BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **19**, No 3
Sofia • 2019 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2019-0031

Identification of Risk Factors for Early Childhood Diseases Using Association Rules Algorithm with Feature Reduction

Indah Werdiningsih[1], *Rimuljo Hendradi*[1], *Purbandini*[1], *Barry Nuqoba*[1], *Elly Ana*[2]

[1]*Information System Study Program, Department of Mathematics, Airlangga University, Mulyorejo Street, 60115 Surabaya, Indonesia*
[2]*Statistics Study Program, Department of Mathematics, Airlangga University, Mulyorejo Street, 60115 Surabaya, Indonesia*
E-mails: indah-w@fst.unair.ac.id rimuljohendradi@fst.unair.ac.id purbandini@fst.unair.ac.id barry-nuqoba@fst.unair.ac.id elly-a@fst.unair.ac.id

Abstract**:** *This paper introduces a technique that can efficiently identify symptoms and risk factors for early childhood diseases by using feature reduction, which was developed based on Principal Component Analysis* (*PCA*) *method. Previous research using Apriori algorithm for association rule mining only managed to get the frequent item sets, so it could only find the frequent association rules. Other studies used ARIMA algorithm and succeeded in obtaining the rare item sets and the rare association rules. The approach proposed in this study was to obtain all the complete sets including the frequent item sets and rare item sets with feature reduction. A series of experiments with several parameter values were extrapolated to analyze and compare the computing performance and rules produced by Apriori algorithm, ARIMA, and the proposed approach. The experimental results show that the proposed approach could yield more complete rules and better computing performance.*

Keywords**:** *Early childhood diseases, PCA, Medical record, Apriori Algorithm.*

1. Introduction

Children are susceptible to diseases, especially new born babies and any one in early childhood with the age ranging from zero to six years old [1]. Treatment of diseases in early childhood is particularly important because mortality is a serious risk. Early childhood mortality rate each year is about 12.4 million [2]. Symptoms that often appear in early childhood diseases are fever, cough, and diarrhea [3, 4].

Identification of early childhood diseases at an early stage is extremely crucial for administering proper treatment and stimulating recovery, but diagnostic systems for early childhood diseases are often time-consuming and prone to errors. Moreover, clinical decisions pertaining to disease diagnosis are mostly based on the intuition and experience of medical experts rather than the wealth of empirical data hidden in databases. The former often poses a risk of misdiagnosis and mistreatment. Moreover, patients are usually advised to make a number of tests, which are often inefficient or unnecessary in diagnosing the disease.

Health care systems generate a huge amount of data containing hidden knowledge that is inaccessible by traditional methods. The adaptable data mining that can be used in medical studies [5], is therefore one workable solution. Data mining allows valuable information searching in large volumes of data, using automatic or semi-automatic exploration and analysis. Large quantities of data are explored and analyzed in order to discover meaningful patterns and rules.

Today’s data are far different from those in the past. Data has various kinds of structures so that data processing is needed [6]. Data mining is a process for extracting knowledge from data [7]. Descriptive data mining is one of the major data mining types that will find human interpretable pattern from a messy set of data, whose processing involve clustering, association rule discovery and sequential pattern discovery. Meanwhile, association rule mining is a data mining process used to find rules that may govern associations and causal objects between item sets. As for Apriori algorithm, it is a method used to find relationship patterns between one or more items in a dataset. Apriori association technique has been proven to be effective in finding various trends in healthcare databases [8]. It is well known for its representativeness in data mining [9], and it deals with item sets called transactional data [10]. Considering the relevancy of association rule mining in the rapidly changing medical databases, and thus in disease diagnosis, this paper attempts to provide a more efficient technique to identify risk factors for early childhood diseases by mining association rules from the dynamically changing medical databases.

A patient medical record generally consists of a myriad of features, so feature reduction is very important to identify the most significant risk factors related to each disease [11]. Feature reduction has been an active and fruitful field of research in machine learning and data mining. Feature reduction is a dimensionality reduction technique used to reduce irrelevant data and to increase accuracy [12]. Principal Component Analysis (PCA) is one popular statistical technique aiming to reduce data dimensions without losing any important information [13, 14]. PCA basically converts and decomposes a large number of uncorrelated variables into a smaller number of correlated variables and deductible data dimensions [15]. PCA has several advantages such as reducing data redundancy, complexity, database size, and noise and it can be used to discover correlation between variables [15].

Previous research introduced an efficient technique to identify the symptoms and risk factors for three adverse diseases: cardiovascular disease, hepatitis and breast cancer, in terms of rare association rules [16]. The results obtained shown that association rule used could achieve remarkable performance gains in terms of execution time. However, rule association rule used involved some computational overhead due to splitting and merging of nodes. Besides that, previous research used less number of features, while the medical record has many features. Therefore, this study used feature reduction to identify risk factors in early childhood diseases.

