Pattern Recognition 116 (2021) 107932

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Adaptive, Hybrid Feature Selection (AHFS)

Zsolt János Viharos[a,b,*,*], Krisztián Balázs Kis[a], Ádám Fodor[a,c], Máté István Büki[a]

[a] Institute for Computer Science and Control (SZTAKI), Centre of Excellence in Production Informatics and Control, Eötvös Loránd Research Network (ELKH), Research Laboratory on Engineering and Management Intelligence, Intelligent Processes Research Group, H-1111, Budapest, Hungary, Kende u. 13–17., Hungary

[b] John von Neumann University, Faculty of Economics, Department of International Economics, Kecskemét, H-1117, Izsáki u. 10., Hungary

[c] Eötvös Loránd University, Department of Software Technology and Methodology, Budapest, H-1117, Pázmány P. sny 1/C., Hungary

ARTICLE INFO

Article history:
Received 11 April 2020
Revised 18 September 2020
Accepted 3 March 2021
Available online 11 March 2021

MSC:
00-01
99-00

Keywords:
Adaptive
Hybrid Feature Selection (AHFS)
Combination of methods
Statistics
Information theory
Exhausting evaluation

ABSTRACT

This paper deals with the problem of integrating the most suitable feature selection methods for a given problem in order to achieve the best feature order. A new, adaptive and hybrid feature selection approach is proposed, which combines and utilizes multiple individual methods in order to achieve a more generalized solution. Various state-of-the-art feature selection methods are presented in detail with examples of their applications and an exhaustive evaluation is conducted to measure and compare the their performance with the proposed approach. Results prove that while the individual feature selection methods may perform with high variety on the test cases, the combined algorithm steadily provides noticeably better solution.

© 2021 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Most real world modeling problems can be formulated as estimation of some numerical value or classifying a given number of samples. These problems, more often than not, are very complex and can be defined by tens, hundreds and even thousands of variables. The high dimensional data is hard to handle by soft computing methods, but in most of the cases many variables are highly redundant, noisy and/or irrelevant when solving a specific estimation or classification task. The need arises to reduce the dimension of a problem by selecting only the relevant features for a given assignment with relatively fast methods compared to the highly accurate but costly estimation models. Feature Selection (FS) methods are doing exactly that. Some of them are able to perform calculations faster than others but with the price of losing accuracy, while others work the other way around. Generally, there is no universal solution, some methods are more suitable for a given assignment than others.

We propose a new, adaptive and hybrid feature selection approach, which combines and utilizes multiple individual methods in order to achieve a more generalized solution. It minimizes the shortcomings of each incorporated algorithms by choosing dynamically the most suitable one for a given assignment and dataset.

The paper contains seven sections. After the introduction the second section presents different feature selection methods and applications from varying branches. The third section describes the proposed combined algorithm, which is followed by the evaluation part considering comprehensive test on artificial datasets in relation to data distributions, noise level and outliers. In the next paragraph the proposed method is evaluated through many benchmarking datasets according to modelling error and calculation time demends, and in the next section it is compared by the very recent, state-of-the-art feature selection methods. Finally, the conclusion, acknowledgment and literature sections close the paper.

2. Methods and applications of feature selection

Feature selection methods reduce the dimension of a problem by selecting or creating a representative subset of features for a given assignment, making it easier for more (computational cost) demanding algorithms to manage. They can be categorized as filter, wrapper and embedded methods [1]. Miao and Niu gave a structure as a simple tree for positioning various feture selection tech-

* Corresponding author.
E-mail address: viharos.zsolt@sztaki.hu (Z.J. Viharos).

https://doi.org/10.1016/j.patcog.2021.107932
0031-3203/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

niques. Key decision points are the label information as supervised, semi-supervised or unsupervised techniques, while filter, wrapper and embedded methods relate to the type of the search strategy. Filter method relies on general characteristics of the training data to select the most relevant subset of variables without involving any learning algorithm. Wrapper methods use learning algorithm to detect possible interactions between variables, then select the best subset of features. Finally, embedded methods try to combine the advantages of both previous methods. In embedded methods the learning and the feature selection part cannot be separated, furthermore feature selection and evaluation proceed simultaneously. Srivastava et al. published a Review Paper on Feature Selection Methodologies and their Applications in which they showed a comparison table about filter, wrapper and embedded methods [2]. It is described that wrapper methods are usually superior to the other two techniques, however, they have the highest computational requirement. On the other hand the learning method independence of filter methods serve with a more general solution in this aspect. Muñoz-Romero et al. proposed a particularly sophisticated method called Informative Variable Identifier (IVI) with the aim to add interpretability for feature selection. They also compared the performance of their algorithm using various state-of-the-art measures of dependencies among variables [3]. A novel, improved unsupervised feature selection algorithm is presented by Shang et al. using matrix decomposition method as core technique. The kernelization of the local discriminant model was introduced for the necessary handling of non-linearity together with the proposal of a new measurement norm [4]. Also improvements in unsupervised feature selection was proposed by Zhang et al. using guided subspace learning [5].

Feature Selection methods are very useful in many fields which work with high-dimensional data. In applications of computer vision and image processing, the features describe artifacts of the digital image. Zini et al. addressed the problem of structured feature selection in a multi-class classification setting by proposing a new formulation of the Group LASSO method [6]. Their method outperformed the state-of-the-art approaches when tested on two benchmark datasets. Jiang and Li used the Minimal Redundancy Maximal Relevance (mRMR) method for classification of cotton foreign matter using hyper-spectral imaging [7]. They showed the generality of the method by building different learning models on the selected features and achieving similar estimation accuracy.

Feature selection methods are also applied for monitoring and fault diagnosis, where several sensor measurements and other variables describe the actual state of the system. Zhang et al. used feature extraction and selection for multi-sensor-based real-time quality monitoring in arc wielding [8]. In another example the authors used feature selection for high-dimensional machinery fault diagnosis [9]. After selecting the relevant features with a hybrid solution combining multiple methods, they used Radial Basis Function networks for classification and evaluated their approach on two data cases showing that the method is useful for revealing fault-related frequency features.

Time series forecast is used for many things like weather, energy consumption, financial plans, etc. Many variables have to be taken into consideration for an accurate forecast, due to the high complexity of the task. Naturally, feature selection methods are very useful on this field, too. Carta et al. compared feature selection methods using ANNs in Measure-Correlate-Predict (MCP) wind speed methods [10]. Their results showed that the Multi-Layer Perceptron-based wrapper method performed better in every test case, while the filter approach, the Correlation Feature Selection method, proved to be more efficient in terms of computational load and resulted in more model interpretability. Kong et al. did wind speed prediction using reduced support vector machines with feature selection [11]. They successfully selected a smaller, relevant subset of the features, which they used for training a reduced support vector machine and proved its effectiveness through detailed analysis and simulations. Finally Ircio et. al. used an adaptation of existing nonparametric mutual information estimators based on the k-nearest neighbor for selecting a subset of multiple time series [12]. In their experiments they managed to strongly reduce the number of time series while keeping or increasing the classification accuracy.

3. The basics of the novel adaptive, hybrid feature selection (AHFS) method

There are two broad approaches to measure the dependency between two random variables. First of all the correlation-based measures were examined to determine the adequacy of features. The linear correlation coefficient is one of the most known measure [13]. Other measures in this category are variations of this approach. There are several benefits of this measure, it helps to remove near zero correlated features to the target class, in addition it can help to reduce redundancy between selected features.

The other common approach for determining dependencies between variables is the use of information-theory based concepts. In general, a feature is good, if it is relevant to the target class, but it is not redundant to any of the other selected relevant features. Prominent relevancy and reduced redundancy can be achieved with the information theoretic ranking criteria, like *entropy* to measure of uncertainty of a random variable. The *conditional entopy* is the amount of uncertainty left in X when a variable Y is introduced, so it is less then or equal to the entropy of both variables. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called *information gain*, which is also used as a synonym of *mutual information*, which is the expected value of information gain.

Various measures inherited from information theory, like Shannon entropy, joint and conditional entropy, mutual information and symmetrical uncertainty are frequently applied in recent feature selection solutions and applications like FCBF [14] and mRMR [15]. The information-theory based measures can observe higher level correlations as well and they are becoming ever more popular in recent decades [16]. The following algorithms are the most popular for selecting the appropriate set of features.

- Forward feature selection (FFSA) [17]
- Modified mutual information-based feature selection (MMIFS) [18,19]
- Linear correlation-based feature selection (LCFS) [13]
- Fast Correlation Based Filter algorithms (FCBF and FCBF#) [20,21]
- Minimal Redundancy Maximal Relevance (mRMR and mRMR#) [15,22]
- Joint Mutual Information Maximisation and Normalized Joint Mutual Information Maximisation (JMIM and NJMIM) [23,24]
- Euclidian distance-based selection [25]

3.1. Motivation and concept

The previous paragraphs highlight various methods and real life applications of Feature Selection (FS) algorithms proving their importance. Various FS methods exist today, and the review of their basic concepts, theories, applications and benchmarking comparisons mirrors that:

- *In general, no "best of" or "best practice" feature selection solution is given.* The applications and scientific publications mirror that the "best solutions" may largely differ from case to case. FS is applied frequently, but the most promising solutions are dataset and/or application field specific. Wang et al. already formu-

lated that there exists a relationship between the performance of a feature selection algorithm and the characteristics of data sets [26].

