Dimension Reduction Technique for Predictive Modeling

Zhen Zhang[1], Lei Zhang[2], Kendell Churchwell[1], James Veillette[1]
[1]C Spire, 1018 Highland Colony Parkway, Ridgeland, MS 39157
[2]Mississippi State Dept. of Health, 570 E. Woodrow Wilson Dr., Jackson, MS 39216

Abstract

Predictive analytics has been widely used in strategic marketing to uncover actionable information for a range of critical marketing decisions. In today's big data era, advanced technologies and digital processing are generating data in an unprecedented variety, volume and speed for businesses of all kind. Although the big data phenomenon has driven the development of a number of accommodating platforms and analytical algorithms, in the data mining world computational and analytical challenges associated with big data persist, and continue to call for effective techniques to reduce data dimensions.

In this study we concentrate on the reduction of the number of categories in categorical variables for the preparation of inputs for predictive modeling. SPSS MODELER has a Feature Selection node that by default filters out variables with number of categories as a percentage of records greater than 95%. This, however, provides little help because few variables have the number of categories that exceed 95% of the number of records. During our practices of predictive modeling in the telecommunication industry, we find supervised reduction of the number of categories effectively change some otherwise useless categorical variables into contributing predictors that significantly enhances model accuracy.

Key Words: dimension reduction, predictive modeling, supervised learning, algorithm, attributes, categorical variable

1. Introduction

Database are structured to captures facts, attributes and measurements of data elements. High dimensional data carries with it an inherited complexity[1], thus dimension reduction is often needed to prepare data prior to the modeling process. This step not only help to speed up algorithm executions, but also help to improve model accuracy and stability[2, 3, 4]. While the term data dimension has taken on new and ever evolving meanings such as unstructured text data, visual, and audio data, in its traditional sense data dimension refers to the three-fold aspects of data: number of records, number of variables and number of categories in categorical variables.

To prepare a modeling dataset, the first step is to extract an optimal yet manageable number of records from a large database. This step is relatively straightforward and can be achieved through systematic or stratified sampling techniques[5]. The second step is to select an efficient subset of variables from the total available variables while retaining the intrinsic variability and information of the original dataset. Noises and faulty inputs often lead to undermined model performance, therefore removing such un-informative or even miss-informative variables is crucial for archiving an optimal model quality. Techniques for reducing the number of variables have been well established in literature[6, 7, 8]. Most commonly used techniques include Principle Component Analysis (PCA), low variance filtering, feature elimination, etc. To our knowledge little has been reported on the techniques of reducing high number of categories in categorical variables. In our practice of predictive modeling, we often encounter categorical variables that are useless for modeling in its natural form, but can be transformed into modeling blocks that significantly improve model performance. The reduction technique we developed is supported by the concept of supervised learning. Supervised learning here is a borrowed term from the machine learning technology, simply to denote the using of historical data to capture the numeric relationship between each of the categories and the outcome of interest. This captured information is then transformed into a new variable of reduced dimension to be used for model building.

2. Method

2.1 Variable Selection

First, select independent categorical variables that could be potentially predictive of the outcome. It is recommended that these variables have greater than 25 categories. The cutting point 25 here is an empirical number that could vary according to the size of the outcome, as well as the domain of application. For the sake of demonstration, in this study we chose $V_{a.,}$ an independent variable that has 104 categories. Without dimension reduction manipulations, V_a is left out by any commonly used modeling algorithms, such as logistic regression, tree algorithms or neural net, for the prediction of a certain outcome. Yet business knowledge tells that this variable is associated with the outcome.

2.2 Supervised Learning

Next, draw historical samples to be used for supervised learning. If dependent variable is seasonally sensitive, samples should be drawn across seasons. Then estimate the probability of the outcome given each value of the category of $V_{a.}$ This process renders a total of 104 estimates. These estimates capture the learned information of each categories of variable $V_{a,}$ supervised by the outcome variable.

2.3 Normality Test of the Estimates

This dimension reduction technique is bases on the assumption of normality of the estimates. So next step is to conduct Shapiro-Wilk normality test on the 104 estimates. Continue if fail to reject null hypothesis of normality.

2.4 Low Variance Filtering

In addition, the estimates needs to pass the low variance filter to be deemed eligible for further process. Generate mean and standard deviation for the estimates, then calculate coefficient of variation (CV). CV is the ratio of the standard deviation and the mean. Continue if $CV > 0.1$.

