ROC-tree algorithm for stratification of binary classifier sets with varied discrimination threshold

Yuri M. Ganushchak (yga@planet.nl)
Maastricht University Medical Center+

P. J.C. Barenburg
Maastricht University Medical Center+

J. G. Maessen
Maastricht University Medical Center+

P Sardari Nia
Maastricht University Medical Center+

Research Article

Keywords: data aggregation algorithm, ROC curve, c-statistics, hybrid metrics

Posted Date: April 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1419287/v1

Abstract

Binary classifier systems are used in multiple practical situations. Evaluation of diagnostic ability of a binary classifier, as its discrimination threshold is varied, often requires data transformation by performing aggregation operations. One of the most used aggregation methods is division by percentiles which divides the data set at the equal by size subgroups blindly, independently from the structure of data. We developed a ROC-tree algorithm for selection of threshold values, which is a recursive downwards splitting of each group at the two subgroups (branches) by cut-off point of ROC curve. We showed that suggested ROC-tree algorithm allows to define optimal (natural) boundaries and number of groups.

Two methods of data aggregation (percentiles and ROC-tree algorithms) were tested using the dataset 'Credit Card Fraud Detection' (https://www.kaggle.com/mlg-ulb/creditcardfraud). The results of one-vs-one reduction for the assessment of the multiclass classifications were presented as macro-average of hybrid threshold performance metrics. The macro-averages of metrics like Youden index, accuracy, optimized precision, and geometric mean were significantly different between used aggregation algorithms. The differences between macro-average of metrics ROC-tree and quartiles algorithms of stratification were preserved during 10 fold stratified cross-validation procedure.

Using algorithm sensitive to the distribution patterns, e.g., ROC-tree algorithm showed adequate stratification at groups by natural cut-off points determined by the data set composition. This method provides effective aggregation for summarizing or analyzing data in a various fields of sciences. In health care described algorithm allows effective evaluation of mortality causes and quality control specialized medical care by hospitals.

Introduction

Binary classifier systems where its elements are classified into two groups are used in multiple practical situation. These include: medical testing or prognostic (risk prediction) models, quality control, fraud detection, and machine learning and information retrial. However, evaluation of diagnostic ability of a binary classifier as its discrimination threshold is varied often requires data transformation by performing aggregation operations. Aggregating individual observations into groups is used in a various fields of sciences as a form of categorization when the discrete groups (strata) of data are created. Grouped data serves as a convenient means of summarizing or analyzing the data. Identification of discrete groups is one of the most important and difficult tasks of data mining, that is why finding a good classifier and classification algorithm is an important component of data mining.

The selection of this threshold value (possibly subjective) can have dramatic effects on model accuracy.[1] One of the most used aggregation methods is division by percentiles (quartiles as a special case of percentiles division) which divides the data set at the equal by size subgroups independently from the

structure of data. We developed a ROC-tree algorithm for selection of threshold values which is recursive downwards splitting of each group at the two subgroups (branches) by cut-off point of ROC curve.

We hypothesized that opposite to the percentiles division, the ROC-tree algorithm allows to define optimal (natural) boundaries and number of groups.

Materials And Methods

The 'Credit Card Fraud Detection' dataset downloaded from https://www.kaggle.com/mlg-ulb/creditcardfraud was used for the illustration of algorithm. The datasets contains transactions made by credit cards in September 2013 by European cardholders.

As a pre-processing step we used we used ROC based feature selection to handle class imbalance classification problem. The AUC for all possible classifiers variables are presented in appendix, Table 1B.

The Youden Index (Bookmaker Informedness) was used for selection of cut-off points in recursive downwards dividing subgroup into two new subgroups (branches). An area under the curve less than 0.65 in at least one subgroup of iteration was considered as an exit condition while cut-off points and number of subgroups from the previous iteration were taken for further analysis (Fig. 1).

The iterative usage of traditional default threshold of 0.5 as the cut-off generated a four discrete groups (quartiles) with equal number of observations.

