Intelligent Information Systems VII
Proceedings of the Workshop held in
Malbork, Poland, June 15-19 , 1998

Comparative study of feature subset selection techniques for machine learning tasks

Jacek Jelonek, Krzysztof Krawiec, Jerzy Stefanowski

Institute of Computing Science, Poznan University of Technology,
Piotrowo 3A, 60-965 Poznan, Poland
jacek.jelonek@cs.put.poznan.pl
krzysztof.krawiec@cs.put.poznan.pl
jerzy.stefanowski@cs.put.poznan.pl

Abstract. In this paper, we discuss problems of using feature selection techniques to improve classification performance of supervised learning algorithms. We consider two approaches for selecting relevant features, i.e. the Forward Beam Feature Selection algorithm and the n^2-classifier, both equipped with the wrapper model. In the computational experiment we compared them against the classifier working with the complete set of features. The obtained results strongly indicate the usefulness of these techniques.

Keywords: classification systems, feature subset selection, multiclass learning problems

1 Introduction

One of the typical machine learning tasks is to create classification systems on the basis of the set of learning examples, i.e. objects described by their features and assigned to some predefined decision classes. Selecting relevant features is the crucial issue for this task. Ideally, the learning algorithm should use only such features that lead to the best classification performance (usually evaluated by prediction accuracy). Unfortunately, most of the practical learning algorithms perform rather poorly when faced with too many irrelevant features (for more details see [13, 14, 1, 15]). Thus additional methods are necessary to reduce the number of features so as to improve the classification performance.

The feature selection has been already studied in statistics and pattern recognition for at least few decades (see e.g. [5]). However, our interest in this field is coming rather from an integration of Artificial Intelligence tools with data analysis. In particular, in last years we worked on the approach based on the rough set theory (originally proposed by Pawlak [18]). This approach can be used to discover redundant features especially for problems where a limited discernibility of objects by their features which causes their imprecise description. We applied it to many practical problems (see e.g. [19, 21, 8]). The current study is oriented to other methods and follows an increasing interest for the topic of feature selection in machine learning field (see, e.g. discussion in [15] or the comprehensive overview of various methods given in [4]).

Feature selection can be generally considered as a search problem, where each state of the search space represents a subset of features. In the following paper, we employ so called *wrapper model* [13] as a general framework for the search process. In this model, the feature subset selection algorithm conducts a search for the good subset using learning algorithm itself as a part of the evaluation function. Thus, features are selected depending not only on the characteristics of data but also on the learning algorithm used finally to evaluate them. Empirical results [13, 2, 10] have confirmed that this approach performs quite efficiently.

In that framework, the authors propose two new elements. The first contribution consists in proposing a *Forward Beam Feature Selection* as a basic technique for heuristic searching through the space of features. It starts with an empty set of features and sequentially adds the most promising ones. Its specificity consists in considering parallely several subsets and choosing finally the best one. The second contribution refers to proposing a specialized approach, called n^2-*classifier* for solving multi-class learning problems (i.e. learning problems with $n > 2$ decision classes). The n^2-classifier is a multiple classification model specialized to discriminate pairs of decision classes. In previous study [11] we checked that classification performance of this approach is often better than the accuracy of a single classifier. That experiment was performed using the complete set of all features. Now, we extend this concept by introducing a forward feature selection algorithm inside each base classifier to obtain as a result an input feature space dedicated for each pair of decision classes. We hypothesize that both these approaches should lead to higher prediction accuracy than using a single learning algorithm for all features.

The aim of this paper is to compare the performance of the two new proposed techniques of feature selection on the typical machine learning data sets. In particular, we consider data sets defined with more than two decision classes. As a classifier we use an instance based learning algorithm because it does not have an inherent capability of reducing any input features.

The paper is organized as follows. Next section groups review of used methodology. Section 3 presents results of a computation experiment. Discussion of results and conclusions are given in the final section.

