Pairwise *vs* Global Multi-Class Wrapper Feature Selection

HUGO SILVA
Instituto de Telecomunicações
Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisbon
PORTUGAL

ANA FRED
Instituto de Telecomunicações
Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisbon
PORTUGAL

Abstract: Wrapper feature selection methods are typically used in multi-class classification problems to determine which feature subspace maximizes the patterns discriminative potential, with respect to the global multi-class scope. However, in most classification tasks, some classes are more easily discriminated than others due to particularly predictive features. Thus the global class set may stand as a hard restriction when performing feature selection. We propose a class pairwise approach, in which the wrapper feature selection framework is applied with the purpose of determining the feature subspaces with higher discriminative potential for each class pair. This method is shown to provide simpler models, reduced number of features, higher scalability, and in some cases even improve the classification performance.

Key–Words: Feature selection, Data representation, Pattern recognition, Dimensionality reduction

1 Introduction

In pattern classification systems each pattern is usually represented as a feature vector consisting of properties, singularities, or measurements [10, 5]. These can often reach dozens or even hundreds, however not all of the available features may be relevant. Some can be redundant, some can be completely irrelevant, and some can even be misleading. As a result, an initial step in pattern analysis is to determine relevant features for the problem at hand, or determining better, alternative representations of the patterns [19]. The motivations can emerge from different purposes [8], among which the most common are:

(a) dimensionality reduction;

(b) improvement of predictive performance;

(c) facilitate data visualization and understanding.

This task is normally addressed either through feature selection or feature extraction. Feature selection (or reduction) techniques use variable ranking criteria's (filter methods) [6, 2], or the predictive performance of a learning machine (wrapper or embedded methods) [11] to select higher performance individual or groups of features from the original space. Feature extraction techniques use transformations of the original feature space, in order to extract relevant cross-feature information [20, 4]. In this paper we will be focusing on feature selection techniques, although the same base concepts can also be extended to feature extraction.

Numerous methodologies have been exploited over the years to tackle the feature selection problem [11, 12, 6], with different characteristics in terms of search strategy, and subspace evaluation criteria, among others [1, 9]. Typically, feature selection is performed by searching the global multi-class scope of known patterns using a chosen strategy. The purpose is to determine the subspace of features, considered to be the best according to a specified feature subspace evaluation criteria, and which will subsequently be used for pattern classification.

Generally, in multi-class classification tasks some classes are more difficult to classify than others due to the lack of good predictive features for that class [7]. Thus the global multi-class scope may stand as a hard restriction, in the sense that the elected subspace of features is the one which optimizes the evaluation criteria on data belonging to all the classes. Subspaces resulting from this approach have to be elaborate enough to fit the whole set of classes, and the search can be misguided in complex class sets. Furthermore, when new classes are added to the original set, there is no guarantee that the previously determined subspaces remains adequate. Thus it becomes necessary to re-compute new feature subspaces for the new global multi-class scope.

In [7] this problem is explored, and a method is proposed in which a filter feature selection framework [11], is used to determine for each individual class which feature subspace better discriminates it from the remaining classes (*one-vs-all* approach). In this paper we propose to perform feature selection based on class pairwise feature selection and classification in the context of wrapper methods. We show that by using only pairs of classes rather than the global multi-class scope as a search base, we get a better insight into the underlying data models, and promote pattern classification system scalability, while retaining its predictive accuracy.

The rest of the paper is organized as follows. Section 2 presents the class pairwise feature selection approach. Section 3 presents experimental results of the class pairwise approach with a sequential space state search wrapper method. Finally, Section 4 summarizes results and main conclusions.

2 Class Pairwise Feature Selection

2.1 Motivation

Let X denote a $m \times n$ matrix of m patterns and n features belonging to $W = \{w_1, ..., w_k\}$ classes. Each pattern x_i $(1 \leq i \leq m)$, is a feature vector of the form $x_i = [f_{1_{x_i}}, ..., f_{n_{x_i}}]$ belonging to class $w_{x_i} \in W$. From the known set of patterns X, the common approach for feature selection consists of conducting the search on the full class set W, in order to determine the feature subspace with higher discriminative potential.

