Weka for Feature Selection and Its Application for Energy Efficiency in Smart Buildings

First A. Author, *Fellow, IEEE*, Second B. Author, and Third C. Author, Jr., *Member, IEEE*

***Abstract*—The massive collection of data via emerging technologies like the Internet of Things (IoT) requires finding optimal ways to reduce the created features that have a potencial impact on the information that can be extracted through the machine learning process.**

The mining of knowledgde related to a concept is done on the basis of the feature of the data. The process of finding the best combination of features in called feature extraction.

Different type of feature extraction methods are being used. The feature selection algorithm should fit with the offline as weil as on-line mining

***Index Terms*—feature selection, energy efficiency, smart buildings, smart cities**

I. Introduction

Feature Selection (FS) is defined in [1] as the process of eliminating features from the data base that are irrelevant to the task to be performed. It should not be confused with Feature Extraction or Dimensionality Reduction, where new features are created combining the previous. FS facilitates data understanding, reduces the measurement and storage requirements, reduces the computational process time, and reduces the size of a data set, so that model learning becomes an easier process. It also avoids the curse of dimensionality (figure explaining xxx)

A FS method can be seen as the combination of a search strategy for proposing feature subsets with a given evaluator that measures the performance of such candidates. The search space for candidate subsets has cardinality $O(2^w)$, where w is the number of features. A stopping criterion establishes when the feature selection process must finish. Such criterion can be defined as a control procedure that ensures that no further addition or deletion of features does produce a better subset, or it can be as simple as a counter of iterations.

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

FS methods are typically categorized into *wrapper*, *filter* and *embedded*, *univariate* and *multivariate* methods.

- Filter methods apply statistical measures to evaluate the set of attributes independtly [2]–[4]. Correlation with the output, mutual information and Fisher test are some examples. reconstruction error
- Wrapper methods [5] use a predetermined learning algorithm to determine the quality of selected features according to an evaluation metric [6].
- Embedded methods take advantage of their own variable selection process and perform feature selection and model fitting simultaneously [7].
- Multivariate methods evaluate features in batches.
- Univariate methods evaluate each feature independently.

The flow that characterises feature selection is depicted in Fig. I

Filter methods do not incorporate learning. Wrapper methods use a learning machine to measure the quality of subsets of features without incorporating knowledge about the specific structure of the classification or regression function, and can therefore be combined with any learning machine. In embedded methods the learning part and the feature selection part can not be separated, for example, Weston et al. [2000] measure the importance of a feature using a bound that is valid for Support Vector Machines only thus it is not possible to use this method with other models.

A. Abbreviations and Acronyms

Feature Selection: FS

II. Feature Selection in Weka

The free software licensed suite Waikato Environment for Knowledge Analysis (Weka) [8] contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.

In Weka, FS is implemented with the `weka.attributeSelection` (w.aS) package using two components of the `w.aS.AttributeSelection` (w.aS.AS) class : the search strategy (`w.aS.ASSearch` abstract class) and the evaluators (`w.aS.ASEvaluation` abstract class).

This structure allows users and programmers to configure a multitude of different methods for FS, both filter and wrapper, univariate and multivariate.