This study incorporates feature reduction technique and association rule mining technique. Apriori algorithm was used to generate the pattern sets. PCA was used as a feature reduction technique to generate the factors contributing to early childhood diseases. As for the association rule mining, we developed an efficient pattern mining technique that can proficiently derive new set of patterns and rare rules from updated databases with faster execution time, without any loss of information. To the best of our knowledge, this is the first attempt to generate a complete set of association rules from dynamically changing medical databases to identify risk factors for early childhood diseases. To summarize, the major contribution of this paper is the identification of factors contributing to early childhood diseases and the efficient generation of the sets of patterns and association rules with updated threshold values, all while designing a new proposed approach that facilitated these two processes.

In what follows, the methods of feature reduction and association rule mining are presented, continued with the description of the research results and discussion, which is then followed by the inferred conclusions.

2. Principal Component Analysis (PCA)

In general, the PCA technique transforms *n* vectors $(x_1, x_2, \dots, x_i, \dots, x_n)$ from a *d*-dimensional space to *n* vectors $(x'_1, x'_2, \dots, x'_i, \dots, x'_n)$ in a new, d'-dimensional space as [17]

$$x'_i = \sum_{k=1}^{d'} a_{k,i} \boldsymbol{e}_k, d' \leq d, \tag{1}$$

where $\boldsymbol{e}_k$ are the eigenvectors corresponding to the d' largest eigenvalues for the *scatter matrix* **S**. *The principal components* of the original data set are the projections of the original vectors x_i on the eigenvectors $\boldsymbol{e}_k$, which is denoted $a_{k,i}$. Both *d* and d' are positive integers, and the dimension d' cannot be greater than *d*. The *d*×*d* scatter matrix **S** for the original data set $(x_1, x_2, \dots, x_i, \dots, x_n)$ is defined as

$$\mathbf{S} = \mathbf{E}\left[x_i x_i^{\mathrm{T}}\right]\boldsymbol{,}\ i = 1, \dots, n, \tag{2}$$

where $\mathbf{E}\left[x_i x_i^{\mathrm{T}}\right]$ is the statistical expectation operator applied on the outer product x_i of its transpose. The representation shown in (1) minimizes the error between the original and transformed vectors. This is illustrated by considering the variance of the principal components given by [18]:

$$\sigma^2(\boldsymbol{e}_k) = \mathbf{E}\left[{a_{k,i}}^2\right] = {\boldsymbol{e}_k}^{\mathrm{T}} \mathbf{S} \boldsymbol{e}_{k,} \tag{3}$$

where $\boldsymbol{e}_k$ represents the *d*×1 vector $\boldsymbol{e}_k = \left[\boldsymbol{e}_{1,k} \boldsymbol{e}_{2,k} \dots \boldsymbol{e}_{d,k}\right]^{\mathrm{T}}$.

It is evident that the variance of the principal components is a function of the magnitude of the components of the vectors $\boldsymbol{e}_k$. At the local maxima and minima for the variance function in (3), the following relationship exists:

$$\sigma^2(\boldsymbol{e}_k + \boldsymbol{\delta e}_k) = \sigma^2(\boldsymbol{e}_k). \tag{4}$$

That equation is satisfied [18] when

$$(\boldsymbol{\delta e}_k)^{\mathrm{T}}\, \mathbf{S} \boldsymbol{e}_k - \boldsymbol{\lambda}(\boldsymbol{\delta e}_k)^{\mathrm{T}} \boldsymbol{e}_k = 0, \tag{5}$$

where $\boldsymbol{\lambda}$ is a scaling factor. This leads to

$$\mathbf{S}\boldsymbol{e}_k = \boldsymbol{\lambda} \boldsymbol{e}_k. \tag{6}$$

Equation (6) can be considered as an eigenvalue problem with nontrivial solutions only when $\boldsymbol{\lambda}$ is the eigenvalue of the scatter matrix. Thus, the associated vectors $\boldsymbol{e}_k$, $k = 1, \ldots, d'$, are the eigenvectors. If the condition $d' < d$ is satisfied, then the above representation also reduces the dimensionality of the vectors. The error in representation of the original data set $(x_1, x_2, \ldots, x_i, \ldots, x_n)$ due to the reduction in the number of dimensions to d' is given by [18]:

(7) $$E_{d'} = 0.5 \sum_{i=d'+1}^{d} \lambda_k,$$

where λ_k are the eigenvalues of the scatter matrix **S** corresponding to the eigenvectors $\boldsymbol{e}_k$. As can be seen in (7), using eigenvectors corresponding to the largest eigenvalues would give the smallest error in the representation, resulting in the variance being maximum in the direction of the eigenvectors. Also, the variance in the directions of the eigenvectors $(e_1, e_2, \ldots, e_i, \ldots, e_n)$ decreases in the same order when

(8) $$\lambda_1 > \lambda_2 > \cdots > \lambda_k > \cdots > \lambda_d.$$

This means that features with the largest variance because of their changing defect condition can be identified by examining their directionality. This property of principal components was explored for the feature selection presented in this study.

