Proceedings of the 2014 Federated Conference on
Computer Science and Information Systems pp. 345–354

DOI: 10.15439/2014F507
ACSIS, Vol. 2

Key Risk Factors for Polish State Fire Service: a Data Mining Competition at Knowledge Pit

Andrzej Janusz[*], Adam Krasuski[†], Sebastian Stawicki[*], Mariusz Rosiak,
Dominik Ślęzak[*‡] and Hung Son Nguyen[*]

[*]Institute of Mathematics, University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
{janusza,slezak,son,stawicki}@mimuw.edu.pl
mariusz.rosiak@gmail.com

[*]Section of Computer Science, The Main School of Fire Service
ul. Słowackiego 52/54, 01-629 Warsaw, Poland
krasuski@inf.sgsp.edu.pl

[†]Infobright Inc.
ul. Krzywickiego 34, lok. 219, 02-078 Warsaw, Poland

***Abstract*—In this paper we summarize AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service which was held between February 3, 2014 and May 5, 2014 at the Knowledge Pit platform http://challenge.mimuw.edu.pl/. We describe the scope and background of this competition and we explain in details the evaluation procedure. We also briefly overview the results of this analytical challenge, showing the way in which those results can be beneficial to one of our other projects which is related to the problem of improving firefighter safety at a fire scene. Finally, we reveal some technical details regarding the architecture and functionalities of the Knowledge Pit competition platform, which we are developing in order to facilitate solving of practical problems that require advanced data analytics.**

***Keywords*-data mining competition, risk factors, attribute selection, EWID system**

I. Introduction

INCIDENT DATA REPORTING SYSTEMS (IDRS) are used by public safety services across the globe to gather information about the incidents which required their actions. The information is gathered in order to calculate statistics within the groups of incidents, identify peculiar cases and to improve the procedures [1]. Results of a thorough analysis of incident reports can also be utilized by decision support systems to increase safety of firefighters at a fire scene [2].

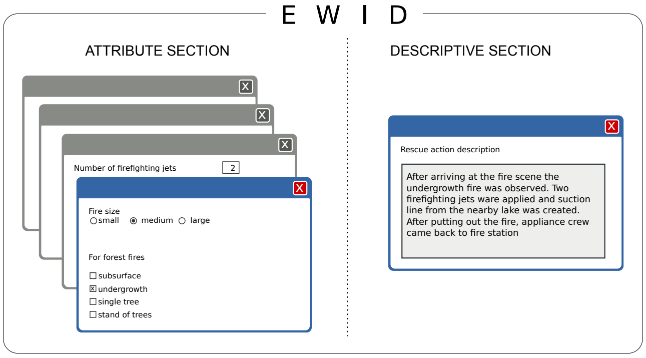
EWID is an example of such a reporting system. It is used by the State Fire Service of Poland [3]. A report submitted to the system by Incident Commander (IC - a coordinating officer) after a fire and rescue action (F&R) consists of two parts: a quantitative description, where facts regarding the action are expressed by numerical or categorical characteristics and a description in a natural language. The first part is often called the attribute section and the second is the descriptive section.

The attribute section is represented in a form of structured and quantified characteristics. Among over 500 attributes, it contains information about incident type, its severity or size and resources involved in the response. The descriptive section can be treated as an extension to the attribute section. It contains a natural language description of probable causes, conditions at the event scene and the course of the action. Figure 1 depicts a chunk of a report submitted to the EWID system.

It is assumed that the descriptive section should contain all the relevant information which could not be expressed in the attribute section. However, due to the fact that there are no instructions regarding what information is relevant in a context of a particular incident type, the descriptive section sometimes contains irrelevant and useless fragments of text. A quality of the textual descriptions in the system also variates, depending on a personality and attitude of IC who writes the report. For instance, a content of the part devoted to the course of the action may range from very useful information concerning the consecutive decisions of IC, applied techniques and their consequences, to very cursory and ambiguous sentences such as: *rubbish lit*.

On the other hand, the attribute section is unable to reflect all information regarding a very large spectrum of possible incidents. All the above mentioned shortcomings make it challenging to extract useful information from the EWID reports, especially when this information is only indirectly related to the set of characteristics from the attribute section [4]. One example of a task that requires such information is the problem of recognition of risk factors which affect the possibility of a serious injury or death among firefighters and other people involved in various incidents. A similar problem, i.e. the identification of threats, has been already investigated by several researchers [3], [5], [6].

In the research presented in this paper we address the above mentioned challenge. We decided to ask the machine learning community to identify characteristics extracted from the EWID reports, which are useful for predicting whether any people were harmed during a given incident. For this



Number of firefighting jets 2

Fire size
○small ● medium ○ large

For forest fires
□subsurface
☒undergrowth
□single tree
□stand of trees

Rescue action description

After arriving at the fire scene the undergrowth fire was observed. Two firefighting jets ware applied and suction line from the nearby lake was created. After putting out the fire, appliance crew came back to fire station

Fig. 1. The chunk of the EWID report.

purpose, we devised a data mining competition, titled *AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service*. A form of this challenge was similar to the competitions which our team organized in the past [7], [8] on the TunedIT platform [9]. This time, however, we organized it on a novel web platform called Knowledge Pit (www.challenge.mimuw.edu.pl) hoping that the participants will be able to enrich our understanding of the EWID data and point at attributes that describe the most relevant information to the stated problem.

In the following sections we reveal details regarding the architecture of Knowledge Pit (Section II) and then, in Section III, we describe the proposed challenge. Next, in Sections IV and V, we present an overview of results obtained by participants of the competition and analyze those results with respect to the semantic types (i.e. the meaning) of attributes that were frequently appearing in the submitted solutions. Finally, we conclude the paper by drawing our plans for a continuation of this study.

II. The Knowledge Pit Platform

Knowledge Pit is a platform created to support organization of data mining challenges. It is designed in a modular way, on top of an open-source e-learning platform Moodle.org [10], to follow the best practices of a software development. Therefore, the platform with its current modules, including user accounts, challenges and resources management subsystems, time and calendar functionalities, communications features (i.e. forums and messaging subsystems), and a flexible interface for connecting automated judging services prepared to evaluate contestants' submissions, is conceptually ready to introduce new features or enhance the existing ones.