- According to the definition of feature selection *this Data Mining (DM) tool is applied before a training algorithm.* On rare occasions, having the results of feature selection (so, having the list of selected features) a theoretical model is built (e.g. using equations), furthermore, feature selection is seldom combined/integrated with learning [27]. Consequently, in general it is a preliminary step of a training algorithm that is based on the same data set.
- Feature selection *is applied for two main reasons*:
  - Reducing the number of features *significantly decreases the computational requirements of the following modelling solution and also the sensor requirements of the given applications.*
  - The elimination of irrelevant information (irrelevant features) *results more accurate models (decreases the noise).*

  It has to be mentioned that in another aspect e.g. in technical applications, like failure detection and forecast, as components of diagnostics and supervision, it is valuable having, in some degree redundant information e.g. to detect non-conform situations [28] or to overcome the failures of sensors or any other data processing components. It is advantageous also when incomplete data arise [29].
- Roughly speaking *feature selection algorithms consist of three main calculation components:*
  - The first part is *the calculation of one (or rarely more) metrics* using the given data set for *having some numerical evaluation of the individual variables or variable sets.* Typically, two aspects are evaluated: redundancies among variables and the correlations between the individual variables and the (later on) estimated (target) variable.
  - The second part is a *search algorithm* that applies the above measure/metrics *to determine the order and/or selected set of features.* There are many such solutions, one could differentiate among them whether they are greedy (*like Sequential Feature Selection, SFS*) or somehow optimized algorithms.
  - The third part is the applied *modeling methodology*. In filter methods it is fully separated from the feature selection component, however in wrapper and embedded methods it is integrated on different levels.

  The scientific literature represents great variety of combinations of the applied measures and search algorithms.

Not all of the available/possible feature selection metrics (measures) were introduced above, only the most frequently applied (because of their superiority over other methods), meaning that a new solution is valuable if it could exploit the advantages of any of the given or later introduced FS techniques. Based on this idea, a ***hybrid*** solution is proposed in the paper which **combines the given, available (supervised) feature selection techniques** that have their own specific, but fixed feature evaluation measures/metrics. The proposed methodology can be extended easily with any of novel FS methods and metrics, so, any alternative feature selection technique can be a part of the proposed solution, too. This is one aspect of the hybridity. Since there is no general, "best" FS algorithm, which indicates that no "best" feature measure/metrics is given, moreover, because the main aim of FS is to support the building of a learning model after it's usage, **it utilizes the applied learning model in its mathematical algorithm.** (The author's implementation uses the MultiLayer Perceptron (MLP) modell, however, aside from some special techniques, any other learning models can be applied here.) This is the second aspect of the hybridity. Since many combinations of applied measures and search solutions are published in the state-of-the-art literature, the proposed algorithm applies the simple but frequently used Sequential Forward Selection (SFS) technique. It is a feed-forward calculation method that extends the already selected set of features with only one additional variable in its iteration steps. In this aspect it is a greedy algorithm, naturally, later on the proposed solution can be improved substituting it with a more suitable and generalized search solution but this additional research direction is beyond the scope of the current paper. According to the above state-of-the-art statements, there is no general, unique feature selection methodology, it has to be always adapted to the given application and dataset, consequently, a novel solution is needed that is *adaptive*. The main aim is to ensure with the adaptivity of the proposed algorithm at each iteration step of the applied Sequential Forward Search iteration. At a certain SFS step a set of already selected variables (features) are given and the method evaluates each possible extension of this dataset by one additional variable. In the state-of-the-art solutions it uses (only) one feature selection measure for selecting an additional variable for the final extension, so, the *state-of-the-art algorithms iterate in the space of the variables.* ***Adaptivity*** of the proposed algorithm is realized in such a way that **at an individual step of the feature selection algorithm it iterates not only in the space of the variables but in the space of available features selection techniques, too. This is the core idea of the paper.** Since there is no "best of" feature selection measure/metrics, choosing one from these candidates can only be realized by using the applied learning method as an independent evaluation tool. It means, each such candidate model configuration is built up and the model having the smallest estimation error specifies which one variable has to be selected as the current extension. This motivations and concepts led to the novel **Adaptive, Hybrid Feature Selection (AHFS)** algorithm, introduced, described and evaluated in the paper.

3.2. Search strategy: Sequential forward selection

Filter, wrapper and embedded methods apply a search strategy for the selection of the feature order. Guyon and Elisseeff presented various feature selection possibilities for finding a subset of variables for building up a good predictor [1]. Such methods are for variable ranking, elimination of redundant variables, variable subset selection, Nested Subset Methods, forward selection and backward elimination. Consequently, there are various possible solutions for the applied search methodology. Sequential Forward Selection (SFS) was selected, however, almost all the other possible techniques can be applied inside the proposed AHFS methodology. The SFS algorithm is a bottom-up search procedure which start from an empty set and gradually adds features to the current feature set. The decision of selection depends on a predetermined evaluation function. At each iteration, the feature to be included in the feature set S is selected among the remaining available features in feature set F, so this extended set S should produce maximum value of the criterion function used [25].

The simplicity and speed of the SFS make a compromise between high-dimensional search spaces, slow evaluation and the execution time demanded of the algorithm.

3.3. Modeling method: Multi-Layer perceptron (MLP)

Artificial Neural Networks (ANNs) are powerful computational models, which can be utilized for solving complex estimation and classification problems due to their robustness and capability of high level generalization. An ANN implements the functionality of the biological neural networks by building up a network of autonomous computational units (neurons) and connecting them via weighted links defined by the first pioneers W. S. McCulloch and W. Pitts [30]. One of the most popular and widespread ANN model

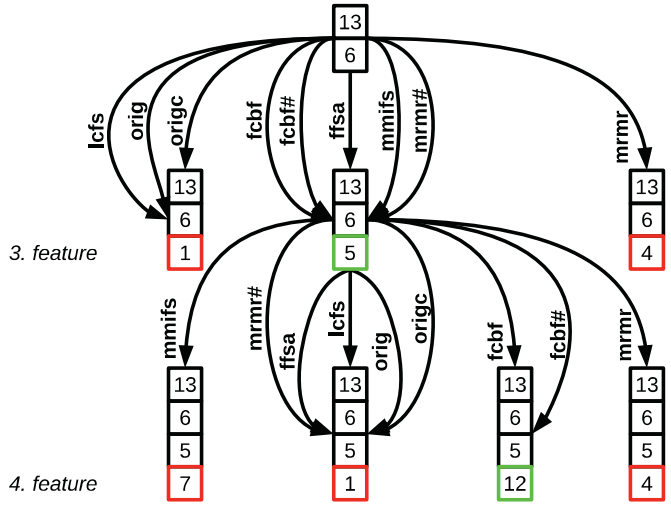

Fig. 1. Operation graph of the proposed algorithm depicting two steps.

type is the MLP [31], another concept which became popular and used widely nowadays is deep learning [32].

The MLP model is used in the algorithm as part of the evaluation, meaning that in every step, where a feature is evaluated, an MLP model is trained to determine how much error the feature yields - compared to others.

3.4. The proposed adaptive, hybrid feature selection (AHFS) algorithm

In the previous chapters the described feature selection methods are described as efficient algorithms which use interdependence of features together with the dependence to the given class. However, their performance is perceptibly diverse through different datasets and assignments. The proposed algorithm combines the different object functions/measures by giving every algorithm the chance to suggest the next possible best feature in the sequential forward selection algorithm. However, this concept is independent of the currently applied search method.

In order to find the best feature subset/order the algorithm uses forward selection technique with sequential search strategy. The termination criterion is the reaching of a k-size subset, prescribed before run, e.g. by a human expert. The proposed algorithm uses predetermined feature selection methods. There are two important phases:

- First one is the feature selection part, which *iterates through the predetermined feature selection algorithms*, invoking them with the already selected feature set S. Every algorithm propose one feature as the next potential best feature to be selected according to its own measure. The candidate features are collected to S_p feature set in every iteration.
- Having candidates for an additional feature inherited from the different feature selection methods, the aim of the second phase is to select one from them, so, in the next phase *it iterates through the promising additional feature space*. This selection is based on the (highest) accuracy of the trained artificial neural network models.