The comparison of classification algorithms was made using methods similar to the evaluation of multiclass classification. Similar to the assessment of the multiclass classification algorithms in machine learning, the one-vs-one reduction was used (Appendix A, Fig A1). Where applicable, the derivations of the 2*2 confusion matrix are presented as their macro-averages of post-hoc procedure results (one-vs-one pairwise comparison). A macro-average is the average of a metric computed independently for each class while treating all classes equally. The confusion matrix for binary classification is presented in Fig A2 (Appendix A).

The 10-fold stratified cross-validation procedure, where each fold has the same proportion of observations with the class outcome value, was used for internal validation of classification algorithm. The capabilities of algorithms were estimated as the average of performance metrics. [2]

The list of variables used in the study and their equations are presented in Appendix A.

The R 3.6.3 for Windows with RStudio 1.2.5033 [3] and standard packages with libraries 'lattice', 'readr' were used for the classification of data and calculation of derivates of contingency tables for the comparison of classification algorithms.

Results

Credit Card Fraud Detection' data set presents transactions that occurred in two days, where was 492 case of frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. The Imbalance ratio, which value lies within the [0, ∞] range, having a value IR = 1 in the balanced case was close to 0 (IR = 0.002) and imbalance coefficient (δ = -0.997) with expected values within the [−1 , 1] range, and 0 value for the perfectly balanced classes.

Iteration of ROC curve procedure through 28 discrimination parameters (Appendix B. Table 1B) uncover several variables with high AUC. Variable V4 (Figure 1, AUC 0,938) was used as a classifier for the ROC-tree downwards splitting. The density plot (Figure 2) shows that the distribution of cases with and without fraud by V4 parameter are different. Two-sample Kolmogorov-Smirnov test confirmed that V4 distribution is different in cases with and without fraud (D = 0.7664, p-value < 2.2e-16). However, Bhattacharyya distance for V4 case is 0.626 and Bhattacharyya coefficient (a measure of the amount of overlap between two statistical samples or populations) is 0,535.

Both methods of stratification creates 4 groups (Tables 1 and 2) with statistically significant differences with expected distributions. However, projection of cut-off point at the density chart (Fig.3) illustrates the fact that in case of quartiles algorithm most of fraud+ cases concentrated in group 4. ROC-tree algorithm provides more "fair" spreading cases with the highest density of fraud+ in the third group (Figure 3)

Table 1

Stratification at 4 groups using ROC-tree algorithm

	gr1	gr2	gr3	gr4	Total
Fraud+	14	60	204	214	492
Fraud-	148625	112038	22844	808	284315
Total	148639	112098	23048	1022	284807

X^2 (3, N= 284807) =26558, p < 0.0001

Table 2

Stratification at 4 quartiles groups.

	gr1	gr2	gr3	gr4	Total
Fraud+	2	11	22	457	492
Fraud-	71200	71190	71180	70745	284315
Total	71202	71201	71202	71202	284807

X^2 (3, N= 284807) =1213, p < 0.0001

Distribution of cases with and without fraud by V4 using ROC-tree algorithm or quartiles division presented at the figure 4. The results of one-vs-one reduction for the assessment of the multiclass classifications are presented as macro-average of performance metrics (Table 3).

Table 3

Performance metrics in one-vs-one comparison groups*

Metric	algorithm	gr4_3	gr4_2	gr4_1	gr3_2	gr3_1	gr2_1	macro-avg
Youden index	ROC-tree	0,478	0,774	0,933	0,603	0,803	0,381	0,66 ± 0.21
	Quartiles	0,456	0,478	0,497	0,167	0,417	0,346	0,39 ± 0.12
Accuracy	ROC-tree	0,958	0,992	0,995	0,831	0,867	0,570	0.869 ± 0.161
	Quartiles	0,503	0,503	0,503	0,500	0,500	0,500	0.502 ± 0.002
Sp	ROC-tree	0,966	0,993	0,995	0,831	0,867	0,570	0.870 ± 0.162
	Quartiles	0,502	0,502	0,502	0,500	0,500	0,500	0.501 ± 0.001
OP	ROC-tree	0,651	0,873	0,966	0,794	0,829	0,396	0.751 ± 0,202
	Quartiles	0,192	0,182	0,173	0,357	0,206	0,243	0.226 ± 0.069
GM	ROC-tree	0,703	0,881	0,966	0,801	0,901	0,680	0.822 ± 0.114
	Quartiles	0,692	0,700	0,707	0,577	0,677	0,650	0.667 ± 0.048

Sp - specificity; OP - optimized precision; GM – geometric mean.

