Algorithms for Feature Selection: An Evaluation

Douglas Zongker
Department of Computer Science
Michigan State University
East Lansing, Michigan, USA
zongker@cps.msu.edu

Anil Jain
Department of Computer Science
Michigan State University
East Lansing, Michigan, USA
jain@cps.msu.edu

Abstract

A large number of algorithms have been proposed for doing feature subset selection. The goal of this paper is to evaluate the quality of feature subsets generated by the various algorithms, and also compare their computational requirements. Our results show that the sequential forward floating selection (SFFS) algorithm, proposed by Pudil et al.*, dominates the other algorithms tested. This paper also illustrates the dangers of using feature selection in small sample size situations. It gives the results of applying feature selection to land use classification of SAR satellite images using four different texture models. Pooling features derived from different texture models, followed by a feature selection results in a substantial improvement in the classification accuracy. Application of feature selection to classification of handprinted characters illustrates the value of feature selection in reducing the number of features needed for classifier design.*

1. Introduction

The problem of feature selection is to take a set of candidate features and select a subset that performs the best under some classification system. This procedure can reduce not only the cost of recognition by reducing the number of features that need to be collected, but in some cases it can also provide better classification accuracy due to finite sample size effects [2]. There has been a resurgence of interest in applying feature selection methods due to the large numbers of features encountered in the following types of problems: (1) Applications where data taken by multiple sensors are fused. (2) Integration of multiple models, where all the parameters from the different models can be used for classification; and (3) Data mining applications, where the goal is to recover the hidden relationships among the features.

The goal of this paper is to evaluate the performance of various feature selection methods on some synthetic data sets. Several well-known and some recently proposed feature selection algorithms have been implemented and tested. Based on these results, the sequential forward floating selection (SFFS) method introduced in [7] has been found to be extremely powerful. This method has been applied to large datasets in two different application domains. Experimental results indicate that feature selection can not only eliminate a large number of redundant features, but also avoid the curse of dimensionality.

2. Feature Selection Algorithms

Let Y be the original set of features, with cardinality n. Let d represent the desired number of features in the selected subset X, $X \subseteq Y$. Let the feature selection criterion function for the set X be represented by $J(X)$. Without any loss of generality, let us consider a higher value of J to indicate a better feature subset. Formally, the problem of feature selection is to find a subset $X \subseteq Y$ such that $|X| = d$ and

$$J(X) = \max_{Z \subseteq Y, |Z|=d} J(Z).$$

A taxonomy of all the available feature selection algorithms into broad categories is presented in Figure 1. We first divide methods into those based on statistical pattern recognition (SPR) techniques, and those using artificial neural networks (ANN). The SPR category is then split into those guaranteed to find the optimal solution and those that may result in a suboptimal feature set. The suboptimal methods are further divided into those that store just one "current" feature subset and make modifications to it, versus those that maintain a population of subsets. Another distinction is made between algorithms that are deterministic, producing the same subset on a given problem every time, and those that have a random element which could produce different subsets on every run. Some representative feature selection algorithms are listed beneath each leaf node in the tree.

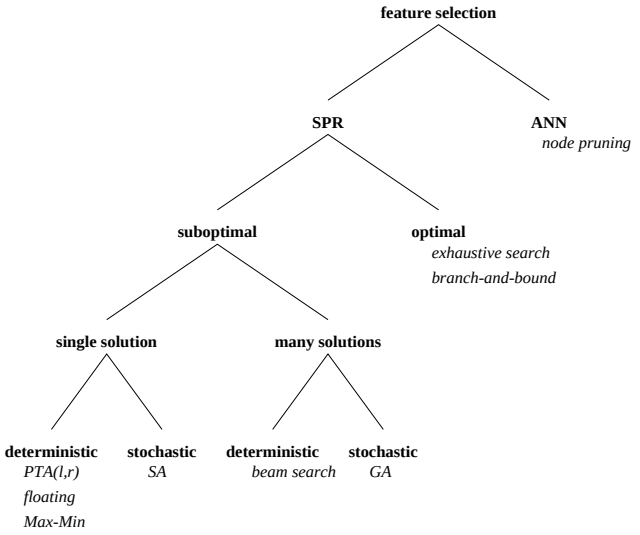

Figure 1. A taxonomy of feature selection algorithms.