A more detailed architecture overview requires to describe two main parts of the system. All elements that are available to the users interested in participating in a data mining competition, together form a web user interface. To fulfill this requirement, Knowledge Pit utilizes a very popular solution stack Apache/MySQL/PHP – a set of software components that is sufficient to provide web solutions ranging from simple to complex ones [11], [12], [13]. The second part of the system concerns competition handling from the point of view of evaluating the submitted solutions. This functionality is separated from the remaining part of the platform to cope with the requirement for high flexibility (with regard to a programming language or a framework, parallelization of expensive calculations, etc.) of the judging software setup. The general architecture of the system is presented in Figure 2.

From the point of view of Knowledge Pit system there are several roles which can be assigned to a user – a guest, a contest participant or a contest organizer role. Therefore, the front-end engine consists of several modules which provide the functionalities to the users, depending on their role in a given moment. The main modules of the system are as follows:

- user management and user privileges
- challenge maintenance
- challenge Leaderboard
- challenge submissions manager
- calendar
- forum module
- internal messaging system
- chat
- private resources (files) repository

The above modules can be thought of as pieces of software that implement specific elements of the system. When combined, they constitute the higher-level features described below in this section.

A. User interface

Knowledge Pit implements users and user groups management, an advanced privileges support and an enhanced context handling, e.g. a user can be a guest in a given challenge, a participant in other and also a creator and manager of another one. The site administrators can manually promote or demote users access corresponding to any of the given contexts, e.g. a context of the page a user can browse. This means that the administrators can grant privileges in a local context (e.g. a forum of a specific contest, a chat, etc.) leaving the access privileges to the other parts of the site unchanged. Moreover, a special registry is used to administer the users and the user groups. If necessary, a new user type or users group can be created with selected privileges granted, thereby facilitating the task of managing large number of users in some particular contexts.

Each user is assigned to a set of assets such as a private file repository, a dedicated calendar with adjustable scope and event levels, a public profile shown to others in contexts of chats or forums, and a personalized site appearance – a set of settings that allows to adjust the way how the site looks, e.g. a user can hide or move specific parts of menus and navigation modules, or use predefined site themes, all accordingly to his own choice.

Each site visitor can view a calendar on which events are displayed accordingly to the access level and the site context. The calendar is fully customizable and has events ordered according to scopes:

- global

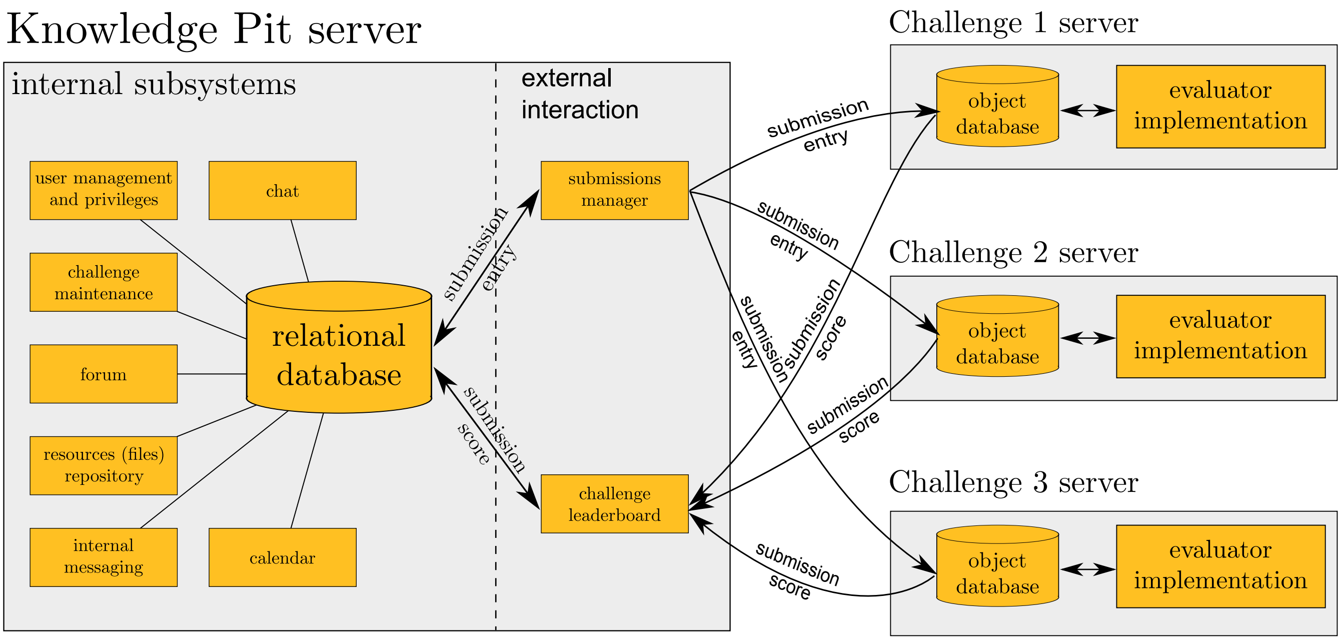


Fig. 2. A system architecture of the Knowledge Pit web platform.

- competition
- group
- user

Apart from events generated automatically with respect to each of the categories mentioned above, users can add their own events and bind them to the given levels. This functionality may facilitate cooperation between users of Knowledge Pit that decide to work in teams.

Calendars can be exported to standard exchange formats (to be imported into other calendar solutions) or can be subscribed to by RSS. This helps users to stay up to date and somehow automate their time management.

Each user has a dedicated storage space (or private resources repository) with an access and management abilities available via the web interface.

Users can communicate using the built-in chat and forum engines. They can exchange messages at various levels according to the context they are enrolled in. Each competition includes a dedicated internal chat and a forum. There are also global forums and chat rooms available to the site users, which can be enabled or disabled by the competition managers. This is yet another idea on how the Knowledge Pit platform can be useful – it may help scientists to get know new people with the same interests and stimulate their cooperation.