Employing a search strategy S with the purpose of selecting relevant features from the initial feature space $F = [f_1, ..., f_n]$ would identify the subspace F^* computed as the best given the feature subspace evaluation criteria J. Considering all classes W, and despite the adopted framework being wrapper or filter, what J ranks is how well a given subspace fits the global multi-class scope of known patterns. Although with lower dimension than F, the selected subspace F^* (and consequently built models), is determined inherently regarding how well patterns from all classes W are distinguished, when represented through F^*. Our hypothesis is that even lower dimensional subspaces (and consequently less complex models), can be obtained if the search is performed only on subsets of W rather than the global scope.

Additionally, determining the best subspace from the global multi-class scope W imposes that, whenever a new class w_{k+1} is added to the problem, the best subspace is re-determined for the new multi-class scope $W \cup \{w_{k+1}\}$[1]. This results from the fact that a feature subspace F^* evaluated as the best for a set W of classes, is not guaranteed to remain the best, nor to perform as well on $W \cup \{w_{k+1}\}$ classes without exploration of the new global multi-class scope [3]. For the same reason, heuristic methods of determining a suitable subspace for the new set, such as using previously rejected or unused features [2] to update F^*, are not guaranteed to produce adequate results. If subsets of W are used to perform the search, adaptation of the pattern classification system to new classes becomes more flexible.

In the next section we describe a method of feature selection that works by determining relevant feature subspaces only for subsets of classes, rather than for the global multi-class scope.

2.2 Search Procedure and Classification

In terms of feature selection, in a multi-class problem the most elementary subset containing sufficient information to discriminate between classes, possible to construct from an initial set W of k classes, is composed by two elements (the unitary class set only provides individual description information [16, 18]). Thus, instead of performing feature selection using the global class set, we propose to use these elementary two class subsets (hence the term class pairwise feature selection) and search for the best feature subspace for each pair.

Algorithm 1 lists the generic class pairwise feature selection algorithm. As inputs we have a set of X patterns, belonging to a set W of k classes, a search strategy S, and a feature subspace evaluation criteria J. From W, every possible distinct pair $\{w_a, w_b\}$ $(1 \leq a, b \leq k \wedge a \neq b)$ of classes is formed. For each class pair $\{w_a, w_b\}$, we compute the best feature subspace by applying the search strategy S considering only the patterns $x_i \in X$, for which the corresponding class w_{x_i} belongs to the pair $(w_{x_i} \in \{w_a, w_b\})$. J is used as subspace evaluation criteria in the search strategy. As a result $k(k-1)/2$ differentiated feature subspaces $F^*_{\{w_a, w_b\}}$ are produced, corresponding to the individual subspaces computed as the best for each class pair. This approach is more demanding in terms of spatial representation, however the tradeoff comes from model simplification, and scalability of the pattern classification system, as discussed subsequently.

By reducing the abstraction level in which the data set is analyzed to only two classes with class pairwise feature selection, the ability to use a single classifier is lost. Instead, the classification depends on individual decisions produced from each class pair using the corresponding feature subspace. There are several strategies to address the problem of combining decisions from multiple classifiers in multi-class

[1] the same applies to the case where a class $w_k \in W$ is removed

Algorithm 1 Generic class pairwise feature selection algorithm.

```
Require:
  $X$ - labeled data set
  $W$ - class set
  $S$ - search strategy
  $J$ - subspace evaluation criteria
Ensure:
  $F^*$ - set feature subspaces
  for $\{w_a, w_b\} \in W : a \neq b$ do
    $D = \{x_i \in X : w_{x_i} \in \{w_a, w_b\}\}$
    $F^*_{\{w_a,w_b\}} = S(D, J)$
  end for
```

classification [5]. For class pairwise feature selection we propose majority voting of decisions of individual classifiers [13], which was shown to hold effectiveness in class pairwise classification, when compared to probabilistic approaches [17]. Each individual classifier votes the class to which a pattern is most likely to belong, and the assigned class is the one most individual classifiers vote for.

2.3 Model Simplification

Consider Figure 1(a) showing a data set with four classes $W = \{w_1, ..., w_4\}$, where each class is modeled by a gaussian distribution. Patterns from each class are represented in a 3-dimensional feature space $F = [f_1, f_2, f_3]$. No projection to a lower dimensional space provides adequate multi-class separability, since every 2-dimensional projection overlaps two classes (Figures 1(b)-1(d)), and every 1-dimensional projection overlaps three classes.

