Comparing the Performance of Winsorize Tree to Other Data Mining Techniques for Cases Involving Outliers

Chee Keong Ch'ng

Abstract: ***Winsorize tree is a modified tree that reformed from classification and regression tree (CART). It lays on the strategy of handling and accommodating the outliers simultaneously in all nodes while generating the subsequence branches of tree. Normally, due to the existence of outlier, the accuracy rate of most of the classifiers will be affected. Therefore, we propose winsorize tree which could resist to anomaly data. It protects the originality of the data while performing the splitting process. In this study, winsorize tree was compared to other classifiers. The results obtained from five real datasets indicate that the proposed winsorize tree performs as good as or even better compare to the other data mining techniques based on the misclassification rate.***

Keywords: winsorize tree algorithm; outlier; gini index; misclassification rate; classification; classification and regression tree; winsorized tree.

I. INTRODUCTION

Outliers

Areal-world data is never perfect as it normally includes a certain amount of exceptional values termed as outliers. For instance, data entry error, violations of integrity constraints. sensor failure, experimental error and human errors which may cause serious problem in statistical analysis (Abedjan et al, 2016; Pit-Claudel, Mariet, Harding, & Madden, 2016). Various scientific disciplines regularly come across 'outliers' in their data (Aguinis,Gottfredson, Joo. 2013; Bakker & Wicherts 2014). In fact, there are many statistical definitions of outlier but it depends on the underlying distribution of the variable in the data. According to Bluman and Allan (2000), outlier can be defined as an extremely high or low value in a data which has potential to influence the statistical analysis. Singh and Upadhyaya (2012) defined that outliers are patterns in data that do not comply to a well-defined notion of normal behavior. Young, Valero-Mora and Friendly (2006) asserted that the values beyond bound or distribution that is drastically affecting on the analysis and become the most challenging and pervasive in an organization (Aguinis, Gottfredson, Joo, 2013). Besides, outlier has also been defined as observation which substantially differs from what it supposes to be (Hair et., 1992). Having the anomaly value in data is problematic as it could distort the original behavior of the data and it might disproportionate the effect on statistical results in classification or prediction (De Veaux & Hand, 2005).

Revised Manuscript Received on June 22, 2019.

Chee Keong Ch'ng, School of Quantitative Sciences, College Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

In such problem, outlier must be detected and handled in order to improve the consistency and transparency of practices (Aguinis, Gottfredson, Joo, 2013). Generally, outlier detection and novelty detection are both used for distortion. Outlier detection is the training data that contain outliers which has been detected as far from normally observations whereas novelty detection is identifying an anomaly that a machine learning does not detect during training (Pimentel et al, 2014). Simply ignoring or removing the outlier could bring bias to the analysis (Ch'ng, Mahat, 2014). Therefore, outlier detection methods must be applied so that we could know the variabilities, characteristics and skewness affected by the extremities. There are single construct techniques and multiple construct techniques in outlier detection methods. In single construct technique, the most classical and popular methods are stem and leaf, boxplot, schematic plot analysis, percentage analysis and standard deviation analysis; in multiple construct techniques, the most common methods are scatter plot, q-q plot, p-p plot, Euclidean distance, Mahalanobis distance, K-clustering, Hosmer and Lemeshowgoodness-of-fit test etc (Aguinis, Gottfredson, Joo, 2013). Once the data has been detected, the next process is to handle it before carrying out any further analysis. There are many approaches that are used to treat the anomaly. The easiest one is by removing or truncating the outliers. However, simply eliminating the outlier would violate the nature of data because some outliers can be legitimate and sometimes it could be the most interesting ones. Therefore, retaining the outlier are normally preferable by most of the researchers. The common techniques have been widely used are winsorization, modification, transformation, estimation etc. Some propose robust approaches could reduce the influence of outlier (Yuan & Zhong, 2008; Zhong & Yuan, 2011).

Outlier handling in classification

Engles and Theusinger (1998) insisted that preprocessing should be carried out before generating model. Data preprocessing is mainly referring to three directions which are data cleansing (treatment of outliers, noises, etc), altering the dimensionality of the data (transformation, attribute generation, filtering, etc) and altering the data quantity (by selecting, sampling, balancing the available data records). However, having a clean data is too academic andsometimes it is not realistic especially in real world application. Moreover, most of the current approaches are carrying double tasks which are preprocessing and model generation (Ch'ng& Mahat, in press).

