TOWARDS A COMBINED APPROACH TO FEATURE SELECTION

Camelia Vidrighin Bratu and Rodica Potolea
Technical University of Cluj-Napoca Computer Science Department, Baritiu St.,Cluj-Napoca, Romania

Keywords: Feature Selection, Classification, Baseline Accuracy, Combined Method.

Abstract: Feature selection is an important step in any data mining process, for many reasons. In this paper we consider the improvement of the prediction accuracy as the main goal of a feature selection method. We focus on an existing 3-step formalism, including a generation procedure, evaluation function and validation procedure. The performance evaluations have yielded that no individual 3-tuple (generation, evaluation and validation procedure) can be identified such that it achieves best performance on any dataset, with any learning algorithm. Moreover, the experimental results suggest the possibility of tackling a combined approach to the feature selection problem. So far we have experienced with the combination of several generation procedures, but we believe that the evaluation functions can also be successfully combined.

1 INTRODUCTION

In supervised learning, a set of instances, each having a number of features, is given, and the goal is to evaluate some objective function, and optimize it. A common objective function is the prediction accuracy, and in this paper we are considering it as the evaluation criterion. It is a well known fact that no classifier can perform remarkably well on any data set (neither real, nor artificial). Hence, the classifier should be selected and adjusted according to the data particularities. Previously, (Moldovan, 2007) we have introduced a method for establishing the baseline accuracy for any problem domain. Thus, the choice of a specific learning scheme for a certain problem is further justified only if its performance is better than that of the system proposed there. The new system was evaluated on several classical benchmark datasets, and its performance was compared to that of its component classifiers. Moreover, comparative evaluations have been carried out with ensemble learning methods in order to emphasize the advantages of the chosen methodology. The results obtained have confirmed the assumptions related to the selective superiority of classifiers and have shown that the system's accuracy is, overall, higher than that of the individual classifiers. Stability across different learning problems was also improved.

Supposing that, for a given dataset, the best classifier has been identified, the assumption of monotonic behaviour – that more instances and more attributes improve the performance – is not generally valid. This means that further refinement is required in order to ensure the best possible performance of a given classifier. Identification of the optimal feature subset becomes thus a prerequisite which allows a classifier to improve its performance. As the search for such a subset can be driven in various ways, in this paper we focus on evaluating several feature selection methods. We have started this work with the intention of developing a methodology to select the optimal feature subset for any data set. However, the experimental results have shown that no such single, all-purpose method could be designed, as the feature selection mechanism is strongly related to the particularities of the dataset, in a similar way the learning scheme is. Therefore we extended our work on more selection mechanisms, combining the results with different classifiers, in the attempt to identify the best combination for each individual dataset. Just like in the case of classifiers, our goal is to device a methodology for determining the baseline accuracy for the feature subset selection function, to represent a starting point in the attempt to develop a better method on a given particular dataset.

Although the final purpose of any data mining effort is to solve a real-world problem, and our work also focuses on that, any new approach needs to be first validated on benchmark data, to allow for comparative evaluations. All the tests presented in

Vidrighin Bratu C. and Potolea R. (2008).
TOWARDS A COMBINED APPROACH TO FEATURE SELECTION.
In *Proceedings of the Third International Conference on Software and Data Technologies - ISDM/ABF*, pages 134-139
DOI: 10.5220/0001878401340139

this paper have been performed on benchmark data, from the UCI Machine Learning Repository (UCI).

The rest of the paper is organized as follows: section 2 describes the feature selection method. The experimental work is presented in section 3, while section 4 contains our conclusions and further work.

2 FEATURE SELECTION

In this information age, the volume of data which is gathered for a certain problem is usually immense. The monotonic assumption (i.e. classification on a set of features should perform better than on any subset it contains), has proven to be false for real-world datasets. In many applications, the size of a dataset is so large that learning might perform extremely poorly on the full feature set. This introduces the need to restrict the learning process to an optimal subset of predictor attributes. If the cost is taken into account, it is even more desirable to determine the optimal subset of features that are relevant to the target concept to reduce the evaluation cost. Moreover, if the running time is considered, by reducing the search space (due to the reduced number of features), the running time of the learning algorithm is also improved.