2 Methodology

2.1 Wrapper approach

In this study we intend to consider the methodologies for finding the smallest subset of features leading to a higher classification accuracy than the set of all features. As it is discussed in literature, the feature selection can be reduced to a search problem where one selects a good feature subset based on a selected evaluation measure. Each state in the search space represents a subset of possible features. It is usually a partially ordered space where each state has a child differing on exactly one feature. Following the idea presented by Kohavi [13], to design a feature selection algorithm one has to define three following components: *search algorithm* (technique that looks through the space of feature subsets), *evaluation function* (used to evaluate examined subsets of features), and *classifier* (a learning algorithm that uses the final subset of features). These elements can be integrated in two ways. They are called *filter* and *wrapper* approaches, respectively [13]. In the filter approach, features are selected as a pre-processing step before a classifier is used. Features are selected (i.e. filtered) basing on properties of data itself independently of the learning algorithm used as a classifier. In the wrapper approach, the search algorithm conducts a search for a good subset of features using the classifier itself as the evaluation function. Evaluation is usually done by estimating classification accuracy. It is claimed [13, 1] that the wrapper model is superior because it takes into account the biases of the learning algorithm in order to get a feature subset. It should result in a better estimate of accuracy on unseen data than the evaluation function in the filter model whose bias differs from that of the classifier.

The major limitation of the wrapper approach is the computational cost resulting from using the learning algorithm to check each feature subset. This leads to choosing a simple and relatively fast learning algorithm. For these reasons it is not reasonable, to use e.g. artificial neural network that requires tuning many parameters. Other reasons for choosing the classifier are connected with an assumption that the classifier should not have an internal ability to select features while performing the learning process, i.e. it should neither remove redundant features itself nor modify them (for example by weighting).

This is why we have decided to choose a non-parametric instance-based learning algorithm [2] inside all models discussed in the further section of this study. We have implemented a *Nearest Neighbor* algorithm (1-NN) with euclidean distance measure. In this approach all training instances are stored. The algorithm simply finds an instance nearest to the new one, retrieves its class and predicts this as the class of the new instance. It should be also noticed that such an algorithm has been already successfully used in the wrapper model context (see e.g. [2, 10]).

The next issue concerns constructing the search algorithm. Since the exhaustive search is of exponential complexity, it is more efficient to perform the heuristic search. Commonly employed algorithms are *backward elimination* and *forward feature selection*. Former starts with all attributes and successively removes the one that its elimination improves performance. The second starts with an empty set of features and successively

adds the one with the best performance. There are also known more sophisticated variants of both approaches that combine performing either adding or removing operations as to get the best subset see, e.g. SFFS procedure described in [20]. There are also other variants as e.g. BEAM BSS [1] that starts with random looking for first initial subsets and then performing backward moves. One should remember that the best feature subsets can be improved if we make assumptions on evaluation function. The most common assumption (often used in statistical approaches) is monotonically what leads to use e.g. branch-and-bound techniques. However, the monotonicity assumption do not hold for many machine learning algorithm and classification accuracy, hence the mentioned selection techniques are heuristics [13].

In the proposed approach we decided to follow the forward selection way. This motivation is somehow related with our previous successful experience [9] in medical applications.

2.2 FBFS algorithm

The forward selection technique in the simplest version consists in a choice of the best single feature (according to the improvement of evaluation function) to be added in each iteration. The process of adding features is stopped when no improvement of evaluation function is observed. Although the simplest version of the greedy choice is locally optimal (in the sense of the chosen evaluation function) for separately added features, it may not be globally optimal for the final subset of features. For these reasons, in [9] we decided to extend the basic technique by introducing a *beam search* strategy. The beam search consists in maintaining a fixed-size collection of the best performing subsets of features. This collection is updated in each iteration of the forward selection. Thus, the algorithm does not consider one way of adding features only but also analyses a limited number of alternative possibilities. Additionally, in our implementation we allow the algorithm to perform trial next moves in case of admissible small worsening of the evaluation function. It is done to check possible other good solution in the neighborhood of the currently found subset. We have already checked that the beam search should decrease the greedy character of a simply best-first strategy and lead to better performance of the classifier based on the chosen features (for more details, see [9])

2.3 The n^2-classifier

The n^2-classifier belongs to the group of multiple classification models dedicated to solve multiclass learning problems (i.e. with n decision classes where $n > 2$). The n^2-classifier was defined by us in [10]. Below we give its short description.

