Variable Reduction for Predictive Modeling with Clustering

Robert Sanche, and Kevin Lonergan, FCAS

Abstract

Motivation. Thousands of variables are contained in insurance data warehouses. In addition, external sources of information could be attached to the data contained in data warehouses. When actuaries build a predictive model, they are confronted with redundant variables which reduce the model efficiency (time to develop the model, interpretation of the results, and inflate variance of the estimates). For these reasons, there is a need for a method to reduce the number of variables to input in the predictive model.

Method. We have used *proc varclus* (SAS/STAT®) to find clusters of variables defined at a geographical level and attached to a database of automobile policies. The procedure finds cluster of variables which are correlated between themselves and not correlated with variables in other clusters. Using business knowledge and $1\text{-}R^2_{ratio}$, cluster representatives can be selected, thus reducing the number of variables. Then, the cluster representatives are input in the predictive model.

Conclusions. The procedure used in the paper for variable clustering quickly reduces a set of numeric variables to a manageable reduced set of variable clusters.

Availability. *proc varclus* from SAS/ STAT® has been used for this study. We found an implementation of variable clustering in R, function varclus, while we did not experiment with it.

Keywords. variable reduction, clustering, statistical method, data mining, predictive modeling.

1. INTRODUCTION

Over the last decade, insurance companies have gathered a vast amount of data in their data warehouses. Some of this information is well-known by the actuaries because it is used for other purposes, e.g. pricing of the policy. Also, there are many sources of external data (demographics, financial, meteorological...) available from vendors. The external sources are typically not as familiar to the actuary as the data from the data warehouses. This vast amount of information is available to create a predictive model. The objective of the predictive model could be to improve the pricing or reserving process, but also to analyze profitability, fraud, catastrophe, and any insurance operation. This amount of information from multiple sources provides numerous variables for the modeling project contemplated.

When a modeling project involves numerous variables, the actuary is confronted with the need to reduce the number of variables in order to create the model. The variables have sometimes an unknown behavior with the objective of the modeling project. In addition, when there is a multitude of variables, it becomes difficult to find out the relationship between variables.

Too many variables reduce the model efficiency. With many variables there is a potential

of overfitting the data. The parameter estimates of the model are destabilized when variables are highly correlated between each other. Also, it is much more difficult to have an explainable model when there are many variables. Finally, creating models with all possible combinations of variables is exhaustive, but this approach would take indefinite time when there are thousands of variables. An intermediate approach to the exhaustive search would also take a lot of time and some combinations of variables could be overseen.

Suppose you want to reduce the number of variables to a smaller set of variable clusters for efficiency, you can use variable clustering. Variable clustering provides groups of variables where variables in a group are similar to other variables in the same group and as dissimilar as possible to variables in another group.

1.1 Research Context

This paper addresses the initial stage of every predictive modeling project performed by an actuary, i.e. variable selection. Then, the variables selected would become inputs to predictive modeling techniques, such as, linear regression, generalized linear model, a neural network, to name a few.

A technical description of the variable clustering algorithm, *proc varclus*, is included in the SAS/STAT® User's Guide.[1] The method is not found in many textbooks on multivariate techniques, it mostly started as an implementation in statistical software.[2]

This paper is focused on variable clustering, but the example could be used, for example, in the context of complement to territorial relativities for automobile insurance. This complement would be obtained from a predictive model based on variables defined at some geographical level. The variables were selected using variable clustering on multiple sources of information, usually not used in pricing, attached to an automobile policy database. If the objective of the predictive model is to predict cost by territory, it makes sense to use fact (demographics, consumer expenditure, weather ...) variables selected from the variable clustering on the multiple sources, defined at some geographies (e.g. county), to complement territorial relativities.

The example provided in the paper is a simplification of a variable reduction problem. Many more variables would be clustered in a real life study.

