A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers

Article *in* International Journal of Computer Applications · October 2017

DOI: 10.5120/ijca2017915495

A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers

Kedar Potdar
Student, Computer Engineering
WIEECT, Worli
Mumbai, INDIA, 400 018

Taher S. Pardawala
Student, Computer Engineering
WIEECT, Worli
Mumbai, INDIA, 400 018

Chinmay D. Pai
Student, Computer Engineering
WIEECT, Worli
Mumbai, INDIA, 400 018

ABSTRACT

In classification analysis, the dependent variable is frequently influenced not only by ratio scale variables, but also by qualitative (nominal scale) variables. Machine Learning algorithms accept only numerical inputs, hence, it is necessary to encode these categorical variables into numerical values using encoding techniques.

This paper presents a comparative study of seven categorical variable encoding techniques to be used for classification using Artificial Neural Networks on a categorical dataset. The Car Evaluation dataset provided by UCI is used for training. Results show that the data encoded with Sum Coding and Backward Difference Coding technique give highest accuracy as compared to the data pre-processed by rest of the techniques.

Keywords

Machine Learning, Statistical Learning, Artificial Neural Networks, Data Preprocessing.

1. INTRODUCTION

A good understanding of data is essential for accurate analysis. Before proceeding to the actual analysis, the data is processed to aid algorithms and to improve efficiency. Data variables generally fall into one of the four broad categories: nominal scale, ordinal scale, interval scale, and ratio scale. [1]

Variables on nominal scale have no quantitative value, i.e. they are purely qualitative variables. They are categories or classifications. Nominal measurement is using categorical levels of variables. Examples of nominal data are variables such as gender (male, female) and marital status (married, unmarried, divorced, separated) simply denote categories. [2]

Ordinal refers to order in measurement. An ordinal scale indicates direction, in addition to providing nominal information. For these variables the ordering exists but the distances between the categories cannot be quantified. Low/Medium/High; or Faster/Slower are examples of ordinal levels of measurement. [3]

Interval scales provide information about order, and also possess equal intervals. An example of an interval scale is temperature, either measured on a Fahrenheit or Celsius scale. A degree represents the same underlying amount of heat, regardless of where it occurs on the scale.

In addition to possessing qualities of nominal, ordinal and interval scales, a ratio scale has absolute zero. Using ratio scale permits comparisons between values of variables.

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can change when exposed to new data. [4] Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. [5]

In both, regression and classification analysis, categorical variables are widely used. However, Machine Learning algorithms accept only numeric values as input. To use categorical data for Machine Learning purposes, the data needs to be encoded into numeric values such that each categorical feature is represented with a number. [6]

In this paper, a comparative study of seven categorical variable encoding techniques to be used for classification using Artificial Neural Networks (ANN) on a categorical dataset is presented. The data is sourced from the ‘Car Evaluation Data Set’, made available by the University of California: Irvine.

In Section II, the categorical encoding techniques to be used are described in-depth. Section III describes the Machine Learning technique of ‘Artificial Neural Networks’ that is employed in this study. Section IV describes in brief the dataset used and its attributes. Section V covers the methodology adopted to perform this experiment. Section VI describes the results and analysis of our experiment and Section VII concludes it.

2. CATEGORICAL VARIABLE ENCODING TECHNIQUES

2.1 One Hot Coding

One Hot Coding is the most widely used coding scheme. It compares each level of the categorical variable to a fixed reference level. One hot encoding transforms a single variable with n observations and d distinct values, to d binary variables with n observations each. Each observation indicating the presence (1) or absence (0) of the dichotomous binary variable. [7]

2.2 Ordinal Coding

In ordinal encoding, an integer is assigned to each category, provided the number of existing categories are known. It does not add any new columns to the data, but implies an order to the variable that may not actually exist. [8]

2.3 Sum Coding

Sum coding compares the mean of the dependent variable for a given level to the overall mean of the dependent variable over all the levels. That is, it uses contrasts between each of the first $k - 1$ levels and level k in this example, level 1 is compared to all the others, level 2 to all the others, and level 3 to all the others. [9]

2.4 Helmert Coding

Helmert Coding compares each level of a categorical variable to the mean of the subsequent levels.

