Improving performance of Memory Based Reasoning model using Weight of Evidence coded categorical variables
Vinoth Kumar Raja, Vignesh Dhanabal and Dr. Goutam Chakraborty, Oklahoma State University

ABSTRACT
Memory based Reasoning (MBR) is an empirical classification method which works by comparing cases in hand with similar examples from the past and then applying that information to the new case. MBR modeling is based on the assumptions that the input variables are numeric, orthogonal to each other, and standardized. The latter two assumptions are taken care by Principal Components’ transformation of raw variables and using the components instead of the raw variables as inputs to MBR. To satisfy the first assumption, the categorical variables are often dummy coded. This raises issues such as increasing dimensionality and overfitting in the training data by introducing discontinuity in the response surface relating inputs and target variables.

The Weight of Evidence (WOE) method overcomes this challenge. This method measures the relative response of the target for each group level of a categorical variable. Then the levels are replaced by the pattern of response of the target variable within that category. SAS® Enterprise Miner’s Interactive Grouping Node is used to achieve this. By this way the categorical variables are converted into numeric. This paper demonstrates the improvement in performance of an MBR model when categorical variables are WOE coded.

A credit screening dataset obtained from SAS Education that comprises of 25 attributes for 3,000 applicants is used for this study. Three different types of MBR models were built using SAS® Enterprise Miner’s MBR node to check the improvement in performance. The results showed the MBR model with WOE coded categorical variables performed best based on misclassification rate. Using this data, when WOE coding was adopted the model misclassification rate decreased from 0.382 to 0.344 while the sensitivity of the model increased from 0.552 to 0.572.

INTRODUCTION
The objective of this study is to demonstrate the improvement in performance of Memory Based Reasoning (MBR) model when categorical variables are Weight of Evidence (WOE) coded instead of dummy coding.

LITERATURE REVIEW
MEMORY BASED REASONING
Memory based Reasoning (MBR) is based on reasoning from memories of past experience. MBR modeling applies this information in classifying (or) predicting new record by finding neighbors similar to it. Hence this approach is case based instead of explanation based. MBR consists of two operations. First, to find the distance between case in hand and the all other observations in order to find its neighbors. Second, combining the results from the neighbors to arrive at the response. In SAS® Enterprise Miner, the MBR node uses K-nearest neighbor algorithm where neighbors are determined by shortest Euclidean distance. To combine the results of the neighbors, democracy method is used. Hence, the response of the target for these neighbors are taken as votes that act as posterior probability for the response of case in hand.
Figure 1. An overview of Memory Based Reasoning

Figure 1 shows the observations that are close to case in hand, their target values and their distance from case in hand. If we consider K as 3, then the nearest neighbors are observations with IDs 108, 12 and 7. And the target values are Y, N and Y respectively. Thus, the posterior probability for the response of case in hand to be Y is 2/3 or 67%. When the target variable is interval, the average of the K-nearest neighbors is calculated as the prediction for case in hand.

MBR ASSUMPTIONS

MBR modeling is based on the assumptions that the input variables are numeric, orthogonal to each other and standardized. The latter two assumptions are taken care by Principal Components' transformation of raw variables and using the components instead of the raw variables as inputs to MBR. But to satisfy the first assumption, the categorical variables are dummy coded. This raises issues such as increase in dimensionality and overfitting in the training data by introducing discontinuity in the response surface relating inputs and target variables.

WEIGHT OF EVIDENCE CODING

Weight of Evidence (WOE) is the method of combining evidence in support of hypothesis i.e., it measures the relative response of target variable for each group level of a categorical variable. Thus the levels are replaced by the pattern of response of target variable within that category. By this way the categorical variables are converted into numeric variables. To compute the WOE, SAS® credit scoring’s Interactive grouping node is used.

\[
WOE_i = \log \frac{P(X = x_i | Y = 1)}{P(X = x_i | Y = 0)} \quad \text{for } i = 1, 2, \ldots, I
\]
DATA PREPARATION

The data used in this study is CS_ACCEPTS, a credit screening dataset obtained from SAS Education. The dataset contains 3,000 applicant instances and has 25 variables that were considered important to distinguish credit-worthy customers from non-credit-worthy. The 25 attributes comprises of sixteen categorical and nine continuous variables, where the levels of nominal variables ranges from 2 to 9. The basic requirement for MBR is it requires exactly one target variable, but that can be binary, nominal or interval variable. In our data, GB is the only target variable and it is binary. The variables Profession, Product, Residence ID had missing values and were imputed using the tree surrogate method. The important requirement for MBR is to satisfy the three assumptions of numeric, orthogonal and standardized. Principal Component’s node in SAS® Enterprise Miner is used to generate numeric, orthogonal and standardized variables that would be used as input for MBR modeling.

MODELING

Data is partitioned in the ratio 70:30 prior modeling. Three models are built - MBR with numeric variables only, MBR with numeric and dummy coded categorical variables and MBR with numeric and WOE coded categorical variables. In the MBR node of SAS® Enterprise Miner, we have used the default settings: RD-Tree method to store and retrieve the nearest neighbors and k as 16. As this is a comparative study, same settings are used for all the models.

