cells and c_2 abnormal cells. Hence an abnormal slide would typically be distinguished from normal by the presence of a small number of abnormal cells in addition to the majority of visually normal cells [6]. As discussed in [2,6] there is a simple two-stage solution to this problem where a conventional individual-based classifier is first used to detect abnormal cells, say where $x>6$ in Fig. 1, and then the slide is classified abnormal if the number of abnormal cells exceeds a fixed proportion of the total cells analysed. However, there are at least two problems with this approach:

1. It is not as simple as it might first appear, as two-stage recognition systems must be optimised holistically to maximise their performance [7] and
2. The second stage of this approach only utilises the relatively small number of cells actually detected as abnormal. That is, it is only estimating $P(c_l|\mathbf{X}, x>6)$. Clearly, as Fig. 1 shows, there are potential differences between c_1 from c_3 across the whole range of feature values, in this case from $x>4$, and so utilising more samples from the whole of $P(c_l|\mathbf{X})$ is intuitively more robust.

Therefore, one can conclude that methods, such as compound and group-based classification, that operate in a single stage utilising all of the relevant samples and context are worthy of further study.

Empirical results presented in [2] have shown that knowing a group of unlabeled samples belong to the same, but unknown, class is useful additional *a priori* knowledge that can be utilised by GBC to obtain error rates below the 'optimal' Bayes error. However, for this prior knowledge to be utilised the application domain must allow the test data to be grouped into homogenous subsets. Two such applications, discussed in [2] and developed further in this paper, are Pap smear screening: where each slide contains a sample of cells obtained from a single patient; and plant species classification, such as the Iris data set [8]: where samples can be grouped so that they come from a single, unknown, plant/species.

In [2], group-based classification was implemented using a hypothesis testing framework, in which the class label decision was based on a sample statistic, specifically the mean squared Euclidean distance and then a test statistic of similarity, specifically the F-test ratio. In this paper, we extend the basic ideas behind group-based classification to one of the most well-known non-parametric individual-based classification techniques, namely k-nearest neighbour (k-NN) and then explore under what conditions the pooling of information required for group-based classification outperforms simple voting procedures based on individual-based classification.

In the following we justify why k-NN is an appropriate choice for extension to group-based classification. In any pattern classification problem where the underlying probability density functions are unknown or difficult to estimate, the use of non-parametric techniques to classify the data is desirable [4]. One

attractive non-parametric technique, which is a well established method in the pattern recognition literature, is *k*-NN [10]. The principal idea of *k*-NN is conceptually simple, in that a test sample is assigned a class label according to the most frequently occurring class label among its *k*-nearest neighbours. A proximity measure, such as Euclidean distance, is used to define the *k*-nearest neighbours to each test sample. Despite its strength as a simple non-parametric individual-based classifier [4], there has been considerable effort improve and extend *k*-NN. For example, by: assigning different weights to every *k*-NN in inverse proportion to their distance from the test sample [11–14]; reducing the effect of outliers [15] by classifying test samples based on the class of their nearest local mean vector (LMV) [16]; selecting a subset of the most discriminatory features [17–19] or weighting features according to their discriminatory power [20]; selecting an optimal value of *k*, with the lowest estimated error rate [21–24]; and reducing computational complexity and memory usage by using Kd-trees [25,26] and hashing functions [27,28].

From the above, it is clear that *k*-NN is both a well-known and widely applied non-parametric classifier. In addition, there is a body of work addressing the commonly cited limitations of the approach. Therefore, extension of *k*-NN to group-based classification should be of wide interest in a number of application areas and also complements our previous formulation in a hypothesis testing framework [2]. In this paper, we will investigate a variety of approaches, based on existing extensions of *k*-NN, that are intuitively suitable for implementing non-parametric group-based classification.

2. Group-based nearest neighbours

Table 1 illustrates a number of existing non-parametric individual-based classification techniques, based on the *k*-NN principle. First, is the simple *k*-NN voting rule [10,29], which is based on the Euclidean distance between $\mathbf{x}$, an n-dimensional test sample $\mathbf{x}=(x^1,\ldots,x^n)^T\in\mathbf{X}_{\mathrm{TE}}$, and the k-nearest neighbours, $\mathbf{X}_{\mathrm{KNN}}$, in the training set, $\mathbf{X}_{\mathrm{TR}}\subset\mathbf{X}$, where $\mathbf{X}=\mathbf{X}_{\mathrm{TR}}\cup\mathbf{X}_{\mathrm{TE}}$ and normally $\mathbf{X}_{\mathrm{TR}}\cap\mathbf{X}_{\mathrm{TE}}=\emptyset$. Each training sample $\mathbf{x}^j\in\mathbf{X}_{\mathrm{KNN}}$ contributes a single vote towards v_l, the votes for each class, based on their class labels

$$v_l=\sum_{j=1}^{k}\mathbf{I}_l(\mathbf{x}^j)$$

where $\mathbf{I}_l(\cdot)$ is the indicator function

$$\mathbf{I}_l(\mathbf{x})=\begin{cases}1 & \text{if } (\mathbf{x}^j\in\mathbf{X}_{KNN},c_l=l)\\ 0 & \text{otherwise}\end{cases}$$

In this way, the test sample, $\mathbf{x}$, is given a class label according to a majority vote amongst the *k*-nearest neighbours, i.e., $c_l=\arg\max(v_l)$.

An extension to this simple voting scheme is the distance weighted (DW) *k*-NN [11]. This is similar to the conventional *k*-NN, in that the nearest neighbours are determined in the same manner. However, here the votes of each of the neighbours of $\mathbf{x}$ are weighted according to their distance from $\mathbf{x}$. In this way, the weight of the rth nearest neighbour in $\mathbf{X}_{\mathrm{KNN}}$, w^r is determined as follows:

$$w_r=\begin{cases}\dfrac{d^k-d^r}{d^k-d^1} & \text{if } d^k\neq d^1\\ 1 & \text{if } d^k=d^1\end{cases}$$

where d^k, d^r and d^1 represent the distance to: the kth neighbour (furthest away), the rth (current) neighbour and the first (closest) neighbour, respectively. Assuming there are p votes from neighbours of class 1, and q votes from neighbours of class L, then the weighted votes for $\mathbf{x}$ from each class are, $d_1=\sum_{r=1}^{p}w_r$ and $d_L=\sum_{r=1}^{q}w_r$. Finally, $\mathbf{x}$ is assigned a class label based on the class with the maximum weight, i.e., $c_l=\arg\max(d_l)$.