Feature selection is a research area with remarkable results. Important efforts are made to find better methods for feature subset selection, in order to make classifiers more efficient. However, most methods offer different answers when determining whether a particular feature is or not relevant. As presented in (Nilsson, 2007), there are several definitions of relevance, depending on the end goal of the analysis. This results in different feature selection problems, for which Nilsson offers very comprehensive definitions. He also presents the relation between the different feature subsets.

Another good overview of the available formalisms is presented in (John, 1994) and (Kohavi, 1997). The last argues that, from a purely theoretical standpoint, whether a feature is relevant or not is not as important. Instead, the optimal set of features for a given inducer and problem should be analyzed, taking into the account the inducer's heuristics, biases and tradeoffs. Although a relation between the optimal feature subset and relevance is provided, one should focus on finding optimal features, rather than theoretically relevant features.

A lot of work has been carried out to develop feature selection methods, in order to satisfy the demand for obtaining good machine learning techniques. There exist two main categories of feature selection algorithms: filter methods, which base their selection on the properties of the data distribution alone, and wrapper methods, which employ an empirical risk estimate for a certain inducer to evaluate a feature subset. Representative for the first category are: FOCUS (Allmualim, 1997), RELIEF (Kira, 1992), LVF (Liu, 1996), statistical methods based on hypothesis tests, e.t.c. Wrappers have been first introduced in (Kohavi, 1997).

(Dash, 1997) contains a comprehensive overview of many existing methods. They are classified using two criteria: generation method and evaluation measures. Fifteen possible categories are identified, and representative methods are discussed. A similar survey, but which uses a 3-dimension classification on existing methods is presented in (Molina, 2002). There, algorithms are classified using three criteria: search organization, generation of successors and evaluation function.

Among the many feature aspects considered by feature selection methods, we focus on identifying those features which characterize the target concept best. Therefore, we are not interested in the theoretical classification of features in weak/strong relevant, or relevant/irrelevant/redundant/correlated. In real-world situations, relevant/ irrelevant features are unknown a priori. Therefore, many candidate features are introduced assuming they represent the domain better. However, some of them fail to characterize the target concept, so that the overall result is the degradation of the objective function. A mechanism to make an initial selection on a new real-world data set is required. **Such a mechanism includes a generation procedure, an evaluation function, and a validation procedure**.

Definition: The **feature selection** process selects the minimal subset of features, considering the prediction accuracy as evaluation function.

So, after selecting the feature subset, we consider each selected feature as (strong) relevant, and rejected features as irrelevant (with no further refinement).

Definition: The **generation procedure** is a search procedure that selects a subset of features (F_i) from the original feature set (F), $F_i \subseteq F$.

There are many search methods available today, from greedy hill climbing search, to genetic or random search methods. Each has its advantages and disadvantages. For the purpose of this paper, we have evaluated 4 of the best known search procedures, on 11 data sets (in more detail later).

Since the particularities of each dataset strongly influence the choice of the best feature subset, there is no general method for partitioning the dataset in relevant/irrelevant features. Exhaustive search in the

attribute space may provide better results on a new dataset, but its application in problems with more than a few attributes is intractable due to complexity issues.

Definition: The **evaluation function** measures the quality of a subset obtained from a given generation procedure.

As the optimal features subset depends on the evaluation function, the process of selecting the appropriate evaluation function is dependent on the particular initial data set. An evaluation function measures the capability of a feature (subset of features) to distinguish among different class labels.

We have considered the prediction accuracy as the evaluation function (the wrapper method). A subset is considered to better characterize the data set if and only of it improves the prediction accuracy.

Definition: The **validation procedure** tests the validity of the selected subset through comparisons obtained from other feature selection and generation procedure pairs.