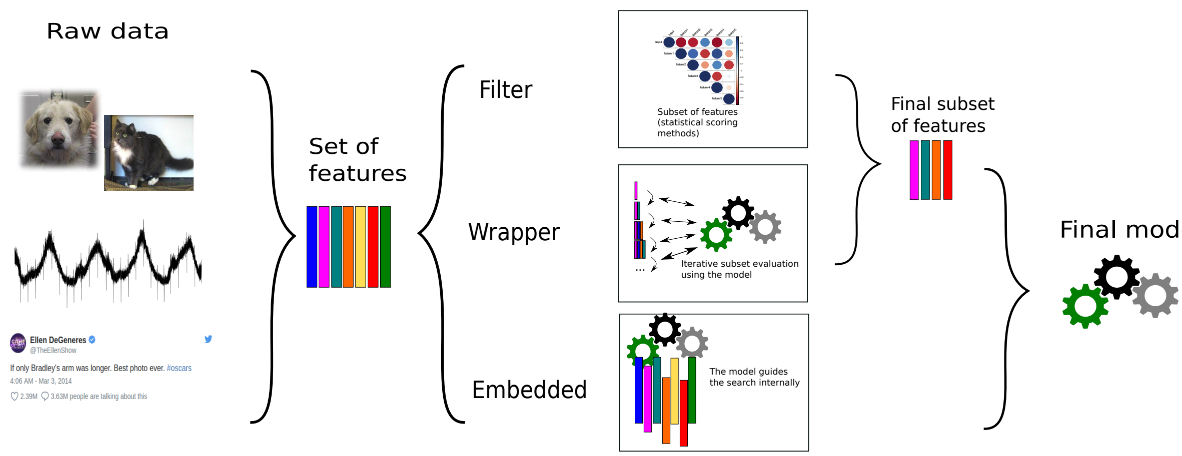

Fig. 1. The feature selection flow

Univariate methods are configured by evaluators with names ending in *AttributeEval* and multivariate methods are configured by evaluators with names ending in *SubsetEval*.

For multivariate wrapper FS methods, the *w.aS* package has the `w.aS.AS.WrapperSubsetEval` class which evaluates attribute sets by using a learning scheme with cross-validation and a performance measure.

For univariate wrapper FS methods, the `w.aS.AS.ClassifierAttributeEval` class evaluates the worth of an attribute by using a user-specified classifier, cross-validation and a performance evaluation measure to use for selecting attributes.

Since the FS and classification processes must be executed in bach mode, Weka offers the class `weka.classifiers.meta.AttributeSelectedClassifier` which is a meta-classifier where dimensionality of data is reduced by attribute selection before being passed on to a learning algorithm.

Y QU PASA CON LOS FILTER? xxx

III. Weka FS Algorithms. Methodology

In this section we analyse the search strategies, evaluators, regression methods and performance metrics that have been considered.

A. Search strategies

The search strategy will determine an interesting feature subset to train the model.

For univariate FS methods, *Ranker* method [9] is required. *Ranker* method ranks attributes by their individual evaluations. A threshold, or the number of attributes to retain, allows reducing the attribute set.

xxx mete rollo sobre ranker

For multivariate FS methods, there exist two approaches: deterministic search strategies (which xxx) and probabilistic algorithms.

- Deterministic search strategies: Greedy Stepwise [10]. `GreedyStepwise` performs a greedy forward or backward search through the space of attribute subsets, stopping when the addition (forward direction) or deletion (backward direction) of any of the remaining attributes results in a decrease in evaluation, thus, it has no backtracking capability.
- Probabilistic algorithms: `MultiObjectiveEvolutionarySearch` [11] and `PSOSearch` [12].

`MultiObjectiveEvolutionarySearch` uses evolutionary computation where two objectives are optimized: the first one is chosen by the evaluator, and it is to be maximized, while the second one is the attribute subset cardinality, and it is to be minimized. The final output is given by the non-dominated solutions in the last population having the best fitness score for the first objective.

`PSO` optimizes a problem moving individuals (particles) around in the search-space according to the particle's position and velocity, influenced by its local best known position and the best known positions.

B. Evaluators

The subsets of features are evaluated in order to know whether they are interesting or the search must continuate.

We considered the multivariate filter evaluator *ConsistencySubsetEval* [13]. *ConsistencySubsetEval* scores a subset of features as a whole, by projecting the training instances according to the attribute subset, and considering the consistency of class values in the obtained instance sets. As far as univariate filter evaluators are concerned, *RelieffAttributeEval* [14] and *PrincipalComponents* [15] were considered. *RelieffAttributeEval* evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. Can operate on both discrete and continuous class data. *PrincipalComponents* performs a principal components analysis and transformation of the data. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data (default 95%). Attribute noise can be filtered by transforming to the principal components space, eliminating some of the worst eigenvectors, and then transforming back to the original space.

