Categorical Data Transformation Methods for Neural Networks

Article · January 2008

Source: OAI

CITATIONS
2

READS
3,713

3 authors, including:

Huanjing Wang
Western Kentucky University
48 PUBLICATIONS **781** CITATIONS
SEE PROFILE

Guangming Xing
Western Kentucky University
43 PUBLICATIONS **219** CITATIONS
SEE PROFILE

Categorical Data Transformation Methods for Neural Networks

Huanjing Wang
Department of Computer Science
Western Kentucky University

Guangming Xing
Department of Computer Science
Western Kentucky University

Kairui Chen
Integrated Science & Technology
Marshall University

Abstract

Data mining is the process of analyzing and exploring large dataset from different perspectives in order to extract hidden predictive and useful information - information that can be used to increase revenue, cut costs, or both. There is a need to pre-process the data to make it easier to mine for knowledge. Data preprocessing includes data cleaning, data transformation and data reduction. This study addresses data transformation, which transformed categorical data to numerical data. In this paper, we proposed a data transformation method, information probability, and used neural network to predict motor vehicle injury accident with information probability in traffic safety domain. Experimental results show the significant improvement achieved by the proposed method. Accurate results of such data analysis can be useful for traffic safety engineer or policy maker to set up preventive countermeasure.

General Terms

Algorithms, Performance

Keywords

Data transformation method, variable selection, neural network, motor vehicle injury

1. INTRODUCTION

Traffic crashes kill or injure a lot of people in the world every year. Police investigators collect traffic crash information. In 2006, there were estimated 5,973,000 police-reported traffic crashes in united states; in which 2,575,000 people were injured; an average of 117 people died each day in motor vehicle crashes — one every twelve minutes [6]. The cost of motor vehicle injuries and fatalities has a great impact on society. Researchers are interested in discovering motor vehicle injury patterns. The particular problem domain addressed in this paper is traffic safety domain.

This study used data from CARE (Critical Analysis Reporting Environment), which is an award-winning statistical analysis software developed at The University of Alabama [7]. The dataset selected for the study contains Alabama traffic accident data for the year of 2006. The traffic dataset has 139,780 records, and each record contains 225 categorical variables. A variable named Injury is used as the target variable in this study. Variable Injury only has two attribute values, 0 and 1. The value of 0 corresponds to the non-injury accident ; while the value of 1 corresponds to the injury accident. In the original dataset, 22.61% of cases have output of injury and 77.39% of cases have output of non-injury. According to the variable definitions for the CARE dataset, all data are categorical data. For example, variable “Day of Week” has seven attribute values, “Monday”, “Tuesday”, “Wednesday”, “Thursday”, “Friday”, “Saturday” and “Sunday”. However, Backpropagation Neural Network (BPN), which will be used to predict motor vehicle injuries, relies on numerical input. This paper will address the problem of data transformation, which will transform categorical data to numerical data. The numerical data will feed to Backpropagation Neural Network.

The objective of the study was to develop neural network model that could automatically predict motor vehicle injury. Accurate results of such data analysis can be useful for traffic safety engineer or policy maker to set up preventive countermeasure.

The remainder of the paper is organized as follows. Section 2 reviews existing variable selection techniques, and then uses Sum Max Gain Ratio to reduce the number of variables of original dataset. Section 3 reviews existing data transformation method and proposes a new data transformation method, called information probability. Section 4 reviews the Backprapagation Neural Network and describes the algorithm used in the study. Finally section 5 discusses experimental results, which solve a typical traffic safety problem. The section also presents conclusions of the study.

2. VARIABLE SELECTION

With the information increases explosively, data mining techniques are frequently employed to discover previously unknown, valid patterns and relationships in large datasets. The quality of data affects the success of

data mining on a given learning task. If information is irrelevant or noisy, then knowledge discovery during training time can be ineffective [3]. Variable selection (or feature selection) is a process of keeping only useful variables and removing irrelevant and noisy variables. It is always used as a data mining preprocessing step, particularly for high-dimensional data. The dataset we studied has the characteristic that each record contains 225 categorical variables. The target variable is Injury. It is noted that not all variables are relevant to injury related crashes.

The existing variable selection methods include Chi-squared, Cramer's V Coefficient, Information Gain, Correlation Coefficient, Sum Max Gain Ratio (SMGR) [10], etc. SMGR is strongly correlated with other variable ranking methods and performs well at variable ranking with less runtime cost than other traditional approaches. SMGR is used in this study to remove the unrelated variables, which are not statistically significant to target variable. As a result of performing the variable selection approach, the number of injury related causal variables was reduced to 20. Table 1 lists all variables selected for further study.