2.5 Polynomial Coding

Polynomial coding is a form of trend analysis that looks for linear, quadratic and cubic trends in the categorical variable. This type of coding system should be used only with an ordinal variable in which the levels are equally spaced.

2.6 Backward Difference Coding

In this coding system, the mean of the dependent variable for one level of the categorical variable is compared to the mean of the dependent variable for the prior adjacent level.

2.7 Binary Coding

In binary coding, first the categories are encoded as ordinal, then those integers are converted into binary code, then the digits from that binary string are split into separate columns. [2]

3. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANN) is a group of interconnected nodes that uses a computational model for information processing. It changes structure based on external or internal information that flows through the network. ANN can be used to model a complex relationship between inputs and outputs and find patterns in data. [3] The output of ANN is determined by characteristics of the features and the weights associated with the interconnections among them. The connections between nodes are modified in the training process to adapt the network to desired outputs. [7]

The neural network gains the experience initially by training the system to correctly identify pre-selected examples of the problem. The response of the neural network is reviewed and the configuration of the system is refined until the neural networks analysis of the training data reaches a satisfactory level. In addition to the initial training period, the neural network also gains experience over time as it conducts analyses on data related to the problem. Classification using ANN is one of the most dynamic research and application areas. ANN is widely used for classification purposes because of its ability to generalize and map input-output relations based on existing data. [10]

4. CAR EVALUATION DATASET

The Car Evaluation dataset used for this study was obtained from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). It was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, M. Bohanec, V. Rajkovic: Expert system for decision making. [12]. The data includes information of 1728 cars used for the study. Each record has 7 attributes, which are described in Table 1. [11]

Table 1. Attribute names and the values

Attribute Name	Attribute Value
Buying	v-high (very high), high, med, low
Maint	v-high (very high), high, med, low
Doors	2, 3, 4, 5-more (5 or more)
Persons	2, 4 , more (4 or more)
Lug_Boot	small, med, big
Safety	low, med, high
Car	unacc (unacceptable), acc (acceptable), good, vgood (very good)

4.1 Car

The Car attribute is the dependent variable that defines the acceptability of a given car. It is based on 4 attribute values from unacceptable (unacc) to very good (vgood).

4.2 Buying

The Buying attribute defines the cost of buying or the buying. price for a given car. It has 4 attribute values from very high (v-high) to low where low defines cars that cost less and high is for cars with high buying price.

4.3 Maint

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

4.4 Doors

The Doors attribute defines the number of doors that are present in a given car. It has 4 attribute values ranging from 2 to 5 or more (5-more) where 2 defines 2 doors and 5-more is for 5 or more doors.

4.5 Persons

The Persons attribute defines the capacity in terms of persons to carry by a given car. It has 3 attribute values ranging from 2 to more where 2 defines a capacity of 2 persons and more defines a capacity of 4 or more.

4.6 Lug_Boot

The Lug Boot attribute defines the size of luggage boot for a given car. It has 3 attribute values from small to big where small defines a luggage boot of small size and big defines a luggage boot which is big in size.

4.7 Safety

The Safety attribute defines the estimated safety of the car for any given car. It has 3 attribute values from low to high where low defines a car with low estimated safety and high defines a car with high estimated safety. [13]

5. METHODOLOGY

The current research was developed in three stages: variable encoding, ANN modeling and, finally, analysis and comparison between the 7 categorical encoding techniques.

Table 2. Encoding techniques and the number of input columns

Encoding Technique	Number of Input Columns
One Hot Coding	21
Ordinal Coding	6
Sum Coding	21
Helmert Coding	21
Polynomial Coding	21
Backward Difference Coding	21
Binary Coding	12

5.1 Categorical Variable Encoding

The Car Evaluation data is encoded using the 7 encoding techniques mentioned above. The original dataset had 6 input attributes and 1 output attribute.

5.1.1 Encoding of Input Attributes

After encoding, the number of columns change as per the encoding technique used. Table 2 lists the dimensionality of the dataset after each encoding technique is applied.

5.1.2 Encoding of Output Attributes.

The neural network to be trained was used for classification task. The output attribute, 'car', takes 4 values viz. unacc (unacceptable), acc (acceptable), good and vgood (very good) indicating the acceptability level of the car. This nominal data also needs to be converted into numeric data to train the neural network model.