Note that the variables used in the example have some intuitive relation to automobile

[1] SAS/STAT® 9.1 User's Guide p. 4797
[2] **Pasta paper 205**

insurance cost, although generally the variables presented to the variable clustering procedure are not previously filtered based on some educated guess. All the demographics, consumer expenditure, and weather variables are used in the clustering analysis. Filtering of variables is typically done after the variable clusters have been created. When there is a multitude of variables, it is more difficult to recognize irrelevant variables than to recognize redundant variables. A variable is considered irrelevant if it is not predictive for the specific predictive model. When the actuary deals with unknown data, a large number of the variables turn out to be irrelevant. A variable is redundant when it is highly correlated with another potential variable.

1.2 Objective

More and more actuaries use advanced statistical methods to create insurance models. This paper provides a tool; variable clustering, that can be added to the arsenal of the actuarial miners. Traditionally, PCA have been used for variable reduction by creating a set of components (weighted linear combinations of the original variables) which are difficult to interpret.

Typically, in the clustering literature, there is a rule for selecting the cluster representative, the $1\text{-}R^2_{ratio}$. Business knowledge from subject matter expert should also complement this rule to guide the selection of variables. For this reason, someone could decide to use more than one variable per cluster. Even though the clustering procedure provides diagnostic measures, there are reasons for using more than one variable per cluster. One of them is that the maximum number of clusters is a parameter provided by the user of the procedure. Also, for communication to users of the predictive model, an alternate variable may provide a better intuitive interpretation of the model than the cluster representative.

We should point out that the variable clustering works only with numeric variables. However, there are ways to convert categorical variables into numeric variables. For example, the hamming distance converts categorical variables into a numeric variable. Conversion of categorical variables is not covered in this paper.

We suggest options (*centroid* without *cov*) to the procedure of variable clustering which turn out to produce a scale-invariant method. Otherwise it would probably be necessary to rescale the ranges of the variables (with *proc standard*).

1.3 Outline

The remainder of the paper proceeds as follows. Section 2 will provide an overview of

clustering and more precisely the variable clustering. We will describe shortly the variable clustering algorithm used in this paper. Section 3 will provide an example of variable reduction in the context of automobile insurance. We will use variable clustering and will explain how variables can be selected to reduce their number. In section 4, we conclude the study. In Appendix A, we include an example of the SAS code and in Appendix B we include the procedure's output.

2. CLUSTERING

2.1 Clustering

"**Cluster Analysis** is a set of methods for constructing a sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual"[3]

In general, the goal of a cluster analysis is to divide a data set into groups of similar characteristics, such that observations in a group are as similar as possible and as dissimilar to observations in another group. Variable clustering, however, does not divide a set of data; instead it splits a set of variables with similar characteristics using a set of subject data.

Clustering is an unsupervised learning technique as it describes how the data is organized without using an outcome[4]. As a comparison, regression is a supervised learning technique as there is an outcome used to derive the model. Most data mining techniques are supervised learning techniques. Unsupervised techniques are only useful when there is redundancy in the data (variables).

At the basis of clustering is the notion of similarity. Without supervision, there is no response to say that occurrence *a* is similar to occurrence *b*. If there was a response associated with each occurrence; it could be used to compare *a* and *b* responses to induce similarity between both. Similarity: Two occurrences are similar if they have common properties.

For example, one occurrence is a car, another occurrence is a motorcycle and the last occurrence is a bicycle. First, lets say we have only number of wheels as a property. Then we would cluster the motorcycle and the bicycle since they have the same number of wheels. However, if we add the number of cylinders and fuel consumption, then the motorcycle is

[3] Everitt 1998
[4] Hastie p.2

more similar to the car. Similarity can be measured by distance measure (Euclidian distance, Manhattan or city block distance ...) or correlation type metrics.

There are two classes of clustering methods:

- Hierarchical: This class of clustering produces clusters that are hierarchically nested within clusters from previous iterations. This is the most commonly used clustering technique.
- Partitive: This class of clustering divides data in clusters by minimizing an error function of the distance between the observation vectors and the reference vectors (centroid - initial guess). This clustering technique requires elaborate selection of parameters and evaluation of the error function for all possible partition is impractical.

