QUANTITATIVE METHODS IN ECONOMICS
Volume XX, No. 3, 2019, pp. 209 – 216

Received: 04.07.2019
Accepted: 14.10.2019

ANALYSIS OF NOVEL FEATURE SELECTION CRITERION BASED ON INTERACTIONS OF HIGHER ORDER IN CASE OF PRODUCTION PLANT DATA[1]

Mateusz Pawluk https://orcid.org/0000-0003-4969-4237
Faculty of Mathematics and Information Science
Warsaw University of Technology, Poland
e-mail: m.pawluk@mini.pw.edu.pl
Dariusz Wierzba https://orcid.org/0000-0001-5829-453X
Faculty of Economic Sciences, University of Warsaw, Poland
e-mail: dariusz.wierzba@fulbrightmail.org

Abstract: Feature selection plays vital role in the processing pipeline of today's data science applications and is a crucial step of the overall modeling process. Due to multitude of possibilities for extracting large and highly structured data in various fields, this is a serious issue in the area of machine learning without any optimal solution proposed so far. In recent years, methods based on concepts derived from information theory attracted particular attention, introducing eventually general framework to follow. The criterion developed by author et al., namely IIFS (Interaction Information Feature Selection), extended state-of-the-art methods by adopting interactions of higher order, both 3-way and 4-way. In this article, careful selection of data from industrial site was made in order to benchmark such approach with others. Results clearly show that including side effects in IIFS can reorder output set of features significantly and improve overall estimate of error for the selected classifier.

Keywords: feature selection, Interaction Information Feature Selection, interactions of higher order, filter methods, information theory

JEL classification: C13, C14, C38, C44, C52

[1] The application of presented method is in scope of Research and Development project aimed at developing an innovative tool for advanced data analysis called Hybrid system of intelligent diagnostics of predictive models. The project POIR.01.01.01-00-0322/18 is co-financed by the National Center for Research and Development in collaboration with scientists from Warsaw University of Life Sciences and The Jacob of Paradies University.

https://doi.org/10.22630/MIBE.2019.20.3.20

INTRODUCTION

Nowadays, there exist multiple cases when applying feature selection to data is of critical importance – to name a few: computer vision [Zhang et al. 2018], genomic analysis [Xing et al. 2001] and natural language processing. In the context of machine learning this is an essential problem, which should be addressed at the very beginning in order to improve interpretation of a given model. Another advantage of using dimensionality reduction is a superior ability to estimate quantitatively hidden relations in data between inputs and the output. In real-life applications simplified models are often better than others, in particular they predict unseen data with lower prediction errors. Furthermore, models with decreased complexity usually take shorter time to optimize and deploy.
One of promising approaches for dealing with feature selection is set of methods based on mutual information. In particular, it is highly desirable to use such tools when there may exist complex, possibly nonlinear, dependencies in data. Due to the fact that every information-theoretic criterion belongs to a group of independent filters, no assumption of specific predictive model is made within the feature selection process. Mutual information methods have already showed successful application to classification and regression tasks.
The goal of this study is twofold – to present up-to-date overall framework for feature selection problem from the point of view of information theory and to benchmark recent achievement in the field, i.e., IIFS criterion, against one of typical business cases. In this article the proposed structure is as follows. Section „Related work" describes similar findings and introduces most common criteria based on information theory, i.e., CIFE, JMI, MIFS and MRMR. Next section „Selected data" shows the sensor data acquired from the wine factory which presents the reader a practical example of the problem. In section „Empirical study" we conduct a series of experiments including benchmark of methods for feature selection. Finally, obtained comparative outcomes are depicted in a later section „Results of research method" with closing remarks in section „Summary".

RELATED WORK

In this article we consider sequential forward feature selection methods as iterative processes. Let F be a full set of available features and S an empty output set. In each step one can compute the score for every candidate according to chosen criterion. The winning feature is usually found as one with the highest score. Afterwards, best candidate is subtracted from F and added to S. Due to this fact, such methods follow greedy approach,seeking for (sub)optimal solution in reasonable amount of time. To the best of authors' knowledge, there is minimal research devoted to heuristics proposal decreasing amount of computational burden in information-theoretic criteria.

CIFE (Conditional Infomax Feature Extraction) [Lin, Tang 2006] is using following metrics

$$J_{CIFE}(X_k) = MI(X_k, Y) + \sum_{j \in S}\left[MI\left(X_j, X_k|Y\right) - MI\left(X_j, X_k\right)\right]. \quad (1)$$

Here and in later criteria, X_k denotes candidate feature, which currently belongs to F and Y target variable. First term is responsible for evaluation of main effect: mutual information between analyzed feature and given output. Second compound term considers how much information would be added after selection of certain candidate with respect to previously chosen features, when condition on target is introduced.