Another variant of the *k*-NN classifier is the local mean vector (LMV) classifier [16]. This differs from the *k*-NN and DW *k*-NN in that the *k*-NNs are determined for each class, l. In this way, the local mean vector of every class, $\boldsymbol{\mu}_l$ is determined for the *k*-NNs of each class, such that $\boldsymbol{\mu}_l=(1/k)\sum_{r=1}^{k}\mathbf{x}_l^r$, where $\mathbf{x}_l^r$ denotes the rth neighbour among the *k*-NNs with class label c_l. Next, the distance, y_l, between the test sample, $\mathbf{x}$, and each class local mean vector, $\boldsymbol{\mu}_l$, is calculated. Finally, $\mathbf{x}$ is given a class label based on minimum distance between $\mathbf{x}$ and the respective local mean vector, i.e., $c_l=\arg\min(y_l)$.

All three of the conventional *k*-NN-based techniques, classify a single (individual) test sample, $\mathbf{x}$, one at a time and so can be considered examples of individual-based classification schemes. There are two obvious options for extending these methods to classify a group of test samples $\mathbf{X}_{\mathrm{TE}}=\{\mathbf{x}^1,\ldots,\mathbf{x}^{N_{\mathrm{TE}}}\}$ under the assumption that we have prior knowledge that the test samples, $\mathbf{X}_{\mathrm{TE}}$, all belong to *same, but unknown*, class.

For example, the conventional *k*-NN classifier can be extended to group-based classification by either:

1. Classifying each test sample individually and then enforcing an additional constraint that all, N_{TE}, samples in $\mathbf{X}_{\mathrm{TE}}$ have the same class label based on a majority vote of the individual class labels, c_l^i, of each test sample, $\mathbf{x}^i\in\mathbf{X}_{\mathrm{TE}}$. We refer to this two-stage approach as a (row-wise) *voting* scheme in Table 2;
2. Alternatively, we could pool the counts of the number of training samples of each class label in the *k*-nearest neighbours, $\mathbf{X}_{\mathrm{KNN}}$, for each individual test sample in $\mathbf{X}_{\mathrm{TE}}$. We refer to this approach as a (column-wise) *pooling* scheme in Table 2.

The final classification decisions are now made identically for every sample in $\mathbf{X}_{\mathrm{TE}}$ by finding the class that has the most pooled nearest neighbours, i.e., $\mathbf{c}_l=\arg\max(v_l^{\mathrm{p}})$, or has won the most votes from the individual test samples, i.e., $\mathbf{c}_l=\arg\max(v_l^{\mathrm{v}})$.

The extension of these group-based pooling and voting schemes to the distance weighted ***k*-NN** and local mean vector are shown in Tables 2 and 3. Therefore, we have can identify six potential group-based classifiers, all based on variations on the nearest neighbour principle. It should also be noted that in Table 3 we have used vector notation for the class label, $\mathbf{c}_l$, to emphasise that the same class label is given to all N_{TE} test samples in $\mathbf{X}_{\mathrm{TE}}$.

Clearly, Table 2 provides some insight into how almost any classification scheme can be extended to group-based classification. In fact, the voting schemes outlined in Table 2 are all variations of the fixed proportion classifiers commonly used in applications such as automated Pap smear screening [6]. Therefore, in this paper we are interested in comparing the efficacy of pooling and voting approaches, as well as determining what, if any, benefit is gained by utilising the prior knowledge that a group of test samples belong to same, but unknown, class.

Table 1
Three common nearest neighbour (NN) approaches to individual-based classification (IBC): conventional *k*-NN, distance weighted (DW) *k*-NN and local mean vector (LMV).

Test sample	Class l metric			Individual-based classification		
				***k*-NN**	**DW *k*-NN**	**LMV**
$\mathbf{x}$	v_l	d_l	y_l	$c_l=\arg\max_l(v_l)$	$c_l=\arg\max_l(d_l)$	$c_l=\arg\min_l(y_l)$

Table 2
Three group-based classifiers (GBC) based on variations of the nearest neighbour rule: group-based (GB) *k*-NN, group distance weighted (GDW) *k*-NN and the group local mean vector (GLMV).

Test set, $\mathbf{X}_{\mathrm{TE}}$	Sample x, class l metric			GB k-NN	GDW k-NN	GLMV
$\mathbf{x}^1$	v_l^1	d_l^1	y_1^1	$c_l^1 = \arg\max_l(v_l^1)$	$c_l^1 = \arg\max_l(d_l^1)$	$c_l^1 = \arg\min_l(y_l^1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}^{N_{\mathrm{TE}}}$	$v_l^{N_{\mathrm{TE}}}$	$d_l^{N_{\mathrm{TE}}}$	$y_1^{N_{\mathrm{TE}}}$	$c_l^{N_{\mathrm{TE}}} = \arg\max_l(v_l^{N_{\mathrm{TE}}})$	$c_l^{N_{\mathrm{TE}}} = \arg\max_l(d_l^{N_{\mathrm{TE}}})$	$c_l^{N_{\mathrm{TE}}} = \arg\min_l(y_l^{N_{\mathrm{TE}}})$
	$v_l^{\mathrm{p}} = \sum_{i=1}^{N_{\mathrm{TE}}} v_l^i$	$d_l^{\mathrm{p}} = \sum_{i=1}^{N_{\mathrm{TE}}} d_l^i$	$y_l^{\mathrm{p}} = \sum_{i=1}^{N_{\mathrm{TE}}} y_l^i$	$v_l^{\mathrm{v}} = \sum_{i=1}^{N_{\mathrm{TE}}} c_l^i$	$d_l^{\mathrm{v}} = \sum_{i=1}^{N_{\mathrm{TE}}} c_l^i$	$y_l^{\mathrm{v}} = \sum_{i=1}^{N_{\mathrm{TE}}} c_l^i$
	(Column-wise) pooling schemes			**(Row-wise) voting schemes**		

Table 3
The final group-based classification is based on either a pooling or voting scheme for the group-based (GB) *k*-NN, group distance weighted (GDW) *k*-NN and the group local mean vector (GLMV).