It is composed of $(n^2 - n)/2$ base binary classifiers. The main principle of the n^2-classifier is the discrimination of each pair of the classes: (i, j), $i, j \in [1..n], i \neq j$, by an independent binary classifier C_{ij}. Each base binary classifier C_{ij} corresponds to a pair of two classes i and j only. Therefore, the specificity of the training of each base classifier C_{ij} consists in presenting to it a subset of the entire data set that contains only examples coming from classes i and j. The classifier C_{ij} yields a binary classification indicating

whether a new example, $\mathbf{x}$, belongs to class i or to class j. Let us denote by $C_{ij}(\mathbf{x})$ the classification of an example $\mathbf{x}$ by the base classifier C_{ij}. In following description of the n^2-classifier we assume that $C_{ij}(\mathbf{x}) = 1$ means that example $\mathbf{x}$ is classified by C_{ij} to class i, otherwise ($C_{ij}(\mathbf{x}) = 0$) $\mathbf{x}$ is classified to class j.

An algorithm providing final classification assumes that a new example $\mathbf{x}$ is applied to all base classifiers C_{ij}. As a result, their binary predictions $C_{ij}(\mathbf{x})$ are computed. The final classification should be obtained by a proper aggregation of these predictions. The simplest aggregation rule is based on finding a class that wins the most pairwise comparisons. The classification performance of base classifiers could be diversified because they have been trained on different pairs of classes. So, it is necessary to estimate the their credibility. It can be done in several ways. In this study we assume that with each classifier C_{ij} we associate a credibility coefficient P_{ij} defined in following way:

$$P_{ij} = \frac{v_i}{v_i + e_j}$$

where e_j is a number of misclassified examples from class j, and v_i is a number of correctly classified examples from class i. The computation of the credibility coefficients is performed during the learning phase the n^2-classifier.
The whole process of classification of classification of particular example $\mathbf{x}$ can be described by the following steps:

1. Apply $\mathbf{x}$ to all base classifiers C_{ij}, and obtain their predictions $C_{ij}(\mathbf{x})$.
2. For each class i, $i \in [1..n]$, compute weighted sum:

$$S_i = \sum_{j=1, i \neq j}^{n} P_{ij} \cdot C_{ij}(\mathbf{x})$$

3. Return $argmax_i(S_i)$ as a final classification (break ties arbitrary in favor of the class that comes first in the class order).

The quite similar approach was independently introduced by Friedman [6]. Then it was extended and experimentaly studied in [7, 16]. Hastie and Tibshirani called this classification model as *classification by pairwise coupling*. As it has been indicated in experimental studies [6, 7, 11] such integration of binary classifiers performs usually better than the single multiclass classification model.

We observed in our previous experiments [10, 11] that the n^2-classifier based on a classifier with an inherent capability of reducing the influence of irrelevant features (decision trees or neural networks in our study) usually gave higher classification accuracy than a single classification model that did not have such capability (e.g. k-NN). This observation led us to a hypothesis that the use of dedicated subsets of features for discriminating each pair of decision classes could lead to higher accuracy than using all features for classifying all classes at once.

Therefore we extended the n^2-classifier by connecting learning phase of its base binary classifiers with the technique of feature selection. We decided to implement it with the wrapper model where the search was performed by forward selection algorithm. The

binary classifier was the k-NN. Let us stress that practically the wrapper model is now repeated $(n^2 - n)/2$ times in order to get dedicated input feature spaces for all pairs of classes. Then, these sets of features are used to construct the final n^2-classifier.

3 Experiments

The aim of this experimental study is to compare the usefulness of different techniques of feature selection discussed in this paper. The k-NN algorithm was as a base classifier in all experiments. We compare its performance

1. as single classifier used for all features without any reduction,
2. as single classifier used inside wrapper model with Forward Beam Features Selection (FBFS) algorithm,
3. in n^2-classifier with wrapper model and FFS algorithm applied inside its each binary classifier.

