ADVANCED TRANSFORMATIONS FOR NOMINAL AND CATEGORICAL DATA INTO NUMERIC DATA IN SUPERVISED LEARNING PROBLEMS

[1]Eftim Zdravevski [2]Petre Lameski [3]Andrea Kulakov
Department of Information Systems, Faculty of Computer Sciences and Engineering
University Ss. Cyril and Methodius
Skopje, Macedonia
[1]eftim@finki.ukim.mk, [2]lameski@finki.ukim.mk, [3]andrea.kulakov@finki.ukim.mk

ABSTRACT

In the last decade machine learning has gained substantial interest in industry and has been applied to almost all areas for which digital data is present. Very often the available data is multivariate and contains both numeric and categorical (i.e. nominal) features. However, many machine-learning algorithms do not natively support categorical features. This is one of the reasons why the data needs to be pre-processed before a machine-learning algorithm can use it. The most common technique for transformation of nominal features into numeric is by generating dummy (binary) variables for all different values of the nominal features. Few of the drawbacks of this technique are that: it does not optimally exploit the predictive potential of the data and it can slow down many algorithms because of the potentially large number of features it can generate. In this paper we present the results of our research based on applying a new technique for data transformation that is based on the weight of evidence (WoE) parameter. We have tested the WoE technique on binary and multiclass classification problems and the results show significant improvements over the technique that generates dummy variables.

Keywords: Weight of evidence, WoE, data transformation, nominal features, categorical features, numeric features, smoothing, multiclass supervised learning, dummy variables

I. INTRODUCTION

Classification problems are a subset of all data mining problems. The goal is to train an algorithm using the available data to be able to classify new and unseen data into two or more categories (i.e. classes). The available data can have origin from multiple sources and usually contains mixed types of data. Therefore there is a common procedure for data mining, as defined in [16], that defines data mining as a process consisting of the following phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. The data preparation phase is usually the most time consuming phase as it can take more than half of the time of the project. The main reason for this is because the nature of the data dictates which transformations are most suitable. Additionally, data can contain outliers, noise, errors or missing values that need to be properly addressed.

Data transformations can be used to normalize the distribution of the values of a feature, or to transform into a more suitable form for processing, usually into numeric values. Pre-processing can be performed in different manners and as a result different trained templates (e.g. classification models) with different performance can be built. Most challenging types of data regarding data transformations are text and categorical (i.e. nominal) types. 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 [1], [2], [3] and [4] are described some methodologies for data transformation. When numeric data (continuous or discrete) needs to be transformed, the range of available transformations is wider. These transformations have been extensively researched and successfully applied on various data coming from different sources. This paper addresses the transformations for nominal data, which are not as extensively researched. For this kind of data the values of a certain feature are not ordered and cannot be uniquely mapped into numeric values. In fact, the choice of transformation to be applied is made by an expert with experience in the area from which the data originates. The problem with this approach is that is subjective and it cannot be automated.

One of the most widely used techniques for transformation of nominal features is the one that generates dummy variables (features). The main advantage of this technique is that it does not depend of the origin or the nature of the data and it can be easily implemented. By using this technique from a nominal feature that has n different values n new dummy features are generated. Each of these artificial features can have a value of 0 or 1, depending on the occurrence of a particular value of the original feature. This approach has been published for the first time in [5] and was mainly used in regression analysis. However over time, it has been added to many software packages as a common stage before applying various machine-learning algorithms. When the number of nominal features and the number of different values they can have is small, this transformation leads to good performance of the algorithms. The problems arise when there are a lot of nominal features that can have

many different values. These kinds of situations lead to rapid increase of the number of generated dummy variables, which in turn, slows down machine-learning algorithms. In fact, the memory requirements or time complexity of algorithms can expand to that degree that they cannot be executed in a reasonable time on the computers that we have today. This issue can be partially addressed by discarding some of the generated features based on their predictive power. By doing that, some potentially useful information in the discarded features is consciously thrown away.