Group-based classifier	*Pooling scheme*	*Voting scheme*
GB *k*-NN	$\mathbf{c}_l = \arg\max_l(v_l^{\mathrm{p}})$	$\mathbf{c}_l = \arg\max_l(v_l^{\mathrm{v}})$
GDW *k*-NN	$\mathbf{c}_l = \arg\max_l(d_l^{\mathrm{p}})$	$\mathbf{c}_l = \arg\max_l(d_l^{\mathrm{v}})$
GLMV	$\mathbf{c}_l = \arg\min_l(y_l^{\mathrm{p}})$	$\mathbf{c}_l = \arg\min_l(y_l^{\mathrm{v}})$

Table 2 also highlights a relationship between the proposed group-based classifiers and the well-known component classifiers, such as Bagging ([4], Section 9.5.1). In general, component classifier techniques utilise multiple classifiers to label a single test sample, while here our group-based classifiers utilise a single classifier to label multiple test samples.

The paper is organised as follows: Section 3 presents the experimental methodology used to evaluate the performance of the proposed group-based nearest neighbour classifiers both on synthetic and real-world data sets. Section 4 presents the experimental results and Section 5 discusses their significance and implications. Finally, Section 6 offers some conclusions and suggestions for future work.

3. Experimental methodology

In this section we first describe the classification algorithms we have chosen to evaluate in comparison to the group-based nearest neighbour techniques proposed in Section 2. Next, we describe the re-sampling techniques used both to determine the number of nearest neighbours, k, used throughout the experiments and for estimating the error rate for each classifier. In particular, for the group-based classifiers we evaluate the effect of varying the group size, s. In its simplest form this is the number of samples in the test set, $\mathbf{X}_{\mathrm{TE}}$, assumed to belong to the same, but unknown, class. Finally, we briefly describe the synthetic and real-world data sets used for the empirical evaluation.

It should be noted that we have specifically chosen to evaluate and compare the performance of the classifiers using error rate estimated on an independent test set. Error rate is a conceptually simple and meaningful metric that is appropriate for data sets where the class priors and misclassification costs are either unknown or assumed equal [30,31]. In addition, the Bayes decision rule is known to minimise the probability of error and is therefore of both theoretical and practical interest [32].

3.1. The algorithms

While we are primarily interested in the relative performance of the three *k*-NN variants, and their extension to group-based classification using either voting or pooling, we also need to benchmark their performance against some commonly used individual-based classifiers that classify one single test sample at a time. In this paper we have chosen to use empirical Bayes [4,32], *k*-NN, and nearest neighbour (NN) [29]:

- The empirical Bayes (EB) classifier is a well established probabilistic classifier based on the direct application of Bayes rule. The class label is determined as the class with the maximum posterior probability based on the class priors and class-conditional probabilities. Here, we use a Gaussian multivariate parametric model and estimate the proportion, mean and covariance directly from the training data for each class ([32], Section 9.2). Clearly, on the synthetic data (described below), which is drawn from a Normal population, performance approaches that of the ideal Bayes classifier as the number of samples increases.
- Direct comparison of the group-based classifiers with *k*-NN allows us to estimate the benefit of our variants that classify test samples in groups. However, it does require us to select a value for k; which we discuss in detail in the next section. Therefore, we also use the nearest neighbour (NN) algorithm (i.e., $k=1$) both to observe the effect of selecting k and because as the number of samples approaches infinity, the error rate of NN is bounded by twice the ideal Bayes error [4,29].

The NN and empirical Bayes classifiers then give us a reasonable indication of the ideal Bayes error rate that an individual-based classifier could hope to achieve on a specific data set.

3.2. Selection of k

We select the 'optimal' value of k, based on the minimum empirical error rate observed on an independent test set. Specifically, for the *k*-NN classifier we utilise a *nested* 10-fold cross-validation technique, where each training set is further subdivided, using 10-fold cross-validation, so that the error rate on this *inner* test set can be estimated for various (odd) values of k (for $k=\{1, 3, 5, \ldots, 15\}$). For the *k*-NN classifier this 'locally optimal' value of k (which we shall refer to as k^*) can then be used to classify the test data in the *outer* test set. The average error rate on the *outer* cross-validation test data is then used to form an unbiased estimate of the true error rate of *k*-NN on the whole data set. In this way, each fold of the *outer* cross-validation may utilise a different value of k^*.

Rather than applying this strategy of nested cross-validation to find a locally optimal value of k for all of our proposed NN classifiers, we decided for both conceptual and computational simplicity, to use the value of k^* determined above for all of the group-based k-NN classifiers described in Section 2. Clearly, this is sub-optimal as it does not account for the interaction between the neighbourhood size, k, and the group size, s (described in Section 3.3). However, it does mean that the degenerate version of the group-based k-NN, i.e., with a group size $s=1$, has identical performance to the conventional k-NN classifier. In addition, using the same value of k^* for all group-based classifiers allows us to directly compare their performance independent of k. Therefore, we believe that evaluating the performance of the proposed group-based classifiers with a potentially (pessimistically) biased value of k^* is acceptable.

3.3. Error rate estimation

In this paper, we exclusively use *stratified* 10-fold cross-validation [33] which ensures that the class prior probabilities are maintained throughout the training and test partitions in all experiments (including those in Section 3.2). However, as we must form our test data into homogeneous subsets to perform group-based classification some minor modifications must be made. Therefore, for every class, $l=\{1, 2, \ldots, L\}$, we divide the samples into ten partitions, as per conventional cross-validation. Then for each fold of the cross-validation, the test set of class l, $\mathbf{X}_{\mathrm{TE}}^{l}$, is sub-sampled into subsets of (group) size s. To prevent ties in the group-based classification (for the two-class case) we make s odd, i.e., $s=\{1, 3, 5, \ldots, S\}$, where S is the maximum group size, determined as the minimum of 15 and the number of samples in $\mathbf{X}_{\mathrm{TE}}^{l}$ (that is, the cardinality of $\mathbf{X}_{\mathrm{TE}}^{l}=N_{\mathrm{TE}}^{l}$), i.e., $S=\min(N_{\mathrm{TE}}^{l},15)$. For example, in the experiments with the Iris data set, $s=\{1, 3, 5\}$ are evaluated, since the maximum size of the test set is only $S=5$. The sub-sampling of $\mathbf{X}_{\mathrm{TE}}^{l}$ into subsets of size s produces a total number of subsets equal to the number of combinations of choosing s from N_{TE}^{l}, that is

$$\mathrm{comb}(\mathbf{X}_{\mathrm{TE}}^{l},s)=\frac{N_{\mathrm{TE}}^{l}!}{(N_{\mathrm{TE}}^{l}-s)!s!}$$

Obviously, when N_{TE}^{l} is large, the number of combinations can be very large and so we limit our evaluation to a maximum of 100 randomly selected subsets. Finally, we estimate the average error rate over the 10-fold cross-validation for each group size, s. In this way, groups of test samples, of various sizes, are presented to a group-based classifier and given a homogenous class label $\mathbf{c}_l$, where the dimensionality of $\mathbf{c}_l=s$.

