© 2022 г. С.А. САЛТЫКОВ, канд. тех. наук
(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

The phenomenon of screening complementarity of features[1]

A constructed example is described, showing that the use of only There are no information-theoretic definitions and methods for calculating complementarity. allows you to catch complementarity escaping that occurs precisely when subsequent use of the greedy algorithm. A when eliminating the shielding element then the accuracy of the model becomes higher. It is proposed to construct an algorithm the prospects of selecting indicators that take this effect into account are justified. development of such an algorithm.

1. Introduction

In machine learning, it is quite common to find a situation where some two both indicators individually can very poorly predict the value of the target variable, and together, they can predict very well. This property of indicators is called complementarity. The idea that indicators can be complementary to each other, and complementarity as such is a new systemic quality that cannot be completely reduced to relevance and redundancy, was known for a long time [1, 2]. However, only a few years later decades after that, algorithms for selecting indicators began to appear, taking into account complementarity and thereby showing better results [3]. In recent years years of understanding that when selecting indicators, you should not focus on the dyad "relevance-redundancy as it was before [4], and on the triad "relevance — redundancy-complementarity"increased and was finally explicated [5,6,7]. Despite this, large reviews of the selection of indicators take into account the complementarity of the- This is reflected rather incompletely and briefly, sometimes even in earlier versions. reviews [8] are more detailed than in later ones [9]. More frequently encountered information-theoretic formalization of the phenomenon complementarity is not the only possibility. There are also formalizations using rank correlations [10, 11] and building two-tiered decision trees [12, 13].

2. The phenomenon of screening complementarity of indicators

Moreover, it can be shown that the formalization of the phenomenon of complementarity through in some cases, constructing a two-tiered tree is even preferable to a theoretical one.- informational formalizations. Namely, when, after selecting indicators, the model is based on the selected indicators are constructed using a "greedy"algorithm. In this case, "at the junction"the complementarity of indicators and the"greed "of the model training algorithm appear in a number of ways: new phenomena that are missing if the model is built using a full-processor algorithm. One such emerging phenomenon is the possibility of screening complementary effect of mutual reinforcement of indicators

[1]

by a certain third indicator, which we will call screening. It can be shown that if the accuracy of this third this indicator will be greater than the accuracy of each of the two complementary indicators by However, if they are both separate, but less than their joint accuracy, then the synergetic complementary method is the effect will be shielded, which means that the model will be built with less accuracy than if it had been this screening metric would not exist at all. In other words, removing the escape metric increases the accuracy. the trained model. Conversely, adding a certain metric to the data can lead to to reduce the accuracy, if the added indicator turns out to be a screening one. somewhat counterintuitive, because when using full-processor algorithms, this is not the case. occurs. For example, if we add a new metric to the data that is linearly regressed the model can either become more accurate, or remain the same accuracy, if the added the indicator will not be relevant or informative at all. But linear regression the model cannot become less accurate when adding a metric. And when using For "greedy"learning algorithms, this is possible, and this should be taken into account. Moreover, you can construct an example that shows that the use of only information-theoretic definitions and methods for calculating complementarity are not allowed "catch"complementarity escaping that occurs precisely during the subsequent search. using a greedy algorithm. And when eliminating the screening indicator the accuracy of the model becomes higher. This is demonstrated on a synthetic dataset. And this opens up opportunities for designing and thoroughly testing the selection algorithm. indicators that no longer take into account the above-mentioned triad, but the "relevance"quartet — redundancy — complementarity — complementarity screening".

3. Иллюстративный пример про недостатки теоретико-информационной дефиниции

Вот ЗДЕСЬ, собственно, сам иллюстративный пример. Его суть заключается в том, чтобы показать, что

$$JMI_{Score} = \sum_{i=1}^{N} R(x_i) - \sum_{i=1}^{N} \sum_{j=i+1}^{N} Iz(x_i, x_j) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} C(x_i, x_j, z)$$

$$R(x) = MI(x, z)$$

$$Iz(x, y) = MI(x, y)$$

$$C(x, y, z) = MI(xy, z) - MI(x, z) - MI(y, z) + MI(x, y)$$

$$JMI_{Score}^{Ekr} = \sum_{i=1}^{N} R(x_i) - \sum_{i=1}^{N} \sum_{j=i+1}^{N} Iz(x_i, x_j) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} C(x_i, x_j, z) - \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k=1;k\neq i,j}^{N} E(x_i, x_j, x_k, z)$$

$$E(x, y, e, z) = \begin{cases} MI(xy, z) - \max\limits_{\{n=1,..,N;n\neq i,j\}} (MI(xn, z), MI(yn, z)), & \text{если } e \text{ экр. комп. пару } x, y \\ 0, & \text{в противном случае} \end{cases}$$