B. Definition and management of competitions

Data mining challenge support and maintenance are the main objectives of the site. Each contest is an individual entity in the system and it requires time and care to be well defined and created. There are two ways to achieve the goal of starting the competition. The site administrators create, run and hook a dedicated evaluator to a competition, due to theirs responsibility, knowledge and appropriate privileges, bypassing a standard acceptance procedure. The second (more common) way involves the operation of a regular user who prepares a project of the competition using the forms and tools available in Knowledge Pit. Then, a request to the site administrators is sent to revise the proposal of the contest. If it is accepted, a new hidden challenge is automatically created and the user obtains manager privileges to it. All standard modules are initialized by the default values of parameters. That includes data description pages, data files folder, news panel and submissions upload interface. Initially, a newly created challenge is hidden from other users until all necessary information is filled in and the dates are defined. After providing the necessary data (including a task description, input data files, etc.), the challenge may be published and becomes visible to anyone visiting the site.

C. A course of a competition

To each challenge there are associated three important dates:

1) a start date
2) a submissions end date
3) a contest end date

The first two dates define the actual time in which participants may compete by submitting their solutions to the challenge task. During this period the users are supposed to upload and manipulate their solution files using the available submission manager. The submitted files are automatically pre-evaluated by a dedicated service (e.g. a script or an external application that are compatible with a Knowledge Pit evaluator protocol), accordingly to the settings defined by the challenge authors. The users may upload multiple result data files among which they mark one as their target solution that would take part in the final evaluation. The pre-evaluation is meant to generate scores that are placed on the competition's Leaderboard. The preliminary scores may serve as a rough estimation of how good the submitted solution is in comparison to results of the other users and therefore may stimulate the competitive spirit of the contest. The second period which ends the contest is the time needed to finally evaluate users' submissions (previously marked by the participants as their target solutions) and determine the winners. Each submission is scored and accordingly placed in a table called Final Leaderboard which is published within the challenge summary.

Organizers of a competition may require additional reports that describe solutions provided by the participants. In this case a competition manager may set a condition which restrains the final evaluation to the submissions with an attached report.

Each challenge, if it is already published and its visibility is not restricted by the organizers, is available to any site user. In that case, a guest has an access to the contest's general information, including a task description, a list of important dates, and the Leaderboard. Unless contest access restrictions are in effect, every user can enroll in a chosen challenge accessing all the contest's resources, including the provided data files.

D. Evaluation of the uploaded solutions

The above description refers only to the features available via the web user interface. It also presents the general flow of events starting from the contest organization, the users enrollment, the submission of solutions, their preliminary and final evaluation, ending on the publication of the challenge summary and announcement of the winners. However, not much was said about the method of defining the evaluators.

It would be very difficult to build and share a general evaluator for all possible types of data mining competitions. Usually, the stated data mining tasks are very different in many aspects, including:

- the category of performed data analysis (clustering, classification, multi-label classification, etc.)
- the solutions representation (a single file vs. multiple files placed together in one archive file) and their formats (multicolumn answers, the way of describing clusters, etc.)

Another important aspect of the data mining solution evaluation is that it often requires a lot of resources (memory, CPU time, disc I/O or database connections), e.g. when it is associated with a predictive model creation. This could result in malfunction of the competition platform due to its resources limitation. In Knowledge Pit the responsibility of the evaluation is delegated to the competition organizers. They need to provide an object database and a working evaluator. The responsibility of Knowledge Pit is limited to interaction with the object database where all the solutions are uploaded and stored. The submission scores are downloaded from it and propagated to internals of the system. The evaluator may be implemented in any suitable programming language, as a script, a stand alone compiled application or a utilization of available libraries. The only requirement is that it should maintain correct protocol of information exchange by means of changing the objects inside the database in a predefined way. The proposed flow of responsibilities frees Knowledge Pit from the things which it cannot cope with in a generic way. It also gives the organizers a very flexible method of expressing their data mining task in a form of a fully customizable evaluation procedures. For example, in AAIA'14 Data Mining Competition which is described in the following sections, MongoDB [14] was utilized as the object database and the evaluation system was implemented in the R programming language [15].

III. The Task Description

The Knowledge Pit was inaugurated with AAIA'14 Data Mining Competition which took place between February 3, 2014 and May 7, 2014. In this challenge the focus was on the feature selection problem and the data came from the public safety domain.

Our team obtained a data set containing nearly 260,000 reports from the EWID system. The reports corresponded to actions carried out by the Polish State Fire Service within the city of Warsaw and its surroundings (the Mazovia district) in years 1992 – 2011. We preprocessed a subset of this data and transformed it into a table in which each of the reports is described by nearly 12,000 attributes. Additionally, we distinguished three target attributes that correspond to information whether in the described incident there were casualties among firefighters, children or other involved people, respectively. The task in AAIA'14 Data Mining Competition was to identify attributes that can be used to robustly assign the reports to the corresponding decisions labels. We hoped that participants would come up with solutions which improve our understanding of the risk factors associated to various types of accidents.

The competition data set was provided to participants in two different formats. The first one was a traditional tabular representation of data as a comma-separated values file. Each row of this file represented a single EWID report and, in the consecutive columns, it contained values of its characteristics (the attributes). The attributes in this table could be divided into two groups. The first one contained the features extracted from the quantitative part of the report and the second group corresponded to a document-term matrix obtained from the natural language description sections. In total, the training data available to participants contained descriptions of 50,000 incident reports. Each report was characterized by 11,852 conditional attributes. All the attributes were discrete and only a few had more than two possible values. We thought about those attributes as indicators of the risk factors corresponding to the incidents.

The same data set was made available in a sparse matrix format as an EAV file [16]. In every row, the file contained exactly three integer numbers: an identifier of an object, an identifier of an attribute and the corresponding attribute value.

Since the EAV file stored exactly the same information as the traditional tabular representation of the data, this file was provided only for convenience of participants.

To each of the reports from the training data there were also assigned values of three binary decision attributes. The first decision attribute indicated incidents in which occurred a serious injury or death of one of the firefighters or members of the rescue team. The second decision attribute indicated cases for which there were children among the injured people. The third decision identified situations where any civilians were hurt. Values of those decision attributes were made available for all participants of the competition in a separate file.

