International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

The Comparisons of Four Splitting Rules for Fitting a Classification Tree with Simulation and an Application Related to Albuminuria Data in Type 2 Diabetes Mellitus

Imran KURT OMURLU[a*], Mevlut TURE[a], Mustafa UNUBOL[b], Merve KATRANCI[a], Engin GUNEY[b]

[a]*Adnan Menderes University, Medical Faculty, Department of Biostatistics, Aydın, Turkey*

[b]*Adnan Menderes University, Medical Faculty, Division of Endocrinology, Aydın, Turkey*

[*]*Email: ikurtomurlu@gmail.com*

Abstract

The objective of this study was to compare the performances of splitting rules for predicting an ordinal response with simulation and a real data set. In the case of simulations, we compared across the methods using different sample sizes and the number of independent variables by employing the Monte Carlo simulation method. In the real data application, an analysis was performed with 265 cases. The results showed that the performances of the generalized Gini with the linear and quadratic costs of misclassification were better suited for analysis based on the gamma ordinal association measure and misclassification error rate than the other approaches. According to the gamma ordinal association measure, the generalized Gini (linear and quadratic) to the major risk factors determined for albuminuria in type 2 diabetes mellitus patients showed a slightly better performance than the other approaches. The predictive capability of splitting rules based on generalized Gini for predicting an ordinal response can be used for different sample sizes, number of independent variables and potential future suitable classification data problems. Consequently, our study will move towards choosing the generalized Gini (linear or quadratic) as the splitting rule and evaluate the data by using the Classification Trees (CT) in future studies, focusing on predicting an ordinal response.

Keywords: albuminuria; classification trees; simulation; Gini; splitting rule

---

* Corresponding author.
E-mail address: ikurtomurlu@gmail.com

1. Introduction

The dependent variable is called an ordinal because it indicates an ordering of responses. Specific methods such as Classification Trees (CT) are used to analyze the ordinal response variable with only a few possible values. Tree-based methods are the most flexible and powerful data analysis tools available to explore complex data structures [1]. As the CT is inherently non-parametric no assumptions are made with respect to the underlying distribution of the values of the predictor variables. Thus, the CT can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either an ordinal or non-ordinal structure [2]. The CT is a popular class prediction method for medical data classification. This analysis tool is especially useful for modelling diseases that have multiple contributing factors for predicting the possible effective risk factors in patients.

The CT measures the impurity of a split at a node by defining an impurity measure or splitting rule. Generally, the Gini index and classification error are used to measure the degree of the impurity. The CT is built in accordance with the splitting rule that performs the splitting of the learning sample into smaller sections [3]. Four splitting rules (ordinal impurity, ordered twoing and generalized Gini with linear and quadratic costs of misclassification) are recommended when the response variable is an ordinal scale. These methods are used to identify an optimal CT for the ordinal response variable. The rpartOrdinal R package developed by [4] implements the ordinal splitting rule methods mentioned.

In our study, we addressed the problem of the splitting rule for fitting a CT on condition that the response variable was an ordinal. The gamma ordinal association measure and misclassification error rate were used to select the best splitting rule both on simulation data and real data.

2. Material and Methods

$\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{ip})$ is a p-dimensional vector of the independent variables. $w = w_j$ represent the cases in class J which denotes the number of categories in the response variable [4].

2.1. Generalized Gini index

To measure the node impurity when the response is nominal, the Gini index is

$$i(t) = \sum_{l \neq k} \mathrm{P}(w_k|t)\mathrm{P}(w_l|t) \qquad (1)$$

where k and l are categories of the response variable and $\mathrm{P}(w_j|t)$ are the proportions of cases given the root node (t) for $j = 1, \dots J$ [2].

The equation for the Gini index can also be expressed as [5].

$$i(t) = \sum_k \mathrm{P}(w_k|t)[1 - \mathrm{P}(w_k|t)] = 1 - \sum_k \mathrm{P}^2(w_k|t) \qquad (2)$$

When the response is an ordinal the Gini index is inadequate and does not make use of the advantage to measure the impurity of the node [4,6]. To measure node impurity for an ordinal response, the generalized Gini impurity function is

$$i_{GG}(t) = \sum_{l \neq k} C(w_k|w_l)\mathrm{P}(w_k|t)\mathrm{P}(w_l|t) \quad (3)$$

$C(w_k|w_l)$ is the cost of misclassifying a class *l* case as belonging to class *k*. The cost of misclassifying is equal to 1 for all $l \neq k$ [2,4].

$s_1 < s_2 < \ldots < s_j$ denotes the scores which are assigned to the ordered categories of the response variable. By using these scores, the misclassification cost can be defined as the absolute differences between the pairs of scores. This transformation enables us to describe the linear and quadratic costs of the misclassification [7].