In this paper we present a new technique for data transformation based on the Weight of evidence parameter. It produces better results for classification problems with two ore more classes where the data contains nominal features. The original formulation of this technique proposed in [6], and later described in [7] and [1], can be applied to data sets that have only two classes and that conform to some preconditions. In [8] are described some enhancements to this technique so it can be used even when the preconditions are not met. In [9] this technique is additionally enhanced by providing theoretical foundations so it can be applied for numeric data that is can be present in the test data, but not present in the training data. All of these enhancements threat binary classification problems.

The rest of this paper describes an extension of this method so it can be used on multiclass datasets. The last section shows the results that were obtained by using the proposed transformation.

II. WEIGHT OF EVIDENCE

The idea for the proposed transformation originates from real life, even though it has solid mathematical formulation. 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. 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 inter dependencies, as it is described in [10]. 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.

In statistics Weight of Evidence (WoE) is defined as quantitative method for assessing the evidence in support of a hypothesis. Basic formulation of this parameter as it is described in [1], [6] and [7] is given by (1):

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

Where SN and SP are defined with (2) and (3), respectively:

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

$$SP = \sum_{i=1}^{n^A} P_i^A \quad (3)$$

Equation (1) defines the weight of evidence (WoE) of the *i*-th value of the variable *A*, where N_i^A is the number of data points (i.e. instances) that were labelled as negative, and P_i^A is the number of data points that were labelled as positive for the *i*-th value of the variable *A*. *SN* is the total number of negatively labelled data points, *PN* is the total number of positively labelled data points in the training set, and n^A is the number of different values for the variable *A*. These parameters are calculated during the pre-processing stage and are independent of which algorithm is to be applied later on.

The enhancements proposed in [8] allow WoE to be used even when the preconditions are not met. Basically, a small number of data points are artificially added during the calculation of the WoE parameters, so that the preconditions are met. The addition of artificial data points is performed with respect to the overall class distribution. The benefits of the approximation are:

- WoE can be computed for all attributes and all values in the data set, meaning that WoE could be used to transform the nominal attributes into numeric.
- The computed WoE could be used for binning of some values of the original attributes.
- The number of features can be reduced, thus improving the performance of many machine-learning algorithms.
- 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 transformation can be applied to numeric and nominal types of features.

The enhancements described in [9] enable the transformation to be applied to numeric and nominal

types of features, even if they were not present in the training dataset.
The rest of this paper presents the results on the test that were performed using this transformation.

BINARY CLASSIFICATION WHERE THE DATASETS CONTAIN NOMINAL AND NUMERIC FEATURES

Fig. 1 shows the results that were obtained using the proposed transformation on the PAKDD 2009 dataset [11]. This dataset contains 11 numeric features and 9 nominal features, and all 50000 instances are labeled with one of two classes. We have transformed this dataset using two transformations. First we have applied the proposed transformation to all 20 features, which produced a transformed dataset with 20 numeric features. The second transformation generated dummy features from all values of the 9 nominal features. The total number of features in the transformed dataset was more than 1000.
Later on both transformed datasets various machine learning algorithms were trained using 10-fold cross validation. The performances were compared in terms of AUC ROC, described in [12]. The chart on Fig. 1 shows significant improvements of the same algorithms when applied on the different datasets. In fact, some of the algorithms were not able to produce results in the case of the dataset that has dummy features, because it has too many features.

