IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 8, August 2016

Effectiveness of Variable Transformation in Customer Marketing Using Classification Model

Hyeuk Kim

Assistant Professor, Department of Applied Statistics, Hoseo University, Asan, Korea

Abstract: We usually focus on how to apply data mining techniques to data. However, it is important to manipulate data such as cleaning observations, dealing with missing values, and transforming variables. The pre-processing and transformation steps in KDD affect the performance of a model much. We compare the performances of various classification methods based on the different data set in the paper. Several data sets are generated from a real bank data. The differences in data sets are which variables are transformed. The results show that variable transformation improves the performance of a model. Additionally, the more transformations for the appropriate variables happens the more improvement in the performance of a model.

Keywords: Classification, variable transformation, customer marketing, performance comparison, neural networks

I. INTRODUCTION

Data mining is one of the procedures in KDD. KDD is the abbreviation for knowledge discovery in databases [1]. KDD is the process to find the pattern which is logical, new, useful, and understandable from data. The KDD process consists of selection, pre-processing, transformation, data mining, and interpretation which are described below.

Step 1: Selection transforms data into target data.
There is too much information in real world. Therefore, we need to construct the data set with relevant information. The irrelevant information is ignored even though we collect.
Step 2: Pre-processing transforms target data into pre-processed data. The data set is usually nasty and unorganized unlike the data set in a textbook. It has to pass data cleaning and pre-process step. Noise is erased and the strategy for missing values is determined.
Step 3: Transformation transforms pre-processed data into transformed data. The third step handles data reduction and projection. It is about manipulating variables such as dimensionality reduction and variable transformation which is an essential point in the paper.
Step 4: Data mining extracts patterns from transformed data. Data mining is the analysis of data with large size in order to find patterns and rules [2]. There are four major techniques in data mining. A prediction is to predict an event which will occur in future. A classification is to classify instances into one of the predefined classes.

An estimation is to evaluate events and assign the score to each event. The score is commonly the probability that the corresponding event will happen. Last, clustering is to group instances into one of several clusters. The instances in the same cluster are similar to each other and the instances in the different clusters are not similar to each other.

Step 5: Interpretation or evaluation finds knowledge from patterns. We interpret the patterns which is found through the previous steps. Visualization is an important tool in the step. It is becoming more and more important as big data is widely available. The knowledge from the patterns are used in many ways. It is used directly, or it is merged into another knowledge for other system. Also, we make a decision for a certain problem from the knowledge. The KDD is not a singular and one-way procedure. The iteration of the procedure is common.

Most of researchers and practitioners focus on data mining step. Many techniques have been developed and applied to various data. However, the steps before data mining are more important in a certain view. There is the report that practitioners use more than 50% work time for cleaning data, handling missing values, and manipulation of variables in data.
The outcomes become totally different based on what pre-processing and transformation steps are performed even though the same data mining technique is applied to the same data set. ‘Garbage in, garbage out’ rule is directly applied in data mining since its performance is especially data-dependent.

The structure of the paper is as follows. In the second section, we describe how to handle missing values and outliers and how to transform variables. We focus on an appropriate transformation of variables since the aim of the paper is to show the importance of transforming variables.
We introduce several classification methods and performance evaluation measures which are used to real data in the third section. We compare the performances of the different variable transformations with a real bank marketing data in the fourth section. In the last section, we make a conclusion.

DOI 10.17148/IJARCCE.2016.58118

II. PRE-PROCESSING AND TRANSFORMATION

Data in real world is commonly not organized and the variables in data are inconsistent. It happens because of the bias from an observer, the error by an operator, or the error of measurement tools. It has to be fixed since it affects the performance badly. The easiest approach for handling missing values is to delete it. It is the simple method, but the disadvantage of this approach is that there is a huge loss of data since we have to delete the observation if it has just one missing variable in the corresponding observation. The general approach for dealing with missing values is to impute the missing data with replacement values. An average value for the corresponding variable is the simplest replacement value for the missing value. An outlier is an observation that is far from other observations.

We search for the cause of an outlier carefully. It occurs because of the characteristics of data even though it is rare that the observation appears. However, it happens due to a measurement error in many cases. An outlier has to be manipulated since it worsens the performance of the model. We do not define an outlier on experiment in the paper. We reduce the effects of outliers by variable transformation since it is hard to delete an outlier under the exact criterion. A variable is not always transformed. It is transformed when the distribution of the variable is not a common shape. The performance of the model is good when the distributions of its variables are a symmetric and normal-like shapes. One example is to transform the variable by applying a square root or a logarithm when its distribution is skewed.