JMI (Joint Mutual Information) [Yang, Moody 1999] is criterion expressed as

$$J_{JMI}(X_k) = |S|MI(X_k, Y) + \sum_{j \in S}\left[MI\left(X_j, X_k|Y\right) - MI\left(X_j, X_k\right)\right]. \quad (2)$$

Note that CIFE and JMI differ slightly only in the first term. Authors argued that for providing variability of main effect during appropriate selection process a multiplication factor is needed. Therefore, |S| represents cardinality of set in this case, which gives basic intuition of lowering second compound term influence in favor of the main effect in further algorithm iterations.

MIFS (Mutual Information Feature Selection) [Battiti 1994] has form of

$$J_{MIFS}(X_k) = MI(X_k, Y) - \sum_{j \in S} MI\left(X_j, X_k\right). \quad (3)$$

This is one of the simplest, yet very popular method, which does not incorporate complex dependencies on target variable (cf. equation 1). Author assumed that for current selection of best feature there is important need to reduce relevancy term expressed as $MI(X_k,Y)$ by redundancy term over already selected features. It can be seen as a penalization of main effect when introducing new candidate does not improve overall information gain due to dependencies with earlier chosen features.

MRMR (Minimum-Redundancy Maximum-Relevance) [Peng et al. 2005] is presented as

$$J_{MRMR}(X_k) = MI(X_k, Y) - \frac{1}{|S|}\sum_{j \in S} MI\left(X_j, X_k\right). \quad (4)$$

Here, main modification related to redundancy term was proposed. When the selection process proceeds further it is more difficult to find relevant features, thus, setting scaling factor to reciprocal of |S| increases influence of main effect. Observe that in equation 3 we had also factor equal to 1.

IIFS (Interaction Information Feature Selection) [Pawluk et al. 2019], the major contribution in recent research of feature selection, states such task in following way

$$J_{IIFS}(X_k) = MI(X_k, Y) + \sum_{j \in S} II\left(X_j, X_k, Y\right) + \sum_{i,j \in S: i<j} II\left(X_i, X_j, X_k, Y\right). \quad (5)$$

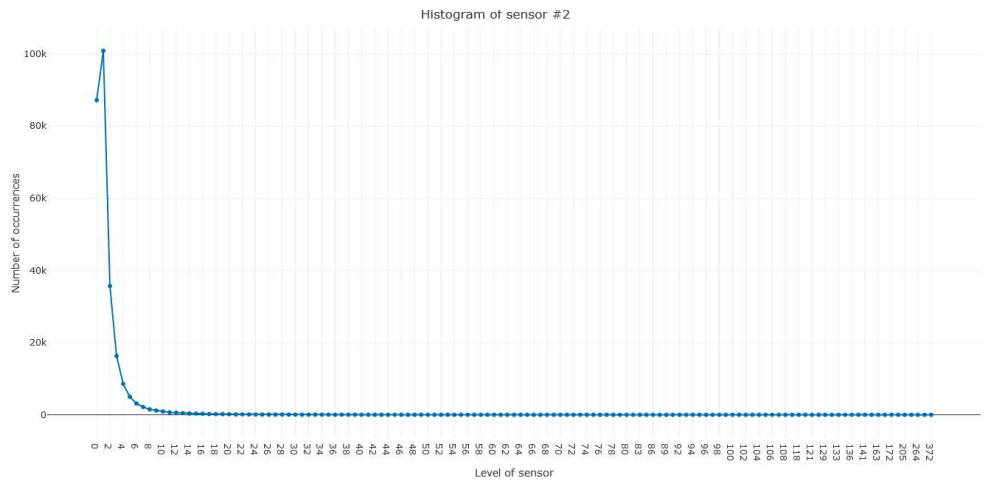
We introduced novel concept in term of feature selection – m-way interaction information of order m>3 [cf. Jakulin, Bratko 2004]. Remaining part of criterion is after basic transformations equivalent to equation 1.

Our goal was to explore higher order interaction term, which would take account of feature pairs in S, X_k and Y, whether it improves approximation of final score. Experiments showed that IIFS criterion obtained competitive results with traditional methods (CIFE, JMI, MIFS, MRMR) and can be successfully used in real-world scenarios. We believe that interaction term of high order might exist in complex dataset, such that IIFS criterion can address this case at the cost of increased amount of complexity. For the purpose of clarification, simple positive 3-dimensional interaction is XOR problem, when Y does not depend on X_1 and X_2, marginally, but jointly on the cartesian product of $X_3=(X_1 \times X_2)$. In this situation if we assume X_1 and X_2 are binary output variables and $Y=XOR(X_1, X_2)$ is binary output variable, then mutual information terms are as follows: $MI(X_1,Y)=0$, $MI(X_2,Y)=0$ and interaction information: $II(X_3, Y)=\log(2)>0$.