If $C(w_k|w_l) = |s_k - s_l|$ transformation is used, the generalized Gini impurity function with linear cost of misclassification is expressed as

$$i_{GG_{linear}}(t) = \sum_{k=1}^{J}\sum_{l=1}^{J}|s_k - s_l|\mathrm{P}(w_k|t)\mathrm{P}(w_l|t) \quad (4)$$

If the $C(w_k|w_l) = (s_k - s_l)^2$ transformation is used, the generalized Gini impurity function with the quadratic cost of misclassification is expressed as

$$i_{GG_{quadratic}}(t) = \sum_{k=1}^{J}\sum_{l=1}^{J}(s_k - s_l)^2\mathrm{P}(w_k|t)\mathrm{P}(w_l|t) \quad (5)$$

2.2. Ordinal impurity

Ordinal impurity is defined by $F(w_j|t) = \sum_{k=1}^{j}\mathrm{P}(w_k|t)$, where the cumulative distribution function of the response variable is taking the place of $\mathrm{P}(w_k|t)$ in (1):

$$i_{OI}(t) = \sum_{j=1}^{J} F(w_j|t)\big(1 - F(w_j|t)\big) \quad (6)$$

Ordinal impurity derives an ordinal response classification tree that does not require the assignment of the costs of misclassification [4,5].

2.3. Ordered twoing

The twoing index splits the categories of the response variable into two superclasses (C_{ij}), and then finds the best split on the independent variable based on the two superclasses [8]. The ordered twoing which is used for

the ordinal response variables is a modification of the twoing index. In the ordered twoing index, only the adjoining categories can be grouped [9].

The superclasses C_{ij} isare defined as

$$C_{ij} = \begin{cases} 1 & w_i = 1,\ldots,j \\ 0 & w_i = j+1,\ldots,J \end{cases} \quad (7)$$

For the node *t*, split *s* and superclasses C_j, the ordered twoing function is defined as

$$\Phi(s,t,C_j) = 2P_L P_R \left(P(C_j|t_L) - P(C_j|t_R)\right)^2 \quad (8)$$

where $P_L = P(t_L) / P(t)$ is the proportion of cases in *t* sent to the left child node, and $P_R = P(t_R) / P(t)$ is the proportion sent to the right child node. The best split *s* is chosen to maximize the value of $\Phi(s,t,C_j)$ based on the independent variables [2,4].

2.4. Description of the simulation methods

Our objective in this study was to compare the gamma statistics and error rates from the ordinal impurity, ordered twoing and generalized Gini with linear and quadratic costs of misclassification. We conducted simulation studies based on p = 8, 12, 16 and 20 independent variables. The independent variables were generated from the multivariate normal distribution related to each other. Generating the multivariate data involved both the low and high correlations between the independent variables. The ordinal dependent variable with three categories was generated based on the function of the independent variables.

We compared across the methods using different sample sizes (n = 250, 500, 750, 1000) using the Monte Carlo simulation method. We did 1000 replications for each model using the methods employed by R 2.9.0. The data was then analyzed using the rpartOrdinal package [4]. The ten-fold cross validation was used to determine the best splitting rule in predicting an ordinal response. Briefly, this process involves splitting up the dataset into 10 random segments and using 9 of them for the training set and the 10th one as a test set for the algorithm.

The gamma statistics and misclassification error rates were obtained from each method for 1000 replications and the mean of the gamma statistics and misclassification error rates were recorded for each sample size.

2.5. Albuminuria data

Urinary albumin loss can be categorized into three classes based on the quantity of albumin lost. Nonalbuminuria is defined as a urinary albumin loss of 0 - 30 mg/24 hours, microalbuminuria as 30-300 mg/24 hours and macroalbuminuria as ≥ 300 mg/24 hours [10]. Diabetic nephropathy develops in 20 - 40% of diabetic individuals [11]. Microalbuminuria, a reversible phase of diabetic nephropathy, is an important finding

characterized by 30 - 299 mg/L excretion of albumin in 24-hour urine [12]. Microalbuminuria is the primary predictor for the development of diabetic nephropathy in patients with type 2 diabetes [11,13]. We analyzed the albuminuria data in type 2 diabetes mellitus from authors in [14]. Overall there were 9 predictors and 265 cases (data from 187 nonalbuminuria, 57 microalbuminuria and 21 macroalbuminuria patients) in the data. In all the cases, weight (kg), blood urea (mg/dl), serum albumin (mg/dl), age (year), creatinine clearance (CrCl) (mL/min), fasting plasma glucose (mg/dl), post-prandial plasma glucose (mg/dl), HbA1c (mg/dl) and urine creatinine (mg/dl) were assessed and documented. We drew k = 10 cross validation sets from the albuminuria data.

The gamma statistics and misclassification error rates were obtained from each method for a ten-fold cross validation and the mean of the gamma statistics and misclassification error rates were recorded for each of the CT algorithms.

3. Results

3.1. Simulations

We simulated the data generated by running for each of the four splitting rules in order to fit a classification tree to predict an ordinal response. The values averaged for the 1000 simulations were reported in Figs. 1-8 using varying sample sizes and number of independent variables. In the simulation results, the gamma statistics were between 0.1975 - 0.9547 for n = 250, 0.4767 - 0.9720 for n = 500, 0.525 - 0.974 for n = 750 and 0.653 - 0.948 for n = 1000. The misclassification error rates ranged between 0.236 - 0.584 for n = 250, 0.176 - 0.506 for n = 500, 0.173 - 0.493 for n = 750 and 0.238 - 0.444 for n = 1000.

