ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/261351467

Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms

Conference Paper · July 2011

DOI: 10.1109/IJCNN.2011.6033219

CITATIONS
21

READS
2,128

3 authors:

Eftim Zdravevski
Ss. Cyril and Methodius University in Skopje
157 PUBLICATIONS **1,426** CITATIONS

SEE PROFILE

Petre Lameski
Ss. Cyril and Methodius University in Skopje
102 PUBLICATIONS **926** CITATIONS

SEE PROFILE

Andrea Kulakov
Ss. Cyril and Methodius University in Skopje Macedonia
85 PUBLICATIONS **762** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Feature engineering of time series data View project

PhD thesis View project

All content following this page was uploaded by Eftim Zdravevski on 12 July 2015.

The user has requested enhancement of the downloaded file.

Weight of evidence as a tool for attribute transformation in the preprocessing stage of supervised learning algorithms

Eftim Zdravevski[1], Petre Lameski[1], Andrea Kulakov[2]

[1]NI TEKNA – Intelligent Technologies, Negotino, Macedonia
{eftim.zdravevski, petre.lameski@ni-tekna.com}
[2] Ss Cyril and Methodius University, Faculty of Computer Sciences and Engineering, Skopje, Macedonia
{andrea.kulakov@finki.ukim.mk}

Abstract

Transformation of features is a common task in the data preprocessing stage while solving data mining and classification problems. Many classification algorithms have preference of continual attributes over nominal attributes, and sometimes the distance between different data points cannot be estimated if the values of the attributes are not continual and normalized. The Weight of Evidence has some very desirable properties that make it very useful tool for the transformation of attributes, but unfortunately there are some preconditions that need to be met in order to calculate it. In this paper we propose a modified calculation of the Weight of Evidence that overcomes these preconditions, and additionally makes it usable for test examples that were not present in the training set. The proposed transformation can be used for all supervised learning problems. At the end, we present the results from the proposed transformation and discuss the benefits of the transformed nominal and continual attributes from the PAKDD 2009 dataset. The results show that the proposed transformation contributes towards a better performance in all tested classification algorithms than the method that generates dummy (i.e. binary) variables for each value of the nominal attributes.

Keywords: data transformation, data preprocessing, weight of evidence, information value, feature selection

1 Introduction

In almost every classification task, the data preprocessing phase is the most time consuming, because it is closely related to the data itself, and as a result it can be applied in different ways. When first presented with a data set, many statisticians and analysts think that they are able to use the data directly, but this is unfortunately not the case. The first step should be to analyze the data, and then to transform them into something usable. Data transformations are used to normalize the distribution of the values of an attribute.

Although there are general guidelines about how to process and transform specific kinds of data, the same transformations are not applicable for all attributes, even if they are of the same data type.

In [Anderson, 2007] and [Witten and Frank, 2005] are given some of the methodologies for data transformations. The Weight of Evidence (WOE) is one of the tools used for the transformation of nominal attributes into continual attributes in labeled data sets i.e. in supervised learning.

WOE has some very desirable properties that make it a very useful tool for transformation of attributes, but unfortunately there are some preconditions that need to be met in order to calculate the WOE. In this paper we will present an extended applicability of WOE, which is achieved by adding an insignificant number of data points that does not change the overall distribution of the data set, but on the other hand, they facilitate its calculation even in cases when the preconditions are not met. Namely, each of the preconditions will be analyzed and discussed, and for each case whenever they are not satisfied, an approximation of WOE will be proposed. This will facilitate the calculation of WOE for arbitrary types of attributes and values. Not being able to transform the nominal attributes into continual ones can be restrictive for the applicable classification algorithms, because some of them demand continual attributes that could be normalized.

In the Results section, the applicability of the proposed transformation for some of the attributes in the PAKDD 2009 data set will be illustrated.