SELECTED DATA

For purpose of later experiment, data from the wine factory were acquired. During the process of alcohol production the sensors located in various points recorded concentration of components, sending independently information in a uniform format. We assumed that obtained input variables had the meaning of specified levels (categorical type) and denoted individual substances. Additionally, the target was set to binary variable (0/1 – bad/good wine quality according to sommelier's grade). We received approximately 150-200 thousands of observations with 14 unknown features registered. One remark related to all distributions of sensors was their common charactistics. In Figure 1 we depict histogram of selected sensor, which is skewed right.

Figure 1. Distribution of sensor #2



Source: own elaboration

EMPIRICAL STUDY

We conducted experiment in following configuration [cf. Brown et al. 2012]. At the beginning, inputs were discretized according to equal-frequency binning with 2 bins before process of feature selection. Evaluation of methods for dimensionality reduction was made using 20% hold-out set: data were divided into two subsets, each having 80% and 20% of observations, respectively. Afterwards, we employed criteria described in previous section using former sample. The size of such set was suitable for case of feature selection, because of common approach for information-theoretic techniques. These algorithms at the lowest level compute measure of entropy, which relies strongly on frequency counts if plugin estimator is chosen. In detection of most informative subset of sensors the number of inputs up to 13 was considered, i.e., we ran processing scenario with selection of 13 output features for each considered criterion. Subsequently, latter sample was utilized to assess performance of selected classification model for currently obtained set of output features having cardinality from 1 to 13. Firstly, we used simple kNN classifier with 3 neighbors, due to the fact that such method does not make any assumption on data without depending on particular criterion. Furthermore, this model is based on similarity function of euclidean distance. According to kNN classifier, we have increased chance of improvement for model evaluation when input dataset consists in only relevant features without redundant ones. Following this approach, value of used metrics can be competitive to other models metrics (including cases when models are complex units), showing that kNN is simple, yet fast and effective alternative to them (we recall ockham's razor as widely adopted way of thinking). Secondly, all subsets of features were applied to 10-fold cross-validation scheme in order to estimate classifier's error in more robust way. Finally, estimation of model's error was done based on metrics capable of dealing with imbalanced data, i.e., Balanced Error Rate (BER) [Tharwat 2018]

$$BER = 1 - 0.5 \cdot (specificity + sensitivity). \tag{6}$$

Here, specificity and sensitivity are true negative rate (proportion of all negatives that are correctly predicted as such) and true positive rate (similarly, proportion of all positives that are correctly predicted as such), respectively, and have forms of

$$specificity = \frac{TN}{N} = \frac{TN}{TN+FP}, \tag{7}$$

$$sensitivity = \frac{TP}{P} = \frac{TP}{TP+FN}. \tag{8}$$

Table 1 explains above measures in more detail and presents confusion matrix.

Table 1. Example of confusion matrix

	True positive condition	True negative condition
Predicted positive condition	TP = true positive	FP = false positive
Predicted negative condition	FN = false negative	TN = true negative

Source: own elaboration

RESULTS OF RESEARCH METHOD

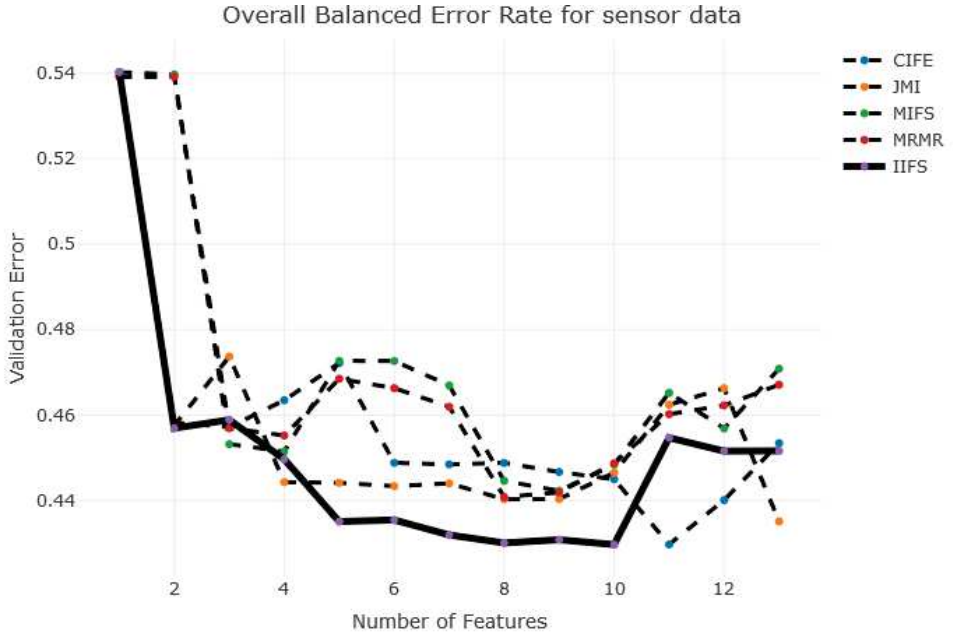
Final results of study are depicted in Table 2 and plotted in Figure 2, respectively.