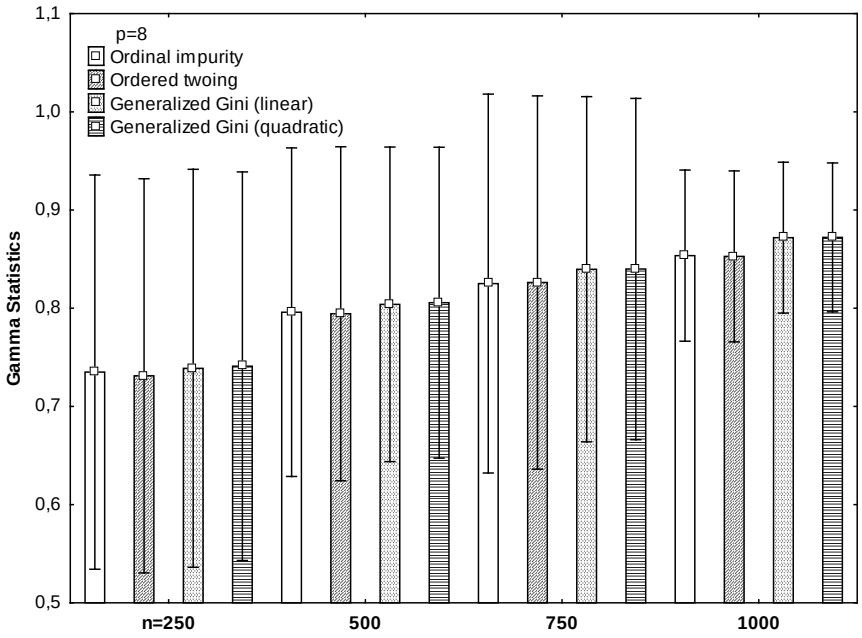

Fig. 1. gamma statistics obtained from 1000 Monte Carlo simulation for p=8 according to four splitting rules and sample size

As evident from the Figures, the generalized Gini with the linear and quadratic costs of misclassification had the biggest gamma statistics of all the sample sizes and the number of independent variables. As the sample size increased, the gamma statistics gradually increased, depending upon the number of independent variables. Similarly, the generalized Gini (linear and quadratic) was found on the smallest misclassification error rate for all of the conditions. While the sample size increased, the misclassification error rate gradually decreased, depending upon the number of independent variables.

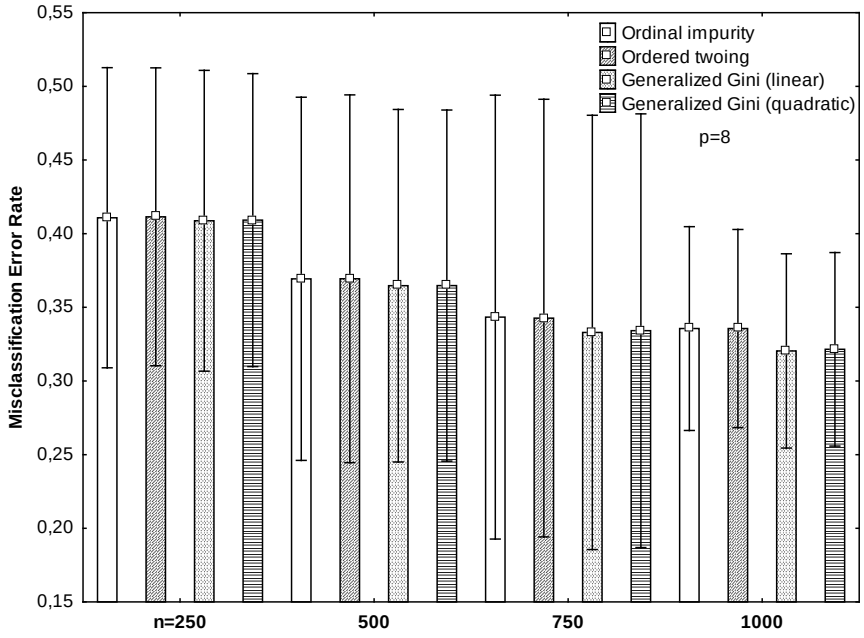

Fig. 2. misclassification error rate obtained from 1000 Monte Carlo simulation for p=8 according to four splitting rules and sample sizes

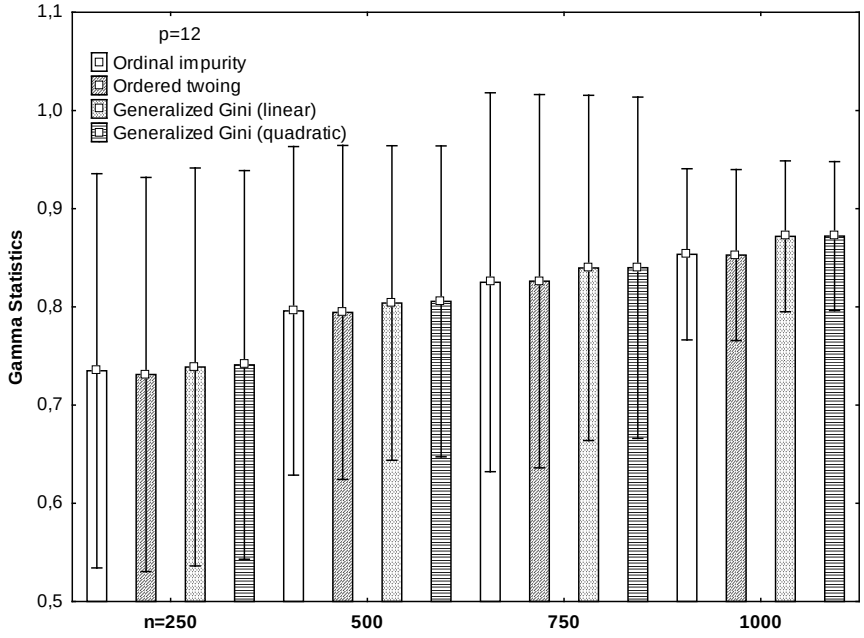

Fig. 3**.** gamma statistics obtained from 1000 Monte Carlo simulation for p=12 according to four splitting rules and sample sizes

