Computers, Materials & Continua
DOI:10.32604/cmc.2021.016766
Article

Tech Science Press

CARM: Context Based Association Rule Mining for Conventional Data

Muhammad Shaheen[1,*] and Umair Abdullah[2]

[1]Faculty of Engineering & Information Technology, Foundation University, Islamabad, Pakistan
[2]Department of Computer Science, Barani Institute of Information Technology, Rawalpindi, Pakistan
[*]Corresponding Author: Muhammad Shaheen. Email: dr.shaheen@fui.edu.pk
Received: 11 January 2021; Accepted: 14 March 2021

Abstract: This paper is aimed to develop an algorithm for extracting association rules, called Context-Based Association Rule Mining algorithm (CARM), which can be regarded as an extension of the Context-Based Positive and Negative Association Rule Mining algorithm (CBPNARM). CBPNARM was developed to extract positive and negative association rules from Spatio-temporal (space-time) data only, while the proposed algorithm can be applied to both spatial and non-spatial data. The proposed algorithm is applied to the energy dataset to classify a country's energy development by uncovering the enthralling interdependencies between the set of variables to get positive and negative associations. Many association rules related to sustainable energy development are extracted by the proposed algorithm that needs to be pruned by some pruning technique. The context, in this paper serves as a pruning measure to extract pertinent association rules from non-spatial data. Conditional Probability Increment Ratio (CPIR) is also added in the proposed algorithm that was not used in CBPNARM. The inclusion of the context variable and CPIR resulted in fewer rules and improved robustness and ease of use. Also, the extraction of a common negative frequent itemset in CARM is different from that of CBPNARM. The rules created by the proposed algorithm are more meaningful, significant, relevant and insightful. The accuracy of the proposed algorithm is compared with the Apriori, PNARM and CBPNARM algorithms. The results demonstrated enhanced accuracy, relevance and timeliness.

Keywords: Association rules; context; CBPNARM; non-spatial data; CPIR; support; confidence; interestingness

1 Introduction

It is an information age as all is transferred to computers and the use of the information system has become a necessity of life. Knowledge extraction from data takes place through the data mining process. Data mining is a step-by-step process that begins with data analysis, classification/prediction and finding trends and patterns [1,2]. A variety of data mining techniques are used for classifying, extracting association rules, clustering, and regression analysis. The accumulation of data in databases using different devices produces a pool of data that serves as

a foundation for knowledge generation. The Size of the data and the reliability of knowledge extraction are directly proportional to one another. With the advent of internet technologies and community applications, millions and trillions of users are generating data every minute and the growth of the repositories storing this data is exponential. As a result, human dependence on data has also increased. Numerous challenges in text mining, web analytics and knowledge discovery have emerged [3]. The discovery of knowledge from databases is a non-trivial process in identifying logical, understandable and innovative patterns from the data [4]. Knowledge extracted through data mining can take different forms, such as rules, clusters, decision trees, classes, rough sets and many others [5–7].

Data mining prepares the data for processing by recovering the erroneous and blank data fields that are then stored in the warehouse and finally applying algorithms to it [1]. Data mining leads to classes, clusters, rules and predictions [8]. It can be applied to different datasets, including educational data [9,10], spatial data [7], satellite data [2,11], scientific experiments [12], biological data [13]. Association rule mining is used to discover the fascinating interdependencies between the set of variables and reveals a hidden pattern in the set of data and variables concomitant with high frequencies. A comprehensive review of the association rule extraction algorithms is provided by AI et al. [14]. Wu et al. [5] emphasized the importance of negative association rules which were not taken into consideration in the mining of association rules before that. A typical association rule of shape ($X \rightarrow Y$) is positive if it indicates a presence association between X and Y. ($X \rightarrow \neg Y$) is a negative association rule if the presence of X assures the absence of Y in the database. Many studies have been carried out to mine positive and negative association rules from different datasets [15–18].