To properly address this problem by searching the global multi-class set W, it would be necessary to consider the full space F in order to obtain adequate class separability. With class pairwise feature selection however, a single feature is sufficient to separate between each class pair. The initial space F of three features is therefore modeled by six simpler individual spaces with only a single feature. From Figure 1 we can see that by class pairwise feature selection: $F^*_{\{w_4,w_1\}} = [f_3]$, $F^*_{\{w_2,w_1\}} = F^*_{\{w_2,w_4\}} = [f_2]$, and $F^*_{\{w_3,w_1\}} = F^*_{\{w_3,w_2\}} = F^*_{\{w_3,w_4\}} = [f_1]$. Simpler feature spaces facilitate data visualization and understanding, which are two of the purposes of feature selection [8], and in some domains can also provide classification improvements.

These benefits become more relevant when the ratio between the number of features and the number of classes is higher. Depending on the strategy, the complexity of the domain can misguide the search [14]. By performing a search only on portions of the class domain rather than on the global multi-class scope, like in class pairwise feature selection, problems arising from the structure of the feature domain, like the lack of predictive features for a particular class, have a lesser impact on the search results improving the pattern classification system performance.

Table 1: Dataset characterization.

database	classes	features	patterns
Wine Recognition	3	13	178
Yeast Cell Cycle	5	17	384
Iris	3	34	351
SAT	6	36	2000

2.4 System Scalability

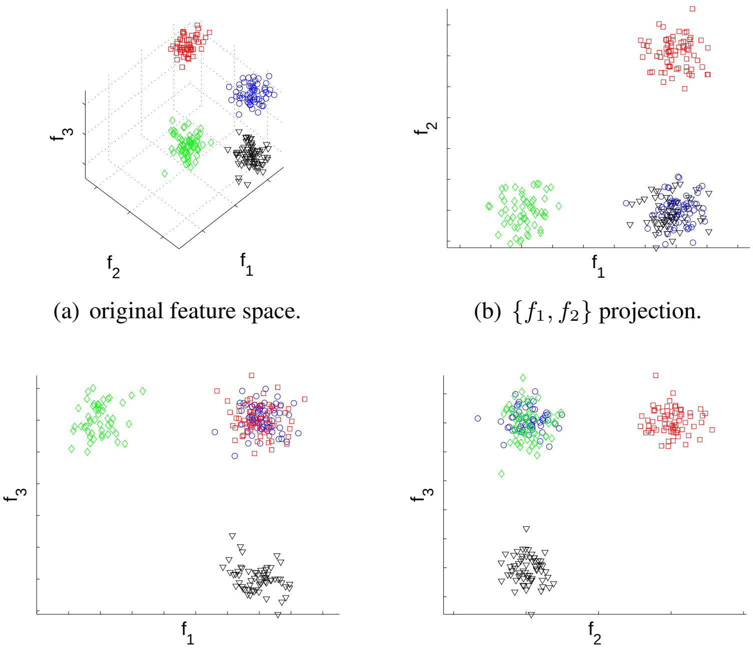
Scalability of the pattern recognition system is another benefit of class pairwise feature selection. In benchmark scenarios or non/slow evolutive domains, feature selection can be performed only once since the need to adapt the resulting feature space to changes in the initial domain is not likely to emerge. However real world domains are evolutive by nature, and sometimes highly mutable in terms of class set expansion or reduction.

If feature selection is performed using the global multi-class scope of classes W, when a new class w_{k+1} is added the system must be retrained on the new global multi-class scope $W \cup \{w_{k+1}\}$. Without exploring the state space for the new set of classes, there is no guarantee that the subspace of features learned for W will hold, nor that it remains the one with better performance for pattern classification. Some of the discarded features can be important, and some of the previously selected features may no longer be useful. This is what forces the feature subspace used by the pattern classification system to be retrained as new classes are added.

In class pairwise feature selection adaptation of the pattern recognition system is incremental. Since the search is performed using only pairs of classes, the adaptation to a new class w_{k+1} consists on determining the individual subspaces $F^*_{\{w_a,w_{k+1}\}}$ $(1 \leq a \leq k)$ that best describe all the pairs containing the new and each of the already existing classes.

