Сучасний стан наукових досліджень та технологій в промисловості. 2022. № 3 (21)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

UDC 004.415.25
DOI: https://doi.org/10.30837/ITSSI.2022.21.039

O. SOLOVEI

NEW ORGANIZATION PROCESS OF FEATURE SELECTION BY FILTER WITH CORRELATION-BASED FEATURES SELECTION METHOD

The **subject** of the article is feature selection techniques that are used on data preprocessing step before building machine learning models. In this paper the focus is put on a Filter technique when it uses Correlation-based Feature Selection (further CFS) with symmetrical uncertainty method (further CFS-SU) or CFS with Pearson Correlation (further CFS-PearCorr). **The goal** of the work is to increase the efficiency of feature selection by Filter with CFS by proposing a new organization process of feature selection. The **tasks** which are solved in the article: review and analysis of the existing organization process of feature selections by Filter with CFS; identify the routs cause the performance degradation; propose a new approach; evaluate the proposed approach. To implement the specified tasks, the following **methods** were used: information theory, process theory, algorithm theory, statistics theory, sampling techniques, data modeling theory, science experiments. **Results.** Based on the received results are proved: 1) the chosen features subset's evaluation function couldn't be based only on CFS merit as it causes a learning algorithm's results degradation; 2) the accuracies of the classification learning algorithms had improved and the values of determination coefficient of the regression leaning algorithms had increased when features are selected according to the proposed new organization process. **Conclusions.** A new organization process for feature selection which is proposed in current work combines filter and learning algorithm properties in evaluation strategy which helps to choose the optimal feature subset for predefined learning algorithm. The computation complexity of the proposed approach to feature selection doesn't depend on dataset's dimensions which makes it robust to different data varieties; it eliminates the time needed for feature subsets' search as subsets are selected randomly. The conducted experiments proved that the performance of the classification and regression learning algorithms with features selected according to the new flow had outperformed the performance of the same learning algorithms built with without applied new process on data preprocessing step.

Keywords: Correlation-based Feature Selection (CFS); symmetrical uncertainty (SU); Pearson Correlation (PearCorr); merit; accuracy; determination coefficient.

Introduction

When data which is gathered for pattern recognitions or machine learning models includes a lot of observations and features then it became difficult to perform effective data visualization; data mining or to build a machine learning model with high accuracy and performance. Therefore, sampling and feature selections methods are developed to cope with high-dimensional datasets [1].

Feature selection is a widely used instrument to remove irrelevant and redundant information from the dataset to avoid overfitting and reduce memory usage and computational costs. The goal of feature selection is to choose an optimal feature subset according to predefined evaluation criterion [2]. The recent trend to have a small number of samples in dataset versus a lot of features may cause problems to machine learning algorithm regarding learning performance therefore feature selection process plays increasingly import role while building machine learning model.

Analysis of the current state of the problem and methods of its solution

The techniques to select features for machine learning model are specified as: wrappers, embedded, filters, dimensionality reduction and hybrid [3]. A different Fast Correlation-Based Feature Selection (FCBFS) algorithm for filter had been considered in the study [4]. It proposes to use a threshold value δ, which is identified by user e.g. for dataset with N features and class C when $merit_{i,c}$ measures the correlation between a feature F_i and the class C then F_i is added in subset if $\vee F_i \in S'$, $1 < i \le N$, $merit_{i,c} > \delta$. Formed in such way subset is processed the 2nd time in order to retain only predominant feature. After one round of filtering features, algorithm takes the remaining features as a new subset and repeat starting to add a new feature. The algorithm stops when there is no feature to be removed. The worst case could be none features are removed. FCBF's performance for ten datasets had been compared with wrapper for two learning algorithms C4.5 and naive Bayes and because FCBF improved the accuracy of both learning algorithms FCBF was concluded as practical for feature selection for classification of high dimensional data.

The idea to use Hybrid approach according to which evaluation strategy uses a filter method and learning algorithm had been considered in several studies [5–7]. In work [5], it was proposed to form feature subsets using FBS with CFS and then to use dominance-based rough set approach (DBRSA) to select the final feature set. As DBRSA is an extension of the classical rough set approach (CRSA) which utilizes a decision tree, so the final feature subset is selected by the means of the learning algorithm.

Adaptive Hybrid Feature Selection methodology (AHFS) is proposed in work [6]. It utilizes the fact that there is no «best of» metric/method to select a feature subset and the choice of the metric/method can be realized by using the applied learning algorithm. AHFS uses SFS to form a subset in each iteration and iterates through possible evaluation methods to assign feature subset a set of ranks corresponding to the method. The final feature subset is selected by artificial neural networks model.

Hybrid feature selection that significantly reduces dimensionality of features was proposed in work [7]. The approach uses the combination of ReliefF and Principle Component Analysis (PCA) algorithms which are applied in the following sequence: 1) features are weighted by ReliefF and a candidate feature subsets are formed from features which weights exceed threshold; 2) PCA is applied on the candidate feature subset to reduce the dimension.

Highlight of the earlier unresolved parts of the general problem. Aim of the study

FCBF algorithm proposes to solve the first suggested in this paper problem by adding to evaluation strategy of feature selection a hyper parameter threshold δ, which value to be decided by user. It makes an algorithm dependent on the dataset and considering possible data variety the chosen value of δ may not be optimal.

Hybrid approaches [5–7] have a common idea that evaluation strategy couldn’t be based on CFS method only but requires additional evaluation criteria which is a learning algorithm. However, in studies [5–7] are left not considered the dependencies between a predefined learning algorithm and selected by evaluation strategy a final feature subset.

This paper proposes to change the organization process of feature selection by Filter with CFS that will take into consideration the dependencies between a predefined learning algorithm and chosen feature subset. The experiment tests of the new organization process will be performed for Filter with CFS-SU method – when dataset has features with discrete values and for Filter with CFS and Filter with CFS-PearCorr method – when dataset has features with discrete values.

Materials and methods

All Feature selection techniques have four steps in common:

1. Starting point – selects the feature from which to begin the feature selection.

2. Search organization – specify the algorithm for feature subset identification. Covers a type of search: exhaustive; complete; random and heuristic. A search can start by adding a new feature to an initially empty set and then a feature subset is expanded with one additional variable in each iteration step is called Sequential Feature Selection (SFS) or add all features and start removing irrelevant or redundant features (backward elimination) or Best First Selection (BFS) – the search is started with the most predictive feature according to chosen metric and then in each iteration step – the most predictive subset is expanded with a feature.

3. Evaluation strategy – specifies how a goodness of feature subset to be evaluated. It can be independent of the machine learning algorithm (common for filter technique) or by performance metrics of the learning algorithm (common for wrappers).

