Proceedings of the Federated Conference on Computer Science and Information Systems pp. 225–229
DOI: 10.15439/2016F90
ACSIS, Vol. 8. ISSN 2300-5963

Massively Parallel Feature Extraction Framework Application in Predicting Dangerous Seismic Events

Marek Grzegorowski
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw,
Banacha 2, 02-097 Warsaw, Poland, Email: M.Grzegorowski@mimuw.edu.pl

***Abstract*—In this paper we introduce an automated mechanism for knowledge discovery from data streams. As a part of this work, we also present a new approach to the creation of classifiers ensemble based on a wide variety of models. Furthermore, we describe an innovative, highly scalable feature extraction and selection framework designed to work with the MapReduce programming model and the application of designed framework to build an ensemble of classifiers which takes into account both the quality and the diversity of individual models. The effectiveness of the solution has been verified through a participation in an open data mining competition which concerned the problem of predicting periods of increased seismic activity causing life-threatening accidents in coal mines. The submitted solution obtained the highest AUC score of all the solutions uploaded by 106 participating research teams.**

I. Introduction

FOR SOME time, we can observe a massive shift in technology that makes sensors are more and more available and common. On the other hand, we can notice a significant grow in popularity of stream analytics as well as a decline in prices of data storage. One of the main beneficiaries of the aforementioned changes are monitoring, threats detecting and decision supporting systems. An exemplary application of which could be active monitoring of coal extraction [21] to provide protection for people underground or supporting fire commanders in decision making [14].

Nowadays, an increasing number of business technology objectives is related with comprehensive data analytics. Meanwhile, many researchers recognize feature engineering [8] as a major step in the process of knowledge discovery, necessary to obtain good results of the analysis. In this paper we propose an innovative approach to data analysis that, in order to provide high quality assessment, implies creation of an ensemble [3] of classifiers using a wide variety of models based on various subsets of attributes, in this way, resulting not only in enhancement of the quality of indications, but also minimizing the impact of concept drift on the final evaluation of results.

The continuous collection and analysis of multiple reading streams from a large network of sensors located underground raises a problem of long lasting and usually very complex preparation of data. Therefore, in order to simplify and speed up the feature extraction we introduced a novel framework designed to process streams of numerical readings from multiple sensors. The developed framework is ready to operate in production environments and, hence, is tailored for incremental processing of the emerging data based on the sliding window technique and the concept of parallelization presented in [5].

In the following sections we present the extension of our former solution [6], [7] with additional features and describe modification of the architecture allowing the work both with incremental data as well as with highly-scalable batch computations via MapReduce [2] programming model. The assessment of the solution was carried out on the basis of real life problems related to the streaming data [22].

The effectiveness of the framework has been confirmed in the analysis of several significantly different problems within the data analysis competitions. The first, concerned the recognition of the activity and posture of firefighter based on readings from multiple motion and vital sensors as a part of AAIA'15 Data Mining Competition: Tagging Firefighter Activities at the Fire Scene[17]. The second concerned the prediction of dangerous concentration of methane in the atmosphere of the mine sidewalks as a part of IJCRS'15 Data Challenge: Mining Data from Coal Mines[10]. The third concerned the prediction of increased seismic activity in mines as a part of AAIA'16 Data Mining Challenge: Predicting Dangerous Seismic Events in Active Coal Mines [9]. In all competitions, the results were very promising. It is worth noting that in the case of the seismic activity analysis, the solution based on the elaborated framework received the highest score in terms of Area Under ROC Curve (AUC) measure, equal to 93.96%, while in the methane concentration level analysis it achieved the second highest result with AUC equal to 94.73%.

The paper is organized as follows. In Section II we present the description of the data challenge problem. In Section III and IV we provide detailed information about the elaborated feature engineering framework, including insights of feature extraction and selection. Next, in Section V, we describe the conduct of the experiments and resulted ensemble model. Finally, in Section VI we summarize the work.

