AIM in Unsupervised Data Mining

Luis I. Lopera González, Adrian Derungs, and Oliver Amft

Contents

Abstract

This chapter explores the differences between association rules extracted using the likelihood mining criterion (LMC) and rules extracted by using frequent item-set rule mining (FRM). LMC provides a change in perspective for rule selection, from a measure of frequency in the dataset to a measure of relationship between the rule items. For illustration, this chapter presents the evaluation of qualitative differences between LMC and FRM rules with three examples: (1) a basic rule mining scenario to illustrate LMC properties, (2) an analysis relating socioeconomic information and chemical exposure data, and (3) mining behavior routines in patients undergoing neurological rehabilitation. Results show that LMC is

L. I. Lopera González (✉) · A. Derungs · O. Amft
Friedrich Alexander University Erlangen-Nuremberg, Erlangen, Germany
e-mail: luis.i.lopera@fau.de; adrian.derungs@fau.de; oliver.amft@fau.de

N. Lidströmer, H. Ashrafian (eds.), *Artificial Intelligence in Medicine*,
https://doi.org/10.1007/978-3-030-58080-3_300-1

capable of extracting rare rules and does not suffer from support dilution. Furthermore, LMC focuses on the individual event generating processes, while FRM focuses on their commonalities.

Keywords

Likelihood · Likelihood mining criterion · Min-support · Association rule mining · Frequency rule mining · Bayesian rule mining · Support dilution · Mining medical datasets · Routine mining

1 Introduction

Association rules can model a process by describing the relationship between variables. In a dynamic process, for example, a rule states that a change on an input will cause a change on an output. As the process evolution is stored in a dataset, association rule mining (ARM) can extract the original relationship between the process inputs and outputs. The common approach for ARM is frequent association rule mining (FRM). Rules extracted by FRM, e.g., $X \rightarrow Y$, have support and confidence greater than a user-specified min-support and min-confidence thresholds [1]. Formally, support of an item-set X is defined as the number of transactions T in the database D containing item-set X divided by the number of transactions in the database. Equation 1 shows the support of a rule $X \rightarrow Y$. The rule confidence is defined using support in Eq. 2.

$$Supp(X \rightarrow Y) = \frac{|\{(X \cup Y) \subseteq T_k, (X,Y) \neq \emptyset, \forall T \in D\}|}{|D|} \quad (1)$$

$$Conf(X \rightarrow Y) = \frac{Supp(X \rightarrow Y)}{Supp(X)} \quad (2)$$

The following thought experiment illustrates one of FRM's limitations. Suppose all supermarket receipts from the winter holidays are used for rule mining. One would expect to see rules that associate the ingredients used for winter holiday meals. Now consider a year worth of receipts from the same supermarket. In the light of the additional data, the winter holiday meal ingredients would not have enough support to be extracted. In other words, the minimum support (min-support) threshold used to extract rules in a small dataset will not work in an extended version of the dataset due to the definition of rule support. In this chapter, min-support's dependency on the dataset size is called support dilution.

A dataset can be viewed as the collection of items sampled from multiple generating processes. However, FRM has the implicit assumption that all items are generated at the same rate. In practice, processes can generate items at different rates. For example, people in Germany occasionally buy white sausages, but when they do, they buy wheat beer too. So the rule "if white sausage then wheat beer" is a rare rule when compared to frequent rules, e.g., "if milk then eggs." FRM can extract rare rules by using a low min-support threshold. Unfortunately, spurious item associations may create unwanted rules that FRM's threshold cannot eliminate. Filtering out unwanted rules for FRM has been addressed in the past [2, 3]. However, the processing required to separate rules does not generalize [4].

The field of medicine advances by capturing evidence that supports a given hypothesis. As the means to collect data surpasses the current capacity to analyze it, automation is required to extract new knowledge. ARM as a tool provides investigators the ability to sift through data repositories automatically. However, FRM methods are inadequate to extract useful knowledge from medical repositories, if the interest is association quality and not how frequent a certain observation occurs in a dataset. Therefore, this chapter reviews the use of likelihood as a means to select rules with quality measured independently of the dataset size. In particular, the minimum likelihood mining criteria (LMC) is explored and compared to min-support FRM. LMC has been proposed to replace min-support as primary rule selection criteria in ARM [5].

This chapter is organized as follows: Sect. 2 presents an overview of ARM algorithms. Sect. 3 describes LMC and illustrates rule selection differences between FRM and LMC by extracting all

atomic rules, i.e., rules of the form $X \rightarrow Y$, where $|X| = |Y| = 1$, from five datasets commonly used in ARM literature. Sect. 4 through 6 illustrate LMC and FRM in three ARM applications: a synthetic timeseries, a dataset linking socioeconomic variables with health-related chemical exposure information, and a dataset of daily behavior routine annotations of patients with hemiparesis. In each setting LMC and FRM extracted rules are compared and evaluated for quality and usefulness.

2 Association Rule Mining

ARM algorithms can be generalized to have two stages: (1) search for useful item-sets, and (2) search for adequate rules within these item-sets [6–8]. In the context of ARM, useful item-sets are categorized as passing a min-support threshold and adequate rules depended on a secondary rule selection criterion. As the problem of support dilution became apparent [9], many approaches looked to circumvent the min-support threshold or to define new rule interest metrics, but always relied on support properties to search the lattice of rule candidates. In this section, algorithms that rely on support to select rules are grouped under FRM methods.

Conceptually, there is no correlation between rule frequency and rule interest, as pointed out by Li et al. [10]. Therefore, alternative rule interest metrics have been proposed, inspired by probability theory, where rules are created based on the relationship between the rule and its conforming items. In this section, algorithms that use probability theory to select rules are grouped under Bayesian rule mining (BRM) due to their use of Bayesian concepts to select interesting rules. LMC is considered a primary rule selection criteria class under BRM. Figure 1 illustrates a summarized taxonomy of ARM algorithms, by mining methodology, primary rule selection criteria, and search type.

The rest of this section provides a brief overview of ARM literature grouped into FRM- and BRM-based approaches.