4. Stopping criterion – a rule to decide when to stop searching the feature subsets.

The current study is focused on filter and wrapper techniques, therefore the details for other techniques are omitted on purpose.

Wrapper method is aimed to select the feature subset that will ultimately provide a better estimate of accuracy. Wrapper uses a predefined machine learning algorithm to evaluate the quality of a selected feature subset. In forward selection, it calculates the accuracy of adding a new unselected feature to the subset and according to received accuracy decides to keep or remove the feature. Wrapper method produces good feature subsets because estimated performance of the learning algorithm is the best heuristic for measuring the goodness of feature subsets, however it computationally inefficient and it does not scale well to large datasets.

Filter – independent of any learning algorithm because its evaluation strategy is based on different statistical measures. When filter uses one of the methods: Low Variance [8], Fisher Score [9], T-score [10], distance measure [11], Chi-square [12], information gain [13], Gini index [14] and others [15] then the features are ranked and a selection strategy, e.g. "select best n_features" is applied to extract the final set of features to be used. Filter methods had been proved as fast and effective while capturing the relevance of features to the target, therefore filter methods are chosen when dataset has a big number of samples. However, filters with the mentioned methods cannot discover redundancy among features whereas redundant features along with irrelevant feature negatively affect the speed and accuracy of learning algorithm [16]. Another point, which makes filters technique weaker is that the features are ranked with no consideration for learning algorithm but different algorithms may perform better or worse for the same feature subsets [17]. In wok [18] was introduced a new Filter with CFS method it returns the set of features from which simultaneously are removed both irrelevant and redundant features. CFS finds feature subset that is useful to predict the target variable and do not strongly interact with other features. The rule is formalized as specified in equation 1

$$CFS\_score(S) = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}\ , \qquad (1)$$

where CFS score is an evaluation score (further «merit») of subset S with k features; $\overline{r_{cf}}$ – the average of the correlations between feature subset and target variable; $\overline{r_{ff}}$ – the average feature-feature correlation. In eq. 1 the numerator indicates the predictive power of the feature set while the denominator shows how much redundancy the feature set has. Correlations $\overline{r_{cf}}$ and $\overline{r_{ff}}$ can be calculated using symmetrical uncertainty formula when features in dataset have discrete values; or Pearson's correlation, for features with continuous values. When Filter uses CFS_PearCorr then $\overline{r_{cf}}$ from eq. 1 is the average of absolute values of Pearson correlations between feature subset and target variable; $\overline{r_{ff}}$ – the average of absolute values of feature-feature Pearson correlation. As a result of conducted experiments of features selection based on CFS score (eq. 1) in study [18] had been concluded that while solving a classification problem with discrete features dataset – CFS-SU can be used as a standard. The effectiveness of CFS-SU was evaluated by comparing it with wrapper which uses predefined learning algorithms: naive Bays, C4.5. The Filter with CFS-SU was organized as:

1. Starting point – from the 1st feature add/remove features one by one.

2. Search organization – features are added to subset with SFS technique until merit shows increasing value five time in sequence. The process is repeated 50 times. The resulted subsets are ordered by merit in descending order.

3. Evaluation strategy – consists of steps: 1) to merge 1st and 2nd best subsets; 2) to calculate the merit of the new composed subset; 3) if the new merit is within 10% of the merit of the best subset then accept the new best subset; 4) form a new subset by merging the best subset with next not used subset from the list with 50 subsets.

4. Stopping criterion – repeats steps 2–4 until the condition in step 3 isn't met.

The following results from the comparison had been shared: accuracy of naive Bayes with feature subset selected by CFS-SU had shown a degradation for one dataset: audiology (au) from 80.24% to 75.55%. Accuracy of C4.5 had shown a degradation for five datasets: mushroom (mu) from 99.59% to 99.37%, audiology (au) from 78.48% to 77.14%, soybean (sb) from 89.16 to 86.80, horse-colic(hc) from 84.02 to 78.79, king-rook vs. king-pawn (kr-vs-kp) from 99.16 to 94.13.

Potentially, there are several problems caused the accuracy's degradation:

1. SFS with CFS-SU merit as evaluation technique doesn't return the feature subset which is good for predefined learning algorithm.

2. Merging approach used for final feature subset works well for naive Bayes but caused the performance degradation for C4.5 because added redundancies to the final subset and algorithm C4.5 isn't efficient when correlated feature are included. This fact means that an evaluation of "goodness" of feature subset is impacted by the chosen learning algorithm but that impact wasn't taken into consideration.

Further, in this paper, those assumptions will be verified.

Study results and their discussion

The characteristics of datasets for which a performance degradation had been captured in study [18] are specified in table 1. The number of observations (n_observations) and number of features (n_features) are different compared to raw datasets from UCI Machine Learning Repository due to performed features engineering: 1) "one hot encoding" and "label encoding" – in order to adapt categorical values to CFS-SU method; 2) data clean up – observations with missing values were removed in case the majority of feature's values are missing.

Table 1. *The characteristics of datasets with discrete and categorical features*

Datasets	n_observations	n_features	Machine learning task	Feature Type			Missing values, Y/N?
				Continuous	Discrete	Categorical	
audiology	194	39	Multi classification			+	Y
horse-colic	299	12	Binary classification		+	+	Y
mushroom	8124	96	Binary classification			+	Y
king-rook vs. king-pawn	3196	35	Binary classification			+	N
soybean	306	35	Multi classification		+	+	Y

The learning algorithms: 1) naive Bayes was chosen in study [18] because, its classification accuracy is negatively affected by present of the redundant features as break the assumption that feature values are independent given the class; 2) C4.5 – is a popular learning algorithm for solving a classification task therefore an accuracy of C4.5 obtained as a result of experiments with feature selection approaches can be considered as a benchmark. In the current study we will continue to use those learning algorithms to compare the results and evaluate the correctness of the new proposals.

To study the result of feature evaluation strategy based on CFS-SU merit, we will visualize and analyze relationships between 1) accuracy of learning algorithm and CSF-SU merit; 2) number of features included in subset vs CSF-SU merit. To visualize the relationship line for each dataset from table 1 we do: 1) randomly select N feature subsets $\vec{S'}$; 2) calculate CSF-SU merit and accuracy of learning algorithms for each $S_i^{'}$, $1 \le i < N$.

The results are presented on fig. 1 – fig. 5.