Table 1. Selected variables using SMGR

Number	Name
1	INJURIES VEHICLE C
2	HIGHEST OCCUPANT SEVERITY
3	ACCIDENT SEVERITY
4	NUMBER OF PEDESTRIANS
5	AMBULANCE ARRIVAL DELAY
6	CAUSAL VEHICLE CATEGORY
7	INJURIES VEHICLE 2
8	SAFETY EQUIPMENT - DRIVER C
9	FIRST HARMFUL EVENT
10	CONDITION - DRIVER C
11	TYPE TEST GIVEN - DR C
12	PRIME HARM EVENT - DR C
13	DAMAGE SEVERITY VEH C
14	TOWED VEHICLE C
15	SAFETY EQUIPMENT, DRIVER 2
16	TYPE TEST GIVEN, DR 2
17	PRIME HARM EVENT, DR 2
18	HAZARDOUS CARGO, VEH 2
19	CONTRIB DEFECT, VEH 2
20	UNIT NUMBER, UNIT 2

3. DATA TRANSFORMATION

All variables in the traffic dataset are categorical variables, a variable that has mutually exclusive ("named") groups that lack intrinsic order. Categorical variables classify data into categories. Categorical variables may be described as nominal, ordinal, interval, or ratio. Nominal variables have attribute values that have no natural order to them (e.g., County – Tuscaloosa, Mobile, Walker, etc.). Ordinal variables do have a natural order but not the difference between values. (e.g., Letter Grade – A, B, C, D, F). Interval variables are created from intervals on a contiguous scale, the difference between two values is meaningful. (e.g., Age of Driver – 22-24, 25-34, 35-44, etc). A ratio variable, has all the properties of an interval variable, but also has a clear definition of 0.0 (e.g., Weight).

No matter what kind of categorical data, numbers are assigned to the different categories for each variable in the traffic dataset. For example, variable "Event Location" has seven attribute values: "Null", "On roadway", "Off roadway", "Median", "Driveway", "Private road/property", and "Intersection". We might assign a "0" for "Null", a "1" for "On roadway", a "2" for "Off roadway", a "3" for "Median", a "4" for "Driveway", a "5" for "Private road/property", and a "6" for "Intersection".

The traffic dataset will feed to a Backpropagation Neural Network whose input is in the range of 0 to 1. Therefore, a data transformation method is needed to transform categorical data to numerical data.

The goal of the data transformation is to change data into another form. The problem of transformation method can be defined as follows. Let x be an original data and y be a transformed data. The data transformation method f can be expressed as:

$$y = f(x) \quad (1)$$

The transformation comes in many forms: Discretization, the process of putting values of a continuous set of data into a discrete number of possible states; Scaling, the process of changing one form of numerical data into another form of numerical data; and coding, the process of transforming categorical data to numerical data. The popular coding methods are 1-to-N binary encoding [8], the thermometer encoding [8], and frequency based encoding (FBE) [5]. The formula for frequency based encoding is described as:

$$FBE(V_{k,i}) = \frac{f_{i,1}}{f_{i,0} + f_{i,1}} \quad (2)$$

Where $f_{i,0}$ is the number of non-injuries occurred with attribute value i of variable V_k
$f_{i,1}$ is the number of injuries occurred with attribute value i of variable V_k

The frequency based encoding only considers frequency for each individual attribute value. In real world application, one attribute value of one particular variable may occur a few times, while another attribute value of the same variable may occur thousand of times. They may have same transformed data. Assume variable V_k has 6 attribute values, $V_{k,0}$, $V_{k,1}$, $V_{k,2}$, $V_{k,3}$, $V_{k,4}$ and $V_{k,5}$. Computation results are shown in table 2. Attribute value $V_{k,1}$ and $V_{k,2}$ have same frequency based encoding. Figure 1 displays the histogram for frequency based encoding.

Table 2. Example of frequency based encoding

Variable V_k:	**Injury**	**Non-injury**	**Frequency based encoding**
$V_{k,0}$	2	0	1.0
$V_{k,1}$	5	5	0.5
$V_{k,2}$	1000	1000	0.5
$V_{k,3}$	1000	0	1.0
$V_{k,4}$	50	50000	0.001
$V_{k,5}$	0	3995	0

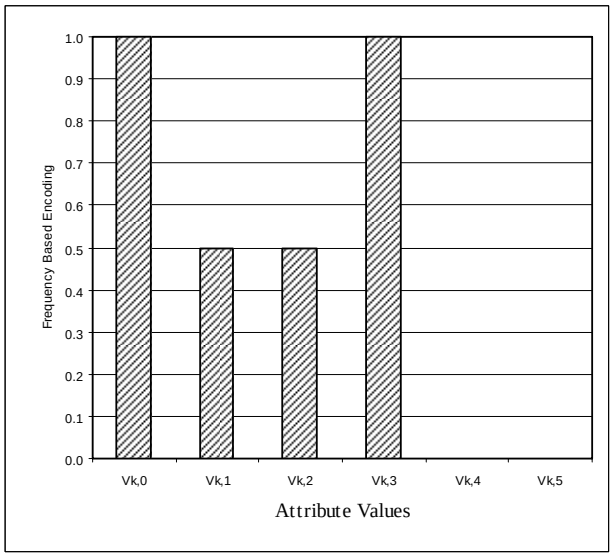

Figure 1. Histogram for frequency based encoding

Given traffic categorical dataset, instead of using existing data transformation methods, a new data transformation method called information probability is defined as below. We first compute probability difference (PD):