Not manyexistence modelsare robust towardsoutliers.Dervilis et al. (2016) proposed a robust regression for outlier detection by exploring and visualizing structural health monitoring (SHM) data as a tool in investigating and monitoring the characteristics of outlier whilst removing it from the data. Dervilis et al. (2014b, 2015) and Rousseeuw, Hubert and Aelst (2008) proposed a high-breakdown method that is robust against the outliers where it can deal with a substantial fraction of outliers in the data (Rousseeuw&Hubert, 2013). Decision tree is another classification technique that is quite robust towards the outliers as it could isolate them during the construction of tree (Breimen et al., 1984; John, 1995; Shouman, Turner & Stocker, 2011). However, simply ignoring the outliers would destabilise the estimation. Therefore, Ch'ng and Mahat (2014) demonstrated the construction of CART is deviated due to the influence of outliers. Therefore, ignoring the outlier could result in wrong estimated values hence producing different structure of trees. At worst, a future object may be misallocated into its class. The most common way is to prune the tree accordingly to reduce the complexity of the tree classifier, hence improves the predictive accuracy or reduces the misclassification rate with smaller size of tree. John (1995) introduced a robust C4.5 algorithm to improve the results. This method incorporates with the pruning scheme to fully remove the effect of outliers before regenerating the decision tree model using the reduced training set. With the idea of clustering, Kyung, June, Dao and Nam (2011) proposed a decision tree-based clustering algorithm to overcome the weaknesses of outliers in data. This method has been applied in HMM-based speech synthesizer training where the outliers must be removed during the growing phase of tree. It has been proven that this method could produce awell-balanced speech quality irrespective to a sentence. Ch'ng & Mahat (In press) produced winsorize tree algorithm which is robust to outliers in the data.Winsorize tree can manage data well especially dealing with outliers during the splitting process in every single node. In other words, winsorize tree can perfectly resist to the outliers and produce an accurate and a right size of tree by performing winsorize gini purity index in every node recursively during the construction of the tree model without involving pruning process. By comparing among few decision tree methods, winsorize tree has been proven outperformed. This paper is the extension of the paper in Ch'ng & Mahat (In press). The outperformed winsorize tree is compared to the other classifiers such as neural network, traditional decision trees (gini and entropy) and logistic regression by using five different data sets. The purpose of this study is to determine the reliability and accuracy of winsorize tree compare to some other classification methods.

II. DATA MINING TECHNIQUES

Neural Network

Neural network is a computer system that mimic on the human brain and nervous system that is used for information processing. Self-learning within network can derive complex and important information whilst recognize patterns from a data. It alsointerprets sensory data through a kind of machine perception, labeling or clustering raw input. In an artificial neural network, simple artificial nodes, called "neurons", "neurodes", "processing elements" or "units", are connected to form a network (Russell, 1991).

Logistic Regression

Logistic regression is used to find the best fitting a statistical model that uses a logistic function to model a binary dependent variable.The goal of logistic regression model to describe data and to explain the relationship between one dependent dichotomous variable and any type of measurement scale in independent variables. Multinomial logistic regression is usually reserved for the case when the dependent variable has three or more categories, such as win, draw, or lose. Logistic regression is more versatile in most of the situation as it does not assume that independent variables are normal distributed.

Decision tree

Decision tree is a supervised learning algorithm that uses the idea of divide and conquer to classification and prediction. It breaks down the data into smaller subsets whilst associates with the increment of the tree size. The topmost of decision tree is called parent node. It splits into two or more branches which is connected to the subsequence child nodes. The process is repeating until it reaches the final node called leaf where it holds the class label. There are many types of decision tree algorithms such as ID3 (Quinlan, 1986), CART (Breimen et al.), and C4.5 (Quinlan, 1993) and Kass (1975). Different algorithm uses different measurement of selecting the best splitting criterion. The most popular metrics are entropy, information gain, gain ratio and gini index.

C4.5

C4.5 is an algorithm which was developed by Ross Quinlan (1993). It is an extension of Quinlan's ID3 algorithm. C4.5 generates decision trees which can be used for classification and therefore C4.5 is often referred to as statistical classifier. This method can deal with both continuous and discrete attributes and also with the missing values and pruning trees after construction. C5.0 is the commercial successor of C4.5 because it is a lot faster, more memory efficient and used for building smaller decision trees.C4.5 performs by default a tree pruning process. This leads to the formation of smaller trees, more simple rules and produces more intuitive interpretations.