3.4. Experimental data

The proposed group-based classification techniques are tested on both synthetic and real-world data. The synthetic data sets are briefly described in Section 3.4.1, while the descriptions of the real-world data sets, namely, the Pap smear and Iris data sets are presented in Sections 3.4.2 and 3.4.3, respectively.

3.4.1. Synthetic data

We briefly describe three commonly used Gaussian data sets, namely the I-I, I-Λ and I-4I data sets originally developed by Fukunaga [15]. Notably, each data set has a different level of 'difficulty' with calculated Bayes error rates of 10%, 1.9% , and 9% for I-I, I-Λ, and, I-4I, respectively. Each data set consists of an 8-dimensional data vector with 1000 samples per class. In these synthetic data sets, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors for class 1 and class 2, respectively. Meanwhile, $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are the corresponding covariance matrices for each class. The values of the $\boldsymbol{\mu}_l$ and $\boldsymbol{\lambda}_l$ are given in Table 4, where I_8 is the 8×8 identity matrix, and for the I-Λ data set, the $\boldsymbol{\mu}_2$and $\boldsymbol{\lambda}_2$ are provided in Table 5.

Table 4
The synthetic data sets I-I, I-4I and I-Λ.

Data set	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\lambda}_1$	$\boldsymbol{\lambda}_2$
I-I ($\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$, $\boldsymbol{\lambda}_1=\boldsymbol{\lambda}_2$)	0	[2.56, 0, …, 0]	I_8	
I-Λ ($\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_2$, $\boldsymbol{\lambda}_1\neq\boldsymbol{\lambda}_2$)		[μ_1, …, μ_8] (Table 5)	I_8	[λ_1, …, λ_8] (Table 5)
I-4I ($\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$, $\boldsymbol{\lambda}_1\neq\boldsymbol{\lambda}_2$)	0		$4I_8$	

Table 5
Parameter values of the I-Λ data set.

Dimension i	1	2	3	4	5	6	7	8
μ_i	3.86	3.10	0.84	0.84	1.64	1.08	0.26	0.01
λ_i	8.41	12.06	0.12	0.22	1.49	1.77	0.35	2.73

3.4.2. Pap smear data

The purpose of the experiments with the real-world data sets is to evaluate the use of group-based classification in practical applications. The first real-world data set used in this paper is the Pap smear data set available previously detailed in [34,35] (and available from http://fuzzy.iau.dtu.dk/download/smear2005). The data set consists of 917 individual cells, each represented by a 20-dimensional feature vector. The original data set is labelled into seven classes. However, for this work we have simplified the data set to a two-class problem: normal (superficial squamous epithelia; intermediate squamous epithelial; and columnar epithelial) versus abnormal (mild squamous non-keratinising dysplasia; moderate squamous non-keratinising dysplasia; severe squamous non-keratinising dysplasia; and squamous cell carcinoma in situ intermediate). In this way, we created a data set consisting of 242 normal and 675 abnormal cells. Since our experiments are conducted using stratified 10-fold cross-validation, each partition consists of approximately 24 normal and 67 abnormal cells. In addition, we utilised only the ten most discriminatory features, as selected in [35], namely: nucleus area (N); cytoplasm area (C); N/C ratio; nucleus brightness; cytoplasm brightness; nucleus shortest diameter; nucleus longest diameter; nucleus perimeter; nucleus position; and maxima in nucleus. Prior to the experiments, we applied a normalisation transform to ensure all feature measurements were zero mean, unit variance (as per [36]).

The aim of group-based classification is to determine a single class label for a group (set) of test samples. Therefore, on this data set group-based classification simulates the situation where a group of cells have come from a single subject and so should be given a homogeneous class label: either normal or abnormal. However, it should be pointed out that that this data set was not collected with group-based classification in mind and so we have no knowledge of which cells came from which subjects. In addition, typically there would be many 1000 s of cells on a Pap smear slide and these, if classified individually, may appear to be normal and/or abnormal. Therefore, it is important that the interpretation of the class label should now be made at the group, not cell, level. That is, the group label indicates whether the sample of cells is from a normal or abnormal slide [2].

3.4.3. Iris data

The second real-world data set used in this paper is the Iris data set [8], which is one of the most commonly used data sets in

the pattern recognition literature. The Iris data set consists of 150, four-dimensional, sample vectors where each sample is represented by: petal length; petal width; sepal length; and sepal width. There are three classes: Setosa; Versicolor; and Virginica, with 50 samples per class. Since our experiments were conducted using stratified 10-fold cross-validation, each test partition consists of five samples per class. Similarly to the Pap smear data, the test set of the Iris data is organised so that all samples have the same, but unknown, class label. Again, as the Iris data set was not gathered with group-based classification in mind, we assume that each group of test samples are homogeneous, i.e., as if they have been collected either from the same plant, or an assortment of plants that are of the same, but unknown, species.

4. Results

We tested the performance of the proposed group-based classification techniques (GB k-NN, GDW k-NN and GLMV) with both pooling and voting schemes and compared their performances with three individual-based classifiers, namely, EB, k-NN and NN. It should be noted that for the individual-based classifiers, only a single test sample is classified at a time, while for the group-based classifiers the test set (group) size, s, is varied.