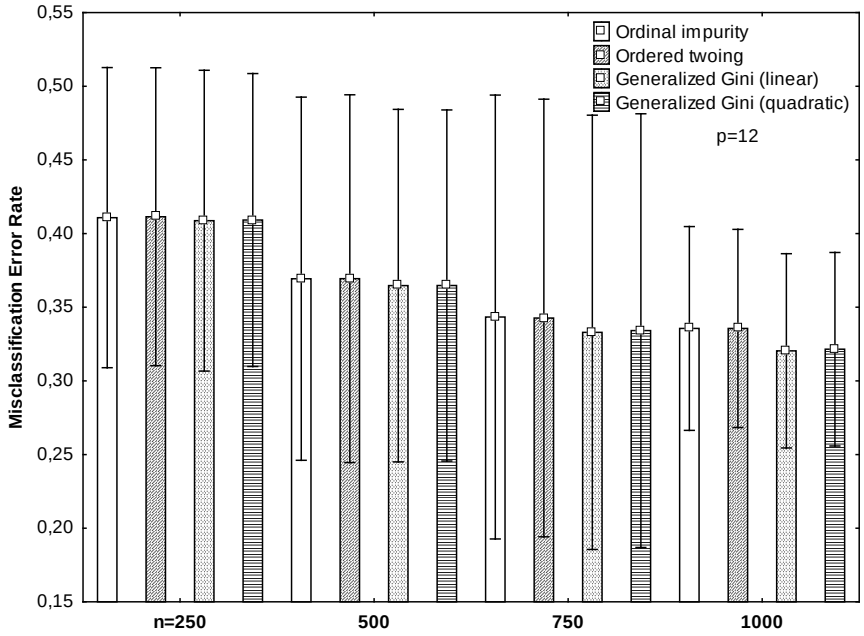

Fig. 4. misclassification error rate obtained from 1000 Monte Carlo simulation for p=12 according to four splitting rules and sample sizes

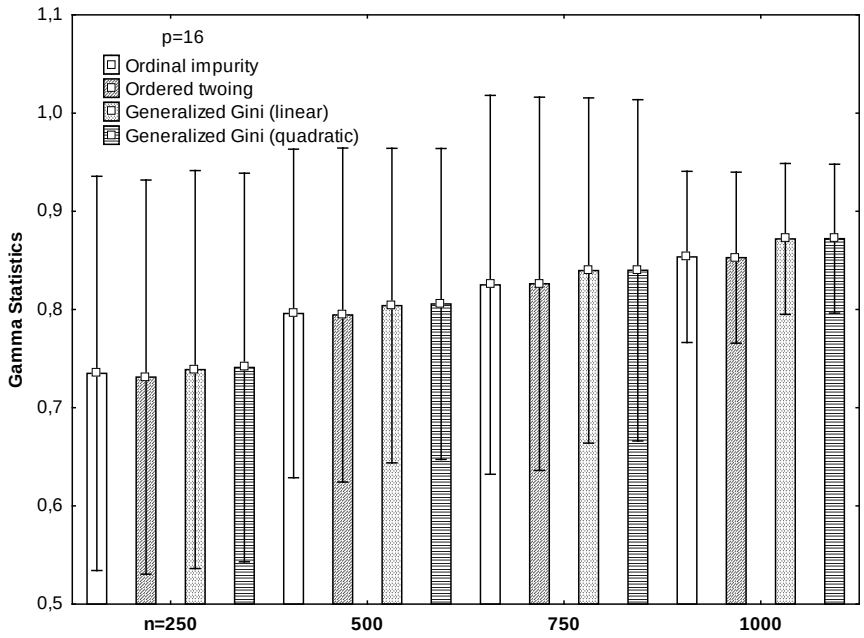

Fig. 5. gamma statistics obtained from 1000 Monte Carlo simulation for p=16 according to four splitting rules and sample sizes

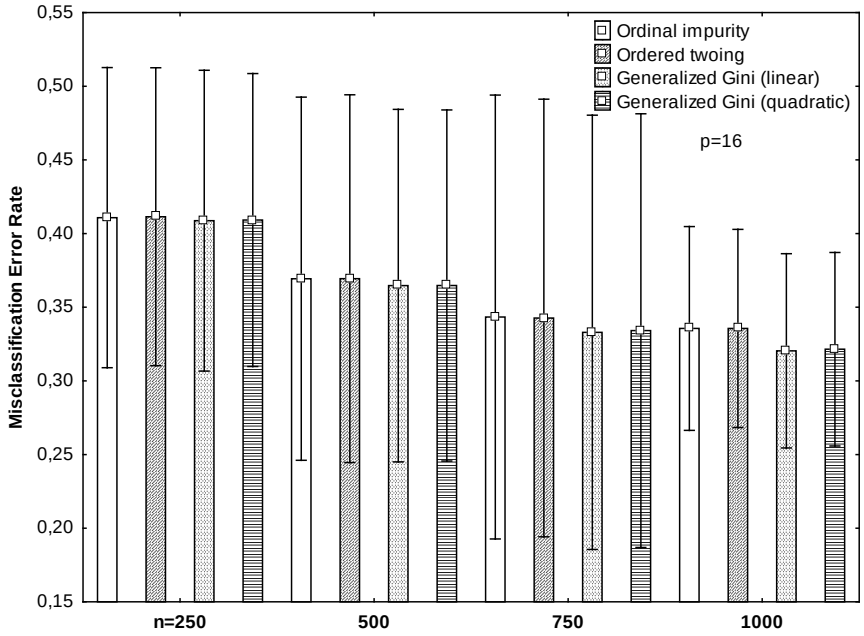

Fig. 6. misclassification error rate obtained from 1000 Monte Carlo simulation for p=16 according to four splitting rules and sample sizes

