Performing Effective Feature Selection by Investigating the Deep Structure of the Data

Marco Richeldi and **Pier Luca Lanzi**†

CSELT
Centro Studi E Laboratori Telecomunicazioni S.p.A.
Via G. Reiss Romoli 274 - Torino, ITALY
(Marco.Richeldi@cselt.stet.it Lanzi@elet.polimi.it)

Abstract

This paper introduces ADHOC (Automatic Discoverer of Higher-Order Correlation), an algorithm that combines the advantages of both filter and feedback models to enhance the understanding of the given data and to increase the efficiency of the feature selection process. ADHOC partitions the observed features into a number of groups, called factors, that reflect the major dimensions of the phenomenon under consideration. The set of learned factors define the starting point of the search of the best performing feature subset. A genetic algorithm is used to explore the feature space originated by the factors and to determine the set of most informative feature configurations. The feature subset evaluation function is the performance of the induction algorithm. This approach offers three main advantages: (i) the likelihood of selecting good performing features grows; (ii) the complexity of search diminishes consistently; (iii) the possibility of selecting a bad feature subset due to overfitting problems decreases. Extensive experiments on real-world data have been conducted to demonstrate the effectiveness of ADHOC as data reduction technique as well as feature selection method.

Introduction

Feature selection plays a central role in the data analysis process since irrelevant features often degrade the performance of algorithms devoted to data characterization, extraction of rules from data, and construction of predictive models, both in speed and in predictive accuracy. The interest in the feature selection problem is intensifying because of the pressing need of mining volume data warehouses, which usually contain a large number of features (for example, in finance, marketing, and product development applications). Indeed, it is quite a hard task to filter irrelevant features out during the warehouse construction process.

Feature selection algorithms that have appeared in the literature can be categorized in two classes, according to the type of information extracted from the training data and the induction algorithm (John, Kohavi, & Pfleger 1994) Feature selection may be accomplished independently of the performance of the learning algorithm used in the knowledge extraction stage. Optimal feature selection is achieved by maximizing or minimizing a criterion function. Such approach may be referred to as the *filter feature selection model.* Conversely, the effectiveness of the *feedback feature selection model* is directly related to the performance of the concept discovery algorithm, usually in terms of its predictive accuracy. (John, Kohavi, & Pfleger 1994) argued that feedback models are preferable for feature selection algorithms and supported their claims with empirical evidence. However, the literature do not address some important issues. First of all, it is not clear which is the best starting point for the search of a good subset of features. Starting the search on the whole set of original features usually turns out to be unfeasible due to combinatorial explosion when the number of features is not limited. An alternative might be start with the features used by a decision tree algorithm. Second, current feature selection algorithm do not help to answer a basic question that arises in a number of data analysis tasks, that is whether there exist some fundamental dimensions which underlie the given set of observed features. This is a major drawback in marketing applications, for example, in which gaining an insight of the deep structure of the data is as important as achieving a good generalization performance.

The attempt to address these open issues are the basis of our research work. In this paper we introduce a statistical algorithm, called ADHOC (Automatic Discoverer of Higher-Order Correlations), that combines the advantages of both filter and feedback feature selection models to enhance the understanding of the given data and increase

† Current address: Politecnico di Milano, Milano, Italy

the efficiency of the feature selection process. Two empirical analysis on real-world data have been conducted to demonstrate the effectiveness of ADHOC as data reduction technique as well as feature selection method. Experimental results are presented and discussed in the last section of the paper.

Data reduction in ADHOC

Factor Analysis (FA), Principal Component Analysis (PCA) and Cluster Analysis (hereafter designated as Statistical Data Reduction Techniques or SDRTs) are well established procedures that are effective in many domains. But the set of mathematical assumptions on which they rely diminish their applicability in a number of machine learning and data mining applications. This is mainly due to the following factors: SDRTs fit a linear model to the data; are suitable to handle numeric features only; are often fooled by spurious or masked correlation; the outcome of SDRTs is rarely easy to interpret.

Current statistical techniques may not represent an optimal solution to the data reduction issue in the data mining framework. The ADHOC algorithm provides a different approach to data reduction that overcomes some of the problems which degrade the performance of pure statistical techniques. ADHOC accomplishes data reduction in four stages: (i) Detection of linear and non-linear direct associations among the original features, (ii) Detection of indirect associations among features by investigating higher-order correlations, (iii) Clustering of related features to discover the hierarchy of concepts underlying the data, (iv) Selection of the most informative partition of the features.