2.1 FRM

High-utility item-set mining is the general approach to extracting useful item-sets. The idea behind high-utility item-set mining is that a utility is given to each item in a transaction. Then, high-utility item-sets are extracted by summing the utility of each item-set across all transactions and comparing them to a min-utility threshold [11]. The advantage of high-utility item-set mining is that items bought occasionally, e.g., white sausages and wheat bear, usually have high utility and are extracted. It has been shown that frequent item-set mining is a special case of high-utility item-set mining, where the transaction utility per item is one and min-utility threshold becomes min-support [12]. Nguyen et al. [13] illustrate multiple approaches to association rule mining using high-utility item-set mining. The challenge with high-utility mining is that the utility of an item is not always available as in the supermarket case. Therefore, the chapter's scope is limited to compare ARM based on LMC with FRM without utility.

In general, the two-stage approach to ARM requires min-support as primary rule selection criterion, and a secondary selection criterion such as min-confidence threshold. Seminal work in ARM, like the Apriori [1] or ECLAT [14] algorithms, introduced the min-confidence threshold as the main mechanism for creating association rules. As ARM was used to solve real-world problems, the threshold on confidence was producing rules with no prediction power [15]. Therefore, several interest metrics were introduced, such as lift [16], conviction [17], relative confidence [18], information gain [19], and others [20]. Alternatively, rule grammars and machine learning approaches have been presented as alternatives of the downward-closure property of support to restrict the rule search space, and extract more interesting rules. For example, Padillo et al. [21] used rule grammar and map reduce to optimize the mining process in large datasets. In timeseries, Guillam-Bert et al. [22] proposed the TITARl algorithm. They used decision trees to improve rule description and temporal specificity. However, these interest metrics were used as

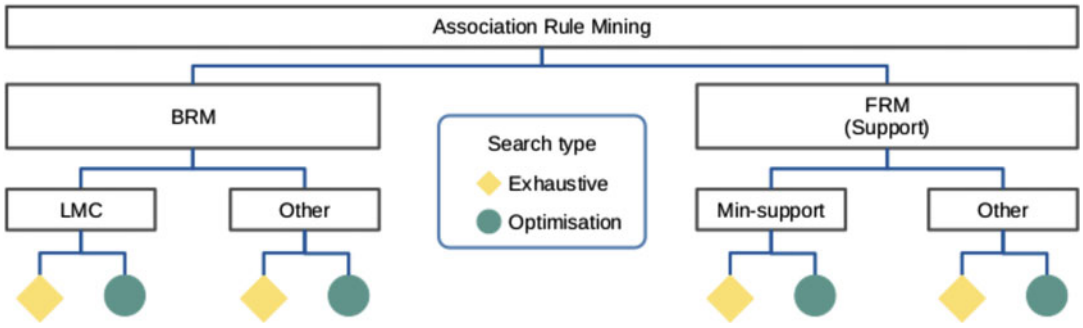

Fig. 1 Extended taxonomy of association rule mining. FRM and BRM rule mining methods represent different concepts related to the rule selection process. FRM employs support properties while the yet less established BRM uses probabilities. In this chapter, LMC and min-support FRM are compared. LMC is considered a primary rule selection criteria class under BRM

secondary rule selection criterion, and depended on item-sets extracted using a min-support threshold.

Rare rules are defined as the rules with support between a min-rare-support and min-support thresholds [23], where min-rare-support $<$ min-support. In the rare-rule community, the available work focuses on resource-efficient extraction. For example, Tsang et al. [24] proposed the RP-Tree structure to facilitate rare rule extraction. Liu and Pan [3] proposed RP-VRIM, an extension RP-Tree that uses vertical layout to reduce dataset scans. Borah and Nath [9] proposed the SSP-Tree that enables the search of frequent and rare pattern combinations simultaneously and considers dynamic datasets.

Selecting the correct value for min-support is difficult. Thus, several approaches proposed rule search mechanisms that omit the min-support threshold. For example, *OPUS* [25] is a frequent item-set exhaustive search algorithms that does not use the downward-closure property to traverse the item lattice. Webb [26] presented an extension that converted *OPUS* into an ARM algorithm. Bashir et al. [6] proposed an exhaustive search algorithm for frequent item-set mining, which starts by selecting n item-sets. Then, their algorithm selects the smallest support values of the chosen item-sets, and prunes the search space using the downward-closure property. Fournier-Viget and Tseng [27] proposed the TNR algorithm where the top n nonredundant rules were extracted. Both TNR and Bashir et al.'s algorithms required min-confidence to be specified.

Support dilution is a common ARM problem in dynamic databases. Cheung et al. proposed FUP [28], the first algorithm to consider efficiently updating extracted knowledge after adding new transactions to a database. Tobji and Gouider [29] extended FUP for different user-given support thresholds after adding transactions. Aqra et al. [30] proposed the Aprior algorithm to improve rule maintenance under append, update, and remove operations over the dataset. All these methods have in common that the min-support threshold needs to be manually updated in order to maintain interesting rules.

Although Tian et al. [31] proposed the use of probabilistic metrics of rule interest based on Bayes theorem, their methodology is based on a frequent item-set extraction, and proposes the Bayesian confidence and Bayesian lift as secondary rule selection criteria.

2.2 BRM

Bayesian methods for ARM are still less frequently investigated, nevertheless offer elegant features and complementary properties to FRM. The following review summarizes key contributions to BRM.

Li et al. [10] introduced local support as primary rule selection criteria in medical datasets. Local support is an approximation of the

likelihood probability based on data observations. Gay and Boullé [32] have used Bayes and Information theory to select rules with the best classification power. Their proposed *level* sets a boundary for extracting interesting rules, with a preference for simpler models, i.e., shorter rules. However, the *level* does not have a monotonic or anti-monotonic property that can be exploited to minimize the search space. Therefore, their methodology looks for locally optimal rules. In contrast to *level*, LMC has the anti-monotonic property [5], which reduces rule search space. Lopera G. et al. [33, 34] used increasing belief in the recursive application of Bayes theorem as rule selection criterion. Increasing belief can be simplified to a threshold on the likelihood. LMC differs from increasing belief by using a fixed minimum likelihood rather than the rule's premise probability as threshold value.

When LMC is used to replace min-support in FRM algorithms, existing interest metrics can be used as secondary rule extraction criteria, including Bayesian confidence and lift. As LMC extracts rules from the entire support range, min-support and min-rare-support thresholds can be defined after mining rules to label LMC rules as rare, e.g., following the FRM definition of rare rules. Although LMC requires data representations like SSP-Tree to maintain item-set counts, in dynamic datasets, LMC does not require the maintenance of thresholds between dataset updates. Unlike min-support, LMC does not depend on the dataset size.