II. AAIA'16 Data Challenge Problem Description

Providing safety of miners working underground is the fundamental requirement for the coal mining industry in Poland. Coal mining companies are obligated by the law to introduce many safety measures to secure proper working conditions of their underground personnel. The task in the competition was to devise a reliable prediction model for detecting periods of increased seismic activity that could endanger miners.

More precisely, the tasks of the data challenge was to predict likelihood of the 'warning' label for the records from the test

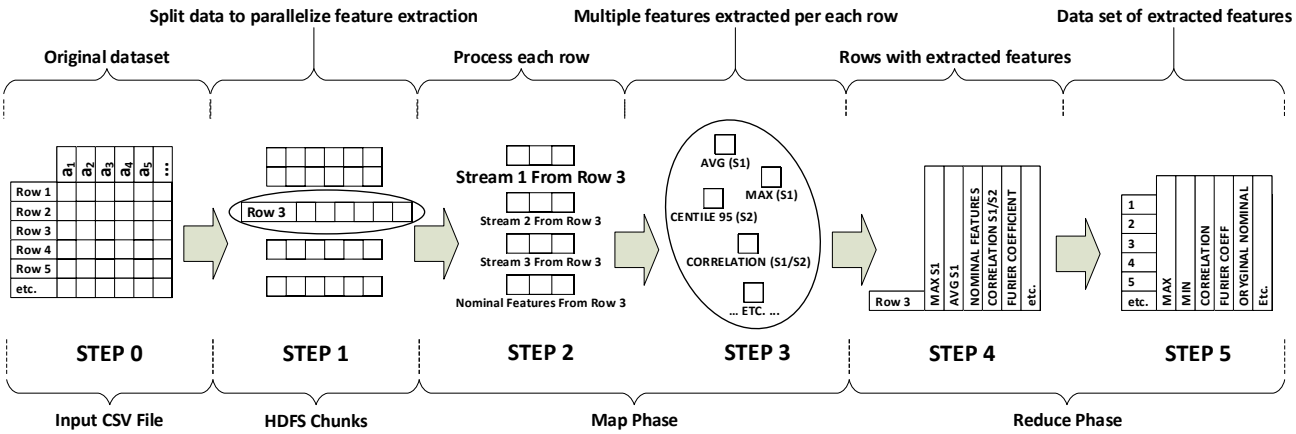

Figure 1. This diagram presents the high level overview of the feature extraction process broken down into individual steps. The curly braces at the top of the diagram indicate goals achieved in each processing step. The curly braces at the bottom of the diagram show how the individual processing steps were implemented. The 'original dataset' in STEP 0 corresponds to dataset provided by organizers of AAIA'16 Data Mining Challenge: Predicting Dangerous Seismic Events in Active Coal Mines. Features $a_1, a_2, a_3, ..$ correspond to attributes presented in Table I. STEP 1 was designed to partition of the original dataset into individual rows in order to parallelize calculations - this step was implemented by internal mechanisms of the MapReduce framework. STEP 2 was to split each row into nominal and stream data.In STEP 3 the feature extraction framework was applied to each data stream (e.g. time series containing 24 values corresponding to the average energy of the most active geophone - see Table I) and all features described in Tables II and III were created. In STEPs 4 and 5 all features, the nominal attributes as well as attributes newly extracted from streams, were put together.

set. The real number corresponding to the predicted likelihood was not expected to be in any particular range, however, higher numerical value should have indicated a higher chance of the 'warning' label. The final assessment of solutions were calculated using the Area Under the ROC Curve (AUC).

The data set, which was provided by Research and Development Centre EMAG, consisted of hourly aggregated readings from seismic sensors that count the number of seismic bumps perceived at longwalls and measure their total energy. Data records were composed of 24 consecutive hours of such readings coupled with the most recent assessments of the conditions at longwalls made by mining experts. All the attributes of the data set are described in Table I. The data sets were well prepared, cleaned of malformed and erroneous values, without missing attributes.