III.CLASSIFICATION METHODS

Classification is the method of learning a function to predict for a test set by classifying predefined classes. A training set is used to induce learning procedure. The purpose of classification is to learn the model with minimum of an error. A classification method usually works well for data set of which target variable is binary or nominal, and it performs badly when the type of a target variable is an ordinal categorical data [3].
Many novel classification methods have been developed so far. Among them, we describe the methods which are used for an experiment in the paper.

A. Logistic regression

We cannot use a linear regression model when a target variable is binary. A linear regression model for the data with a binary target variable has several problems. One problem is that the value of a target variable is binary, but the value of a predicted target variable is not binary. Another problem is the distribution for a target variable does not correspond to the proposed distribution for a target variable in a linear regression model. The distribution for a target variable follows Bernoulli distribution when a target variable is binary, but a target variable has a normal distribution since a target variable is supposed to be continuous A logistic regression is the logit transformation of a linear regression model to overcome the above problems.

$$\log \frac{P(y=1|x_1,x_2,\dots,x_p)}{1-P(y=1|x_1,x_2,\dots,x_p)} = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$$

This is the general formula for a logistic regression. Both left term and right term have real values. We extract the estimated equation for a posterior probability from the previous formula.

$$\hat{P}(y=1|x_1,x_2,\dots,x_p) = \frac{\exp(a+b_1x_1+\cdots+b_px_p)}{1+\exp(a+b_1x_1+\cdots+b_px_p)}$$

The acquired posterior probability is used to classify observations into predefined classes. A posterior probability has the value between 0 and 1, so the corresponding posterior probability is compared with an appropriate cut-off value.

B. Gradient boosting

J. H. Friedman has developed a gradient boosting model [4]. It is an ensemble of single classifiers such as decision trees. It follows a boosting algorithm in constructing classifiers and minimizes the values for various loss functions. Its performance is generally very high among the classification methods.

Calibrated boosted trees which belongs to the type of gradient boosting, even though they are a little different from gradient boosting, shows the best performance over other classification algorithms overall [5]. An extended research also supports the results in [5]. Boosted decision trees perform best when the dimension of the data set is up to 4000 [6].

C. Decision tree

It is one of the basic algorithms in classification. CART [7] is the most popular algorithm in decision trees. It has been developed in 1984. It is the abbreviation for classification and regression trees. There are various types of decision trees. CART uses Gini index for calculation of an impurity, and C4.5 algorithm [8] uses an entropy index in calculating an impurity.

C4.5 algorithm is a successor of ID3 (Iterative Dichotomiser 3) [9]. Both algorithms have been developed by J. R. Quinlan. CHAID stands for chi-squared automatic interaction detection and uses chi-square for an impurity.

A decision tree has several advantages. It builds the rules that are understood easily. It can handle both continuous and categorical variables and shows the degree of the importance for each variable. However, a decision tree performs badly when a target variable is continuous or the analysed data belongs to time-series data.

DOI 10.17148/IJARCCE.2016.58118

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 8, August 2016

D. Neural network

A neural network is the classification method that is inspired from biological nervous systems. It consists of many interconnected neurons. The neurons in the training mode is trained for particular input rules. The neurons in the using mode make a decision when input information matches a certain rule which is defined in the training phase. It searches for the parameters which show the optimal performance in estimation. The parameters are updated by the iterative process for all observations. The advantage of a neural network is its performance. Its performance is known to be good even though no free lunch theorem [10] is applied. However, it cannot figure out the relationship between the input variables and the target variable unlike a decision tree. The model with the above status is called a black-box model. There is a possibility for a neural network algorithm to model overfitting and it takes long time for learning due to the complexity of computation.

We evaluate the performance of the classification model after it is learned from a training data. The misclassification rate is a fundamental measure for evaluation, but there are a large number of performance evaluation measures. Many measures are developed since we have to consider several factors for evaluating the performances of the classification methods. We introduce five measures for evaluating the performance of the classification algorithmsbriefly.

E. Misclassification rate

A confusion matrix is the essential tool in evaluation of the classification models.

		Predicted Class	
		positive	Negative
Actual Class	positive	n_{++}	n_{+-}
	negative	n_{-+}	n_{--}

Table 1 A confusion matrix

We derive many measures from a confusion matrix. The misclassification rate is the basic ratio whose equation is as follow.

$$\text{Misclassification Rate} = \frac{n_{+-} + n_{-+}}{n_{++} + n_{+-} + n_{-+} + n_{--}}$$

Most of measures have high values if the classification method works well, but the low value of a measure means the good performanceof a model for a misclassification rate.