Analysis of direct association between features. Input of the algorithm is a training set of feature-valued examples. In the first stage, ADHOC measures direct pairwise association between features by comparing the outcome of two non-parametric (distribution-free) statistical analysis, namely, correlation analysis and chi-square analysis. Measurement of the (linear or non-linear) dependence between any pair of features are normalized in the range [-1, 1] and collected in a matrix called the *first order dependence matrix*. Unlike SDRTs, ADHOC can handle both numeric and symbolic features. Numeric features are automatically discretized with the algorithm described in (Richeldi & Rossotto 1995) if they need to be compared with symbolic features to estimate possible dependence. ADHOC selects the most appropriate test from a set of available statistics for any given pair of features automatically. For example, correlation between a real-valued feature and an ordinal discrete-valued feature is estimated by applying a Stuart's Tau c test.

Analysis of indirect association between features. ADHOC identifies groups of features that are equivalent measures of some factor. It can be regarded, therefore, as a clustering technique. However, the mechanism underlying the formation of clusters is very different from the one employed by cluster analysis of features.

SDRTs rely on the analysis of direct correlation between features to perform data summarization. Their goal is to obtain factors that help to explain the correlation matrix of the features. But correlation may not provide a reliable measure of the association between two features, as it does not take effects of other features into account. Spurios correlations may occur, or correlations may be imposed artificially or masked by other features. In this case, indirect relationships between features need to be investigated, since direct associations, which are measured by correlation, do not convey enough information to discover the deep structure of the data.

ADHOC search for indirect association is based on the concept of feature *profile*. The profile of a feature F denotes which features F is related to and which ones F is not related to. For example, let A, B, C, D, E, and F be six features that characterize a given data set. Also, let 0.2, 0.1, -0.8, 0.3, and 0.9, be estimates of the direct relationships between F and A, B, C, D, and E, respectively. F's profile is defined as the vector <0.2, 0.1, -0.8, 0.3, 0.9, 1.0>. Features which have similar profiles provide different measurement of the same concept (data dimension) for they are equally related (unrelated) to the rest of the features. If the converse were true, two concepts would be related in two contrasting ways at the same time, a very much unlikely situation in nature. Comparing feature profiles may yield more reliable an estimate of true association than a direct measure of association, such as correlation, provided the cardinality of the feature profile is not too small (at least 4). Since components of the profile vector express correlations, comparing feature profiles may be viewed as correlating correlations. The result of the comparison has been named *2nd-order correlation* in (Doyle 1992), to stress out the double application of correlation. Accordingly, standard Pearson's correlation coefficient is named *1-st order correlation*. A statistical test, called R_{sim}, was designed to estimate profile similarity. Higher-order correlations between features are computed by recursive application of the R_{sim} statistics. Nth-order correlations result in a matrix called the *Nth-order dependence matrix*. By examining the Nth-order dependence matrix, one can determine the strength of relationship between features, and group those features that appear to be related. The recursive process halts when the profile similarity of features in each cluster goes over a predefined threshold or a given number of iterations have been done. Predictor variables may be partitioned into four different categories of clusters. They are called *Positive_Concept*, *Negative_Concept*, *Undefined*, and *Unrelated_Features*, respectively, and reflect the different typology and strength of dependences that may exist between a set of features. Features that share very similar profiles, i.e., that appear to contribute to the same dimension of the phenomenon, are grouped into a cluster of type Positive_Concept. Features related by a negative association to other features are assigned to a Negative_Concept-type cluster. Features which appear not to influence or to

be influenced significantly by the rest of the features form the Unrelated_Features cluster. The Undefined cluster contains all the remaining features which can not be assigned to one of the other three types of clusters.

The analysis can then be repeated on each group of features in turn. The aim is refining the classification, as for the Positive_Concept clusters, or identifying relationships that could be masked by other features, as for the Undefined cluster. Cluster refinement is terminated when cluster cardinality goes below to a predefined value.

As a result, ADHOC returns a hierarchy of clusters which would resemble the hierarchy of concepts that characterize the observed phenomenon. A test of homogeneity of content is then applied to every level of the hierarchy to determine a good factorization of features.