In the second case FBFS was implemented with beam consisting of 10 independent paths. In the third case due to computation costs we discard the beam option and stayed with only one path.

3.1 Data sets

All computation experiments have been performed on the typical benchmark machine learning data sets. As we wanted to analyse the performance of n^2-classifier we decided to work multiclass data sets Some characteristics of the employed data sets are summarized in Table 1. Most of them are well known machine learning problems and are available in the Machine Learning Repository at the University of California at Irvine [17]. The Primary Tumor data was provided by M. Skolic and M. Zwitter from University Medical Center, Institute of Oncology, Ljubljana.

Table 1. Data sets used in the experiments

No. Data set	Number of examples	Number of classes	Number of attributes
1. Automobile	159*	6	43
2. Ecoli	336	8	7
3. Glass	214	6	9
4. Iris	159	3	4
5. Meta-data	528	5*	43
6. Primary Tumor	339	21	17
7. Soybean-large	542*	14*	35
8. Vowel	990	11	10
9. Yeast	1484	10	8
10. Zoo	101	7	16

Some of the studied data sets have been slightly modified - what is indicated by asterisks. Modifications concern the choice of decision attributes for two problems, i.e. for *Automobile* data set we have used the first ("symboling") attributes, and the *Meta-data* set is characterized by continuos decision attribute which has been discretized using thresholds: 6, 13, 20 and 50, thus giving five classes. Moreover, as in the experiments, we have decided to omit the possible problem of handling missing values. Some of these modifications consist in removing examples or attributes if they contain too many missing values (*Automobile* and *Soybean* data sets). In the case of the *Meta- data* and the *Primary Tumor*, missing values have been replaced by the most frequent values. We transformed *Automobile* and *Meta-data* sets replacing nominal attributes by binary word code "1-of-n" (where n is number of values of a nominal attribute). Therefore, a number of the attributes increased up to 43 in both cases.

3.2 Results

The classification accuracy was estimated by stratified version of 10-fold cross-validation technique, i.e. the training examples were partitioned into 10 equal-sized blocks with similar class distributions as in the original set. For performing validation techniques inside wrapper approach used for n^2-classifier we observed higher variance of accuracy between folds than we expected comparing to single classifier. Therefore we decided to repeat the 10-cross-validation 10 times when we evaluated the feature subset and then we calculated the average accuracy. This rule was also used to obtain results for single k-NN and the FBFS model. Classification accuracies for all three compared models is presented in table 2. For each average accuracy we present its standard deviation. The improvements of FBFS model and n^2-classifier vs. single k-NN are calculated as difference of average accuracies presented together with confidence intervals. This was calculated using a t-test for two-paired difference of means, assuming the confidence level equal to 0.95. An asterik indicates that the difference for compared classifiers and given data sets is not statistically significant. All other differences are significant.

Table 2. Classification performance of compared classifiers and feature selection approaches

No.	Name of data set	Accuracy k-NN (%)	Accuracy FBFS (%)	Accuracy n^2 (%)	Improvement FBFS (%)	Improvement n^2 (%)
1.	Automobile	77.7 ± 0.9	92.5 ± 1.3	94.8 ± 1.7	14.8 ± 1.5	17.1 ± 1.3
2.	Ecoli	81.3 ± 0.5	81.3 ± 0.5	87.1 ± 1.3	0.0 N/A	5.8 ± 1.0
3.	Glass	68.8 ± 0.8	78.5 ± 0.7	82.4 ± 1.6	9.7 ± 0.8	13.6 ± 1.1
4.	Iris	93.9 ± 0.5	95.8 ± 1.0	96.1 ± 0.7	1.9 ± 0.9	2.2 ± 0.8
5.	Meta-data	40.6 ± 1.6	55.8 ± 1.4	60.1 ± 2.9	15.2 ± 1.2	19.5 ± 1.8
6.	Primary Tumor	33.4 ± 1.2	38.1 ± 1.5	43.8 ± 3.3	4.7 ± 0.9	10.4 ± 2.4
7.	Soybean-large	89.9 ± 0.4	93.7 ± 0.6	95.0 ± 1.2	3.8 ± 0.5	5.1 ± 0.9
8.	Vowel	98.9 ± 0.2	99.2 ± 0.2	99.4 ± 0.2	0.3 ± 0.2	0.5 ± 0.1
9.	Yeast	52.8 ± 0.7	52.8 ± 0.7	56.1 ± 1.1	0.0 N/A	3.3 ± 0.6
10.	Zoo	96.9 ± 0.6	98.0 ± 0.0	97.0 ± 0.0	1.1 ± 0.4	0.1* ± 0.4