Приведем иллюстративный пример, показывающий, что использование метрик, основанных на взаимной информации (как совместной, так и условной), не позволяет «ловить» эффект экранирования комплементарности показателей. Более того, это будет происходить вне зависимости от того, какой способ перебора наборов показателей будем выбран: даже при переборе всех наборов заданного размера такой эффект будет наблюдаться. Это происходит именно из-за того, что для каждого из наборов используется «жадный» метод построения дерева. И действительно, если бы для каждого из набора показателей использовался бы полный перебор всех возможных деревьев, то оценки, которые дают метрики, основанные на взаимной информации, были бы справедливы. Но из-за жадности алгоритмов построения деревьев такие метрики могут давать неадекватный результат.

№	X	Y	E	N	Z
1	0	0	0	0	0
2	1	0	0	0	1
3	1	0	0	0	1
4	0	0	1	0	0
5	0	0	1	0	0
6	0	1	0	0	1
7	0	1	0	0	1

В итоге если всё аккуратно посчитать, то получатся примерно следующие результаты:

Threesome	Tree Score	JMI
XYE	0.86	1.12
XYN	1.00	0.99
XEN	0.86	0.59
YEN	0.86	0.59

4. When it is recommended to use dimensionality reduction

Many seminal reviews [2011, 2017] and the textbook [2020] argue that the main reasons to use feature selection are (I) to reduce computational and storage costs (II) to create a cleaner, more interpretable, compact model (III) to improve performance metrics in general and learning accuracy in particular. But how can dimensionality reduction in general and feature selection in particular affect accuracy?

The reasons why irrelevant and/or redundant features can lead to a decrease in accuracy are almost always not given, but sometimes still the logical chain is explicated "Also, with a large number of features, learning models tend to overfit, which may cause performance degradation on unseen data."[2017] Thus, the only reason for possible accuracy degradation with sufficient time and computational resources when using superfluous features is indicated by overfitting. Therefore, it seems plausible that in the absence of overfitting under conditions of sufficient time and computational resources, features are not superfluous: in an ideal situation, adding features will not reduce accuracy, but it may not lead to improved accuracy.

The latter thesis is almost never stated explicitly, apparently because it is considered self-evident. Sometimes, however, this idea is expressed [2011] (in a slightly distorted form): "In theory, increasing the size of the feature vector is expected to provide more discriminating power.

Consequently, if we have a dataset where the amount of overfitting is negligible, and we have enough time and computational resources, then it means that dimensionality reduction can not be used. From a practical point of view for an engineer, this means that investing time in learning and implementing dimensionality reduction methods in this case is not worthwhile.

However, we will show that this is wrong. And that one can sometimes lose a very significant amount in accuracy by not using feature selection, even if we have a dataset containing orders of magnitude more instances than features (so overfitting will be insignificant), and we have any sufficient amount of time and computational resources.

5. Description of experiments and main results

The main results are.

First we find statistically significant screening triples for each dataset. For ARR we found 161 such triples, for BOS 11 triples, for DIA only one triplet. In total, there are 174 triples of screening. For each of the triples, we know which pair of complementary features is being screened by which feature; how accurate the model is on the three features and on the two features, and, accordingly, what the screening factor is.

How do we determine statistically significant triples? We randomly divide the dataset into training and test samples, and for this division we separately count the accuracy of two and three features. We accumulate them. Then we count the confidence intervals for each of the two accuracies. When they stop overlapping, it means that we broke through the statistical significance, and the triple is statistically significant.

How do we calculate the screening coefficient for a statistically significant triplet? Since we already have confidence intervals for each of the calculations - for two features and for three features - so there are two expected value. We simply divide one expected value by the other and get the screening coefficient.