Table I
THE TABLE PRESENTS ATTRIBUTES OF THE MAIN DATA FILES.

no.	feature description
1	id of the main working site where the measurements were taken
2	total energy of: seismic bumps, major seismic bumps, destressing blasts and all types of bumps registered in the last 24h
3	latest progress in the mining from, both, left and right side
4	latest seismic, comprehensive and seismoacoustic (standard and alternative method) hazard assessments made by experts (a/b/c/d): a - no hazard; b - moderate hazard; c - high hazard; d - dangerous
5	maximum yield from the last meter of the small-diameter drilling
6	depth at which the maximum yield was registered
7	five time series containing 24 values (one per hour 1..24) each corresponding to a number of seismic bumps with energy in the following ranges: $(0, 10^2]$, $(10^2, 10^3]$, $(10^3, 10^4]$, $(10^4, 10^5]$ and $(10^5, Inf)$ aggregated per hour (1..24)
8	five time series containing 24 values (one per hour 1..24) each corresponding to sum of energy of registered seismic bumps with energy in the following ranges: $(0, 10^2]$, $(10^2, 10^3]$, $(10^3, 10^4]$, $(10^4, 10^5]$ and $(10^5, Inf)$ aggregated per hour (1..24)
9	four time series, each containing 24 values (one per hour 1..24) corresponding to the number of: seismic bumps, rock bursts, destressing blasts and to energy of the strongest seismic bump
10	four time series, each containing 24 values (one per hour 1..24) corresponding to maximum activity, maximum energy, average activity and average energy of the most active geophone
11	four time series, each containing 24 values (one per hour 1..24) ccorresponding to the maximum difference and average difference in, both, activity and energy registered by the most active geophone

In total, the training file contained 133150 records provided in a tabular format with 541 columns. The label indicates whether a total seismic energy perceived during 8 hours after the period covered by a data record exceeded the warning threshold of $5 * 10^4$ Joules. There were 2963 examples labeled as 'warning' and 130187 'normal' cases in the training set.

III. FEATURE EXTRACTION

In the course of feature extraction we generated a large number [26] of potentially relevant characteristics [16], [23] and applied the feature selection in the next step. The feature extraction was based on the sliding window [24] method and was configured to accept on its input a data set containing readings from multiple streams [5]. According to the submitted configuration each stream, stored in a row of the csv file (training and test set), was divided into three non-overlapping frames. During the process of moving a sliding window through the time series a number of aggregating functions were applied. The Table II presents features extracted form a single time series.

The statistics indicated in Table II were supplemented by Kendall's correlation between each pair of data streams for every row in csv. Furthermore, because there were more than one window generated for each time series we extracted inter-window statistics, that is, a set of values that express