4 Discussions and final remarks

The paper addresses problem of feature selection in supervised learning tasks. We considered two approaches for selecting relevant feature that improve predictive accuracy, i.e. forward beam feature selection algorithm and the n^2-classifier model for solving multiclass learning problems. We explained motivations for using these approaches together with the wrapper model. Then we compared them against the classifier using the non-reduced set of features.

Let us summarize results obtained in this comparative study. In a case of applying wrapper model with FBFS algorithm we observed statistically significant improvements of classification accuracy in 8 of 10 data sets while in two remaining ones there were no difference. Similarly, for the n^2-classifier approach the increase of classification accuracy was noticed in 9 of 10 data sets. For the *Zoo* data the improvement was non-significant.

Generally, the n^2-classifier performed the best in all experiments. It somehow confirms our initial hypothesis that creating of feature subspaces dedicated for discrimination of pairs of classes is more promising than using the same set of features for classifying all decision classes. Let us remark that in our previous experiment [11] with using the n^2-classifier built on k-NN without any feature selection we did not obtain many improvements for analysed data sets. On the other hand such significant improvements were previously observed for the n^2-classifier created using decision tree algorithm (that has an ability to select relevant feature while building a tree structure).

All our results strongly indicate that the problem of feature selection is of primary importance for machine learning tasks. It seems especially important in such practical problems where one searches for classification models in large data sets with many feature without filtering them in preprocessing phase e.g. by human experts. A good choice of features may not only improve predictive accuracy but also results in smaller models which are easier to understand and interpret.

In our experiments we did not report the decrease of the number of features. However, such results were described in the context of the wrapper model with feature selection applied to real problem with a large number of features, see e.g. our experiments with histological images [9], Kohavi experiments with decision trees [13], case-based classification of cloud types [1], or our experiments with IBL algorithms [22].

The performed experiments were focused on FBFS and FFS search algorithms. Other search strategies could be also successfully used in this context. For instance, Aha [1] describes good performance of specific versions of backward elimination connected with random looking for starting solutions (so called random BEAM BSS). Stefanowski and Urbaniak already checked that these techniques are quite efficient on typical machine learning data sets [22].

One should notice that the better of forward or backward results should not be taken as the best performance possible from the wrapper model. More comprehensive search strategies could examine a larger part of the search space and might even yield improved performance. Such an attempt was undertaken and the results are reported in [12]. In this approach genetic algorithm is used as feature selection engine inside the n^2-classifier.

In out experiments we verified the usefulness of the n^2-classifier with embedded feature

selection technique. However the idea is general and therefore more complex attribute transformation can be used instead of simple feature selection methods. We expect additional improvement of classification results in case of real life machine learning problems in which domain-dependent constructive induction operators can be applied.

However, the considered approaches for feature selection are more demanding from the computational and memory point of view than simple approach without selecting features. Feature selection inside wrapper model increases computational costs mainly due to the number of features while the n^2-classifier approach is characterized by higher computational complexity due to the number of classes. Therefore it is worth to perform further research on effectiveness of the n^2-classifier approach for larger data sets with many decision classes.

Acknowledgements
The computational experiments have been partially carried out at the Poznan Supercomputing and Networking Center affiliated to the Institute of Bioorganic Chemistry at the Polish Academy of Sciences. The authors wish to acknowledge support from research grant no. 43-295/DS.