The first group of methods begin with a single solution (a feature subset) and iteratively add or remove features until some termination criterion is met. These "sequential" methods can be divided into two categories, those that start with the empty set and add features (the "bottom-up," or "forward" methods) and those that start with the full set and delete features (the "top-down," or "backward" methods). We have implemented the following well-known sequential algorithms:

SFS	Sequential forward selection
SBS	Sequential backward selection
GSFS($\cdot$)	Generalized sequential forward selection
GSBS($\cdot$)	Generalized sequential backward selection
PTA(l, r)	Plus l-take away r
SFFS	Sequential forward floating selection
SFBS	Sequential backward floating selection
MM	Max-Min search

The two "floating" selection methods are described in Pudil *et al.* [7]. All of the other methods are given in detail in Kittler [3]. Siedlecki and Sklansky [9] introduced the use of genetic algorithms (GA) for feature selection, a technique that is also evaluated in [1]. The branch-and-bound feature selection algorithm, proposed by Narendra and Fukunaga [6], can be used to find the optimal subset of features much more quickly than exhaustive search. One drawback is that the branch-and-bound procedure requires the feature selection criterion function to be monotone, i.e. the addition of new features to a feature subset can never decrease the value of the criterion function. We know from the curse of dimensionality phenomenon that in small sample size situations this may not be true.

Mao *et al.* [4] use a multilayer feedforward network with a backpropagation learning algorithm for pattern classification. They define a *node saliency* measure and present an algorithm for pruning the least salient nodes to reduce the complexity of the network after it has been trained. The pruning of input nodes is equivalent to removing the corresponding features from the feature set. The node-pruning method simultaneously develops both the optimal feature set and the optimum classifier.

3. Experimental Results

We have compared different feature selection algorithms in terms of classification error and run time on a 20-dimensional, 2-class Gaussian data set which was used by Pudil *et al.* [7]. Our criterion function for assessing the "goodness" of a feature subset was the Mahalanobis distance between the class means—the larger the Mahalanobis distance, the better the feature subset. Maximum likelihood estimates of the covariance matrix and mean vectors were computed from the data. A total of thirteen feature selection algorithms, listed in Table 1, were evaluated and compared. Execution times reported are processor ticks (0.01 second)

SFS	SBS	GSFS(2)
GSBS(2)	GSFS(3)	GSBS(3)
SFFS	SFBS	PTA$((1),(2))$
PTA$((1),(3))$	PTA$((2),(3))$	
Branch-and-Bound	Max-Min	

Table 1. Feature selection algorithms used in experimental evaluation.

spent in user space on a SUN SPARCserver 1000. Ten randomly generated data sets, each with 1,000 patterns per class were tested and the averages of the runs are reported. Figure 2 shows the results. The solid line in each figure indicates the optimal result for each target feature subset of size d, obtained using the branch-and-bound method.

The following conclusions can be drawn based on these empirical results:

- The Max-Min algorithm, while very fast, gives poor results compared to the other algorithms.
- The SFS and SBS algorithms have comparable performance, but show *nesting* problems. (For instance, the optimal 3-subset is not contained in the optimal 4-subset.) The forward method is faster than its backward counterpart, as expected. This is also true of the generalized methods (GSFS and GSBS).
- The floating methods (SFFS, SFBS) show results comparable to the branch-and-bound algorithm and are, for the most part, faster than it.