Selection of the best feature subset

The problem of feature selection involves finding a good subset of features under some objective function, such as generalization performance or minimal use of input features. In our opinion, a feature subset cannot be truly informative and, consequently, good performing on unseen cases, unless it contains at least one feature which contribute to define every dimension underlying the data. Moreover, if there exist n important concepts that contribute to the target concept, and a feedback model identifies a feature subset with less than n features which achieves the best predictive performance, it is very likely that the subset overfit the data. On the other hand, in the very unlikely case in which data has no structure, every feature can be regarded as reflecting a single concept and the search would start from the entire set of features. The search for the best feature subset in ADHOC is based on the above considerations. The second step of the algorithm consists of selecting at most one feature from each of the factor, i.e., dimension of the data, that has been discovered in the data reduction step. As a consequence, feature subsets that reflect all the problem dimensions are formed, and search efficiency strongly increases. We investigated several search heuristics to select the smallest number of features from each factor (group of features which reflect the same data dimension). Among the others, genetic algorithms (GAs) turned out to be an excellent fit to this task (Vafai & De Jong 1992). Experimental studies were conducted by forcing the GA to select at most one feature from each factor, in order to focus the search on the best performing, least-sized feature subset which covers all the data dimensions. The feature subset evaluation function was the generalization performance of the induction algorithm C4.5. To fairly estimate the prediction accuracy of the learning algorithm, a k-fold cross-validation technique was applied. The training data set was broken into k equally sized partitions. The learning algorithm was run k times; in each run k-1 partitions were used as training set and the other one as test set. The generalization accuracy was estimated by averaging the error rate on the test set over all the k runs. Results, which are summarized and discussed in the next section, confirmed the intuition that GAs are able to find highly predictive, in many cases nearly-optimal, feature subsets.

Experimental results

We carried out an extensive empirical analysis in order to evaluate the effectiveness of ADHOC. We selected 14 real-world datasets featuring different types of problematic features, i.e., interacting, redundant and irrelevant features in different measures. Some of the datasets were drawn from the U. C. Irvine Repository (Murphy & Aha 1994), others from the StatLog Repository (Michie, Spieghalter, & Taylor 1994) and the COCOMO data set from (Bohem 1981). The experiments were carried out as follows. Real-valued features were discretized using the algorithm described in (Richeldi & Rossotto 1995) when necessary. The selection of the best number of data dimensions was left to the algorithm. The second step of ADHOC was performed by running C4.5 as induction algorithm and using the pruned trees. As a consequence, C4.5 was also used as term of comparison for the accuracy of the resulting feature subsets. To estimate the generalization performance of feature subsets, 10-fold cross-validation was used. ADHOC was first run on the training data; then the test set was used to evaluate the performance of the best feature subset learned by GA. The tables in this section report the average over the 10 runs.

Table 1 shows that the performance of feature subsets discovered by ADHOC improves C4.5 on 11 out of 14 domains. In particular, five times the improvement is significant at the 95% confidence level and twice at the 90% confidence level. ADHOC's performance is worse than C4.5's on the remaining three domains, in one of which, namely Segment, the degradation was significant at the 95% confidence level. Table 1 reports also the cardinality of the output feature subsets. Lack of space makes it impossible to list the factors as were discovered by ADHOC in the data reduction step for each dataset. However, we refer the interested reader to (Richeldi & Rossotto 1996) for a description of the results which were attained for two of the most interesting domains, namely German and COCOMO.

A second empirical analysis was conducted to evaluate the performance of the data reduction algorithm that was introduced above, hereafter designated ADHOC-DR. The test was made by comparison with the performance of factor analysis (FA) and cluster analysis (CA) of features. Basically, we run ADHOC on the same domains employed in the previous analysis two more times. The first step of ADHOC was modified to replace ADHOC-DR with FA and CA in turn. The second step of ADHOC was left unchanged, so that the discovered set of factors were used as starting point for the search of feature subsets carried out by the GA. Of course, since FA and CA cannot handle symbolic features, we had them to work on the

same input correlation matrix that was used to feed ADHOC-DR. This was the best way to make a fair comparison among the three methods. Moreover, we run FA by using different factor extraction and rotation methods, then reporting the best performance result in case more alternative factor sets were discovered.

Table 2 summarizes comparison results. It can be noticed that both FA and CA could not process 6 out of the 14 datasets due to multicollinearity among the features. ADHOC-DR outperformed statistical data reduction techniques in all the remaining domains. The improvement was significant over the 95% confidence level in 3 out of 8 domains. Further analysis showed that each of the three datasets is characterized by quadratic relationships among features which cannot be discovered by statistical methods based on linear models. These results support the claim that investigating higher-order correlation may well overcome some of the problem of statistical techniques devoted to data reduction.