$$PD(V_{k,i}) = \frac{f_{i,1}}{N_1} - \frac{f_{i,0}}{N_0} \quad (3)$$

Where $f_{i,0}$ is the number of non-injuries occurred with attribute value i of variable V_k
$f_{i,1}$ is the number of injuries occurred with attribute value i of variable V_k
N_0 is the total number of non-injuries
N_1 is the total number of injuries

Equation 3 can be rewritten to equation 4

$$PD(V_{k,i}) = (f_{i,1} - N_1 \times \frac{f_{i,0}}{N_0}) / N_1 \quad (4)$$

$f_{i,1} - N_1 \times \frac{f_{i,0}}{N_0}$ can be explained as Max Gain[7]. Max Gain is used to express the number of cases that could be reduced if the subset frequency (experimental subset, injury) was reduced to its expected value (control subset, non-injury). Information probability becomes the probability of potential number of cases that could be reduced if subset frequency (injury) was reduced to its expected value (non injury). The range of information probability is between -1 and 1. The input of neural network is between 0 and 1. The probability difference is converted to information probability.

$$IP(V_{k,i}) = \frac{PD(V_{k,i}) + 1}{2} \quad (5)$$

Table 3 shows the information probability of each attribute value for the variable V_k. It is noted that information probability is different even if frequency based encoding is same. For example, Attribute value $V_{k,1}$ and $V_{k,2}$ have different information probability, but different frequency based encoding. Figure 2 displays the histogram for information probability.

Table 3. Example of information probability

Variable V_k:	**Injury**	**Non-injury**	**Information probability**
$V_{k,0}$	2	0	0.5
$V_{k,1}$	5	5	0.501
$V_{k,2}$	1000	1000	0.734
$V_{k,3}$	1000	0	0.743
$V_{k,4}$	50	50000	0.058
$V_{k,5}$	0	3995	0.464

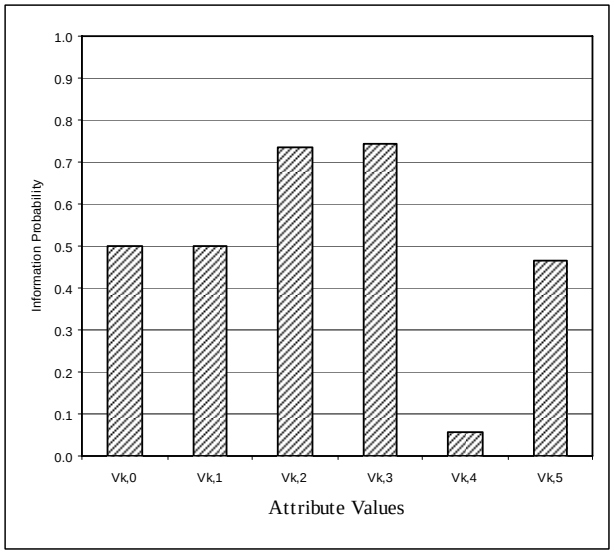

Figure 2. Histogram for information probability

Implementation of the above frequency based encoding and information probability can be illustrated by the following procedure.

```
1:  Procedure DataTransformation(Data, r, c)
2:  Input: two dimensional array: Data, data has r rows
    and c columns, the last column of array Data is target
    variable Injury
3:  Output: arrays DataFBE, DataIP contain transformed
    data
4:  N0 = 0
5:  N1 = 0
6:  max = 0
7:  for j=1 to c-1 do
8:      for i=1 to r do
9:          if (Data[i][j] > max)
10:             max = Data[i][j]
11:         endif
12:         if Data[i][c] = 0
13:             X0[j][Data[i][j]] = X0[j][Data[i][j]] + 1
14:             if j = 1
15:                 N0=N0+1
16:             endif
17:         else
18:             X1[j][Data[i][j]] = X1[j][Data[i][j]] + 1
19:             if j = 1
20:                 N1=N1+1
21:             endif
22:         endif
23:     endfor
24: endfor
25: for j=1 to c-1 do
26:     for t=0 to max
27:         if X0[j][t]>0 or X1[j][t]>0
28:             FBE[j][t] = X1[j][t]/(X0[j][t]+X1[j][t])
29:             IP[j][t] = (X1[j][t]/N1 – X0[j][t]/N0+1)/2
30:         endif
31:     endfor
32: endfor
33: for j=1 to c-1 do
34:     for i=1 to r do
35:         DataFBE[i][j] = FBE[j][Data[i][j]]
36:         DataIP[i][j] =IP[j][Data[i][j]]
37:     endfor
38: endfor
```