3. Association rule

Let D be the task-relevant data. T_{id} is a set of database transactions and each transaction T is a set of items. A set of items is $I= \{I_1, I_2, \ldots, I_m\}$. An itemset containing k-items is a k-itemset. If a k-itemset satisfies minimum support (Min_sup) then it is a frequent k-itemset, denoted by L_k. The first step in Apriori algorithm is to generate a set of k-itemsets candidates, denoted by C_k. Frequent itemsets are the candidate itemsets satisfying the minimum support. The descriptions of the algorithm are as follows [19]:

1) Assume a minimum support threshold (Min_sup) and a minimum confidence threshold (Min_conf) [18].

2) Scan the dataset, candidate 1-itemsets, C_1 and determine the number of occurrences of each item. The set of frequent 1-itemsets L_1 is then determined, consisting of those candidate 1-itemsets in C_1 having minimum support. Candidate 2-itemsets C_2 is generated using $L_1 \propto L_1$.

3) Scan the dataset again, frequent 2-itemsets L_2 is then determined, consisting of those candidate 2-itemsets in C_2 having minimum support. Candidate 3-itemsets C_3 is then generated by $L_2 \propto L_2$.

4) Repeatedly scan the dataset. The support count of each candidate in C_{k-1} is compared to Min_sup, then L_{k-1} is generated and C_k is generated using join $L_{k-1} \propto L_{k-1}$ until there are no more candidate item sets.

Joining and pruning actions are used to find the frequent item sets. The two steps are as follows:

a) *The joining step*. To find L_k, C_k is generated by joining L_{k-1} with itself if member l_1 and member l_2 are joined.

b) *The pruning step*. The members of C_k may not be frequent. Determining the count of each candidate in C_k is done by scanning the database, and a candidate k-itemset is deleted using L_{k-1} so that it will result in the determination of L_k.

4. Proposed approach

The proposed identification of early childhood diseases method consists of three parts, i.e., pre-processing, feature reduction using PCA, and identification of risk factors using association rule. Each part is described in details as it is follow.

4.1. Pre-processing

The first part was pre-processing the data. The data used was obtained from a hospital in Surabaya and collected from interviews and documentation. The early childhood diseases were classified into 16 diagnoses and 38 symptoms [20].The data of each patient consists of dates of treatment, age, sex, weight, height, body temperature, and any records of experiences of the 38 symptoms of 16 early childhood diseases. The 16 diseases are: cough, pneumonia, severe pneumonia, diarrhea, mild dehydration diarrhea, severe dehydration diarrhea, persistent diarrhea, severe diarrhea, common fever, severe fever, measles, measles with severe complication, measles with complication, fever may be Dengue Haemorrhagic Fever (DHF), and fever is not DHF. The patient's data are grouped based on [21] and can be seen in Table 1.

Table 1. The Patient's Grouping

Age (month)	Age feature	Weight (kg)	Weight feature	Height (cm)	Height feature
0-12	U_1	3-9	B_1	49-76	T_1
13-24	U_2	10-12	B_2	77-87	T_2
25-36	U_3	13-14	B_3	88-96	T_3
37-48	U_4	15-16	B_4	97-103	T_4
49-60	U_5	17-19	B_5	104-110	T_5
61-72	U_6	19-24	B_6	111-116	T_6

4.2. Feature reduction

The second part was to do feature reduction. PCA was used for feature reduction. Correlation between the features then became known, and so features with the smallest correlation value were eliminated [13]. The features used by PCA were only those relevant to early childhood diseases. Four steps of the feature reduction were outlined below.

1. Determination of features to be analyzed

There were 43 features. These features were: age = X_1, sex = X_2, weight = X_3, height = X_4, body temperature= X_5, flu = X_6, cough = X_7, fever = X_8, diarrhea = X_9, paling = X_{10}, dizziness = X_{11}, nausea = X_{12}, puking = X_{13}, inability to drink or suckle = X_{14}, vomiting = X_{15}, seizures = X_{16}, unconsciousness = X_{17}, fast breathing = X_{18}, breathing difficulty = X_{19}, stridor = X_{20}, liquid or soft defecating = X_{21}, hollowed eyes = X_{22}, poor abdominal skin turgor = X_{23}, restlessness = X_{24}, fussiness/ irritability = X_{25}, abnormal thirst = X_{26}, diarrhea for 14 days or more = X_{27}, blood in feces = X_{28}, stiff neck (child cannot nod until chin reaches chest) = X_{29}, rash = X_{30}, red eyes = X_{31}, turbidity on the cornea = X_{32}, mouth ulcer = X_{33}, festering eyes = X_{34}, fever for 2 to 7 days = X_{35}, high and continuous sudden fever = X_{36}, heartburn = X_{37}, red spots = X_{38}, puke mixed with blood/ coffee like = X_{39}, black feces = X_{40}, bloody nose and gums = X_{41}, infection = X_{42}, purulent eyes = X_{43}.

2. Creation of covariance matrix

Covariance matrix was created by calculating the covariance values between features. For each feature *X*, the value of the relationship with itself and other features was calculated.

3. Calculation of eigenvalues and eigenvectors

After generating the covariance matrix, the eigenvalues and eigenvectors were calculated. .