Shaheen et al. [7] introduced a variable called context that can essentially be used to mine valid positive and negative association rules. The context variable can produce valid but false rules that qualify the support value criteria and are included in the final rule set. For example, the higher selling rate of sanitisers in summer produce a rule (summer $\rightarrow$ High_sanitiser_sale) whereas the actual reason for the increase in the sale of sanitisers in the last two summers was the spread of coronavirus. Thus the spread of the coronavirus in this example is a context variable. It may not be permanently stored as an attribute in the database, or ignored. These variables are context variables that can affect the validity of the retrieved association rules. CBPNARM [7] uses context variables and has given very good results in terms of the number of rules, confidence and interestingness. The use of context variable for mining association rules can also be cited in other studies, but the definition of context seems confined to time and location [19].

The CBPNARM algorithm was developed for the extraction of Spatio-temporal association rules, which was applied only to spatial data. Spatial data differs from conventional data in that it relates directly or indirectly to a location on earth. Spatial data attributes combine to represent an image that is drawn on the geographic information system (GIS) or other similar information systems [20]. Attributes that are not spatial are represented by non-spatial attributes and are known as characteristic data. A context-based algorithm for non-spatial data is also required which can be applied to non-spatial GIS data and any other dataset established through conventional data procedures. Apriori algorithm is the one most commonly used for exploring positive association rule mining on these datasets.

Apriori algorithm proposed by Agarwal et al. [21] is used to derive the relationship between frequent items of transactional databases. An Apriori association rule is written as (Antecedent $\rightarrow$ Consequent) and can be elaborated as "if antecedent happens, it is more likely that consequent happens." The selection criteria for a rule in the final rule set differ depending on the algorithms.

The most common are support, confidence, lift, interestingness measure, dependency, etc. Apriori uses support, confidence and lift to select rules [22]. The Apriori algorithm only looks for positive association rules. An exceptionally sheer number of rules is mined when the database is considered for extracting positive and negative association rules. Different pruning measures are proposed by Wu et al. [5] to reduce the number of positive and negative association rules, thus increasing the prospects for outcomes. The context variable in CBPNARM also served as a pruning technique to reduce the final set of association rules. The value of the context can sometimes lead to violating the validation criteria of the association rule for which they are either reckoned or pruned and are not included in the final rule set. The influencing factor, that is to say, the context variable may alter the value of another variable, which may cause the final rule to change [7,12]. Given the context variable, the patterns and rules generated may be more accurate and meaningful.

The proposed algorithm is implemented for sustainable energy development indicators. Sustainability in the energy sector is the primary need of almost every country in the world. The commission on sustainable development has provided a list of indicators [23] that were refined by the International Atomic Energy Agency (IAEA) for its use in evaluating sustainable energy development [24]. These sustainability indicators are used in many studies to assess energy development [8], energy security [25], environmental impacts [26], energy poverty [27], energy consumption and relationships with one another [28]. A classification mechanism for a country's energy development is developed by Shaheen et al. [8] using only quantifiable indicators. The algorithm proposed in this study is also implemented for the same dataset. The algorithm applied to the sustainable energy indicators returned association rules which defined the covarying sustainability metrics. Depending on the value and extent of the covariance between these indicators, a decision-maker can develop an optimum plan to ensure the sustainability of the energy sector.

This paper is intended to develop an algorithm for exploring positive and negative context-based association rules for conventional/characteristic data as an extension to the CBPNARM algorithm. The accuracy of the proposed methodology is compared with Apriori, CBPNARM at the methodological level and is also compared to sustainable energy development, categorized at the application level. The contribution made in this study is given below:

1) CBPNARM algorithm was designed for spatial data only. CARM is the algorithm proposed in this paper which can be applied to non-spatial or conventional numeric and ordinal data.
2) The algorithm is applied to energy datasets to mining rules for energy sustainability.
3) CPIR is not used in the CBPNARM algorithm as the complexity of CBPNARM became greater after CPIR when the results were not remarkable. CPIR is added to the proposed algorithm.
4) The extraction of negative frequent items in the CARM differs from that of CBPNARM.
5) Four CARM algorithm cases given in the pseudo-code differ from CBPNARM.