Fig. 1 – on picture (a) the highest accuracy 83.67% corresponds to merit 0.145, calculated with 27 features (picture (c)) and the highest merit 0.181 corresponds to lower accuracy 79.59%. On picture (b) the highest accuracy 81.63% corresponds to merit 0.119, calculated for 27 features and the highest merit 0.18 corresponds to lower accuracy 73.47%. Therefore, for audiology dataset, the height merit of $S_i^{'}$ doesn't correspond to the best accuracy of both learning algorithms; the number of features can be reduced from 39 to 27 for naive Bayes and C4.5 algorithms.

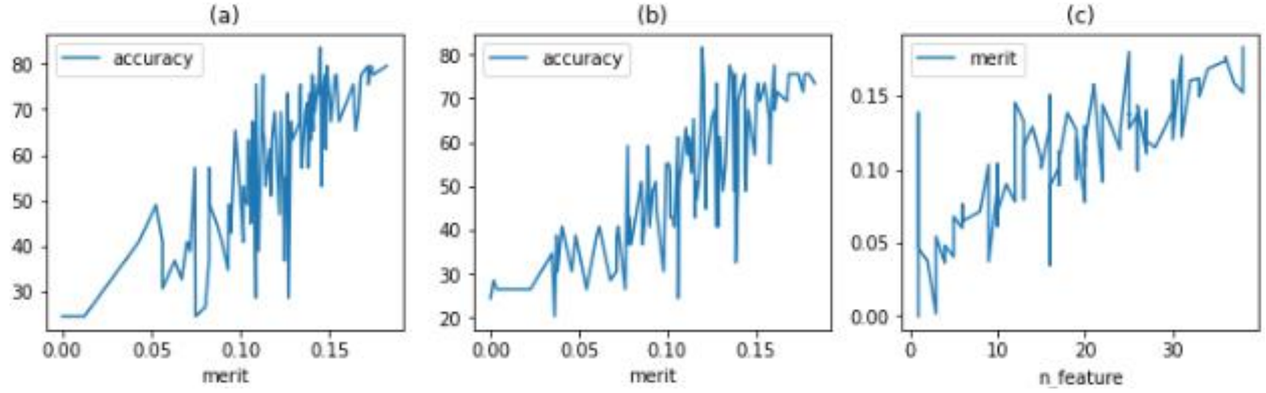

Fig. 1. (a) – accuracy of naive Bayes vs merit of au dataset with $S_i^{'}$ and target variable; **(b)** – accuracy of C4.5 vs merit of au dataset with $S_i^{'}$ and target variable; **(c)** – merit vs number of features in $S_i^{'}$ for au dataset

Fig. 2 – on picture (a) the highest accuracy 82% corresponds to merits 0.016 calculated with 6 features (picture (c)) and the highest merit 0.03 corresponds to lower accuracy 60%. On picture (b) the highest accuracy 85.28% corresponds to merit 0.023, calculated for 2 features (picture (c)) and the highest merit 0.027 corresponds to lower accuracy 76.58%. Therefore, for horse-colic dataset, the highest merit of $S_i^{'}$ doesn't correspond to the best accuracy of both learning algorithms and the number of features can be reduced from 12 to 6 for naïve Bayes and from 12 to 2 for C4.5.

Fig. 3 – on picture (a) the highest accuracy 99.7% corresponds to merit 0.027, calculated with 62 selected features (picture (c)) and the highest

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

merit 0.03 corresponds to lower accuracy 98.13%. On picture (b) the highest accuracy 100% the 1st time corresponds to merit 0.021, calculated for 52 features and that level of accuracy is kept with further growing merits. Therefore, for mushroom dataset, the height merit of $S_i^{'}$ doesn't correspond to the best accuracy of both learning the number of features can be reduced from 96 to 62 for naïve Bayes and from 96 to 52 for C4.5.

Fig. 4 – on picture (a) the highest accuracy 93.24% corresponds to merit 0.008, calculated with 25 selected features (picture (c)) and the highest merit 0.014 corresponds to lower accuracy 67.83%. On picture (b) the highest accuracy 99.62% corresponds to merit 0.009, calculated with 35 features and the highest merit 0.01 corresponds to lower accuracy 90.99%. Therefore, for kr-vs-kp dataset, the height merit of $S_i^{'}$ doesn't correspond to the best accuracy of both learning algorithms, the number of features can be reduced from 35 to 25 for naïve Bayes algorithm and no feature reduction is expected for C4.5.

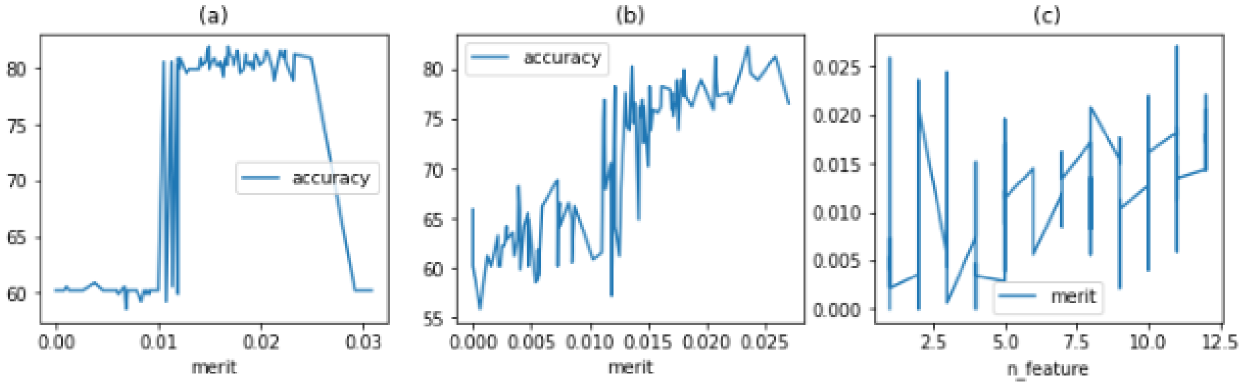

Fig. 2. (a) – accuracy of naive Bayes vs merit of hc dataset with $S_i^{'}$ and target variable; **(b)** – accuracy of C4.5 vs merit of hc dataset with $S_i^{'}$ and target variable; **(c)** – merit vs number of features in $S_i^{'}$ for hc dataset