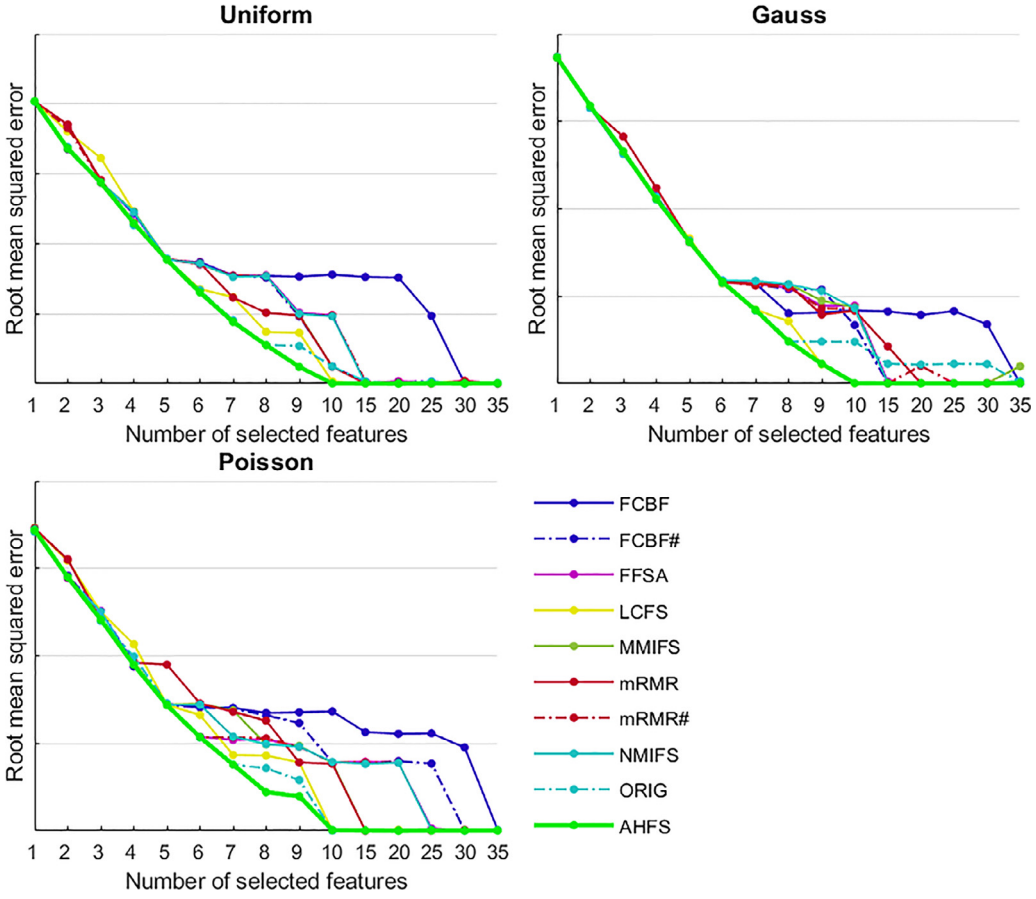

Fig. 2. Model error as a function of the number of selected features related to all the three chosen distributions.

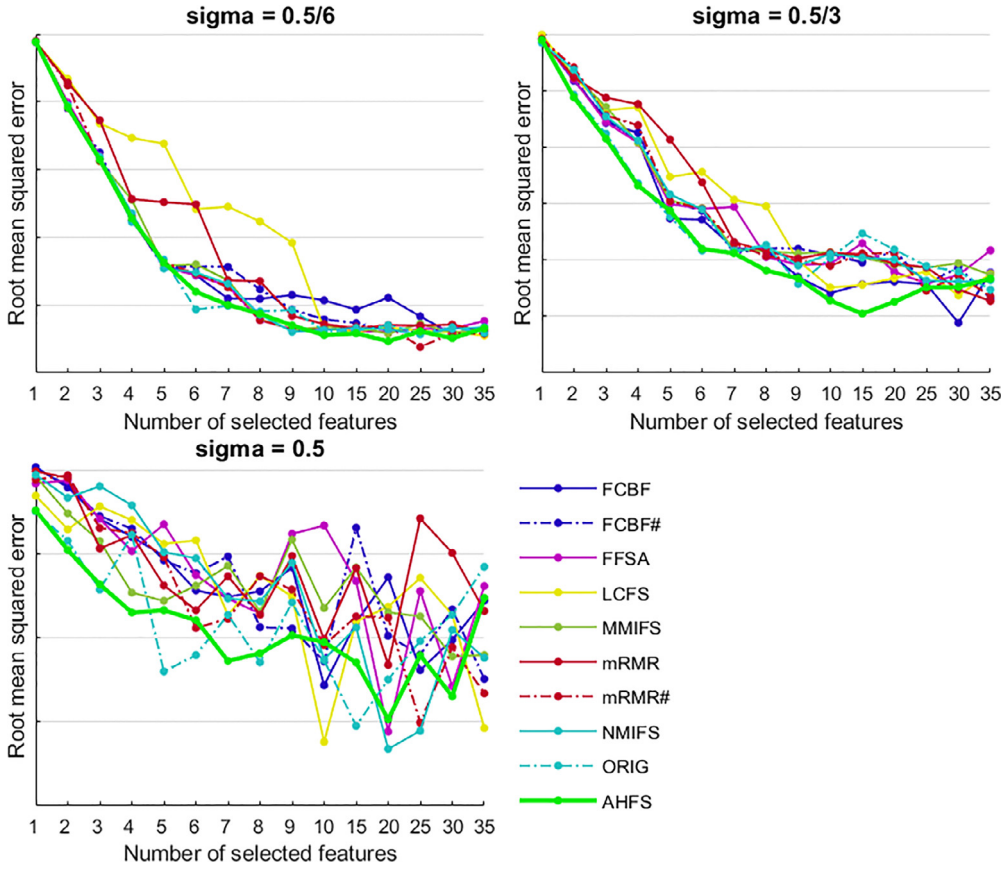

Fig. 3. Model error as a function of the number of selected features related to all the three noise levels.

Features from set S and 1 candidate feature from set S_p will be used as inputs, as well as the target variable as a single output. In this way as many models as unique features are in S_p are be generated. As S_p is not a multiset, this quantity is just the same as $|S_p|$. In order to evaluate the performance of a candidate features, the MLP models (described in paragraph 3.3) with the generated configurations are trained. In every iteration only one feature is selected and added to the output set S, until the expected number of features is reached.

The operation of the proposed algorithm on Housing dataset [33] is demonstrated on Fig. 1. The nodes of the graph contain the indices of features as numbers with frames using different color variations. The meaning of colors vary depending on the state of the examined variable. Features with black frames are the already selected feature set S in the current state, while the colorful ones (red, green) are the candidate features in set S_p. Green frame signs the best feature, which has the smallest estimation error compared to the other possible variables in red frames. Directed edges represent the transition between different states of feature subsets, which are in this case the different predetermined feature selection methods.

In addition to the best subset of features, *the proposed algorithm gives a collection of the (in the given search step) best feature selection methods per variables as result*. These lists contain the names of the algorithms, which have selected the examined variable.

4. Evaluation

Four different directions of tests were applied to comprehensively evaluate and compare the capabilities of the proposed AHFS algorithm:

- Known linear functions were used for test data generation and additional effects were given to this data by varying its distributions, noise and outlier levels for analyze their effects on the performance of the AHFS.
- Known non-linear functions were used for test data generation by Gaussian distribution with added middel level noise and outliers.
- Various test were performed on real, well-known benchmark data-sets from the UCI Machine Learning repository and on some other real data-sets, collected by the authors during various industrial applications.
- Finally, AHFS was compared on a small and also on a big data-set to some other, very recent highest level state-of-the-art feature selection algorithms.

4.1. Evaluation on artificial data sets with known effects

4.1.1. Experiments on linear functions with varying distribution, noise level and outliers

In order to provide much more insight into the advantages and challenges of the proposed data analysis approach an evaluation through artificial datasets with known and tunable effects were carried out. During these tests the performance of the proposed algorithm (AHFS) and its components with respect to the evolution of model error was compared while the number of selected features was increased.

At first, data from three different source distributions were generated. 1000-by-15 feature matrices consisting of entries whose value came from the specified distributions independently upon the value of other cells were created. 10 out of the 15 features were used to create the target column by applying a linear function with random coefficients to the chosen features (let us call

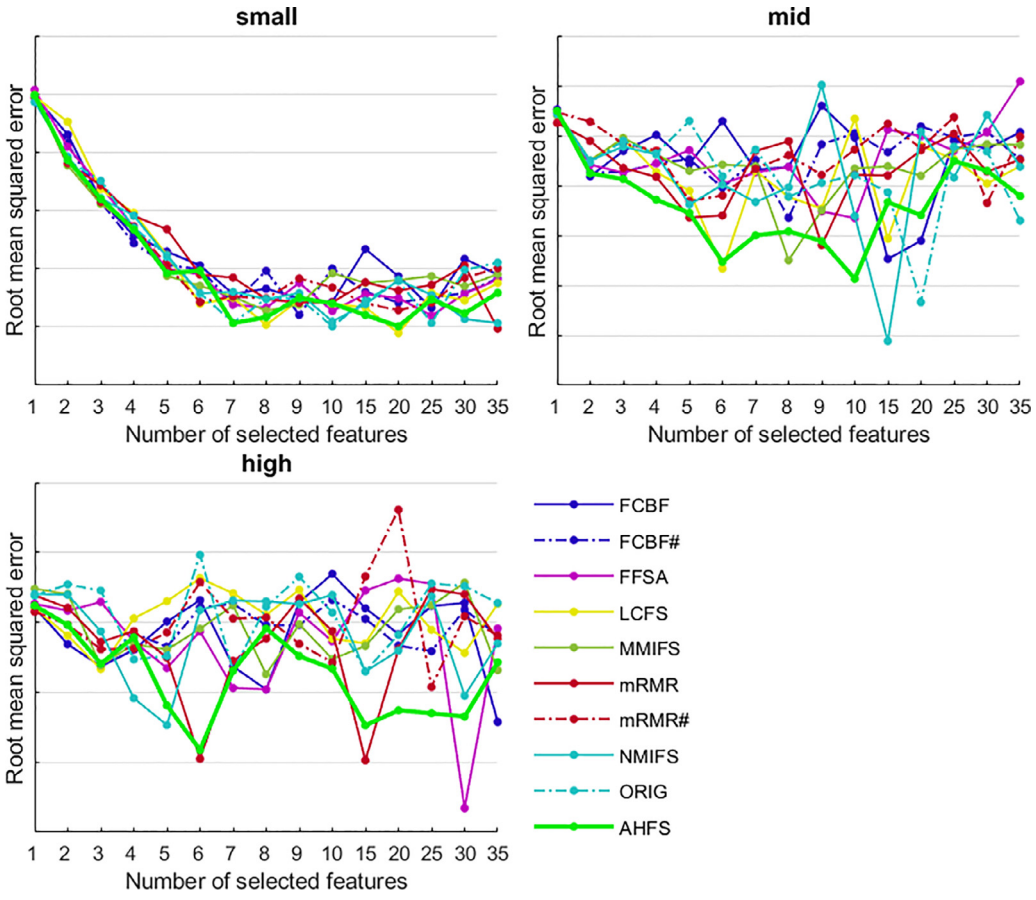

Fig. 4. Model error as a function of the number of selected features related to all the three outlier levels.