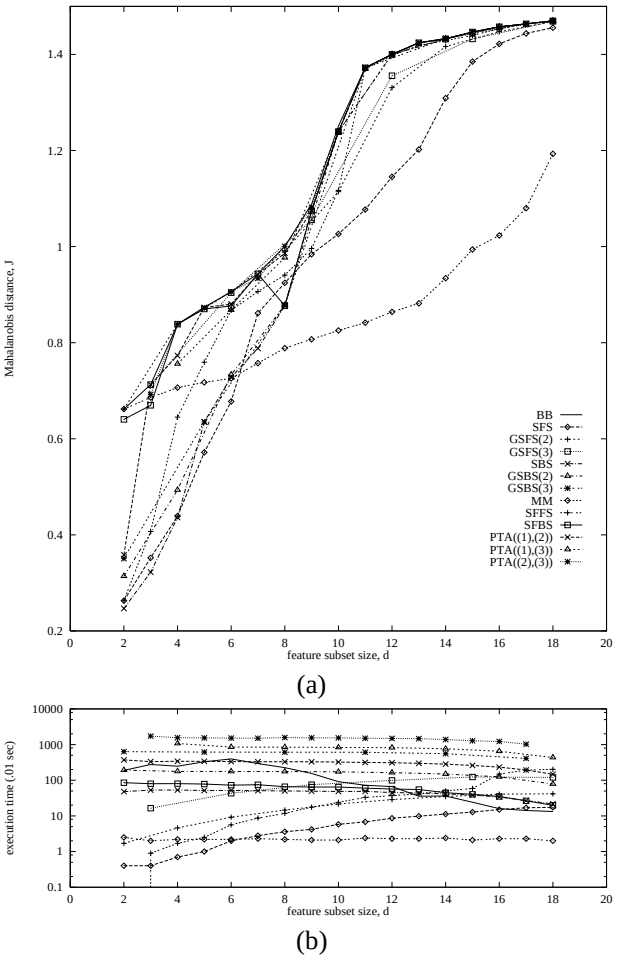

Figure 2. Performance of some feature selection algorithms: (a) criterion value, (b) execution time.

- The PTA$((l),(r))$ methods, while generally giving near-optimal performance, are far slower than the branch-and-bound method.

  Overall, the floating methods perform better than their non-floating counterparts, giving near-optimal results with reasonable execution times.

4. Effect of Training Set Size

How reliable are the feature selection results in the presence of small amounts of training data? In the case where Mahalanobis distance is used as the criterion, the error arising from estimating the covariance matrix can lead the feature selection process astray, producing inferior results (relative to the true distributions) on independent test data even if the selected subset is optimal for the given training data [8]. This phenomenon, which is related to the *curse of dimensionality*, is illustrated by running the feature selection algorithm on varying amounts of training data drawn from known distributions. Trunk [11] used the following simple example to illustrate the curse of dimensionality. The two multivariate Gaussian class-conditional densities, are given below:

$$p(\mathbf{x}|\omega_1) \sim N(\boldsymbol{\mu}, I) \qquad p(\mathbf{x}|\omega_2) \sim N(-\boldsymbol{\mu}, I) \qquad (1)$$

where

$$\boldsymbol{\mu} = \left[\begin{array}{cccc} \frac{1}{\sqrt{1}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \cdots \end{array} \right]^t . \qquad (2)$$

and I denotes the identity matrix. Note that for these class-conditional densities, the optimal d-feature subset is the first d features.

Various size data sets, ranging from 10 to 5,000 training patterns per class, were generated from the two 20-dimensional distributions (equations (1) and (2)). For each training set size, five data sets were generated, and the results averaged. The *quality* of each selected feature subset was calculated by taking the number of commonalities in the resulting subset when compared with the optimal subset of the true distribution: features that were included in both sets, and features that were excluded from both sets. This count was divided by the number of dimensions, and that value was averaged over values of d from 1 to 19 inclusive to give a final quality value for the set. Note that this value is *not* a measure of the classification error, but a measure of the difference between the subset produced by a feature selection method and the ideal feature subset. The average quality for each training set size for the branch and bound and SFS methods is shown in Figure 3.

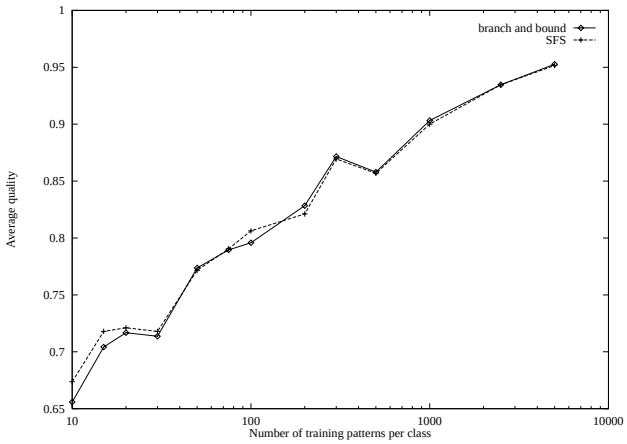

Figure 3. Quality of selected feature subsets as a function of the size of training data.