The objective of the validation procedure is to identify the best performance that could be obtained in the first two steps of the method for a given data set, i.e. to identify the selection method which is most suitable for the given dataset and classification method. As a consequence, the minimal feature subset is selected. All features from the subset are considered relevant to the target concept. Moreover, the classification method performs the best, so it is to be considered for further classifications. In order to perform the validation we have employed 4 different learning algorithms: Naïve Bayes (Cheeseman, 1995), AdaBoost.M1 (Freund, 1997), a PANE method (Onaci, 2007) and a method for determining the baseline accuracy of a dataset, based on the theory of Dempster-Shafer (DST) (Moldovan, 2007). This was also done with the purpose of studying the connection between the selected subset of features and the final learning scheme: is the attribute subset universally good or there is a strong connection between it and the employed learning algorithm?

Our vision on the feature selection process assumes a very strong connection between the different steps of the method. Moreover, since there is no generic best approach to feature selection (no single combination of steps yields the best solution on a random dataset), there appears the need to consider several approaches at once on a new real-world problem. Although this might not lead to the best possible performance, it guarantees improvement over the initial dataset, due to its stability. Thus, it offers an effective starting point for the evaluation of a new problem.

3 EXPERIMENTAL WORK

In the evaluation part we concentrate on one **feature selection** method, the wrapper subset evaluation method, which uses the learning algorithm as evaluation function. This method has been reported to obtain more significant accuracy improvements than other feature selection approaches (Hall, 2003). As **generation procedures**, we focused on 4 well known methods: greedy forward selection and backward elimination, and forward and bidirectional best-first search.

Forward selection starts with an empty subset of attributes. At each step, it tries to add a new attribute to the subset, by evaluating the worth of the subset with the added attribute, using some numeric measure of the expected performance of the dataset. The best attribute according to this measure is selected, and the procedure continues with the new subset. The search stops when no further improvement can be found. Backward elimination works in a very similar way: it starts with the full feature set and, at each step, tries to eliminate an attribute. Best-first search is slightly more sophisticated. It does not terminate when the performance drops, but instead keeps a list of best partial attribute subsets evaluated so far, so that it can backtrack.

We have employed J4.8 (revision 8 of the more popular C4.5 algorithm (Witten, 2005)) as **evaluation function**. For **validation** we used 4 different learning methods. In the trials we have performed we experimented with the WEKA implementation of the wrapper subset evaluation and search methods (Witten, 2005).

The datasets we experimented on have been taken from the UCI Machine Learning Data Repository. We selected datasets for which the baseline accuracy (determined with the method presented in (Moldovan, 2007)) was around 70-90%, because these possess the highest capacity for improvement. Most of the datasets are two-class problems, with both nominal and numeric attributes.

A first set of tests was conducted to establish whether a certain search method performs better than the others when used with wrapper subset evaluation. We used forward (gsf) and backward (gsb) greedy stepwise search, and forward (bfs) and bidirectional (bfs_bid) best first search. Ten fold cross-validation was performed, which generated

Table 1: Accuracy levels of J4.8 on attribute subsets resulted from wrapper subset evaluation with different search methods.

Dataset	Initial Attrib.	Initial Accuracy	bfs Attrib.	bfs Acc.	gsb Attrib.	gsb Acc.	gsf Attrib.	gsf Acc.	bfs_bid Attrib.	bfs_bid Acc.
Australian	14	86.2	6	85.10	8	87.43	1	85.51	5	84.97
Breast-cancer	9	73.68	4	75.67	4	75.67	3	75.60	4	75.67
Cleve-detrano	13	76.63	7	79.84	5	78.64	5	77.28	5	78.86
Crx	15	84.93	5	85.87	6	86.32	4	85.49	6	85.36
German	20	71.72	10	73.82	10	74.12	8	73.85	7	74.86
Heart	13	76.16	4	83.19	7	80.19	4	83.19	5	82.00
Hepatitis	19	78.05	3	83.59	10	82.28	3	83.59	4	83.45
Labor	17	78.38	6	80.17	7	79.90	4	81.63	6	80.17
Lymphography	18	76.46	6	82.90	8	82.20	4	81.23	6	82.90
Pima diabethes	8	73.82	3	74.26	3	75.73	3	74.26	3	74.26
Tic-tac-toe	9	83.43	7	82.96	6	81.44	1	69.94	6	81.44

selection percentages for each individual attribute. These percentages were used to quantify the "worth" of an attribute. The final feature subset was composed of those attributes having percentages above the average. The feature subset selected by each individual method was then presented to J4.8 for classification (validation), and the accuracy of the learned model was evaluated.