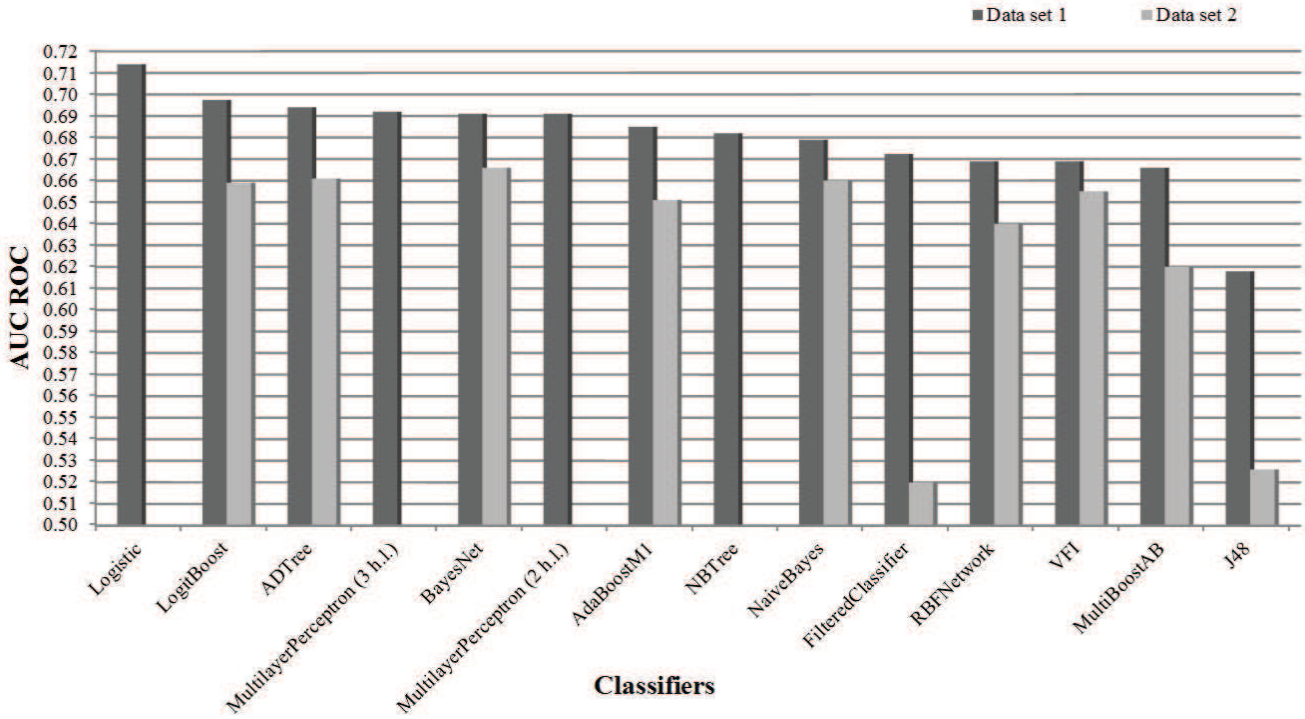

Fig 1. Results on the PAKDD 2009 dataset.

MULTICLASS CLASSIFICATION WHERE THE DATASETS CONTAIN NOMINAL AND NUMERIC FEATURES

The main contribution of this paper is the extension of the WoE technique for multiclass classification problems. In order to be able to do that, the algorithm needs to overcome the requirements of the basic definition of the WoE parameter given with Eq. 1. Namely, the latter equation demands the dataset to have only 2 classes. This can be done if the multiclass classification problem can be represented as a set of binary classification problems. In [13] and [14] is described a similar approach, named one-vs-all or one-vs-the rest, that has been used for generalization of many machine-learning algorithms that natively support only two classes (e.g. support vector machines). Using the one-vs-all technique, WoE transformation can be generalized for multiclass problems using the Algorithm 1. All steps in Algorithm 1 are repeated for each of the *n* classes.

Algorithm 1:

Foreach class C_i do

- Temporary label with *class 1* all instances that were originally labelled with class C_i

- Temporary label with *class 2* all instances that are not labelled with C_i
- Calculate the WoE parameters for all instances using the temporary labels (*class 1* and *class 2*)
- Transform all *m* nominal features using the calculated WoE parameters in the previous step. This produces *m* new numeric features.
- Add the *m* new numeric features to the transformed dataset.
- Change the temporary classes of all instances into their original classes

End

With this algorithm from *n* nominal features and *m* classes *m* x *n* new numeric features are generated. The same algorithm can be applied for transformation of numeric attributes in the original dataset as well.