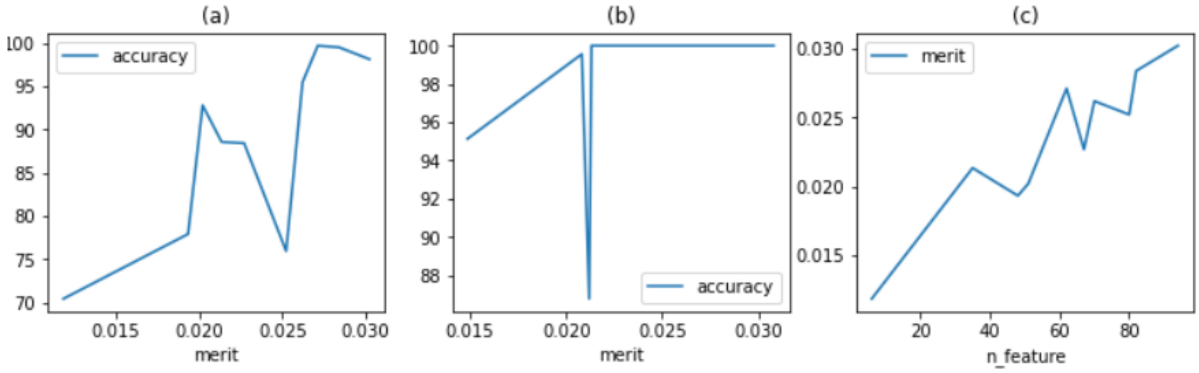

Fig. 3. (a) – accuracy of naive Bayes vs merit of mu dataset with $S_i^{'}$ and target variable; **(b)** – accuracy of C4.5 vs merit of mu dataset with $S_i^{'}$ and target variable; **(c)** – merit vs number of features in $S_i^{'}$ for mu dataset

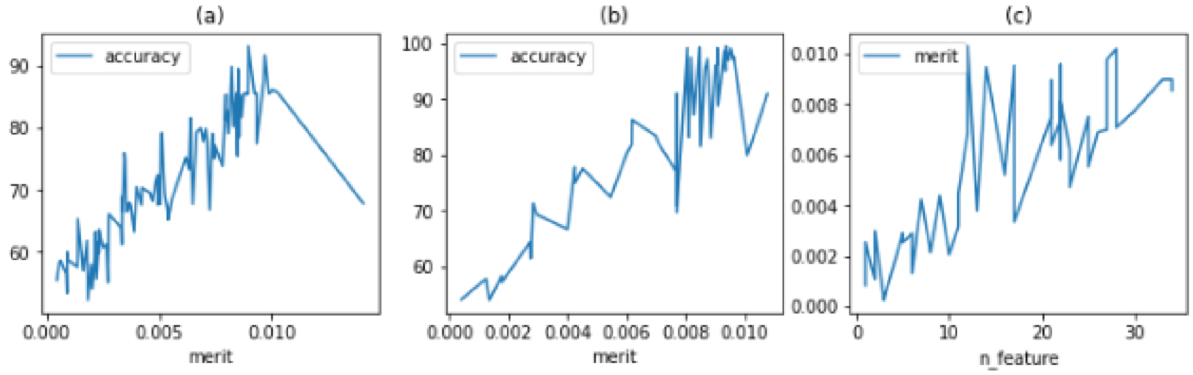

Fig. 4. (a) – accuracy of naive Bayes vs merit of kr-vs-kp dataset with $S_i^{'}$ and target variable; **(b)** – accuracy of C4.5 vs merit of kr-vs-kp dataset with $S_i^{'}$ and target variable; **(c)** – merit vs number of features in $S_i^{'}$ for kr-vs-kp dataset

Fig. 5 – on picture (a) the highest accuracy 93.51% corresponds to merit 0.255, calculated with 23 selected features (picture (c)) and the highest merit 0.3 corresponds to lower accuracy 88.31%. On picture (b) the highest accuracy 89.61% corresponds to merit 0.281, calculated with 27 features. Therefore, for soybean dataset, the height merit of $S_i^{'}$ doesn't correspond to the best accuracy of both learning algorithms, the number of features can be reduced from 35 to 23 for naïve Bays and from 35 to 27 for C4.5.

The summary from the conducted analysis (fig. 1 – fig. 5):

– a direct correlation between CFS-SU merit and accuracy of learning algorithm exists, however which value of merit corresponds to the beast accuracy couldn't be formalized. As a result: evaluation strategy can't be based only on CFS-SU merit value.

– both naive Bayes and C4.5 algorithms works better with different number of features, therefore feature subset selected with CFS-SU only without knowing a learning algorithm may also negatively impact the accuracy of the built model.

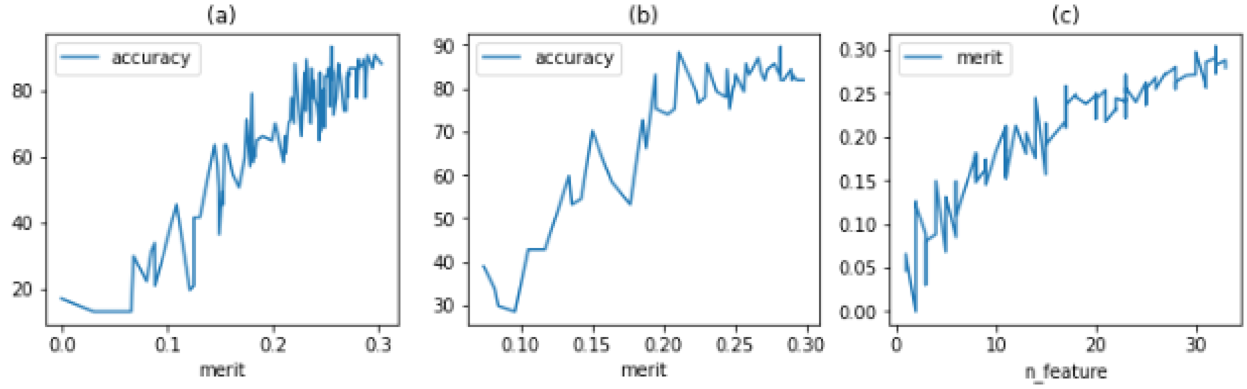

Fig. 5. (a) – accuracy of naive Bayes vs merit of soybean dataset with $S_i^{'}$ and target variable; **(b)** – accuracy of C4.5 vs merit of soybean dataset with $S_i^{'}$ and target variable; **(c)** – merit vs number of features in $S_i^{'}$ for soybean dataset

To tackle the identified problems, the following improvements in organization of feature selection process by Filter with CFS method are proposed:

1. Starting point: randomly select N feature subsets S'; calculate CFS merit (M_i) for each $S_i^{'}$, $1 \le i < N$; sort a vector with $S_i^{'}$ by M_i in ascending order.

Note: As a result of the sorting the higher accuracy will likely correspond to subset from the 2nd part of a vector $\overrightarrow{S'}$.

2. Search strategy: is not required.

3. Evaluation strategy: calculate accuracy $A_{S_i^{'}}$ of learning algorithm with subset $S_i^{'}$, where $i = N/2$ and calculate accuracy $A_{S_N^{'}}$ of learning algorithm with subset $S_N^{'}$.