* table includes metrics with meaningful values and significant differences between macro-averages (t-test, $p < 0.05$) are presented in the table.

All confusion table metrics presented in the Table 3 were significantly higher in case of ROC-tree algorithm classification.

In order to compare the performance of stratification algorithms the 10-fold stratified cross-validation procedure was completed for each algorithm. The original data were randomly partitioned into 10 equal sized subsamples (folds) so that each partition contains roughly the same proportions of the two types of class labels (fraud+, fraud-) (Table 4) and similar density distributions (Figure 5).

Table 4

Number of cases and classifier* mean ± std per fold

	fraud -			fraud +		
fold	n	mean	std	n	mean	std
1	28432	-0,0076	1,4012	50	4,562	3,020
2	28432	-0,0077	1,4009	50	4,528	2,995
3	28432	-0,0077	1,4007	49	4,624	2,896
4	28432	-0,0078	1,4001	49	4,606	2,896
5	28432	-0,0080	1,3993	49	4,579	2,893
6	28431	-0,0078	1,3987	49	4,557	2,885
7	28431	-0,0079	1,3984	49	4,538	2,879
8	28431	-0,0080	1,3983	49	4,498	2,839
9	28431	-0,0080	1,3980	49	4,478	2,841
10	28431	-0,0081	1,3979	49	4,447	2,841

* V4 was selected as classifier

Procedure of 10-fold cross-validation was done at the next way. Of the 10 folds, a single fold was reserved as the validation data for testing the model, and the remaining 9 subsamples were used as training sets of data. The cross-validation process is then repeated, with each of the 10 folds used exactly once as the validation data. The results from the folds were averaged to produce a single estimation.

The results of cross validation metrics included in the study are presented in the tables 2b ….16b, Appendix B. The differences between macro-average of metrics ROC-tree and quartiles algorithms of stratification were preserved during cross-validation procedure. However, the relative bias and mean square error of algorithm were statistically not different from 0 (One Sample t-test) and did not differ in ROC-tree vs Quartiles groups for all metrices included in the study. Additionally, computing confusion table metrics in control folds through cross-validation procedure of quartiles algorithm in 10% failed in comparison gr 3 vs 2, 3 vs 1 and 20% in comparison gr 2 vs 1 for GM, OP, and Youden index. This effect can be caused apparent to unsensitivity of quartiles algorithm to the distribution of fraud -/fraud + cases and concentration of most of fraud+ in fourth group (Figure.3).

Discussion

We evaluated the quality of two classification (aggregation) algorithms: ROC-tree and division at quartiles. The universal nature of the aggregation task allows to use for the demonstration of the algorithm ‘Credit Card Fraud Detection’ dataset downloaded from https://www.kaggle.com/mlg-ulb/creditcardfraud. This dataset contains much more cases than any available medical dataset. ‘Credit Card Fraud Detection’ preserves the imbalance structure inherent to the medical data. Furthermore, using

dataset distant from the healthcare allows to avoid unnecessary discussion around acceptability of predictive scores (e.g. Euroscore, syntax score, CSA-AKI, Charlson comorbidity index, et cetera).

Classification methods are used in various fields of biological and medical sciences as a form of categorization when the discrete groups (strata) of data are created. Classification is one of the most important and difficult tasks of data mining, which is why finding a good classifier and classification algorithm is an important component of data mining. Classification into several tiers is the further step in the organization and understanding data. For example: division at high, medium, and low risk, based on scores of the patient cohort is an important step in the organization and understanding clinical contexts.[4]

One of the most often used algorithm for division dataset into tiers is division at percentiles with creation of strata with similar number of cases or usage of early predefined cut-off points are traditional specially in medical investigations.