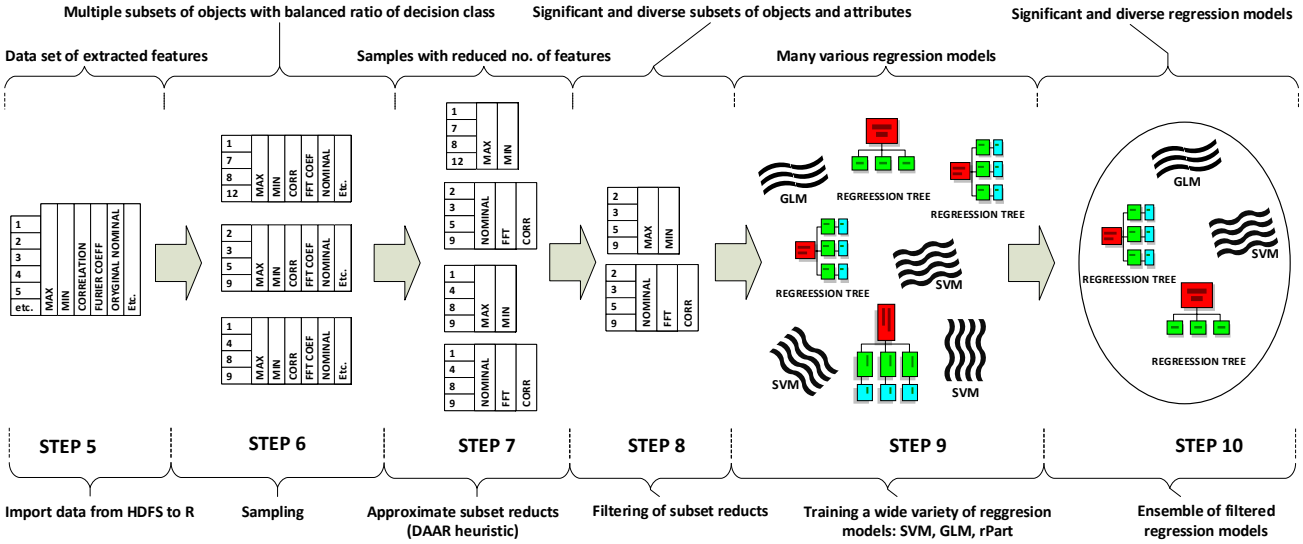

Figure 2. This diagram presents the course of feature selection, regression models training and construction of the final ensemble of regression models broken down into individual steps. The curly braces at the top of the diagram indicate goals achieved in processing steps. The curly braces at the bottom of the diagram show how each processing step was implemented. All the processing steps from 6 to 10 were implemented in R environment for statistical computing. Consecutive processing steps were designed to: STEP 6 - was to drew a number of random samples of objects with balanced decision classes. In STEP 7 the reduced attribute subsets were calculated - this step was implemented basing on the concept of approximate decision reducts derived from the theory of rough sets. In STEP 8, a number of attribute subsets were obtained. Each subset was derived by merging 2 or 3 reducts, however, only significantly different subsets were maintained for the purpose of model training in the following phase of processing. In STEP 9 previously obtained subsets of objects and attributes were used to train regression models based on selected algorithms: rPart, SVM, GLM. Lastly, in STEP 10 the most important models were used to form an ensemble. In the process of models selection for the purpose of ensemble construction we took into consideration, both the quality of the reggresion model as well as the degree of diversity in relation to already selected models. The course of steps 9 and 10 has been described as Algorithm 1.

the changes between pairs of same statistics in consecutive sliding windows. The inter-window stats are presented in Table III.

In order to optimize time-effectiveness of the experimentation process the framework implementation was modified so that it could be run in several modes, including: incremental stream processing[5], single-threaded mode and the map-reduce mode. In order to allow multi-mode framework operation, the feature extraction mechanism was isolated as a separate tool with respect to the runtime environment.

The solution for the competition was calculated in batch mode and launched on a cluster of Docker[2] containers with installed Hadoop[3] software library as an implementation of MapReduce protocol. The scheme of the whole feature extrction process is depicted in Figure 1. In the "Map" phase (compare steps 2 and 3 in Figure 1) every data row was divided into: sub-streams of numerical readings from various sensors, set of nominal and aggregated features and, in case of the training set, also label. Labels, nominal and aggregated attributes were transferred to the "Reduce" phase unchanged while the numerical streams were subjected to an additional feature extraction described above. In the "Reduce" phase

Table II
THE TABLE PRESENTS FEATURES WHICH ARE CALCULATED TO REPRESENT THE TIME SERIES IN A SLIDING WINDOW.

feature	description
max	a maximum value of the readings in the window
min	a minimum value of the readings in the window
maxMinDiff	a difference between the max and min
mean	a mean value of readings in the window
percentileX	a Xth percentiles for the readings, where: $X \in \{2,5,10,15,20,25,30,50,70,75,80,85,90,95,98\}$
percentiles5Diff	a subtraction of the percentiles 95% and 5%
stdDev	a standard deviation of the readings
variance	a variance deviation of the readings
fftCoeffSet	a set containing first 5 Fourier transform coefficients
kurtosis	a Kurtosis measure [1]
skewness	a measure of the asymmetry

Table III
INTER-WINDOW STATISTICS THAT EXPRESS THE CHANGES BETWEEN A PAIR OF EQUIVALENT FEATURES IN CONSECUTIVE WINDOWS.