3 Likelihood Mining Criterion (LMC)

When considering data supporting a hypothesis, Bayes' theorem is a useful tool to measure the change in belief when new data becomes available. In general, Bayes' theorem states the relationship between posterior and likelihood scaled by the prior of the data and the hypothesis. The posterior is defined as the probability of the hypothesis H being true given the data D. The likelihood is the probability of observing D given that H is true. The priors are the respective probabilities of D and H. Equation 3 illustrates the relationship.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \tag{3}$$

The ARM task can be restated as finding the association such that an item-set D provides evidence to an item-set H in the form $D \rightarrow H$. Using this new formulation, and approximating probabilities form data, it follows that confidence is equivalent to the posterior $P(H\backslash D)$, and local support is equivalent to the likelihood $P(D\backslash H)$. In this chapter, LMC is designed to find atomic rules when the likelihood is at least 50%, describing associations that are better than random choice for the item in H. Equation 4 illustrates LMC for a rule of the form $a \rightarrow b$, where a and b are items in the dataset.

$$\text{LMC} : P(a|b) \geq 0.5. \tag{4}$$

3.1 LMC–FRM Comparison

Difference between FRM and LMC mined rules were evaluated using five datasets commonly used in ARM literature: the Chess, Accidents, Retail, Mushrooms, and T40I10D100K available at http://fimi.ua.ac.be/data/. A miner extracted all possible atomic rules from each dataset, and illustrates which rules pass FRM and LMC criteria. To extract all atomic rules, the miner scanned each dataset once and recorded all pairwise combinations, keeping counts of rule appearances, and individual item appearances. For example, given a transaction $T_k = \{a,b\}$, atomic rules $a \rightarrow b$ and $b \rightarrow a$ were extracted. Subsequently, rules were selected by applying LMC after each dataset was scanned. Any possible rule in the dataset follows that $Conf(r) \geq Supp(r)$, as the $Supp(X) \leq |D|$ generating a triangular rule region in the support/confidence plane. Figure 2 illustrates the triangular rule region with the boundary support = confidence. Furthermore, Fig. 2 illustrates the FRM region delimited by the min-support,

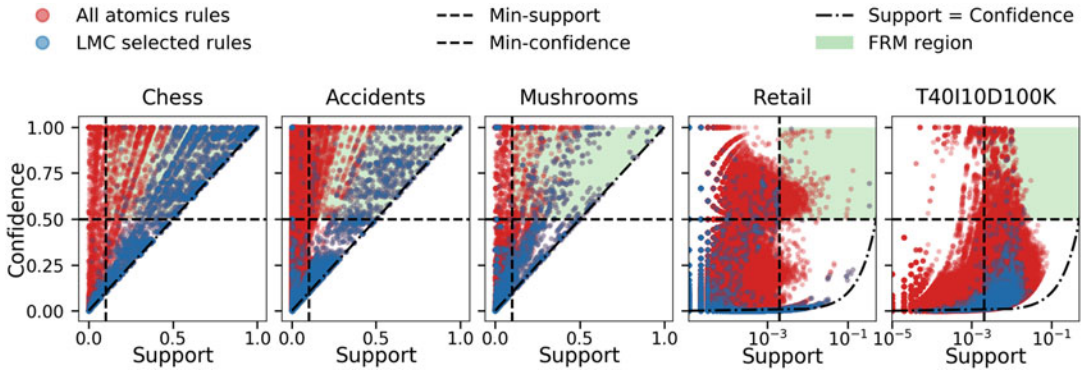

Fig. 2 Distribution of all possible atomic rules in each dataset. Rules in the FRM region pass the min-support and min-confidence thresholds. LMC rules are highlighted, and depending on the dataset, LMC rules can come from any location on the triangular rule region. Log scale is used to display the support axis for the Retail and T40I10D100K datasets

min-confidence thresholds, using the following values for best visualization: min-confidence = 0.5, min-support = 0.1, and min-support = 0.002 for the Retail and T40I10D100K datasets. Fig. 2 also shows all the atomic rules available in each dataset, and highlights rules that pass LMC. Depending on the dataset, LMC can extract rules from any part of the triangular rule region, where FRM is bound to the rules made available by the min-support and min-confidence thresholds. Although all LMC rules might not be of use, the exploration of space provides the user with information about the application and helps to guide the knowledge extraction process, and recorded all pairwise combinations, keeping counts of rule appearances, and individual item appearances. For example, given a transaction $T_k = \{a, b\}$, atomic rules $a \rightarrow b$and $b \rightarrow a$were extracted. Subsequently, rules were selected by applying LMC after each dataset was scanned. Any possible rule in the dataset follows that $\text{Conf}(r) \geq \text{Supp}(r)$,as the $\text{Supp}(X) \leq |D|$ generating a triangular rule region in the support/confidence plane. Figure 2 illustrates the triangular rule region with the boundary support = confidence. Furthermore, Fig. 2 illustrates the FRM region delimited by the min-support, min-confidence thresholds, using the following values for best visualization: min-confidence = 0.5, min-support = 0.1, and min-support = 0.002 for the Retail and T40I10D100K datasets. Figure 2 also shows all the atomic rules available in each dataset, and highlights rules that pass LMC. Depending on the dataset, LMC can extract rules from any part of the triangular rule region, where FRM is bound to the rules made available by the min-support and min-confidence thresholds. Although all LMC rules might not be of use, the exploration of space provides the user with information about the application and helps to guide the knowledge extraction process.

4 Basic Rule Mining Example

This example compares LMC and FRM miners. The goal is to cluster items from a synthetic timeseries according to their generating process, including the extraction of rare rules.

4.1 Methodology

A timeseries generator simulates processes emitting common and rare items. Using an observation window of fixed size, two miners extracted all atomic rules created by pairing the first item of the observation window with all remaining items. Then, each miner is applied a primary and secondary rule selection criteria and graphs are created using the resulting rules. When independent subgraphs formed, each subgraph was considered

an item cluster that represented a generating process. For primary rule selection criteria one miner used LMC. For comparison, the FRM miner used a min-support threshold that was set using heuristics to 0.1. The selected heuristic assumed all items were equally likely to appear, i.e., 1/7, and rounding down to one decimal point.