It is worth noting that, by its nature, the provided data set was highly dimensional. The total number of conditional attributes corresponded to the number of distinct words in the textual part of the reports (after lemmatization), plus several hundreds of attributes from the quantitative part. Additionally, the data was sparse since only a small fraction of the attributes had a non-zero value for a particular report. On top of that, all three decision attributes were highly imbalanced – the positive classes corresponded to relatively rare events. The proportions of the positive cases for the rescuers, children and civilians were ≈ 0.004, ≈ 0.007 and ≈ 0.059, respectively. There was also a separate test data set which was used for the evaluation of submissions. It had similar characteristics to the training data but it was not available for the participants during the competition.

The competitors were asked to indicate sets of attributes that allow to accurately classify the incidents using an ensemble of Naive Bayes models [17], [18] and upload their solutions using the on-line submission system. We required that in each solution there were exactly ten attribute sets. The sets were ought to contain at least three integer numbers corresponding to indexes of attributes from the training data set. There was no upper limit for the number of attributes indicated in a single set, however, the evaluation system penalized solutions that use a large number of features.

The submitted solutions were evaluated on-line and the preliminary results were published on the competition Leaderboard. The preliminary score was computed for each submission on a random subset of the test set, which was fixed for all participants. This subset corresponded to approximately 10% of the test data. The final evaluation was performed after completion of the competition using the remaining part of the test data. Those results were also published on-line. In order to be considered for the final evaluation, each participating team had to provide a short report describing their approach.

Quality of the submissions was assessed by measuring performance of a classifier ensemble composed of Naive Bayes models. Those models were constructed using attribute sets indicated by the submitted solution, separately for each decision attribute. An output of the ensemble was computed by averaging probabilities of the positive classes returned by the individual Naive Bayes models. During the evaluation, all the training data was used for the construction of the models. The performance of a single ensemble was measured by taking Area Under the ROC Curve (AUC) [17], [18] of the probability predictions for the corresponding decision attribute and the result was averaged over all three decision attributes. Finally a penalty was applied for using a large number of conditional attributes.

In more details, if we denote by:

s – a submitted solution,
$|s|$ – a total number of attributes used in the solution (counted with repetitions),
$AUC_i(s)$ – Area Under the ROC Curve (AUC) of a classifier ensemble for the i-th decision attribute,

then the quality measure used for the assessment of submissions can be expressed as:

$$score(s) = F\left(\frac{1}{3}\sum_{i=1}^{3} AUC_i(s) - penalty(s)\right)$$

where the penalty is equal to:

$$penalty(s) = \left(\frac{|s| - 30}{1000}\right)^2$$

and the function F is defined as:

$$F(x) = \begin{cases} x & \text{for } x > 0 \\ 0 & otheiwise \end{cases} .$$

All the data sets utilized in the competition, including the test set with the corresponding decision values, were made available after completion of the challenge at the competition web page: http://challenge.mimuw.edu.pl/contest/view.php?id=83. We are convinced that the public availability of the data will facilitate future research in this area by other members of the machine learning community.

IV. Results of the Competition

AAIA'14 Data Mining Competition attracted many skilled participants from around the world. In total there were 116 registered teams, from which 57 actively participated in the challenge by submitting at least one solution to the stated task. We received nearly 1,300 solutions and 290 of those submissions obtained a score higher than 0.94. Additionally, 46 teams provided a short report describing their approach.

The participants utilized diverse machine learning techniques in order to come up with their final attribute sets. A large share of the solutions was devised by combining the attribute filtering approach for reducing the initial feature subset with well known wrapper-based techniques. The final solutions were commonly tuned using evolutionary algorithms or the hill climbing method. The best results, however, were obtained by using algorithms optimized specifically for finding attribute sets that improve the AUC of Naive Bayes prediction models.

The wide spectrum of solutions submitted by the participants during the challenge makes it possible to perform a comprehensive study of the factors that have the biggest impact on predictions of the positive classes in the data. However, since

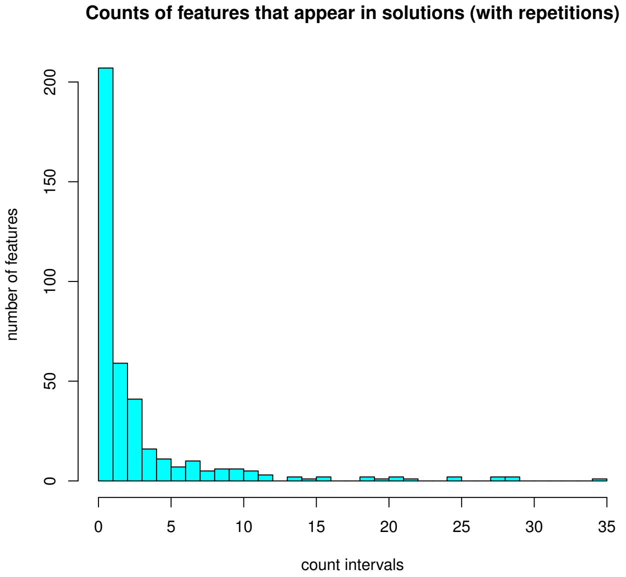


Fig. 3. Frequencies of individual attributes in the final solutions submitted by the twenty top-scored participants of AAIA'14 Data Mining Competition. The attributes were counted with every repetition. The most frequent attribute was present 35 times in the considered solutions. It corresponded to the term "during".