The results are presented in table 1. Although accuracy improvements can be observed for all but one dataset, there is no single method that constantly boosts the performance on every dataset. As reported in (John, 1997), greedy stepwise search may halt prematurely, leaving forward selection with too few attributes, and backward elimination with too many. Best first search (both forward and bidirectional) are somewhere in between, selecting usually more attributes than forward selection and fewer than backward elimination. Like in the case of classifiers, attribute selection methods also exhibit a selective superiority, which makes the problem of appropriately choosing the selection scheme be very important.

Next we wanted to verify whether the evaluation function used in the wrapper selection introduces a significant bias, or the feature subset that was selected and validated with a certain procedure is successful with other learning methods. Therefore, we evaluated the accuracies of the selected subsets with four other learning methods: Naïve Bayes, AdaBoost.M1, a PANE method and a method for determining the baseline accuracy of a dataset, based on the theory of Dempster-Shafer (DST).

Naïve Bayes (Cheeseman, 1995) is a simple probabilistic classifier. It employs Bayes' theorem under strong independence assumptions. It is naïve because, in practice, the independence assumptions usually don't hold.

AdaBoost.M1 (Freund, 1997) employs an ensemble method, by combining several weak classifiers through voting; the resulting composite classifier generally has a higher predictive accuracy than any of its components. Each distinct model is build through the same learning mechanism, by

Table 2: The accuracy of Naïve Bayes on the attribute subsets selected previously, with different search methods.

Dataset	Initial Acc	bfs Acc	gsb Acc	gsf Acc	bfs_bid Acc
Australian	77.35	86.68	76.71	85.51	84.78
Breast-cancer	73.16	74.87	74.87	74.97	74.87
Cleve_detrano	83.73	84.06	83.89	83.30	83.44
Crx	77.86	86.00	78.87	85.45	85.41
German	75.16	73.70	74.70	73.72	73.28
Heart	83.59	82.26	81.00	82.26	83.00
Hepatitis	83.81	83.67	85.89	83.67	85.3
Labor	93.57	89.63	92.23	89.23	89.63
Lymphography	84.90	77.00	77.65	79.52	77.00
Pima diabethes	75.75	75.68	76.72	75.68	75.68
Tic-tac-toe	69.64	71.14	73.12	69.94	73.12

Table 3: The accuracy of AdaBoost.M1 on the attribute subsets selected previously, with different search methods.

Dataset	Initial Acc	bfs Acc	gsb Acc	gsf Acc	bfs_bid Acc
Australian	84.64	85.55	85.35	85.51	85.26
Breast-cancer	72.38	73.58	73.58	74.41	73.58
Cleve_detrano	83.30	84.12	82.91	83.20	83.99
Crx	84.80	85.61	85.48	85.49	85.64
German	71.27	71.74	72.57	72.49	71.8
Heart	81.59	84.85	80.56	84.85	84.52
Hepatitis	81.37	81.80	79.95	81.80	80.72
Labor	88.37	86.43	90.70	87.67	86.43
Lymphography	74.98	75.72	75.72	74.84	75.72
Pima diabethes	74.92	74.80	74.43	74.80	74.80
Tic-tac-toe	72.72	72.34	71.35	69.94	71.35

varying the distribution of examples in the training set.

After each boosting phase, the weights of the misclassified examples are increased, while those for the correctly classified examples are decreased.

Table 4: The accuracy of DST on the attribute subsets selected previously, with different search methods.