The timeseries generator mixed two processes: (1) a common process $p_r(t)$ that frequently emitted random items and (2) a rare process $p_c(t)$ that occasionally emitted a specific pattern of items denoted as a chain. The process $p_r(t)$ sampled vocabulary $V_r = 0,1,2,3$ using a uniform distribution. In contrast, $p_c(t)$ used vocabulary $V_c = 10,11,12$ to emit the chain $10 \rightarrow 11 \rightarrow 12$. Chain items were always emitted in the same order, but the timing between items varied uniformly, sampled from the integer interval [1, 10] items. The timeseries generator filled the gaps between $p_c(t)$ emissions with items from $p_r(t)$, resulting in a dense timeseries. Additionally, the timeseries generator used 1000 sampled items from $p_r(t)$ and 20 chains from $p_c(t)$. $p_c(t)$ chains were uniformly distributed throughout the timeseries and could not overlap. Equation 5 shows an excerpt of a generated timeseries *ts,* with the items emitted from $p_c(t)$ highlighted.

$$ts = [\bullet\bullet\bullet, 0, 2, 3, \mathbf{10}, 2, 0, 0, \mathbf{11}, 2, 2, 1, 1, \mathbf{12}, 2, 2, 2, \bullet\bullet\bullet] \quad (5)$$

Following FRM's two-step process for rule mining, to improve the subgraph separation, the FRM miner used the following secondary rule selection criteria: (1) a min-confidence threshold of 0.5, matching the threshold on *P(a\|b)* defined in LMC, (2) a selection criterion based on the Bayesian factor, and (3) selecting rules with the highest confidence for each conclusion. Except for the min-confidence threshold, the secondary rule selection criteria were chosen to avoid additional thresholds. Equation 2 shows the estimation of rule confidence where $X = \{a\}$ and $Y = \{b\}$ confidence. The Bayesian factor was estimated using Eq. 6, where the rule $a \rightarrow b'$ denotes atomic rules in the miners final rule set A that have premise a and items other than b as conclusion.

$$\frac{P(b|a)}{P(b'|a)} = \frac{Supp(r)}{Supp(a \rightarrow b')}$$
$$Supp(a \rightarrow b') = \sum_{\forall a \rightarrow x \in \Lambda x \neq b} Supp(a \rightarrow x) \quad (6)$$

4.2 Evaluation

In accordance to the generator's characteristics, performance was evaluated by grouping extracted rules into the following categories: (1) R_r contained all possible atomic rules that use V_r items, ($|R_r|$: $|V_r|^2 = 16$) (2) R_c contained all atomic, time-ordered, decompositions of the chain $10 \rightarrow 11 \rightarrow 12$, i.e., $10 \rightarrow 11$, and $11 \rightarrow 12$, ($|R_c|$: 2) (3) R_{rc} contained atomic rules of the form $i \rightarrow j$, where $i \in V_r$ and $j \in V_c$, ($|R_{rc}|$: $|V_r| * |V_c| = 12$) (4) R_{cr} contained atomic rules of the form $j \rightarrow i$, where $i \in V_r$ and $j \in V_c$, ($|R_{cr}|$: $|V_r| * |V_c| = 12$), and (5) R_{cv} contained atomic rules which were created from all possible pairwise combination of V_c items and are not in R_c, ($|R_{cv}|$: $|V_c|^2 - 2 = 7$).

Process separation occurred when miners only extracted rules from the categories R_r, R_c, and R_{cv}, as no rules bind items from the two generating processes. Rules in R_{cv} were not generated by $p_c(t)$. Thus, they are considered a separate category. Rules in R_{cr} and R_{rc} connect the items from $p_c(t)$ and $p_r(t)$ and no process separation is possible. R_{cr} and R_{rc} were defined as independent categories to evaluate LMC's effect on item association between frequent and rare items, when considering their position in the rule. The mining performance metric quantified the extraction rate for a rule category R with size |R| as shown in Eq. 7, where A is the mined rule set.

$$\text{Extraction rate} = \frac{|\forall r \in A \cap R|}{|R|} * 100[\%] \quad (7)$$

A parametric search looked for observation window sizes in the range between [2, 500] items in one item increments. One hundred timeseries were generated for each observation window size. The following evaluation steps used a window size that minimized the chances of extracting rules from categories R_{rc}, R_cr, and

R_{cv}. The selected window size also ensured that the miners always mined all rules from the R_c and R_r categories.The secondary rule selection methods required the generation of a new batch of one hundred timeseries. Each miner processed the new timeseries and produce results using the secondary rule selection criteria. The performance evaluation measured the average number of times items were correctly separated into generating processes $p_r(t)$ and $p_c(t)$, respectively.

The evaluation found that the FRM miner could not retrieve the items from V_c as their support was around 0.01. Therefore, any atomic rule from $p_c(t)$ will also not pass the min-support threshold of 0.1 due to support's downward-closure property. In addition, sampling extra items from $p_r(t)$ caused $p_c(t)$ chain's support to dilute. In contrast, LMC does not depend on the number of items in the timeseries. Therefore, LMC always retrieved *the* $p_c(t)$ chain.

Figure 3 illustrates how the rule categories were extracted as a function of the observation window size. Using an observation window larger than the expected item timing of $p_c(t)$ of five samples, LMC always extracted all rules from the R_r and R_c categories, which are needed to separate items according to their generating processes. Rules in R_{cv} were extracted when at least two partial chains were seen by the observation window. An observation window size in the range

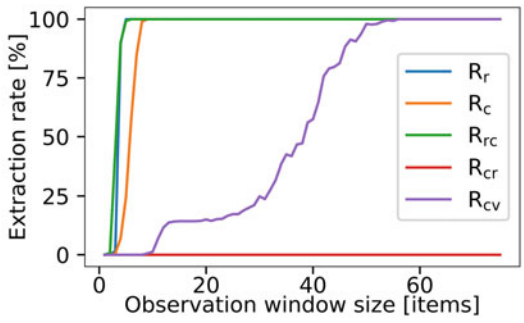

Fig. 3 Search results of observation window sizes for all atomic rule categories. LMC extracts all rules from the R_r, R_{rc}, and R_c categories when the observation window size is greater than the expected value of $p_c(t)$'*s* item timing. As the observation window grows, LMC extracts rules from R_{cv}, which are innocuous for the generating process separation task

[5, 8] prevents the extraction of R_{cv} rules. Rules from R_{rc} meet LMC because the rule support is similar to the conclusion support. Thus, $P(X|Y) \approx 1 > 0.5$ and R_{rc} rules are always extracted. In contrast, R_{cr} will never pass LMC and are ignored because the rule support is smaller than the conclusion support, $\approx 20/250$, which illustrates that under LMC, rare premises cannot associate with frequent conclusions. In the LMC miner, the generating process separation task needed to remove R_{rc} rules from the final set A using a secondary rule selection criterion. The min-confidence threshold always correctly separated the items into generating processes. The Bayesian factor only selected $p_c(t)$ rules and no items from $p_r(t)$ are grouped. Finally, the best confidence per conclusion criterion failed to separate items into two generating processes, because $p_c(t)$ item 10 is always associated as conclusion with $ap_r(t)$ item.

