2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Continuous Variables Segmentation and Reordering for Optimal Performance on Binary Classification Tasks

Paulo J. L. Adeodato, Domingos S. P. Salazar, Lucas S. Gallindo, Ábner G. Sá, and Starch M. Souza

***Abstract*—It is common to find continuous input variables with non-monotonic propensity relation with the binary target variable. In other words, when taken as propensity scores, these variables do not generate unimodal Kolmogorov-Smirnov Curves. However, these variables possess highly discriminant information which could be explored by simple classifiers if properly preprocessed. This paper proposes a new method for transforming such variables by detecting local optima on the KS2 curve, segmenting and reordering them to produce a unimodal KS2 on the transformed variable. The algorithm was tested on 4 selected continuous variables from the benchmark problem of Loan Default Prediction Competition and the results showed significant improvement in performance measured by both the AUC_ROC and Max_KS2 metrics for 3 different Artificial Intelligence algorithms, namely Linear Discriminant Analysis, Logistic Regression and MultiLayer Perceptron.**

***Index Terms* — Continuous variables´ transformations, Weight of evidence, Binary decision, Monotonic propensity.**

I. Introduction

MUCH scientific investment has been made towards the development of effective Artificial Intelligence (AI) algorithms along the years [1]. The input data preprocessing however has not received that much attention. Furthermore, that effort has been mostly focused on the syntactic role of making the input data format compatible to the AI techniques requirements [2,3].

Usually, from the syntactic point of view, categorical variables either have to be converted to numerical values or encoded in a binary representation, for some AI techniques while continuous values have to be normalized and/or discretized in equidistant intervals [2,3].

Slightly more sophisticated approaches take into consideration the continuous input data distributions by normalizing its values according to limiting quantiles and/or discretizing them in equally spaced quantile intervals [2,3].

A lot more sophisticated approaches are based on the information gain in the classification task. CHAID is one of such approaches widely used for this purpose [4].

An approach developed by SAS claims to optimize the data segmentation of continuous variables [5] but that cannot be checked since it is kept as industrial secret. Weight of Evidence (WoE) is a statistical approach which also segments continuous input variables weighted by a metric very useful for logistic regression [6]. Both of these approaches are related to the algorithm proposed here by the segmentation of the continuous input variables according to a form of univariate information gain based on the target variable.

This paper presents a new approach for transforming the continuous input variables using the Kolmogorov-Smirnov Curve as information gain metrics towards the target variable. The local optima on the KS2 curve are detected, the variable is segmented and reordered in decreasing order of the information gain in each segment, thus producing new transformed **continuous** variable with maximum unimodal KS2.

This paper is organized in 5 more sections. Section II presents the related approaches for data preprocessing and describes the Kolmogorov-Smirnov (KS2) metrics for performance assessment of binary classifiers. Section III details the proposed algorithm for segmentation and interval reordering in decreasing information gain. Section IV presents the ROC Curve as another performance metrics for binary classification, the three algorithms used as classification techniques and the benchmarking data set used in the experimental project. Section V shows the results and their interpretation confirming the proposed approach produces statistically significant improvement. Section VI summarizes the paper´s main contributions, considers its impacts and discusses limitations and potential ways to solve them.

II. Related Methods for Preprocessing

Preprocessing is an important stage in any application of artificial intelligence algorithm and is a compulsory chapter on any data mining book [2,3]. General texts describe several preprocessing techniques mostly aimed at the syntactical role of making input variables data formats compatible with those of the knowledge extraction algorithms and a few standard techniques to enrich the data.

On data mining solutions, preprocessing required for syntactical compatibility affects both continuous and categorical input data variables. From the syntactical point of view, continuous input data variables need either to be discretized for some classification techniques or just to be normalized for some other techniques.

This paper focuses on the mathematical semantics that can be embedded on continuous input variables through data distributions´ based transformations. Some of these mathematical semantics preprocessing techniques (*e.g.* Principal Component Analysis) deal only with the input data

P. J. L. Adeodato is with the Centro de Informática da Universidade Federal de Pernambuco and NeuroTech S.A., Recife, Brazil (phone: +55-81-2126-8430; fax: +55-81-2126-8438; e-mail: pjla@cin.ufpe.br).