2 Indicators for Sustainable Energy Development

The importance of energy is vigorous in eliminating scarcity and elevating the standard of human life [29]. The world has acknowledged that sustainable energy development is important. In 2005, the Commission for Sustainable Development (CSD) recognized the role of the energy sector in the sustainable development of a country [23]. A list of 30 energy sustainability indicators was finalized. These indicators are classified into three categories that are essential ingredients for sustainability; (1) social domain (2) economic domain and (3) ecological domain. The social

domain of sustainability indicators is divided into equity and health as shown in Tab. 1. Equity is about equitable access and the availability of all the energy resources at an affordable price. Health covers safe access to energy by caring for accidents in the fuel cycle and eradicating problems related to air pollutants, etc. The social domain indicators selected for this study are placed in the first section of Tab. 1.

Table 1: Indicators for sustainable energy development [23]

S. No.	Name of indicator	Data required
Social domain		
1	Share of households (or population) without electricity or commercial energy	Population (no energy), total population
2.	Share of household income spent on fuel and electricity	Income spent on energy, total income
	Domestic use of energy classified with respect to the income group and fuel mix	Energy use per household, Household income, Corresponding fuel mix
	Accident fatalities per energy produced by fuel chain	Annual fatalities by fuel chain, Annual energy produced
Economic domain		
3.	Energy use per capita	Energy use, Total population
4.	Energy use per unit of GDP	Energy use, GDP
5.	Efficiency of energy conversion and distribution	Losses in electricity generation, transmission and distribution
6.	Reserves-to-production ratio	Proven recoverable reserves, Total energy production
7.	Resources-to-production ratio	Total estimated resources, Total energy production
8.	Value added by energy in industrial sector	Use of energy in industry, Value added
9.	Value added by energy in agriculture	Use of energy in agriculture, Value added
10.	Value added by energy in service sector	Use of energy, Value added
11.	Value added by energy in household	Use of energy in household, Value added
	Value added by energy in transport	Use of energy in transport, Value added
	Fuel shares in energy and electricity	Primary energy supply and final consumption by fuel type, Total primary energy supply and final consumption
12.	Non-carbon energy share in energy and electricity	Non-carbon energy supply and final consumption, Total primary energy supply and final consumption
13.	Renewable energy share in energy and electricity	Renewable energy supply and final consumption, Total primary energy supply and final consumption
14.	End-use energy prices by fuel and by sector	Energy prices with and without tax
15.	Net energy import dependency	Energy imports, Total primary energy supply
16.	Stocks of critical fuels per corresponding fuel	Stocks of critical fuel, Critical fuel consumption

(Continued)

Table 1: Continued

S. No.	Name of indicator	Data required
Ecological domain		
17.	Greenhouse-gas emissions by energy products	Greenhouse-gas emissions resulting from energy products and its use, Total population, GDP
	Ambient concentrations of air pollutants in urban areas	Concentration of pollutants in air
	Air pollutant emissions from energy systems	Air pollutant emissions
18.	Pollutant expulsions in liquid wastes from energy systems	Pollutant expulsions in energy liquid wastes
	Soil area where acidification exceeds critical load	Affected soil area, Critical load
19.	Rate of deforestation attributed to energy use	Forest area at two different times, Biomass utilization
20.	Ratio of waste generated in energy production to energy obtained	Amount of generated waste from the source, Total energy production from the source
21.	Ratio of waste properly disposed of total generated solid waste	Disposed solid waste, Total solid waste
22.	Ratio of solid radioactive waste to units of energy produced	Units of solid radioactive waste, Energy produced
23.	Ratio of solid radioactive waste awaiting disposal to total generated solid radioactive waste	Solid radioactive waste awaiting disposal, Total solid waste

The economic domain of sustainability indicators can be divided into consumption, production patterns and security of supply. The indicators related to the consumption and production of energy include energy use per GDP per capita, energy supply efficiency, energy production, etc. The ecological domain covers the impacts of energy-related indicators of atmosphere, water and land [30]. IAEA [23] did not consider the institutional dimension of sustainability because the data associated with this aspect was unquantifiable. The report also suggested some auxiliary statistics that measure demographics, wealth, economic development, transportation, urbanization, etc. These measures include GDP per capita, population, shares of sectors in GDP, distance travelled per capita, freight transport, income inequality, floor area per capita and manufacturing value. The commission also recommended the analysis of time-series data, the preparation of data for analysis and the interpretation of the discourse of the data collected for that purpose. This study specifically followed the recommendations of the report and proposed an algorithm for such an assessment. CBPNARM being specifically for spatiotemporal data mining does not adapt exactly, the need for the problem.

