Pre-Processing For Neural Model Design In A Real Industrial Problem

Silvia Cateni, Valentina Colla, Alessandro Maddaloni, Antonella Vignal

Scuola Superiore Sant'Anna, TeCIP Institute, ICT-COISP Center, Pisa, Italy.

silvia.cateni{valentina.colla, alessandro.maddaloni, antonella.vignali}@santannapisa.it

***Abstract* - In the last years, the artificial neural networks have been effectively applied to several industrial problems in order to improve knowledge and get a deeper insight into correlations among different factors which affect production processes. In many applications Neural Networks are applied to predict the relationship between available input process variables and the target to be forecasted. Data pre-processing is an important step in developing a neural network application, which could affect the accuracy and the results of the developed models and applications. In the present paper an approach is proposed for data pre-processing, concerning a particular application related to the steel production. Such approach is tested on one row dataset coming from a real industrial context and the obtained results demonstrate the effectiveness of an accurate and appropriate pre-processing phase.**

Keywords - pre-processing; data mining; industrial data; variable selection; outlier detection

I. INTRODUCTION

In the last years, the quality of European steel has constantly enhanced, giving reliable steel products which are very competitive within world-wide markets. A consequence of this advance is a progressively increasing demand for customers, requesting for smaller and narrower tolerance intervals for significant quality attributes such as material properties. Online measurement systems of the material properties already exist but they are not located in a stage of process which allows providing their results back into the automation. Two popular online measuring systems for material properties, which are commercially available, are the Impulse Magnetic Process Online Controller (IMPOC) [1-3] and the Harmonic Analysis Coil Online Measurement System (HACOM).

The project entitled "Novel automatic model identification and online parameter adaptation for supporting the industrial deployment of model-based process control," (AutoAdapt), which is funded by the Research Fund for Coal and Steel (RFCS) of the European Union, aims at enhancing the diffusion of automation concepts controlling material properties online within Hot Strip Mills (HSM) through the enhancement of model identification procedures based on accurately physical models which can be appropriately adapted online. Additionally, an accurate measurement technology, which is directly capable to identify material properties at the HSM and also in subsequent process stages is necessary and the selection of identifiable models is central [4].

This paper presents part of the work which has been developed within the above mentioned project. In particular, an effective procedure is presented, which starts from a real raw dataset and outputs a prediction of a mechanical property of the final steel product by performing an adequate and necessary preprocessing phase which leads to the design of computationally efficient and accurate prediction models, that can be exploited within a control approach and that can be adapted by exploiting data, once it becomes available.

The importance of data pre-processing for the successful application of machine learning approaches has been widely highlighted in literature, such as, for instance, in [5-11].

The paper is organized as follows: Sec. II provides a short introduction to the considered industrial problem, Sec. III describes in detail the adopted pre-processing and design approach; the obtained results are then shown and discussed in Sec. IV and finally, in Sec. V, some concluding remarks are provided.

II. THE INDUSTRIAL APPLICATION

In the production of flat steel products, the Hot Rolling Mill (HRM) is the process which reduces the thickness of the steel slab after its reheating in the reheating furnace and provides a longer and flat semi-finished product called Hot Rolled Coil (HRC).

The HRM is composed of different sub-processes: the roughing mill, which performs the first gross thickness reduction, and the finishing mills, which provides the HRC with the desired final thickness. The HRC is finally cooled and coiled to be stored.

The HRC is further processed in the Cold Rolling Mill (CRM), which produces the Cold Rolled Coil (CRC), that is sold to manufacturers producing a wide variety of goods. The HRM is therefore an intermediate but fundamental stage for the determination of the final mechanical properties of the products.

A subsequent process called Hot Dip Galvanizing (HDG) is applied to the CRC where the strip is coated with zinc to prevent it from rusting. Also, the HDG can affect the final mechanical properties of the final product, since other intermediate and important sub-processes are performed,

such as strip reheating and cooling, respectively before and after the zinc pot and the following skinpass rolling process.