4. Determination of principal components

The eigenvectors above were used to determine the Principal Components. All Principal Components were identified by multiplying the eigenvectors with X features.

5. Calculation of loading values

Loading is the correlation between features and principal components. Loading gives an indication of how much original variable influences the generation of new variables. The higher the value of loading means that the original variable is more influential on the formation of new variables. Loading can be calculated by the next equation. The loading value which is considered a cut-off value is 0.5 [22]:

$$l_{ij} = \frac{w_{ij}}{s_j}\sqrt{\lambda_i}, \tag{9}$$

where l_{ij} is the loading value from *j*-feature to *i*-principal component; w_{ij} is the weight from *j*-feature against *i*-principal component; λ_i is the eigen value from *i*-principal component; S_j is the standard deviation value from *j*-variable.

4.3. Identification

The final part was to identify early childhood diseases. Association rule mining is a procedure to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. Mining Association Rules operates are using a two-step approach.

1. Frequent item sets generation. Frequent item sets generation is a process of generating all item sets whose support ≥ minsup, calculating support and confidence value of each item set [13]. Transaction data is shown in Table 2. Support (*s*) is a fraction of transactions that contain both *X* and *Y*. Confidence (*c*) measures the frequency of items in *Y* appears in transactions that contain *X*. Based on Table 2, Support (*s*) value and confidence (*c*) values can be calculated. Examples of *s* and *c* calculation are as follows:

Table 2. Transaction data

T_{ID}	Symptoms
1	cough, fever
2	cough, cold, diarrhea, vomiting
3	fever, cold, diarrhea, paling
4	cough, fever, cold, diarrhea
5	cough, fever, cold, paling

$$s = \frac{\sigma\,(\text{fever, cold, diarrhea})}{|T|} = \frac{2}{5},$$
$$c = \frac{\sigma\,(\text{fever, cold, diarrhea})}{\sigma\,(\text{fever, cold})} = \frac{2}{3},$$

where σ is the frequency of occurrence of an item set, and $|T|$ is the number of transaction.

Apriori Algorithm was used to generate early childhood diseases patterns.

Apriori Algorithm

```
Step 1. Let $k$=1
Step 2. Generate frequent item sets of length 1
Step 3. Repeat until no new frequent item sets are identified
    i. Generate candidate item sets with length $k$+1 from frequent item sets with length $k$
    ii. Prune candidate item sets containing subsets of item sets with length $k$ that are infrequent
    iii. Count the support of each candidate by scanning the database
    iv. Eliminate infrequent candidates, leaving the frequent ones
```

2. Rule generation. Rule generation is a process of generating high confidence rules from each frequent item set, where each rule is a binary partitioning of frequent item sets. The steps of rule generation were: list all possible association rules, compute support and confidence for each rule, and prune rules that not to reach the minsup and min_conf thresholds.

5. Experimental results

The data collected consist of 3000 items, obtained from a hospital and a health center in Surabaya. The collection techniques were interviews and document analysis. The interviews were to obtain information about risk factor from early childhood diseases. Based on the results of the interviews, it was discovered that early childhood diseases were closely associated with specific symptoms, but the patterns of and the factors that influence early childhood diseases could not yet be determined. The document analysis helped unravel the diagnosis of early childhood diseases and also the risk factors that influence the early childhood diseases. Based on the 16 diagnoses of early childhood diseases, the risk factors were age, sex, weight, height, body temperature, and the 38 symptoms (Table 3).

Table 3. Pre-processing result

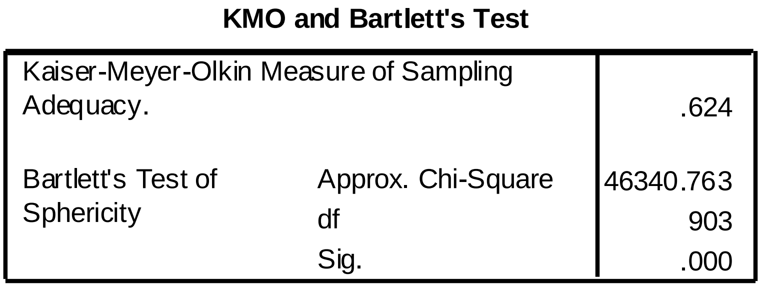
Date of treatment	Age (month)	Weight (kg)	Height (cm)	Sex	Body temperature (^{o}C)	Symptoms
16 August 2016	20	10 (B2)	79 (T2)	Female	37.6	fever, cough
6 September 2016	24	12(B2)	80(T2)	Female	39.2	fever, cough, paling
24 October 2016	16	11(B2)	78(T2)	Male	39	fever, cough, paling
11 March 2017	7	8(B1)	66(T1)	Male	38.4	fever, flu, cough
24 May 2017	24	13(B3)	76(T1)	Male	37.9	fever, flu

The first part was to input patient data and conduct the pre-processing. The data used at this stage were dates of treatment, age, weight, height, sex, body temperature, and the 38 symptoms. The Pre-processing result can be seen in Table 3. Following this pre-processing was the feature reduction process.