If $A_{S_i^{'}}$ is less than $A_{S_N^{'}}$ then save N; increment i by $(N-i)\, div\, step$; ENDIF

IF $A_{S_i^{'}}$ is higher than $A_{S_N^{'}}$ then save i; decrement N by 1; ENDIF

IF $A_{S_i^{'}}$ is equal to $A_{S_N^{'}}$ then save i; increment i by 1; ENDIF

4. Stopping criterion: repeat Evaluation until (i is less than N) or $((N-i)\, div\, step > 0)$ The above steps 1–4 are formalized on fig. 6 and a calculation example to illustrate as the "best" feature subset can be chosen is illustrated in table 2 for horse-colic dataset with 10 feature subsets.

In table 2. CFS-SU merits are marked by bold to show how start and end indices had been moved: end index N is always decremented by constant 1 in order to not miss the feature subset which fits the best and likely correspond to the higher merit; start index i is incremented by $(N-i)\, div\, step$ to make a search more computationally efficient. Step value δ is a hyper parameter.

Time complexity of algorithm (fig. 2) depends on values of two parameters: N and δ and doesn't depend on dataset dimension. In worth scenario its time complexity is $O\left(\frac{N}{2\delta}\right)$. At the same time, the algorithm (fig. 2) eliminates "search strategy" from commonly used feature selection process which reduces over all time and simplifies the process.

To understand whether the new process can be used when feature is selected by Filter with CFS and other method from the specified in [15], we include in the experiments the datasets (table 3) which features have continuous values.

```
Input: $\overrightarrow{S'}$; dataset; $\delta$
Output: Index of feature subset which predicts the best accuracy
calculate $\vec{M}$
sort $\overrightarrow{S'}$ by $\vec{M}$ in ascending order
$i := int\left(len\left(\overrightarrow{S'}\right)/2\right)-1$
$N := len\left(\overrightarrow{S'}\right)-1$
Divide dataset on train/test.
WHILE True
    Fit learning algorithm with train data with features $\overrightarrow{S_i'}$;
    Calculate $A_{S_i'}$;
    Fit learning algorithm with train data with features $\overrightarrow{S_N'}$;
    Calculate $A_{N_i'}$;
    IF $A_{S_i'} < A_{N_i'}$ THEN $index := N$; $i := i + int\left(\left(N-i\right)/\delta\right)$ ENDIF
    IF $A_{S_i'} > A_{N_i'}$ THEN $index := i$; $N := N-1$ ENDIF
    IF $A_{S_i'} = A_{N_i'}$ THEN $index := i$; $i := i+1$ ENDIF
    IF $i \geq N$ or $(int\left(\frac{N-i}{\delta}\right)$ ==0) THEN Break ENDIF
ENDWHILE
```

Fig. 6. Algorithm of the new organization process for feature selection by Filter with CFS

Table 2. *Calculation example to illustrate algorithm fig. 6 for horse-colic dataset*

Merit	0.002	0.003	0.005	0.007	**0.0076**	0.013	0.0158	0.0162	0.0175	**0.0215**
Accuracy					54.67					72.00
Merit	0.002	0.003	0.005	0.007	0.0076	0.013	**0.0158**	0.0162	0.0175	**0.0215**
Accuracy							85.33			72.00
Merit	0.002	0.003	0.005	0.007	0.0076	0.013	**0.0158**	0.0162	**0.0175**	0.0215
Accuracy							85.33		70.67	
Merit	0.002	0.003	0.005	0.007	0.0076	0.013	**0.0158**	**0.0162**	0.0175	0.0215
Accuracy							85.33	72.00		

Table 3. *The characteristics of datasets with continuous and categorical features*

Datasets	**n_observations**	**n_features**	**Machine learning task**	**Feature Type**			**Missing values, Y/N?**
				Continuous	**Discrete**	**Categorical**	
automobile	205	26	Regression	+		+	Y
forecast order	60	12	Regression	+	+		N
rental building	372	108	Regression	+	+		N
boston house prices	506	13	Regression	+	+	+	N
Computer hardware	209	9	Regression	+		+	N

Experiments on continuous data follow a similar methodology as was applied for dataset from table 1. The only difference is the learning algorithms – three algorithms representing diverse approaches to learning are chosen for experiments with continuous values: a decision tree learner for regression (C4.5 for regression); Linear Regression (LR) – a linear predictor function is used to fit a prediction model; Locally Weighted Linear Regression (LWR) – non-linear learning algorithm for fitting a regression surface to data through multivariate smoothing. The quality of the built models is evaluated by determination coefficient (further R^2 score). The visualization of the relationships between R^2 score of C4.5, LR, LWR algorithms and CFS_PearCorr merit (fig. 7 – fig. 9) shows similar tendency as on fig. 1 – fig. 5, i.e. the highest value of merit doesn’t correspond to the highest value of R^2 score which means that an evaluation strategy can’t be based only on CFS_PearCorr merit and new proposed process is applicable to be applied to get better quality of machine learning model.

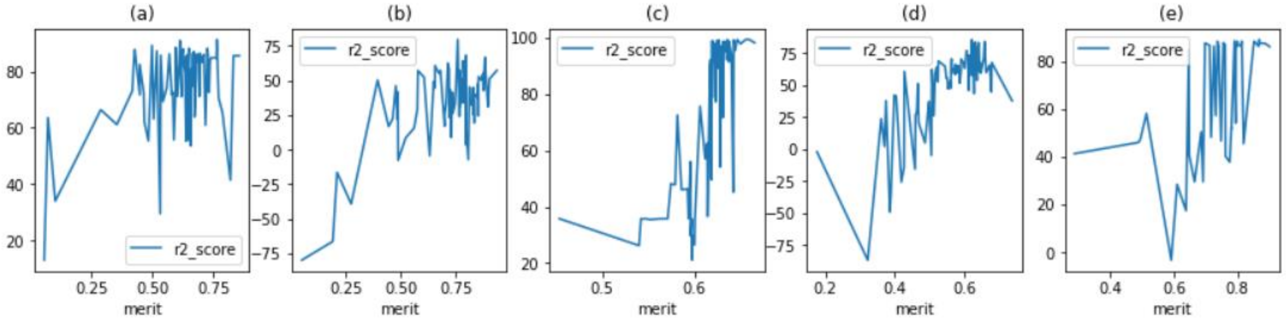

Fig. 7. R^2 score of C4.5 for regression vs CFS_PearCorr merit with **(a)** – automobile dataset; **(b)** – forecast order dataset; **(c)** – rental building dataset; **(d)** – boston house prices dataset; **(e)** – computer hardware datasets