Every steel plant produces several kind of steels, that are grouped by grade, and each grade group holds its own characteristic chemical composition and target mechanical properties. Thus any kind of developed model needs to be flexible enough to be adapted to the needing of a specific steel produced as well as to the different types of products.

III. PRE-PROCESSING AND MODEL DESIGN

The main objective of this procedure is the prediction of an important mechanical property (target), which is measured by the IMPOC system, starting from process parameters.

Measurements of available dataset have been collected during Hot Dip Galvanizing (HDG) process and they are associated to longitudinal positions on the coil with a spatial interval of ten meters.

The analysis has been centred on three different steel grades called IF, HSLA1 and HSLA2 in order to test the proposed methods in several cases.

An overview of the procedures applied to the initial dataset in order to predict the target variable is schematically depicted in Figure 1

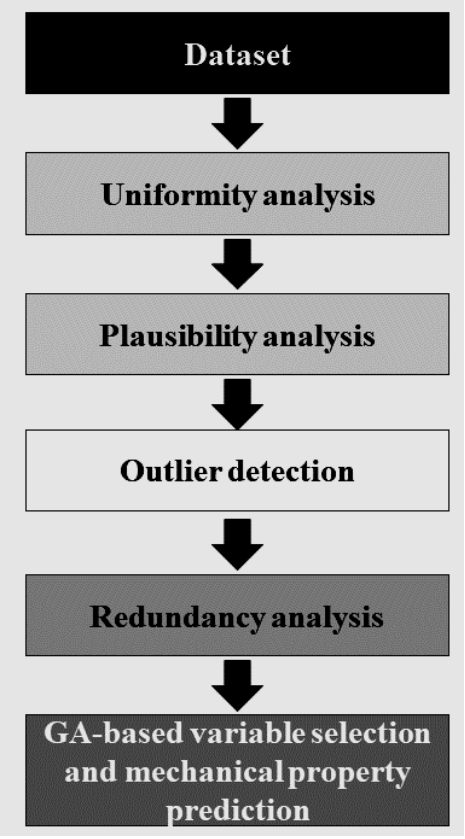

Figure 1. General procedure scheme.

A. Variable selection based on process knowledge.

Due to the large number of measured variables, there is the need to reduce such number by selecting those variables which mostly affect the considered target. A first variable selection stage has thus been performed based on the process knowledge of the technicians and plant managers.

B. Uniformity analysis.

A coil is defined as *uniform* if the chemical analysis of coils belonging to the same steel code can be considered almost constant. From a statistical point of view, this means that the value w_j of the content of each chemical element which is significant in the determination of the steel grade differs from its mean value η_j less than twice the related standard deviation σ_j, namely $\eta_j - 2\,\sigma_j < w_j < \eta_j + 2\,\sigma_j$. This definition of chemical analysis uniformity is statistically justified when the considered values of the content of chemical elements are approximately normally distributed. This statement has been verified through normal probability plot, a graphical technique used to validate the assumption of gaussianity of the distribution of a stochastic variable. This method compares data considering two measures: the original data and the theoretical normally distributed data. When the two measures agree, the normal probability plot appears to be linear, meaning that data distribution can be approximated by a normal distribution.

C. Plausibility analysis.

The plausibility analysis has been applied in order to create prediction models starting from process parameters lying within acceptable value ranges. To this purpose, a plausibility range or plausible values for each process parameter has been defined with the help of the technical personnel, in order to have plausibility check on process variable measurements during data pre-processing. Another check regards the possibility to have missing values on some process variable measures on some positions of the coils, due to measurement errors: all rows or coil positions where at least one value is missing or not plausible are removed and if more than 30% of a process variable samples are not plausible, that variable is eliminated.

