A VARIABLE GROUPING METHOD BASED ON GRAPH THEORETIC TECHNIQUES

KATI VIIKKI

DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES

UNIVERSITY OF TAMPERE

REPORT A-2001-1

UNIVERSITY OF TAMPERE
DEPARTMENT OF COMPUTER AND
INFORMATION SCIENCES
SERIES OF PUBLICATIONS A
A-2001-1, FEBRUARY 2001

A VARIABLE GROUPING METHOD BASED ON GRAPH THEORETIC TECHNIQUES

KATI VIIKKI

University of Tampere
Department of Computer and Information Sciences
P.O.Box 607
FIN-33014 University of Tampere, Finland

ISBN 951-44-5051-5
ISSN 1457-2060

A Variable Grouping Method Based On Graph Theoretic Techniques

K. Viikki (`kati.viikki@cs.uta.fi`)
Department of Computer and Information Sciences, University of Tampere

Abstract. This paper presents a variable grouping method that is based on techniques of graph theory. An association matrix is calculated and the properties of the graph induced by the matrix are employed in variable grouping. The method gives a deeper insight into data. Furthermore, it can be utilised in feature subset selection for machine learning and statistical methods.

Keywords: Variable grouping, Graph theory, Measures of association, Feature subset selection

1. Introduction

Consider a data set collected in an experimental study or retrieved from a database of an information system. The data set consists of cases representing concrete or abstract objects of the application domain. Certain characteristics or properties of these objects have been targets of interest and measurement, that is, numeric or symbolic values have been attached to the variables corresponding to them. The variables have been measured using nominal, categorical, interval, or ratio scales that determine statistical methods applicable to them (Freund and Simon, 1992). The information contained in the data set can be extracted using descriptive statistics (Freund and Simon, 1992), measures of association (Freund and Simon, 1992), or more sophisticated data analytic techniques (Sharma, 1996) such as regression analysis, analysis of variance, discriminant analysis, principal component analysis, and cluster analysis. Information or knowledge extraction methods have been developed also in the field of machine learning; examples of such methods are nearest neighbour techniques, decision tree methods, neural networks, and Bayesian learning (Mitchell, 1997).

The definition of the knowledge representation, that is, variables that are used to describe cases, is an essential task when models are formed from data sets using machine learning or statistical methods. Although a larger set of variables usually carries more information than a smaller one, it is not necessarily beneficial for model forming. First, simple models are easier to interpret and, thus, preferable. Second, increasing the number of variables may result in a decreased accuracy. Decision tree algorithms, for example, have been reported to produce weaker results with too many variables (Kohavi and John, 1997; Viikki et al., 1999). Even if the increase in the number of variables does not decrease the accuracy, a method may be sensitive to correlated or dependent variables. Examples of such methods are regression analysis (Weisberg, 1985) and Naive Bayes classifiers (Kohavi and John, 1997; Mitchell, 1997).

In some application areas, there are a large number of attributes available for model forming. Similarly, when data are retrieved from databases for knowledge extraction there are usually a considerable number of available variables. Still another example of the wealth of variables is provided by studies of experimental sciences. If the phenomenon to be studied is not well known, a large number of potentially relevant variables may be measured. The abundance of available variables makes the task of variable or feature subset selection (Kohavi and John, 1997) difficult and time-consuming, which calls for automated methods.

This paper presents a variable grouping method that can be utilised in feature subset selection. The method is based on measures of association and techniques of graph theory. An association matrix is calculated and the properties of the graph induced by the matrix are used to form variable groups. The objective is to find groups of related variables that describe some phenomenon from different points of view. The method gives a deeper insight into data: If the data set is large with respect to the number of variables, the mere task of finding related variables can be tedious. Further, the variable groups can be used to guide variable selection for model forming. When variables are selected, the aim is to maximise the information contained by the selected variables without including correlated variables and redundant information. Selecting only one variable from the variable group approximates this provided that the selected variable widely 'represents' the information of the entire group. Alternatively, a set of independent variables found from the variable group can be selected.