5.1. Feature reduction

The first step of feature reduction was to look at the Kaiser-Meyer-Olkin (KMO) and Bartlett's test value to ensure that the dataset meets the requirements for PCA analysis. KMO and Bartlett Test results can be seen in Fig. 1. Bartlett's Test value was 46340.763 at significant 0.000 which means correlation between variables is very significant, and KMO value was 0.624 which shows adequacy of samples so that the dataset used meets the requirements for PCA analysis.

The next step was to do implement PCA analysis. PCA analysis began by calculating the correlation value between variables. There were two ways in determining the relationship between variables, namely calculating the correlation value (correlation matrix) between variables, and calculating covariance (covariance matrix) of all existing variables [23]. After that, a correlation matrix was formed. It was used to calculate PCA by looking at eigenvalues of each variable. New features (principal component) formed based on eigenvalues that were higher than 1. The results of eigenvalues and variance calculation can be seen in Table 4.



KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.624
Bartlett's Test of Sphericity	Approx. Chi-Square	46340.763
	df	903
	Sig.	.000

Fig. 1. KMO and Bartlett test result

Table 4 shows that there are 16 principal components having eigenvalue higher than 1. After the principal components were formed, the next step was factor rotation using Varimax rotation. The results of factor rotation can be seen in Table 5.

Table 5 describes the loading factor, correlation between original features and principal components formed by PCA. The features chosen are these, whose loading factor value is above 0.5 that is considered capable of explaining the influential features. Features that have loading factor below 0.5 are considered not or less influential. Based on the results obtained, 20 features were generated from PCA, i.e., weight, height, sex, flu, cough, fever, diarrhea, stridor, blood in the feces, vomiting, seizures, unconsciousness, inability to drink or suckle, hollowed eyes, fussiness/ irritability, abnormal thirst, turbidity on the cornea, fever of 2-7 days, poor abdominal skin turgor, diarrhea of 14 days or more, and breathing difficulty. These features were the ones having influence in early childhood diseases, which were then used for the generation of early childhood diseases patterns.

Table 4. The results of eigenvalues and variance

Component	Initial eigen			Component	Initial eigen		
	Total	% of variance	Cumulative %		Total	% of variance	Cumulative %
1	4.787	11.131	11.131	11	1.251	2.908	53.855
2	2.708	6.298	17.429	12	1.224	2.846	56.702
3	2.547	5.924	23.353	13	1.134	2.637	59.339
4	2.281	5.305	28.658	14	1.073	2.496	61.835
5	1.968	4.578	33.236	15	1.043	2.426	64.261
6	1.684	3.917	37.153	16	1.026	2.387	66.648
7	1.604	3.731	40.884	17	0.981	2.282	68.931
8	1.552	3.610	44.494	…	…	…	…
9	1.476	3.432	47.926	…	…	…	…
10	1.299	3.021	50.947	43	0.049	0.114	100.000

Table 5. The results of factor rotation using Varimax

X_i	Components					
	1	2	3	…	...	16
X_1	0.023	0.391	0.380	…	…	0.026
X_2	0.291	0.703	–0.029	…	…	0.036
X_3	0.158	0.619	0.078	…	…	–0.024
X_4	0.355	0.674	–0.045	…	…	–0.009
…	…	…	…	…	…	…
…	…	…	…	…	…	…
X_{43}	0.463	0.061	–0.566	…	…	–0.056

5.2. Identification of risk factors

After the feature reduction, the next in line was the generation of risk factors patterns using Apriori algorithm, processing the features sifted in the reduction stage. Three different sets of experiments were conducted to measure the significance of the proposed approach compared to other pattern-mining approaches. The evaluation was carried out by comparing the item sets and the generated rules, as well as the execution time spent by each algorithm under the different conditions. After the generation of frequent patterns was completed, we compared the performance of our approach with the existing pattern mining techniques.

5.2.1. Experimental evaluation

This subsection is aimed to carry out a comparative performance evaluation of the proposed approach compared to other approaches in analyzing the early childhood disease data. As mentioned previously the data had already been sifted in terms of the number of item sets and generated rules, and also the execution time.

Item sets and rules generated. The efficiency of the proposed approach compared to existing techniques can be evaluated in terms of the number of item sets and generated rules. The use of PCA can reduce the features that will be used for identification, which gives flexibility to users to generate any sets of desired patterns. Thus, PCA is capable of extracting complete sets of item sets and rare rules without any loss of information. Table 6 depicts the respective numbers of item sets and rules generated by the proposed approach and other association rule mining techniques.

Table 6. Rule and item sets generated

Algorithm	Frequent Item sets	Rare Item sets	Rules
Apriori Algorithm	272	–	3662
ARIMA	–	956	4897
The proposed approach	140	764	1810

Regarding the association rule mining, the Apriori algorithm generated only the frequent item sets and hence resulting only the frequent association rules. The second technique is Apriori-Rare algorithm [24], also known as ARIMA. ARIMA was only able to generate the rare item sets and rare association rules. Different from these two techniques, the proposed approach could generate the complete sets (the frequent and the rare item sets) as well as frequent association rules and rare association rules [16].