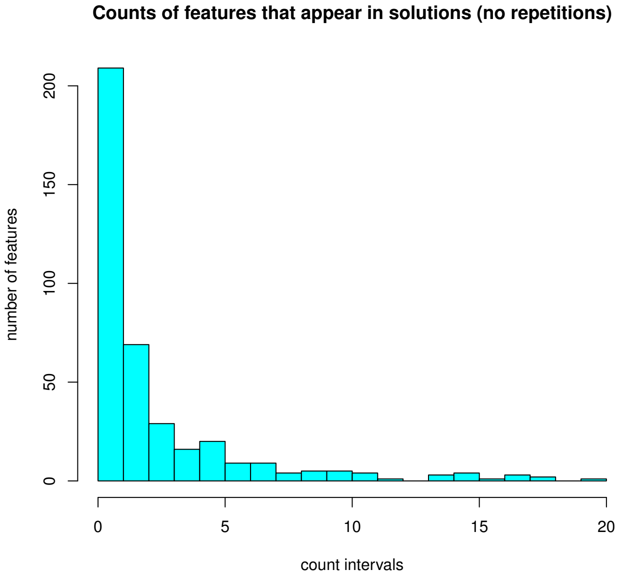


Fig. 4. Frequencies of individual attributes in the final solutions submitted by the twenty top-scored participants of AAIA'14 Data Mining Competition. The attributes were counted without repetitions (an attribute is counted once for every solution in which it appears). The most frequent attribute was present in all of the considered solutions. It corresponded to the term "during".

the considered phenomena are complex and diverse, and there was a short period between the submissions of the solutions and the preparation of this paper, the analysis described here is limited to a set of hypotheses. Those hypotheses try to explain some of the observed regularities, however, we need to stress that most of them require further investigation to be supported with strong statistical evidences. Therefore, all the explanations which we present in the following sections are just a starting point for a more detailed analysis leading to a better understanding of dangerous situations and the related threats faced by firefighters.

A. Analysis of frequent features

We start our analysis with reviewing the features which most commonly appeared in the top-scored solutions. Each of the features in question corresponds to a specific attribute from the attribute section, a word or a word-set (see Section III) from the descriptive section of the reports.

We analyzed the attributes indicated by the final solutions submitted by the twenty participants who obtained the highest ranks on the Final Leaderboard. In total (counted with repetitions) they constituted a set of 1,538 attributes. The number of different attributes in this set is 413. In most of the cases an attribute was used only in a single solution. However, there is a small subset of features that are more frequently used than the others. We focused our investigation on these attributes and verified their relation with the occurrence of serious injuries. Figures 3 and 4 depict the frequency of individual attributes, computed by considering every occurrence of attributes and by counting the solutions in which an attribute is present, respectively. In the second case, an attribute is counted only once for every solution in which it appears, regardless of the number of its appearances in that solution.

We analyzed the most frequent attributes with regard to their relation to the fire safety domain. We arranged them in several groups. Each of those groups defines a different type of relation to the cases of serious injuries or deaths in incidents.

The first group consists of features that correspond to nouns related to the injured parties. Incident Commanders who report the casualties tend to use a set of specific terms for describing the victims. Those terms are not used to refer to other people present at the emergency scene. In particular, examples of such nouns are: "a boy", "a kid", "a girl", "male", "female" or "a private". The usage of these words in a report indicates that something happened to the described person. There is also some interesting regularity in this group: the term "boy" appears nearly three times more often than the word "girl". The other interesting thing observed among the attributes in this group is that the terms referring to the common participants of the interventions change in situations when they get injured. For example, a firefighter who performs his/her activities according to the schedule, is usually anonymous (e.g. "two firefighters set the ladder"). However if something wrong happens, a full rank, the name and surname are reported in the descriptive section of the report. In the most of cases the injuries concern lower-ranked firefighters. Therefore, attributes represented by terms such as "str" (in Polish it is an abbreviation for "a private") were often selected as good predictive features.

TABLE I
THE FEATURE SETS SUBMITTED BY THE WINNERS OF AAIA'14 DATA MINING COMPETITION.

#Sub	First place	Second place	Third place
1	action_people_evacuation, to_sting, longitude, thought, ignition, action_people_release, losses_overall_dollars, hospital_name, private	local_threat_type_road_transport, kid, knee-joint, coal, head, to_transport, observation, bite	palm, during, head, action_people_evacuation, icicle, bar, man
2	road, observation, hand, man, hit, transport, rinse	private, action_people_evacuation, to_drive_away, passenger, warm, carbon_monoxide_poisoning, first_fire_engine_drove_km	thumb, withdraw, blast_off, losses_overall_dollars, grating, boy, observation
3	year, right, to_transport, leg, kid, during, used_equipment_chainsaw, to_go_somewhere, light, hospital_name, daughter, ankle	resources_fire_rescue_unit, person, ankle, condition, foot, action_used_extinguishing_aid_attack, hospital_name	girl, to_bite, carbon_monoxide_poisoning, to_lead, to_sting, action_people_release, to_hit_a_pedestrian, local_treat_medium
4	action_door_opening, decease, compartment, allotment garden, to_transport	light, used_equipment_chainsaw, to_faint, carbon_monoxide, to_sting, corpse, apply	kid, resources_fire_rescue_unit, hospital_name, driver, volume_of_incident_scene, finger, daughter
5	carbon_monoxide, oxygen_therapy, local_threat_cause_gas_device_fault, local_threat_medium, extinguishing_in_attack, burn, victim	kid_or_kids, eye, minibar, street_name, officer_name, latitude	work, apparatus, drive, hospital_name, deceased, light, action, delay
6	knee, head, rescuers_fire_rescue_unit_people, district	suffer, leg, grating, deceased, knee, extricate, palm	kid, make_of_a_car, injury, hospital_name, extract, hand, morning
7	action_smoke_extraction, palm, latitude, local_threat_type_road_transport, person, connector_or_a_switch, hospital_name, state_property, deceased	truck, to_set_fire, star_or_a_whistle, to_hit_or_run_over_someone, longitude, extract, compartment	truck_type, ankle, to_slip, private, hospital_name, to_swell, to_bite, corpse
8	resources_police_cars, team, technical, kid, water, to_drive_a_driver_or_a_steering_wheel, to_lead, cause_of_a_local_threat_act_of_terror, to_do_or_break	action_people_release, to_hit_or_crash, agricultural, immediately, department_branch_or_division, withdraw, boy	year, knee, passenger, suffer, personal_details, foot, firefighter
9	explosion_any_type, delay, face, homeless, grating, forearm	during, hand, to_fall_asleep, local_threat_medium, functioning, explosion_any_type, oxygen_therapy	leg, to_force, to_hit_or_crash, oxygen_therapy, coal, to_hit_or_run_over_someone, carbon_monoxide, to_wash, technical
10	corpse, girl, to_hit_or_run_over_someone, suffer, boy, burn_down	darkness, to_hit_or_knock_someone_off, ankle, to_twist, man, action_inside_chimneys, delay	team, local_threat_cause_careless_driving, orthopedic, face, to_carry_or_transport, mean_of_transport, compartment, person

The second group is related to descriptions of the injuries and mostly consists of names of human body parts. In this group, the most commonly used words are: “leg”, “palm”, “hand”, “side”, “body”, “foot”, “twinkle”, etc.