2 Weight of evidence

Every day we make decisions based on the probability of some event to occur. Some situations are more trivial, as well as the decisions associated with them. For example, one can decide whether to take an umbrella based on how the weather looks like, or based on the weather forecast. Other decisions require information from multiple sources and are more complex. Regardless of the complexity of the situation, usually the probability of an outcome is far from empirical as it depends on more facts, which could have complex interdependencies [Chater and Oaksford, 2008]. For each decision we determine the circumstances that are

associated with it and the weight of the facts. Basically, this maps the risk associated with a particular choice or a fact on a linear scale, which aids the human brain in assessing the risk. This usually is done with the parameter named Weight of Evidence (WOE) which is discussed in more detail in [Smith et al., 2002] and [Anderson, 2007]. This is a fairly simple parameter, but yet it has a good mathematical background, which makes it a great tool for assessing the relative risk based on the available information. In binary classification problems WOE could be defined as:

$$WOE_i^A = ln\left(\frac{N_i^A/_{SN}}{P_i^A/_{SP}}\right) = ln\left(\frac{N_i^A}{P_i^A}\right) - ln\left(\frac{SN}{SP}\right) \quad (1)$$

where *SN* and *SP* are defined with eq. (2):

$$SN = \sum_{i=1}^{n} N_i^A \quad \text{and} \quad SP = \sum_{i=1}^{n} P_i^A \quad (2)$$

Eq. (1) defines the weight of evidence (WOE) of the *i*-th value of the attribute *A*, where N_i^A is the number of data points that were labeled as negative, and P_i^A is the number of data points that were labeled as positive for the *i*-th value of the attribute *A*. *SN* is the total number of negatively labeled data points, *PN* is the total number of positively labeled data points in the training set, and *n* is the total number of values for the attribute *A*.

The second part in eq. (1) illustrates that WOE is consisted of two components: a variable component which relates to the group of data points that have a particular value for the attribute whose WOE is being computed, and a constant component which relates to the whole sample i.e. to the training set. These numbers are calculated during the preprocessing phase of the data and do not depend on the classification algorithm that is going to be used.

Obviously from eq. (1) the values for $N_i^A$and P_i^A have to be different than zero, and given that they represent counts, these constraints transform to $N_i^A > 0$ and $P_i^A > 0$. Later in this section an implementation that overcomes these constraints will be described.

The following example illustrates the calculation of WOE:

Example 1*:* Let a binary training set contain 10000 negatively labeled and 40000 positively labeled data points, making a total of 50000 data points. Let one of the attributes be "Age" (in years) and let there be 70 different values for this attribute. We want to determine the WOE for the "Age" attribute when its value is 20 (years), given that there are 1000 data points that have the value 20 for the attribute "Age", from which 700 are negatively labeled, and 300 are positively labeled concerning the classification outcome. All needed parameters are stated in Table 1, and by using the eq. (1) and (2) we can calculate WOE_i^{Age}, which is illustrated by eq. (3).

$$WOE_i^{Age} = ln(700/300) - ln(40000/10000) = 0.8473 - 1.3863 = -0.539 \quad (3)$$

The obtained value shows that the particular group, to which the WOE is applied, is with a higher risk than the average, i.e. it has a negative weight of evidence. The WOE for any group with average odds is zero, because the constant and the variable portion of eq. (1) would be approximately the same.

Table 1. Values of the parameters for the calculation of WOE in Example 1

Parameter	**Description**
SP=10,000	Constant for the whole data set
SN=40,000	Constant for the whole data set
n=70	Constant for the whole data set
$P_i^{Age} = 300$	Applies only for the value 20 (years) of the attribute "Age"
$N_i^{Age} = 700$	Applies only for the value 20 (years) of the attribute "Age"

The WOE has a linear relationship with the logistic function, which makes it a well-suited tool for the transformation of an attribute when logistic regression is used.

There are few main reasons why WOE is a useful measure:

- It provides an easy and intuitive estimate of the relative risk of the different values of a particular attribute and points to the more risky groups of values.
- It can be used as a very practical tool for easy transformation of the attributes from one type to another one. This is particularly useful for transforming multivalued non numerical (i.e. nominal) attributes in numerical attributes that would have continual values (the WOE values).
- After the transformation of the attributes, the groups of values with similar relative risk could be easily noticed. This property could be used for binning multiple values into fewer groups. Using this property, the attribute "Age" in Example 1 could be reduced so it would have significantly less different groups. Using this binning, one group would correspond to multiple values that have similar relative risk, and the group would be represented by the average WOE of the values in that group.

- The information value (i.e. the predictive power) of an attribute could be estimated using the WOE of its values.