In the two-class classification task, the Receiver Operating Characteristic (ROC) curve is one of the most widely used tools to assess the performance of algorithms.[5, 6] The area under the receiver operating characteristic curve (AUC) (also referred to as the c statistic) is by far the most popular index of discrimination ability[7] ROC curves have an attractive property: they are insensitive to changes in class distribution. The ROC curves are independent of the proportion of positive to negative instances in a test set.[8]

Several researchers have investigated the application ROC curves not only as a metrics of classification successes. Ferri et al. (2002) altered decision trees to use the AUC-ROC as their splitting criterion.[9, 10] Another example of binary decision tree construction algorithm based at c-statistics is developed by Hossain et al. (2008). These authors used an AUC measure to select a node based on its classification performance and then used the misclassification rate to choose a split point.[11] In our study, we adapted the idea of ROC-tree as a form of tree which divides the classification process at a number of smaller steps which are intuitive and generally easily interpretable.[12] However, we used the Youden index (Bookmaker Informedness) for the determination of the optimal cut-off point. The misclassification rate as a complement of accuracy (one can be calculated from the other) can be misleading when the data are imbalanced,[13] because of the dominating effect of the majority class.[5]

The Youden index, in contrast to the accuracy, directly includes a true positive and a true negative rate. This index is recognized as suitable performance metrics of the classification of imbalanced datasets.[14]

The selection of performance metrics is another issue considered in this study. Accuracy and error rate, sensitivity and specificity are the most often used metrics for summarizing the performance of classification models. Comparing different classifiers using these measures is easy, but it has many problems such as the sensitivity to imbalanced data and ignoring the performance of some classes.[13, 15–17] Class imbalance is one of the significant issues which affect the performance of classifiers[18] The determination of the most suitable performance metrics is a major issue in the classification of class

imbalanced datasets.[14, 18] In imbalanced datasets, not only is the class distribution skewed, the misclassification cost is often uneven too. The minority class examples are often more important than the majority class examples.[5]

It is recommended to consider a combination of different measures instead of relying on only one measure when dealing with class-imbalance data.[13] Hybrid threshold metrics, such as the Geometric Mean[14, 19] or the Bookmaker Informedness[14] showed to be useful as performance metrics for imbalance datasets. The F-measure (harmonic mean) is also recommended as the measure in this case.[19] However, it still completely ignores true negatives which can vary freely without affecting the statistic. [20] The Matthews correlation coefficient (MCC) described as least influenced by imbalanced data.[13]

In our study, we used hybrid measures for comparison of classification algorithms. The macro-average of the Youden index as a metric of discriminative power[21] was significantly higher for the ROC-tree algorithm in the one-vs-one comparison (Table 3). Also, other hybrid threshold metrics such as optimized precision, geometric mean had difference with higher values of macro-averages for the ROC-tree algorithm in the one-vs-one comparison.

The “reproducibility” of cut-off points and metrices were tested by the 10-folds cross-validation which is more stable extension of split-sample validation.[2, 22] In this case cut-off points were determined in nine of the ten and testing in one of the ten, which is repeated ten times. In this way, all cases have served once to test the model. The performance is commonly estimated as the average of all assessments.[2] The cut-off points derived using the full dataset are accepted as unique and can be used for further evaluation.[23]

Study limitations. Extending the number of studied datasets could increase the power of derived conclusions. The power of conclusions could also be increased by including more known confusion table derivates which could lead to the selection of most effective combination of classification performance metrics. We defined an optimal cut-off point in ROC analysis using the Youden index. However, a comparision of stabilty of cut-off points computed by other known methods could help in selecting optimal metrics for the determination of the splitting point.

The effects of sampling techniques such as down-sampling with reducing the number of samples in majority class and the assessment the differences in proportion of minority class in datasets were not evaluated in our study. However, these methods are known and recognized as effective in machine leaning fields. To some extent, the development of *‘failure to rescue’* as a quality indicator[24, 25] is an example of down-sampling in health care.

In our study, the metrics in the one-vs-one comparison of classes were computed independently for each class and then their averages were compared. These macro-averages treated all classes equally. The combination of this approach with micro-average, which aggregates the contribution of all classes, to compute the average metric, could be effective in the evaluation of the effect of the individual classes.

Conclusion

Using algorithms sensitive to the distribution patterns, e.g. ROC-tree algorithm showed a better stratification at groups by natural cut-off points determined by the data set composition which is more convenient for summarizing or analyzing data in a various fields of sciences. In health care described algorithm allows effective evaluation of mortality causes and quality control specialized medical care by hospitals.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the Kaggle repository, https://www.kaggle.com/mlg-ulb/creditcardfraud

Competing interests

The authors declare that they have no competing interests.