D. S. P. Salazar is with the UAEADTec da Universidade Federal Rural de Pernambuco, Recife, Brazil (e-mail: domingos_salazar@hotmail.com).

L. S. Gallindo is with the Universidade Federal de Pernambuco, Recife, Brazil (e-mail: lgallindo@gmail.com).

A. G. Sá and S. M. Souza are with Centro de Informática da Universidade Federal de Pernambuco, Recife, Brazil (e-mails: abner.gomes.sa@gmail.com and sms2@cin.ufpe.br).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

distributions while others associate them with the target variable, as a way to improve the information gain of the original input variables with respect to a specific metrics.

This section presents the two most related approaches for producing a transformed continuous input variable aiming at maximizing the univariate information gain of the original variable. As supervised learning algorithms, care should be taken with data sampling to prevent over-fitting due to the use of a posteriori information at the input variables.

A. Weight of Evidence (WoE)

The Weight of Evidence (WoE) is a preprocessing approach for enriching the information gain developed focusing on the application of logistic regression for binary decision problems. It consists on a measure of how important a given evidence, e, is for the realization of a given hypothesis, h. If the hypothesis and the evidence are independent events, then the WoE is assigned to zero. Formally, the WoE is defined as the formula [6]:

$$WoE(h:e) = log\frac{O(h|e)}{O(h)}, \quad (1)$$

where $O(h)$ and $O(h|e)$ are the prior and posterior odds of the hypothesis h respectively, defined by the formulas:

$$O(h) = \frac{p(h)}{1-p(h)}, \quad (2)$$

$$O(h|e) = \frac{p(h|e)}{1-p(h|e)}. \quad (3)$$

A straightforward application of the WoE in the continuous variable segmentation for binary classification problems is to calculate the $WoE(h:e)$ for the hypothesis $(target = 1)$. After normalizing the continuous input variable, x, and partitioning the $(0,1)$ interval into subsets $\{b_1, b_2, \dots, b_n\}$, the evidence e_i is defined as the sentence $(x \in b_i)$, for $i = 1, \dots, n$. In that case, the different evidences are flags denoting if the value of the input variable is in a given segment. Then, the WoE for the different evidences e_i, for $i = 1, \dots, n$ are calculated. Finally, by grouping the variable segments, b_i, that resulted in similar WoE, it is obtained a discrete segmentation of the continuous variable. It is expected that segments with neighbor WoE values should behave similarly as predictors of the $target$ variable, therefore they are grouped together.

However this WoE procedure has the drawback of producing a discrete/categorical set of values for the continuous input variable, therefore resulting in an information quantization loss.

The WoE specific implementation of StatSoft [7] looks for three basic relationship profiles on the original input variables, namely monotonic, quadratic and cubic.

Other problems that affect the application of the WoE are related to the requirements of 1) Normality of the reference data, 2) Independence of samples, and 3) Homogeneity of variance [8] which are not always found.

B. Binning optimization (SAS®)

The approach developed by SAS® claims to optimize the data segmentation (binning) of continuous variables [5]. It proposes an algorithm that claims to have the following features: 1) minimum and/or maximum percentage of observations per bin (good, bad, and total), 2) minimum and/or maximum bin width, 3) minimum and/or maximum number of bins, 4) minimum aggregate WoE difference between bins and 5) monotonicity of aggregate bin WoE values (increasing, decreasing, or a combination).

The original idea behind this framework is the mathematical programming model structure and implementation of constraints. The model is built by creating a ordered set, $N(x)$, of the values of a continuous characteristic variable, x.