Execution time. To analyze the effect of threshold variations, all datasets were used for testing. The support threshold was varied gradually to generate the sets of frequent and rare patterns under different threshold values. The minimum support threshold was set from 5% up to 25%. The performance evaluation of the proposed approach compared to other pattern mining algorithms is illustrated in Figs 2 and 3.

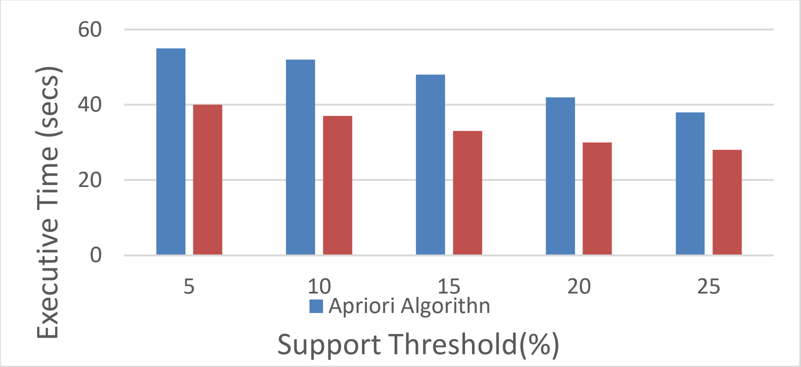


Fig. 2. Execution time for frequent pattern generation from dataset

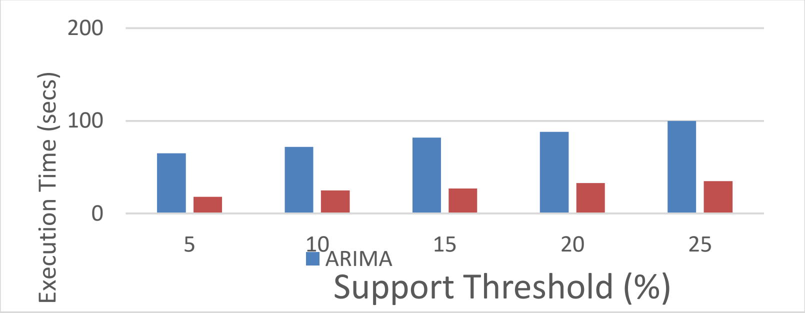


Fig. 3. Execution time for rare pattern generation from dataset

5.2.2. Rule analysis

A detailed analysis of the rules generated from early childhood dataset is presented in this section. The rules generated by the Apriori, ARIMA, and the proposed approach were compared using the number of MinFreq, MinRare, and MinConf values. The medical significance of the rules was evaluated using the interestingness measures of support and confidence. For the rule generation, minFreq value was fixed at 15%, minRare value was fixed at 1%, and minimum confidence and minConf were fixed at 80%. Rules generated by Apriori Algorithm, ARIMA and the proposed approach are illustrated in Tables 7-9.

Table 7 presents the rules extracted by Apriori Algorithm. The rules generated have high support and confidence values, and contain only frequent items in the antecedent and the consequent part. They represent rules of the form frequent → Frequent.

Table 7. Rules generated from dataset by Apriori Algorithm

Rule	Antecedent	Consequent	Support	Confidence
1	Weight = B_2, or height = T_2, and symptoms = fever, flu, cough	Class = low	0.25	0.95
2	Sex = Female, and symptoms = fever, cough	Class = low	0.20	0.90
3	Sex = Male, and symptoms = flu, cough, diarrhea	Class = low	0.20	0.90
4	Sex = Male, Weight = B_3 or height = T_3, and symptoms = hollowed eyes, blood in the feces, diarrhea, fever	Class = medium	0. 10	0.80
5	Weight = B_2 or height = T_2, and symptoms = cough, flu, vomiting, breathing difficulty	Class = medium	0.10	0.85
6	Sex = Female, Weight = B_1 or height = T_1, and symptoms = diarrhea of 14 days or more, fever, stridor	Class = medium	0.10	0.85
7	symptoms = flu, fever, vomiting, inability to drink or suckle	Class = medium	0.25	0.95
8	Weight = B_2 or height = T_2, and symptoms = Cough, flu, turbidity on the cornea	Class = low	0.20	0.85

Table 8 illustrates the rules generated by ARIMA, which are of low support value but of high confidence value. ARIMA generated two categories of rare item sets: one having only rare items and another having a combination of frequent and rare items. It failed to generate rare item sets of the third category that contain only frequent items. The generated rules thus either contain only rare items in the antecedent and consequent part or a combination of frequent and rare items.