We use the wrapper *WrapperSubsetEval* [5] for multivariate FS methods and *ClassifierAttributeEval* [16] for univariate FS methods

C. Models and metrics

Wrapper methods are used in conjunction with the predictors *RandomForest* [17], *IBk* [18] and *LinearRegression* [19], and with the metrics *root mean squared error* (*RMSE*) and *mean absolute error* (*MAE*) [20] both in univariate and multivariate scenarios.

RandomForest is an *ensemble learning* method which constructs a forest of random trees with controlled variance, for classification or regression purposes. *IBk* is a simple instance-based learner that uses the class of the nearest k training instances for the class of the test instances and it is also valid for regressin. *LinearRegression* uses the *Akaike* criterion for model selection, and is able to deal with weighted instances.

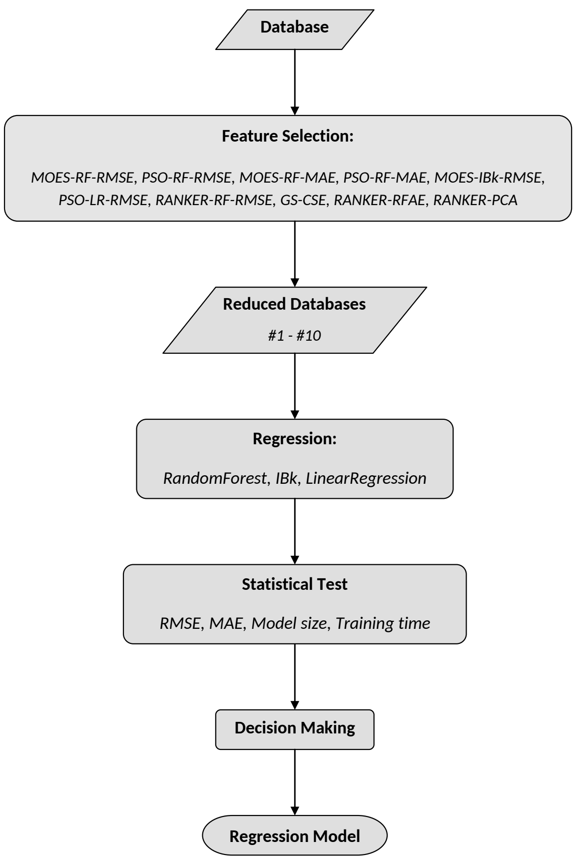

Fig. 2. Methodology for Feature Selection for regression.

We have followed the methodology shown in the Figure 2 to perform FS. We have systematically applied 10 different FS methods for regression shown in Table I and graphically in Figure 3. In Table I, *Database #Id* denotes the identifier of the reduced database generated with each FS method. Each FS method is the result of a specific choice among the search strategy, the evaluator, and the performance metric (for wrapper FS methods). We considered for this research seven wrapper FS methods and three filter FS methods. Among them, seven FS methods are multivariate and three FS methods are univariate. Table VIII shows the parameters used for each FS method.

IV. Use Case: Energy Efficiency

Feature Selection is part of the analytics process and there is an enormous quantity of use cases where is applied nowadays. Almost in any scenario where machine learning is used, feature selection is involved.

For this work, we are going to compare the results on energy consumption prediction scenarios using Weka feature selection approach and the ones that we have carried out in previous works in order to asses the suitability of the discussed Weka feature selection process.

Energy consumption in buildings is of special interest since they are one of the largest consumers of primary energy; for example in European Union countries, energy consumption in buildings represents around 40% of the total energy consumption. Attaining their efficiency is, therefore, an important goal since it can yield economical savings, reduce greenhouse gas emissions and alleviate energy poverty [21].

The reference building in which the proposed procedure has been carried out in order to select features for generating accurate building models is the Chemistry Faculty of the University of Murcia, which is a building used as a pilot for the H2020 ENTROPY project[1].

The data that is used in order to build and train our baseline corresponds to 1 year's worth of data, from February 2016 to February 2017.

A. Considered Features

Weather and the working hours have been proved to be useful in the task of energy consumption prediction [22].

However, due to the big amount of data that is available nowadays, the decision of which data source to use is reduced to the scientist subjective perception.

Not only there are three meteorological stations surrounding the Chemistry Faculty building, but also there exists many online services and API such as Weather Underground that might incorporate both more information but also more instability to the system.