these as informative features in the next paragraphs), while the remaining 5 feature are independent, random values. The chosen distributions were Uniform on the interval [0, 1], Poisson with lambda parameter 50, and Standard, normal distribution.

The results can be seen in Fig. 2 showing the model errors according to the number of selected (input) features. It is obvious that AHFS can immediately find every informative feature resulting in a strictly monotonously decreasing model error which is saturated after all the informative features have been found. The majority of other algorithms can't accomplish this or even if they can, they resulted in much higher model errors. This difference is especially outstanding in case of the Poisson distribution. *It can be seen that the proposed algorithm (AHFS) outperforms all of the individual (incorporated) methods in all data distribution cases.*

In the second case, the effect of adding noise to the data to different extents were tested. Hereafter, for the sake of simplicity uniform distribution was used during the data generation process and after creating the target, it was normalized linearly to the [0, 1] range. The noise distribution was Gaussian with three different values of their standard deviation (*sigma*, σ): 0,5; 0,5/3; 0,5/6 (0.5 is the half of the total data range). Noise was applied to all the entries of the feature matrix and the target column as well. The results are represented in Fig. 3. According to the graphs, AHFS is much better against any other of the individual (component) algorithms in respect to the model error. This fact is outstandingly justified by the graph belonging to the "sigma = 0.5/3" and "sigma = 0.5/6" (smaller noise) cases. *It can be seen that AHFS performs significantly better than the other algorithms on all noise levels.*

Interesting insights can also be aquired by exposing the uniformly generated, linear dataset to *various outlier levels*. Results are shown in Fig. 4.

Different portions of the entries of these matrices were randomly selected and uniformly distributed noise generating the outliers (with parameters specified below) was added to the chosen values. Three different noise levels were labeled as "small", "mid(le)" and "high". The parameters corresponding to these labels respectively are the following: 2% ±[1, 1.5]; 5% ±[1.5, 2]; 10% ±[2, 3]. The percentages corresponds to the chosen portions of the matrices and the "±" sign refers to the process of generating outliers. Namely, that after choosing a value to be ruined, the sign of the uniformly distributed value in the specified range (specified in []) was then added to the cell of the matrix that was chosen randomly. The original data were pre-normalized to [0,1] range before, so, the e.g. +[1,1.5] transforms the given data clearly to outside the original, complete data range, so, the outlier level is significant. *The benefits of AHFS is slighter than that was in the previous experiments but the advantage is still obvious for handling outliers.*

4.1.2. Experiments on non-linear dependencies

Finally, the non-linear benchmark regression problem called Friedman1 was used. The results are illustrated by Fig. 5. In this problem there are 5 informative features which are in a highly non-linear relationship to the target, but 3 more features which didn't have anything to do with the target were put in. Two different scenarios were distinguished: one, in which the data-set wasn't exposed to any disturbing effect (and, with uniform distribution), and another, in which noise (with sigma = 0.5/3) and outliers (on "mid" level) with Gauss distribution we applied in the same time.*As it*

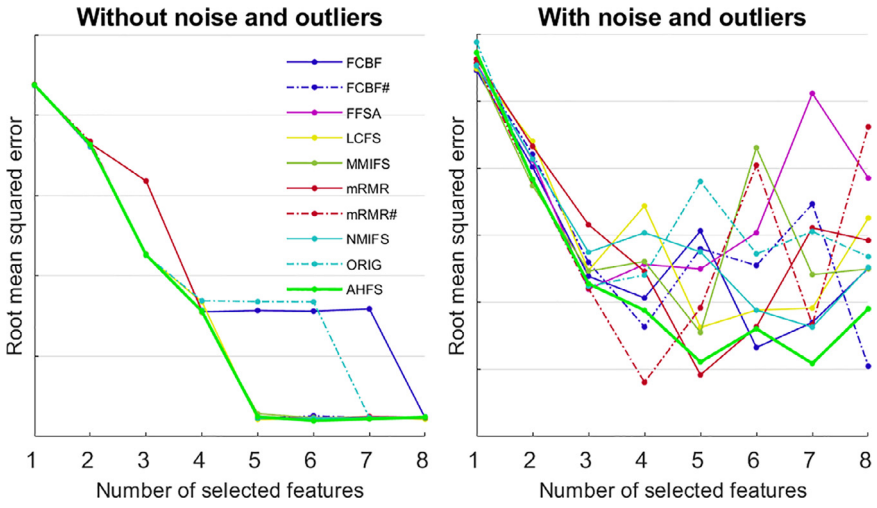

Fig. 5. Model error as a function of the number of selected features related to the two variants of the Friedman1 regression problem.

was seen in the experiments, the proposed algorithm outperforms the other ones in case of high non-linear dependencies as well.

4.2. Evaluation on becnhmarking datasets

There are various applied test cases, the UCI machine learning repository is the probably most frequently applied by the Artificial Intelligence/Machine Learning community [33]. *Iris* as the most frequently used, well-known dataset was selected for having as one of the first classification test assignment. For regression oriented tests another dataset named *Housing (also named as Boston)* was selected as a public benchmark case. In order for having noise free dataset together with also noisy data from the same domain *Calculated cutting* and *Measured cutting* is applied, respectively. *Wind turbine monitoring* and *Situation detection* during special machining are test cases for higher data amount and for high data complexity/variety, with various levels of noisy data, incorporating party redundancy, high non-linearity, outliers, non-uniform data distribution and many other disturbing, industrial real-life effects. The datasets can be described as: *Calculated cutting*: this test case consists of different machine setting, cutting tool, monitoring and product quality parameters of metal cutting (turning); *Measured cutting*: this test case also consists of cutting parameters, but in contrast to the previous one, this was collected from real measurements; *Iris*: this is one of the most popular public benchmark database [33], which is used frequently for comparing different Machine Learning methods; *Housing*: another well-known, regression benchmark dataset, "Housing", was also selected [33], which describes the properties of some suburban real estate of Boston; *Wind turbine SCADA*: this test case contains measured values that describe the detailed state of wind turbines and they were gathered from the wind turbine SCADA; *Wind turbine monitoring*: this test case utilizes the high frequency data of the many monitoring sensors inside a working wind turbine; *Situation detection during special machining*: this dataset was built from high frequency monitoring parameters of a special machining process over multiple experiments. Table 1 summarizes their sizes and types.

4.2.1. Measured cutting

This simple and small test case is inherited from real metal (steel) cutting measurements (preformed by the corresponding author). There are three assignments regarding this dataset.

In Fig. 6 each line of the diagrams describes the performance of a single feature selection method, where the x axis shows the number of features used for building the model and the y axis shows the related model error. For the sake of simplicity, only the first 50 features were selected for each assignment, if the dataset contains at least that many features. If there are less than 50 features, then every variable is selected.

Fig. 7 shows the operation of the proposed algorithm. This table is a simplified version of the operation graph shown in Fig. 1 at Section 3.4. Each row correspond to a feature selection method, while the columns denote the number of selected features which is the equivalent of the step number the algorithm is currently at. The gray cells, with plus sign inside them, mark the methods that choose the best feature in that step, while the white cells, with minus sign, mean that those methods choose other features that proved to be less desirable during the model based evaluation. The operation diagram is useful to see how the algorithm works, and how diverse the selection of the individual methods is, moreover, it represents that in many cases multiple feature selection methods choose the same, best parameter, while others select worse ones in terms of model accuracy.

The first assignment is the estimation of surface roughness R_a. *The proposed algorithm preformed slightly better than the individual methods* (Fig. 6 (a)). In many cases the same features are selected by each feature selection methods as shown in Fig. 7 (a). The main component of the cutting force was estimated in the second assignment. The selection of the first two features were diverse, but *the AHFS selected the variable with minimal error value* proposed by LCFS, too (Fig. 6 (b)). The last assignment for this dataset is the cutting temperature estimation. Fig. 6 (c) shows that *the greedy evaluation affects the selection slightly at step 3, but before the third and after the fourth variable the AHFS does not suffer from this inconvenience*.