Table 8. Rules generated from dataset by ARIMA

Rule	Antecedent	Consequent	Support	Confidence
1	Weight = B_4 or height = T_4, and Symptoms = inability to drink or suckle, seizures	Class = medium	0.010	0.85
2	Symptoms = flu, cough, diarrhea	Class = low	0.025	0.90
3	Symptoms = Hollowed eyes, blood in the feces, diarrhea	Class = medium	0.020	0.80
4	Weight = B_2 or height = T_2, and symptoms = cough, flu, vomiting	Class = low	0.025	0.85
5	Symptoms = Hollowed eyes, fussiness/ irritability, thirst, stridor	Class = medium	0.02	0.85
6	Weight = B_3 or height = T_3, and symptoms = flu, turbidity on the cornea	Class = low	0.01	0.80
7	Weight = B_3 or height = T_3, and symptoms = cough, poor abdominal skin turgor	Class = low	0.025	0.95
8	Sex = Female, and symptoms = fever, flu, hollowed eyes	Class = low	0.020	0.85

Table 9 presents the rules generated by the proposed approach. The proposed approach managed to generate the complete sets of rare item sets and also all the two categories of rare association rules discussed in the earlier section.

Table 9. Rules generated from dataset by proposed approach

Rule	Antecedent	Consequent	Support	Confidence
1	Symptoms = inability drink or suckle, vomiting, stidor	Class = low	0.20	0.90
2	Sex = Male, Weight = B_2, symptoms = hollowed eyes, fussiness/ irritability, abnormal thirsty, stridor, diarrhea	Class = medium	0.25	0.95
3	Weight = B_2, and symptoms = cough, flu, turbidity on the cornea	Class = low	0.15	0.85
4	symptoms = blood in the feces, diarrhea, fever	Class = medium	0.20	0.90
5	symptoms = hollowed eyes, fever of 2-7 days, poor abdominal skin turgor	Class = medium	0.10	0.90
6	Sex = Male, Weight = B_2, and symptoms = blood in the feces, fever, child cannot drink or suckle, vomiting	Class = medium	0.20	0.95
7	Symptoms = diarrhea of 14 days or more, stridor	Class = medium	0.25	0.85
8	Weight = B_3, and symptoms = cough, breathing difficulty, vomiting	Class = low	0.20	0.80

6. Discussion

Previous research showed that association rule used achieved remarkable performance gain in terms of execution time [16]. However, association rule used involved some computational overhead due to splitting and merging of nodes. Therefore, this study used Apriori algorithm because it could reduce the number of candidates by pruning. The pruning made Apriori algorithm to have good performance [19].

Besides that, previous research used less number of features, while the medical record has many features. Therefore, this study used feature reduction to identify risk factors in early childhood diseases. PCA was used for feature reduction because PCA can reduce data redundancy, complexity, database size, and noise and can be used to discover correlation between variables [15].

This study incorporates feature reduction technique and association rule mining technique. The experiments results shown that the proposed approach had good performance and reduce the feature as shown in Fig. 2 and Fig. 3.

Fig. 2 compares the performance of the two algorithms under the conditions of different support levels in processing the 3000 data items. When the support level is high, both algorithms are less time-consuming because PCA can reduce the features, so that the number of iterations is small. The proposed approach takes less time to discover frequent itemsets and shows much greater efficiency than the Apriori algorithm.

Fig. 3 compares the performance of the two algorithms under different support levels in processing the 3000 data items. When support is increased, the time increases. This is because rare association rules have a low support level and a high confidence level as opposed to the general association rules which are characterized by a high support and a high confidence level. However, ARIMA and the proposed approach takes less time, because the proposed approach has a small number of iterations. The proposed approach takes less time and performs with much greater efficiency than ARIMA.

The proposed approach achieves remarkable performance gain in terms of execution time. The use of PCA can reduce the features that will be used for identification, which gives flexibility to users to generate any sets of desired patterns. Thus, PCA is capable of extracting complete sets of item sets and rare rules without any loss of information.

7. Conclusion

The increasing mortality rate every year due to early childhood diseases has become a major concern worldwide. Computational intelligence techniques like rare association rule mining facilitate the intensification of decision-making and medical diagnosis by analyzing the rare correlations between different patient characteristics and diseases. The study, therefore, aimed to identify the factors contributing to early childhood diseases' diagnosing with the appearing of 38 symptoms. This paper introduces an efficient rare association rule mining technique for generating the sets of significant rare association rules from medical record datasets. A detailed and thorough analysis of the extracted rare association rules was run to assist the medical experts in diagnosis of these diseases.

The efficiency of the proposed approach compared to other pattern mining approaches was evaluated using a widely used medical record dataset. The comparative result analysis indicates the pre-eminence of the proposed approach over existing frequent and rare pattern mining approaches and its efficacy in generating significant rare association rules pertaining to medical diagnosis.

The proposed approach has achieved remarkable performance gain in terms of execution time. The use of PCA can reduce unnecessary features to aid identification, which gives flexibility to users to generate any sets of desired patterns. Thus, PCA is capable of extracting the complete sets of item sets and rare rules without any loss of information. In our future endeavors, we will attempt to investigate other issues related to dynamic mining of association rules.