Then, the logarithm of the posterior odds, $logO(h|e_i)$, of the hypothesis, h, $(target = 1)$ is calculated. In this case, the evidence, e_i, is assigned to the sentence $(x = n_i)$, where n_i denotes the i^{th} value of the ordered set $N(x)$, mentioned above. A binning procedure is to select particular values of N to denote the limits of the bins that will partition the continuous variable set. In that sense, the binning set, B, may be defined a $B(x) = \{b_1, b_2, \dots, b_n\}$, a partition of N, where b_i is an interval.

Similarly to the WoE approach, the next step is to calculate the aggregate logarithm of the posterior odds, $logO(h|e_{bi})$, for the hypothesis $(target = 1)$ and evidence $(x \in b_i)$. Finally, the optimization binning problem is stated as the

$$B^* = argmin_{B,z} \sum_{b \in B} \sum_{i \in b} S_i |logO(h|e_b) - z_b|, \quad (4)$$

where S_i is the number of observations of the value $n_i \in N(x)$ and z_b is an aggregate value of a decision variable for the bin b [4]. This optimization is subject to the constraints such as the bin size, the number of bins, and monotonicity $(z_{b+1} \geq z_b)$. The resulting optimized binning B^* is claimed to have all five desired properties mentioned in the beginning of the section.

This algorithm, however, by being preserved as a industrial secret, is not amenable to scientific verification due to lack of details in its published version.

III. Proposed Preprocessing Algorithm

This paper proposes an algorithm to transform a continuous input data variable into another which is optimal in terms of information gain. The algorithm segments it and reorders its segments to generate another continuous input variable that maximizes the gain in relation to the binary target variable.

Its main advantage compared to the previous two approaches is that it transforms a continuous variable into another continuous with optimal gain instead of producing a categorical one through optimal binning. Therefore, it does not produce quantization error in the variable segmentation, as the previous approaches do in the binning process for categorization.

The specific metrics chosen for information gain used in the proposed algorithm is the Kolmogorov-Smirnov (KS) Curve between the data distributions of the two classes of the target variable in binary decision problems [9]. It has been chosen because it facilitates the detection of the segmentation points (maxima and minima) and offers two metrics for gain.

Despite having been conceived as a non-parametric test for

measuring adherence of distributions to data, in credit risk assessment, the Kolmogorov-Smirnov maximum distance (Max_KS2) is used for assessing the classifier quality by measuring the dissimilarity between the data distributions along the score for the two classes. In binary classification problems, when used for measuring adherence in fitting distributions to data, the Kolmogorov-Smirnov is called KS1. When used for measuring the classifier discriminating ability, it is called KS2. To apply the KS2 to compare classifiers along all possible values of their score ranges instead of only their maximum KS2 values, an important transformation needs to be made: the KS2 curve is plotted with the normalized rank position along the abscissa axis. That is; after having the sample sorted by the score the abscissa represents the quantiles or fraction of the data sample. As a consequence, the Kolmogorov-Smirnov curve offers another important metrics; the area under the curve (AUC_KS2) [10].

Fig. 1 illustrates the KS2 curve for a continuous variable with a non-monotonic gain profile with the maximum and minimum marked for segmentation.

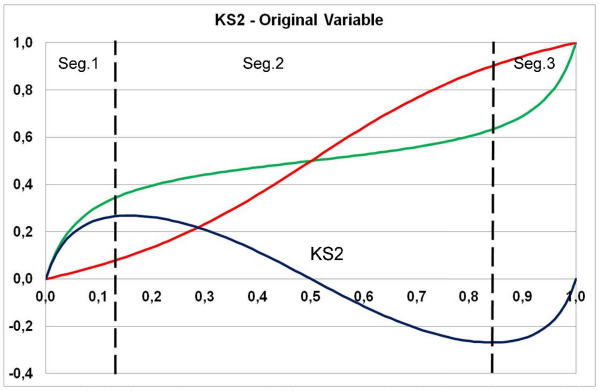

Fig. 1. Original continuous variable´s CDFs for the two classes and the KS2 curve with segmentations at its maximum and minimum.