4.2.2. Situation detection during special machining

The situation detection during special machining test case consists high frequency monitoring parameters of a special machining process. The dataset contains over 10 000 samples and more than 1100 variables. The assignment is a binary classification problem: estimating whether the situation has happened or not.

According to Fig. 8, the ORIG method performed extremely poor compared to the other methods. Furthermore, the proposed AHFS algorithm provides the best performance but it is worth noting, that the MMIFS method holds the same model errors as the AHFS for the first 10 selected features, although it loses accuracy compared to the AHFS above 10 features. This is mirrored in the selection table in Fig. 9, too. It is also worth noticing, that this as-

Table 1
Dataset properties.

	Samples	Dimension	Category
Calculated cutting	450	9	Industry
Measured cutting	120	7	Industry
Iris	150	7	Biology
Housing	506	14	Socioeconomy
Wind turbine SCADA	839	57	Energy
Wind turbine monitoring - Oil pressure	1857	998	Energy
Wind turbine monitoring - Oil temperature	1886	866	Energy
Wind turbine monitoring - Bearing temperature	1876	951	Energy
Situation detection during special machining	> 10,000	> 1,100	Industry

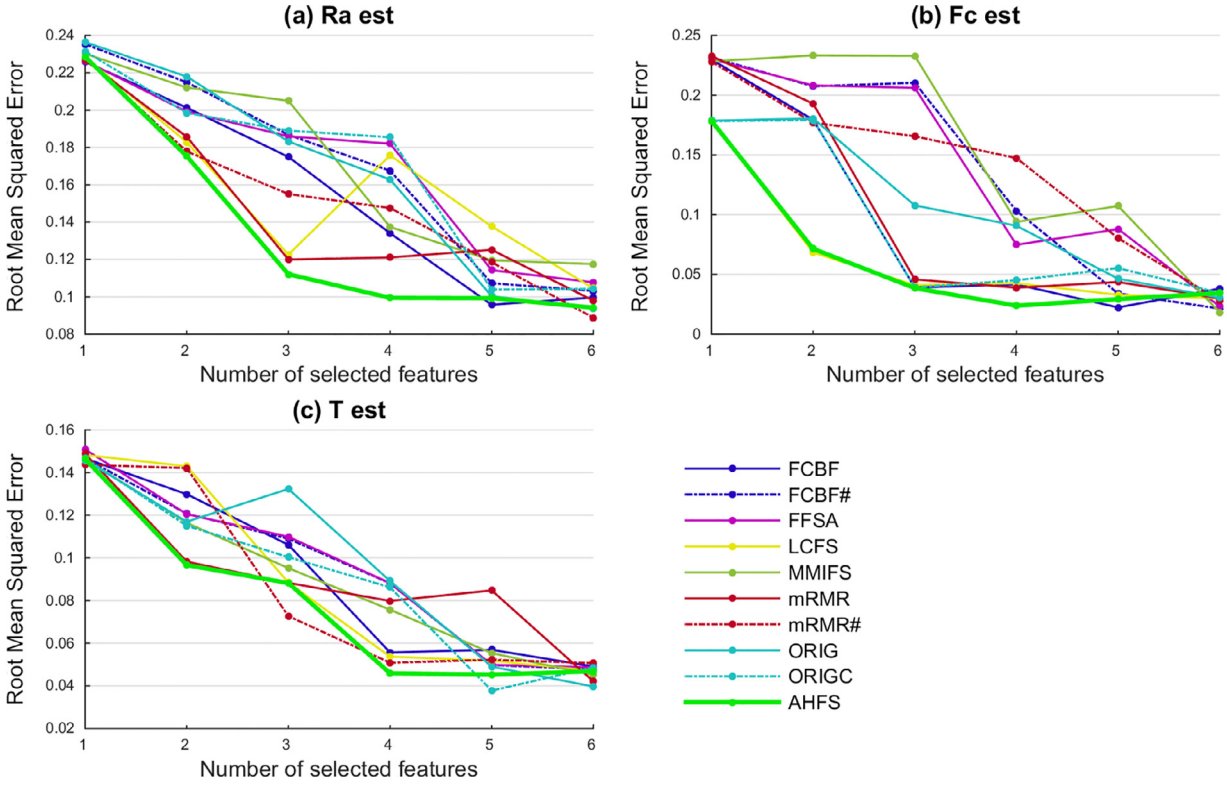

Fig. 6. Performance on Measured cutting dataset: (a) R_a est, (b) F_c est and (c) T est.

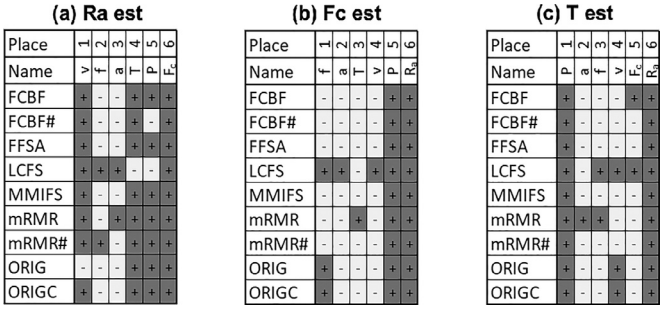

Fig. 7. Algorithms selected on Measured cutting dataset: (a) R_a est, (b) F_c est and (c) T est.

signment's selection matrix is sparse, meaning that at most of the steps, only one feature selection method provided the best feature.

It has to be emphasized that the proposed algorithm was already successfully applied in the industrial collaboration of the authors before writing this paper. The models prepared according to the results of the paper are already incorporated in the control system of the related machines and works well (detects difficult identifiable situations) on the shop floor, in the daily production.

4.3. Comprehensive evaluation

The previous sections detailed the performance results of the individual dataset tests and in the end of this section a comprehensive evaluation is given to summarize the results of the individual dataset tests and highlight the main characteristics of the proposed algorithm. *To give an overall performance evaluation, an average of the individual feature selection algorithms' performances has been calculated and presented*. The modeling error values, measured at the evaluation of individual estimation assignments, had to be normalized using a linear scale into the range from 0 to 1 in order to compare them to each-other and to enable the calculation of an average performance.

As detailed before, some datasets have less than 50 features and in their case, naturally, the algorithms selected less features and consequently, less model was built for evaluation. As consequence, the average performance was calculated from less and less assignments as the number of selected features grows. For example the average performance of the first 2 selected features is calculated from all of the assignments, because every dataset has at least 2

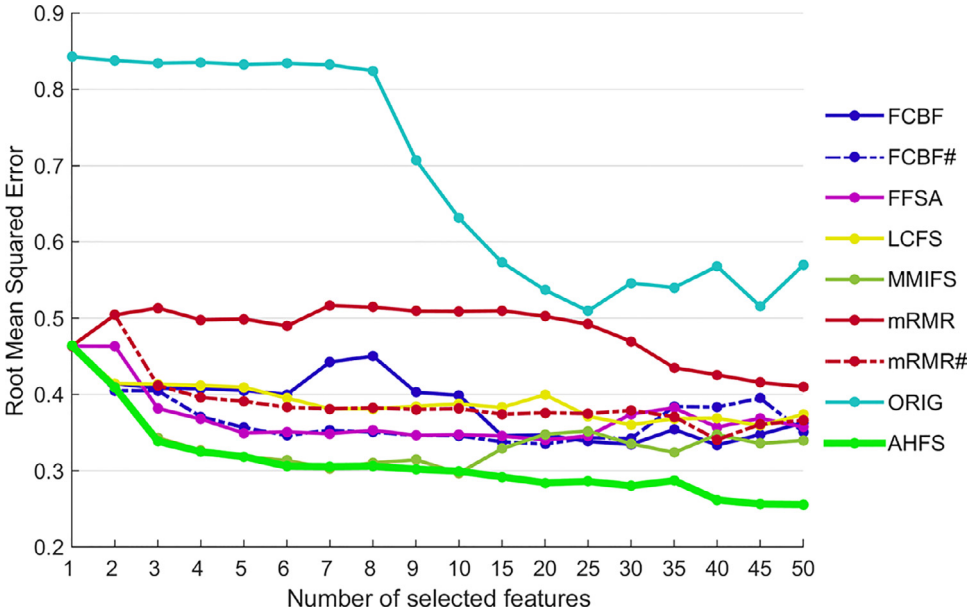

Fig. 8. Performance on Situation detection during special machining dataset: SM situation est.