F. AUC

The full abbreviation is AUROC (Area Under the Receiver Operating Characteristic Curve). AUC (Area Under the Curve) denotesthe area under the curve on ROC chart. Therefore, the classification model whose performance is expressed by a red line works best among four models in Figure 1.

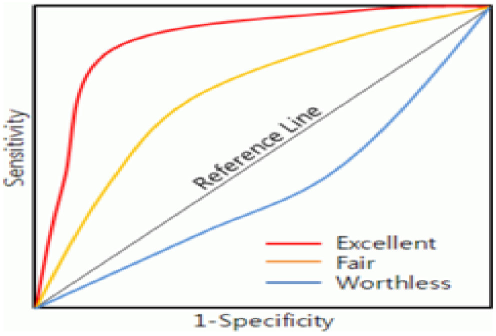

Fig. 2 ROC chart

A horizontal axis denotes 1-specificity and a vertical axis denotes sensitivity in ROC chart. The classification model whose AUC is 1 works perfectly since its specificity is one and its sensitivity is also one. A reference line means the model whose decision is determined randomly, so the model whose curve is below a reference line or AUC is less than 0.5 stands that it is worthless.

G. Kolmogorov-Smirnov statistic

The follow two measures are not derived from a confusion matrix. Kolmogorov-Smirnov statistic is a nonparametric measurewhich denotes the distance between a specific distribution and a reference distribution. The distribution of a reference function means the performance of a random decision. Therefore, the classification model works better as the Kolmogorov-Smirnov statistic is larger.

H. Top 10% cumulative response rate

It is a top 10% percentile for the cumulative response rate. We are especially interested in the performance of a part of customers in marketing. Therefore, a top 10% cumulative response rate or a top 20% cumulative response rate are generally used. Additionally, a cumulative gain or lift chart is used frequently. It is the cumulative response rate divided by the overall response rate.

IV.EXPERIMENT

The data is related with direct marketing campaigns of a Portuguese bank. Phone calls are used as a tool for direct marketing campaigns [11]. The aim is to construct the predictive model that determines whether a client subscribes a term deposit or not. We do not perform the direct marketing campaigns to all clients. We make phone calls to the customers who have a high probability to subscribe a term deposit. Therefore, a banker is able to save time and expense.

Subject	Value
The number of observations	45,211
The number of the input variables	16
Time range	May 2008 - Nov. 2011

Table 2 the description of a Portuguese bank data

DOI 10.17148/IJARCCE.2016.58118

We investigate observations for pre-processing. We do not need to consider how to handle missing values since there are no missing values in the data set. It is arbitrary whether an observation is an outlier or not. The careless decision for outlier worsen the performance of a classification method. So, we do not delete outliers and transform the variables to minimize the effects of outlies even though there are some outliers.

Table 2 describes the 16 input variables in the data set.

Name	Description	Data type
Age	Age	Integer
Job	Type of job	Categorical data
Martial	Martial status	Categorical data
Education	Education	Categorical data
Default	Credit status (default)	Binary data
Balance	Average bank balance	Integer
Housing	Housing loan	Binary data
Loan	Personal loan	Binary data
Contact	Contact communication type	Categorical data
Day	Last contact day of year	Integer
Month	Last contact month of year	Categorical data
Duration	Last contact duration, in sec	Integer
Campaign	Number of contacts performed during this campaign	Integer
Pdays	Number of days that passed by after the client was last contacted from a previous campaign	Integer
Previous	Number of contacts performed before this campaign	Integer
Poutcome	Outcome of the previous marketing campaign	Categorical data

Table 2 the description of the input variables in Portuguese bank data

We explore the distribution of each variable and transform the corresponding variable if needed. SAS Enterprise Miner is used for variable transformation and applications of the classification methods.

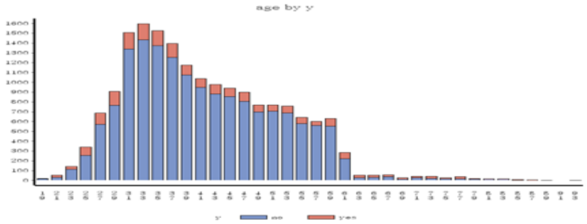

Fig. 2the distribution of Age variable

Age variable denotes the customer's age and its range is between 19 and 93. It is transformed since the frequencies between 25 and 61 are very high. We make four categories and classify the customers' ages into one of four categories. The distribution after transformation is described in Figure 3.