B. Other Approaches

xxx

- PCA - por tramos (maana tarde y noche) [22]
- MIC - por horas y por semanas [23]
- Raw data - por das [21]

[1]http://entropy-project.eu/

Database #Id.	*FS method*	*Name*	*Search strategy*	*Evaluator*
#1	Wrapper Multivariate	*MOES-RF-RMSE*	*MultiObjectiveEvolutionarySearch*	*RandonForest (RMSE)*
#2	Wrapper Multivariate	*PSO-RF-RMSE*	*PSOSearch*	*RandonForest (RMSE)*
#3	Wrapper Multivariate	*MOES-RF-MAE*	*MultiObjectiveEvolutionarySearch*	*RandonForest (MAE)*
#4	Wrapper Multivariate	*PSO-RF-MAE*	*PSOSearch*	*RandonForest (MAE)*
#5	Wrapper Multivariate	*MOES-IBk-RMSE*	*MultiObjectiveEvolutionarySearch*	*IBk (RMSE)*
#6	Wrapper Multivariate	*PSO-LR-RMSE*	*PSOSearch*	*LinearRegression (RMSE)*
#7	Wrapper Univariate	*RANKER-RF-RMSE*	*Ranker*	*RandonForest (RMSE)*
#8	Filter Multivariate	*GS-CSE*	*GreedyStepwise*	*ConsistencySubsetEval*
#9	Filter Univariate	*RANKER-RFAE*	*Ranker*	*ReliefFAttributeEval*
#10	Filter Univariate	*RANKER-PCA*	*Ranker*	*PrincipalComponents*

TABLE I
PROPOSED FEATURE SELECTION METHODS FOR REGRESSION.

C. Weka Feature Selection Results

For this study, there are 51 features that have been studied. Descrbelas xxx: unas vienen de IMIDA otras del Weather Underground

Once FS is made, the next step is to perform regression with the reduced and original databases using *RandomForest*, *IBk* and *LinearRegression*. The Table II shows the parameters used for the regression methods. In order to detect over-fitting and prediction ability, the regression models have been evaluated in both full training set and 10-fold cross-validation over 10 iterations. Tables III and IV show the evaluation in full training set for the *RMSE* and *MAE* metrics respectively. The TablesV, VI and VII show the evaluation in 10-fold cross-validation (10 iterations) for the metrics *RMSE*, *MAE*, *Serialized_Model_Size* and *UserCPU_Time_testing*[2] respectively. The result of the experiment has been analysed through a *paired t-test (corrected)*, with 0.05 significance (being *#1#2* the test base in Tables **??**, and *RandomForest* in Tables VI and VII). For each result, a mark $*$ denotes that the result is statistically worse than the test base; similarly, a mark v denotes a statistically better result, and no mark denotes no statistically meaningful difference. Note that the reduced databases *#1* and *#2* are the same, as are the reduced databases *#3* and *#4*, so they appear together in all the tables.

Database #Id.	*RandomForest*	*IBk*	*LinearRegression*
#1#2	4.2267	0.0	66.8625
#3#4	4.2734	15.8069	66.8642
#5	4.6102	0.0	69.5321
#6	9.3867	0.0	54.7665
#7	5.8233	0.0	62.1653
#8	6.8531	0.0	56.583
#9	10.9122	7.0239	56.5407
#10	16.6868	0.0	58.8544
Original	10.0615	0.0	54.6545

TABLE III
RMSE WITH FULL TRAINING SET.

Database #Id.	*RandomForest*	*IBk*	*LinearRegression*
#1#2	2.2503	0.0	51.4108
#3#4	2.2696	4.7964	51.3844
#5	2.438	0.0	53.8089
#6	5.8176	0.0	39.4351
#7	3.273	0.0	44.8639
#8	4.005	0.0	40.8774
#9	5.9238	1.5123	41.52
#10	10.892	0.0	41.2654
Original	6.2188	0.0	39.3108

TABLE IV
MAE WITH FULL TRAINING SET.