feature	description
maxDiff	a difference between *max* stats in the consecutive sliding windows
meanDiff	a difference between *mean* stats in the consecutive sliding windows
minDiff	a difference between *min* statistics in the consecutive sliding windows
percentileXDiff	a difference between *Xth* percentile statistics in the consecutive sliding windows, where: $X \in \{5,25,50,75,95\}$

[2]See https://www.docker.com
[3]See http://hadoop.apache.org

(steps: 4 and 5 in Figure 1) all the attributes obtained for each row were combined back together.

IV. Feature Selection

The experimentation was significantly affected by the fact that the training data set was very unbalanced, rows labeled as 'warning' represented only 2,3% of all objects. Therefore, in the first step we drew a number of random samples that contained between 10 000 and 20 000 objects out of 133 150 all objects of the training set. Created samples differed in the number of objects of "warning" class (minority class), each contained a minimum of 1000 and a maximum of 2000 objects of this class. Objects within a particular sample were unique, however they could be repeated between different samples.

The generated data samples were randomly divided into two disjoint groups: group A - containing object subsets for the purpose of the feature selection, group B containing samples for the purpose of training of regression models. It should be noted that in order to use the chosen feature selection algorithms, before generating the 'A' samples the training set was subjected to discretization of a numerical attributes using local (univariate) version of the algorithm described in [18].

The selection of attributes [12] was carried out on the basis of a filter method derived from the theory of rough set[19]. Using the R[4] language and environment for statistical computing with installed RoughSets [20] package we calculated approximate decision reducts[13] with DAAR [11] heuristic. Approximate decision reducts are relatively small, thus, we decided to merge few as a single attribute subset. As a result of feature selection process (depicted in Figure 2) we prepared a number of significantly different attribute subsets.

V. Models Training and Ensambling

The task of maximizing AUC measure, posed by the organizers of the competition, prompted us to apply regression algorithms to identify the probability of the 'warning' label. The final solution is based on the concept of building an ensemble of diverse regression models which interpret 'warning' label as 1, while 'normal' label as 0. To provide a diversity of models we trained them on various subsets of attributes, what was important in order to provide a variety of regression models because the analyzed data set had few very dominant attributes. An additional effect of using different features for different models was to protect the ensemble against significant concept drift [1] on the part of the attributes between the training and test sets. This approach was also expected to protect the model against over-fitting [15] and, hence, against the significant decrease in the quality of prediction on the test set.

The machine learning were conducted on pre-prepared samples of objects with reduced number of attributes. The ultimate ensemble consisted of 8 significantly different regression models which were calculated with three various algorithms, including: regression trees (calculated by the algorithm from the rPart[5] package), SVM regressor (computed using the algorithm from the e1071[6] package) and the glm[7] function from R language to fit a generalized linear model. The following list presents models included in the final ensemble:

- Five simple regression tree models calculated with rPart.
- Two SVM models with different kernel functions
  - SVM_1 - regression, kernel: linear, cost: 1, gamma: 0.1, eps: 0.1 , Number of Support Vectors: 2968
  - SVM_2 - regression, kernel: radial, cost: 1, gamma: 0.07143, eps: 0.1, Number of Support Vectors: 7171
- One generalized linear model

Algorithm 1: The construction of regressors ensemble.

Data:
- *attSubsets* - pre-calculated subsets of attributes - approximate reducts
- *objectSamples* - pre-calculated object samples
- *testSet* - test set
- *regressionAlgorithms*, default: { rPart, SVM, glm }
- *allowedAttempts*, default: 3
- *minimalQualityTreshold* - minimal quality treshold