In opposition to the useful characteristics of WOE, there are few that are, in fact, drawbacks. This is something that needs to be addressed properly before applying the WOE transformation to an attribute:

- WOE does not consider the proportion of data points with a particular value of an attribute, only the relative risk. In *Example 1*, $P_i^{Age} = 300$, $N_i^{Age} = 700$ and $WOE_i^{Age} = -0.539$. The same value for WOE would be obtained if $P_i^{Age} = 3$ and $N_i^{Age} = 7$, even though these examples are very different in terms of the proportion of data points from the whole data set. This issue has to be addressed with other statistical techniques, among which is a technique named information value, covered in the next section.
- WOE measures discriminability of a single attribute, but would not capture the discriminability of an attribute in combination with another. For instance, an attribute X may generally have a low WOE for its value A, but when a second attribute Y has a value B, the combination X=A and Y=B may have high WOE and become very useful for classification. This means, that giving X=A a low WOE, should be done carefully, because sometimes this is not appropriate. A possible solution to this is to first detect the interacting attributes with an appropriate method, and then to properly model the known interactions. Non-parametric methods, such as classification trees and neural networks, are well suited for identifying interactions and dealing with them. One can use classification trees to identify the interactions, and afterwards to use another classification algorithm to build a predictive model [Thomas *et al*., 2002]. Modeling interactions can be done either by segmenting the data into groups and using different classification model for each group; or by generating interaction attributes and using one model for the whole data set.

3 Information value

In order to estimate the predictive power of a particular attribute, the measure named information value could be used [Anderson, 2007]. It is a useful measure because it can be computed in the preprocessing phase and can be used for feature selection by discarding attributes that have very low information value. Eq. (4) can be used for its calculation. Here F^A is the information value of attribute A, whereas the definition of the other parameters is the same as in eq. (1) and (2).

$$F^A = \sum_{i=1}^{n} \left[\left(\frac{N_i^A}{SN} - \frac{P_i^A}{SP} \right) \times WOE_i^A \right] \quad (4)$$

The values of F^A are always positive and can be greater than 3 for very predictive attributes. Attributes with information value less than 0.1 are usually considered as week, while those attributes with information value greater than 0.3 are sought after, and are likely to be used in the scoring models. Please note, that weak attributes may provide value in combination with others; or have individual values that could provide predictive power as dummy variables.

The first drawback of WOE mentioned at the end of the previous section is addressed with the information value in the following manner. In eq. (4) the first term of the product containing N_i^A and P_i^A, tends to zero when they are very small, implying that the whole product for the particular value of i, would tend to zero, regardless of the value of WOE_i^A. This, in fact means that such particular case will not have significant influence on the whole sum in eq. (4) i.e. on the information value of attribute A.

The information value is sensitive to the way how the attribute is grouped, and to the number of groups, but it will provide the same result, irrespective of how the values are ordered. However, it can be difficult to interpret, because there are no associated statistical tests. As a general rule, it is best to use the information value and/or chi-square test to assess individual attributes.

4 Modified calculation of WOE

As we have described previously in section 2, WOE has some properties that are well-suited for analyzing the attributes and each of their values, for attributes' transformation, or for estimating the attributes' information values. Unfortunately, there are some restrictions that need to be overcome in order WOE to be computable. The following two subsections describe these restrictions, and propose appropriate adjustments that can lead to successful and accurate calculation of WOE.

4.1 Unsatisfied preconditions

As it was mentioned previously when we defined WOE, the constraints $P_i^A \neq 0$ and $N_i^A \neq 0$ need to be satisfied in order to calculate WOE with eq. (1). If these conditions are not met for some values of an attribute, then the WOE of that attribute would not be computable. But if we want to use the weight of evidence as a tool for transforming or binning the attributes, or to estimate attributes' information values, we need the WOE of all values of the attribute. This implies that some adjustments have to be introduced so that the WOE can be calculated even when the preconditions are not met. These kinds of situations are listed below.