3 Experimental Results

As listed in Algorithm 1 the class pairwise search consists in determining for each pair of classes which sub-



(a) original feature space.

(b) $\{f_1, f_2\}$ projection.

(c) $\{f_1, f_3\}$ projection.

(d) $\{f_2, f_3\}$ projection.

Figure 1: Illustration of a four classes and three features gaussian problem in which benefits can arise by using class pairwise feature selection. Class labeling is as follows: $\circ - w_1$; $\Box - w_2$; $\Diamond - w_3$; $\nabla - w_4$

space of features performs best. There are several algorithms which can be used for feature subspace selection [12]; for simplicity we used wrapper sequential forward search. This method starts from an initially empty space $F^* = []$; in each iteration F^* is updated by selecting the subspace with better criteria value J^*, from all subspaces formed by the currently best subspace F^*, and each of remaining available features, until convergence. The criteria in a wrapper framework is the predictive accuracy of a classifier, for which we used the k-nearest neighbor, and the naive Bayes, being the classifier used during feature selection the same used for classification.

We evaluated the class pairwise feature selection algorithm on two real world sets of small size ($n < 20$ features), and two of medium size ($20 \leq n < 50$ features) characterized in Table 3. All sets were preprocessed to ensure the removal of patterns containing missing values, and nominal values were converted to discrete numerical values. After preprocessing the sets were split into a first set of training data and a second set of test data, each set exclusively containing 50% of the available patterns (randomly selected). In the feature selection phase we used the first set to select the individual feature subspaces for each class pair, and the second set to evaluate the classifier performance. In the classification phase we used the second set to train each individual classifier on its computed subspace, and the first set to evaluate the overall performance of the method. It is important to enhance that this way, although only two sets are used, the classifier performance is always assessed with data which is unknown to the classifier, that is, not using during its training. Since differentiated feature subspaces are produced, the assigned class depends on individual decisions computed for each class pair. For this purpose, we employ majority voting as described in section 2.2, in which individual classifiers vote the class to which a pattern is most likely to belong. The assigned class is the one the majority of individual classifiers vote for.

Tables 2 and 3 list the mean classification error, and mean subspace size, respectively for global multi-class scope and class pairwise feature selection, computed over 50 runs using the procedure as described before. Bracketed values are the corresponding standard deviations.

From the results it is easily noticeable that class pairwise feature selection consistently reduces the number of necessary features by a factor of roughly $2 : 1$ when compared to the results obtained by using the global multi-class scope for all datasets, as we can see in Table 3. This confirms what was stated in Section 2.3 regarding the achievement of simpler models. As we can see in Table 2, results revealed improvements of the classification performance for the

Table 2: Mean classification error.

global multi-class feature selection				
	wine	**yeast**	**iris**	**sat**
1-nn	10.00 (6.29)	29.20 (3.02)	6.21 (3.16)	13.00 (0.73)
3-nn	8.71 (3.52)	28.00 (2.46)	4.99 (1.92)	13.10 (0.89)
bayes	3.69 (2.59)	27.80 (2.65)	3.92 (2.13)	14.80 (0.96)
pairwise multi-class feature selection				
	wine	**yeast**	**iris**	**sat**
1-nn	9.24 (2.73)	27.50 (2.45)	6.35 (3.15)	12.70 (0.82)
3-nn	7.87 (2.25)	25.90 (2.85)	4.96 (1.83)	13.20 (0.78)
bayes	5.80 (3.07)	26.70 (2.52)	3.97 (2.01)	15.00 (0.86)

Table 3: Mean subspace size.

global multi-class feature selection				
	wine	**yeast**	**iris**	**sat**
1-nn	4.46 (1.40)	10.20 (2.70)	2.20 (0.81)	22.70 (4.66)
3-nn	4.70 (1.93)	10.20 (2.73)	2.18 (1.00)	19.10 (4.33)
bayes	6.26 (1.94)	8.02 (2.07)	2.48 (0.89)	10.10 (2.89)
pairwise multi-class feature selection				
	wine	**yeast**	**iris**	**sat**
1-nn	2.23 (0.71)	5.25 (0.76)	1.40 (0.27)	9.42 (1.09)
3-nn	1.89 (0.49)	4.68 (0.77)	1.39 (0.34)	8.98 (1.14)
bayes	3.01 (0.87)	4.32 (0.70)	1.49 (0.31)	7.42 (0.99)