References

1. Aha D.W., Bankert R.L. Feature Selection for Case-Based Classification of Cloud Types: An Empirical Comparison. In Proceedings of the AAAI-94 Workshop on Case-Based Reasoning. Seatle, AAAI Press, 1994, 106-112.
2. Aha D.W., Kibler E., Albert M.K.: Instance-based learning algorithms. Machine Learning, 6, 1991 37-66.
3. Chan, P.K., Stolfo, S.J.: Experiments on multistrategy learning by meta-learning. In Proceedings of the Second International Conference on Information and Knowledge Management, 1993, 314-323.
4. Dash M., Liu H.: Feature selection for classification. Intelligent Data Analysis, 1 (3), 1997, (electronic journal http://www.elsevier.com/locate/ida).
5. Devijver P.A., Kittler J.: Pattern recognition: A statistical approach. Englewwod Cliffs, NJ. Prentice-Hall, 1982.
6. Friedman, J.H.: Another approach to polychotomous classification, Technical Report, Stanford University, 1996.
7. Hastie, T., Tibshirani R.: Classification by pairwise coupling, Proc. NIPS97.
8. Jelonek J., Krawiec K., Slowinski R., Szyma J. Rough set reduction of features for picture-reasoning. In T.Y. Lin, A.M. Wildberger (eds) Soft Computing, The Society for Computer Simulation, San Diego, 1995, 89-92.
9. Jelonek J., Stefanowski J., Feature subset selection for classification of histological images. Artificial Intelligence in Medicine, 9, 1997, 227-239.
10. Jelonek, J., Stefanowski J.: Using n^2-classifier to solve multiclass learning problems. Technical Report, Poznan University of Techonology 1997.
11. Jelonek, J., Stefanowski J.: Experiments on solving multiclass learning problems by n^2-classifier. Accepted to 10th European Conference on Machine Learning, Chemnitz, Germany, April 1998.

12. Jelonek, J., Komosinski M.: Using n^2-classifier in constructive induction. Accepted for workshop "Upgrading Learning to the Meta-Level: Model Selection and Data Transformation", 10th European Conference on Machine Learning, Chemnitz, Germany, April 1998.
13. John G., Kohavi R., Pfleger K., Irrelevant features and the subset selection problem. In Proceedings of the Eleventh International Machine Learning Conference. New Brunswick, NJ. Morgan Kaufmann, 1994, 121-129.
14. Kohavi R., Sommerfield D. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. Proceedings of the First International Conference on Knowledge Discovery and Data Mining. Montreal, AAAI Press, 1995, 192-197.
15. Langley P. Elements of Machine Learning. Morgan Kaufmann, San Mateo C.A., 1996.
16. Mayoraz, E., Moreira, M.: On the decomposition of polychotomies into dichotomies, Proc. 14th Int. Conf. Machine Learning, July 1997, 219-226.
17. Murphy, P.M., Aha, D.W.: UCI Repository of machine learning databases. University of California at Irvine, Department of Information and Computer Science [URL: http://www.ics.uci.edu/ mlearn/MLRepositoru.html].
18. Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer, Dordrecht 1991.
19. Pawlak Z., Slowinski K.,Slowinski R.: Rough classification of patients after highly selective vagotomy for duodenal ulcer. Int. Journal of Man-Machine Studies, 24, 1988, 413–433.
20. Pudil P., Novovicova J., Kittler J.: Floating search methods in feature selection. Pattern Recognition Letters, 15, 1994, 1119-1125.
21. Stefanowski J., Slowinski K.,: Rough set theory and rule induction techniques for discovery of attributes dependencies in medical information systems. In Komorowski J., Zytkow J.(eds.) Principles of Knowledge Discovery. Proceedings of the First European Symposium: PKDD'97, Trondhiem, Norway, Springer LNAI no. 1263, Springer Verlag, 1997, 36–46
22. Stefanowski J., Urbaniak Sz.: Using case-based learning in decision support systems. In Proceedings of Intelligent Information Systems VI, Zakopne June 97, IPI PAN Press, 159-168.