D. Outlier Detection.

An outlier is defined as an observation that deviates from the rest of the available data [12, 13]. Outliers can be caused by measurement errors or by a particular unexpected condition or phenomenon affecting the considered process or system. Thus, depending on the application, an outlier can be a sample to be discarded or to be highlighted, as it represents a rare anomaly to be identified [14]. Outliers detection is an important step of data mining as well as it is a useful pre-processing stage in many applications belonging to different

contexts such as financial analysis, fraud detection, network robustness analysis, network intrusion detection [15, 16]. Traditional outlier detection approaches can be categorized into four main classes: distribution-based, distance-based, density-based and clustering-based. Distribution-based is a statistical method that considers an object as an outlier if it does not fit well with a standard distribution; the concept of distance-based approach was introduced by Knorr and Ng in 1998 [17] as *An object O in a dataset T is a DB(p,D)-outlier if at least fraction p of the objects in T lie greater than distance D from O*; density based-approach identifies an outlier on the base on how isolated the data is with respect to the surrounding neighbourhood; finally the clustering-based approach classifies as outlier an object which does not belong to any cluster after a clustering operation: moreover if a cluster is quite different to other clusters the data that falls in it is considered as an outliers [14, 15]. The approach which is proposed here combines these different approaches through a Fuzzy Inference System (FIS), similarly to what proposed in [18-21]. The four features are fed in input to the FIS that outputs an outlierness degree belonging to the range [0, 1]. The FIS is a Mamdami type and three fuzzy sets have been stated for each feature (labeled low, medium and high) and Gaussian-shaped membership functions have been used with the exclusion of distribution function that is a binary variable, thus two fuzzy singleton have been adopted. The FIS output is a variable in the range [0, 1] that if it is close to one, the corresponding pattern is labeled as outlier. In this industrial context outlier detection has been applied on each coil independently from other coils, being some process settings depending on coil basic features.

E. Redundancy analysis.

Redundant variables are recognized by using the *dominating set algorithm*, which derives from graph theory [22, 23]. The proposed algorithm creates a graph, where every variable defines a vertex and two vertices are connected by an edge if the related variables have a linear correlation [24] greater than 0.95 and a p-value (correlation significance) lower than 0.05. The adopted procedure detects the minimal dominating set of the graph, namely the minimum vertex subset *D* such that all the remaining vertices have a connection with at least an element of *D*. Redundant variables are associated with those vertices of the graph that are not in the minimal dominating set [25].

F. Genetic Algorithm (GA)-based Variable Selection and model design.

Variable reduction is a substantial preliminary step of the development of Artificial Neural Network (ANN)-based models [26, 27]. An inappropriate selection of the input variables can deteriorate the accuracy of ANN in the training phase [5, 28]. Commonly we think that ANNs are able to identify redundant and noisy variables during the training stage and this credence leads the designers to use all available variables believing to include more information and also improve the performance of the developed model. Nevertheless, the number of input parameters which are included in the ANN influences the computational burden and the training time [29]. Additionally, another significant aspect, which is largely dealt in literature, is the *curse of dimensionality*, which states that if the dimensionality of a model increase linearly, the total volume of the modelling process domain increase exponentially [30]. The variable selection methods can be classified in three main groups: filters, wrappers and embedded approaches. **Filter approaches** can be considered as a pre-processing phase, as they are independent on the learning algorithm [31, 32]. The variables subsets are created by considering the relationship between input and output. The main advantage of the filters is their simplicity and speed and also they are suitable for treating with large datasets [33-35]. **Wrapper approaches**, introduced by Kohavi et al. in [36], estimate the performance of the developed learning machine in order to choice a subset of variables considering their predictive power. Wrappers adopt the learning algorithm as a black box making these approaches remarkably universal. The most obvious wrapper approach is the *exhaustive search*, also called *brute force method*, which analyses all combinations of input variables. A common wrapper approach is the *Greedy Search* strategy, which gradually creates the variables subset by adding or removing single variables from a primary set. Finally, other variable selection procedures based on genetic algorithms are proposed in literature in order to obtain a good performance in a reasonable time and one of them are proposed in this work. **Embedded methods** execute the variable selection as portion of the learning phase and are naturally specific of a detailed learning machine [37, 38]. Classic examples of embedded methods include classification trees, random forests [39]. The main advantage of embedded methods is their inclusion in the learning algorithm which considers the relevance of the variables. In recent times some hybrid variable selection methods have been proposed in order to exploit their advantages by overcoming their shortcomings [25, 40-42]. Finally, a significant phase to be considered treating with variable selection is the stability of the result [43-45]. The stability of a variable selection approach is commonly defined as the robustness of the variables selected associated to variations in the training sets produced from the same creating distribution [40, 46-48].