Place	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Name	Var-1011	Var-0114	Var-0139	Var-1020	Var-0151	Var-0003	Var-1010	Var-1001	Var-0129	Var-0912	Var-0150	Var-0149	Var-0138	Var-0001	Var-0719	Var-0429	Var-0441	Var-0105	Var-0731	Var-0602	Var-0002	Var-0440	Var-1022	Var-0730	Var-0004	Var-1021	Var-0729	Var-0140	Var-0005	Var-0419	Var-0012	Var-0720	Var-0439	Var-0035	Var-0006	Var-0710	Var-0007	Var-0008	Var-0011	Var-0009	Var-0010	Var-0428	Var-0709	Var-0013	Var-0014	Var-1000	Var-0722	Var-0627	Var-0430	Var-1013
FCBF	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FCBF#	+	+	-	-	-	-	-	-	-	-	-	+	-	-	+	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	-	+	-	-	+	+	-	-	+
FFSA	+	-	-	-	-	-	-	+	+	-	+	+	+	-	-	+	-	-	+	-	-	+	-	+	-	+	-	-	-	+	-	+	-	-	+	-	-	-	-	-	+	+	-	-	+	-	-	-	-
LCFS	+	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	+	-	-	+	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-
MMIFS	+	+	+	+	+	-	-	-	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	+	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
mRMR	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-
mRMR#	+	-	-	-	-	-	+	+	+	-	+	-	-	-	-	-	-	-	+	-	-	-	+	+	-	+	-	-	-	+	-	+	-	-	+	-	-	-	-	-	-	+	+	-	-	+	-	-	-	-
ORIG	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	-	-	+	-	-	-	+	-	-	-	-	-	-	+	-	+	+	-	+	+	-	+	+	-	-	-	-	-
ORIGC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Fig. 9. Algorithms selected on Situation detection during special machining dataset: SM situation est.

features to select, but in the case of the first 30 features, only the assignments of the wind turbine SCADA/monitoring and the situation detection datasets were used because all the other have less than 30 variables in total.

Fig. 10 shows the average of the individual feature selection performances measured on each assignment. Each line describes the performance of a single, individual feature selection method, where the x axis shows the number of features used for building the model and the y axis shows the normalized model error.

The overall performance diagrams in Fig 10

mirrors, that the proposed algorithm significantly outperforms the individual methods in general. Furthermore, the biggest difference reveals itself in the case of the first 4 to 25 selected features which means that the new method finds the most important features earlier than the other methods.

Table 2 mirrors the overall performance (model accuracy) improvement of the proposed algorithm compared to the individual, state-of-the-art methods.

rows of the table denote the base of the comparison, in the first row (MEAN), the proposed algorithm is compared to the mean performance of the individual methods; in the second row (MIN), it is compared to the minimum of the individual method errors for each feature number; and in the third row (MMIFS), it is compared to the best individual method, which provided the best model accuracy in general. The columns indicate that the model errors of the FIRST N features were averaged in the comparison. So, for example the first column of the first row shows that the average model error of the individual methods is 210% of the model error of the proposed AHFS algorithm, when considering of the first 5 features. In other words the average model error, calculated on the first 5 features selected by the proposed algorithm, is approximately half of the average model error of the individual methods. The percentages vary from 139% to 304% with the overall average

Table 2
Comparison of the proposed algorithm with the individual methods.

	FIRST 5	FIRST 10	FIRST 30	FIRST 50
MEAN	**183%**	**218%**	**236%**	**304%**
MIN	139%	152%	153%	167%
MMIFS	147%	156%	156%	183%

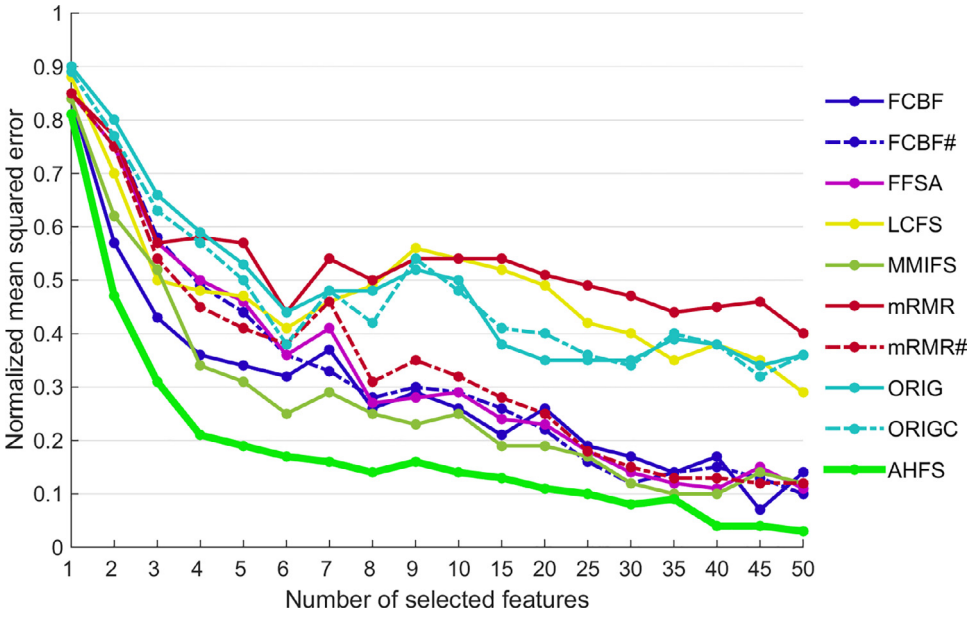

Fig. 10. Average performance of the proposed algorithm and the individual methods.

of 183%, which concludes that ***the AHFS nearly doubles the accuracy (resulting in around half value for the related modeling error) compared to the individual methods, making it a superior feature selection algorithm***. It is worth mentioning that the individual algorithms are well-known and widely applied, best methods.

4.3.1. Calculation time performance

The previous paragraphs evaluated the proposed AHFS algorithm in the modelling accuracy point of view, the current one *evaluates it's computational requirement*. Dongmei Mo and Zhihui Lai proposed a robust jointly sparse regression for effective feature selection [34]. Their experimental results indicate that the proposed method can outperform the locality based methods (LPP, OLPP, FOLPP), the joint sparsity learning methods (JELSR) and the L1-norm based methods (SLE, LPP-L1) with strong robustness. However, the reported computational complexity is much higher than the traditional methods, such as PCA, LPP and RR, in this aspect this scientific result is similar to the AHFS results. Moreover, Zhao et. al. presented a classified nested equivalence class (CNEC)-based approach to calculate the information-entropy-based significance for feature selection using rough set theory [35]. Their goal was to increase the computational efficiency of the information-entropy-based significance measure and showed that their solution is indeed greatly improved the runtime of multiple feature selection algorithms that use rough set theory, which can be beneficiary in further developments of the AHFS method.

Fig. 11 *represents the running time demands of datasets in a comprehensive way, considering the number of samples (X, horizontal axis) and features (Y, vertical axis) displayed on a logarithmic scale, while the sizes of the circles represent the calculation time requirement*. For simplicity, when multiple assignments are defined on the same dataset, the average computation time of the assignments is calculated and plotted. It is worth noticing, that in the case of Wind turbine SCADA, the proportion of training time is significantly higher, than in the case of other datasets. This characteristic of the proposed method is the consequence of: (a) if the dataset has low dimension and a smaller number of samples then the computation time of model building and feature selection is in the same order of magnitude, (b) if the number of features is considerably high (like more than 500), then computational time of the mutual information matrix (required by most of the individual methods) becomes higher than the time taken by model training. *Taking everything into account, the algorithm has bigger time demand, between 171% and 7571%, so an average of 2246% of the mean over all the other, individual algorithms*.

Because of the already high computational times of the AHFS algorithm, the JMIM and NJMIM methods were discarded as they are significantly slower compared to the other information theory based solutions, moreover, they don't add enough value in terms of potential feature candidates, to justify their high computational demands.

The complexity can be linear to the number of iterations in a random search, but experiments show that in order to find best feature subset, the number of iterations required is mostly at least quadratic to the number of features. The main reason is that most existing subset search methods demand the analysis of pairwise correlations between all features (named F-correlation). With quadratic or higher time complexity in terms of dimensionality these algorithms do not have strong scalability to deal with extreme high dimensional data.

The runtime complexity of the proposed AHFS method highly depends on the complexity of other FS algorithms used as submodules. However, runtime requirements and also complexity of the AHFS algorithm can be reduced in the future by switching the independent, interchangeable modules.

4.4. Comparison with other state-of-the-art feature selection techniques

This paragraph focuses on the comparison of the proposed AHFS to the family of Relief-based feature selection methods [36]*, they are really recent developments of the field, moreover, they are state-of-the-art algorithms*. Relief calculates a feature score for each feature. Using this score, the most important features of a dataset can be ranked and selected. The feature scoring is based on the identification of feature value differences between nearest neighbor (NN) instance pairs. In case of feature value difference, the score is decreasing (increasing), if the corresponding class labels are the same (different). The original Relief can deal with discrete and continuous attributes and it is limited to only two-class problems. How-