The objective of cluster analysis (Sharma, 1996) is to form disjoint subsets or clusters by grouping elements that are highly similar to each other into the same cluster and elements that have low similarity to each other into different clusters. Statistical packages, such as SPSS (SPSS, 2000), provide tools for clustering variables. However, these tools typically use only one measure of association in clustering, which causes problems with mixed data sets having variables of different measurement scales. The demands set by the mixed vertigo data set (Viikki et al., 1999) used in our earlier studies led to the development of the variable grouping method reported in this paper. The need for detailed information about variable groups (variables compressing the information of the entire variable group and sets of independent variables) affected also the development of the method. Techniques of graph theory have been utilised in the field of cluster analysis for decades; a recent example is a clustering algorithm based on high-connectivity in similarity graphs by Hartuv and Shamir (2000).

2. Variable grouping method

In this section, the variable grouping method is described. The reader is assumed to be familiar with the basic concepts of graph theory as given, e.g. in (Even, 1979).

2.1. Association matrix

In order to form variable groups, the presence or absence of the relation between two variables has to be known for all variables. Further, the degrees of the present relations have to be known. There are various measures of association and their tests of significance such as the phi coefficient, the Cramér's V coefficient, the Spearman's rank-order correlation coefficient, and the Pearson's product-moment correlation coefficient (Siegel and Castellan, 1988). Generally speaking, the choice of the measure depends on the scales and distributions of the variables involved and the number of cases.

The variable grouping algorithm presented in Table I takes as its input a data set D and a list of the measurement scales of its variables. The degrees of associations between the variables are calculated using appropriate measures of association and stored in the association matrix M. The matrix can be seen as an undirected graph in which vertices represent variables and degrees of association over some threshold value establish connections between them. The time complexity $T(D)$ for computing the association matrix for the data set D depends on the number of variables and cases and on the measures of association used in the matrix. Because the matrix is symmetric and has 1's along the main diagonal, only the upper triangle matrix has to be calculated. Hence, we have

$$T(D) = \sum_{i=1}^{v-1} \sum_{j=i+1}^{v} f_{ij}(M_{ij}), \tag{1}$$

where i and j are variable indices, v is the number of variables, and $f_{ij}(M_{ij})$ is the time complexity for calculating the measure of association used in the element M_{ij}.

Let n be the number of cases. The time complexity for measures of association typically varies from $O(n)$ to $O(n^2)$. For example, the

Table I. The variable grouping algorithm.

Algorithm	VG
Input:	A data set D, a list of the measurement scales of its variables, and threshold values.
Output:	Variable groups, and the corresponding degrees of variables, cliques, and independent sets.

1. Compute the association matrix M inducing a graph G.
2. Find the connected components (variable groups) of the graph G with the depth-first search.
3. For each connected component
   3.1 Find the cliques.
   3.2 Find the independent sets.

Pearson's product moment correlation coefficient has the time complexity of $O(n)$. Association measures that require sorting of cases (e.g. the phi coefficient, the Cramér's V, and the Spearman's rank-order correlation coefficient) have the time complexity of $O(n \log n)$ or $O(n^2)$ depending on the implementation of the sorting process. Hence, the time complexity for computing the association matrix typically varies from $O(v^2 n)$ to $O(v^2 n^2)$. Notice however, that it might be possible to have cases for which all the variables do not have values. This can decrease the time needed in a practical situation, but does not affect to the worst case time bounds.

2.2. Variable groups

Related variables form variable groups. Two variables can be related to each other directly or via other variable(s). A structure formed by related variables corresponds to a connected component of the graph induced by the association matrix. The connected components are found by traversing the graph using the depth-first search. The time complexity is linear on the size of the graph but quadratic on the number of variables.

From the viewpoint of feature subset selection, the structures formed by dependent (respectively independent) variables in the connected components are of interest to us. A set of variables in which all elements are dependent from each other forms a clique. Cliques can overlap, and a vertex with the largest number of cliques to which it belongs is called a dominating variable of a connected component. Notice that a dominating variable is not necessarily unique. If we want to select only one variable from a connected component for the model forming, a dominating variable is a reasonable choice because it represents, in a way, the information of the entire connected component. Alternatively, we may want to choose from a connected component the maximum number of variables that are independent among themselves. The corresponding structure in the graph is a maximum independent set in which all elements are not directly connected to each other.