5 Census and Chemical Exposure Database Mining

The following example compares LMC and FRM miners in a database mining task. The database is the publicly available dataset from Huang et al. [35]. The dataset comprises US census tract information from the American Community Survey (ACS) 5-year summary files for the 2010–2014 period. Moreover, the dataset contains health-related chemical exposure data generated from the 2011 National-Scale Air Toxics Assessment (NATA), specifically air pollutant exposure concentration.

5.1 Methodology

Huang et al. reported results for two mining scenarios: (1) mining rules using the socioeconomic variables as premises and chemical exposure variables as conclusions (S→C), and (2) mining rules within the socioeconomic dataset (S→S). Huang etal. Used min-support of 0.1 and lift ≥ 1 as thresholds for FRM-based rule selection. Here their atomic rule findings are replicated with an exhaustive search, min-support, and lift FRM

algorithm and compared the extracted rules to LMC results. Lift is defined in Eq. 8

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Supp}(X \rightarrow Y)}{\text{Supp}(X) \times \text{Supp}(Y)} \quad (8)$$

Huang et al. categorized variables as follows: socioeconomic scores were divided into deciles, chemical variables into quartiles, and age group used the ranges: (0–20), (20–30), (30–35), (35–38), (38–40), (40–50), and (50–150). The poverty score was computed as the percentage of population per track that had a ratio between income and poverty level below 1.5. Additionally, deciles seven through ten were combined. Education score was calculated as the population percentage per track that had demonstrated education beyond high school level. Deciles eight through ten of the education score were merged. Finally, the race score was calculated as the percentage of non-White population per track.

5.2 Evaluation

Huang et al. [35] analysis provided a relevance criterion to interpret the extracted rules within their field. Therefore, the evaluation goal was to measure how many of Huang et al.'s criteria were extracted by the LMC miner.

For any newfound rule, the odds ratio (OR) was computed using a 95% confidence interval (CI). The CI was calculated using 10000 runs of bootstrapping, following Huang et al. [35]'s evaluation of rule relevance.Table 1 lists the mined rules from Huang et al. [35] S → C scenario, with their respective support and lift. The LMC miner extracted only five rules from Huang et al.'s 13 rules.

Table 2 lists mined association rules in the S → S scenario. LMC found two out of the six rules reported by Huang et al. Additionally, LMC found three rules, highlighted in Table 2, that did not pass the min-support and lift criteria from Huang et al.

Table 3 shows the odds ratio (OR) and estimated 95% confidence interval (CI) for rules exclusively extracted by LMC in the S → S scenario. The OR analysis showed that the new rules had higher OR than the rules from Huang et al., whose OR ranged from 1.75–3.56. Rules that passed LMC had the largest OR in both mining scenarios. Thus, LMC rules are more likely to appear in a repeat experiment, and therefore LMC rules may be deemed more desirable.

LMC rules provide meaningful insight, in particular on rarely, but consistently occurring relations, which may provide application experts new hypotheses to investigate. For example, as seen in the S → S scenario of the database mining examples, LMC provided additional rules that hint to a predominantly White ageing population (Race score = 1 → Age group 40–150), and to a correlation between low poverty score and high education levels (poverty score = 1 → education score = 7 or 8).

6 Rehabilitation Routine Mining

The following example illustrates how LMC can be used to interpret patient behavior during stays at a day care rehabilitation center. LMC and FRM miners try to classify patients into physically active and sedentary groups. However, results show that FRM represents the cohort's average behavior and thus fails to assign patients to groups.

6.1 Methodology

The rehabilitation routine mining examples used activity labels from the longitudinal stroke rehabilitation study of Derungs et al. [36]. The study was approved of the Swiss Cantonal Ethics Committee of the canton Aargau, Switzerland (Application number: 2013/009). There were 11 patients in the study, aged 34–75 years, among them five female and four used a wheelchair. In addition, data from a patient excluded from the original rehabilitation study [36] was added to the dataset, for a total of 12 patients.

Patients visited the day care center for approximately 3 days per week over 3 months to

Table 1 LMC's extraction of socioeconomic and chemical exposure association rules (S → C). LMC mined five rules out of the 13 extracted using FRM. LMC rules had the highest lift

Rules	Support	Lift
Race score = 1 → Diesel = Q1	0.144	1.783
Race score = 1 → Butadiene = Q1	0.142	1.805
Race score = 1 → Toluene = Q1	0.138	1.746
Race score = 1 → Benzene = Q1	0.130	1.653
Race score = 1 → Acetaldehyde = Q1	0.126	1.596

Table 2 LMC extracted five rules in the S → S scenario. Two rules correspond to the FRM mined rules with highest lift, and three rules (in bold) did not pass Huang et al. [35] min-support threshold

Rules	Support	Lift
Age group = 40–50 → Race score = 1	0.172	1.585
Race score = 1 → Age group = 40–50	0.172	1.585
Poverty score = 1 → Education score = 8	0.038	3.919
Poverty score = 1 → Education score = 7	0.034	3.210
Race score = 1 → Age group = 50–150	0.015	2.227

participate in individual and group training sessions, socialize with others, and follow personal activity preferences. Some training sessions available to patients were physiotherapy, ergotherapy, and training in the gym. Patients performed activities of daily living, including walking, eating and drinking, setting the table, writing, and making coffee. Behavior of each patient was recorded for up to 8 h on 10 days at the center by two observers accompanying patients. In addition, body motion was recorded using inertial sensors attached to wrists, upper arms, and tight positions. During the observation time, the examiners annotated patient activities using a customized annotation tool on a smartphone, resulting in a total of 16,226 activity labels. Therapists scored patients for their ability to execute activities of daily living independently using the Extended Barthel Index (EBI) [37]. The EBI consists of 16 categories. Each category receives a score within the range zero to four, where zero means that the patient requires full support, and four means the patient can live independently.