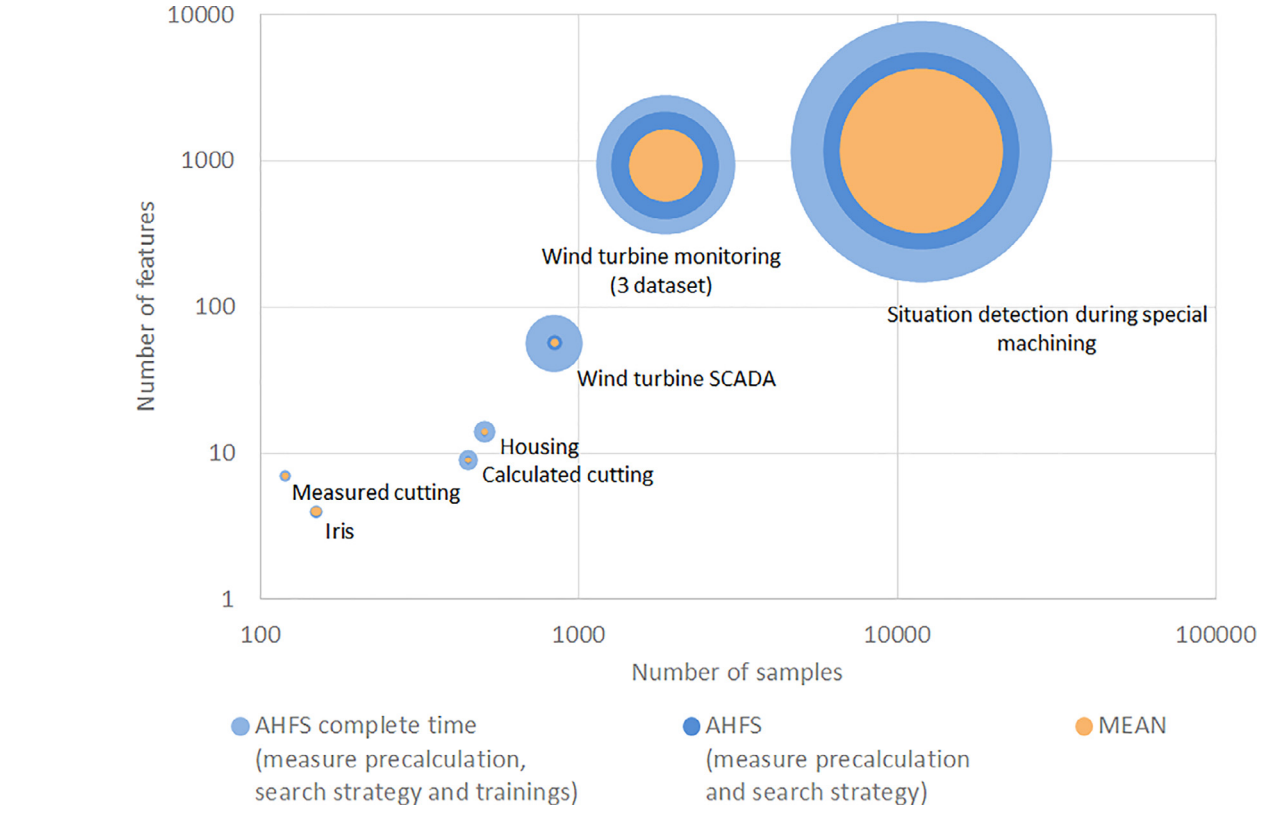

Fig. 11. Comprehensive feature selection time performance considering different dataset sizes.

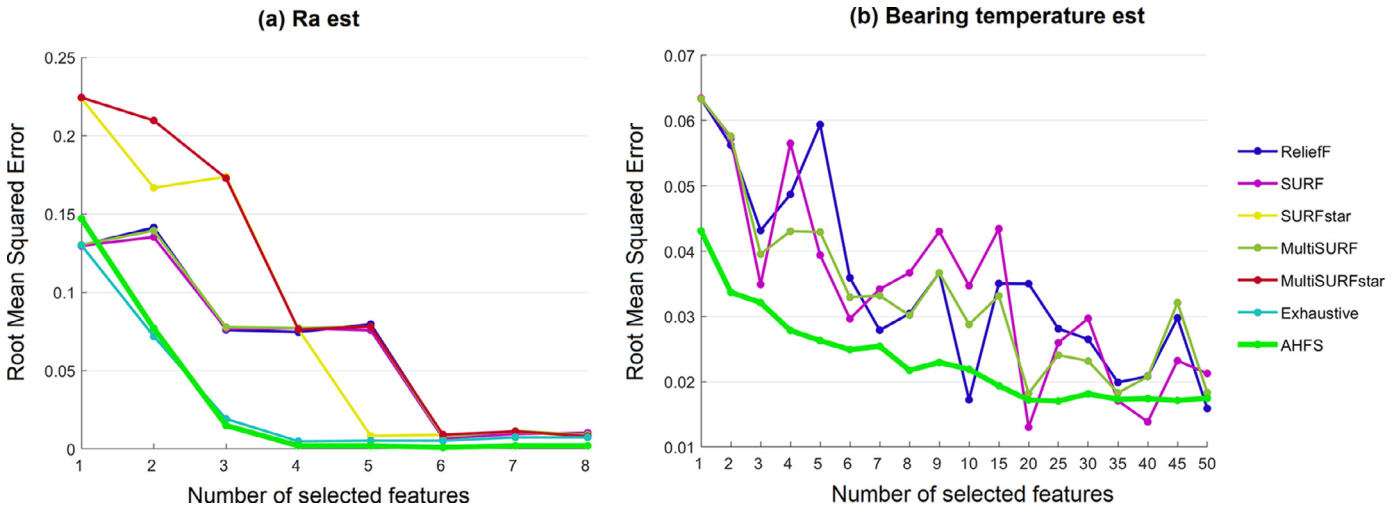

Fig. 12. Performance as model error on (a) the Calculated cutting dataset: R_a est(imation) and (b) Wind turbine monitoring: Bearing temperature est(imation). In case of (b) for the SURFstar and MultiSURFstar methods their RMSEs are much higher, therefore, they are excluded from the figure (they are far above the current scale). Also worth mentioning, that the Exhausive search is unfeasable with higher number of features, like in case (b).

ever, multiple extensions have been proposed to deal with noisy, incomplete data and it is also adapted to work with regression.

A freely available Relief Based Algorithm Training Environment (ReBATE) was applied to fairly test and compare the performance of the proposed AHFS algorithm with this widely used, recent, state-of-the-art feature selection methods. ReBATE has been implemented with five core Relief-based Algorithms (RBAs): ReliefF, SURF, SURFstar (SURF*), MultiSURF, MultiSURFstar (MultiSURF*). These feature selection methods were explicitly developed for noisy regression tasks, and were tested on numerous real-world problems, they are among the most recent and best performing state-of-the-are feature selection algorithms.

Furthermore, in order to compare the results to "the ideal" results, AHFS was compared with an Exhaustive search in case of a smaller dataset (Calculated cutting, R_a est). All possible subsets are selected and evaluated during this kind of feature selection, the best result for each subset is selected and shown in Fig. 12 (a). Exhaustive search represents the best possible results with exceptional high computational time requirement. With high number of features, the execution time explodes, so, in such cases this kind of evaluation is unfeasible.

As presented in Fig. 12 *the proposed AHFS method serves with the smallest modeling error outperforming other state-of-the-art feature selection methods*. The feature subsets produced by AHFS are more consistent and therefore more reliable even in smaller subsets than its competitors'. It is worth mentioning that the performance of AHFS is the same as the exhaustive search in Fig. 12 (a), however its required execution time is just a fraction of what exhaustive search requires. In the given case of Calculated cutting, R_a est, exhaustive search has 9525% times bigger calculation time demand, than the proposed AHFS algorithm. *As conclusion, it was measured that the proposed AHFS algorithm outperformed the recent state-of-the-art feature selection algorithms, moreover, it serves with the same modeling accuracy as the ideal exhaustive search algorithm, but the computational demand of AHFS is smaller in several magnitudes. These results also prove the superiority of the proposed AHFS algorithm.*

It shall be mentioned that thanks to the flexible structure of the proposed AHFS algorihtm, these other, state-of-the-art feature selection algorithms can be easily included in the AHFS solution.

5. Conclusions

A novel Adaptive, Hybrid Feature Selection (AHFS) approach has been presented, which chooses a combination of most suitable feature selection methods for a given problem in order to achieve the best feature order (aiming to build up the most accurate model), e.g. because there exists no general, "best of" or "best practice" feature selection solution in Machine Learning field. *Adaptivity* of the proposed algorithm is realized in such a way that at an individual step of the feature selection algorithm it iterates not only in the space of the variables but in the space of available features selection techniques, too. This is the core idea presented in the paper. A double level *hybrid* solution is proposed in the paper because the introduced algorithm combines the given, available feature selection techniques and also it utilizes the applied learning model in its mathematical algorithm.

Different feature selection methods were presented in detail with examples of their applications and an exhausting evaluation has been carried out to measure and compare the performance of them to the proposed approach.

Evaluation and comparison experiments were performed on artificial data sets with known effects having (simple) linear dependencies with included independent variables, together with varying distributions, noise levels and outlier disturbances. AHFS was compared also on artificial, but highly non-linear dataset ruined with Gaussian data distribution, middle level noise and outliers. *Independently from the linearity, distribution, noise and outlier presence AHFS consequently showed its superiority over the individual, state-of-the-art feature selection algorithms proving its robustness against such challenging effects.*

Test on real-life benchmarking and industrial datasets proved that while the individual feature selection methods may perform badly on one or more of the test cases, the combined AHFS algorithm steadily provides noticeably better solution. In comparison, modelling accuracy improvement percentages vary from 139% to 304% with the overall average of 183%, which concludes that the *AHFS nearly doubles the accuracy (resulting in around half value for the related modelling error) compared to the individual methods, making it a superior feature selection algorithm.*

Since the AHFS must calculate all of the measures used by all of the incorporated feature selection algorithms, its computational requirement is always higher than the requirements of the other algorithms, individually and together, as well. Moreover, it incorporates model training steps which also have more significant time demand in many cases. The required computational need is varying between 171% and 7571% compared to the average need of the previous, individual solutions, all in all in average its time requirement ratio is 2246%.