Dataset	Initial Acc	bfs Acc	gsb Acc	gsf Acc	bfs_bid Acc
Australian	83.63	83.69	83.28	85.26	85.98
Breast-cancer	71.02	74.76	74.76	74.80	74.76
Cleve_detrano	76.34	78.17	79.45	74.50	78.91
Crx	83.79	84.27	84.34	82.42	85.36
German	69.99	69.69	69.92	67.82	68.64
Heart	76.09	81.17	77.23	81.17	80.90
Hepatitis	81.53	83.19	80.69	83.19	81.15
Labor	80.00	77.00	79.10	77.50	77.00
Lymphography	77.36	77.32	79.11	75.89	77.32
Pima diabethes	72.06	70.68	70.34	70.68	70.68
Tic-tac-toe	87.63	83.30	82.34	69.89	82.34

Table 5: The accuracy of PANE on the attribute subsets selected previously, with different search methods.

Dataset	Initial Acc	bfs Acc	gsb Acc	gsf Acc	bfs_bid Acc
Australian	85.36	86.28	85.25	85.29	85.89
Breast-cancer	72.23	74.00	74.00	74.41	74.00
Cleve_detrano	76.42	79.27	78.77	81.19	81.67
Crx	85.33	85.46	86.06	86.19	85.42
German	71.36	74.14	74.02	75.03	74.56
Heart	78.14	82.57	80.12	82.57	81.87
Hepatitis	80.04	84.17	84.37	84.17	83.54
Labor	79.19	82.64	79.73	79.88	82.64
Lymphography	77.9	78.87	78.90	80.21	78.87
Pima diabethes	74.5	75.27	75.22	75.27	75.27
Tic-tac-toe	84.31	81.40	79.60	69.42	79.60

The PANE method (Onaci, 2007) focuses on improving the performance of symbolic classifiers (such as decision trees, or rule learners), by using as preprocessing step a neural network ensemble. The aim is to improve the performance and the stability of the classification process, while keeping its transparency.

The DST method (Moldovan, 2007) focuses on establishing the baseline accuracy of a dataset, such as to allow the initial assessment of a dataset. It uses belief functions and the plausible reasoning from the Dempster-Shafer Theory to combine the predictions of different learning schemes.

Table 6: The accuracy of J4.8 on the attribute subsets selected with the combination method.

Dataset	Initial Attrib.	Initial Accuracy	Combination Accuracy	Combin. Attrib.
Australian	14	86.2	84.83	5
Breast-cancer	9	73.68	75.67	4
Cleve-detrano	13	76.63	82.88	5
Crx	15	84.93	86.25	8
German	20	71.72	73.88	9
Heart	13	76.16	83.19	4
Hepatitis	19	78.05	83.18	7
Labor	17	78.38	81.63	4
Lymphography	18	76.46	82.90	6
Pima diabetes	8	73.82	74.26	3
Tic-tac-toe	9	83.43	75.08	3

The results for this second batch of tests are shown in tables 2-5. Although no selection method yields always the best improvement, the bfs and gsb-based methods perform constantly very well. Moreover, the attribute subset that obtains the best improvements depends strongly on the learning scheme. It is known that for the Bayesian classifier, the independence of the attributes is of utmost importance (Cheeseman, 1995), while the decision trees prefer fewer and more relevant attributes. Thus, the choice of the selection scheme, such that it produces the optimal attribute subset for a given problem and learning algorithm is of interest.

Our current efforts focus on combining the selections made with different search methods. In this direction we have already performed experiments with three search methods: forward and backward greedy stepwise and forward best first search. In this version also cross-validation was employed, and we considered the votes of each search method uniformly, resulting in the selection of those attributes which "gathered" a percentage above the average of the three different methods. The results for this experiment are presented in table 6 below. With the exception of two datasets, significant accuracy improvements are observed. The accuracy levels are similar to the highest rates achieved by the individual selection strategies in table 1, or higher. Also, the combination method improves the stability of the selection schemes.

4 CONCLUSIONS

Experience has proved to us that feature selection is a very important step in the data mining process. Although there are many good approaches for selecting a feature subset, there appears to be no

such thing as a globally best feature selection method, or a globally optimal feature subset. No individual 3-tuple (generation, evaluation and validation procedure) can be identified such that it achieves best performance on any dataset, with any learning algorithm. However, due to the particularities of the attributes selected by individual inducers, we expect that the tuples using the same inducer in the evaluation and validation steps will perform better than combined tuples.