Classification and Regression tree (CART)

CART is a recursive partitioning method that builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). CART was introduced by Breiman in 1984. For classification, gini index is used as selecting the splitting attribute. Regression-type problem is generally predicting or forecasting the values of a continuous variable from one or more continuous and categorical predictor variables in a given time period.

Winsorize tree

Winsorize tree is robust method that is reformed from CART. Ch'ng & Mahat (In press) proposed a strategy in decision tree usingwinsorize tree algorithm which would be able to simultaneously handle and accommodate the outliers during the process of constructing the tree.The advantage of this method is it could resist the abnormal data set and protect the original information of the data. It automatically investigates, detects, penalises and accommodates the suspicious value in all nodes to reduce the effect of contaminated data before performing gini purity measurement (for attribute selection) using the original data.

In other words, every node performs outlier inspection before computing the metrics (gini index). In this extent, this approach can avoid losing the originality of data. The process is done recursively in every node until the threshold is achieved. Besides, pruning process is not required in this study as the tree can stop before overfitting. Previous research showed thatwinsorize tree algorithm is capable to produce a comparable or even betterresults when comparing to other decision tree models (Ch'ng & Mahat, in press). Therefore, this study compares winsorize tree to other data mining models by measuring the misclassification rate.

III. METHODOLOGY

Winsorize tree

The algorithm has been discussed in Ch'ng and Mahat (In press). The idea of winsorize tree algorithm lays on the strategy of handling and accommodating the outliers during the process of developing the tree simultaneously. In other words, winsorize tree integrates the preprocessing and the splitting process in all nodes.In general, winsorize tree are divided into five parts which are data inspection, outlier handling, gini impurity measurement, tree construction and evaluation. All steps are repeated until it reaches theleaf node. The algorithms are as below:

All variables in adata are screened by using boxplot.

$$L_k = Q_{1k} - 1.5 \times IQR_k$$
$$U_k = Q_{3k} + 1.5 \times IQR_k$$

Outlier region = $R < L_k$ or $R > U_k$

The outlier is winsorized and accommodated

$$R_{wk} = \{r[x_{(r+1)}],\ \textstyle\sum_{i=r+1}^{n-r} x_i\ , r[x_{(n-r)}]\}$$

Gini index and weighted average are computed

$$G_{wk} = 1 - \sum_j [p(j/t)]^2$$

$$weighted\ average, G_{wsplit} = \sum_{i=1}^{k} \frac{n_i}{n} G_{wk}$$

$$Info\ gain, \Delta(t_w) = i(n) - G_{wsplit}$$

where $i(n)$is the gini score in parent node. The best split is the one provides the highest information gain. Every time when the process repeats for the subsequence nodes, the original data will be reused instead of maintaining the winsorize data. Then, the process above is repeated in every node until the threshold for stopping rules are achieved.

Finally, we test for error estimation.

$$Misclassification\ rate = (T - C)/T$$

where T represent the total number of objects and C represent the correct classified objects in a test set.

Data

Indians Liver Patient Dataset (ILPD)

Indians Liver Patient Dataset (ILPD) was collected from north east of Andhra Pradesh, India. Many researches used the data for comparative analysis and trying to improve in prediction accuracy (Ramana, Babu & Venkateswarlu, 2012). The data contains 583 observations. The data has 10 independent variables and a dependent variable with two groups. There are 441 male patients and 142 females in record. 416 of the patients have liver problem and 167 have no liver problem in the group.

Pima Indians

The data set contains the collection of medical diagnosis report of 768 observations and 9 variables with two dependent variables on the status of diabetes, either positive (P) or negative (N) of getting diabetes. There are 500 patients from the negative group and the remaining are from positive group were being tested glucose levels. (Smith, Everhart, Dickson, Knowler & Johannes, 1988).The variables used for distinguishing those suffer with diabetes are number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml),body mass index ($weight\ in\ kg/(height\ in\ m)$^2), diabetes pedigree function, age (years), class variable (P or N).