The proposed transformation has been tested on the Annealing dataset, described in [15]. This dataset contains 798 instances described with 9 numeric and 29 nominal attributes labelled with 5 classes. The instances are distributed in classes as: 8, 88, 608, 60 and 34. First the dataset was transformed with the proposed WoE transformation and dataset 1 is obtained. Then the dataset was transformed by generating dummy variables from all nominal variables and dataset 2 was obtained. Both datasets were tested using a feed forward back propagation neural network. The training and test partitions of the datasets were obtained using k-fold cross validation and 2, 4, 6, 8 and 10 were used as k values. Because the instances of the partitions are chosen randomly, the performance can vary and may not be consistent. Therefore, the whole process were repeated 10 times for each value of k and each dataset. We have calculated average accuracy from the accuracy obtained from the 10 repetitions. Fig. 2 shows that with the proposed transformation there is an improvement from 5 to 10% for each value of k.

во прецизноста помеѓу двата методи и се движи од 5 до 10% во зависност од k.

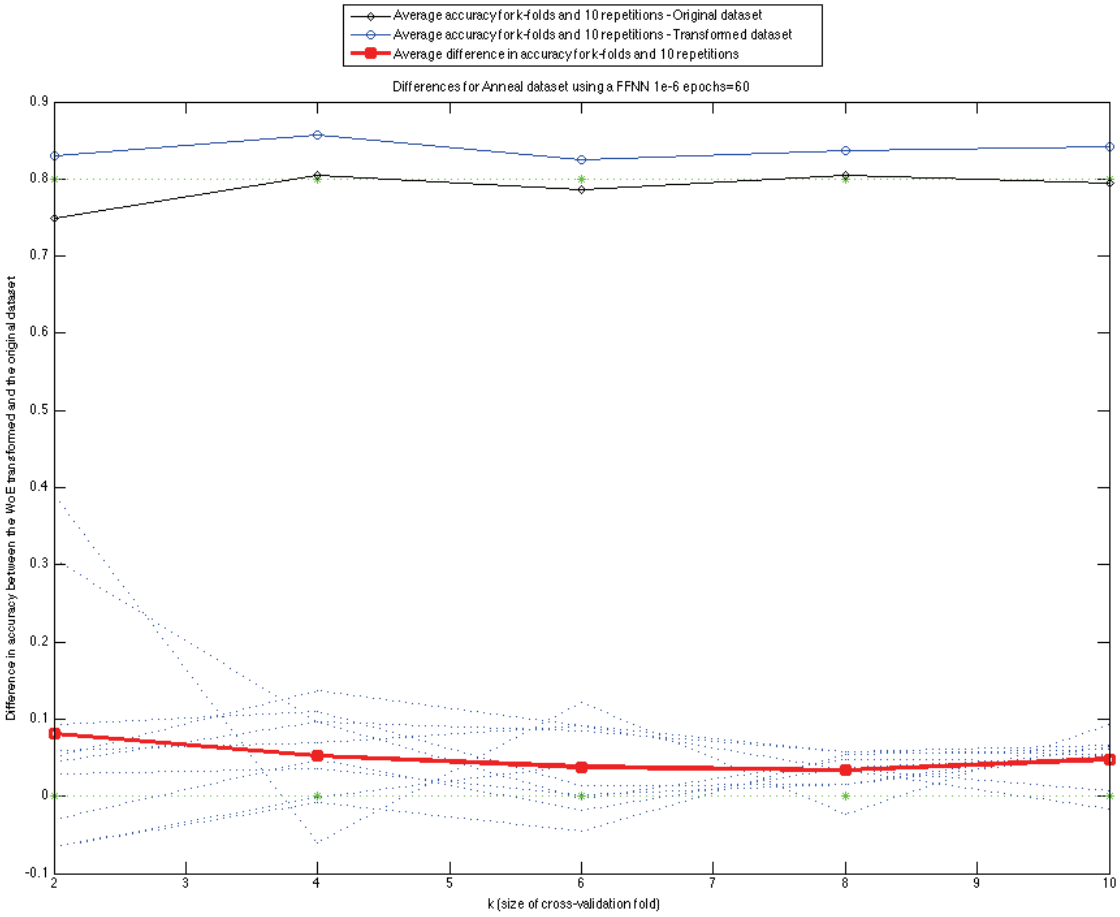

Fig 2. Results on the Annealing dataset

Conclusion

This paper addressed the issues with data transformation of nominal features. We have proposed a transformation that relies on the WoE parameter, but in addition it overcomes the constraints of the original definition of WoE, and enhances it so it can be applied to multiclass problems. The proposed transformation was applied to two datasets that contain both numeric (discrete and continual) and nominal features. At the same time, the datasets were transformed using the most common transformation for nominal features – generation of dummy features. Then we have compared the performance of a neural network algorithm using the two transformations and the results are very promising because the proposed transformation shows significant improvement.

However, there are situations when the proposed technique can generate more features than the one that introduces dummy variables and these kinds of situations should be further investigated. Also the current research has been performed using only a few machine-learning algorithms. In the future the transformation should be tested with other algorithms and other datasets and to compare our results with other published results on the same subject.

Acknowledgement: This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University in Skopje, Macedonia.