Moreover, the experimental results suggest the possibility of tackling a similar approach to the one in (Moldovan, 2007). There, because of the high degree of stability (smaller variations than single classifiers across several datasets), the system can be used to establish the baseline accuracy for a certain dataset. In a similar manner, the selections of several generation methods can be combined in order to achieve higher stability and (possibly) improved performance. The evaluations performed so far in this direction have yielded promising results. However, work still has to be done, to perfect the method, and try new combination approaches. Here we only experimented with a number of different generation procedures, in a manner similar to the ensemble learning methods. The evaluation functions can also be combined. To do that, however, you need a more sophisticated approach. One that seems appropriate is the one used to establish the baseline accuracy of a dataset, using the Dempster-Shafer theory.

The feature selection process can be considered for data imputation as well. By switching the target concept from the class to a particular feature which is incomplete, we can efficiently predict the missing values using only the optimal feature subset which characterizes the particular attribute. This is another current concern in our work.

Also, to enhance cost-sensitive learning, the feature selection mechanism could be modified such as to consider a cost-sensitive evaluation function, instead of the prediction accuracy. This is something we haven't tackled yet, but the idea seems promising.

ACKNOWLEDGEMENTS

The authors wish to thank to Dan Bratucu, Cristian Botau and Adrian Cosinschi for their contributions to the implementation and for running part of the tests.

REFERENCES

Almuallim, H., Dietterich, T. G., 1997. "Learning with many irrelevant features", In *Proceedings of Ninth National Conference on AI,* pp. 547-552.

Cheeseman, P., Stutz, J., 1995. Bayesian classification (AutoClass): Theory and results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining.* Menlo Park, CA: AAAI Press, pp. 153–180.

Dash, M., Liu, H., 1997. Feature Selection for Classification. In *Intelligent Data Analysis 1, 131–156.* INSTICC Press.

Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. In *Journal of Computer and System Sciences*, 55(1):119–139.

Hall, M. A., Holmes, G., 2003. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. In *IEEE Transactions on Knowledge and Data Engineering*, v.15 n.6, p.1437-1447.

Kira, K., Rendell, L. A., 1992. "The feature selection problem - Traditional methods and a new algorithm", *In Proceedings of Ninth National Conference on AI,* pp. 129-134.

Kohavi R., John, J. H., 1997, "Wrappers for feature subset selection", *Artificial Intelligence*, Volume 7, Issue 1-2.

John, G.H., 1997. *Enhancements to the Data Mining Process*. PhD Thesis, Computer Science Department, School of Engineering, Stanford University.

John, G.H., Kohavi, R., Pfleger, K., 1994. Irrelevant features and the subset selection problem. In *Proceedings of the Eleventh International Conference on Machine Learning*, 121–129.

Liu, H., Setiono, R., 1996. "A probabilistic approach to feature selection–a filter solution", *In Proceedings of International Conference on Machine Learning*, pp. 319-327.

Moldovan, T., Vidrighin, C., Giurgiu, I., Potolea, R., 2007. "Evidence Combination for Baseline Accuracy Determination". *Proceedings of the 3rd ICCP*, 6-8 September, Cluj-Napoca, Romania, pp. 41-48.

Molina L. C., Belanche L., Nebot, A., 2002. "Feature Selection Algorithms: A Survey and Experimental Evaluation", *In Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*.

Nilsson, R., 2007. *Statistical Feature Selection, with Applications in Life Science*, PhD Thesis, Linkoping University.

Onaci, A., Vidrighin, C., Cuibus, M., Potolea, R., 2007. "Enhancing Classifiers through Neural Network Ensembles". *Proceedings of the 3rd ICCP*, 6-8 September, Cluj-Napoca, Romania, pp. 57-64.

Witten, I., Frank E., 2005. *Data Mining: Practical machine learning tools and techniques*, 2nd edition, Morgan Kaufmann.