The proposed approach recommends segmentation to produce a monotonic sequence of decreasing information gain. However, due to imprecision of the segmentation process, the less reordering the better for the transformed variable information gain. As shown in Fig. 2, segment 3 goes to position 1 instead of 2 to prevent the separation of segments 1 and 2 already in the original sequence only being shifted. Any ordering among increasing KS2 segments and among decreasing KS2 segments make the transformed variable have the same Max_KS2 which is a robust metrics against any monotonic data transformation.

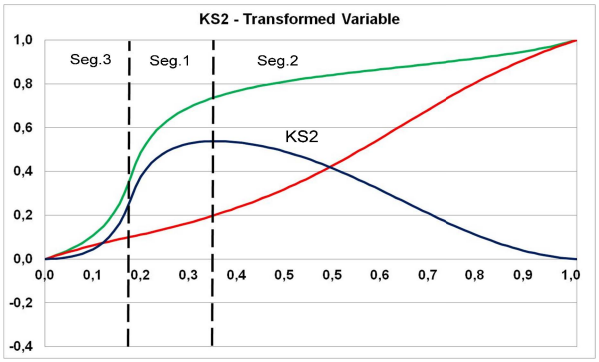

Fig. 2. Transformed continuous variable´s CDFs for the two classes and the KS2 curve with segmentations at its maximum and minimum.

The algorithm for each continuous variable can be summarized as follows:

1. Take a data sample for segmentation purposes
2. Perform bootstrapping *n* times on this data set
3. Build the average KS2 (smooth) for the continuous variable
4. Detect the KS2 maxima and minima for segmentation
5. Measure the information gain for each segment
6. Place all segments with increasing information gain on the left hand side followed by all segments with decreasing information gain, in their original order, thus producing a monotonic continuous variable
7. Apply the segmentation and reordering to the modeling and testing data sets (independent)
8. Measure the Max_KS2 for this new variable.

After applying this algorithm to all continuous variables, the transformed continuous input space is ready for training and testing the classifiers.

The next section presents the experimental procedure for validating the proposed approach.

IV. Experimental Project

The experimental project to test the algorithm used the Loan Default Prediction Competition as a benchmark problem [11], constrained to 4 continuous variables, 3 different classifiers, (Linear Discriminant Analysis, Logistic Regression and MultiLayer Perceptron) and 2 performance metrics (AUC_ROC and Max_KS2).

A. Benchmarking data set

The problem chosen to validate the approach was the competition organized by Kaggle for Imperial College London [11] this year. The problem concerned predicting the loss caused by loan default and was reduced to a binary decision problem in this paper by labeling "default" the examples which caused any loss (different from zero).

The problem had hundreds of continuous variables, mostly with non monotonic propensity towards the binary target variable. In this study, only four variables (features) were selected based on different multimodal KS2 behavior profiles. Fig. 3 below shows the KS2 curves of the four variables both in their original and transformed forms. It is clear that the transformation increases their Max_KS2 significantly.

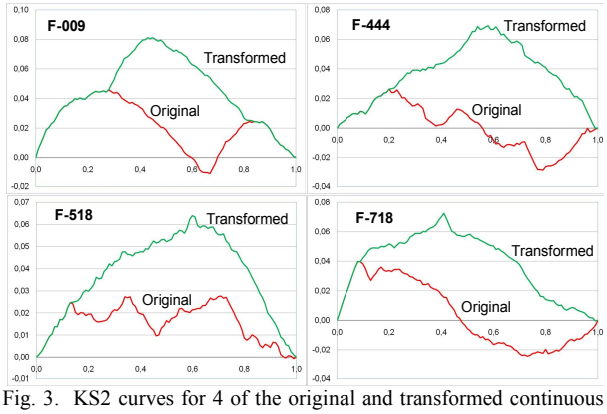

Fig. 3. KS2 curves for 4 of the original and transformed continuous variables from the benchmark, namely F-009, F-444, F-518 and F-718.

B. Classification algorithms

In decision support systems, for controllability and impact simulation, the binary decision is the result of applying a decision threshold on a propensity (or risk) score. This was the main criterion for choosing the classifiers together with their popularity in binary decision making.