Result:
- *ensemble of regression models*

```
/* Initialization of variables                    */
ensemble ← ∅;  weakAttempts ← 0
alg ← regressionAlgorithms.removeFirst
while TRUE do
    a1, a2 ← attSubsets.drawAndRemoveTwo
    b1, b2 ← objectSamples.drawAndRemoveTwo
    /* Every model is trained and
       validated on various samples               */
    model ← alg.trainAndEvaluate(a1, b1, a2, b2)
    score ← model.score(testSet)
    /* The ensemble is expanded if the
       model meets the quality threshold
       and there is no similar model              */
    if model.evaluation > minimalQualityTreshold ∧
     ¬ensemble.containsSimilar(model, score) then
        ensamble ← ensamble ⋃ {model ⊕ score}
    else
        weakAttempts ← weakAttempts + 1
        if weakAttempts < allowedAttempts then
            continue;
        if regressionAlgorithms ≠ ∅ then
            alg ← regressionAlgorithms.removeFirst
              weakAttempts ← 0
        else end of experimentation
            break;
return ∑_{s∈ensemble.scores} s;
```

[4]See https://www.r-project.org

[5]See https://cran.r-project.org/web/packages/rpart

[6]See https://cran.r-project.org/web/packages/e1071

[7]See https://stat.ethz.ch/R-manual/R-patched/library/stats/html/glm.html

The whole phase of machine learning were carried out in the R statistical environment. The ensemble was successively extended with new models based on one of three designated algorithms, starting from the the rPart algorithm which in the initial assessment achieved the most promising results. In each step of the experiment: a sample of objects from those available in group 'B' and a subset of the attributes form those obtained in the phase of feature selection were drawn. The prepared subset of the training set was used to train a single regression model which was added to the ensemble under two conditions. First, the evaluation of the results had to exceed the satisfactory quality threshold. Second, the results of the regression for the test set had to be significantly different from any of the models already added to the ensemble. A detailed description of the experiment is presented in the Algorithm 1.

VI. Summary

The paper introduces an automated framework of extraction and selection of attributes designed to work with big data using MapReduce programming model. The article presents the proof of concept application of the framework to build an ensemble of classifiers based on a simple heuristic indicating the extension of the ensemble which takes into account both the quality and the diversity of the ultimate solution. The effectiveness of the developed solution has been verified by the participation in an open knowledge discovery competition in which it obtained the highest score in terms of AUC (93.96%) of all solutions submitted by 106 participating research teams.

VII. Acknowledgements

This research was partially supported by Polish National Centre for Research and Development (NCBiR) grant PBS2/B9/20/2013 in frame of Applied Research Programme.

References

[1] M. Boullé. Tagging fireworkers activities from body sensors under distribution drift. In Ganzha et al. [4], pages 389–396.

[2] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[3] T. G. Dietterich. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 1–15, London, UK, UK, 2000. Springer-Verlag.

[4] M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors. *2015 Federated Conference on Computer Science and Information Systems, FedCSIS 2015, Lódz, Poland, September 13-16, 2015*. IEEE, 2015.

[5] M. Grzegorowski. Scaling of complex calculations over big data-sets. In D. Ślęzak, G. Schaefer, S. T. Vuong, and Y. Kim, editors, *Active Media Technology - 10th International Conference, AMT 2014, Warsaw, Poland, August 11-14, 2014. Proceedings*, volume 8610 of *Lecture Notes in Computer Science*, pages 73–84. Springer, 2014.

[6] M. Grzegorowski and S. Stawicki. Window-Based Feature Engineering for Prediction of Methane Threats in Coal Mines. In Yao et al. [25], pages 452–463.

[7] M. Grzegorowski and S. Stawicki. Window-Based Feature Extraction Framework for Multi-Sensor Data: A Posture Recognition Case Study. In Ganzha et al. [4], pages 397–405.

[8] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, Mar. 2003.

[9] A. Janusz, M. Sikora, Ł. Wróbel, and D. Ślęzak. Predicting Dangerous Seismic Events: AAIA16 Data Mining Challenge. In M. Ganzha, L. A. Maciaszek, and M. Paprzycki, editors, *Proceedings of FedCSIS 2016*. IEEE, 2016. In print September 2016.