There are two approaches to hierarchical clustering:

- Agglomerative
    1. Start with each observation as its own cluster
    2. Compute the similarity between clusters
    3. Merge the clusters that are most similar
    4. Repeat from step 2 until one cluster is left
- Divisive
    1. Start will all observations assigned to one cluster
    2. Compute the similarity between clusters
    3. Split the cluster that are least similar
    4. Repeat from step 2 until each observation is a cluster

2.2 Variable Clustering

The procedure used in this paper for variable clustering is both a divisive algorithm and iterative algorithm. The procedure starts with a single cluster and recursively divides existing clusters into two sub-clusters until it reaches the stopping criteria, producing a hierarchy of disjoint clusters.

As mentioned previously, the procedure starts with all variables in one cluster. Based on the smallest percentage of variation explained by its cluster component a cluster is

chosen for splitting. The chosen cluster is split in two clusters by finding the first two principal components and assigning each variable to the component with which it has the higher correlation. The assignment follows a hierarchical structure with the approach presented in this paper. The clustering stops when the maximum number of clusters is attained or reached a certain percentage of variation explained.

3. VARIABLE CLUSTERING EXAMPLE

After the multiple sources of data (demographics, consumer expenditures, meteorological ...) are attached to the auto policy database, variable clustering can be performed to reduce the number of variables. The SAS code is included in Appendix A. The rule dictates to select the variable with the minimum $1\text{-}R^2_{\text{ratio}}$ as the cluster representative. The $1\text{-}R^2_{\text{ratio}}$ is defined below.

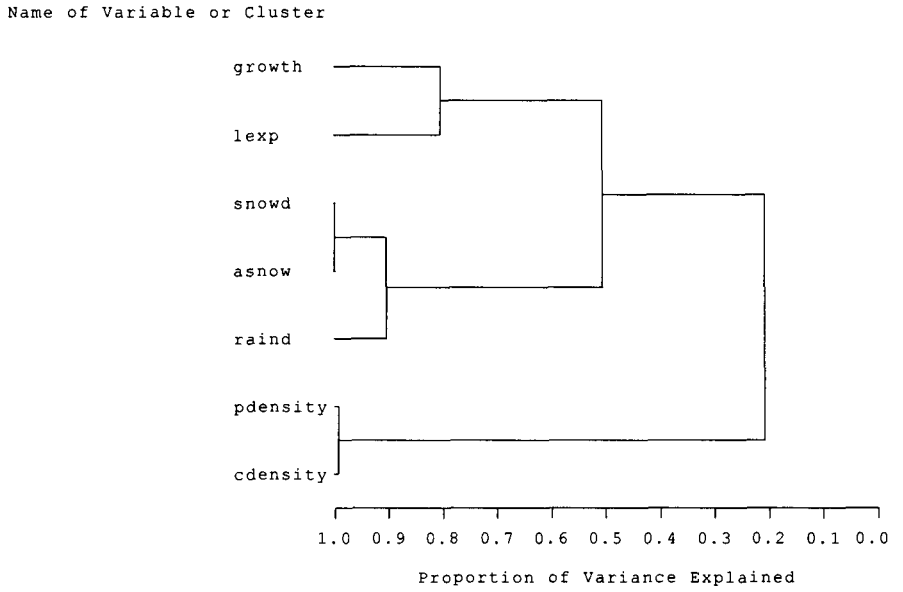
$$1\text{-}R^2_{\text{ratio}} = (1\text{-}R^2_{own})/(1\text{-}R^2_{nearest}) \tag{3.1}$$

Intuitively, we want the cluster representative to be as closely correlated to its own cluster ($R^2_{own} \to 1$) and as uncorrelated to the nearest cluster ($R^2_{nearest} \to 0$). Therefore, the optimal representative of a cluster is a variable where $1\text{-}R^2_{\text{ratio}}$ tends to zero.

Below, we include an extract of the output from *proc varclus* (see Appendix B for additional output from the procedure) with three clusters. Based on the $1\text{-}R^2_{\text{ratio}}$ we should select variables snowd, cdensity, and lexp as cluster representatives.