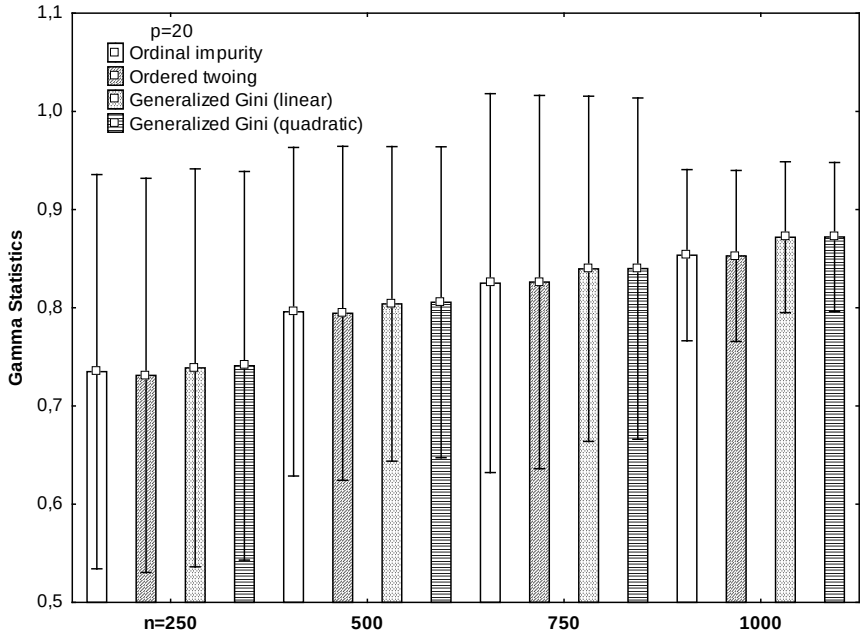

Fig. 7. gamma statistics obtained from 1000 Monte Carlo simulation for p=20 according to four splitting rules and sample sizes

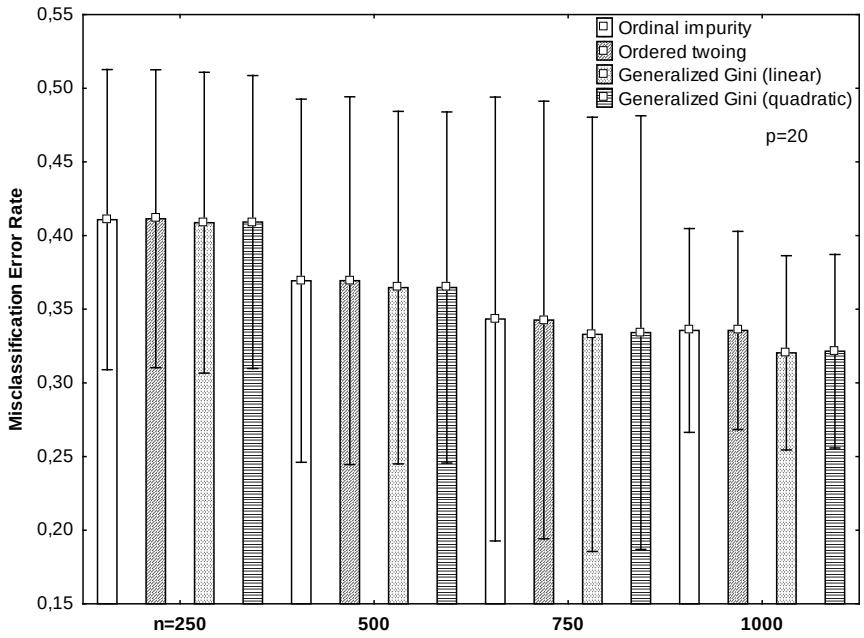

Fig. 8. misclassification error rate obtained from 1000 Monte Carlo simulation for p=20 according to four splitting rules and sample sizes

Performances of ordinal impurity and ordered twoing were almost similar for the four sample sizes. As a result the simulations showed that the generalized Gini with linear and quadratic costs of misclassification had a slightly better predictive performance for all the four sample sizes.

3.2. Evaluation conducted on albuminuria data

We proposed to discover the risk factors in the management of the albuminuria, based on the cross validation method. We used the gamma statistics and misclassification error rate for the ten-fold cross validation to monitor the prediction performance of the methods. In our study, we reported a research where several splitting rules were used for predicting the nonalbuminuria, microalbuminuria and macroalbuminuria groups based on the weight, blood urea, serum albumin, age, CrCl, fasting plasma glucose, post-prandial plasma glucose, HbA1c and urine creatinine variables.