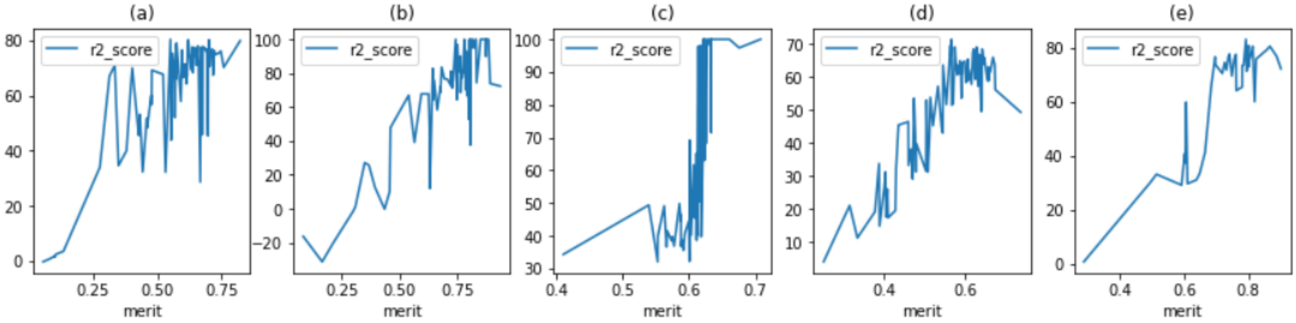

Fig. 8. R^2 score of Linear Regression vs CFS_PearCorr merit with **(a)** – automobile dataset; **(b)** – forecast order dataset; **(c)** – rental building dataset; **(d)** – boston house prices dataset; **(e)** – computer hardware datasets

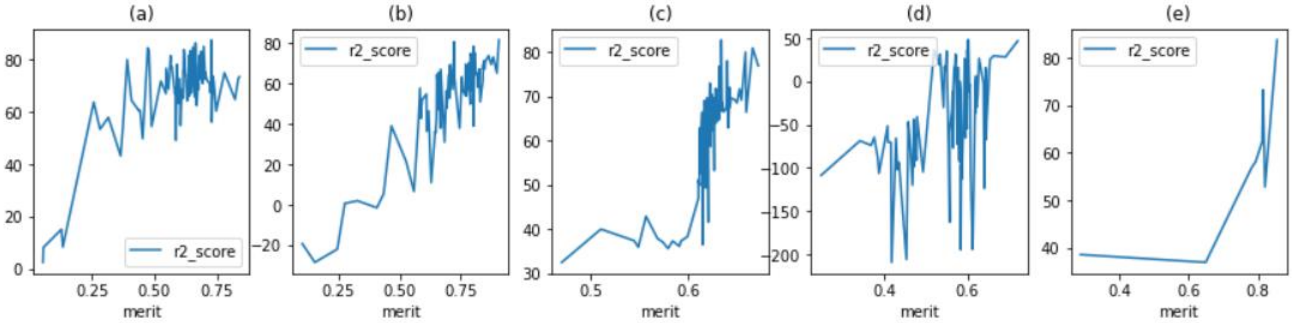

Fig. 9. R^2 score of Locally weighted linear regression vs CFS_PearCorr merit with **(a)** – automobile dataset; **(b)** – forecast order dataset; **(c)** – rental building dataset; **(d)** – boston house prices dataset; **(e)** – computer hardware datasets

In our experiments for datasets from table 1 the algorithm (fig. 2) had been ran with $\delta = 4$ and $N = 100$ and the received results are included in table 4. Columns: "all features, %" record the accuracy of naive Bays and C4.5 learning algorithms which were trained on dataset with all features. Columns: "Filter CFS_SU, %" record the accuracy of naive Bays and C4.5 learning algorithms which were trained on dataset with feature selected by approach from work [18]. Columns: "novel Filter CFS_SU, %" record the accuracy of naive Bays and C4.5 learning algorithms which were trained on dataset with feature selected by proposed process (fig. 2). Columns "Number of selected features" record the number of features selected by process (fig. 2) with regard to learning algorithm.

Accuracies in table 4 show that there is an improvement in the performance of naive Bays and C4.5 learning algorithms when features are selected according to the proposed process (fig. 2).

Different numbers of selected features for naive Bays and C4.5 learning algorithms proved that the same feature subset fits differently to learning algorithms and evaluation strategy of the feature selection process must take this into consideration.

The results of tests of the new flow (fig. 2) for datasets from table 3 is recorded in table 5. Columns: "all features, %" record R^2 score of C4.5, LR and LWR learning algorithms which were trained on dataset with all features. Columns: "CFS_PearCorr, %" record R^2 score of C4.5, LR and LWR algorithms which were trained on dataset with feature selected by proposed process (fig. 2). Columns "Selected features" record the number of features selected by process (fig. 2) with regard to the learning algorithm.

R^2 scores in table 5 show that there is an improvement in the performance of C4.5 and LR algorithms when features are selected according to the proposed process (fig. 2) for all dataset. R^2 scores of LWR algorithm had improved for three datasets included in the test, however, showed the degradation for "boston house pricing" and "computer hardware" datasets. The degradation happened because the values of hyper parameters of LWR algorithm are changing when feature subset is changed therefore if new flow to be used with LWR algorithm, it should have an additional step "selection of hyper parameters" before steps 7 and 9 on fig. 2.

Table 4. *Accuracy of learning algorithms for classification*

Datasets	naive Bays				C4.5			
	all features, %	Filter CFS_SU, %	novel Filter CFS_SU, %	Number of selected features	all features, %	Filter CFS_SU, %	novel Filter CFS_SU, %	Number of selected features
au	77.55	75.55	83.67	27	71.43	77.14	81.63	27
hc	80	88.76	82	6	76.6	78.79	85.28	4
mu	94.49	97.53	99.7	62	99.7	99.37	100	52
kr-vs-kp	85.48	90.20	93.24	25	99.62	94.13	99.62	35
sb	88.31	91.18	93.51	23	83.12	86.80	89.61	27

Table 5. R^2 *score of learning algorithms for regression*

Datasets	C4.5			LR			LWR		
	all features, %	CFS_ PearCorr, %	Selected features	all features, %	CFS_ PearCorr, %	Selected features	all features, %	CFS_ PearCorr, %	Selected features
automobile	84.35	90.55	6	77.41	81.55	8	78.52	87.55	9
Forecast order	38	71.66	7	100	100	6	55.5	81.51	2
Residential building	90.91	99.48	25	98.23	100	28	66.29	82.83	23
Boston house prices	76.17	86.12	10	68.4	70.95	11	91.9	50.5	9
computer hardware	87.89	88.5	5	80.45	83.06	5	92.22	85.53	2

P-value of paired t-test for accuracies from table 4 are recorded in tables 6 and 7 for naive Bayes and C4.5 learning algorithms correspondingly. P=0.0049 indicates statistically significant difference in accuracies of naive Bayes algorithm is obtained when features are selected by the proposed process vs accuracies of naive Bayes algorithm with all features.