Funding

this work was not supported by any funding

Authors' contributions

YG: Conceptualization, Methodology, Writing – original draft;

PB: Data curation, Writing-original draft;

JM: Supervision, Writing – review & editing

PS: Project administration, Supervision, Writing – review & editing

All authors read and approved the final manuscript.

References

1. Freeman E and Moisen G. A Comparison of the Performance of Threshold Criteria for Binary Classification in Terms of Predicted Prevalence and Kappa. *Ecological Modelling*. 2008; 217: 48-58.
2. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2009.
3. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing,, 2018.
4. Wang X, Wang F, Hu J and Sorrentino R. Towards actionable risk stratification: A bilinear approach. *Journal of Biomedical Informatics*. 2015; 53: 147-55.
5. Weng CG and Poon J. A New Evaluation Measure for Imbalanced Datasets. 2008, p. 27-32.
6. Swamidass SJ, Azencott C-A, Daily K and Baldi P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics*. 2010; 26: 1348-56.
7. Wu YC and Lee WC. Alternative performance measures for prediction models. *PLoS One*. 2014; 9: e91249.
8. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006; 27: 861-74.
9. Hand DJ and Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*. 2001; 45: 171-86.
10. Cèsar Ferri, Peter Flach and Hernández-Orallo J. Learning Decision Trees Using the Area Under the ROC Curve. *Proceedings of the 19th International Conference on Machine Learning*. Morgan Kaufmann, 2002, p. 139 - 46.
11. Hossain MM, Hassan MR and Bailey J. ROC-tree: A Novel Decision Tree Induction Algorithm Based on Receiver Operating Characteristics to Classify Gene Expression Data. *SDM*. 2008.
12. Han J, Kamber M and Pei J. *Data Mining: Concepts and Techniques*. 3d ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. , 2011.
13. Akosa JS. Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data. *Proceedings of the SAS Global Forum*. 2017.
14. Luque A, Carrasco A, Martín A and de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*. 2019; 91: 216-31.
15. Tharwat A. Classification assessment methods. *Applied Computing and Informatics*. 2018.
16. Sokolova M, Japkowicz N and Szpakowicz S. *Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*. 2006, p.1015-21.
17. Amin A, Anwar S, Adnan A, et al. Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. . *IEEE Access*. 2016: 7940-57.

18. Potolea R and Lemnaru C. A Comprehensive Study of the Effect of Class Imbalance on the Performance of Classifiers. *ICEIS (1)*. 2011, p. 14-21.
19. Hossin M and M.N S. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*. 2015; 5: 01-11.
20. Powers D and Ailab. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol*. 2011; 2: 2229-3981.
21. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950; 3: 32-5.
22. Berrar D. Cross-Validation. 2018.
23. Faraggi D and Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Stat Med*. 1996; 15: 2203-13.
24. Farjah F, Backhus L, Cheng A, et al. Failure to rescue and pulmonary resection for lung cancer. *J Thorac Cardiovasc Surg*. 2015; 149: 1365-71; discussion 71-3 e3.
25. Johnston MJ, Arora S, King D, et al. A systematic review to identify the factors that affect failure to rescue and escalation of care in surgery. *Surgery*. 2015; 157: 752-63.