The values averaged over the ten-fold cross validation were reported in Table 1 and Figs. 9-10 under four splitting rules. As evident, the mean of the gamma statistics ranged between 0.3060 and 0.6796. The mean of the misclassification error rates hovered between 0.2989 and 0.3849. The maximum gamma statistics value in determining the effect of the risk factors in albuminuria in type 2 diabetes mellitus patients was 0.6796 for the generalized Gini (quadratic) while the minimum misclassification error rate value was 0.2989 for the generalized Gini (linear) (Table 1).

Table 1. Gamma statistics and misclassification error rates obtained from ten-fold cross validation of four splitting rules for albuminuria data

	Gamma Statistics			
Fold	Ordinal Impurity	Ordered Twoing	Generalized Gini (Linear)	Generalized Gini (Quadratic)
1	0.7857	0.5152	0.8519	0.9231
2	0.5029	0.3613	0.7935	0.8163
3	0.1849	0.0661	0.4884	0.5929
4	0.3021	0.1414	0.5782	0.5880
5	0.5069	0.3549	0.7205	0.7149
6	0.4938	0.3532	0.6599	0.6542
7	0.5133	0.3345	0.6646	0.6592
8	0.5000	0.3375	0.6346	0.6449
9	0.4531	0.3001	0.5998	0.5942
10	0.4262	0.2953	0.6002	0.6083
Mean±sd	0.4669±0.1553	0.3060±0.1238	0.6591±0.1066	0.6796±0.1104

	Misclassification Error Rate			
Fold	Ordinal Impurity	Ordered Twoing	Generalized Gini (Linear)	Generalized Gini (Quadratic)
1	0.1538	0.3462	0.1154	0.0769
2	0.3462	0.3077	0.1923	0.2308
3	0.5385	0.5385	0.4615	0.4231
4	0.3846	0.4231	0.3077	0.3462
5	0.3462	0.4231	0.3077	0.3846
6	0.2692	0.2692	0.3462	0.3462
7	0.2692	0.3846	0.2692	0.2692
8	0.3077	0.2692	0.3077	0.2692
9	0.4615	0.5000	0.4231	0.4615
10	0.3871	0.3871	0.2581	0.2258
Mean±sd	0.3464±0.1074	0.3849±0.0907	0.2989±0.1011	0.3033±0.1128

sd: standard deviation

Findings were similar for each of the simulations and the real data set. As evident from the Figures, the generalized Gini with the linear and quadratic costs of misclassification took on the biggest gamma statistics (Fig. 9). Similarly, the generalized Gini (linear and quadratic) were found on the smallest misclassification error rate (Fig. 10).