4.1. Synthetic data

Figs. 2–4 show the estimated error rate as a function of group size for each of the synthetic data sets. These figures show that for the three individual-based classifiers there is a general trend for decreasing error rate from NN (highest) to k-NN (middle) to EB (lowest), with EB forming a close estimate to the calculated Bayes error for this data (as per Section 3.4.1). The reduced error rate of k-NN compared to NN can be explained by the use of nested cross-validation to select an appropriate value of k^* on each data set. While the actual value varies for each fold of the cross-validation, the median value of k^* were 13, 5 and 3 on the I-I, I-Λ and I-4I data sets, respectively. These values are in broad agreement with previous results in [16] . In addition, the degenerate case of $s=1$, shows that the GB k-NN and traditional k-NN have the same error rate (as they should). However, as the group size, s, increases, all of the group-based techniques demonstrate a decrease in error rate, with the exception of the I-4I data set, where the error rate initially increases, not falling below the initial error rate ($s=1$) until $s>5$. In addition, on all three data sets, as s increases, the error rate of the group-based techniques falls below the estimated Bayes error rate (for an individual-based classifier). The only exceptions to this trend are the GB k-NN and GDW k-NN classifiers on the I-4I data set. Here only the GLMV technique with $s=15$ demonstrates an error rate below the estimated Bayes rate. For the I-I and I-Λ data sets the estimated error rate approaches zero as the group size increases ($s>11$ and 5, respectively).

We observe that the most challenging classification task is on the I-4I data set, where both classes have the same mean (zero), but different covariance matrices. Clearly, the nearest neighbour paradigm is not well suited to this type of problem and the group size has to be significantly larger before the error rate approaches that of EB. However, even though I-4I is a difficult problem, all the group-based techniques eventually show a decreasing error rate as the group size increases.

Table 6 illustrates the number of times either a pooling (P) or voting (V) scheme produced the lowest error rate on each of the data sets, with '–' denoting a tie. Table 6 shows over the three data sets the pooling scheme performs significantly better than voting scheme. However, as the group size gets larger, the error rate typically aproaches zero and so both schemes perform equally. Table 6 also shows that the group-based local mean vector (GLMV) tends to produce the lowest error rate on all three data sets, except when $s<7$ on the I-I and I-4I data sets. The highest error rate is approximately equally likely from the group distance weighted or group-based k-NN on all three data sets.

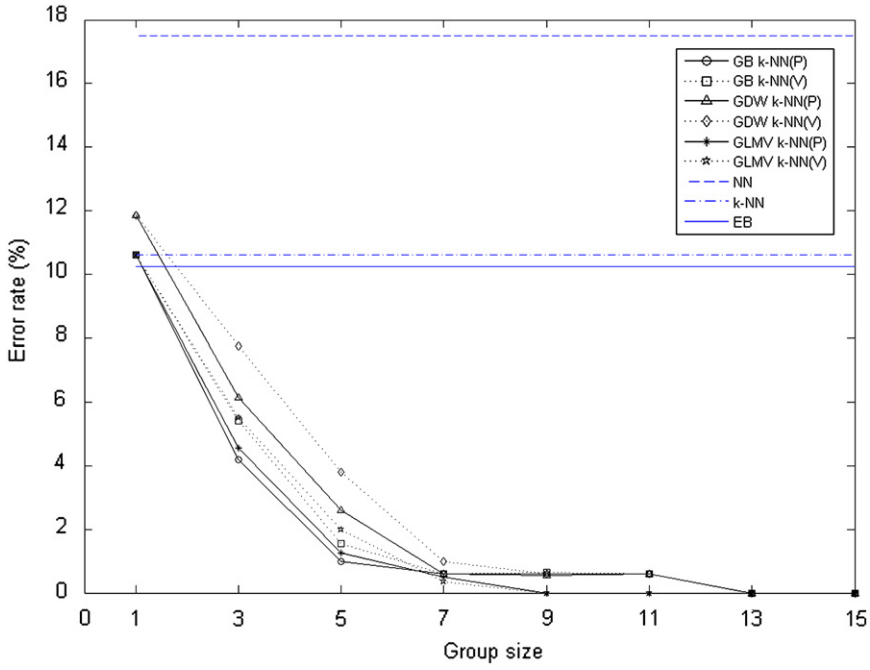

Fig. 2. Error rate versus group size for the I-I data set.

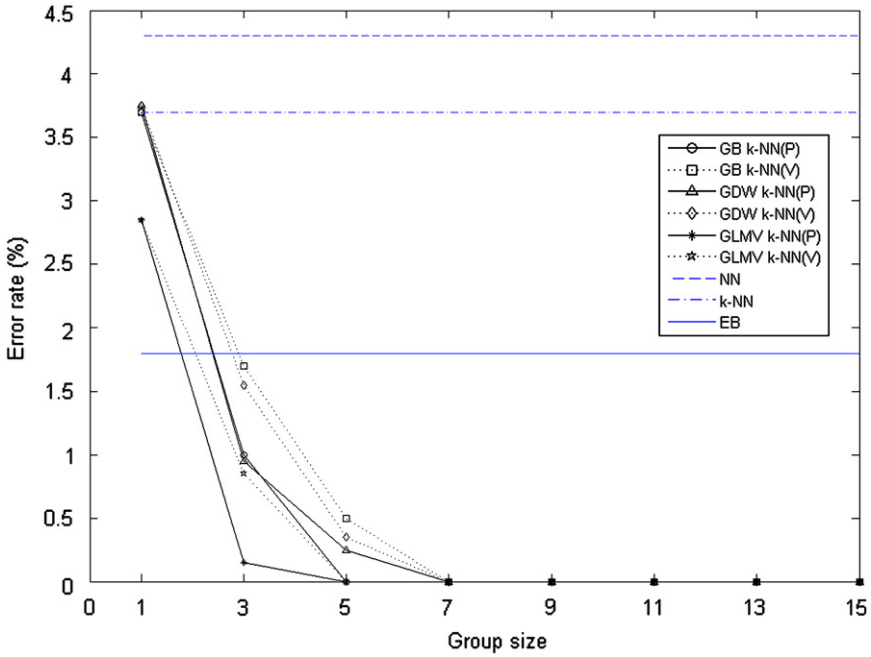

Fig. 3. Error rate versus group size for the I-Λ data set.