Case 1: *The number of positively labeled data points is zero ($P_i^A = 0$) and the number of negatively labeled data points is zero ($N_i^A = 0$).* There are no data points with the i-th value of the attribute A, so we assume that WOE_i^A is zero, meaning that this value will have no impact on any transformations nor will change the calculation of some other parameters that are dependent on WOE. In fact, this value

could be even deleted from the possible set of values for the current attribute, but since it does not have any effect on anything, it could be retained in the set of possible values in case some data points from the training data set have the *i*-th value of attribute *A*, as well as for future analysis.

Case 2: *The number of positively labeled data points is zero ($P_i^A = 0$) and the number of negatively labeled data points is greater than zero ($N_i^A > 0$).* There are no positively labeled data points, and only negatively labeled data points with the *i*-th value of the attribute *A*. We propose to add one data point that is labeled as positive, so $P_i^A = 1$, and to add the appropriate number of negatively labeled data points, so the overall ratio of the added data points will be equal to the ratio of the whole data set (*SN/SP*).Equations (5)-(11) define how the number of added data points will be calculated, as well as, the proposed estimate of WOE.

$$\frac{Added\_Positive\_Data\_Points_i^A}{Added\_Negative\_Data\_Points_i^A} = \frac{SP}{SN} \tag{5}$$

Eq. (5) can be transformed as:

$$Added\_Negative\_Data\_Points_i^A = \\ = \frac{SN}{SP} \times Added\_Positive\_Data\_Points_i^A \tag{6}$$

As we mentioned previously, we only add one positively labeled data point, so $Added\_Positive\_Data\_Points_i^A = 1$, and eq. (6) transforms to:

$$Added\_Negative\_Data\_Points_i^A = \frac{SN}{SP} \tag{7}$$

The modified P_i^A and N_i^A that include the added data points are:

$$Modified\_P_i^A \\ = P_i^A + Added\_Positive\_Data\_Points_i^A = 1 \tag{8}$$

and

$$Modified\_N_i^A = \\ = N_i^A + Added\_Negative\_Data\_Points_i^A \\ = N_i^A + \frac{SN}{SP} \tag{9}$$

If N_i^A and P_i^A in eq. (1) are substituted with their modified values, $Modified_N_i^A$and $Modified\_P_i^A$, then eq. (10) is obtained:

$$WOE_i^A = ln\left(\frac{Modified\_N_i^A}{Modified\_P_i^A}\right) - ln\left(\frac{SN}{SP}\right) \tag{10}$$

After substituting eq. (8) and (9) in (10), and after simplifying the expression, the eq.(11) is obtained which denotes the proposed estimation of WOE for cases when $P_i^A = 0$ and $N_i^A > 0$:

$$WOE_i^A = ln\left(\frac{N_i^A \times SP + SN}{SN}\right) \tag{11}$$

Case 3: *The number of negatively labeled data points is zero ($N_i^A = 0$) and the number of positively labeled data points is greater than zero ($P_i^A > 0$).* There are no negatively labeled data points, and only positively labeled data points with the *i*-th value of the attribute *A*. We propose to add one data point that is labeled as negative, so $N_i^A = 1$, and to add the appropriate number of positively labeled data points, so the overall ratio of the added data points will be equal to the ratio of the whole data set (*SN/SP*). Equations (12)-(15) define how the number of added data points will be calculated, as well as, the proposed estimate of WOE.

Eq. (5) can be transformed as:

$$Added\_Positive\_Data\_Points_i^A = \\ = \frac{SP}{SN} \times Added\_Negative\_Data\_Points_i^A \tag{12}$$

We only add one negatively labeled data point, so $Added\_Negative\_Data\_Points_i^A = 1$, $Modified\_N_i^A = 1$ and eq. (12) can be simplified to:

$$Added\_Positive\_Data\_Points_i^A = \frac{SP}{SN} \tag{13}$$

And

$$Modified\_P_i^A = \\ = P_i^A + Added\_Positive\_Data\_Points_i^A \\ = P_i^A + \frac{SP}{SN} \tag{14}$$

When the values for $Modified\_P_i^A$ and $Modified\_N_i^A$ are substituted in eq.(10), then the proposed estimation of WOE for cases when $N_i^A = 0$ and $P_i^A > 0$ is:

$$WOE_i^A = ln\left(\frac{SP}{P_i^A \times SN + SP}\right) \tag{15}$$

This subsection proposed estimations of WOE for the specific cases when it cannot be calculated using the regular formula (eq. (1)), because of unsatisfied preconditions. The estimation algorithm first adds a particular number of data points so that the number of positively and negatively labeled data points for all attributes and all their values is greater than zero. However, the very small number of data points is added with care in regards to their distribution, so

the overall distribution of the data set is not changed at all. The benefits from the proposed estimation of WOE are:

- It would be computable for all attributes and all values in the data set, meaning that WOE could be used to transform the nominal attributes into continual.
- The computed WOE could be used for binning of some values of the attributes.
- Information value of all attributes could be computed, and later it could be used in the feature selection phase.
- Many classification algorithms have preference of continual attributes over nominal attributes, and sometimes the distance between different data points cannot be estimated if the values of the attributes are nominal. The transformed attributes can be compared in terms of WOE.

The estimations could be used for nominal and continual attributes, because the unsatisfied preconditions in general pose a problem for all kinds of attributes.

However, the proposed transformation could potentially lead to incorrect results in the presence of noise. Namely, if the noise is significant, then the estimated risk of a particular value of an attribute could differ from the real risk. However, noisy data pose a serious problem for data mining, in general, and should be properly handled before applying any kind of transformation.

4.2 Unknown values in the training set

Using the eq. (1) and the proposed estimations of WOE in Section 4.1, it can be precisely calculated or estimated accurately enough for all values of all attributes in the training data set. Regardless of the transformations of the attributes, and regardless of the selected features, a classification model can be constructed that will be dependent on the training data set. As a standard procedure, the classification model is validated and tested using some different data sets than the training data set, and it is not uncommon for these data sets to contain new and unknown values. This is even more likely to happen if the classification model is deployed in a production system, or if the validation and test data sets come from different time periods than the training data sets.

If WOE is used to transform nominal attribute into continual attribute, then the new value will not pose any problem, because its WOE will be zero. However, if a continual attribute is transformed using WOE, than it is not so accurate, nor practical to assume that the WOE of some new value of an attribute is zero. The new value may be very similar to some other value of the same attribute that is already present in the data set, and in that case we may approximate that WOE of the new value is the same as WOE of the existing value. In order to use this approximation, we need to define a measure of similarity between values of an attribute. With Algorithm 1 we propose a static, but easily computable method for finding similar values. First, the values of attribute A are ordered ascending and put in a vector of unique values. Afterwards, we compute the differences between the neighboring values (which are ordered) of the vector, and the computed differences are put into a temporary vector. Then the average (denoted by avg_diff^A) and the standard deviation (denoted by $stdev\_diff^A$) of the temporary vector are computed. Finally, if V_{new}^A is the new value of attribute A, then with eq. (16) the interval of similar values is defined:

$$V_{new}^A - \frac{avg\_diff^A + stdev\_diff^A}{2} \leq V_{similar}^A \leq V_{new}^A + \frac{avg\_diff^A + stdev\_diff^A}{2} \quad (16)$$

For each attribute A
  $Values^A$= GetDistinctValues(in A, in training_data set);
  $OrderedValues^A$ = OrderAsceding(in *Values*);
  $Differences^A$ = GetNeighbourDifferences(in *OrderedValues*);
  avg_diff^A = Average(in *Differences*);
  $stdev\_diff^A$ = StDev(in *Differences*);
End for each

Fig. 1. Algorithm for estimation of similar values

If we use the proposed method, depending on the distribution of values across the whole interval of possible values for attribute A and the new value, the number of estimated similar values can vary from none to more than one. Here is how we estimate WOE in these different situations:

- In case when no similar values are found, the estimated WOE would be zero.
- In case when one similar value is found, the estimated WOE is equal to WOE of the similar value.
- If more than one similar value is found, we sum the positively and the negatively labeled data points that have one of the similar values for attribute A, and afterwards we compute WOE using eq. (1) of the similar values. The estimated WOE in general is different than WOE of the similar values and the average of their WOE.

In order to make more accurate estimation of similar values, we can cluster the values of attribute A and compute the boundaries of each cluster, so when new values appear, we can use the WOE of the cluster they belong to. However, this approach is significantly more complicated, and it should be used only in cases when there is a large number of values and when new values appear very often. In credit scoring problems, the attributes that can benefit from the more complicated approach are: income, months in residence, months in the job, payment day etc.