Linear Discriminant Analysis (LDA) consists of a method of dimensionality reduction and feature extraction used in different domains, from early applications in corporate finance [12] to more recent in health data mining [13]. In some cases, it is used as a binary classifier although the method has some long date known limitations [14]. The idea is to optimize class separation by introducing a rotation in the continuous variable space. Formally, in the binary case, one has a set of *D*-dimensional feature vectors $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \dots, \overrightarrow{x_N}\}$, of which N_0 belongs to class "0", and N_1 belongs to class "1". In LDA, one projects all vectors $\overrightarrow{x_i}$ onto a line, by taking following inner product

$$y_i = \vec{w} \cdot \overrightarrow{x_i}, \quad (6)$$

where $\vec{w}$ is a constant vector. The goal is to find the vector $\vec{w}$ that maximize the separation of the scalars y_i. The concept of separation depends on some criterion function, $J(\vec{w})$. To account for both the mean and dispersion of the classes, in order to obtain a $\vec{w}$ that separates well samples from different classes, Fischer [15] introduced a linear functional to characterize this optimization problem, $J(\vec{w})$:

$$J(\vec{w}) = \frac{|\vec{w} \cdot (\overrightarrow{\mu_0} - \overrightarrow{\mu_1})|}{\widetilde{s_0}^2 + \widetilde{s_1}^2}, \quad (7)$$

where the mean vector of each class, 0 and 1, is defined by $\overrightarrow{\mu_0}$ and $\overrightarrow{\mu_1}$ respectively, and the distance between the resulting mean scalars is given by $|\vec{w} \cdot (\mu_0 - \mu_1)|$. The variables $\widetilde{s_0}^2$ and $\widetilde{s_1}^2$ denote the scatter, a measure equivalent to the variance, defined for $i \in \{0,1\}$ as

$$\widetilde{s_i}^2 = \sum_{y \in Class\ i} \left(y - \vec{w} \cdot \overrightarrow{\mu_i}\right)^2. \quad (8)$$

This optimization problem has a closed form, $\vec{w}'$, that can be obtained though differential calculus [15] and its solution is:

$$\vec{w}' = S^{-1}(\overrightarrow{\mu_0} - \overrightarrow{\mu_1}), \quad (9)$$

where the matrix *S* is obtained uniquely by the bilinear form relation $\widetilde{s_0}^2 + \widetilde{s_1}^2 = \vec{w}^T S\, \vec{w}$.

Logistic Regression has been successfully applied to binary classification problems, particularly to credit risk assessment. It does not require a validation set for over-fitting prevention and presents explicitly the knowledge extracted from data in terms of the coefficients (β) indicated in Equation (10), statistically validated by their significance (p).

The logistic regression technique is well-suited to study the behavior of a binary dependent variable based on a set of p independent variables x_p (explanatory features).

The logistic regression model can be expressed by the logit function

$$\log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p, \quad (10)$$

where $\pi(x)$ is defined as

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}, \quad (11)$$

and $x_1, \ldots, x_p$ are the explanatory variables.

MultiLayer Perceptron (MLP) neural network is a non-linear classifier composed of fully connected layers of neurons with weights in each connection [16]. In this paper, the MLP architecture had a single hidden layer with 5 neurons and was trained with the error back-propagation algorithm. The weights were initialized with values randomly drawn from a uniform distribution between -10^{-4} and $+10^{-4}$ for symmetry breaking, trained with a learning rate of 0.01 and stopped by cross-validation on the hold out data set.

C. Performance metrics

Considering that for general applications in decision support systems the operating point (decision threshold) is not fixed, the performance metrics should assess a feature of the classifier itself. Therefore, the area under the ROC curve [17] and the maximum KS2 distance are among the most widely accepted performance metrics for binary classifiers. Despite known drawbacks of these metrics in relation to costs [18], this paper leaves the cost issue for BI tools to help human experts analyze the optimal operation point (decision threshold).