Figures

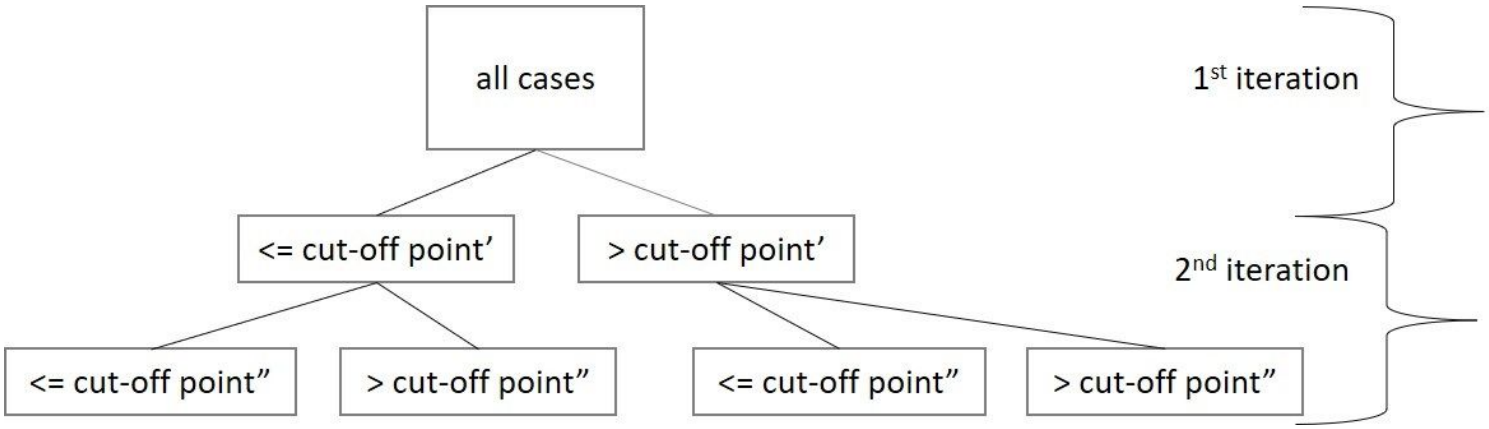

Figure 1

Flowchart shows ROC-tree algorithm of data aggregation. The division of all data at the two subgroups by Youden index followed by the second round of division at the two sequential subgroups.