The next group represents the attributes that describe activities undertaken by firefighters when they faced an injured person. This group consists of attributes from the attribute section such as: “action_people_evacuation”, “localizing_people”, “oxygen_therapy”, etc. or words (mostly verbs) from the descriptive section such as: “to transfer” (to an ambulance), “to transport”, “observation”, “to cut”, “to open”, etc.

Another group represents features related to terms which describe a cause of the injuries or fatalities. In this group we find the following words: “intoxication”, “to hit” (a pedestrian), “to get” (a stroke), “sprain”, “to twist” (an ankle), “to slip”, “bite”, “bump”, etc.

All the groups described above, are examples of attributes which were used by ICs in order to address matters related to an injury or death. They can be useful for a post-incident analysis of the causes, since the identification of such key phrases may boost performance of information retrieval systems that work on EWID data. However, those terms alone do not reveal any specific risk factors related to the fire safety domain. The knowledge resulting from the identification of those attributes does not have a direct impact on the safety of firefighters and incident victims. Moreover, it does not reveal interesting aspects of the rescue actions, apart from the words or phrases which are used in the reports in order to describe the casualties or fatalities.

There is, however, one group of features which are likely to correspond to important risk factors. By obtaining information regarding those factors during a real-life F&R action we may potentially improve the safety of involved people. This group consists of attributes corresponding to terms such as: “carbon monoxide”, “darkness”, “single-family terraced buildings”, “mart”, “electrocution”, “bite off” and some specific geographical coordinates. A further analysis of a role and a context of these attributes in the reports may shed light on the factors that affect the possibility of serious incidents. Nevertheless, a thorough investigation is required in order to explain their role in the generation of the unwanted events.

B. Analysis of frequent attribute sets

Due to the fact that usefulness of knowledge obtained by the analysis of individual attributes was limited, we performed an additional investigation of frequent attribute sets. In this analysis we distinguish global and local sets of attributes.

As the first type we consider the whole attribute sets that correspond to individual models in the submitted solutions. The second type refers to subsets of attributes that commonly co-appear in the top-scored solutions.

In our first attempt, we analyzed the global feature sets, i.e., those which turned out to have the best predictive abilities for the whole data. These global feature sets were submitted by the winners of the competition. Table I gives the names of attributes from the sets submitted by the three best teams.

In those groups we indicated a few interesting types of sets that should undergo a further analysis. The first one can be summarized by terms or quantitative attributes such as: "activities_opening_doors", "decease", "compartment", "allotment", "garden", "to transport". Features from this set were often present in reports describing incidents caused by homeless or youngsters who illegally occupy cottage-gardens. This may indicate that there is a considerably large fraction of fatalities resulting from fires started in such conditions.

Another interesting type of attribute sets contains terms such as: "carbon monoxide", "oxygen therapy", "caused_by_heating_device_fault", "extinguishing", "fatality". This group may indicate that a large number of deaths is caused by carbon monoxide poisonings or fires started as a result of malfunctioning heating devices.

The next of the interesting feature set types can be characterized by terms: "explosion", "corps", "face", "homeless", "rate" and "forearm". It seems that there is a considerable number of incidents that involve homeless and some explosions. This set is very difficult to explain without a deeper analysis of the reports describing specific incidents.

A different attribute set type can be represented by the terms: "light", "used_equipment_chainsaw", "wood", "wane", "body", "to sting" and "girl". Combination of those terms often indicates a subset of incidents related to light injuries caused by an inappropriate usage of sharp tools, such as a chainsaw.

There is also a type of attribute sets which may be related to the incidents that happens after a nightfall, in a situation when somebody or something fell into a hole or a chimney. This set is represented by the attributes: "nightfall", "man", "cat", "twist", "shorten", "hit", "action_inside_chimneys".

The last of the identified types of interesting attribute sets is once again related to the problem of carbon monoxide poisonings. However, if the terms "oxygen therapy" or "carbon monoxide" appear along the terms such as "coal", "technical" or "functioning" it may indicate that the problem of poisoning is often related to malfunctioning coal furnaces.

C. Analysis of the local attribute sets

The analysis of the attribute sets submitted by the winning teams was an attempt to identify the most significant factors that have an impact on the occurrence of the cases from the positive decision classes. However, due to a large diversity of interventions of Fire Services – ranging from fires, through road traffic accidents, to natural disasters – finding the globally most affecting features is a very complex task. Therefore, we need to face a problem of finding attribute sets which have an impact on subclasses of incidents such as fires in residential buildings [3]. To accomplish this challenging task we applied a frequent item set mining technique, i.e. the *Apriori* algorithm [19].

We computed frequent attribute sets from the solutions submitted by the twenty top-scored participants (i.e. every line in the solution files was treated as a transaction) and we ranked them according to their support. The utilization of Apriori resulted in finding millions of frequent attribute sets. Due to our limited human processing abilities, we reduced the number of the sets for the analysis to the top 351 with the highest value of the support. Below we present a few examples of the interesting attribute sets which were revealed by this analysis.

As in the previous analysis, the most commonly appearing attribute sets are related to the expressions used by IC in order to report the injuries or fatalities. However, as in the previous cases, we were able to identify a few interesting attribute chunks. All of them should be further analyzed by experts from State Fire Service. Examples of such sets include: "used_equipment_chainsaw" and "light" – it indicates that there is a group of incidents related to an unfortunate use of a chainsaw by fire-fighters. Even though it seems reasonable that in the most of such cases the inflicted injuries are superficial, those results indicate that a proper handling of this type of tools should be better stressed during the firefighter training.

Another group consists of terms: "firefighter" and "sprain". It may indicate that there is a significant number of limb injuries during the rescue actions. The next of the interesting attribute sets is composed of terms: "firefighter", "releasing people" and "bite" which may indicate that there is a number of cases where firefighters are bitten by animals during a rescue activities. The last example of a common attribute set is "to slip" and "hand". It may be considered similar to the group of attributes related to limb injuries. It requires a further investigation in order to be associated with a specific type of firefighter actions.