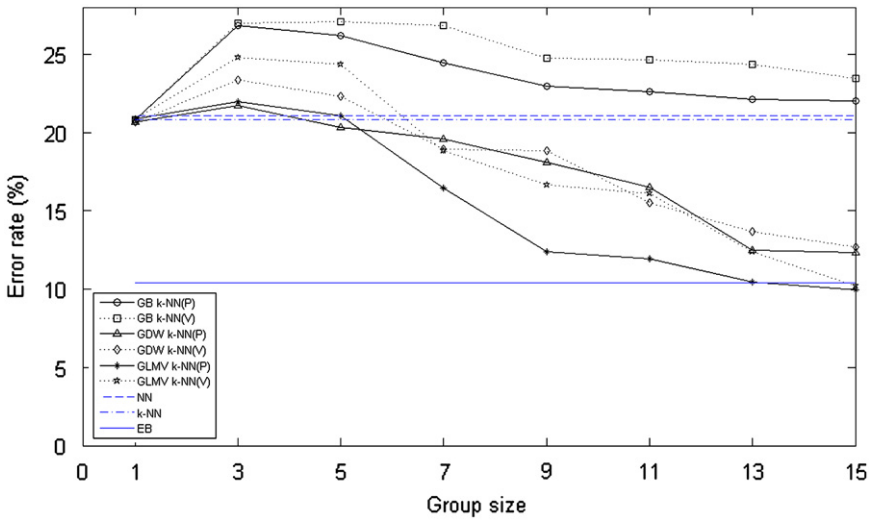

Fig. 4. Error rate versus group size for the I-4I data set.

4.2. Pap smear data

Fig. 5 illustrates the estimated error rate as a function of group size, s, for the Pap smear data and Table 7 the corresponding summary of lowest error rate for the pooling versus voting group-based classifiers. Fig. 5 shows that, in contrast to the synthetic data sets, the empirical Bayes classifier has the highest error rate, with k-NN the lowest. Here, the nested cross-validation selected a median value of $k^*=5$. Fig. 5 also shows that all of the proposed group-based techniques perform better than the individual-based classifiers when the group size is larger than one. Indeed, with the exception of the voting GB k-NN, all group-based classifiers achieve a zero error rate when the group size is larger than five. Table 7 indicates that in general the pooling schemes perform better than the voting schemes when the group size is less than seven. The exception to this is for GDW k-NN where the voting scheme achieves the lowest error rate for group size three. Again, GLMV appears to produce the lowest error rate and GB k-NN the highest.

4.3. Iris data

Fig. 6 illustrates the estimated error rate as a function of group size on the Iris data set. Fig. 6 shows that, as per the synthetic data sets, among the individual-based classifiers EB has the lowest error rate with NN next and k-NN the highest. Here, the nested cross-validation selected a median value of $k^*=3$. However, there was a broad range of value selected from $k^*=1$ to 7. Again, Fig. 6 illustrates that the estimated error rates for all of the proposed techniques decrease when the group size

Table 6
Summary of results for the pooling versus voting schemes on the synthetic data sets. Here, 'P' and 'V' indicate whether the pooling or voting scheme has the lowest error rate on this data set, respectively, while '–' indicates a tie.

	Group size (s)						
	3	5	7	9	11	13	15
Data set: I-I							
GB k-NN	P	P	–	P	–	–	–
GDW k-NN	P	P	V	P	–	–	–
GLMV k-NN	P	P	P	–	–	–	–
Total P	3	3	1	2	0	0	0
Total V	0	0	1	0	0	0	0
Lowest error	GB k-NN	GB k-NN	GLMV k-NN	GLMV k-NN	GLMV k-NN	–	–
Highest error	GDW k-NN	GDW k-NN	GDW k-NN	GDW k-NN	GDW k-NN	–	–
Data set: I-Λ							
GB k-NN	P	P	–	–	–	–	–
GDW k-NN	P	P	–	–	–	–	–
GLMV k-NN	P	–	–	–	–	–	–
Total P	3	2	0	0	0	0	0
Total V	0	0	0	0	0	0	0
Lowest error	GLMV k-NN	GLMV k-NN	–	–	–	–	–
Highest error	GB k-NN	GB k-NN	–	–	–	–	–
Data set: I-4I							
GB k-NN	P	P	P	P	P	P	P
GDW k-NN	P	P	V	P	V	P	P
GLMV k-NN	P	P	P	P	P	P	P
Total P	3	3	2	3	2	3	3
Total V	0	0	1	0	1	0	0
Lowest error	GDW k-NN	GDW k-NN	GLMV k-NN	GLMV k-NN	GLMV k-NN	GLMV k-NN	GLMV k-NN
Highest error	GB k-NN	GB k-NN	GB k-NN	GB k-NN	GB k-NN	GB k-NN	GB k-NN

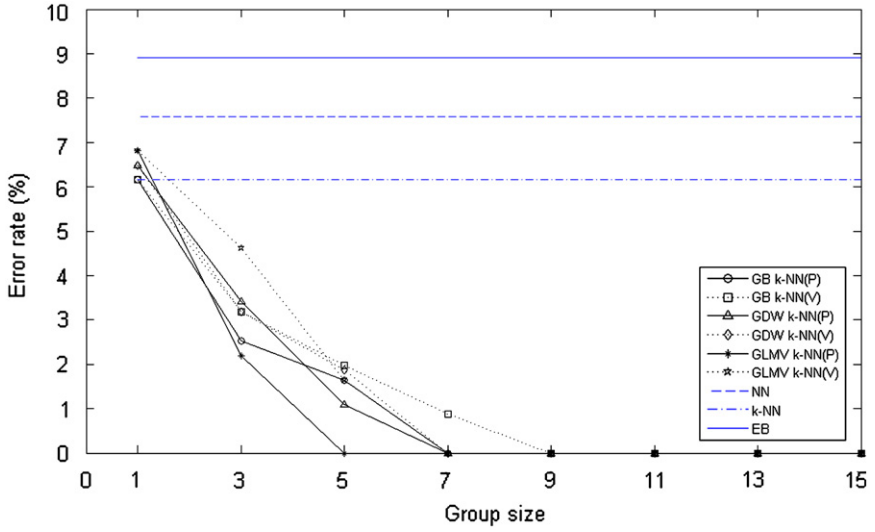

Fig. 5. Error rate versus group size for the Pap smear data set.

Table 7
Summary of pooling versus voting scheme for Pap smear data set.