The performance assessment was carried out in a 30-fold cross-validation process tested with a two-sided paired t-test with a 0.99 confidence level [19].

The ROC Curve [17] is a non parametric tool that represents the compromise between the true positive and the false positive example classifications based on a continuous output along all its possible decision threshold values (the score). In medical scenarios, the ROC curve equivalently expresses the compromise between sensitivity and specificity (actually, 1- specificity). There are two metrics usually extracted from the ROC curves: The minimum distance from the ROC curve to the upper left corner (ideal point) which is a performance indicator constrained to a single operation point and the Area Under the ROC Curve (AUC_ROC) which is used for assessing the performance throughout the whole continuous range of scores [17]. Considering binary decision on continuous range, the bigger the AUC_ROC, the closer the system is to the ideal classifier (AUC_ROC=1). If the ROC curve of a classifier appears above that of another classifier along the entire domain of variation, the former system is better than the latter.

The Kolmogorov-Smirnov maximum distance (Max_KS2) has already been described in Section III.

V. Results and Interpretation

From the several continuous input variables available, only four with the desired non-monotonic property were used in this experiment. Fig.4 shows the KS2 curve for the variable *f-444* before and after being transformed by the proposed

algorithm. The figure shows that the monotonic transformation increases a lot the Max_KS2 distance between the two classes´ cumulative distribution functions.

One could expect that this univariate increase in performance would be transferred to a system that gathered the contributions from all isolate variable with such a feature.

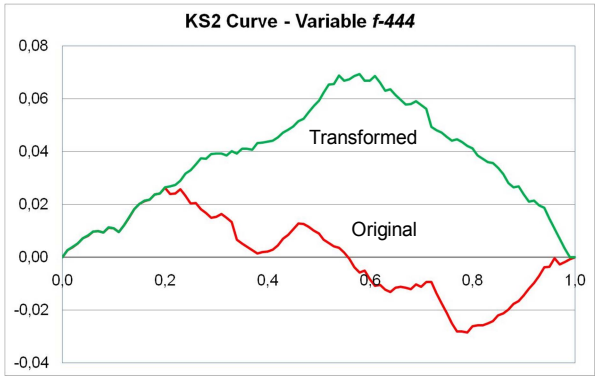

Fig. 4. KS2 curves of the *f-444* variable before and after the transformation.

And that was the idea put forward in this paper. Tables I and II below show the comparisons between each set of input variables before and after transformation for all the three techniques and for both the Max_KS2 and AUC_ROC metrics. The figures in boldface are statistically significant at the 0.01 level (α=0.99), in a 30-fold cross-validation process.

TABLE I
PERFORMANCE COMPARISON BY THE MAX_KS2 METRICS

Max_KS2	Average	StdDev	CoeffVar	p-Value
LDA Improvement	**0,06719**	0,01973	0,25255	0,00000
GLM Improvement	**0,06770**	0,01915	0,24101	0,00000
MLP Improvement	**0,01447**	0,01154	0,79717	0,00000
GLM - LDA	-0,00047	0,00358	17,09273	0,77810
MLP - GLM	**0,07526**	0,01866	0,24792	0,00000
GLM_Seg - LDA_Seg	-0,00047	0,00272	2,43568	0,77810
MLP_Seg - GLM_Seg	**0,01047**	0,01478	1,41145	0,00055

Table I shows the difference of performance measured by the Max_KS2 metrics with significant improvement in all three classifiers caused by the proposed transformation. Despite statistically significant for the MLP neural network, the increase was not large because of its inherent high capability to deal with non-linear problems. However the MLP took much longer to converge on the original variables set than on the transformed variables set.