For this dataset, since the features are all independent with identical variance, only the difference in means along a feature axis is significant. Therefore, any feature selection

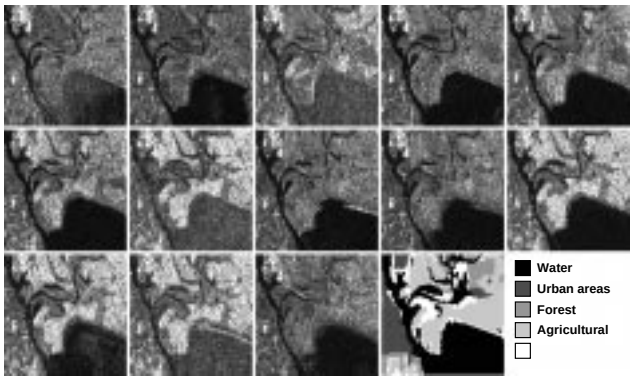

Figure 4. Sample SAR images.

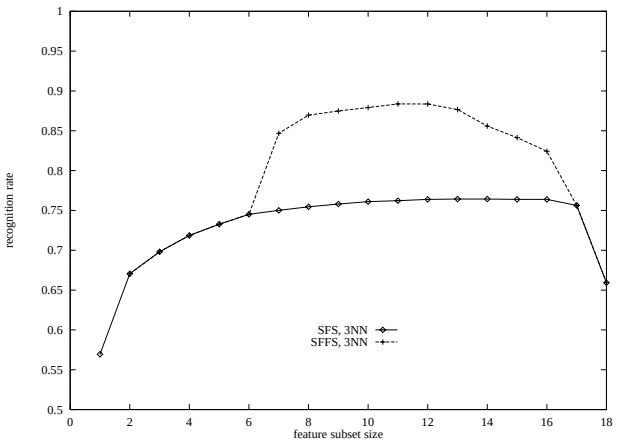

Figure 5. Recognition rates of SFS and SFFS methods on texture feature data.

algorithm should perform well on such a simple classification problem. Indeed, the curve for the SFS algorithm closely matches that of the branch-and-bound algorithm. Note that as expected, the quality of the feature subset for small training sets is low, but improves as the training set size increases.

5. Selection of Texture Features

We have applied various feature selection algorithms to select the best subset of texture features for the problem of land use classification using SAR (Synthetic Aperture Radar) images (see Figure 4). Solberg and Jain [10] have used texture features computed from SAR images to classify each pixel. A total of 18 features per pattern (pixel) were computed from four different texture models: local statistics (5 features), gray level co-ocurrence matrices (6 features), fractal features (2 features), and a lognormal random field model (5 features). Our goal is to determine whether the classification error can be reduced by applying feature selection to this set of 18 features from four different texture models. A similar feature selection study for 2D shape features was reported by You and Jain [12].

We report results for one SAR image (the October 17 image from [10]), containing approximately 22,000 pixels. This data was split evenly to form independent training and test sets. The recognition rate of the 3NN classifier is used as the feature selection criterion. Based on its consistently high performance for the synthetic data in Section 3, we chose to apply the SFFS method to the texture data set. The results of these runs are shown in Figure 5.

The best recognition rate obtained by SFFS was 88.4%, with an 11-feature subset. Notice that the recognition rate does not monotonically increase as the number of features is increased. The feature selection process is not just using the features derived from a single texture model but is utilizing features from different models to provide a better performance. For instance, in every run, the five-feature subset selected contained features from at least three different texture models. The best individual texture model for this data set was the random field model with a classification accuracy of 68.8% [10]. Pooling features from four different models and then applying feature selection increased the classification accuracy.

6. Selection of Handprinted Character Features

We have also applied feature selection to aid the classification of a subset of the NIST SD-3 handprinted character set. The classes correspond to the 26 lowercase letters (see Figure 6). A total of 2,439 patterns are used for training and 1,525 are used for testing. The features used are the 88 contour direction features from Mohiuddin and Mao [5].