The one-hot encoding technique was used to encode the 'car' attribute resulting in 4 columns. In this encoding, for any particular output level, the corresponding attribute is 1 and rest all are 0s.

Thus, a total of 7 datasets were formed corresponding to the 7 encoding techniques. These were then used to train a neural network model for the classification task.

5.2 Artificial Neural Network Modeling

In this study, the Backpropagation algorithm was used to build neural networks for classification task on the Car Evaluation dataset. Seven neural networks were trained for the 7 datasets produced in the encoding phase. Depending on the encoded inputs, the network classified the car into any one of the four output levels. The Lavenberg-Marquardt backpropagation algorithm was used with training for 1000 epochs and learning rate of 0.001.

The data is randomly divided into training, validating and testing data. Accordingly, 70% of data is used for training, 15% for validation and the remaining 15% for testing.

6. RESULT AND ANALYSIS

Classification accuracy of the 7 neural network model was tested and is show in Table 3. As seen, the Sum coding and Backward Difference coding techniques, both, result in a high accuracy of 95%. Furthermore, both these techniques lead to dimensionality of 21 columns each, after encoding. The polynomial encoding technique results in next highest accuracy of 91%, with a dimensionality of 21 columns. Ordinal encoding results in least accuracy of 81%. However, it does not generate any new columns during the encoding process, leading to the final dataset having 6 columns as well.

Table 3. Encoding technique and the accuracy percent

Encoding Technique	Accuracy (Percentage)
One Hot Coding	90
Ordinal Coding	81
Sum Coding	95
Helmert Coding	89
Polynomial Coding	91
Backward Difference Coding	95
Binary Coding	90

7. CONCLUSION

This paper demonstrated and compared the classification accuracy of ANN models applied to categorical data encoded using 8 different encoding techniques. The aim of this study was to find out the most accurate encoding technique for the Used Car dataset. According to prediction results, the Sum Coding and Backward Difference Coding techniques outperform the rest, with an accuracy of 95%. Though not an exhaustive study, the Sum Coding and Backward Difference Coding can be considered more preferable for prediction tasks involving purely categorical data like the Used Car dataset.

8. REFERENCES

[1] "Types of Data & Measurement Scales.", MyMarketResearchMethods (n.d) Retrieved July 2017, from www.mymarketresearchmethods.com/types-of-data- nominal-ordinal-interval-ratio/.

[2] N Gujarati, Damodar, "Basic econometrics.", The McGraw Hill, 2004

[3] K. Potdar and R. Kinnerkar, "A Non-linear Autoregressive Neural Network Model for Forecasting Indian Index of Industrial Production", Proceedings of the IEEE Tensymp 2017, Kochi, India

[4] "What is Machine Learning.", WhatIs (June 2017) Retrieved July 2017, from whatis.techtarget.com/definition/machine- learning.

[5] "Evolution of Machine Learning.", SAS (n.d) Retrieved July 2017, from www.sas.com/en_us/insights/analytics/machine-learning.html.

[6] Gregory Carey, (2003) Coding Categorical Variables, Retrieved July 2017, from psych.colorado.edu/čarey/courses/psyc5741/handouts/Coding %20Categorical%20Variables%202006-03-03.pdf.

[7] Brett Lantz, "Machine Learning with R", Packt Publishing Limited, 2013. ISBN - 978-1782162148.

[8] Von Eye, Alexander, and Clifford C. Clogg, eds. "Categorical variables in developmental research: Methods of analysis." Elsevier, 1996.

[9] "R Library Contrast Coding Systems for Categorical Variables", UCLA (n.d) Retrieved July 2017, from stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/

[10] Saravanan K and S. Sasithra, "REVIEW ON CLASSIFICATION BASED ON ARTIFICIAL NEURAL NETWORKS", International Journal of Ambient Systems and Applications (IJASA) Vol.2, No.4, December 2014.

[11] M. Bohanec and V. Rajkovic, "Knowledge acquisition and explanation for multi-attribute decision making." In 8th Intl. Workshop on Expert Systems and their Applications, Avignon, France. pages 59-78, 1988.

[12] B.Zupan, M.Bohanec, I.Bratko, J.Demsar, "Machine learning by function decomposition." ICML-97, Nashville, TN. 1997 (to appear).

[13] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.