Figure 3. Procedure to Perform Data Transformation

4. NEURAL NETWORK LEARNING

Data mining, also known as knowledge discovery, is the exploration and analysis of a large dataset in order to extract hidden predictive and useful information. Data mining tasks include prediction, clustering, classification, description, etc. We focused on classification in this study. Classification consists of finding a function that predicts one or more discrete variables based on the other variables in the dataset. Existing techniques for classification include decision trees [9], nearest neighbor [1], neural network [2], and so on.

An artificial neural network, or neural networks, is an information processing system that is inspired by the way biological nervous systems (such as the brain) process information. Neural networks can be used to predict and extract patterns (such as motor vehicle crash pattern). There are several architectures and learning algorithms for neural networks. We will use Backpropagation Neural Network [2, 4] in the study. A Backpropagation Neural Network learns by case, that is, we must provide a training set that consists of some input examples and the target (injury in this study) output for each case (record).

The major elements of a Backpropagation Neural Network include inputs (each input corresponds to a single variable), outputs (solution to the problem), connection weights, summation function and activation function. The learning algorithm includes the following procedures:

(1) The network weights are initially set to random value, learning rate, the number of hidden layer and the number of nodes for each hidden layer.
(2) Read in the input data and target output.
(3) Compute the actual net output via the calculations, working forward through the layers. This study used the activation function of:

$$O(x) = \frac{1}{1+e^{-4x+2}} \qquad (6)$$

(4) Compute error and change the weights by working backward from the output layer through the hidden layer.

5. EXPERIMENTAL RESULTS AND CONCLUSION

We used Backpropagation Neural Network to evaluate the error rate on predicting motor vehicle injury for information probability and frequency based encoding. The original data set has 225 categorical variables. Injury is selected as target variable. The remaining 224 variables are reduced to 20 variables using SMGR described in section 2. The new data set only has 21 variables. The neural network has three layers, input layer, one hidden layer and output layer. The number of input layer has 20 nodes; the hidden layer has 10 nodes (average of input nodes and output nodes) and the output layer has 1 node. The training dataset contains 2000 records selected from 2006 Alabama crash data. The training dataset is used to train network. The test dataset contains 3000 records and is used to evaluate the trained network. Table 4 shows the error rate and the average square of deviation. The square of deviation is defined as:

$$SD(V_j) = (D_j - O_j) \times (D_j - O_j) \quad (7)$$

Where D_j is the target output for record j

O_j is the actual net output for record j

Table 4. Error rate and average square of deviation

Data Transformation	Error rate	Average square of deviation
Information probability	4.2%	0.001229
Frequency based encoding	9.3%	0.01317

As seen in the experimental results, the neural network with information probability has lower error rate (5.1% more accurate than frequency based encoding) and lower average square of deviation. Experimental results show the significant improvement achieved by the proposed method. It is noted that the data transformation plays an important role in prediction accuracy. The proposed data transformation method enables the Backpropagation Neural Network to be trained efficiently. Neural network is very powerful and efficient. Further study may involve combining neural network with other data mining techniques, such as decision trees and genetic algorithms. We will use the combination techniques to analyze traffic accident data to save lives in future.

6. REFERENCES

[1] Berry, M. and Linoff, G., *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons, 1997.

[2] Freeman, J. and Skapura, D., *Neural Networks: Algorithms, Applications, and Programming Techniques*, Addison-Wesley, 1991.

[3] Hall, M. A. and Smith, L. A., *Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper*, Proceedings of the Florida Artificial Intelligence Symposium, 1999.

[4] Haykin, S., *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999.

[5] Kauderer H. and Mucha H., *Supervised Learning with Qualitative and Mixed Attributes. Classification, Data Analysis, and Data Highways*, Springer-Verlag Press, pp. 374-382, 1997.

[6] NHTSA, *Traffic Safety Facts*, 2006, www.nhtsa.gov

[7] Parrish, A., Dixon, B., Cordes, D., Vrbsky, S. and Brown, D., *CARE: A Tool to Analyze Automobile Crash Data*, IEEE Computer, 36(6), pp.22-30, 2003.

[8] Pyle, D., *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, 1999.

[9] Quinlan, J. R., *Introduction of Decision Trees*, Machine Learning, pp. 81-106, 1986.

[10] Wang, H., Parrish, A., Smith, R. and Vrbsky, S, *Improved Variable and Value Ranking Techniques for Mining Categorical Traffic Accident Data*, Expert Systems with Applications, 29(4), December 2005.