We again applied the SFFS algorithm to this data set. While our best recognition rates (88.7% with 51 features using 1NN, 89.6% with 71 features using 3NN) did not reach those reported by Mohiuddin and Mao [5] using different, non-nearest neighbor-based classifier methods, this data set illustrates one of the major advantages of performing feature selection—it can dramatically reduce the number of features with only a small drop in recognition rate. For instance, the 1NN recognition rate reaches 87.7% accuracy using only 36 of the 88 features. Over half the features can be culled from the data set for only a 1% drop in the recognition rate!

7. Summary

This paper illustrates the merits of various methods of feature selection. In particular, the practicality of finding the optimal subsets in feature spaces of moderately high dimension using the branch-and-bound algorithm (where the

Figure 6. Sample handprinted characters.

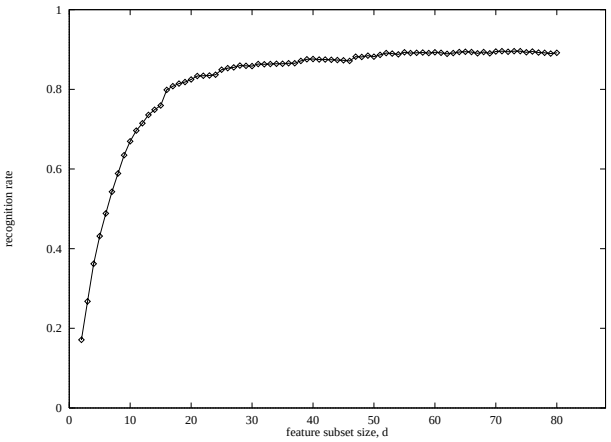

Figure 7. Recognition rates as a function of feature subset size for character data.

monotonicity requirement for the criterion function is satisfied), and the quality of the results given by the floating search methods are illustrated. The floating search methods show a great promise of being useful in situations where the branch-and-bound method can not be used, due to either the nonmonotonicity of the feature selection criterion or computational reasons.

We also show the pitfalls of using feature selection with limited training data. By using feature selection on a classification problem with known distributions and comparing the selected subsets (under finite sample size) with the true ideal subsets, the quality of the selected subset can be quantified. Our experiments with the Trunk distributions show the problems associated with sparse data in a high dimensional space. Results on texture data show that feature selection is useful in utilizing feature derived from different texture models while at the same time avoiding the curse of dimensionality. Experiments on handprinted character data illustrate that a large number of feature can be eliminated without a significant loss of classification performance.

References

[1] F. Ferri, P. Pudil, M. Hatef, and J. Kittler. Comparative study of techniques for large-scale feature selection. In E. Gelsma and L. Kamal, editors, *Pattern Recognition in Practice IV*, pages 403–413. Elsevier Science B.V., 1994.

[2] A. K. Jain and B. Chandrasekaran. Dimensionality and sample size considerations. In P. R. Krishnaiah and L. N. Kanal, editors, *Pattern Recognition Practice*, volume 2, chapter 39, pages 835–855. North-Holland, 1982.

[3] J. Kittler. Feature set search algorithms. In C. H. Chen, editor, *Pattern Recognition and Signal Processing*, pages 41–60. Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, 1978.

[4] J. Mao, K. Mohiuddin, and A. K. Jain. Parsimonious network design and feature selection through node pruning. In *Proceedings of 12th ICPR, Jerusalem*, pages 622–624, 1994.

[5] K. M. Mohiuddin and J. Mao. A comparative study of different classifiers for handprinted character recognition. In E. S. Gelsema and L. N. Kanal, editors, *Pattern Recognition in Practice IV: Multiple Paradigms, Comparative Studies and Hybrid Systems*, pages 437–448. Elsevier Science B.V., 1994.

[6] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, C-26(9):917–922, September 1977.

[7] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, November 1994.

[8] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, March 1991.

[9] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, 10:335–347, November 1989.

[10] A. H. S. Solberg and A. K. Jain. A study of the invariance properties of textural features in SAR images. In *Proc. IGARS Conference*, pages 670–672, Florence, Italy, July 1995.

[11] G. V. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(3):306–307, July 1979.

[12] Z. You and A. K. Jain. Performance evaluation of shape matching via chord length distribution. *Computer Vision, Graphics and Image Processing*, 28:185–198, 1984.