5 Results

In this section we present the experimental results that were obtained using the proposed transformation. We have worked on a problem from a real domain – credit risk assessment on a credit card application. Basically, a retail chain offers credit to potential clients, and they can use it to buy goods in the stores of the retail chain. This problem was the topic of PAKDD 2009 Data Mining Competition [PAKDD, 2009]. There were three available data sets for the competition: a labeled training data set with 50,000 records which are collected during one year; an unlabeled validation data set with 10,000 records which are collected during another year, but with one year gap after the first year; and a test data set with 10,000 records which are collected during a third year, but after a gap of one year after the year of the validation set. In other words, the data sets contain records for credit applications for three years in a period of five years, with one year gap between each set. The competition focused on the credit scoring model's robustness against performance degradation caused by market gradual changes along few years of business operation. Each record in the data sets consist of 30 attributes that describe a credit application, and additionally a binary label that denotes whether the customer had any defaults after the credit was approved, or not. Note that the label is not revealed for the validation and test data sets. We have investigated different techniques for transformation of continual attributes and have chosen the log transformation for some attributes which denote counts (e.g. Age, Months in the job etc.). Some new attributes were generated by clustering their values (e.g. Age group), or by combining nominal attributes (Age group + Residence type, Age group + Profession code, Marital status + Number of dependents etc.). Using iterative attribute selection, some of the original and generated attributes were discarded and a data set with an optimal subset of attributes was obtained. Because the data is collected from relatively long period of time, attribute selection should be done with consideration of attribute's stability over time, i.e. selected attributes should be stable to ensure robustness and usability of the classification models during the following years.

The stability of attributes can be analyzed by comparing the recorded counts of different values of a particular attribute in the different data sets. The following parameters were used to analyze the stability of attributes in respect to the values in the training and validation data sets: Chi-Square test ([Lancaster and Seneta, 2002] and [Rossi, 2002]), correlation coefficient [Rodgers and Nicewander, 1988], and population stability index [Karakoulas, 2004]. To investigate attribute's predictive power the parameter information value [Anderson, 2007] has been used.

Because of the limited space of the paper, here we present the analysis of only two WOE transformed attributes that were included in the classification model. The other nominal attributes were transformed in a similar manner. Tables 2 and 3 show the results from these transformation.

From the three available data sets we will use the labeled training data set, containing 50,000 records. From this data set were derived two new data sets:

- Data set 1, in which the nominal attributes were transformed into continual ones using the proposed WOE transformation. This data set contains 21 continual attributes, from which 11 were continual from the start, 9 are nominal attributes transformed into continual ones using the proposed transformation, and the last attribute is the binary label.
- Data set 2, in which the nominal attributes were transformed into continual ones by generating dummy (i.e. binary) attribute for each value of an attribute. This data set contains the same 11 continual attributes that are also in the other data set, and additionally over 1,000 binary attributes that are obtained by transforming the nominal attributes. The last attribute is the binary label.

Next, testing on the two data sets was performed using some of the classification algorithms implemented in the software package WEKA [Witten and Frank, 2005]. The results were compared using the official performance measure of the competition, AUC ROC, because it is statistically more consistent and discriminating than the accuracy. A formal proof for this is presented in [Ling Jin *et al.,* 2003]. The classification models were evaluated using 10-fold cross validation. Fig. 2 shows a comparison on both data sets of the same classification algorithms, named using the convention from WEKA. The multilayer perceptron is actually a feed-forward back-propagation neural network, and we have tested several configurations of it. On fig.2 are listed only those with 2 and 3 hidden layers. Regarding the computational performance, all cross-validations for the WOE transformed data set were performed in a matter of minutes. However, the significantly greater number of attributes in the second data set has negative impact on the computational performance and the cross-validations, and all classifiers performed significantly slower than on the first data set. Some classifiers (e.g. Logistic regression, Multilayer perceptron, Naïve Bayesian Tree etc.) did not finish at all, due to the computational and memory complexity that is implied by the vast number of binary attributes.

Fig. 2 shows that all classifiers performed better on the WOE transformed data set, than on the other alternative transformation that uses generated binary attributes.