3 Clusters		R-squared with			
Cluster	**Variable**	**Own Cluster**	**Next Closest**	**1-R**2 Ratio**	
Cluster 1	**Raind**	0.5995	0.0426	0.4183	
	Snowd	0.8976	0.0317	0.1058	Choose
	Asnow	0.8940	0.0314	0.1095	
Cluster 2	**Pdensity**	0.9804	0.0228	0.0201	
	Cdensity	0.9804	0.0113	0.0199	Choose
Cluster 3	**Growth**	0.6459	0.0911	0.3896	
	Lexp	0.6459	0.0013	0.3546	Choose

After *proc varclus*, we have created a tree using *proc tree* which shows how the variable clusters are created. The variables are displayed vertically. The proportion of variance explained at each clustering level is displayed on the horizontal axis.



In that example, variables with similar factual attributes were clustered together; weather variables are in the same cluster and density variables are in the same cluster. Even with more variables, similar grouping patterns are observed.

If we consider three clusters; snowd, asnow and raind would all be in one cluster as they are on the same branch of the tree. The variable snowd would be the cluster representative since it has the lowest $1\text{-}R^2_{\text{ratio}}$.

The number of variables has been reduced and, now, we can efficiently create a predictive model to solve the problem at hand using linear regression, GLM[5], or neural network[6].

4. CONCLUSIONS

Given hundreds of variables, in order to create a predictive model the variable clustering

[5] Holler
[6] Francis

procedure runs quickly and produces satisfying results. We were able to reduce the number of variables using this procedure in order to efficiently create a predictive model. An efficient model was defined as followed:

- Interpretable
- Stable
- Timely

With this procedure, the modeling process is sped up significantly. The hierarchies produced by this procedure are easily interpretable with the tree output. Subject-matter experts usually do not have expertise to analyze statistical output in table form, but given the cluster hierarchy in tree output, can easily uncover alternate cluster representatives or eliminate irrelevant input. Other variable reduction techniques (e.g. PCA) do not create interpretable and disjoint clusters.

Appendix A: Code

```
* Example of variable clustering ;

%let varlist= pdensity cdensity growth /* demographics   */
              lexp                     /* expenditures   */
              raind snowd asnow        /* weather        */

proc varclus data='C:\example.sas7bdat' outtree=tree centroid maxc=6;
var &varlist ;
weight exp;
run;

axis1 label=(angle=0 rotate=0) minor=none;
axis2 minor=none order=(0 to 1 by 0.10);

proc tree data=tree horizontal vaxis=axis1 haxis=axis2;
height _propor_;
run;
```

Appendix B: Ouput

Cluster summary:

Cluster summary gives the number of variables in each cluster. The variation explained by the cluster is displayed. The proportion of variance explained is the variance explained divided by the total variance of the variables in the cluster.

Also displayed, is the summary are the R^2 of each variable with its own cluster, its closest cluster, and the $1\text{-}R^2_{\text{ratio}}$

Cluster Summary for 3 Clusters				
Cluster	**Members**	**Cluster Variation**	**Variation Explained**	**Proportion Explained**
1	3	3	2.371253	0.7904
2	2	2	1.960732	0.9804
3	2	2	1.291809	0.6459

Total variation explained = 5.623794 Proportion = 0.8034

3 Clusters		R-squared with			
Cluster	**Variable**	**Own Cluster**	**Next Closest**	**1-R**2 Ratio**	**Variable Label**
Cluster 1	**Raind**	0.5995	0.0426	0.4183	Rain 2
	Snowd	0.8976	0.0317	0.1058	Snow 2
	Asnow	0.8940	0.0314	0.1095	Snow 1
Cluster 2	**Pdensity**	0.9804	0.0228	0.0201	Pop density
	Cdensity	0.9804	0.0113	0.0199	Car density
Cluster 3	**Growth**	0.6459	0.0911	0.3896	Pop growth
	Lexp	0.6459	0.0013	0.3546	Leg expenditures

Standardized scoring coefficients:

The standardized scoring coefficients predict clusters from the variables. If a variable is not in a cluster, then the coefficient is zero. SAS does not provide unstandardized scoring coefficients.