References

John, G. H.; Kohavi, R.; and Pfleger, K. 1994. Irrelevant Features and the Subset Selection Problem. In Procedings of the 11th Int. Conf. on Machine Learning.

Doyle, J. 1992. MCC-Multiple Correlation Clustering. *Int. Journal of Man-Machine Studies* 37, 751-765.

Vafai, H. and De Jong, K. 1992. Genetic algorithms as a tool for features selection in machine learning. In Proceedings of the 4th International Conference on Tools with Artificial Intelligence, 200-203.

Bohem, B. W. 1981. *Software Engineering Economics*. Prentice Hall.

Richeldi, M. and Rossotto, M. 1995. Class-Driven Statistical Discretization of Continuous Attributes. In Proceedings. of the 8th European Conf. of Machine Learning. Springer & Verlag.

Murphy, P. M. and Aha, D. W. 1994. UCI repository of machine learning databases.

Michie, D.; Spieghalter, D. J. and Taylor, C. C. eds. 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood Publ.

Richeldi, M. and Rossotto, M. 1996. Combining Statistical Techniques and Search Heuristic to Perform Effective Feature Selection. To appear in *Machine Learning and Statistics: the Interface*. Taylor, C., and Nakhaeizadeh, R., eds. John Wiley & Sons.

Dataset	C4.5		ADHOC		p-value	Dataset Size	Source
	%acc ± σ	*	%acc ± σ	**			
Anneal	95.6±1.6	19	95.0±2.3	8	0.292	798	UCI
Australian	84.2±4.0	14	86.7±2.8	5	0.014	690	STATLOG
Cocomo	73.8±21.2	18	77.2±21.8	7	0.420	63	BOHEM
CRX	85.0±4.0	15	85.1±6.1	7	0.941	690	UCI
Diabetes	69.6±5.4	8	71.2±5.5	3	0.404	768	STATLOG
German	69.5±5.2	20	74.2±2.5	7	0.018	1000	STATLOG
Glass	66.3±11.6	9	70.5±7.8	4	0.064	214	UCI
Heart	74.8±5.5	13	80.8±6.5	5	0.011	270	STATLOG
Pima	69.6±5.4	8	73.2±3.8	3	0.112	768	UCI
Satimage	85.5±1.3	35	86.6±1.1	6	0.040	6435	STATLOG
Segment	96.4±0.8	19	95.4±1.0	7	0.022	2310	STATLOG
Sonar	60.7±7.2	60	76.0±9.0	16	0.000	312	UCI
Vehicle	70.3±3.3	18	69.6±6.1	7	0.761	846	STATLOG
Vote	95.0±4.0	16	95.7±3.5	5	0.081	435	UCI

Table 1. C4.5 and of ADHOC's predictive accuracy on all the features and on the best feature subset, respectively. * column: no. of original features; ** column: size of the best feature subset. St. dev. given after the ± sign. P-values computed using a two-tailed T test.

Dataset	ADHOC		Factor Analysis		p-value	Cluster Analysis		p-value
	%acc ± σ	No Att.	%acc ± σ	No Att.		%acc ± σ	No Att.	
German	74.2±2.5	7	72.9±3.8	7	0.454	73.3±3.8	7	0.553
Glass	70.5±7.8	4	67.2±10.8	4	0.271	66.3±8.9	4	0.242
Pima	73.2±3.8	3	72.3±4.4	3	0.427	72.0±3.9	3	0.405
Satimage	86.6±1.1	6	85.5±1.4	6	0.021	85.4±1.4	6	0.021
Segment	95.4±1.0	7	94.4±1.2	7	0.045	82.7±2.6	7	0.000
Sonar	76.0±9.0	16	71.2±10.4	13	0.001	69.8±10.8	13	0.005
Vehicle	69.6±6.1	7	66.2±3.8	7	0.142	69.0±3.6	7	0.816
Vote	95.7±3.5	5	95.4±3.4	5	0.343	95.4±3.4	5	0.343

Table 2. Percentage predictive accuracy of ADHOC, Factor Analysis and Cluster analysis. St. dev. given after the ± sign. "No.Att" columns indicate the size of the best performing feature subsets. P-values were computed using a two-tailed T test.