Miners used the start of activity labels as timestamped items and a 20-minute observation window to create candidate rules by pairing the first item of the observation window with remaining items. The miners extracted atomic rules using their respective primary selection criterion. The same secondary criterion was used to filter the resulting rule sets and the remaining rules were assembled into graphs. Each resulting independent subgraph was considered a routine and a study observer assigned a label. The FRM miner's min-support threshold was set to 0.0038 with the goal of selecting the same number of rules as the LMC miner.

6.2 Evaluation

Miners were evaluated by submitting their extracted rules to the same post-processing two stage procedure: (1) secondary rule selection criterion and (2) graph-based routine classification. In the secondary rule selection stage, three methods were evaluated: Bayesian factor (Eq. 6), min-confidence threshold of 0.5, and best confidence per conclusion. With the retained rules a graph was constructed and routines extracted as independent subgraphs. For each mining method, the goal was to find the secondary rule selection criterion that provided a balance between activity label count per graph and the number of independent graphs.

A patient's contribution to the resulting routines was analyzed using a patient exclusion process (PEP), where a patient's data was removed

Table 3 Odds ratio (OR) and confidence interval (CI) for LMC extracted rules in the S → S scenario. The 95% CI was estimated using 10,000 bootstrapping iterations

Rule	OR	Est. 95%	CI
Poverty score = 1 → Education score = 8	11.18	10.49	11.94
Poverty score = 1 → Education score = 7	6.74	6.35	7.17
Race score = 1 → Age group = 50–150	5.68	5.10	6.39

from the dataset and resulting routines were compared with routines mined using all patients.

Based on the type of the majority of activities in the routine, a study observer named LMC routines as socializing, eating, using the phone, and intense and balance training. Whereas, FRM routines were named mobility, eating, and cognitive-motor training. Preliminary results showed that FRM routines lacked emphasis on activities related to socializing.

The *Primary Criteria* column in Fig. 4 shows the resulting graphs based on atomic rules extracted by LMC and FRM miners. Activities in both graphs are hyperconnected, i.e., multiple edges connect activities. However, for FRM, there are nodes with one edge. FRM rules do not describe the flow from one activity to another, but rather, the associations of repeating events, e.g., repetitions of an exercise. In contrast, LMC looks for successive activities, and the respective low count of activity transitions vs exercise repetitions does not affect the rule selection. For both mining algorithms, the hyperconnected graph yielded no useful routine information.

With a Bayesian factor ≥ 1, LMC mined rules focus mostly on self-referencing activities, e.g., walking → walking, resulting in single activity subgraphs. In contrast, the Bayesian factor criterion removed most of the FRM rules. The resulting subgraphs had too few activities to consider them as routines. For FRM rules, the confidence secondary rule selection criterion reduced the graph size, but it was unable to create independent subgraphs. However, with LMC rules, the confidence secondary rule selection criterion selected many self-referencing rules creating two more subgraphs than the Bayesian factor, containing four activities each. Nevertheless, there were too many single activity subgraphs to consider the split as routines. The best balance was obtained between the number of subgraphs and activities per graphs using the best confidence per conclusion criterion. After secondary rule selection, LMC-mined rules yielded five routines, whereas FRM-mined rules yielded only three. Figure 4 illustrates the resulting subgraphs for each mining algorithm and secondary rule selection criteria.

Figure 5 illustrates an example of the changes in routine graphs when removing patients with active and sedentary behavior. For both mining algorithms, when removing one patient, the routine's activity composition varied, but the assignment of routine labels by the study observer did not vary. In the PEP analysis, using the best confidence per conclusion secondary criterion, LMC mined routines that grouped patients into physically active and sedentary groups, as indicated by the study observer. Removing patient ID 10 made the routine *using a phone* disappear. Apparently, patient 10 received calls while playing with the phone. The physically active group contained patient IDs 2,4,6,9, and 10. The physically active group refers to patients following their rehabilitation schedule closely. No relation to activity intensity or EBI score was found. The active patient group consisted of one wheelchair rider, patients with different EBI starting points, and some patients, where the EBI score did not change. When a patient from the active group was removed, the extracted routine number reduced to four and *socializing* was always missing. The result appears counterintuitive, as the sedentary group has been likely involved in socializing, but could be explained by the chosen 20-minute observation window, which causes LMC to focus on transitions between activities of at most 20 min duration. Sedentary patients would perform individual activities for periods longer than 20 min.

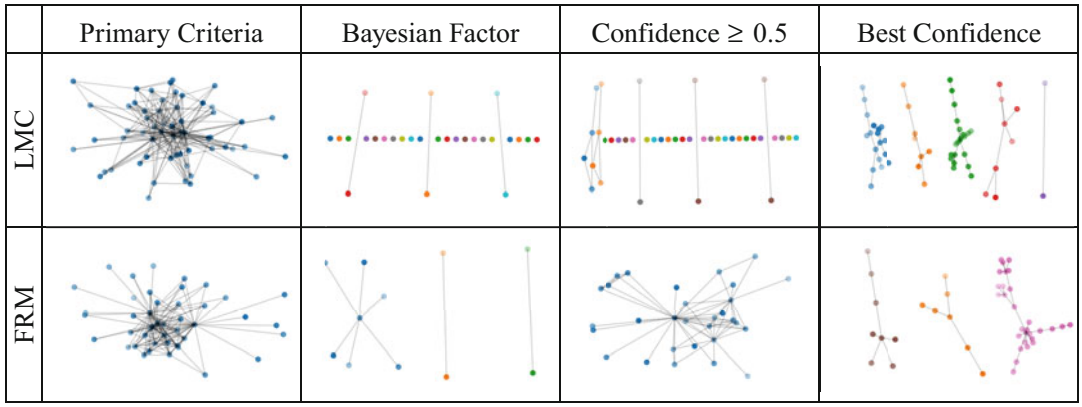

Fig. 4 Graphs constructed using rules derived by LMC, FRM, and different secondary rule selection criterion. Without a secondary rule selection criterion, both LMC and FRM produce a single graph and no useful routine information is extracted. A balance between activity count per graph and number of independent graphs was achieved using the best confidence per conclusion secondary rule selection criterion