Standardized Scoring Coefficients				
Cluster		1	2	3
Pdensity	Pop density	0.000000	0.504982	0.000000
Cdensity	Car density	0.000000	0.504982	0.000000
Growth	Pop growth	0.000000	0.000000	0.622137
Lexp	Leg expenditures	0.000000	0.000000	0.622137
Raind	Rain 2	0.374930	0.000000	0.000000
Snowd	Snow 2	0.374930	0.000000	0.000000
Asnow	Snow 1	0.374930	0.000000	0.000000

Cluster Structure:

The cluster structure gives the correlation between the variables and the clusters.

Cluster Structure				
Cluster		1	2	3
Pdensity	Pop density	-.069069	0.990134	-.151107
Cdensity	Car density	-.082041	0.990134	-.106496
Growth	Pop growth	-.301845	-.204659	0.803682
Lexp	Leg expenditures	-.036435	-.004435	0.803682
Raind	Rain 2	0.774267	-.102212	-.206297
Snowd	Snow 2	0.947393	-.044943	-.177956
Asnow	Snow 1	0.945502	-.056370	-.177070

Inter-Cluster Correlation:

This table provides the correlations between the clusters.

Inter-Cluster Correlations			
Cluster	1	2	3
1	1.00000	-0.07631	-0.21046
2	-0.07631	1.00000	-0.13008
3	-0.21046	-0.13008	1.00000

Cluster 3 will be split because it has the smallest proportion of variation explained, 0.645904, which is less than the PROPORTION=1 value.

Final summary:

Cluster summary and the other tables are listed for each number of clusters up to the maximum of clusters (option *maxc*). This table is listed at the end of the output and summarizes for each number of clusters the total variation and proportion explained by the clusters, the minimum proportion explained by a cluster, the minimum R^2 for a variable and the maximum $1\text{-}R^2_{\text{ratio}}$ for a ratio.

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Minimum R-squared for a Variable	Maximum 1-R**2 Ratio for a Variable
1	1.454308	0.2078	0.2078	0.0007	.
2	3.539308	0.5056	0.3157	0.0300	1.0124
3	5.623794	0.8034	0.6459	0.5995	0.4183
4	6.331985	0.9046	0.7904	0.5995	0.4349
5	6.952514	0.9932	0.9804	0.9804	0.0205
6	6.991782	0.9988	0.9959	0.9959	0.0058

5. REFERENCES

[1] B.S. Everitt , *The Cambridge Dictionary of Statistics*, 1998
[2] David J. Pasta, Diana Suhr, "Creating Scales from Questionnaires: PROC VARCLUS vs. Factor Analysis," *SUGI 29 Proceedings*, 2004, Paper 205-29
[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning*, Springer, 2001
[4] SAS Institute Inc. 2004, SAS/STAT® 9.1 User's Guide, NC: SAS Institute
[5] Holler, "Something Old, Something New in Classification Ratemaking with a Novel Use of GLMs for Credit Insurance," *Casualty Actuarial Society Forum, Winter 1999.*
[6] Francis, "Neural Network Demystified," *Casualty Actuarial Society Forum, Winter 2001.*

Abbreviations and notations

PCA, principal component analysis
GLM, generalized linear model
proc, procedure in SAS

Biographies of the Authors

Robert Sanche is a Consultant with Tillinghast a business of Towers Perrin. He is responsible for predictive modeling projects. Prior to joining Tillinghast, he developed class plans for personal lines automobile using multivariate techniques with Travelers and The Hartford. He has also worked for GMAC Insurance and AXA in personal lines doing data mining and ratemaking respectively. He has degrees in Mathematics (Actuarial Science) and Computer Science (Operations Research) from Université de Montréal.

Kevin Lonergan graduated from Southern Connecticut State University in 1969 with BS, 1972 with MS. He taught mathematics in high school from 1969 to 1980. He has developed a new automobile product at turn of century. ACAS 1982. FCAS 1983.