Yeast Cell Cycle data set for all classifiers, and the Wine data set for the k-nearest neighbor type classifiers. In the remaining cases slightly worse classification performance was obtained, although within the confidence interval bounds. Naive Bayes, when used as classifier on the Wine data set, was the case where results have degraded more. Also interesting to analyze is the refinement of the standard deviation obtained for both the mean classification error, and mean feature subspace size: class pairwise feature selection leads in general to reduced variability.

4 Conclusions

We described the application of class pairwise testing and decision to the problem of feature selection, pointing the main benefits and drawbacks of such approach.

As seen, the main benefits arise from the fact that simpler models can be achieved through this method (due to the reduction of the number of features), and easier adaptation mechanisms can be employed when new classes are added, by simple training of pairwise associations with the new classes, as opposed to the global multi-class approach which requires total retraining of the pattern recognition system involving all the class.

Experimental results not only proved that class pairwise feature selection decreases the number of necessary features, thus permitting more amenable models, with additional insight into the differentiating features, while achieving comparable or even improved classification accuracy.

Acknowledgments

This research was partially funded by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC). The Iris and Wine Recognition databases were obtained from the University of Wisconsin Hospitals, Madison [15].

References:

[1] A. Blum and P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence* 97, 1997, pp. 245–271.

[2] R. Caruana and V. R. de Sa, Benefitting from the variables that variable selection discards, *Journal of Machine Learning Research* 3, 2003, pp. 1245–1264.

[3] T. M. Cover and J. M. Campenhout, On the possible orderings in the measurement selection problem, *IEEE Transactions on Systems, Man and Cybernetics* 9, 1977, pp. 657–661.

[4] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*, Prentice Hall, 1982.

[5] R. O. Duda, P. E. Hart, and D. G. Stork , *Pattern classification, 2nd edition*, John Wiley & Sons Inc, 2001.

[6] G. Forman, An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research* 3, 2003, pp. 1289–1306.

[7] G. Forman, A pitfall and solution in multi-class feature selection for text classification, *ICML '04: Proceedings of the 21st International Conference on Machine Learning*, 2004, pp. 38–46.

[8] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *Journal of Machine Learning Research* 3, 2003, pp. 1157–1182.

[9] M. A. Hall, *Correlation-based feature selection for machine learning*, PhD thesis, Waikato University, 1998.

[10] D. V. Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, parameter estimation and state estimation - an engineering approach using MATLAB*, John Wiley & Sons Inc, 2004.

[11] R. Kohavi and G. H. John, Wrappers for feature subset selection, *Artificial Intelligence* 97, 1997, pp. 273–324.

[12] M. Kudo and J. Sklansky, Comparison of algorithms that select features for pattern classifiers, *Pattern Recognition* 33, 2000, pp. 25–41.

[13] L. Lam and S. Y. Suen, Application of majority voting to pattern recognition: An analysis of its behavior and performance, *IEEE Transactions on Systems, Man, and Cybernetics* 27, 1997, pp. 553–568.

[14] L. C. Molina, L. Belanche, and A. Nebot, Feature selection algorithms: A survey and experimental evaluation, *lsi technical report lsi-02-62-r*, 2002.

[15] D. J. Newman, D. S. Hettich, C. L. Blake, and C. J. Merz, UCI repository of machine learning databases, 1998.

[16] D. M. J. Tax, *One-class classification*, PhD thesis, Delft University of Technology, 2001.

[17] D. M. J. Tax and R. P. W. Duin, Using two-class classifiers for multiclass classification, *Proceedings of the International Conference on Pattern Recognition*, 2002.

[18] D. M. J. Tax and K. R. Müller, Feature extraction for one-class classification, *Proceedings of the ICANN/ICONIP 2003* 2714, 2003, pp. 342–349.

[19] S. Theodoridis and K. Koutroumbas, *Patern recognition*, Academic Press, 1999.

[20] K. Torkkola, Feature extraction by non-parametric mutual information maximization, *Journal of Machine Learning Research* 3, 2003, pp. 1415–1438.