Table 2. Values of errors for CIFE, JMI, MIFS, MRMR and IIFS

Criterion	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13
CIFE	0.540	0.458	0.457	0.463	0.472	0.449	0.448	0.449	0.447	0.445	0.430	0.440	0.453
JMI	0.539	0.458	0.474	0.444	0.444	0.443	0.444	0.440	0.440	0.447	0.462	0.466	0.435
MIFS	0.540	0.540	0.453	0.452	0.473	0.473	0.470	0.445	0.442	0.448	0.465	0.457	0.471
MRMR	0.539	0.539	0.457	0.455	0.469	0.466	0.462	0.441	0.442	0.449	0.460	0.462	0.467
IIFS	0.540	0.457	0.459	0.450	0.435	0.435	0.432	0.430	0.431	0.430	0.455	0.452	0.452

Source: own elaboration

Figure 2. Plots of errors for CIFE, JMI, MIFS, MRMR and IIFS



Source: own elaboration

In Table 2 we summarized errors' estimates for analyzed feature selection criteria and marked region of interest for IIFS, whose values are superior to other methods. Note that, Balanced Error Rate seems to be lowest not only in a range from 5 to 10, inclusively, but also beyond such range. When number of features is smaller than 5 all techniques work in similar way and IIFS does not fail at all. For case of numbers of inputs greater than 10, only CIFE presents better results, but the reduction of features is slight here and selection of these subsets is not reasonable.

In Figure 2 we depicted plot for analyzed feature selection criteria to show methods' behavior in visual way. It can be stated that the most encouraging trend is of IIFS ownership and all other methods cannot overcome its quality.
Summing up, there is a significant improvement of IIFS compared to remaining criteria for subsets having number of features from 5 to 10. This follows the basic assumption of strong need to address interactions existence of higher order in data, therefore, proposed criterion takes full advantage of own approach and includes them in decisive process. Consequently, overall performance is clearly better when using IIFS than other competitors. In such case, BER increases and traditional criteria cannot be used efficiently.

SUMMARY

Summing up, the analysis of real data showed undoubtedly that IIFS criterion works very well in case of complex data, which exhibits interdependent nature. However, tradeoff between complexity and accuracy needs to be examined, because calculation of high-order interactions involves considerable resources. On the other hand, if there are no specific requirements, it is recommended to follow IIFS approach when feature selection is of particular interest. This way, one can obtain better overall results in model development, allowing to be more successive in business scenarios.

ACKNOWLEDGMENTS

The application of presented method is in scope of Research and Development project aimed at developing an innovative tool for advanced data analysis called Hybrid system of intelligent diagnostics of predictive models. The project POIR.01.01.01-00-0322/18 is co-financed by the National Center for Research and Development in collaboration with scientists from Warsaw University of Life Sciences and The Jacob of Paradies University.

REFERENCES

Battiti R. (1994) Using Mutual Information for Selecting Features in Supervised Neural-Net Learning. IEEE Transactions on Neural Networks and Learning Systems., 5(4), 537-550.

Brown G., Pocock A., Zhao M. J., Luján M. (2012) Conditional Likelihood Maximisation: a Unifying Framework for Information Theoretic Feature Selection. Journal of Machine Learning Research, 13(1), 27-66.

Jakulin A., Bratko I. (2004) Quantifying and Visualizing Attribute Interactions: an Approach Based on Entropy. Manuscript.

Lin D., Tang X. (2006) Conditional Infomax Learning: an Integrated Framework for Feature Extraction and Fusion. LNCS Springer, 3951, 68-82.

Pawluk M., Teisseyre P., Mielniczuk J. (2019) Information-Theoretic Feature Selection Using High-Order Interactions. LNCS Springer, 11331.

Peng H., Long F., Ding C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226-1238.

Tharwat A. (2018) Classification Assessment Methods. Applied Computing and Informatics.

Xing E., Jordan M., Karp R. (2001) Feature Selection for High-Dimensional Genomic Microarray Data. ICML Proceedings of the Eighteenth International Conference on Machine Learning, 601-608.

Yang H. H., Moody J. (1999) Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. Advances in Neural Information Processing Systems, 12, 687-693.

Zhang F., Li W., Zhang Y., Feng Z. (2018) Data Driven Feature Selection for Machine Learning Algorithms in Computer Vision. IEEE Internet of Things Journal, 5(6).