References

[1] R. Anderson, "The Credit Scoring Toolkit – Theory and Practice for Retail Credit Risk Management and Decision Automation", Oxford University Press Inc., New York, 2007.

[2] H. Witten, E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques", 2nd Edition, Morgan, 2005.

[3] Ho Chung Wu, Robert Wing Pong Luk, Kong Kam Fai Wong, Kui Lam Kwok, "Interpreting TF-IDF term weights as making relevance decisions", ACM Transactions Journal on Information Systems, Volume 26, Issue 3, June 2008.

[4] Ho Chung Wu, Robert Wing Pong Luk, Kong Kam Fai Wong, Kui Lam Kwok, "Interpreting TF-IDF term weights as making relevance decisions", ACM Transactions Journal on Information Systems, Volume 26, Issue 3, June 2008.

[5] Daniel B. Suits, "Use of Dummy Variables in Regression Equations", Journal of the American Statistical Association, Volume 52, No. 280, pp. 548-551, December 1957

[6] Irving John Good, Probability and the Weighing of Evidence, C. Griffin & Co., London, UK, 1950.

[7] E. P. Smith, I. Lipkovich, K. Ye, "Weight of Evidence (WOE): Quantitative Estimation of Probability of Impact", Department of Statistics, Virginia Tech, Blacksburg, 2002.

[8] Eftim Zdravevski, Petre Lameski, Andrea Kulakov, "Weight of Evidence as a Tool for Attribute Transformation in the Preprocessing Stage of Supervised Learning Algorithms, The 2011 International Joint Conference on Neural Networks (IJCNN), pp. 181-188, San Jose, USA, July 2011

[9] Eftim Zdravevski, Petre Lameski, Andrea Kulakov, "Towards a general technique for transformation of nominal features into numeric features in supervised learning", The 9th Conference for Informatics and Information Technology (CIIT 2012), Bitola, Macedonia

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

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

[12] C. Ling Jin, C. X. Ling, J. Huang, H. 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

[13] E. L. Allwein, R. E. Schapire, and Y. Singer. "Reducing multiclass to binary: A unifying approach for margin classifiers", Journal of Machine Learning Research, Volume 1, pp. 113-141, 2000.

[14] Ryan Rifkin and Aldebaro Klautau, "In defense of one-vs-all classification", Journal of Machine Learning Research, Volume 4, pp. 101-141, 2004

[15] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Annealing] . Irvine, CA: University of California, School of Information and Computer Science.

[16] Shearer C. „The CRISP-DM model: the new blueprint for data mining". J Data Warehousing. 2000: 13—22