Both the maximum clique and the independent set problems are known to be NP-complete (see e.g. Even, 1979). However, if the size of the connected component is relatively small, the cliques and independent sets can be found in an acceptable time. The size of the connected component is, in the worst case, equal to the number of variables. Even if this is not the case, the size of the component may be too large for searching cliques and independent sets. When the size of the component hinders us to find the cliques and independent sets, the following heuristic can be used. Vertices with the largest degrees are probably the most dominating ones, and, thus, we can select one of them. Respectively, vertices with the lowest degrees belong probably to maximum independent sets that can be formed by selecting vertices with the lowest degrees. The number of the vertices included in an independent set must be adjusted experimentally.

Dependability of the variables defines the density of the graph. Variables that all are dependent from each other induce a complete graph, whereas a variable set with a low number of dependencies induces a sparse graph. The threshold value establishing connections between the vertices can be used to regulate the density of the graph. The lower the threshold value is, the denser the graph becomes, and vice versa. In the selection of the threshold value, rules of thumb suggested for the interpretation of association measures can be used. For the Pearsons product moment correlation coefficient, for example, the following rule has been proposed: 0.00-0.29 weak, 0.30-0.49 low, 0.50-0.69 moderate, 0.70-0.89 strong, and 0.90-1.00 very strong (Pett, 1997). If the graph is too dense resulting in too large connected components, the threshold value can be increased. However, the used value should be reasonable from the viewpoint of the application area in question.

The algorithm outputs the found variable groups with detailed information about them: degrees of variables, maximum cliques and independent sets. The output of the algorithm can be utilised by inputting it to a system that forms variable subsets by selecting from each variable group one of the dominating variables or one of the maximum independent sets. The formed variable subsets are, in turn, inputs for a model forming method.

3. Concluding remarks

A variable grouping method based on graph theoretic techniques was presented. The applicability of the method does not restrict to the variable grouping but it can be applied to any elements (cases in data sets), for which an association or a similarity measure can be defined.

Our preliminary studies with medical data (Viikki et al., 2001) suggested that finding the cliques and independent sets in an acceptable time is possible in this application area. The largest connected component found from the total of 104 variables contained 21 vertices. However, one of the future aims is to find guidelines for forming independent sets in a heuristic way by selecting variables with the lowest degrees. The problem of adjusting the threshold value establishing the connections between vertices is also an interesting question. The preliminary results with the medical data (Viikki et al., 2001) showed that the variable grouping method is useful. It found variable groups that were reasonable and informative in the opinion of the medical expert and can be used in the variable subset selection. Future work will include development of a feature subset selection system that utilises the method.

Acknowledgements

I would like to thank Erkki Mäkinen, Ph.D., and Prof. Martti Juhola for their valuable comments and advice. This work was funded by Tampere Graduate School in Information Science and Engineering.

References

Even, S. *Graph Algorithms.* Pitman, London, 1979.

Freund, J. E. and G. A. Simon. *Modern Elementary Statistics.* Prentice Hall, Englewood Cliffs, 1992.

Hartuv, E. and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76:175–181, 2000.

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

Mitchell, T. M. *Machine Learning.* McGraw-Hill, New York, 1997.

Pett, M. A. *Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions.* SAGE Publications, Thousand Oaks, 1997.

Sharma, S. *Applied Multivariate Techniques.* John Wiley & Sons, New York, 1996.

Siegel, S. and N. J. Castellan, Jr. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, Singapore, 1988.

SPSS. *SPSS for Windows*, Release 10.0.7, 2000, http://www.spss.com/.

Viikki, K., E. Kentala, M. Juhola, and I. Pyykkö. Decision tree induction in the diagnosis of otoneurological diseases. *Medical Informatics & The Internet in Medicine*, 24:277–289, 1999.

Viikki, K., E. Kentala, M. Juhola, and I. Pyykkö. Feature subset selection for decision tree induction in the context of otoneurological data: a preliminary study. A manuscript submitted to *10th World Congress on Medical Informatics (medinfo2001)*.

Weisberg, S. *Applied Linear Regression*. John Wiley & Sons, New York, 1985.