In this work, once redundant variables are discarded from dataset, a GA-based variable selection procedure is implemented [49, 50]. The chromosomes are characterized by binary vectors, whose length is equal to the number of input variables, so that each gene is univocally associated to a variable and a unitary value means that the related variable is included in the variable subset associated to the considered chromosome. The GA population is randomly initialized and genetic operators (mutation and crossover) are computed at each generation. In particular, the mutation operator switches

a randomly selected gene, while the crossover operator creates each gene of the son chromosome by randomly selecting genes from the two parent chromosomes. Before applying the variable section procedure, the dataset is split in two subsets: 75% of the data is attributed to the training set, while the remaining 25% of data is used as validation set. The fitness function to minimize is the mean relative prediction error of the model exploiting each variable subset as input and is evaluated on the validation set [51]. The GA terminates when a fixed maximum number of iterations has been reached or a plateau of the fitness function is reached. The winner chromosome provides the subset of input variables related to the best value of the fitness as well as the trained model which exploits that subset. In effect, this variable selection approach belongs to the category of wrapper methods [52, 53], which exploits the learning machine inside the variable selection procedure in order to evaluate the goodness of each variable subset.

In the present case, several models have been created according to coil properties such as steel code and thickness, which are factors which heavily affect the predicted mechanical properties. Coils have been divided in several classes on the basis of the steel code and each class has been divided again in 10 sub-classes corresponding to different thickness ranges, each of width 0.2 mm.

For each sub-model three prediction models have been applied: a polynomial approximation and two 3-layers feed-forward networks ($NN3_1$ and $NN3_2$). The number of neurons in the hidden layers of the NN-based models is empirically determined by means of several tests performed.

Furthermore, according to the specifications provided by skilled personnel of the steelworks, two variables have always been included in the final dataset, due to the need of considering the influence on the mechanical properties prediction. Such variables are not included in the variable selection procedure.

IV. Experimental results

Performance results have been calculated considering the mechanical property named *Tensile Strength* as target. The measurement unit of such variable is MegaPascal (MPa).

The preliminary knowledge-based variable selection stage selected 38 potential input variables, 37 of them belonging to four categories (Temperatures, fan cooler, skinpass, tension leveler) while one variable is the mechanical property to be predicted.

By applying the plausibility analysis and, afterwards, the above mentioned outlier detection approach to the available dataset no variable was discarded but about 30% of coils has been removed from the dataset.

The results obtained by applying the plausibility analysis are shown in Table 1 that are expressed in terms of percentage of measures that are outside the plausibility ranges indicated by the expert personnel.

TABLE I. Results of the plausibility analysis

Steel class	No of coils	Input samples	Not plausible %
IF	2911	668150	29
HSLA1	1101	214411	28.53
HSLA2	1120	202229	35.53

After the plausibility analysis, the outlier detection technique has been applied and results, in terms of percentage of detected outliers, are shown in Table II.

TABLE II. Results of the outlier detection

Steel class	Percentage of outliers
IF	0.67 %
HSLA1	0.75 %
HSLA2	0.85 %

The results of the preliminary uniformity analysis showed that the 72.8%, 77% and 78.6% of the data related to the chemical content of all coils, respectively, for the IF, HSLA1 and HSLA2 steel grades can be considered uniform.