Name	*Parameters*
RandomForest	-P 100 -I 500 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
IBk	weka.classifiers.lazy.IBk -K 1 -W 0
	-A "weka.core.neighboursearch.LinearNNSearch
	-A "weka.core.EuclideanDistance -R first-last""
LinearRegression	-S 0 -R 1.0E-8 -num-decimal-places 4

TABLE II
PARAMETERS OF THE REGRESSION METHODS.

V. ANALYSIS OF RESULTS AND DISCUSSION

Clearly the best results have been obtained with the FS methods *MOES-RF-RMSE / PSO-RF-RMSE* (*#1#2*) and *MOES-RF-MAE / PSO-RF-MAE* (*#3#4*), which show statistically significant differences with respect to the rest of the analysed FS methods for the *RMSE* and *MAE* performance metrics. The Table IX show the selected attributes with these FS methods. All attributes of *#3#4* also belong to *#1#2*, and the performances of the databases are similar in both full training set and 10-fold cross-validation (10 iterations). We can then conclude that the best selection of attributes is the solution *#3#4*.

The following general statements can be derived from the results:

- Wrapper FS methods have shown better performance for *RMSE* and *MAE* metrics than filter FS methods.
- Multivariate FS methods have shown better performance for *RMSE* and *MAE* metrics than univariate FS methods.
- For wrapper FS methods, *RandomForest* has proven more effective than *IBk* and *LinearRegression* based evaluators.
- Run time of *RandomForest* is acceptable for wrapper FS methods setting the number of iterations to 10 (-I 10), and the method is not very sensitive to the variation of its parameters. However, *RandomForest* generates large size regression models.
- *IBk* is very prone to over-fitting.
- *LinearRegression* is very fast and not prone to over-fitting, but it has not been efficient for this problem.

[2]Intel (R) Core (TM) i5-4460 @ 3.20 GHz 3.20 GHz RAM 8.00 GB Operating Systems 64 bits, processor x64.

	Model	*#1#2*	*#3#4*	*#5*	*#6*	*#7*	*#8*	*#9*	*#10*	*Original*
RMSE	*RandomForest*	11.00	11.02	12.23 *	26.04 *	16.13 *	18.82 *	24.02 *	46.02 *	27.59 *
	IBk	21.17	40.04 *	20.82 v	37.06 *	34.47 *	30.24 *	39.08 *	46.79 *	36.95 *
	LinearRegression	66.89	66.88	69.55 *	55.09 v	62.28 v	56.72 v	56.66 v	58.93 v	55.31 v
MAE	*RandomForest*	5.19	5.15	5.70 *	15.93 *	8.51 *	10.64 *	13.54 *	29.98 *	16.82 *
	IBk	10.19	19.91 *	9.99 v	17.31 *	16.17 *	13.74 *	20.07 *	21.21 *	16.69 *
	LinearRegression	51.46	51.44	53.87 *	39.71 v	44.97 v	40.99 v	41.62 v	41.35 v	39.81 v

TABLE V

RMSE AND *MAE* WITH 10-FOLD CROSS-VALIDATION (10 ITERATIONS).