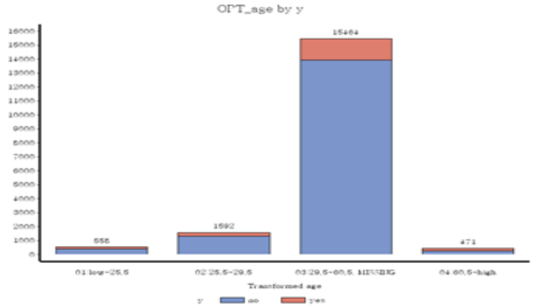

Fig. 3 the distribution of Age variable transformed

Balance variable is the average bank balance of the customer and it is from a negative value to very high value since the customers' economic situations are totally different. Wealth is a common example for a skewed distribution. The square root and the logarithm for a variable are methods for transforming imbalanced data into balanced data. We apply both methods for Balance variable since the distribution of Balance variable is too skewed.

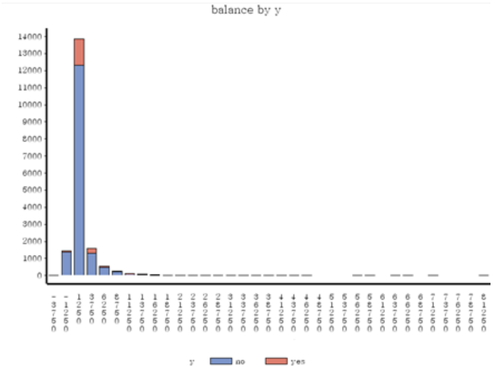

Fig. 4 the distribution of Balance variable

The distribution of Balance variable becomes less skewed after transformation as we see the shape in Figure 5.

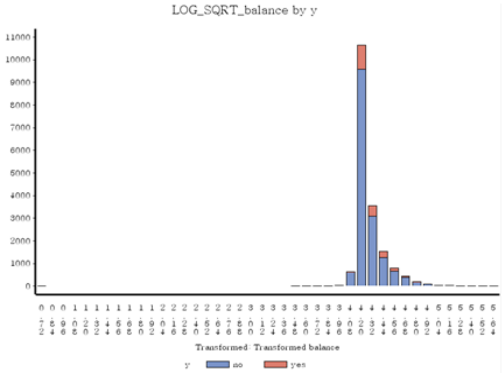

Fig. 5 the distribution of Balance variable transformed

Duration variable means the last contact duration and its measurement is second. Most of the customers talk shortly, but some customers have a conversation for a long time. Therefore, the distribution of Duration variable is right-skewed in Figure 6.

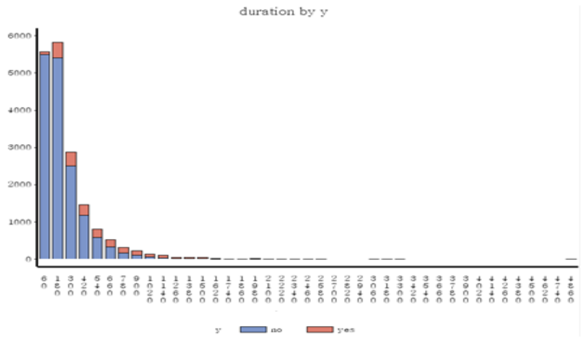

Fig. 6 the distribution of Duration variable

It is transformed using the square root of Balance variable. The degree of imbalance is resolved after the transformation in Future 7

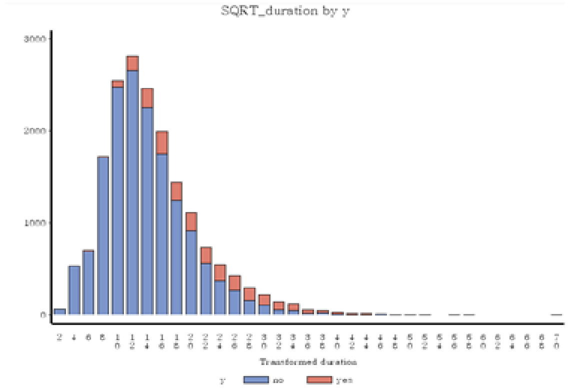

Fig. 7 the distribution of Duration variable transformed

A. Comparison of the performances of the classification methods without variable transformation

We compare the performances of the classification methods without variable transformation as an initial procedure.

Four classification methods are selected for comparison and four measures are used for evaluating the performances of the classification methods. All algorithms and measures are already described in the third section.