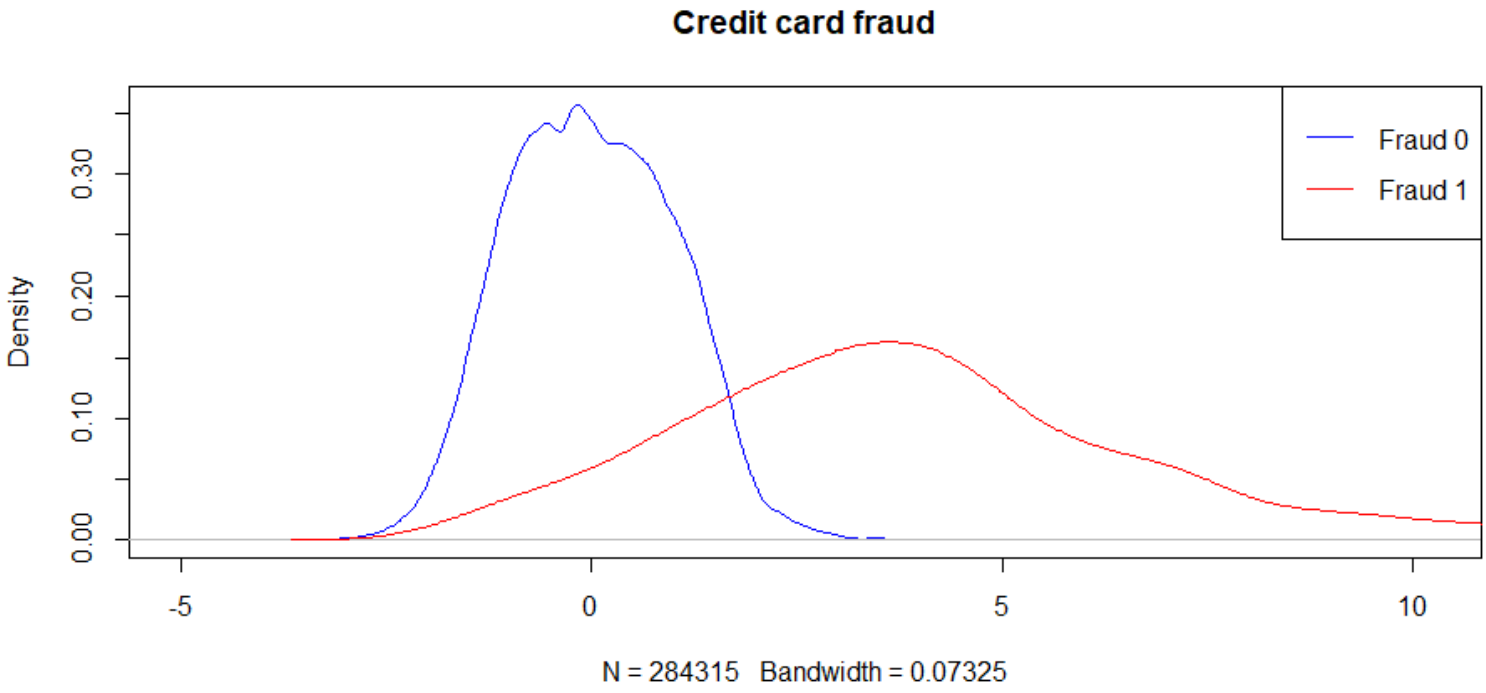

Figure 2

Kernel Density Estimation plot whole data set. V4 (mean ± std 8.32E-13 ± 1,42; minimum -5.68; maximum 16.88).