Breast Tissue

The breast tissue data set is a sample of data that explain about breast cancer diagnosis that had been analyzed and reported by some researchers including Jossinet (1996) and Silva, Marques and Jossinet (2000). The measurements in the data are based on Electrical Impedance Spectroscopy (EIS) which are used to measure the complex impedance properties of a material. In medical practices, the EIS measurement of breast tissue can be used as prescreening for cancerous tissue. Therefore, historical data of EIS gives opportunity to researchers to investigate further about the potential patients of breast cancer hence some early pre-cautions can be taken to minimize its implications on the patients. Breast tissue data set contains nine variables to discriminate 6 classes of tissue.

Iris

Perhaps iris flower data set is one of the best and prominent case of study in pattern recognition literature. The Iris data was collected by Edgar Anderson in year 1936 in which the flowers were classified into 3 different species (Iris Setosa, Iris Virginica and Iris Versicolor). The data consists of 150 examples from each species and four variables namely SepalLength (sepal length), SepalWidth (sepal width), PetalLength (petal length) and PetalWidth (petal width). This data is retrieved from Fisher (1936).

Egyptians Skulls

The changes of skull sizes were recorded between the time periods. The change in skull size is due to the interbreeding of the Egyptians with immigrant population over the years.Four measurements are made of male Egyptian skull which are maximal breadth of skull(mb), basibregmatic height of skull (bh), basialveolar length of skull (bl), and nasal height of skull (nh) from five different time period ranging from 4000B.C to 150 A.D. (Handet al.,1994).

IV. RESULTS

In this paper, we compared 5 classification models to winsorize tree to see their performance by using 5 different data sets. We compared them is classification rate and the results are shown as in Table 1.

Table. 1 Comparing the misclassification rate between winsorize tree and other classifiers

	Data				
Classification models	Indian Liver patient dataset	Pima Indians	Breast Tissues	Egyptian Skulls	Iris
Logistic regression	*0.2881	0.2241	0.5000	*0.7353	0.0938
Neural network	0.3220	0.2414	0.3056	0.7647	0.0938
Neural network (no hidden nodes)	0.3220	0.2413	0.3056	0.7647	0.0938
Winsorize tree	0.3109	*0.1758	*0.2308	0.7568	*0.0000
Decision tree model (gini)	*0.2881	0.2586	0.3056	0.7647	*0.0000
Decision tree model (entropy)	*0.2881	0.2672	0.3333	0.7647	*0.0000

Even though the misclassification rate of winsorize tree is slightly higher compare to logistic regression model and both decision tree models, winsorize tree still outperforms in Pima Indians and Breast Tissues data set with the lowest misclassification rate 0.1758 and 0.2308 respectively. In iris data set, there is no different among all trees. We purposely try on this data as we want to see whether winsorize tree is stable when the data consists of only one outlier or no outlier at all. The result shows that winsorize tree is still comparative to other trees in this situation with the 0.000 misclassification rate. Overall, winsorized tree algorithm is capable to produce a comparable or even better classifier with no data values are excluded along the construction of tree. Moreover, it could resist to any outlier while performing the splitting process in all nodes. Winsorize tree does not require any pruning process too as other type of trees do as it is able to stop generating the branch at the right time with the right size. Therefore, winsorize tree is highly recommended to any classification problem.

V. CONCLUSION

In this paper, we compared winsorize tree to other data mining techniques. Based on the experiments conducted on five data, our method has been proven comparative or even better in any size of data set. Our experiments on the data sets confirmed this claim by producing lower misclassification rate with simpler size of tree. Moreover, winsorize tree algorithm is robust or insensitive towards the data that contains outlier by protecting and retaining the original data.