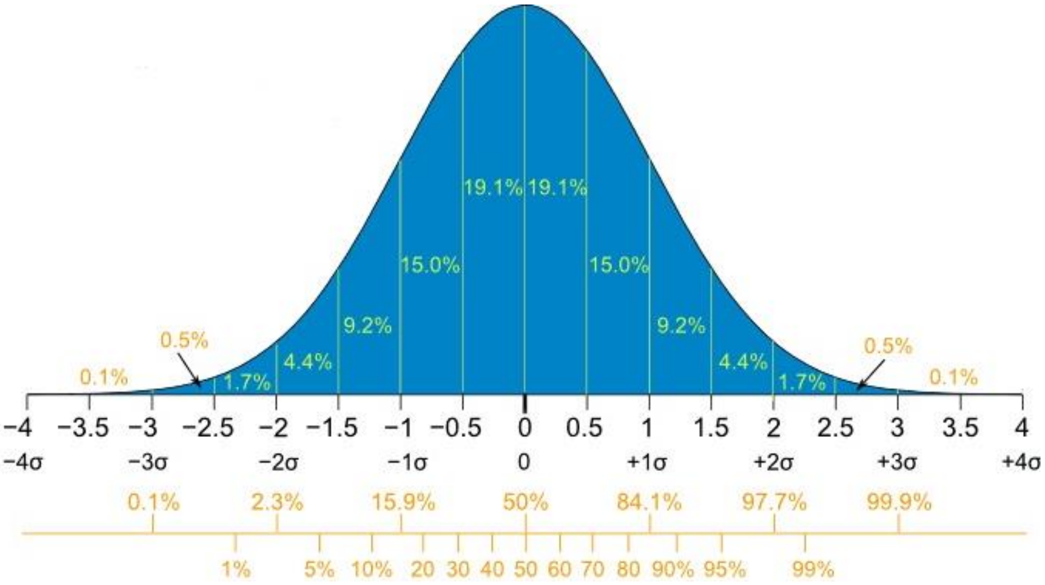
Low CV indicates the lack of relevant information contained in this variable for the predicting of the outcome. The Feature Selection node of SPSS modeler also sets CV equals to 0.1 as the default minimum for low variance filtering, and uses this feature to remove the non-contributing variables. In practice, however, this threshold could be adjusted.

2.5 Constructing the New Variable of Reduced Categorical Dimension

Next, use the estimates' mean, μ, and standard deviation, σ, to construct a new variable, V_b. In this study, we arbitrarily make V_b an ordinal variable of 6 values:

$<\mu-1\sigma$;
$\mu-1\sigma$ to $\mu-0.5\sigma$;
$\mu-0.5\sigma$ to μ;
μ to $\mu+0.5\sigma$;
$\mu+0.5\sigma$ to $\mu+1\sigma$;
$> \mu+1\sigma$.

See chart below for illustration.



2.6 Building models with and without V_b

Finally, draw training and testing datasets using the same stratified sampling method. 50 copies of testing datasets are drawn with a time element for better testing of model stability. Model built without V_b serves as the control model, whereas model built with V_b serves as test model. Then, run 50 copies of testing data through control and test models, record model accuracies.

3. Results

Table 1. Coefficient of Variation for Comparison of Model Stability

	N	Mean	SD	CV
Control Model Accuracy	50	81.1%	1.4%	0.017
Test Model Accuracy	50	81.4%	1.4%	0.017

As shown in Table 1, the mean accuracy for control and test model is 81.1% and 81.4% respectively. The coefficient of variation, which serves as an indirect measure of model stability, are the same for control and test models. This indicates that adding variable V_b into the model does not alter model stability. In addition, Levene's test further confirms homogeneity of variances between the two distributions of model accuracies, as shown in Table 2.

Table 2. Homogeneity of Variance Test Result

	df1	df2	p value
Levene Statistic	1	98	0.919

Table 3 shows the result of t-test for the comparison of mean accuracies of the two models. Test model shows significantly increased accuracy, mean increase is 0.26%.

Table 3. One-sample t-test Result

			95% CI		
	N	Mean	Lower	Upper	p value
Difference between accuracy of control and test models	50	0.26%	0.18%	0.29%	<0.001

4. Discussion/Conclusion

Independent categorical variables with large categories can be transformed into useful predictors through supervised reduction technique.

In practice, the supervised learning of a categorical variable could get stale over time, it is a good practice to always refresh V_b every time the model is refreshed. The value of V_b can be assigned as nominal or ordinal, although generally we find ordinal to be slightly better in terms of predicting. The number of values of V_b can also be optimized with experience.

References

1. Statistical Techniques for Dimension Reduction. UCRL-WEB-201342, 2005
2. A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization. D. Engel, L. Hiittenberger and B. Hamann. OpenAccess Series in Informatics, OASICS Schloss Dagstuhl. 1998.
3. Moden Multidimensional Scaling: Theory and Applications. I. Borg and P. J. F. Groenen. Springer, 2005.

4. A Review of Dimension Reduction Techniques. M. A. Carreira-Perpinan. Technical Report CS-96-09, Department of Computer Science, University of Sheffield, 1997
5. Advanced Sampling Theory with Applications. S. Singh. Kluwer Academic Publishers, 2003
6. Seven Techniques for Dimensionality Reduction. R. Silipo, I Adae, A. Hart and M. Berthold. Open for Innovation KNIME, 2014
7. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 15(6):1373–1396, 2003.
8. J. Ham, D. Lee, S. Mika, and Sch¨olkopf B. A kernel view of the dimensionality reduction of manifolds. In International Conference on Machine Learning, 2004.

Acknowledgements

This work is supported by C Spire, a telecommunication and technology company headquartered in Ridgeland, Mississippi.

Submitting author:

Zhen Zhang, Ph.D., Department of Marketing, C Spire.

1018 Highland Colony Parkway, Ridgeland, MS 39157, USA.

Phone (601) 540-7157

E-mail: zzhang@cspire.com