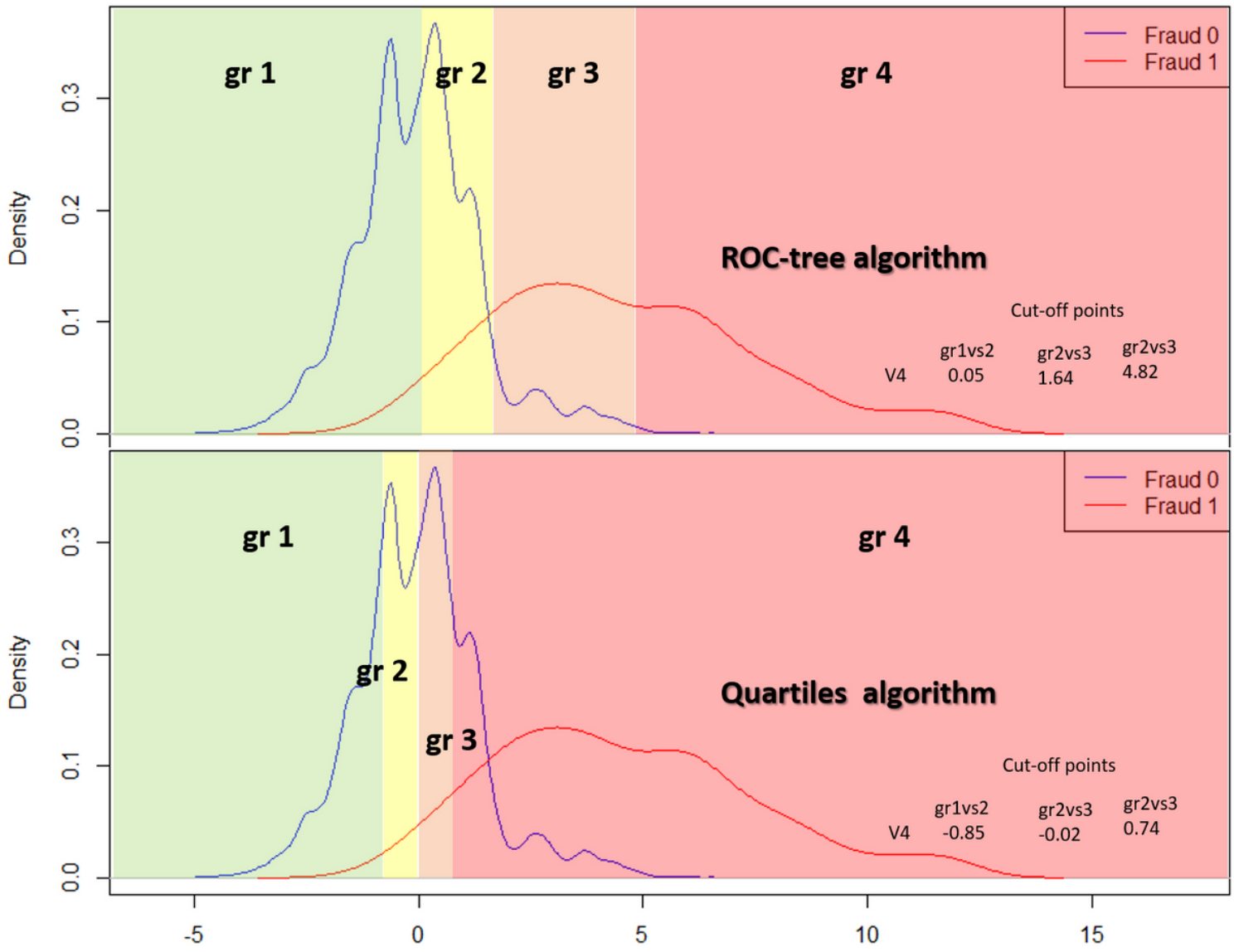

Figure 3

Cut-off points an groups areas projected at the density graph. ROC-tree algorithm (a) provides more “fair” spreading cases with the highest density of fraud+ in the third group

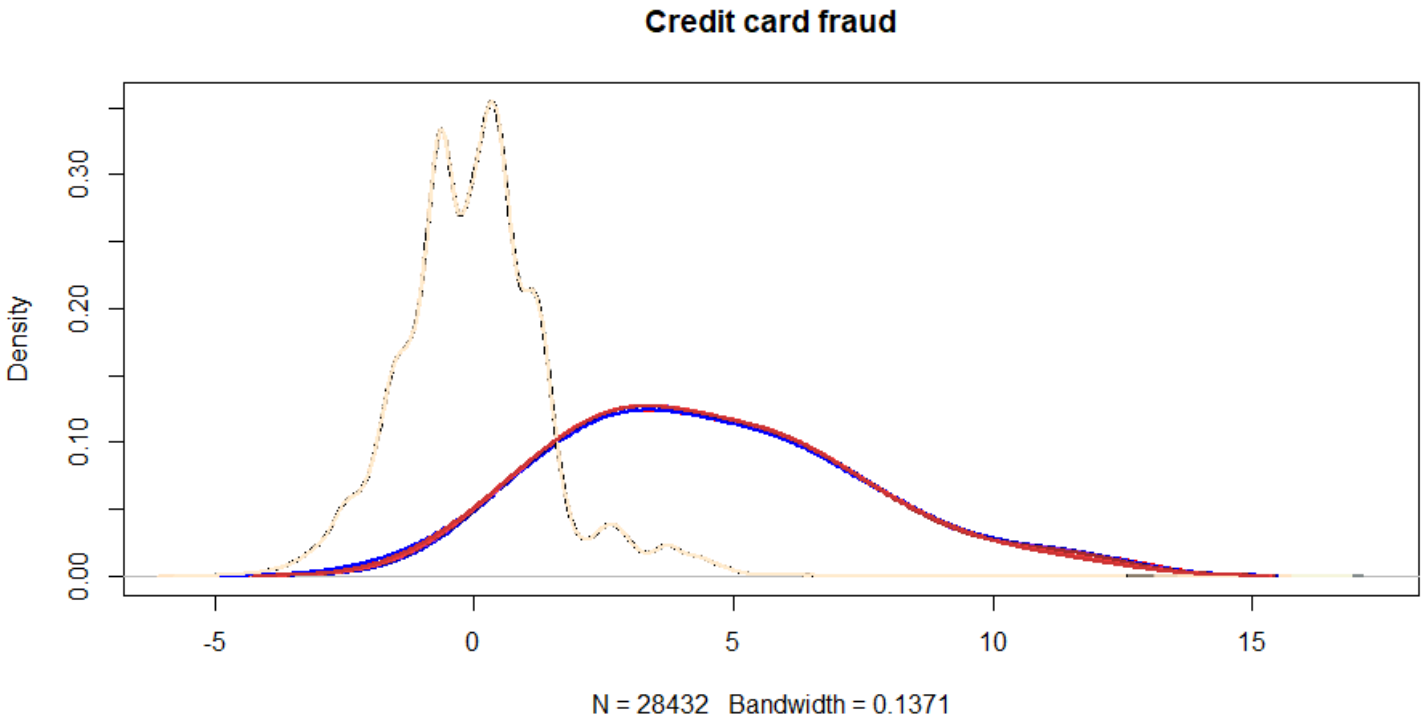

Figure 4

Density plots though 10 generated folds are similar.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixA.docx
- AppendixB.docx
- FigureA1.jpg
- FigureA2.jpg