Starting from reliable data, the redundant variable are removed and finally the above-described GA-based algorithm have been applied in order to select the input variables which mainly affect the considered target. The results of redundancy and selection of variables are shown in Table III depicting how many input variables are selected for each steel code and thickness range. The obtained results are considered satisfactory, as they are in line with the expectation of the technical personnel of the plant.

TABLE III. Number of redundant and selected variables for each steel code and thickness range

Thickness range	IF		HSLA1		HSLA2	
	#Red. Var.	#Sel. Var.	#Red. Var.	#Sel. Var.	#Red. Var.	#Sel. Var.
1	6	21	-	-	-	-
2	6	19	5	20	6	15
3	9	13	8	20	7	10
4	7	15	7	13	4	13
5	5	16	7	15	-	-
6	7	18	6	16	6	12
7	8	16	5	19	6	10
8	5	13	6	20	6	19
9	-	-	-	-	4	12
10	4	11	5	13	6	12

In order to demonstrate the effectiveness of the use of an accurate and appropriate phase of data pre-processing, a comparison was made between the performance of two neural networks, the first trained with the raw data set and

the second trained with a clean data set, namely the dataset, related to same period, cleaned through the above mentioned pre-processing operations.

The results achieved by training the network with the raw dataset and with the "clean" data set are shown in Figure 2, Table IV and Table V in terms of average error ε and standard deviation σ as well as average relative error ε_r and standard deviation σ_r.

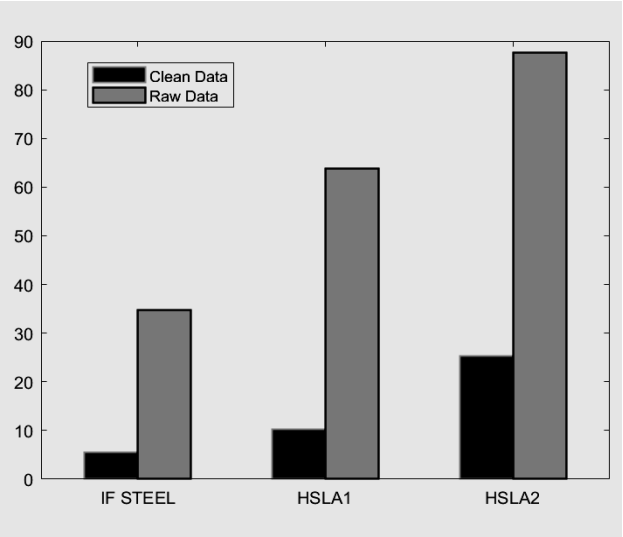

Figure 2. Absolute Error obtained by exploiting clean and raw data.

TABLE IV. RESULTS OBTAINED BY TRAINING THE ANN WITH THE RAW DATASET

Steel class	ε	σ	ε_r	σ_r
IF	34.78	12.11	0.118	0.041
HSLA1	63.75	52.18	0.163	0.134
HSLA2	87.72	111.16	0.193	0.248

TABLE V. RESULTS OBTAINED BY TRAINING THE ANN WITH THE "CLEAN" DATASET

Steel class	ε	σ	ε_r	σ_r
IF	5.51	4.81	0.019	0.016
HSLA1	10.25	9.73	0.026	0.027
HSLA2	23.38	16.74	0.056	0.0441

This study shows the importance and the effectiveness of data preprocessing neural networks modelers. In effects, the performance achieved through the preprocessing of data improves. in average, of a factor higher than 20% with respect to the one achieved by considering the raw data and this results is a very satisfactory. Moreover, by selecting specific variables, the knowledge about the phenomena studied is enhanced, as those parameters that affect the target are highlighted.