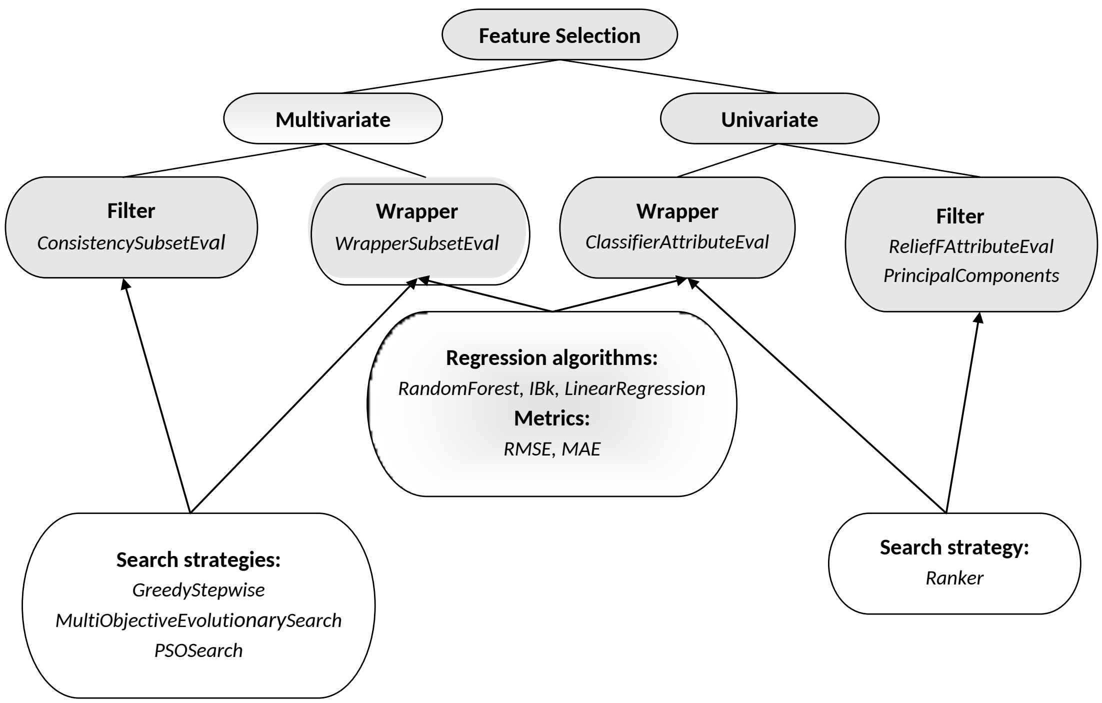

Fig. 3. Organization chart of the proposed Feature Selection methods for regression.

	RandomForest	*IBk*	*LinearRegression*
#1#2	77954766.98	513355.60 v	54118.52 v
#3#4	75704860.26	403181.80 v	53715.00 v
#5	79702492.83	319109.36 v	6339.56 v
#6	91655663.29	1321302.40 v	57163.96 v
#7	82474001.83	586730.40 v	54166.12 v
#8	83994700.42	759559.00 v	7235.44 v
#9	87917552.94	539231.84 v	6418.92 v
#10	103251437.55	539979.40 v	7176.08 v
Original	89395544.24	2128847.60 v	58977.72 v

TABLE VI

Serialized_Model_Size WITH 10-FOLD CROSS-VALIDATION (10 ITERATIONS).

	RandomForest	*IBk*	*LinearRegression*
#1#2	0.18	0.11 v	0.00 v
#3#4	0.19	0.09 v	0.00 v
#5	0.18	0.08 v	0.00 v
#6	0.26	0.23 v	0.00 v
#7	0.22	0.10 v	0.00 v
#8	0.23	0.13 v	0.00 v
#9	0.23	0.09 v	0.00 v
#10	0.23	0.10 v	0.00 v
Original	0.24	0.29 *	0.00 v

TABLE VII

UserCPU_Time_testing WITH 10-FOLD CROSS-VALIDATION (10 ITERATIONS).

ACKNOWLEDGMENT

REFERENCES

[1] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[2] A. G. Karegowda, A. S. Manjunath, M. A. Jayaram, Comparative study of attribute selection using gain ratio and correlation based feature selection, International Journal of Information Technology and Knowledge Management 2 (2) (2010) 271–277.

[3] M. A. Hall, Correlation-based feature selection for machine learning, Tech. rep., University of Waikato (1999).

[4] A. Ahmad, L. Dey, A feature selection technique for classificatory analysis, Pattern Recognition Letters 26 (1) (2005) 43–56.

[5] R. Kohavi, G. H. John, Wrappers for feature subset selection, Artificial intelligence 97 (1-2) (1997) 273–324.

[6] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, New York, NY, USA, 2011.