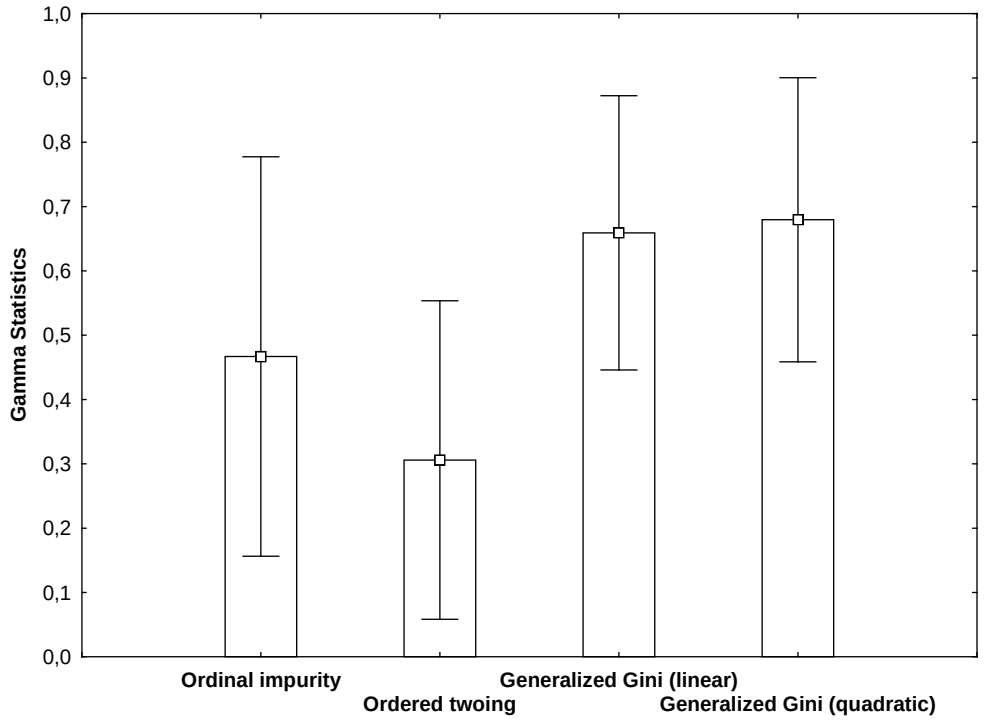

Fig. 9. gamma statistics of four splitting rules for albuminuria data

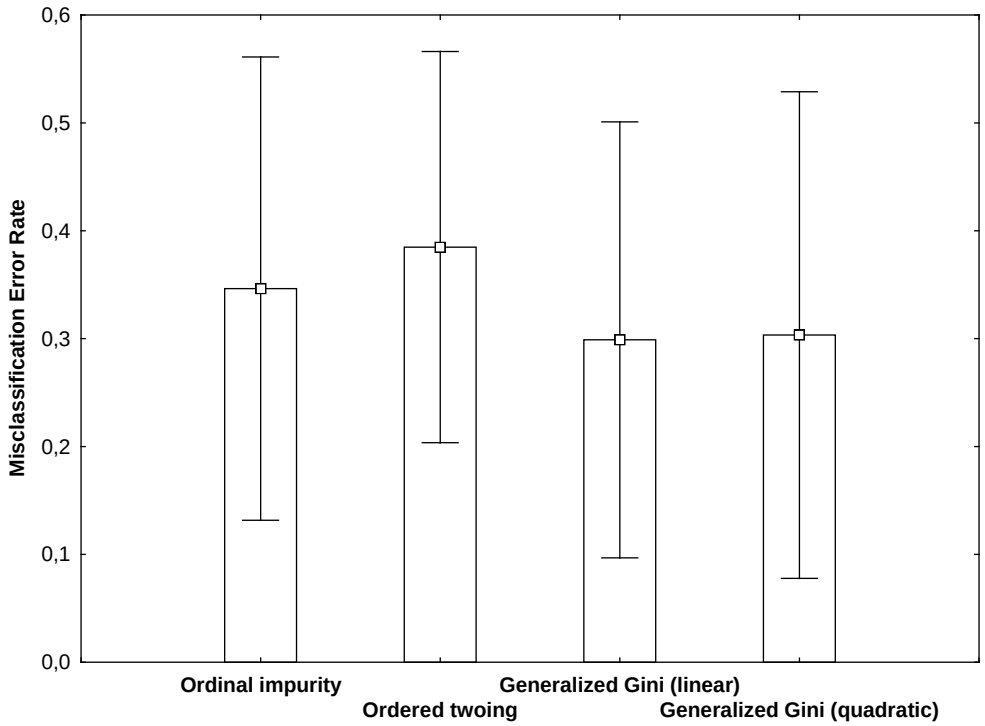

Fig. 10. misclassification error rates of four splitting rules for albuminuria data

4. Conclusion

We tried to compare across the methods with varying sample sizes and number of independent variables by using the Monte Carlo simulation method, as well as discover the risk factors and make decision rules for the management of albuminuria in type 2 diabetes mellitus patients. For this purpose, we evaluated the performance of the methods by using the gamma statistics and the misclassification error rate.

The CT uses recursive partitioning to assess the effect of specific variables on the response variable, thereby ultimately generating groups of patients with similar clinical features. The categorizing of patients into groups with differing characteristics using clinical variables generates a tree-structured model that can be analyzed to assess its clinical utility. Therefore, the CT is more suitable than the classical statistical methods. Additionally, this method probably has the potential to complement the existing statistical models and contribute to the interpretation and presentation of risk in computerized decision support systems.

A few works have been published earlier classifying the ordinal response using the CT. The authors in [15] developed a simple approach, the C4.5-ORD method, showing how standard classification algorithms are applied to the ordinal response. In the process of the classification, which comprises two stages, viz., the training process and testing process, the ordinal dataset with J classes is converted to the J-1 binary dataset. By using the probabilities of the J original ordinal classes, the class which has the highest probability is estimated. They used 29 data sets to compare the results with different methods and demonstrated that the accuracy of the decision trees can be improved by employing the C4.5-ORD method. The author in [16] established a decision tree model, considered a binary tree, by modifying the impurity measure with n-wise and top-k measures. In the study, the top-k enables building the tree based on the top choices of the response variable, and n-wise assumes that each of the ordered categories of the response variable are a discrete choice; therefore, it configures the orderings by making n-wise comparisons for all the categories. They reported that minimal cost-complexity pruning was used to identify the optimum-sized tree, while the area under the ROC curve was used to assess the performance of the classification tree. The authors in [17] proposed an Ordinal Decision Tree (ODT) method which enables the treatment of an ordinal classification as the ordering of a pair of comparisons, because the ordinal classification involves a weak ordering between the classes. Thus, in their method, the original class orderings are not important and are, therefore, not assigned to the elements in the leaf nodes. These orderings are assigned to the dominant branch splits from the node. They termed this induction strategy, the top-slicing strategy. They demonstrated that the ODT produces a simpler model with increased ordinal response prediction accuracy by comparing the results of 5 datasets to which they applied the ODT algorithms and C4.5 method. The author in [18] evaluated the performance of the Neural Networks (NN), decision tree and Logistic Regression (LR) analyses for the classification problem by comparing them based on the number of uncorrelated independent variables, types of independent variables, number of classes in the independent variables, number of classes of dependent variables and sample size on the simulated classification examples. The author in [18] reported that NN revealed the best performance in the case of complex characteristics of the condition by using uncorrelated simulated data, whereas LR performed best in the case when the number of classes of the dependent variable was small, while NN was superior to decision tree and LR in the case when the number of classes of the dependent variable was three or more. The author in [19] evaluated the use of C4.5