Table 2. Transformation of attribute PROFESSION_CODE

Attribute name	PROFESSION_CODE
Transformation	Values of this attribute are substituted by their weight of evidence, as it was proposed

	in section 4.1. WOE has been discretized with step 0.1. The discretization enables easy identification of values with similar relative risk, and can be considered as automatic binning.	
Benefits	The information value has been increased, the number of different values has been significantly reduced and the stability over time of the attribute has been increased (parameters CST and PSI below).	
Attribute type	Before	Nominal
	After	Continual (Discretized)
Number of different values	Before	295
	After	36
Chi-square test(CST)	Before	709.587
	After	140.671
Correlation coefficient (CC)	Before	0.985
	After	0.995
Population stability index (PSI)	Before	0.995
	After	0.997
Information value (IV)	Before	0.183
	After	0.192

	in section 4.1. WOE has been discretized with step 0.1. The discretization enables easy identification of values with similar relative risk, and can be considered as automatic binning.	
Benefits	The number of different values has been significantly reduced, but no compromise has been made in regards to the IV, and the stability over time of the attribute has been significantly increased (parameters CST, CC and PSI below).	
Attribute type	Before	Nominal
	After	Continual (Discretized)
Number of different values	Before	31
	After	13
Chi-square test (CST)	Before	2845.303
	After	543.181
Correlation coefficient (CC)	Before	0.696
	After	0.983
Population stability index (PSI)	Before	0.884
	After	0.994
	Before	0.078
	After	0.077

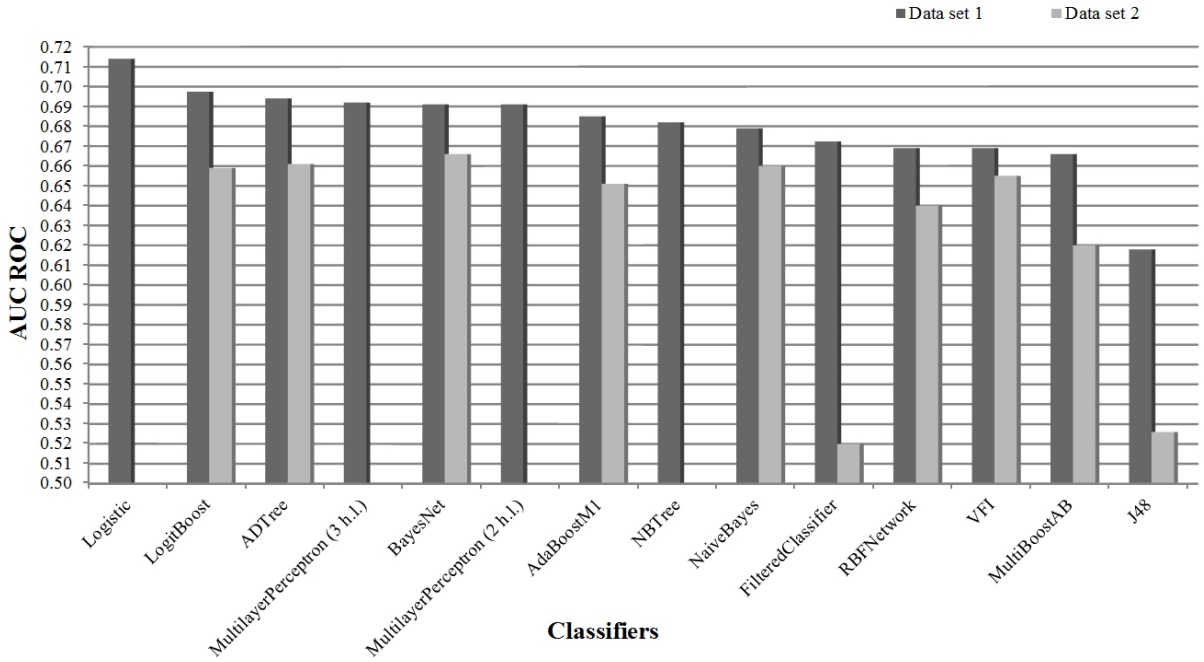

Fig. 2. Performance of the classifiers on both data sets. The left column for each classifier is the performance on Data set 1 (containing WOE transformed nominal attributes) and the right column corresponds to its performance on Data set 2 (containing generated binary attributes for each value of the nominal attributes). Some classifiers could not be evaluated for Data set 2, because of the computational and memory complexity implied by the vast number of binary attributes in Data set 2, hence there are no results for them for Data set 2.