[7] S. Salzberg, C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993, Machine Learning 16 (3) (1994) 235–240. doi:10.1007/BF00993309.
URL `http://dx.doi.org/10.1007/BF00993309`

[8] I. H. Witten, E. Frank, M. A. Hall, Introduction to weka, in: I. H. Witten, E. Frank, M. A. Hall (Eds.), Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), third edition Edition, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann,

Database #Id.	*Parameters*
#1	-E “weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.trees.RandomForest -F 5 -T 0.01 -R 1 -E DEFAULT – -P 100 -I 10 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1” -S “weka.attributeSelection.MultiObjectiveEvolutionarySearch -generations 1000 -population-size 100 -seed 1 -a 0”
#2	-E “weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.trees.RandomForest -F 5 -T 0.01 -R 1 -E DEFAULT – -P 100 -I 10 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1” -S “weka.attributeSelection.PSOSearch -N 20 -I 1000 -T 0 -M 0.01 -A 0.33 -B 0.33 -C 0.34 -S 1”
#3	-E “weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.trees.RandomForest -F 5 -T 0.01 -R 1 -E MAE – -P 100 -I 10 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1” -S “weka.attributeSelection.MultiObjectiveEvolutionarySearch -generations 1000 -population-size 100 -seed 1 -a 0 ”
#4	-E “weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.trees.RandomForest -F 5 -T 0.01 -R 1 -E MAE – -P 100 -I 10 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1” -S “weka.attributeSelection.PSOSearch -N 20 -I 1000 -T 0 -M 0.01 -A 0.33 -B 0.33 -C 0.34 -S 1”
#5	-E “weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.lazy.IBk -F 5 -T 0.01 -R 1 -E DEFAULT – -K 1 -W 0 -A “weka.core.neighboursearch.LinearNNSearch -A “weka.core.EuclideanDistance -R first-last” -S “weka.attributeSelection.MultiObjectiveEvolutionarySearch -generations 10 -population-size 100 -seed 1 -a 0”
#6	-E “weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.functions.LinearRegression -F 5 -T 0.01 -R 1 -E DEFAULT – -S 0 -R 1.0E-8 -num-decimal-places 4” -S ”weka.attributeSelection.PSOSearch -N 100 -I 1000 -T 0 -M 0.01 -A 0.33 -B 0.33 -C 0.34 -R 1000 -S 1”
#7	-E “weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.trees.RandomForest -F 5 -T 0.01 -R 1 -E DEFAULT – -P 100 -I 10 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1” -S “weka.attributeSelection.Ranker -T -1.8E308 -N 10”
#8	-E “weka.attributeSelection.CfsSubsetEval -P 1 -E 1” -S “weka.attributeSelection.GreedyStepwise -T -1.8E308 -N -1 -num-slots 1”
#9	-E “weka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10” -S ”weka.attributeSelection.Ranker -T -1.8E308 -N 10”
#10	-E “weka.attributeSelection.PrincipalComponents -R 0.95 -A 5” -S “weka.attributeSelection.Ranker -T -1.8E308 -N 10”

TABLE VIII

PARAMETERS OF THE FEATURE SELECTION METHODS FOR REGRESSION.

Database #Id.	*Attributes*
#1#2	time, hour, stMO12_IMI_prec, stMU62_IMI_prec, month, day, dow, holiday
#3#4	time, hour, day, dow, holiday

TABLE IX

BEST SELECTED ATTRIBUTES.

Boston, 2011, pp. 403 – 406.

[9] J. Novakovic, Toward optimal feature selection using ranking methods and classification algorithms, Yugoslav Journal of Operations Research 21 (1).
URL http://yujor.fon.bg.ac.rs/index.php/yujor/article/view/364

[10] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 2nd Edition, Prentice-Hall, 2003.

[11] F. Jiménez, G. Sánchez, J. García, G. Sciavicco, L. Miralles, Multi-objective evolutionary feature selection for online sales forecasting, Neurocomputing.

[12] A. Moraglio, C. Di Chio, J. Togelius, R. Poli, Geometric particle swarm optimization, in: Proceedings of the 10th European Conference on Genetic Programming, 2007, pp. 125–136.

[13] H. Liu, R. Setiono, A probabilistic approach to feature selection - a filter solution, in: Proceedings of the 13th International Conference on Machine Learning (ICML), Vol. 96, 1996, pp. 319–327.