P=0.018 (table 7) indicates statistically significant difference in accuracy of C4.5 algorithm is obtained when features are selected by the proposed process vs accuracies of C4.5 algorithm with features selected by Filter with CFS-SU [18].

P-value of paired t-test for R^2 scores from table 5 are recorded in tables 8. P=0.03 indicates statistically significant difference in R^2 score of Linear Regression algorithm.

Table 6. *Accuracies and p-value of paired t-test of naive Bayes*

Accuracy of Naive Bayes						**P-value of paired t-test**
all features, %	77.55	80	94.49	85.48	88.31	0.0049
novel Filter CFS_SU, %	83.67	82	99.7	93.24	93.51	
Filter CFS_SU, %	75.55	88.76	97.53	90.2	91.18	0.49
novel Filter CFS_SU, %	83.67	82	99.7	93.24	93.51	

Table 7. *Accuracies and p-value of paired t-test of naive Bayes of C4.5*

Accuracy of C4.5						**P-value of paired t-test**
all features, %	71.43	76.6	99.7	99.62	83.12	0.07
novel Filter CFS_SU, %	81.63	85.28	100	99.62	89.61	
Filter CFS_SU, %	77.14	78.79	99.37	94.13	86.8	0.018
novel Filter CFS_SU, %	81.63	85.28	100	99.62	89.61	

Table 8. R^2 *score and p-value of paired t-test of regression algorithms*

R^2 score *of C4.5 for regression,%*						**P-value of paired t-test**
all features	84.35	38	90.91	76.17	87.89	0.1
CFS_PearCorr	90.55	71.44	99.48	86.12	88.5	
R^2 score *of Linear Regression,%*						
all features	77.41	100	98.23	68.4	80.45	0.03
CFS_PearCorr	81.55	100	100	70.95	83.06	
R^2 score *of Locally weighted Linear Regression,%*						
all features	78.52	55.5	66.29	91.9	92.22	0.95
CFS_PearCorr	87.55	81.51	82.83	50.5	85.53	

Conclusion and perspectives of further development

This paper has presented new organization process for feature selection by Filter with CFS. The proposed process eliminates a time consuming "search strategy" step which is commonly included in feature selection procedure but is a time consuming and not always efficient. Time complexity of the new process (fig. 2) doesn’t depend on dataset’s dimension which makes it robust to different varieties of datasets which is often visible.

The conducted experiments with five datasets which features have discrete values and two predefined classification algorithms: naive Bayes and C4.5 have shown that by using a new process the performance results of learning algorithms are improved. P-value of paired t-test records statistically significant difference in the accuracies: for naive Bayes when features are selected by the proposed process compared with the accuracies of naive Bayes when all features are included in the model; for C4.5 when features are selected by the proposed process compared with the accuracies of C4.5 when features are selected by Filter with CFS-SU.

The conducted experiments with five datasets which features have continuous values and three predefined regression algorithms: C4.5 for regression, Linear Regression and Locally Weighted Linear Regression have shown that by using a new process the performance results of learning algorithms are improved. It is also noted that the new process should include additional step which is aimed to select values of hyper parameter when it is used with Locally Weighted Linear Regression algorithm.

Future work will be to extend a new approach by applying to different filter methods.

References

1. Guyon, I., Elisseeff, A. (2003), "An introduction to variable and feature selection", *J. Machine Learning Research*, No. 3, P. 1157–1182.
2. Dernoncourt, D., Hanczar, B., & Zucker, J.-D. (2014), "Analysis of feature selection stability on high dimension and small sample data", *Computational Statistics & Data Analysis*, No. 71, P. 681–693. DOI: https://doi.org/10.1016/j.csda.2013.07.012

3. Luan, C., Dong, G. (2018), "Experimental identification of hard data sets for classification and feature selection methods with insights on method selection", *Data Knowl. Eng*. No. 118, P. 41–51.
4. Senliol, B., Gulgezen, G., Yu, L., Cataltepe, Z. (2008), "Fast Correlation Based Filter (FCBF) with a different search strategy", *23rd international symposium on computer and information sciences*, P. 1–4.
5. Yu, L., Liu, H. (2021), "Enhancing Big Data Feature Selection Using a Hybrid Correlation-Based Feature Selection", No. 10, 2984 p. DOI: https://doi.org/10.3390/electronics10232984
6. Alzami, F., Tang, J., Yu, Z., Wu, S., Chen, P., You, J., Zhang, J. (2018), "Adaptive Hybrid Feature Selection-Based Classifier Ensemble for Epileptic Seizure Classification", *IEEE Access*., No. 6, P. 29132–29145. DOI: https://10.1109/ACCESS.2018.2838559
7. Jaina, D., Singhb, V. (2018), "An Efficient Hybrid Feature Selection model for Dimensionality Reduction", *Procedia Computer Science*, No. 132, P. 333–341.
8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. (2011), "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, No. 12, P. 2825–2830.
9. Duda, R., Hart, P., Stork D. (2012), *Pattern classification*, John Wiley & Sons.
10. Mundra, P., Rajapakseab, J. (2016), "Gene and sample selection using T-score with sample selection", *Journal of Biomedical Informatics*, No. 59, P. 31–41. DOI: https://doi.org/10.1016/j.jbi.2015.11.003
11. Tan, H., Wang, G., Wang, W., Zhanga, Z. (2022), "Feature selection based on distance correlation: a filter algorithm", *Journal of Applied Statistics*, No. 49 (2), P. 411–426.
12. Zhai, Y., Song, W., Liu, X., Liu, L. (2018), "A Chi-Square Statistics Based Feature Selection Method in Text Classification", *IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. DOI: https://10.1109/ICSESS.2018.8663882
13. Ircioa, J., Lojo, A., Morib, U., Lozanobc, J. (2020), "Mutual information based feature subset selection in multivariate time series classification", *Pattern Recognition*, No. 108. DOI: https://doi.org/10.1016/j.patcog.2020.107525
14. Sarkar, D., Goswami, S. (2013), "Empirical Study on Filter based Feature Selection Methods for Text Classification", *International Journal of Computer Applications*, No. 6, P. 38–43.
15. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R., Tang, J., Liu, H. (2018), "Feature Selection: A Data Perspective", *ACM Computing Surveys*, No. 50, P. 1–45. DOI: https://doi.org/10.1145/3136625
16. Koller, D., Sahami, M. (1996), "Toward optimal feature selection", *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, P. 284–292.
17. Ahn, E., Mullen, T., Yen, J. (2011), "A two-population evolutionary algorithm for feature extraction: Combining filter and wrapper", *IEEE Congress of Evolutionary Computation (CEC)*. DOI: https://ieeexplore.ieee.org/document/5949692
18. Hall, M. A. (1998), *Correlation-based Feature Selection for Machine Learning*, Ph.D diss. Dept. of Computer Science, Waikato Univ.