REFERENCE

1. Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., & Tang, N. (2016). Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*, *9*(12), 993-1004.
2. Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods, 16, 270–301. doi: 10.1177/1094428112470848.
3. Anderson, E. (1936). "The species problem in Iris". Annals of the Missouri Botanical Garden. *23*(3): 457–509.
4. Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: the power of alternatives and recommendations. Psychological Methods, 19, 409–427. doi: 10.1037/met0000014.
5. Bluman, Allan (2000), Elementary Statistics, brief version, New York: McGraw-Hill
6. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) Classification and regression trees, Chapman & Hall, New York.
7. Ch'ng, C. K. and Mahat, N. (2014) 'Winsorised gini impurity: A resistant to outliers splitting metric for classification tree', International Conference on Quantitative Sciences and Its Applications.
8. Ch'ng, C. K. and Mahat, N. (in press). Winsorize tree algorithm for handling outlier in classification problem. *International Journal of Operational Research.*
9. De Veaux, R. D. and Hand., D. J. (2005) 'How to lie with bad data', Journal of Statistical Science, Vol. 20, No. 3, pp.231-238.
10. Dervilis, N., Antoniadou, I., Barthorpe, R. J., Cross, E. J., and Worden, K. (2016) 'Robust methods for outlier detection and regression for SHM applications', Int. J. Sustainable Materials and Structural Systems, Vol. 2, Nos. 1/2, pp.3–26.
11. Dervilis, N., Cross, E.J., Barthorpe, R.J. and Worden, K. (2014b) 'Robust methods of inclusive outlier analysis for structural health monitoring', Journal of Sound and Vibration, Vol. 333, No. 20, pp.5181–5195.
12. Engels, R., and Theusinger, C. (1998) 'Using a data metric for preprocessing advice for data mining applications', Proceeding of 13th European Conference on Artificial Intelligence, John Wiley & Sons, Chichester. Evans
13. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics. 179-188.
14. Hair, J. F., Anderson, R., Tatham, R. L., and Black, W. C. (1992) Multivariate data analysis with reading (3rd Eds.), New York, Macmillan.
15. Hand, D.J., Daly, F., Lunn, A. D., Mcconway, K. J. and Ostrowski, E. (1994). A Handbook of Small Data Sets, New York: Chapman & Hall, pp. 299-301.
16. John, G. H. (1995). Robust decision trees: removing outliers from databases, KDD-95 Proceeding, CA: AAAI, Menlo Park.
17. Jossinet, J. (1996).Variability of impedivity in normal and pathological breast tissue. Med. & Biol. Eng. & Comput, 34, 346-350.
18. Kyung, H. O., June, S. S., Doo, H. H. and Nam, S. K. (2011). 'Decision tree-based clustering with outlier detection for HMM-based speech synthesis', 12th Annual Conference of the International Speech Communication Association, ISCA, Florence, Italy.
19. Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Process. 99, 215–249 (2014)

20. Pit-Claudel, Z. Mariet, R. Harding, and S. Madden. Outlier detection in heterogeneous datasets using automatic tuple expansion. Technical Report MIT-CSAIL-TR-2016-002, CSAIL, MIT, 32 Vassar Street, Cambridge MA 02139, February 2016.
21. Quinlan, J. R. (1987). Simplying decision tree. International Journal of Man-Machine Studies - Special Issue: Knowledge Acquisition for Knowledge-based Systems, 27(3), 221-234.
22. Quinlan, J. R. (1993). C4.5: Programs for machine learning. USA: Morgan Kaufmann Publishers.
23. Ramana, Bendi, V., Surendra Prasad Babu, M., and Venkateswarlu, N. B., A critical comparative study of liver patients from USA and INDIA: An Exploratory Analysis. Int. J. Comp. Sc. Issues 9(3):506–516, 2012.
24. Rousseeuw, P. and Hubert, M. (2013) 'High-breakdown estimators of multivariate location and scatter', Robustness and Complex Data Structures, pp.49–66, Springer.
25. Rousseeuw, P., Hubert, M. & Aelst, S. V. (2008). High-breakdown multivariate robust method. *Statistical Science*, *23*(1), pp.92–119.
26. Russell, I., Neural Networks in the Undergraduate Curriculum, *Journal of Computing in Small Colleges*, April 1991, *6*(4), April 1
27. Shouman, M., Turner, T. and Stocker, R. (2011) 'Using decision tree for diagnosing heart disease patients', In Proc. Australasian Data Mining Conference (AusDM 11), Ballarat, Australia
28. Silva, J. E., Marques de Sá, J. P., & Jossinet, J. (2000). Classification of Breast Tissue by Electrical Impedance Spectroscopy. *Med & Bio Eng & Computing*, 38, 26-30.
29. Singh, K., Upadhyaya, D.S., 2012. Outlier detection: Applications and techniques. *International Journal of Computer Science, 9*(1), 307-323.
30. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proc Annu Symp Comput Appl Med Care (pp. 261-265). IEEE Computer Society Press.
31. Young, F. M., Valero-Mora, P. M., & Friendly, M. (2006). Visual statistics: seeing data with dynamic interactive graphics, Wiley, New Jersey
32. Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology*, 38, 329-368.
33. Zhong, X., & Yuan, K.-H. (2011). Bias and efficiency in structural equation modeling: Maximum likelihood versus robust methods. Multivariate Behavioral Research, 46, 229-265.