with the bagging and boosting method to predict the total analgesic consumption on medication-Patient Controlled Analgesia (PCA) patient dataset by comparing with the NN, support vector machine, random forest, rotation forest and naïve Bayesian classifiers. In this study, the dependent variables (total analgesic dose and PCA analgesic dose) had three ordinal categories (low, medium and high dose). The performance of the methods was compared by using the ten-fold cross validation. Consequently, decision tree-based learning revealed a better performance level than the other methods in analgesic consumption. The author in [4] compared among the performances of the ordinal splitting rule methods on two different real data sets (birth weight data set and gene expression in the B-cell acute lymphocytic leukemia data set). The performance of the classification methods were evaluated using the five-fold cross validation and gamma statistics. In both data sets, the predictive performance of the ordinal impurity methods was much better than the performance of the ordered twoing, and the generalized Gini with linear and quadratic costs of misclassification. The author in [4] reported that these splitting rule methods based on the CT are used to model the ordinal response variables for high-dimensional data sets such as gene expression data. In the present study, we found that the results were similar for each of the simulation and the real data sets. According to the gamma ordinal association measure and misclassification error rate, the generalized Gini with linear and quadratic costs of misclassification methods performed better for the prediction of the categories of the ordinal response variable than the ordinal impurity and the ordered twoing.

The tree representation in CT shows proximity to the medical reasoning and can help to structure the understanding of prediction. The CT provides a comprehensive analytic framework to reveal the optimal design of the clinical guidelines and health policy for the prevention and management of albuminuria in type 2 diabetes mellitus patients. Therefore, it is an important problem to choose the proper splitting rule and find an optimal tree among the classification trees available.

In our study, we compared the splitting rules by using the simulation and a real data set in order to provide information on the general tendency of the data structures in the data sets and thus help researchers to select the best splitting rule for solving the problems of classification in predicting an ordinal response. However, only limited data on the sufficiency of classification efforts using only one splitting rule is available. Based on these considerations, we suggest that the data should be better explored and processed using high performance modelling splitting rules. In the future researchers should avoid data assessment by using only one splitting rule irrespective of whether the focus is on albuminuria or any other clinical condition.

References

[1] H. Zhang. “Recursive partitioning and tree-based methods” in *Handbook of Computational Statistics Concepts and Methods.* Gentle JE, Härdle W, Mori Y, Ed. Berlin, Germany: Springer; 2004. pp. 814-833.

[2] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone. *Classification and Regression Trees.* New York: Chapman and Hall/CRC, 1984, pp. 93-126.

[3] R. Timofeev. “Classification and regression trees (CART) theory and applications”. M.A. thesis, Humboldt University, Berlin, 2004.

[4] K.J. Archer. "rpartOrdinal: An R package for deriving a classification tree for predicting an ordinal response". *J Stat Softw*, vol. 34, pp. 1-17, 2010.

[5] R. Piccarreta. "Classification trees for ordinal variables". *Computation Stat*; vol. 23, pp. 407-427, 2008.

[6] P. Radaelli, C.G. Borroni, M. Zenga. "Predicting ordinal classes via classification trees," in *Proc*. 58th World Statistical Congress (CPS053), Dublin, 2011, pp. 5197-5203.

[7] G. Galimberti, G. Soffritti, M. Di Maso. "Classification trees for ordinal responses in R: the rpartScore package". *Journal of Statistical Software*, vol. 47, pp. 1-25, 2012.

[8] AnswerTree™ 2.0 User's Guide. United States of America: SPSS Inc, 1998.

[9] K.P. Soman, S. Diwakar, V. Ajay. *Insight into Data Mining: Theory and Practice*. Delhi: Prentice-Hall of India Learning, 2006.

[10] P.E. De Jong, R.T. Gansevoort, S.J. Bakker. "Macroalbuminuria and microalbuminuria: do both predict renal and cardiovascular events with similar strength?". *J Nephrol*, vol. 20, pp. 375-380, 2007.

[11] American Diabetes Association. "Standards of medical care in diabetes". *Diabetes Care*, vol. 33, pp. 11-61, 2010.

[12] D.E. Busby, G.L. Bakris. "Comparison of commonly used assays for the detection of microalbuminuria". *J Clin Hypertens*, vol. 6, pp. 8-12, 2004.

[13] M. Unubol, M. Ayhan, E. Guney. "The relationship between mean platelet volume with microalbuminuria and glycemic control in patients with type II diabetes mellitus". *Platelets*, vol. 23, pp. 475-480, 2012.

[14] E. Guney, M. Unubol, V. Yazak, I. Kurt Omurlu. "Tip 2 diyabetli hastalarda albüminürisiz nefropatiyi yakalayabiliyor muyuz?" in *Proc*. 47. Ulusal Diyabet Kongresi, Antalya, 2011. pp. 123.

[15] E. Frank and M. Hall. "A simple approach to ordinal classification," in *Proc*. 12th European Conference on Machine Learning (EMCL-01), London, UK, 2001. pp. 145-156.

[16] P.L.H. Yu, W.M. Wan, P.H. Lee. "Analyzing ranking data using decision tree," in *Proc*. ECML PKDD, Hong Kong, 2008. pp. 139-156.

[17] J.W.T. Lee and D. Liu. "Induction of ordinal decision tree," in *Proc*. First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002. pp. 2220-4.

[18] Y.S. Kim. "Performance evaluation for classification methods: a comparative simulation study". *Expert Systems with Applications*, vol. 37, pp. 2292-2306, 2010.

[19] Y.J. Hu, T.H. Ku, R.H. Jan, K. Wang, Y.C. Tseng, S.F. Yang. "Decision tree-based learning to predict patient controlled analgesia consumption and readjustment". *BMC Medical Informatics and Decision Making*, vol. 12, pp. 131, 201