V. CONCLUSIONS

The proposed pre-processing approach provides a more informative data subset in a reasonable time and it can be applied to all kind of real datasets. The aim of the proposed approach was to demonstrate the need to perform an adequate pre-processing of industrial data before creating a model. The obtained results show the importance of an effective and accurate pre-processing stage for the design and development of neural network-based models. This consideration can be extended to any machine learning approach, whose results strongly depend on the quality of the exploited data. Future work will involve testing the proposed approach on other industrial cases study.

ACKNOWLEDGMENT

The work described in the present paper has been developed within the projects entitled ``Novel automatic model identification and online parameter adaptation for supporting the industrial deployment of model-based process control," (Ref. AutoAdapt, Contract No. RFSR-CT-2015-00030) that has received funding from the Research Fund for Coal and Steel of the European Union, which is gratefully acknowledged. The sole responsibility of the issues treated in the present paper lies with the authors; the Commission is not responsible for any use that may be made of the information contained therein.

REFERENCES

[1] Scheppe, D.: Impoc - increasing production yield by online measuremenent of material properties. SEAISI Quarterly (South East Asia Iron and Steel Institute) 83(3), 2009, pp.41–46

[2] Van-Den-Berg, F., Kok, P., Yang, H., Aarnts, M., Vink, J.J., Beugeling,W., Meilland, P., Kebe, T., Stolzemberg, M., Krix, D., Peyton, A., Zhu, W., Martinez-De-Guerenu, A., Gutierrez, I., Jorge-Badiola, D., Gurruchaga, K., Lundin, P., Volker, A., Mota, M., Monster, J., Wirdelius, H., Mocci, C., Nastasi, G., Colla, V., Davis, C., Zhou, L., Schmidt, R., Labbe, S., Reboud, C., Skarlatos, A., Svaton, T., Leconte, V., Lombard, P.: In-line characterisation of microstructure and mechanical properties in the manufacturing of steel strip for the purpose of product uniformity control. In: Proc.19th World Conference on Non- Destructive Testing WCNDT 2016 (June 2016)

[3] Van-Den-Berg, F., Kok, P., Yang, H., Aarnts, M.P., Meilland, P., Kebe, T., Stolzemberg, M., Krix, D., Zhou, W., Peyton, A., Martinez-De-Guerenu, A., Gutierrez, I., Jorge-Badiola, D., Mamstrom, M., Volker, A., Duijster, A., Bostrom, H.W.A., Mocci, C., Vannucci, M., Colla, V., Davis, C., Zhou, L., Schmidt, R., Labbe, S., Reboud, C., Skarlatos, A., Leconte, V., Lombard, P.: Product uniformity control - a research collaboration of european steel industries to nondestructive evaluation of microstructure and mechanical properties. In: 22nd International Workshop on Electromagnetic Nondestructive Evaluation, ENDE 2017. Studies in Applied Electromagnetics and Mechanics, vol. 43, 2018, pp. 120–12.

[4] Colla, V., Bioli, G., Vannucci, M.: Model parameters optimisation for an industrial application: A comparison between traditional approaches and genetic algorithms. In: Proceedings - EMS 2008, European Modelling Symposium, 2nd UKSim European Symposium, on Computer Modelling and Simulation. 2008, pp. 34–39.