Received 14.09.2022

Відомості про авторів / Сведения об авторах / About the Authors

Соловей Ольга Леонідівна – кандидат технічних наук, Київський національний університет будівництва і архітектури, доцент кафедри інформаційних технологій проектування та прикладної математики, Київ, Україна; e-mail: solovey.ol@knuba.edu.ua; ORCID: https://orcid.org/0000-0001-8774-7243

Соловей Ольга Леонидовна – кандидат технических наук, Киевский национальный университет строительства и архитектуры, доцент кафедры информационных технологий и прикладной математики, Киев, Украина.

Solovei Olga – PhD (Technical Sciences), Associate Professor, Kyiv National University of Construction and Architecture, Kyiv, Ukraine.

НОВА ОРГАНІЗАЦІЯ ПРОЦЕСУ ВИБОРУ ОЗНАК ЗА ДОПОМОГОЮ ФІЛЬТРА НА ОСНОВІ КОРЕЛЯЦІЇ

Предметом статті є методи вибору ознак, що використовуються на етапі попереднього оброблення даних перед побудовою моделей машинного навчання. Увага надається методу фільтра в разі застосування вибору ознак на основі кореляції (далі CFS) та методу симетричної невизначеності (далі CFS-SU) або кореляції Пірсона (далі PearCorr). **Метою** статті є підвищення ефективності вибору ознак за допомогою фільтра з CFS шляхом нової організації процесу вибору ознак. **Завдання**, які вирішуються в роботі: огляд та аналіз наявної організації процесу виділення ознак фільтром із CFS; визначення причин, що спричиняють погіршення якості моделі; розроблення нового підходу; оцінювання запропонованого підходу. Для реалізації поставлених завдань використовувалися такі **методи**: теорія інформації, теорія процесів, теорія алгоритмів, теорія статистики, методи вибірки, теорія моделювання даних, наукові експерименти. **Результати.** На основі отриманих результатів доведено: 1) функція оцінки обраної підмножини ознак не може ґрунтуватися лише на CFS-оцінці, оскільки це спричиняє погіршення результатів алгоритму навчання; 2) точність алгоритмів навчання класифікації покращилася, а значення коефіцієнта детермінації алгоритмів регресії зросли, коли ознаки обиралися відповідно до запропонованого процесу. **Висновки**. Новий процес організації для вибору ознак, що пропонується в цій роботі, поєднує властивості фільтра та алгоритму навчання в стратегії оцінювання, що дає змогу обрати оптимальну підмножину ознак для попередньо визначеного алгоритму навчання. Обчислювальна складність запропонованого підходу не залежить від розмірів набору даних, що робить його стійким до будь-яких різновидів даних; запропонований процес також дає змогу заощадити час, необхідний для пошуку підмножин функцій, оскільки підмножини обираються у випадковий спосіб. Проведені експерименти довели, що продуктивність алгоритмів класифікації та регресії покращилась, порівняно з продуктивністю тих самих алгоритмів навчання, але без застосування запропонованого процесу на етапі попереднього оброблення даних.

Ключові слова: вибір ознак на основі кореляції (CFS); симетрична невизначеність (SU); кореляція Пірсона (PearCorr); критерій якості; точність; коефіцієнт детермінації.

НОВАЯ ОРГАНИЗАЦИЯ ПРОЦЕССА ВЫБОРА ПРИЗНАКОВ С ПОМОЩЬЮ ФИЛЬТРА НА ОСНОВЕ КОРРЕЛЯЦИИ

Предметом статьи являются методы выбора признаков, которые используются на этапе предварительной обработки данных перед построением моделей машинного обучения. Внимание уделяется методу фильтра при использовании выбора признаков на основе корреляции (далее CFS) и методу симметричной неопределенности (далее CFS-SU) или корреляции Пирсона. **Целью** статьи является повышение эффективности выбора признаков с помощью фильтра CFS-SU путем новой организации процесса выбора признаков. **Задачи**, решаемые в статье: обзор и анализ существующей организации процесса выделения признаков фильтром с CFS; определение причин, вызывающих ухудшение качества модели; разработка нового подхода; оценка предложенного подхода. Для реализации поставленных задач использовались следующие **методы**: теория информации, теория процессов, теория алгоритмов, теория статистики, методы выборки, теория моделирования данных, научные эксперименты. **Результаты**. На основе полученных результатов доказано: 1) функция оценки выбранного подмножества признаков не может базироваться только на CFS-оценке, поскольку это приводит к ухудшению результатов алгоритма обучения; 2) точность алгоритмов обучения классификации улучшилась, а значение коэффициента детерминации алгоритмов регрессии выросли, когда признаки выбирались в соответствии с предложенным процессом. **Выводы**. Новый процесс организации для выбора признаков, который предлагается в данной работе, сочетает свойства фильтра и алгоритма обучения в стратегии оценки, что помогает выбрать оптимальное подмножество признаков для предварительно определенного алгоритма обучения. Вычислительная сложность предлагаемого подхода не зависит от размеров набора данных, что делает его устойчивым к разным разновидностям данных; также предложенный процесс позволяет экономить время, необходимое для поиска подмножества функций, поскольку подмножества выбираются случайным образом. Проведенные эксперименты доказали, что производительность алгоритмов классификации и регрессии улучшилась по сравнению с производительностью тех же алгоритмов обучения, но без применения предложенного процесса на этапе предварительной обработки данных.

Ключевые слова: выбор признаков на основе корреляции (CFS); симметричная неопределенность (SU); корреляция Пирсона (PearCorr); критерий качества; точность; коэффициент детерминации.

Бібліографічні описи / Bibliographic descriptions

Соловей О. Л. Нова організація процесу вибору ознак за допомогою фільтра на основі кореляції. *Сучасний стан наукових досліджень та технологій в промисловості*. 2022. № 3 (21). С. 39–50. DOI: https://doi.org/10.30837/ITSSI.2022.21.039

Solovei, O. (2022), "New organization process of feature selection by filter with correlation-based features selection method", *Innovative Technologies and Scientific Solutions for Industries*, No. 3 (21), P. 39–50. DOI: https://doi.org/10.30837/ITSSI.2022.21.039