***Acknowledgements*:** Authors would like to thank the General Director of Higher Education (DIKTI) who had trusted and funded the research entitled System of Diseases Diagnosis for Early Childhood using PCA and Apriori algorithm with contract number: 004/SP2H/T/LT/DRPM/IV/2017. We express our gratitude for facilitating this research so that it could be completed without any significant difficulties.

References

1. Yanto, B. F., I. Werdiningsih, E. Purwanti. Expert System Application of Early Childhood Diseases Diagnosis Using Forward Chaining Method. – J. Inf. Syst. Eng. Bus. Intell., Vol. **3**, Indonesian Version, 2017, No 1, pp. 61-67.

2. Bank, W. World Development Report: Investing in Health. 1993.
3. Garenne, M., C. Ronsmans, H. Campbell. The Magnitude of Mortality from Acute Respiratory Infections in Children under 5 Years in Developing Countries. – World Health Stat. Q., Vol. **45**, 1992, No 2-3, pp. 180-191.
4. D. O. M., C. Monteiro, J. Akré, G. Clugston. Global Database on Child Growth and Malnutrition the Worldwide Magnitude of Protein – Energy Malnutrition : An Overview from the WHO Global Database on Child Growth. ─ Bull. World Health Organ., Vol. **71**, 2015, No December, p. 2015.
5. Koh, H. C., G. Tan. Data Mining Applications in Healthcare. ─ J. Healthc. Inf. Manag., Vol. **19**, 2011, No 2, pp. 64-72.
6. Demetrovics, J., H. M. Quang, N. V. Ank, V. D. Thi. –An Optimization of Closed Frequent Subgraph Mining Algorithm. – Cybernetics and Information Technologies, Vol. **17**, 2017, No 1, pp. 3-15.
7. Venkatram, K., M. A. Geetha. Review on Big Data &amp; Analytics – Concepts, Philosophy, Process and Applications. – Cybernetics and Information Technologies, Vol. **17**, 2017, No 2, pp. 3-27.
8. Jain, D., S. Gautam. Implementation of Apriori Algorithm in Health Care Sector: Data Mining in Health Care Sector. – Int. J. Comput. Sci. Commun. Eng., Vol. **2**, 2013, No 4, pp. 26-32.
9. Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, A. Verkamo. Fast Discovery of Association Rules. – Advances in Knowledge Discovery and Data Mining, Vol. **12**. 1996, pp. 307-328.
10. Wu, M., H. Sakai. On Parallelization of the NIS-Apriori Algorithm for Data Mining. – Procedia Comput. Sci., Vol. **60**, 2015, No 1, pp. 623-631.
11. Jabbar, M. A., B. I. Deekshatulu, P. Chandra. Heart Disease Classification Using Nearest Neighbor Classifier with Feature Subset Selection. – Anale. Comput. Sci. Ser., Vol. **XI**, 2013, pp. 47-54.
12. Jain, D., V. Singh. Feature Selection and Classification Systems for Chronic Disease Prediction: A Review. – Egypt. Informatics J., 2018.
13. Han, J. J. P., M. Kamber. Data Mining Concepts and Techniques. Third Edition. Elsevir, 2012.
14. Martono, G. H., T. B. Adji, N. A. Setiawan. PCA Implementation for Reducing Factors Influencing Coronary Heart Diseases. – In: Seminar Nasional “Science, Engineering and Technology”, Indonesian Version, 2012, pp. 1-5.
15. Tang, J., S. Alelyani, H. Liu. Feature Selection for Classification: A Review. – Data Classif. Algorithms Appl., 2014, pp. 37-64.
16. Borah, A., B. Nath. Identifying Risk Factors for Adverse Diseases Using Dynamic Rare Association Rule Mining. – Expert Syst. Appl., Vol. **113**, 2018, pp. 233-263.
17. Duda, R. O., P. E. Hart, D. G. Stork. Pattern Classification. Second Edition. Viley, 2001.
18. Haykin, S. Neural Network. 2005.
19. Bennett, S. E., S. Lane, D. Mcmillen. Optimization of Association Rule Mining Apriori Algorithm Using ACO. – Int. J. Soft Comput. Eng., 2016, No 1, pp. 24-26.
20. Indonesian Government of Health. Manajemen Terpadu Balita Sakit (MTBS). Indonesian Version, 2011.
21. Indonesian Government of Health. Decision from Health Ministery of Indonesian Government Number 1995 Years 2010 about Antropometry Standard on Evaluating Children Nutrition. Indonesian Version, 2010.
22. Sharma, S. Applied Multivariate Techniques Subhash Sharma. 1996.
23. Jolliffe, I. T. Principal Component Analysis. Second Edition. – Springer Ser. Stat., Vol. **98**, 2002. 487 p..
24. Romero, C., J. R. Romero, J. M. Luna, S. Ventura. Mining Rare Association Rules from e-Learning Data. – Virchows Arch, Vol. **442**, 2003, No 5, pp. 462-467.

Received: 04.01.2019; Second Version:20.03.2019; Third Version: 30.04.2019; Accepted: 07.06.2019