D. Attribute cluster analysis

After the investigation of attributes and attribute subsets that frequently appear in the best solutions, we decided to check whether there is any redundancy among them. We were also interested in finding pairs of attributes that can be regarded semantically similar in the context of the fire safety. Successful identification of such pairs or groups would be beneficial for the further analysis of the EWID data. It would also be very useful for the risk assessment purposes, in situations when a part of information about an incident is unavailable or unreliable.

In order to find groups of closely related attributes we performed an attribute cluster analysis [20]. Intuitively, two attributes can be considered similar if they often co-occur in the solutions with the same groups of other attributes. However, if a pair of attributes commonly appears in the same sets submitted by the highest scored participants, it means that those features are complement in some way and they should

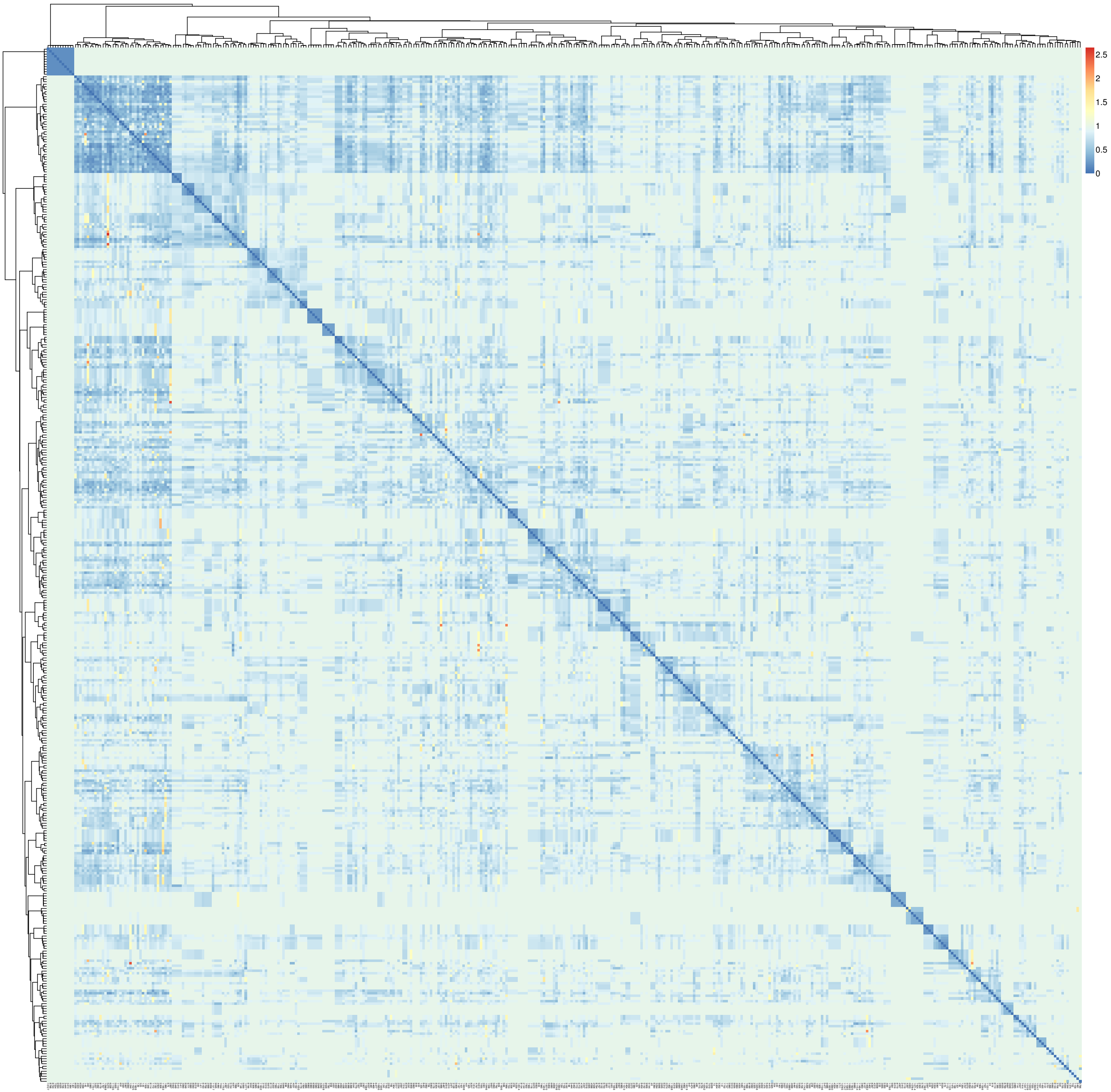


Fig. 5. A co-occurrence-based heat map of the attributes appearing in the top 20 solutions. Rows and columns of the matrix correspond to the attributes and the color of the spots symbolizes their dissimilarity. Additionally, on the top and the left side of the plot, a dendrogram of a hierarchical clustering of the selected attributes is given. The darker squares along the diagonal correspond to clusters of closely related (i.e., potentially exchangeable) attributes.

not be regarded similar. For this reason, we computed the dissimilarity between each attribute pair twofold.

First, we constructed a co-occurrence matrix whose rows and columns correspond to the 413 attributes from the solutions. For every attribute (a row of the matrix), we iterated over the attribute sets in which it was present and increased the matrix entries in the columns corresponding to the co-occurring features, by the inverse of the attribute set cardinality. In this way we constructed a new representation of the attributes. In the second step, we created a dissimilarity matrix taking the values from the co-occurrence matrix and subtracting from them the corresponding values of cosines between the new attribute representations.

Using the dissimilarity matrix we were able to perform the attribute cluster analysis. We did it using the hierarchical agglomerative approach with the Ward's linkage function [21]. The clustering results are depicted by the heat map in Figure 5. The darker spots correspond to pairs of similar attributes. They are likely to be exchangeable in the context of classification and thus can be interpreted as semantically related. In the plot, potential clusters of attributes are represented by dark squares aligned along the diagonal of the matrix. For example, in the first cluster there were attributes such as "type_of_building_standalone_compartment", "one_story_high", "single_family_houses" and "action_inside_buildings_at_ground_floor".