Data set: Pap smear	Group size						
	3	5	7	9	11	13	15
GB k-NN	P	P	P	–	–	–	–
GDW k-NN	V	P	–	–	–	–	–
GLMV k-NN	P	P	–	–	–	–	–
Total P	2	3	1	0	0	0	0
Total V	1	0	0	0	0	0	0
Lowest error	GLMV	GLMV	–	–	–	–	–
Highest error	GLMV	GB k-NN	GB k-NN	–	–	–	–

increases to three and five. All group-based classifiers aproach a zero error rate at a group size of three and five, with both the voting and pooling schemes performing equally well on this data set.

5. Discussion

The results on all five data sets clearly demonstrate that the error rate obtained when classifying test samples individually can be reduced by classifying a set of test samples as a group, i.e., at the same time. These results also show that as the group size is

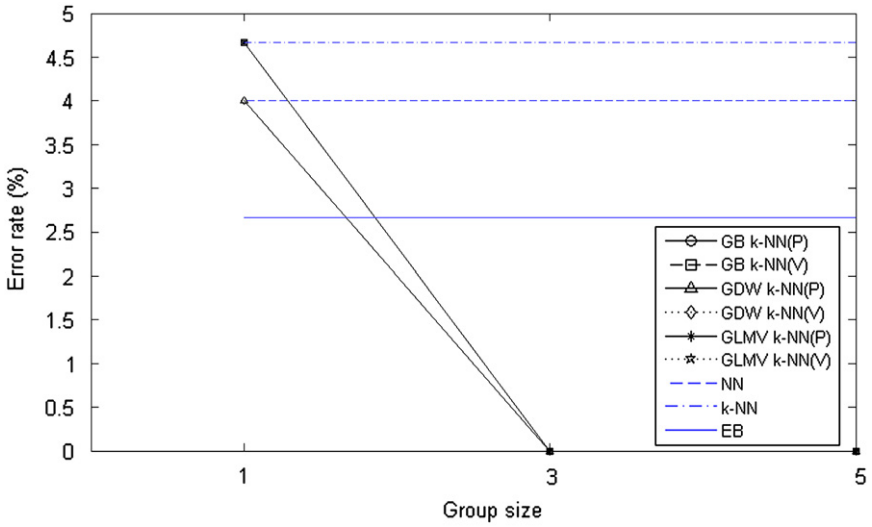

Fig. 6. Error rate versus group size for the Iris data set.

increased the error rate typically continues to decreases until, in four out of the five data sets, an error rate of zero is achieved. This would indicate that the additional prior knowledge that a group of test samples belong to same, but unknown, class label can be effectively utilised to reduce the, apparently optimal, Bayes error rate. That is, on the data sets evaluated here, the *group*-based Bayes error rate would appear to be significantly lower than the traditional (individual) Bayes error rate. These results not only agrees with our previous findings [2], but clearly demonstrate that the additional prior knowledge utilised by group-based classification is the same contextual information that aids compound decision making in general [4].

The contextual prior knowledge contained in the group of test samples, if utilised via a pooling or voting scheme can reduce the influence of 'outliers.' That is, samples that would be on the wrong side of the decision boundary if classified individually can now be classified correctly due to their context. In this way, having a number of test samples in a homogenous group can provide enough information for the test samples, as a whole, to be correctly classified and misclassifications on individual 'outliers' avoided. Indeed, the results indicate that the larger the group size, the better the performance of the group-based classification. In other words, group-based classification depends on the feature statistics of the group as a whole, rather than on individual sample features. A consequence of this observation, is that a group-based classifier does not have a 'decision boundary' in the traditional sense, as class labels are not determined for an individual point in feature space, but rather are based on the statistics of a group of samples in feature space. That is, a specific test sample, at one point in feature space, may be labelled differently depending on the context provided by the remaining samples in the group. This was also illustrated in Fig. 1.

The other finding apparent from the results (on all five data sets) is that the group-based classifiers that utilise pooling are more effective than those that utilise a two-stage approach; classifying each test sample individually and then voting on the class label for the group. This result should be expected as pooling utilises the raw counts, or distances to local neighbours, rather than just final class labels. This information is clearly more informative to the group-based decision and should therefore be preferred. It is interesting to note that pooling has the biggest advantage over simple voting on the most challenging data set (I-4I: same mean, different covariance). In addition, this result would also seem to illustrate a much older result related to the limitations of (simple) rules that use only the *i*th observation to make the *i*th decision for compound decision problems [37].

Over the five data sets evaluated here there is no clear consistency as to which specific form of group-based classification produces the lowest error rate overall. Tables 6 and 7 show that GLMV appears to have the lowest error rate the most often and that GB *k*-NN, GDW *k*-NN tend to have the highest errors rates. One possible explanation for the lower error rate of GLMV is that it utilises the *k*-nearest neighbours from both classes and for group-based classification, where decisions are based on group statistics, this may proffer an (unfair) advantage. Clearly, some methods are better suited to certain types of problems and so, as for individual-based classifiers, there is 'no free lunch' and we must select the group-based classifier that is best suited to the particular classification problem at hand. In addition, while any classifier could be extended to classify groups of samples by either a pooling or voting approach, only in certain application domains is it possible to group the data so that this additional context can be utilised.

6. Conclusions

In this paper, we have extended three variants of the non-parametric nearest neighbour technique to group-based classification (GB *k*-NN, GDW *k*-NN and GLMV). In particular, two different ways of accumulating information about a group of test samples were presented, referred to as pooling and voting schemes, respectively. The proposed techniques were then evaluated for a variety of group sizes using both synthetic and real-world data and their performance evaluated, in terms of average error rate, against three individual-based classifiers (EB, *k*-NN and NN). The results indicate that the proposed group-based classification techniques have the potential to outperform the individual-based techniques, especially as the (group) size of the test set increases. In addition, it is shown that algorithms that pool information from the whole test set perform better than those that apply a two-stage approach, based on a vote of the class labels of individual test samples. These results indicate that the additional prior knowledge that a group of test samples belong to same, but unknown, class label can be effectively utilised to reduce the individual Bayes error rate.

Future work will include the expansion of group-based classification to other machine learning algorithms and the development of the specific details of the relationship between group-based classification and other forms of non-sequential compound classification.