Table 3. Transformation of attribute ID_SHOP

Attribute name	ID_SHOP
Transformation	Values of this attribute are substituted by their weight of evidence, as it was proposed

6 Conclusion and discussion

In this paper we have investigated the possibilities to modify the calculation of the measure Weight of Evidence in order to facilitate its calculation for arbitrary types of attrib-

utes and values. WOE has some properties that make it a useful tool for transformation of attributes, but unfortunately there are some preconditions that need to be met in order to calculate it.

The benefits from the proposed estimation of WOE are: that it would be computable for all attributes and all values in the data set, meaning that WOE could be used to transform the nominal attributes into continual; the computed WOE could be used for binning of some values of the attributes; information value of all attributes could be computed so it can be used as part of the feature selection process; and nominal attributes can be compared in terms of similarity distance in WOE which is quite useful for classification algorithms that require continual attributes. However, the transformation can only be applied for labeled data sets. For unsupervised problems WOE transformation would be unusable, because the data sets are not labeled and an exact relationship between the data and the label cannot be established, which is a must-have prerequisite for WOE. When the proposed transformation is applied to imbalanced data sets, the distribution of the added data points conforms to the overall distribution of the whole data set.

For continual attributes, a lot of different transformations can be applied, like introduction of dummy variables, logarithmic transformations, clustering of attributes into discrete clusters etc. The main problem is processing nominal attributes that have a fairly large number of distinct values, like those described in Tables 2 and 3. Some algorithms could not be tested before applying the proposed transformation, because the distance between values of the nominal attributes could not be estimated, or because the vast number of different values of the nominal attributes.

As a future work the proposed transformation should be tested on other data sets, preferably data sets that are well-researched. In that way, we can marginalize the influence of the attribute selection process, so the obtained results would not depend on it. Also, we have to compare our results that are based on the WOE transformation, with results that would be obtained with some other kinds of transformations.

References

[Anderson, 2007] Raymond Anderson, *The Credit Scoring Toolkit - Theory and Practice for Retail Credit Risk Management and Decision Automation*, Oxford University Press Inc., New York, 2007.

[Chater and Oaksford, 2008] Nick Chater, Mike Oaksford, Eds., The Probabilistic Mind: Prospects for Bayesian Cognitive Science, Oxford University Press, 2008.

[Karakoulas, 2004] Grigoris Karakoulas, Empirical Validation of Retail Credit-Scoring Models, *The RMA Journal*, September 2004.

[Lancaster and Seneta, 2002] H. O. Lancaster, E. Seneta, Chi-Square Distribution. Encyclopedia of Biostatistics, John Wiley & Sons, 2005

[Ling Jin *et al.,* 2003] Charles Ling Jin, Charles X. Ling, Jin Huang, Harry Zhang, "AUC: a Statistically Consistent and more Discriminating Measure than Accuracy", In Proceedings of 18th International Conference on Artificial Intelligence (IJCAI-2003), pp.329-341, Acapulco, Mexico, 2003

[PAKDD, 2009] PAKDD 2009 Data Mining Competition, http://sede.neurotech.com.br/PAKDD2009, retrieved in January 2011

[Rodgers and Nicewander, 1988] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, pp. 59–66, February 1988.

[Rossi, 2002] J. S. Rossi, Chi-Square Test, Corsini Encyclopedia of Psychology. 1. 2010.

[Smith *et al.*, 2002] Eric P. Smith, Ilya Lipkovich, Keying Ye, *Weight of Evidence (WOE): Quantitative Estimation of Probability of Impact*, Department of Statistics, Virginia Tech, Blacksburg, 2002.

[Thomas *et al.*, 2002] Thomas L.C., Edelman D.B., and Crook J.N. *Credit Scoring and its Applications*, Society for Industrial and Applied Mathematics, SIAM Publishing, Philadelphia, PA. 2002

[Witten and Frank, 2005] Ian H. Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufmann, June 2005.

View publication stats