[5] Cateni, S., Colla, V.: The importance of variable selection for neural networks based classification in an industrial context. International Workshop on Neural Networks, WIRN 2015. Smart Innovation, Systems and Technologies, vol. 54, 2015, pp. 363–370.
[6] Cateni, S., Colla, V.: Variable selection for efficient design of machine learning-based models: Efficient approaches for industrial applications. Communications in Computer and Information Science 629, 2016, pp. 352–366.
[7] Vannucci, M., Colla, V., Cateni, S.: An hybrid ensemble method based on data clustering and weak learners reliabilities estimated through neural networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9095, 2015, pp. 400–411.
[8] Cateni, S., Colla, V., Vannucci, M., Vannocci, M.: A procedure for building reduced reliable training datasets from realworld data. Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2014. pp. 393–399.
[9] Bellman, R.: Adaptive control processes: a guided tour. Princeton University Press, 1961.
[10] Sgarbi, M., Colla, V., Cateni, S., Higson, S.: Pre-processing of data coming from a laser-emat system for non-destructive testing of steel slabs. ISA Transactions Vol. 51(1), 2012, pp. 181–188.
[11] Vannucci, M., Colla, V., Sgarbi, M., Toscanelli, O.: Thresholded neural networks for sensitive industrial classification tasks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5517 LNCS(PART 1), 2009, pp. 1320–1327.
[12] Barnett, V., Lewis, T.: Outliers in Statistical Data. 3rd ed., John Wiley & Sons, 1984.
[13] Shetty, M., N.M.Shekokar: Data mining techniques for real time intrusion detection systems. International Journal of Scientific and Engineering Research, Vol. 3(4), 2012.
[14] Patcha, A. Park, J-M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends, computer networks. International Journal of Computer and Telecommunications Networking 51 (12), 2017, pp. 3448–3470.
[15] Aggarwal, C., Yu, P.: Outlier detection for high dimensional data. Proceeding of ACM SIGMOD Conference 1, 2001, pp. 37–47.
[16] Koc, L., Carswell, A.D.: Network intrusion detection using a hnb binary classifier. Proceedings - UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim 2015, pp. 81-85.
[17] Knorr, E.M., Ng, R.: Algorithms for mining distance-based outliers in large datasets. Proceeding VLDB 1, 1998, pp. 392–403.
[18] Cateni, S., Colla, V., Vannucci, M.: A fuzzy logic-based method for outliers detection. In: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2007. pp. 561–566.
[19] Cateni, S., Colla, V., Nastasi, G.: A multivariate fuzzy system applied for outliers detection. Journal of Intelligent and Fuzzy Systems 24(4), 2013, pp. 889–903.
[20] Cateni, S., Colla, V., Vannucci, M.: A fuzzy system for combining different outliers detection methods. Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2009. pp. 87–93.
[21] Cateni, S., Ritacco, A., Iannino, V., Colla, V., Vannucci, M., Dettori, S.: Smart data pre-processing modules and graphical user interfaces for machine learning tasks. International Journal of Simulation: Systems, Science and Technology Vol. 19(5), 2018, pp. 24.1–24.7.
[22] N. Biggs, E. Lloyd, E.R.W.: Graph Theory, vol. Oxford University Press. 1986.
[23] Bondy, J.A., Murty, U.: Graph Theory , Springer, 2008.
[24] Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation basedfilter solution. Proc. of the 20th International Conference on Machine Learning ICML 1, 2003, pp. 856–863.
[25] Cateni, S., Colla, V.: A hybrid variable selection approach for nn-based classification in industrial context. Smart Innovation, Systems and Technologies. Vol. 69, 2017, pp. 173–180.
[26] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Machine Learning Vol. 3, 2003, pp. 1157–1182.
[27] Liu, H., Motoba, H., Setiono, R., Zhao, Z.: Feature selection: an ever evolving frontier in data mining. JMLR: workshop and Conference proceedings 10. The 4th workshop on Feature Selection in Data Mining, 2010, pp. 4-13.
[28] Fausett, L.: Foundamentals of Neural Networks. Prentice Hall 1994.
[29] May, R., Dandy, G., Maier, H.: Review of Input Variable Selection Methods for Artificial Neural Networks. Artificial Neural Networks Methodological Advances and Biomedical Applications, 2011.