TABLE II
PERFORMANCE COMPARISON BY THE AUC_ROC METRICS

AUC_ROC	Average	StdDev	CoeffVar	p-Value
LDA Improvement	**0,05415**	0,01193	0,22247	0,00000
GLM Improvement	**0,05482**	0,01166	0,22242	0,00000
MLP Improvement	**0,00631**	0,00509	0,80682	0,00000
GLM - LDA	-0,00027	0,00078	0,66407	0,99710
MLP - GLM	**0,05958**	0,01247	0,20926	0,00000
GLM_Seg - LDA_Seg	-0,00027	0,00050	46,41815	0,99710
MLP_Seg - GLM_Seg	**0,01344**	0,00820	0,61050	0,00000

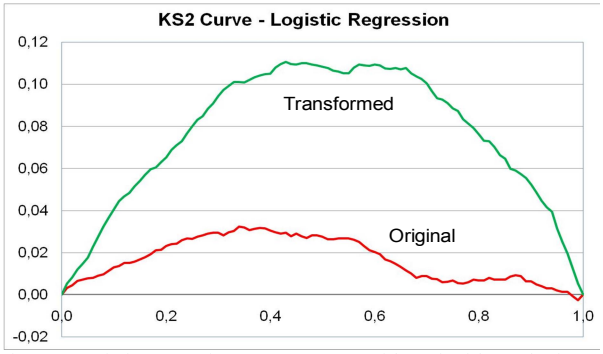

Fig. 5. Logistic regressions´ KS2 curves with and without the input variables´ transformations with the proposed approach.

Logistic regression (GLM) and Linear discriminant analysis (LDA) have no significant difference in performance. The MLP outperforms them both with or without the variables´ transformation but its advantage is much smaller after the variables´ transformation.

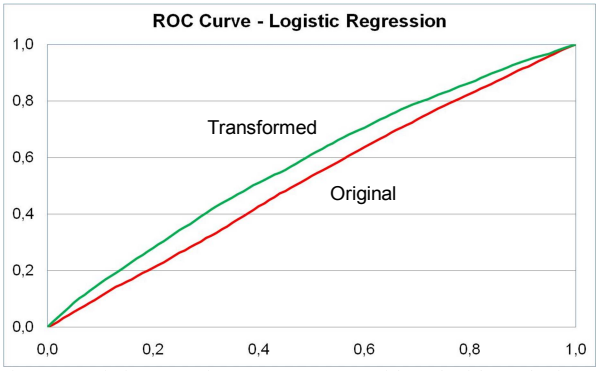

Fig. 6. Logistic regressions´ ROC curves with and without the input variables´ transformations with the proposed approach.

Table II shows the difference of performance measured by the AUC_ROC metrics also with significant improvement in all three classifiers caused by the proposed transformation.

The results are quite similar showing that these performance metrics are highly correlated.

Finally, Fig .5 and Fig. 6 below show the KS2 and the ROC curves respectively for the Logistic regression model before and after the input variables transformations. The overall gain is impressive, increasing over 50% the performance in reference to that of a random decision. Similar curves were produced for the Linear discriminant analysis and for the MLP neural network models.

VI. FINAL CONSIDERATIONS

This paper has presented a new approach for transforming continuous input variables to optimize its information gain measured by the KS2 metrics in binary decision problems without adding quantization errors.

The experiments have shown that the approach produces statistically significant improvement in performance on the binary decision benchmarking problem assessed by both

Max_KS2 and AUC_ROC metrics for the Logistic regression (GLM), Linear discriminant analysis (LDA) and the MLP neural network.

It is important to emphasize that the MLP neural network was able to solve the problem with significantly better performance than the other two techniques on the original non-monotonic variables at the cost of a much longer training time than on the transformed variables. It also benefitted from the variables transformed for significantly producing a monotonic mapping to the target variable, although in a smaller scale.

The large increase in performance for Linear Discriminant Analysis and Logistic Regression, suggests that these techniques are not well suited for problems with input variables presenting non-monotonic relationship with the target variable. That might be the reason for the success of the Weight of Evidence transformation despite its quantization error.

The proposed algorithm needs refinement in detecting the KS2 maxima and minima. At the moment, it requires a lot of repetitions in the bootstrapping process for smoothing the curve to produce a precise segmentation.