Model	MIS	AUC	K-S	CRR
Logistic regression	0.0957	0.905	0.676	58.511
Gradient boosting	0.1098	0.888	0.651	47.826
Decision tree	0.0944	0.809	0.578	57.687
Neural networks	0.0930	0.926	0.713	60.354

Table 3the performances of four classification methods for a Portuguese bank data

There are several abbreviations such as MIS for the misclassification rate, K-S for the Kolmogorov-Smirnov statistic, and CRR for the top 10% cumulative response rate. The measurement for CRR is percent unlike other measures. We use the average rank to compare the classification methods for overall evaluation by four measures. The average rank for each classification method is 2.25, 3.50, 3.50, 3.25, and 1.00, respectively. Therefore, we do not conduct research for a gradient boosting and a decision tree due to their low performances any more.

We apply the transformed data gradually to show the effects of variable transformations. First, we build a logistic regression model trained by the transformed data.

B. Logistic regression model based on transformed data

We construct three transformed data. The first transformed data have the transformed variables for Balance and Previous. Previous variable denotes the number of contacts performed before the campaign. The square root is used for transformation of Previous variable. The second transformed data have the transformed variables from the first data and Age and Duration variables which are transformed. The last transformed data have the most transformed variables. All transformed variables such as Age, Balance, Duration, and Previous variables from the previous data are included and Pday variable is transformed. Pday variable is the number of days that passed by after the client was last contacted from a previous campaign.

Data	MIS	AUC	K-S	CRR
No transformation	0.0957	0.905	0.676	58.511
1st transformed data	0.0958	0.905	0.677	58.659
2ndtransformed data	0.0950	0.911	0.681	59.764
3rd transformed data	0.0942	0.913	0.691	59.985

Table 4 the performances of the logistic regression model based on the different transformation

We use the average rank for comparison among various transformed data like Table 3. The average ranks are 3.63, 3.38, 2.00, and 1.00 respectively. The result is a little different from the result of the model from original data when we apply a logistic regression model on data with the transformed Balance and Previous variables. However, the model performance is improved as we transform more variables. Finally, the performance of the model is the best when five variables are transformed.

C. Neural network model based on transformed data

We follow the same process like the construction of a logistic regression model based on various transformed data. The initial data and three transformed data are applied for building a neural network model and the performances for each data are compared.

Data	MIS	AUC	K-S	CRR
No transformation	0.0930	0.926	0.713	60.354
1st transformed data	0.0938	0.923	0.700	60.354
2ndtransformed data	0.0925	0.928	0.721	62.122
3rd transformed data	0.0910	0.929	0.721	62.343

Table 5 the performances of the neural network model based on the different transformation

The average rank of four measures for each data is 3.13, 3.88, 1.88, and 1.12 respectively. When the tie happens in ranking, we assign the average ranks to two corresponding values. The neural network model which is based on the most transformed data shows the best performance like a logistic regression model.

V. CONCLUSION

Data mining is one of the steps in KDD. KDD is the whole process that we find knowledge from unorganized data by data mining techniques. Processes before applying data mining techniques are important in many cases. The performance of data mining is dependent on how we deal with data. We can improve the performance of data mining algorithms dramatically by cleaning observations, handling missing values, and manipulating variables in data. We make an experiment on a real Portuguese bank data to verify the effects of variable transformation. First, four classification methods are compared for data without variable transformation. We continue to conduct research on a logistic regression model and a neural network model further that show better performances than a gradient boosting and a decision tree. We find that the models based on data with many transformed variables shows the best performances from the extended experiments. It has the tendency that the performance of a model increases as the number of the transformed variables increases.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Magazine, vol. 17(3), 1996.

[2] G. S. Linoff and M. J. A. Berry, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd ed., Wiley, Apr. 2011.

[3] E. Frank and M. Hall, "A simple approach toordinal classification," in Machine Learning: ECML, 2001, pp. 145–156.

[4] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," The Annals of Statistics, vol. 29(5), pp. 1189–1232, Jun. 2001.

[5] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in ICML '06 Proceedings of the 23rd international conference on Machine learning, 2006, pp. 161–168.

[6] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in ICML '08 Proceedings of the 25rd international conference on Machine learning, 2008, pp. 96–103.

[7] L. Breiman, J. H. Friedman, R. A. Stone, and C. J. Stone, Classification and Regression Trees, 1st ed., Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[8] J. R. Quinlan, Programs for Machine Learning, 1st ed., Morgan Kaufmann Publishers, 1993.

[9] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1(1), pp. 81–106, 1986.

[10] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," Neural Computation, vol. 8(7), pp. 1341–1390, Oct. 1996.

[11] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decision Support Systems, vol. 62, pp. 22–31, Jun. 2014.