[30] Haykin, S.: Neural Networks: a Comprehensive Foundation. MacMillman Publishing, 1994.
[31] Mitchell, T., Toby, J., Beauchamp, J.: Bayesian variable selection in linear regression. J. of the Amer. Statistical Ass. Vol. 83, 1988, pp. 1023–1032.
[32] Sun, Y., Robinson, M., Adams, R., Boekhorst, R., Rust, A.G., Davey, N.: Using feature selection filtering methods for binding site predictions. Proc. 5th IEEE International Conference on Cognitive Informatics ICCI ’06, 2006, pp. 566-571.
[33] He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. Advances in Neural Information Processing Systems, 2005, pp. 507–514.
[34] Cateni, S., Colla, V., Vannucci, M.: A fuzzy system for combining filter features selection methods. International Journal of Fuzzy Systems 19(4), 2016, pp. 1168–1180.
[35] Zhang, S., Zhao, Z.: Feature selection filtering methods for emotion recognition in chinese speech signal. 9th International Conference on Signal Processing, ICSP 2008, pp. 1699-1702.
[36] Kohavi, R., John, G.: Wrappers for feature selection. Artificial Intelligence, Vol. 97, 1997, pp. 273–324.
[37] Bo, L., Wang, L., Jiao, L.: Multilayer perceptrons with embedded feature selection with application in cancer classification. Chinese Journal of Electronics Vol. 15, 2006, pp. 832–835.
[38] Breiman, L.: Random forests. Machine learning Vol. 45, 2001, pp. 5–32.
[39] Breiman, L., Friedman, J.H., Olshen, R.A., Stone., C.J.: Classification and Regression Trees. Wadsworth and Brooks, 1984.
[40] Cateni, S., Colla, V., Vannucci, M.: A hybrid feature selection, method for classification purposes. Proceedings - UKSim-AMSS 8th European Modelling Symposium on Computer Modelling and Simulation, EMS 2014, pp. 39–44.
[41] Lee, K.: Combining multiple feature selection methods. Ph.D. thesis, The Mid-Atlantic Student Workshop on Programming Languages and Systems Pace University 2002.
[42] Sebban, M., Nock, R.: A hybrid filter/wrapper approach of feature selection using information theory. Pattern Recognition, Vol. 35, 2002, pp. 835–846.
[43] Kuncheva, L.I.: A stability index for feature selection. IASTED International conference on Artificial intelligence and Application AIA 2007 pp. 95–116.
[44] Loscalzo, S., Yu, L., Ding, C.: Consensus group stable feature selection. Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. ACM 2009, Vol. 1, pp. 567–575.
[45] Novovicova, J., Somol, P., Pudil, P.: A new measure of feature selection algorithms stability. IEEE International Conference on Data Mining Workshops, 2009, pp. 382–387.
[46] Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms. In: Proc. 5th IEEE International Conference on Data Mining (ICDM’05), 2005, pp. 218–225.
[47] Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Information Systems Vol. 12, 2007, pp. 95–116.
[48] Cateni, S., Colla, V.: Improving the stability of sequential forward and backward variables selection. 15-th International Conference on Intelligent Systems design and applications ISDA, 2016, pp. 374-379.
[49] Cateni, S., Colla, V., Vannucci, M.: General purpose input variable extraction: A genetic algorithm based procedure give a gap. 9th International Conference on Intelligence Systems design and Applications ISDA’09, 2009, pp. 1307–1311.
[50] Cateni, S., Colla, V., Vannucci, M.: Variable selection through genetic algorithms for classification purpose. In: IASTED

International Conference on Artificial Intelligence and Applications AIA 2010, pp. 6–11.

[51] Jain, A., Chandrasekaran, B.: Dimensionality and sample size considerations in pattern recognition practice. Handbook of statistics 1 North-Holland, 1982, pp. 835–855.

[52] Cateni, S., Colla, V., Vannucci, M.: A genetic algorithm based approach for selecting input variables and setting relevant network parameters of som based classifier. International Journal of Simulation Systems Science and Technology, Vol. 12(2), 2011, pp. 30–37.

[53] Cateni, S., Colla, V.: Improving the stability of wrapper variable selection, applied to binary classification. International Journal of Computer Information Systems and Industrial Management Applications. Vol. 8, 2016, pp. 214–225.