Despite the visual and intuitive appeal, mathematical formalization is being carried out to prove that the approach is theoretically sound. Furthermore, the approach has to be compared to the WoE approach to assess if the quantization error is significant for Logistic regression (GLM), Linear discriminant analysis (LDA) and the MLP neural network.

ACKNOWLEDGMENT

The authors thank Kaggle and Imperial College London for letting them use data from the Loan Default Prediction Competition they had organized in 2014. The authors also thank NeuroTech S.A. for letting them use the company´s decision support system for running the experimental project.

REFERENCES

[1] A. K. Jain, R. P. W., Duin, J. Mao, J., "Statistical Pattern Recognition: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 4-37, 2000.

[2] I. Witten and E. Frank. "*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*". Morgan Kaufmann, San Francisco, CA, 2005.

[3] J. Han, M. Kamber, and J. Pei. "Data Mining: Concepts and techniques", 3rd ed. Morgan Kaufmann, Waltham, MA, 2012.

[4] J. Magidson, *SI-CHAID 4.0 User's Guide*. Belmont, Massachusetts: Statistical Innovations Inc., ch.1, 2005.

[5] Oliveira I., Chari M., and Haller S.: SAS/OR: Rigorous Constrained Optimized Binning for Credit Scoring. In: *SAS Global Forum 2008 on Data Mining and Predictive Modeling*, SAS Institute Inc., Cary NC, USA 2008.

[6] Good, I. J. "Weight of Evidence: A brief Survey," Bayesian Statistics 2. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, (eds.). North-Holland, Amsterdam, pp. 249-269, 1983.

[7] STATISTICA Formule Guide - Weight of Evidence Module (2003). [Online]. Available: http://documentation.statsoft.com/.

[8] E. P. Smith, I. Lipkovich, and K. Ye: "Weight of Evidence (WOE): Quantitative Estimation of Probability of Impact", *Human and Ecological Risk Assessment*, vol.8, pp. 1585-1596, 2002.

[9] W. J. Conover, "*Practical Nonparametric Statistics*", Third edition, John Wiley & Sons, NY, USA, 1999.

[10] P. J. L. Adeodato, G. C.Vasconcelos, A. L. Arnaud, R. C. L. V. Cunha, D. S. M. P. Monteiro, and R. F. Oliveira Neto. The Power of Sampling and Stacking for the PAKDD-2007 Cross-Selling Problem. Int. Jour. Data Warehousing and Mining, v.4, pp. 22-31, 2008.

[11] Loan Default Prediction Competition (2014, March) Kaggle and Imperial College London [Online]. Available at: https://www.kaggle.com/c/loan-default-prediction.

[12] E. Altman, P. Narayanan. "An International Survey of Business Failure Classification Models,". *Financial Markets, Institutions and Instruments*, vol. 6, no. 2, pp. 1-57, May 1997.

[13] L .J. Hargrove, E. J. Scheme, K.B. Englehart and B.S. Hudgins, "Multiple Binary Classifications via Linear Discriminant Analysis for Improved Controllability of a Powered Prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 1, pp. 49-57, Feb. 2010.

[14] J. Ohlson. "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of Accounting Research*, vol. 18, no. 1, pp. 109–131, Spring 1980.

[15] R. A. Fisher. "The Use of Multiple Measurements in Taxonomic Problems,". *Annals of Eugenics* 7 (2): 179–188, 1936.

[16] D. E. Rumelhart, and L. J. McClelland, "*Parallel Distributed Processing: Explorations in the Microstructure of Cognition*", vol.1, MIT Press, Cambridge, MA, USA, 1986.

[17] F. Provost, T. Fawcett, "Robust Classification for Imprecise Environments." *Machine Learning Journal*, vol.42, n.3. pp. 203-231, March 2001.

[18] D. J. Hand. "Measuring classifier performance: a coherent to the area under the ROC curve". *Machine Learning* vol.77, issue 1,pp. 103–123, 2009.

[19] C. H. Goulden, "*Methods of Statistical Analysis*", second edition, New York: Wiley, pp. 50-55, 1956.