[14] K. Kira, L. A. Rendell, A practical approach to feature selection, in: Proceedings of the Ninth International Workshop on Machine Learning, ML92, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992, pp. 249–256.
URL http://dl.acm.org/citation.cfm?id=141975.142034

[15] H. Abdi, L. J. Williams, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics 2 (4) (2010) 433–459. doi:10.1002/wics.101.
URL http://dx.doi.org/10.1002/wics.101

[16] R. Schafer, Accurate and efficient general-purpose boilerplate detection for crawled web corpora, Language Resources and EvaluationOnline first. doi:10.1007/s10579-016-9359-2.
URL http://link.springer.com/article/10.1007/s10579-016-935

[17] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[18] D. W. Aha, D. Kibler, M. K. Albert, Instance-based learning algorithms, Mach. Learn. 6 (1) (1991) 37–66. doi:10.1023/A:1022689900470.
URL https://doi.org/10.1023/A:1022689900470

[19] X. Yan, Linear Regression Analysis: Theory and Computing, World Scientific Publishing Company Pte Limited, 2009.
URL https://books.google.es/books?id=MjNv6rGv8NIC

[20] C. J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Climate Research 30 (1) (2005) 79–82, cited By (since 1996)149.
URL http://www.scopus.com/inward/record.url?eid=2-s2.0-3044

[21] A. Gonzlez-Vidal, A. Ramallo-Gonzlez, F. Terroso-Senz, A. F. Skarmeta, Data driven modeling for energy consumption prediction in smart buildings, in: IEEE International Conference on Big Data, 2017.

[22] A. González-Vidal, V. Moreno-Cano, F. Terroso-Sáenz, A. F. Skarmeta, Towards energy efficiency smart buildings models based on intelligent data analytics, Procedia Computer Science 83 (2016) 994–999.

[23] T. Vantuch, A. Gonzlez-Vidal, A. Ramallo-Gonzlez, A. F. Skarmeta, S. Mik, Machine Learning Based Electric Load Forecasting for Short and Long-term Period, in: IEEE 4th World Forum on Internet of Things (WF-IoT), 2018.

First A. Author (M'76–SM'81–F'87) and all authors may include biographies. Biographies are often not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author's educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year the degree was earned. The author's major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author's last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (publisher name, year) similar to a reference. Current and previous research interests end the paragraph. The third paragraph begins with the author's title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, it should be of good quality, and professional-looking. Following are two examples of an author's biography.

Second B. Author was born in Greenwich Village, New York, NY, USA in 1977. He received the B.S. and M.S. degrees in aerospace engineering from the University of Virginia, Charlottesville, in 2001 and the Ph.D. degree in mechanical engineering from Drexel University, Philadelphia, PA, in 2008.

From 2001 to 2004, he was a Research Assistant with the Princeton Plasma Physics Laboratory. Since 2009, he has been an Assistant Professor with the Mechanical Engineering Department, Texas A&M University, College Station. He is the author of three books, more than 150 articles, and more than 70 inventions. His research interests include high-pressure and high-density nonthermal plasma discharge processes and applications, microscale plasma discharges, discharges in liquids, spectroscopic diagnostics, plasma propulsion, and innovation plasma applications. He is an Associate Editor of the journal *Earth, Moon, Planets*, and holds two patents.

Dr. Author was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, and the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011.

Third C. Author, Jr. (M'87) received the B.S. degree in mechanical engineering from National Chung Cheng University, Chiayi, Taiwan, in 2004 and the M.S. degree in mechanical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2006. He is currently pursuing the Ph.D. degree in mechanical engineering at Texas A&M University, College Station, TX, USA.

From 2008 to 2009, he was a Research Assistant with the Institute of Physics, Academia Sinica, Tapei, Taiwan. His research interest includes the development of surface processing and biological/medical treatment techniques using nonthermal atmospheric pressure plasmas, fundamental study of plasma sources, and fabrication of micro- or nanostructured surfaces.

Mr. Author's awards and honors include the Frew Fellowship (Australian Academy of Science), the I. I. Rabi Prize (APS), the European Frequency and Time Forum Award, the Carl Zeiss Research Award, the William F. Meggers Award and the Adolph Lomb Medal (OSA).