Figure 3. Types of MBR models used in this study
For Model-1 only continuous variables are used. Model-2 uses both continuous and categorical. The Principal Components node converts the categorical variables to continuous using dummy coding method. In Model-3, there are two stages. Stage 1 is where categorical variables are WOE coded using Interactive Grouping node and thus converted into continuous variables. Stage 2 is where these converted continuous variables along with other continuous variables are fed into Principal Component node prior to MBR modeling. In Model-3, five binary variables: Finloan, Div, Title, Imp_resid and Location and seven nominal variable: Tel, Imp_prof, Car, Regn, Imp_prod, Bureau and Nat are rejected by Interactive Grouping Node because of their low Gini value and Information Value.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Gini Statistic</th>
<th>Information Value</th>
<th>Level for Interactive</th>
<th>Calculated Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>STATUS</td>
<td>22.487</td>
<td>0.203</td>
<td>NOMINAL</td>
<td>Input</td>
</tr>
<tr>
<td>CARDS</td>
<td>17.41</td>
<td>0.135</td>
<td>NOMINAL</td>
<td>Input</td>
</tr>
<tr>
<td>EC_CARD</td>
<td>15.697</td>
<td>0.133</td>
<td>BINARY</td>
<td>Input</td>
</tr>
<tr>
<td>TEL</td>
<td>9.164</td>
<td>0.058</td>
<td>NOMINAL</td>
<td>Rejected</td>
</tr>
<tr>
<td>PROF</td>
<td>9.659</td>
<td>0.056</td>
<td>NOMINAL</td>
<td>Rejected</td>
</tr>
<tr>
<td>CAR</td>
<td>9.455</td>
<td>0.051</td>
<td>NOMINAL</td>
<td>Rejected</td>
</tr>
<tr>
<td>REGN</td>
<td>8.984</td>
<td>0.028</td>
<td>NOMINAL</td>
<td>Rejected</td>
</tr>
<tr>
<td>IMP_PRODUCT</td>
<td>7.36</td>
<td>0.022</td>
<td>NOMINAL</td>
<td>Rejected</td>
</tr>
<tr>
<td>FINLOAN</td>
<td>6.434</td>
<td>0.017</td>
<td>BINARY</td>
<td>Rejected</td>
</tr>
<tr>
<td>DIV</td>
<td>5.025</td>
<td>0.013</td>
<td>BINARY</td>
<td>Rejected</td>
</tr>
<tr>
<td>BUREAU</td>
<td>4.351</td>
<td>0.009</td>
<td>NOMINAL</td>
<td>Rejected</td>
</tr>
<tr>
<td>TITLE</td>
<td>2.928</td>
<td>0.004</td>
<td>BINARY</td>
<td>Rejected</td>
</tr>
<tr>
<td>NAT</td>
<td>2.302</td>
<td>0.004</td>
<td>NOMINAL</td>
<td>Rejected</td>
</tr>
<tr>
<td>IMP_RESID</td>
<td>0</td>
<td>0</td>
<td>BINARY</td>
<td>Rejected</td>
</tr>
<tr>
<td>LOCATION</td>
<td>0</td>
<td>0</td>
<td>BINARY</td>
<td>Rejected</td>
</tr>
</tbody>
</table>

Figure 4. Interactive Grouping Node Statistics

Figure 4 shows that only three categorical variables Status, Cards and EC_Card were selected by Interactive Grouping Node by WOE coding for modeling based on Gini Statistic.

MODEL COMPARISON

To compare model performance of these three models, Model Comparison node is used. The model selection is based on least misclassification rate in the validation dataset. Figure 5 shows the models built for this study using SAS Enterprise Miner.

Figure 5. Models built for this study using SAS Enterprise Miner
In reference to Figure 6, MBR model with WOE coded categorical variables outperforms MBR model with dummy coded categorical variables and MBR model with numeric variables only.

The misclassification rate of Model-3 is less than that of Model-2. This difference indicates the significant improvement in the model performance when categorical variables are WOE coded.

<table>
<thead>
<tr>
<th>Model</th>
<th>Misclassification Rate</th>
<th>KS-Statistic</th>
<th>ROC Index</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBR WOE Coded Categorical Variables</td>
<td>0.344</td>
<td>0.41</td>
<td>0.77</td>
<td>0.572</td>
</tr>
<tr>
<td>MBR with Numeric Variables Only</td>
<td>0.359</td>
<td>0.36</td>
<td>0.75</td>
<td>0.572</td>
</tr>
<tr>
<td>MBR Dummy Coded Categorical Variables</td>
<td>0.382</td>
<td>0.35</td>
<td>0.74</td>
<td>0.552</td>
</tr>
</tbody>
</table>

Figure 6. Model comparison statistics

RESULTS DISCUSSION

- The misclassification rate of MBR model with WOE coded categorical variables is 0.344 and that of MBR model with dummy coded categorical variables is 0.382.
- In other words we can say that there is an increase of 9.94% in model performance when categorical variables are WOE coded instead of dummy coding.
- It is also to note that KS-Statistic, which indicates the separation of two classes of target variable, is greater for MBR model with WOE coded categorical variables than other models.
- Sensitivity of MBR model with WOE coded categorical variables is greater than that of MBR model with dummy coded categorical variables.
- Based on the ROC chart, the area under the curve (AUC) is large for Model-3 when compared to Model-2 which clearly indicates the outperformance of MBR model with WOE coded categorical variables.

Figure 7. ROC charts for model built using SAS Enterprise Miner
CONCLUSION
Two different models, MBR with numeric variables only and MBR with dummy coded categorical variables were used to compare the results of MBR with WOE coded categorical variables. Weight of Evidence coding of categorical variables improves the performance of MBR model significantly. The reasons being elimination of risk in dimensionality increase and the risk of over fitted training data in WOE coding.

REFERENCES

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the authors at:

Vinoth Kumar Raja
Phone: 405-612-5305
E-mail: vinotkr@okstate.edu

Vignesh Dhanabal
Oklahoma State University
Phone: 405-762-3275
E-mail: vignesh.dhanabal@okstate.edu

Dr. Goutam Chakraborty
Oklahoma State University
E-mail: goutam.chakraborty@okstate.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.