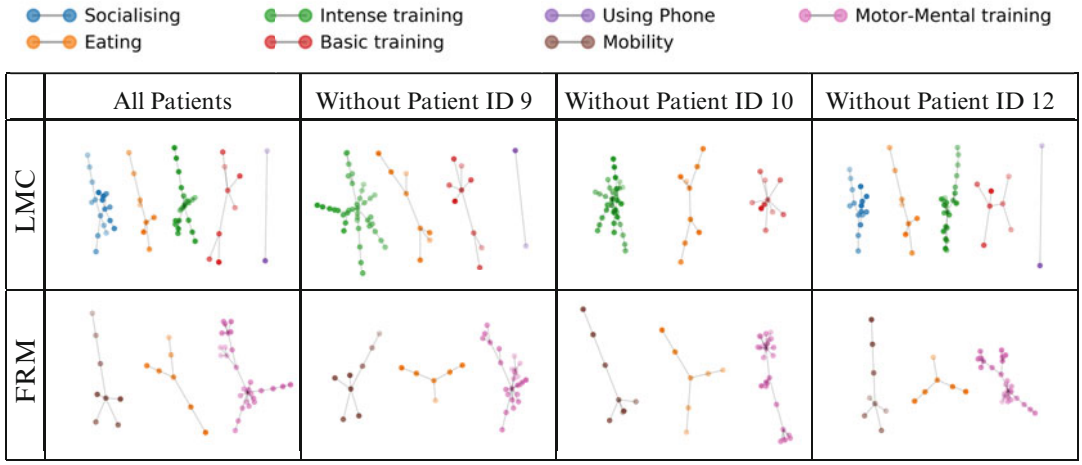

Fig. 5 Routine graphs when patients are removed from the dataset, i.e., PEP method. LMC's focus on rare rules highlights the importance of each individual patient's contribution to the graph representation. With PEP analysis, the LMC miner was able to separate patients into active and sedentary behavior groups. In contrast, the number of FRM-mined routines did not change under PEP analysis and no further insight was derived

Therefore, their socializing activities would not be associated into rules.

For comparison, PEP analysis for FRM-extracted routines using rules with the best confidence per conclusion found that the removal of any one patient did not affect the extracted routines. Therefore, FRM provided no further insight to classify patients into active or sedentary groups.

The rehabilitation routine mining examples illustrated the difference between both association rule mining criteria. Routines mined with FRM did not change during PEP analysis. FRM mined routines that were common to the entire population. With LMC, the routines changed during PEP analysis, grouping patients into active and sedentary groups. Hence, FRM answers the question: Which routines are common among patients?, and LMC answers the question: What types of patients are there?

7 Conclusions

FRM algorithms can be converted to use LMC by simply replacing the min-support threshold. As most algorithms exploit the support's closure property, and LMC has the same computational complexity as the calculation of support or confidence, there is no complexity penalty on any migrated algorithm. Albeit, the extracted rules will be different.

One limitation of LMC is the sporadic association of frequent premises with infrequent conclusions into irrelevant rules. A secondary rule selection criterion can help remove irrelevant rules. However, the secondary rule selection criterion of choice depends on the application. For example, in the basic rule mining example a confidence threshold was used to separate items into generating processes, best rule confidence per conclusion worked for the rehabilitation routine mining task, and for the database mining example, no secondary rule selection criterion was needed.

Part of the motivation to introduce secondary rule selection criteria, other than confidence, is that association rules should provide some predictive power [15]. For LMC, rules have a predictive power of at least 50%. In other words, items in the premise of an LMC rule are colocated with items in the conclusion in at least 50% of the transactions. Although LMC was chosen to have better predictive performance than random chance for atomic rules, the 50% threshold used by LMC might be too restrictive. Lower threshold values might be necessary when considering categorical variables in the conclusion or conjunctive rules.

BRM, and in particular LMC, are not a replacement for FRM. The application should drive the choice of algorithm. For example, suppose a dataset contains symptoms, health-related behaviors, and disease outcomes. FRM is better suited to answer questions like "Which behaviours are most conducive to sickness?". Whereas BRM is better suited to answer questions including "Which symptoms and behaviours correlate to a specific disease?". The difference between ARM branches is summarized as follows: FRM focuses on extracting rules that describe commonalities between generating processes. In contrast, BRM looks for rules that describe each process.

This chapter showed that LMC does not suffer from support dilution, and that LMC is capable of extracting rare rules. The basic rule mining and socioeconomic examples illustrated how LMC extracted frequent and rare rules. In the rehabilitation routine mining examples, LMC was used to mine rules, create routines, and group patients into active and sedentary groups. Only LMC rules provided patient grouping information. As new medical datasets are investigated, LMC is a powerful tool when considering ARM for knowledge extraction.

Acknowledgments The authors are thankful for the permission to utilize the datasets used for illustration in this chapter.

References

1. Agrawal R, Imieliński T, Swami A. Mining Association Rules Between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. SIGMOD '93. ACM; 1993. p. 207–216. Available from: https://doi.org/10.1145/170035.170072.
2. Lopera Gonzalez LI, Amft O. Mining Hierarchical Relations in Building Management Variables. Pervasive and Mobile Computing. 2016;26:91–101. Available from: http://www.sciencedirect.com/science/article/pii/S1574119215001935.
3. Liu S, Pan H. Rare itemsets mining algorithm based on RP-Tree and Spark framework. AIP Conf Proc. 1967 (1):040070. https://doi.org/10.1063/1.5039144.
4. Grabot B. Rule mining in maintenance: analysing large knowledge bases. Comp Indust Eng. 2018; 139:1–15. Available from: https://hal.archives-ouvertes.fr/hal-02134705
5. Li J, Fu AWC, Fahey P. Efficient discovery of risk patterns in medical data. 2009;45(1):77–89. Available from: https://www.sciencedirect.com/science/article/pii/S0933365708000900.
6. Bashir S, Jan Z, Baig AR. Fast algorithms for mining interesting frequent itemsets without minimum support. 2009, Available from: http://arxiv.org/abs/0904.3319.
7. Djenouri Y, Djenouri D, Belhadi A, Fournier-Viger P, Lin JCW. A new framework for metaheuristic-based frequent itemset mining. Appl Intell. 2018;48 (12):4775–4791. Available from: https://doi.org/10.1007/s10489-018-1245-8.
8. Tahyudin I, Nambo H. The combination of evolutionary algorithm method for numerical association rule