In the future we plan to extend this analysis by considering decision rules which may be constructed from the frequent attribute sets. Such rules may compose a useful tool for supporting ICs at a fire ground, which is the main task of our ICRA project [2].

V. Summation

In this paper we focused on introducing a web platform, called Knowledge Pit, created in order to support organization of data mining competitions. On the one hand, this platform is appealing to members of the machine learning community for whom competitive challenges can be a source of new interesting research topics. Solving real-life complex problems can also be an attractive addition to academic courses for students who are interested in practical data mining. On the other hand, setting up a publicly available competition can be seen as a form of outsourcing the task to the community. This can be highly beneficial to the organizers who define the challenge, since it is an inexpensive way to solve the problem which they are investigating. Moreover, an open data mining competition can become a bridge between domain experts and data analysts. In a longer perspective, it may leverage a cooperation between industry and academic researchers.

We also described *AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service* which was the first analytic challenge organized at Knowledge Pit. We presented the scope of this competition and briefly summarized its results. In addition, we discussed the results of our initial analysis of the best of the submitted solutions, highlighting their potential practical applications.

The conducted analysis is by no means complete. In future, the results of the competition will be thoroughly investigated by a team composed of experienced Incident Commanders and data mining experts. We hope that the results of this research, conducted as a part of a larger project ICRA [2], will have a noticeable impact on the fire safety domain. We also hope that our competition will revive a discussion on this topic among researchers with different backgrounds and expertise.

Acknowledgments

This work was partly supported by Polish National Science Centre (NCN) grants DEC-2011/01/B/ST6/03867 and DEC-2012/05/B/ST6/03215 and by National Centre for Research and Development (NCBiR) grant No. O ROB/0010/03/001 in the frame of Defence and Security Programmes and Projects: "Modern engineering tools for decision support for commanders of the State Fire Service of Poland during Fire&Rescue operations in the buildings".

References

[1] H. Johansson, *Decision Making in Fire Risk Management*. Dept. of Fire Safety Engineering, Lund University, 2001.

[2] A. Krasuski, A. Jankowski, A. Skowron, and D. Ślęzak, "From sensory data to decision making: A perspective on supporting a fire commander," *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, vol. 3, pp. 229–236, 2013.

[3] A. Krasuski and A. Janusz, "Semantic tagging of heterogeneous data: Labeling fire & rescue incidents with threats," in *FedCSIS*, 2013, pp. 77–82.

[4] K. Bąk, A. Krasuski, and M. Szczuka, "Searching for Concepts in Natural Language Part of Fire Service Reports," in *Concurrency Specificaton and Programming*, 2013.

[5] B. Gilbert, D. Nichols, B. Aisbett, M. Phillips, M. Sargeant *et al.*, "Fighting with fire: how bushfire suppression can impact on fire fighters' health," *Australian family physician*, vol. 36, no. 12, p. 994, 2007.

[6] M. Zaksek and J. L. Arvai, "Toward improved communication about wildland fire: mental models research to identify information needs for natural resource management," *Risk analysis*, vol. 24, no. 6, pp. 1503–1514, 2004.

[7] M. Wojnarski, A. Janusz, H. S. Nguyen, J. Bazan, C. Luo, Z. Chen, F. Hu, G. Wang, L. Guan, H. Luo, J. Gao, Y. Shen, V. Nikulin, T.-H. Huang, G. J. McLachlan, M. Bošnjak, and D. Gamberger, "RSCTC'2010 discovery challenge: Mining DNA microarray data for medical diagnosis and treatment," in *Proceedings of RSCTC'2010*, ser. LNAI, M. S. Szczuka et al., Ed., vol. 6086. Heidelberg: Springer, 2010, pp. 4–19.

[8] A. Janusz, H. S. Nguyen, D. Ślęzak, S. Stawicki, and A. Krasuski, "JRS'2012 Data Mining Competition: Topical Classification of Biomedical Research Papers," in *Proceedings of RSCTC'12*, ser. LNAI, J.T. Yao et al., Ed., vol. 7413. Springer, Heidelberg, 2012, pp. 417–426.

[9] M. Wojnarski, S. Stawicki, and P. Wojnarowski, "TunedIT.org: System for automated evaluation of algorithms in repeatable experiments," in *Proceedings of RSCTC'2010*, ser. LNAI, vol. 6086. Springer, 2010, pp. 20–29.

[10] J. Cole, *Using Moodle*, 1st ed. O'Reilly, 2005.

[11] Lee and B. Ware, *Open Source Development with LAMP: Using Linux, Apache, MySQL and PHP*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.

[12] E. Rosebrock and E. Filson, *Setting Up LAMP: Getting Linux, Apache, MySQL, and PHP Working Together*. Alameda, CA, USA: SYBEX Inc., 2004.

[13] P. C. Isaacson, "Building a simple website using open source software (gnu/linux, apache, mysql, and python)," *J. Comput. Sci. Coll.*, vol. 19, no. 1, pp. 286–288, Oct. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=948737.948777

[14] E. Plugge, T. Hawkins, and P. Membrey, *The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing*, 1st ed. Berkely, CA, USA: Apress, 2010.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[16] J. Wróblewski and S. Stawicki, "Sql-based kdd with infobright's rdbms: Attributes, reducts, trees," in *RSEISP*, ser. LNCS, M. Kryszkiewicz, C. Cornelis, D. Ciucci, J. Medina-Moreno, H. Motoda, and Z. W. Raś, Eds., vol. 8537. Springer, 2014, pp. 28–41.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[18] T. M. Mitchell, *Machine Learning*, ser. McGraw Hill series in computer science. McGraw-Hill, 1997.

[19] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo *et al.*, "Fast discovery of association rules." *Advances in knowledge discovery and data mining*, vol. 12, no. 1, pp. 307–328, 1996.

[20] A. Janusz and D. Ślęzak, "Rough set methods for attribute clustering and selection," *Applied Artificial Intelligence*, vol. 28, no. 3, pp. 220–242, march 2014.

[21] L. Kaufman, P. Rousseeuw, and E. Corporation, *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley Online Library, 1990, vol. 39.