[10] A. Janusz, M. Sikora, L. Wróbel, S. Stawicki, M. Grzegorowski, P. Wojtas, and D. Ślęzak. Mining Data from Coal Mines: IJCRS'15 Data Challenge. In Yao et al. [25], pages 429–438.

[11] A. Janusz and D. Ślęzak. Random probes in computation and assessment of approximate reducts. In M. Kryszkiewicz, C. Cornelis, D. Ciucci, J. Medina-Moreno, H. Motoda, and Z. W. Ras, editors, *Rough Sets and Intelligent Systems Paradigms - Second International Conference, RSEISP 2014, Held as Part of JRS 2014, Granada and Madrid, Spain, July 9-13, 2014. Proceedings*, volume 8537 of *Lecture Notes in Computer Science*, pages 53–64. Springer, 2014.

[12] A. Janusz and D. Ślęzak. Rough set methods for attribute clustering and selection. *Applied Artificial Intelligence*, 28(3):220–242, 2014.

[13] A. Janusz and D. Ślęzak. Computation of approximate reducts with dynamically adjusted approximation threshold. In F. Esposito, O. Pivert, M. Hacid, Z. W. Ras, and S. Ferilli, editors, *Foundations of Intelligent Systems - 22nd International Symposium, ISMIS 2015, Lyon, France, October 21-23, 2015, Proceedings*, volume 9384 of *Lecture Notes in Computer Science*, pages 19–28. Springer, 2015.

[14] A. Krasuski, A. Jankowski, A. Skowron, and D. Ślęzak. From sensory data to decision making: A perspective on supporting a fire commander. In *2013 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Atlanta, Georgia, USA, 17-20 November 2013, Workshop Proceedings*, pages 229–236. IEEE Computer Society, 2013.

[15] P. Lameski, E. Zdravevski, R. Mingov, and A. Kulakov. SVM parameter tuning with grid search and its impact on reduction of model over-fitting. In Yao et al. [25], pages 464–474.

[16] J. Lasek and M. Gagolewski. The winning solution to the AAIA'15 data mining competition: Tagging firefighter activities at a fire scene. In Ganzha et al. [4], pages 375–380.

[17] M. Meina, A. Janusz, K. Rykaczewski, D. Ślęzak, B. Celmer, and A. Krasuski. Tagging firefighter activities at the emergency scene: Summary of AAIA'15 data mining competition at knowledge pit. In Ganzha et al. [4], pages 367–373.

[18] H. S. Nguyen. On efficient handling of continuous attributes in large data bases. *Fundam. Inform.*, 48(1):61–81, 2001.

[19] Z. Pawlak. Rough sets. *International Journal of Parallel Programming*, 11(5):341–356, 1982.

[20] L. S. Riza, A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. Ślęzak, and J. M. Benítez. Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "roughsets". *Inf. Sci.*, 287:68–89, 2014.

[21] M. Sikora and B. Sikora. Improving prediction models applied in systems monitoring natural hazards and machinery. *International Journal of Applied Mathematics and Computer Science*, 22(2):477–491, 2012.

[22] J. Stefanowski, A. Cuzzocrea, and D. Slezak. Processing and mining complex data streams. *Inf. Sci.*, 285:63–65, 2014.

[23] S. Wawrzyniak and W. Niemiro. Clustering approach to the problem of human activity recognition using motion data. In Ganzha et al. [4], pages 411–416.

[24] A. Wieczorkowska, J. Wroblewski, P. Synak, and D. Ślęzak. Application of temporal descriptors to musical instrument sound recognition. *J. Intell. Inf. Syst.*, 21(1):71–93, 2003.

[25] Y. Yao, Q. Hu, H. Yu, and J. W. Grzymala-Busse, editors. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing - 15th International Conference, RSFDGrC 2015, Tianjin, China, November 20-23, 2015, Proceedings*, volume 9437 of *Lecture Notes in Computer Science*. Springer, 2015.

[26] A. Zagorecki. A versatile approach to classification of multivariate time series data. In Ganzha et al. [4], pages 407–410.