mining optimization. In: Xu J, Hajiyev A, Nickel S, Gen M, editors. Proceedings of the tenth international conference on management science and engineering management. Advances in intelligent systems and computing. Singapore: Springer. 2017;p. 13–23.
9. Borah A, Nath B. Identifying risk factors for adverse diseases using dynamic Rare association rule mining. Expert Syst Appl. 2018;113:233–263. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417418304251.
10. Li J, Fu AWc, He H, Chen J, Jin H, McAullay D, et al. Mining risk patterns in medical data. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining. KDD '05. ACM; 2005. p. 770–775. Available from: https://doi.org/10.1145/1081870.1081971.
11. Erwin A, Gopalan RP, Achuthan NR. Efficient mining of high utility itemsets from large datasets. In: Advances in knowledge discovery and data mining. Springer, Berlin, Heidelberg; 2008. p. 554–561. Available from: https://doi.org/10.1007/978-3-540-68125-0_50.
12. Fournier-Viger P, Lin JCW, Truong-Chi T, Nkambou R. A survey of high utility itemset mining. In: High-utility pattern mining. Cham: Springer; 2019. p. 1–45. https://doi.org/10.1007/978-3-030-04921-8_1.
13. Nguyen LTT, Mai T, Vo B. High utility association rule mining. In: High-utility pattern mining. Cham: Springer; 2019. p. 161–74. https://doi.org/10.1007/978-3-030-04921-8_6.
14. Zaki M. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng. 2000;12(3):372–90.
15. Lin WY, Tseng MC, Su JH. A confidence-lift support specification for interesting associations mining. In: Chen MS, Yu PS, Liu B, editors. Advances in knowledge discovery and data mining, Lecture notes in computer science. Berlin: Springer; 2002. p. 148–58.
16. Brin S, Motwani R, Silverstein C. Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the 1997 ACM SIGMOD international conference on management of data. SIGMOD '97. ACM; 1997. p. 265–276. https://doi.org/10.1145/253260.253327.
17. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD international conference on management of data. SIGMOD '97. ACM; 1997. p. 255–264. https://doi.org/10.1145/253260.253325.
18. Yan X, Zhang C, Zhang S. Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Syst Appl. 2009;36(2):3066–76.
19. Liu L, Wang S, Peng Y, Huang Z, Liu M, Hu B. Mining intricate temporal rules for recognizing complex activities of daily living under uncertainty. Pattern Recogn. 2016;60:1015–28. Available from: http://www.sciencedirect.com/science/article/pii/S003132031630173X
20. Srinivasan V, Koehler C, Jin H. RuleSelector: selecting conditional action rules from user behavior patterns. Proc ACM Interact Mobile Wearable Ubiquitous Technol. 2018;2(1):35:1–35:34. https://doi.org/10.1145/3191767.
21. Padillo F, Luna JM, Herrera F, Ventura S. Mining association rules on big data through mapreduce genetic programming. Integr Comp Aided Eng. 2017;25(1):31–48. https://doi.org/10.3233/ICA-170555.
22. Guillame-Bert M, Crowley JL. Learning temporal association rules on symbolic time sequences. In: Proceedings of the 4th Asian conference on machine learning, ACML; 2012. p. 159–174.
23. Liu B, Hsu W, Ma Y. Mining association rules with multiple supports. In: Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining. KDD '99. ACM; 1999. p. 15–18.
24. Tsang S, Koh YS, Dobbie G. RP-Tree: rare pattern tree mining. In: Data warehousing and knowledge discovery. Berlin, Heidelberg: Springer; 2011. p. 277–88. https://doi.org/10.1007/978-3-642-23544-3_21.
25. Webb GI. OPUS: an efficient admissible algorithm for unordered search. J Artif Intell Res. 1995;3:431–65. Available from: https://www.jair.org/index.php/jair/article/view/10152
26. Webb GI. Efficient search for association rules. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining. KDD '00. ACM; 2000. p. 99–107. Available from: https://doi.org/10.1145/347090.347112.
27. Fournier-Viger P, Tseng VS. Mining top-K NoN-REDUNdant association rules. In: Chen L, Felfernig A, Liu J, Raś ZW, editors. Foundations of intelligent systems, Lecture notes in computer science. Berlin: Springer; 2012. p. 31–40.
28. Cheung DW, Han J, Ng VT, Wong CY. Maintenance of discovered association rules in large databases: an incremental updating technique. In: Proceedings of the twelfth international conference on data engineering; 1996. p. 106–114.
29. Tobji MB, Gouider M. Incremental maintenance of association rules under support threshold change. In: Proceedings of the IADIS international conference on applied computing. IADIS; 2006. Available from: http://arxiv.org/abs/1701.08191.
30. Aqra I, Abdul Ghani N, Maple C, Machado J, Sohrabi SN. Incremental algorithm for association rule mining under dynamic threshold. Appl Sci. 2019;9(24):5398. Available from: https://www.mdpi.com/2076-3417/9/24/5398
31. Tian D, Gledson A, Antoniades A, Aristodimou A, Dimitrios N, Sahay R, et al. A Bayesian association rule mining algorithm. In: 2013 IEEE international conference on systems, man, and cybernetics. IEEE; 2013. p. 3258–3264.
32. Gay D, Boullé M. A Bayesian approach for classification rule mining in quantitative databases. In: Machine learning and knowledge discovery in databases. Berlin,

Heidelberg: Springer; 2012. p. 243–59. https://doi.org/10.1007/978-3-642-33486-3_16.
33. Lopera Gonzalez LI. Mining functional and structural relationships of context variables in smart-buildings [PhD Thesis]. 2018. Available from: https://opus4.kobv.de/opus4-uni-passau/frontdoor/index/index/docId/573.
34. Lopera Gonzalez LI, Derungs A, Amft O. A Bayesian approach to rule mining. 2019. Available from: https://arxiv.org/abs/1912.06432v1.
35. Huang H, Tornero-Velez R, Barzyk TM. Associations between socio-demographic characteristics and chemical Concentrations contributing to cumulative exposures in the United States. J Expos Sci Environ Epidemiol. 2017;27(6):544–50. https://doi.org/10.1038/jes.2017.15.
36. Derungs A, Schuster-Amft C, Amft O. Longitudinal walking analysis in hemiparetic patients using wearable motion sensors: is there convergence between body sides?. Front Bioeng Biotechnol. 2018;6. https://doi.org/10.3389/fbioe.2018.00057/full.
37. Prosiegel M, Böttger S, Schenk T, König N, Marolf M, Vaney C, et al. Der Erweiterte Barthel-Index (EBI)–eine Neue Skala Zur Erfassung von Fähigkeitsstörungen Bei Neurologischen Patienten. Neurol Rehabil. 1996;1:7–13.