References

[1] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 4–37.
[2] N.A. Samsudin, A.P. Bradley, Group-based meta-classification, 19th International Conference on Pattern Recognition, IEEE, Tampa, Florida, 2008.
[3] G.T. Toussaint, The use of context in pattern recognition, Pattern Recognition 10 (1978) 189–204.
[4] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, Inc., New York, 2001.
[5] D. Lewis, Naive (Bayes) at forty: the independence assumption in information retrieval, in: Proceedings of ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE, 1998, pp. 4–15.
[6] B. Nordin, E. Bengtsson, Specimen analysis by rare event, cell population, and/or contextual evaluation, in: H.K. Grohs, O.A.N. Husain (Eds.), Automated Cervical Cancer Screening, IGAKU-SHOIN Medical Publishers, New York, 1994, pp. 44–51.
[7] T.C.W. Landgrebe, P. Paclik, D.M.J. Tax, R.P.W. Duin, Optimising two-stage recognition systems, in: International Workshop on Multiple Classifier Systems, Monterey, California, 2005.
[8] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188.
[9] R. Dybowski, V.A. Gant, P.A. Riley, I. Phillips, Rapid compound pattern classification by recursive partitioning of feature space: an application in flow cytometry, Pattern Recognition Letters 16 (1995) 703–709.
[10] B.V. Dasarathy, Nearest Neighbor Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, 1991.
[11] S.A. Dudani, The distance-weighted *k*-nearest neighbor rule, IEEE Transactions on Systems, Man, and Cybernetics SMC-6 (1976) 325–327.
[12] T. Baily, A.K. Jain, A note on distance-weighted *k*-nearest neighbor rules, IEEE Transactions on Systems, Man, and Cybernetics 8 (1978) 311–313.
[13] J.E.S. MacLeod, A. Luk, D.M. Titterington, A re-examination of the distance-weighted *k*-nearest neighbor classification rule, IEEE Transactions on Systems, Man, and Cybernetics SMC-17 (1987) 689–696.
[14] R.L. Morin, D.E. Raeside, A reappraisal of distance-weighted *k*-nearest neighbor classification for pattern recognition with missing data, IEEE Transactions on Systems, Man, and Cybernetics SMC-11 (1981) 241–243.
[15] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, 1990.
[16] Y. Mitani, Y. Hamamoto, A local mean-based nonparametric classifier, Pattern Recognition Letters 27 (2006) 1151–1159.
[17] L. Jiang, H. Zhang, Z. Cai, J. Su, Evolutional naive Bayes, in: Proceedings of the First International Symposium on Intelligent Computation and its Applications, 2005, pp. 344–350.
[18] R. Kohavi, G. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273–324.
[19] P. Langley, S. Sage, Induction of selective Bayesian classifiers, in: Proceedings of the Tenth National Conference of Artificial Intelligence, 1994, pp. 400–406.
[20] K.K. Han, Text categorization using weight adjusted *k*-nearest neighbor classification, Technical Report, Department of CS, University of Minnesota 1999.
[21] Z. Xie, W. Hsu, Z. Liu, M.L. Lee, SNNB: a selective neighborhood based naive Bayes for lazy learning, in: Sixth Pacific-Asia Conference on KDD, 2002, pp. 104–114.
[22] E. Frank, M. Hall, B. Pfahringer, Locally weighted naive Bayes, in: Conference on Uncertainty in Artificial Intelligence, 2003, pp. 249–256.
[23] L. Jiang, H. Zhang, Z. Cai, Dynamic *k*-nearest neighbor naive Bayes with attribute weighted, in: Third International Conference on Fuzzy Systems and Knowledge Discovery, 2006, pp. 365–368.
[24] L. Jiang, H. Zhang, J. Su, Instance cloning local naive Bayes, in: Eighteenth Canadian Conference on Artificial Intelligence, 2005, pp. 280–291.
[25] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman, A.Y. Wu, An optimal algorithm for approximate nearest neighbor searching in fixed dimensions, Journal of the ACM 45 (1998) 891–923.
[26] J.L. Bentley, Multidimensional binary search trees used for associative searching, Communications of the ACM 18 (1975) 509–517.
[27] J.M. Kleinberg, Two algorithms for nearest neighbor search in high dimensions, in: 29th Annual ACM Symposium on Theory of Computing, 1997, pp. 599–608.
[28] P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, in: 30th Symposium on Theory of Computing, 1998, pp. 604–613.
[29] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1967) 21–27.
[30] A.P. Bradley, ROC curves and the chi-square test, Pattern Recognition Letters 17 (1996) 287–294.
[31] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition 30 (1997) 1145–1159.
[32] W.L. Martinez, A.R. Martinez, Computational Statistics Handbook with MATLAB, Chapman & Hall/CRC, 2002.
[33] E. Alpaydin, Assessing and comparing classification algorithms, Introduction to Machine Learning, The MIT Press, London, 2004.
[34] A. Tsakonas, G. Dounias, J. Jantzen, H. Axer, B. Bjerregaard, D.G. von Keyserlingk, Evolving rule based systems in two medical domains using genetic programming, Artificial Intelligence in Medicine 32 (2004) 195–216.
[35] Y. Marinakis, M. Marinaki, G. Dounias, Particle swarm optimization for pap smear diagnosis, Expert Systems with Applications 35 (2008) 1645–1656.
[36] J. Han, M. Kamber, Data Mining Concepts and Techniques, second ed., Morgan Kauffmann, San Francisco, 2006.
[37] E. Samuel, On simple rules for the compound decision problem, Journal of the Royal Statistical Society, Series B (Methodological) 27 (1965) 238–244.

Noor A. Samsudin received her B.Sc. degree in Computer Science in 1996 from the University of Missouri-Columbia, USA and her Masters degree in Computer Science in 2001 from the National University of Malaysia. She is currently a Ph.D. candidate in the School of Information Technology and Electrical Engineering at The University of Queensland, Australia. Her research interests include pattern classification and machine learning.

Andrew Bradley is a Senior Lecturer in Biomedical Engineering at The University of Queensland. He received his Ph.D. from The University of Queensland in 1996 and since then has held various research positions in Australia, the United Kingdom and Canada. His research interests are currently focused on biomedical applications of pattern recognition in signal and image analysis. He is a Chartered Electrical Engineer and Senior Member of the IEEE.