At the final evaluation stage, AHFS was compared to five, completely independent, recent, well performing Relief-based feature selection methods. *It was measured that the proposed AHFS algorithm outperformed the recent sate-of-the-art feature selection algorithms, moreover, it serves with the same modeling accuracy as the ideal exhaustive search algorithm, but its computational demand is smaller in several magnitudes.*

It has to be emphasized that *the proposed algorithm was already successfully applied in the industrial collaborations of the authors* before writing this paper. The models prepared according to the results of AHFS are already incorporated in the control system of the related machines and works well (detects difficult identifiable situations) on the shop floor, in the daily production.

Future research is needed to reduce the computational time of the AHFS algorithm, which is currently a disadvantage compared to the individual filter methods. Drawbacks can be eliminated by using a different, not greedy search strategy that is able to "think ahead" and, at the same time, reduce the number of required model training. Moreover, it would be very useful having a general metric substituting the model based evaluation, however, this problem seems to be extremely difficult. Finally, the selective involvement (calculation) of the incorporated individual feature selection algorithms at the individual search steps could increase the speed of AHFS significantly, this approach in investigated currently by some of the authors with some first success already.

All in all, the proposed AHFS algorithm already proved to be superior to other state-of-the-art feature selection methods for the reasons that, 1) it is significantly less sensitive to the varying properties of the dataset it is applied to, and 2) it provides a significantly better feature order for model building.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research in this paper was supported by the European Commission through the H2020 project EPIC (https://www.centre-epic.eu/) under grant No. 739592, by the grant of the Highly Industrialised Region in Western Hungary with limited R&D capacity: "Strengthening of the regional research competencies related to future-oriented manufacturing technologies and products of strategic industries by a research and development program carried out in comprehensive collaboration", under grant No. VKSZ_12-1-2013-0038, by the Hungarian ED_18-2-2018-0006 grant on a "Research on prime exploitation of the potential provided by the industrial digitalisation" and by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program.

References

[1] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.
[2] M.S. Srivastava, M.N. Joshi, M.M. Gaur, A review paper on feature selection methodologies and their applications, International Journal of Engineering Research and Development 7 (2013) 57–61.
[3] S. Muñoz-Romero, A. Gorostiaga, C. Soguero-Ruiz, I. Mora-Jiménez, J.L. Rojo-Álvarez, Informative variable identifier: expanding interpretability in feature selection, Pattern Recognit 98 (2020).
[4] R. Shang, Y. Meng, W. Wang, F. Shang, L. Jiao, Local discriminative based sparse subspace learning for feature selection, Pattern Recognit. 92 (2019) 219–230.
[5] Y. Zhang, Q. Wang, D.-w. Gong, X.-f. Song, Nonnegative laplacian embedding guided subspace learning for unsupervised feature selection, Pattern Recognit. 93 (2019) 337–352.

[6] L. Zini, N. Noceti, G. Fusco, F. Odone, Structured multi-class feature selection with an application to face recognition, Pattern Recognit. Lett. 55 (2015) 35–41.
[7] Y. Jiang, C. Li, MRMR-based feature selection for classification of cotton foreign matter using hyperspectral imaging, Comput. Electron. Agric. 119 (2015) 191–200.
[8] Z. Zhang, H. Chen, Y. Xu, J. Zhong, N. Lv, S. Chen, Multisensor-based real-time quality monitoring by means of feature extraction, selection and modeling for al alloy in arc welding, Mech. Syst. Signal Process. 60–61 (2015) 151–165.
[9] K. Zhang, Y. Li, P. Scarf, A. Ball, Feature selection for high-dimensional machinery fault diagnosis data using multiple models and radial basis function networks, Neurocomputing 74 (2011) 2941–2952.
[10] J.A. Carta, P. Cabrera, J.M. Matías, F. Castellano, Comparison of feature selection methods using ANNs in MCP-wind speed methods. a case study, Appl. Energy 158 (2015) 490–507.
[11] X. Kong, X. Liu, R. Shi, K.Y. Lee, Wind speed prediction using reduced support vector machines with feature selection, Neurocomputing 169 (2015) 449–456.
[12] J. Ircio, A. Lojo, U. Mori, J. Lozano, Mutual information based feature subset selection in multivariate time series classification, Pattern Recognit. 108 (2020) 107525, doi:10.1016/j.patcog.2020.107525.
[13] S.-y. Jiang, L.-x. Wang, Efficient feature selection based on correlation measure between continuous and discrete features, Inf. Process. Lett. 116 (2016) 203–215.
[14] B. Senliol, G. Gulgezen, L. Yu, Z. Cataltepe, Fast correlation based filter (FCBF) with a different search strategy, 2008.
[15] Y. Jiang, C. Li, A fault diagnosis scheme for planetary gearboxes using modified multi-scale symbolic dynamic entropy and mRMR feature selection, Mech. Syst. Signal Process. 91 (2017) 295–312.
[16] S. Sharmin, M. Shoyaib, A.A. Ali, M.A.H. Khan, O. Chae, Simultaneous feature selection and discretization based on mutual information, Pattern Recognit. 91 (2019) 162–174.
[17] F. Amiri, M.R. Yousefi, C. Lucas, A. Shakery, N. Yazdani, Mutual information-based feature selection for intrusion detection systems, Journal of Network and Computer Applications 34 (2011) 1184–1199.
[18] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5 (4) (1994) 537–550.
[19] J. Song, Z. Zhu, P. Scully, C. Price, Modified mutual information-based feature selection for intrusion detection systems in decision tree learning, J. Comput. (Taipei) 9 (7) (2014) 1542–1546.
[20] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, 2003.
[21] Y. Liu, J. Zhang, L. Ma, A fault diagnosis approach for diesel engines based on self-adaptive WVD, improved FCBF and PECOC-RVM, Neurocomputing 177 (2016) 600–611.
[22] Y. Jiang, C. Li, MRMR-based feature selection for classification of cotton foreign matter using hyperspectral imaging, Comput. Electron. Agric. 119 (2015) 191–200.
[23] H.H. Yang, J. Moody, Feature selection based on joint mutual information, 1999, pp. 22–25.
[24] H.H. Yang, J. Moody, Data visualization and feature selection: New algorithms for nongaussian data, 1999, pp. 687–693.
[25] P.A. Devijver, J. Kittler, Pattern recognition, a statistical approach, Prentice-Hall International Inc., England, London, 1982.
[26] G. Wang, Q. Song, H. Sun, X. Zhang, B. Xu, Y. Zhou, A feature subset selection algorithm automatic recommendation method, J. Artif. Intell. Res. (JAIR) 47 (2013) 1–34.
[27] Z.J. Viharos, Automatic generation a net of models for high and low levels of production control, 16th IFAC Word Congress, 2005.
[28] Z.J. Viharos, G. Erdös, A. Kovács, L. Monostori, Ai supported maintenance and reliability system in wind energy production, 2010.
[29] Z.J. Viharos, K.B. Kis, Diagnostics of wind turbines based on incomplete sensor data, XX IMEKO World Congress - Metrology for green growth, TC10 on Technical Diagnostics, 2012.
[30] W.S. McCulloch, W.H. Pitts, A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics 5 (1945) 115–133.
[31] P.J. Werbos, Beyond Regression: New Tools for Prediction and Analysis in the Behaviour Sciences, Harward University, Cambridge, 1974.
[32] L. Deng, D. Yu, Deep learning methods and applications, Foundations and Trends in Signal Processing 7 (2014) 1–199.
[33] M. Lichman, UCI machine learning repository, 2013, http://archive.ics.uci.edu/ml.
[34] D. Mo, Z. Lai, Robust jointly sparse regression with generalized orthogonal learning for image feature selection, Pattern Recognit. 93 (2019) 164–178.
[35] J. Zhao, J.-m. Liang, Z.-n. Dong, D.-y. Tang, L. Zhen, Accelerating information entropy-based feature selection using rough set theory with classified nested equivalence classes, Pattern Recognit. 107 (2020) 107517, doi:10.1016/j.patcog.2020.107517.
[36] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: introduction and review, J. Biomed. Inform. (2018) 189–203.

Dr. Zsolt Jánános Viharos, MBA, senior research fellow and project manager of the Institute for Computer Science and Control (SZTAKI) and full researcher and lecturer of the John von Neumann University, in Hungary. In the Research Laboratory on Engineering and Management Intelligence at SZTAKI, he is the leader of the Intelligent Processes Research Group. He is leading various sizes of industrial, and national or European supported R&D projects with durations from some months up to many years. His typical roles are project sponsorship, project management and content leadership for industrial projects. He has 140 scientific publications resulted in 500 independent references, is member of the Boards of Reviewers of the scientific journals: Measurement, Applied Intelligence, Reliability Engineering and System Safety and is member, or chair of various scientific conferences. He is Chairperson of the TC10 - Measurement for Diagnostics, Optimization and Control of the Hungarian National IMEKO Committee. He is member of the IEEE (Institute of Electrical and Electronics Engineers), No.: 93787359, and of the International Society of Applied Intelligence, member of the Production Systems section of the Scientific Society for Mechanical Engineering (GTE) in Hungary and member of the Computer and Automation Committee of the public body of the Hungarian Academy of Sciences (MTA), member of the Hungarian Standards Institution (MSZT). (please, visit: http://www.sztaki.hu/~viharos)