№	set	pairs	type	min	median	max	acc-median	ro	p-value
1	ARR	161	C	1.09	1.23	2.52	0.2117	-0.76	$9.8 * 10^{-32}$
2	BOS	12	R	1.10	1.28	2.81	0.2108	-0.60	$3.9 * 10^{-2}$
3	DIA	1	C	24.77	24.77	24.77	0.0034	–	–
4	ALL	174	C/R	1.09	1.24	24.77	0.2117	-0.76	$2.4 * 10^{-34}$

6. Conclusion

In fact, this changes the entire pipeline in classical machine learning. It was thought that if you don't expect overfitting and have enough computational resources, you shouldn't even look at dimensionality reduction methods. And overfitting is not expected if your number of rows exceeds the number of columns by one or two orders of magnitude (or more), assuming, of course, that you choose the hyperparameters of the model correctly. In other words, it is considered that if you have enough data and computational resources, then spending time and attention on dimensionality reduction methods (feature selection and extraction) is simply overkill.

But it turns out that this is not the case. Non-use of dimensionality reduction methods, even in the case of sufficient data and computational resources, can in some cases lead to a very significant loss of accuracy compared to the use of dimensionality reduction methods. Moreover, these losses will be greater the worse the available features can predict the target variable. In other words, the less 'learned' the subject area is, the more relevant it is to pay attention to methods of dimensionality reduction even if sufficient data and computational resources are available.

СПИСОК ЛИТЕРАТУРЫ

1. *R. Kohavi and G. H. John* Wrappers for feature subset selection. Artificial Intelligence, 97(1-2):273–324, 1997.

2. *I. Guyon and A. Elisseeff.* An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182, 2003.

3. *Meyer P. E., Schretter C., Bontempi G.* Information-theoretic feature selection in microarray data using variable complementarity // IEEE Journal of Selected Topics in Signal Processing. – 2008. – T. 2. – No. 3. – C. 261-274.

4. *Yu L., Liu H.* Efficient feature selection via analysis of relevance and redundancy // The Journal of Machine Learning Research. – 2004. – T. 5. – C. 1205-1224.

5. *Shishkin, A., Bezzubtseva, A., Drutsa, A., Shishkov, I., Gladkikh, E., Gusev, G., Serdyukov, P.* Efficient high-order interaction-aware feature selection based on conditional mutual information // Proceedings of the 30th International Conference on Neural Information Processing Systems. – 2016. – C. 4644-4652.

6. *Singha S., Shenoy P. P.* An adaptive heuristic for feature selection based on complementarity //Machine Learning. – 2018. – T. 107. – No. 12. – C. 2027-2071.

7. *Li, C., Luo, X., Qi, Y., Gao, Z., Lin, X.* A new feature selection algorithm based on relevance, redundancy and complementarity // Computers in biology and medicine. – 2020. – T. 119. – C. 103667.

8. *Kotsiantis S.* Feature selection for machine learning classification problems: a recent overview // Artificial Intelligence Review. – 2011. – T. 42. – No. 1. – C. 157-176.

9. *Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., Liu, H.* Feature selection: A data perspective // ACM Computing Surveys (CSUR). – 2017. – T. 50. – No. 6. – C. 1-45.

10. *Салтыков С.А.* Algorithm of Building Regression Decision Tree Using Complementary Features / Proceedings of the 13th International Conference "Management of Large-Scale System Development"(MLSD). M.: IEEE, 2020. C. https://ieeexplore.ieee.org/document/9247785. DOI: 10.1109/MLSD49919.2020.9247785.

11. *Салтыков С.А.* Корреляция наукометрических показателей из РИНЦ с цитированием по базе Web of Science / Труды 13-й Международной конференции «Управление развитием крупномасштабных систем» (MLSD'2020, Москва). М.: ИПУ РАН, 2020. С. 1677-1684.

12. *Салтыков С.А.* Analysis of Decrease in Accuracy of Two-tier Trees without Using Feature Selection / Proceedings of the 14th International Conference "Management of Large-Scale System Development"(MLSD). Moscow: IEEE, 2021. C. https://ieeexplore.ieee.org/document/9600173. DOI: 10.1109/MLSD52249.2021.9600173.

13. *Салтыков С.А.* Алгоритм отбора показателей для построения двухъярусного дерева решений в задачах объясняемого искусственного интеллекта / Труды 14-й Международной конференции "Управление развитием крупномасштабных систем"(MLSD-2021). М.: ИПУ РАН, 2021. С. 1544-1551.

Фамилия И.О. первого автора, *место работы, должность, город*, адрес электронной почты