The basis for the selection of energy sustainability indicators for this study is identical to that proposed by Shaheen et al. [8], where quantifiable and available indicators were selected. In this study, only indicators for which data are available on online energy portals are selected. The list of selected sustainability indicators is given in Tab. 1. The data for the marked attributes in the grey-shaded boxes in Tab. 1 was not available where such attributes were excluded from the database. Data for 16 of the other 23 attributes was readily available, while the remaining data was derived from the available datasets.

3 Definitions

3.1 Support

Support is a measure of finding the frequency of an itemset in the database [31,32]. The support of an association rule X → Y is 0.6 if X and Y appeared in a transactional database T for 60% times of the total transactions in T. The equation to compute support is given below:

$$Supp\,(X \rightarrow Y) = \frac{|X \rightarrow Y| \therefore X,\ Y \in T \ and\ X \rightarrow Y \subseteq T}{|T|} \quad |X \rightarrow Y| \text{ represents the size of the set containing X and Y.} \tag{1}$$

3.2 Confidence

Confidence is an indication of how often a rule is true [32,33]. The confidence of an association rule $X \rightarrow Y$ is 1 if X appeared in the database 10 times and Y appeared with X in all the transactions. The equation to compute confidence is given below:

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

3.3 Lift

Lift is used to measure the correlation value of the antecedent and consequent of an association rule [31,32]. Lift of an association rule $X \rightarrow Y$ is 1 if X is not correlated to Y. Lift is computed by the equation given below:

$$Lift\,(X \rightarrow Y) = \frac{Supp(X \rightarrow Y)}{Supp\,(X)\,Supp(Y)} \tag{3}$$

3.4 Interestingness

Interestingness is a measure used to find potentially positive and potentially negative item sets from a dataset. A rule $X \rightarrow Y$ is not interesting if its support is lesser than the product of individual supports of X and Y [5].

$$Interest(X \rightarrow Y) = |\,(Supp(X \rightarrow Y)) - \text{Supp (X)Supp(Y)}\,| \tag{4}$$

3.5 CPIR

The conditional-probability increment ratio (CPIR) of a rule is computed based on the dependence of the antecedent and consequent. In an association rule $X \rightarrow Y$, X is positively dependent on Y, if the value of lift of $X \rightarrow Y$ is greater than 1 and negatively dependent if the value is lesser than 1. The dependence when equated as per Eq. (5) returns the value of CPIR [5].

$$\begin{cases} CPIR = \dfrac{P(Y \mid X)\,P(Y)}{1 - P(Y)} & if\ \ P(Y \mid X) \geq P(Y),\ \ P(Y) \neq 1 \\ \dfrac{P(Y \mid X)\,P(Y)}{P(Y)}, & if\ \ P(Y) > P(Y \mid X),\ \ P(Y) \neq 0 \end{cases} \tag{5}$$

3.6 Context

Context is the state of the entity, environment or action that can affect the results of association rule mining. The value of the context variable must be within the normal range to make a matching rule valid. For example, the change in vegetation color in the surrounding area may indicate an emergency below the earth's surface. If the value of the "waterflood" context variable is not normal and is not in normal ranges, then the change in vegetation color may indicate the presence of a volcano. The color, in this example, was changed due to the waterflood so that the waterflood, which in this case is a context variable, whose value for this rule was over the normal range [7]. The value change of the context variable can have four cases that are addressed in [7].

4 Proposed Method

The method proposed for extracting positive and negative association rules in conventional data sets is named CARM and is dependent on support, confidence, interestingness, CPIR and the value of the context variable. This method fetches the rules from the non-spatial datasets. CBPNARM [7] is developed as an extension of [5,34] and is used in some successful studies [12,35]. The proposed algorithm is an extended CBPNARM. A positive association rule $X \rightarrow Y$ is valid if $X \cap Y = \phi$, $Supp(X) > BaseSupVal$, $Supp(Y) > BaseSupVal$ and $Supp(X \cup Y) > BaseSupVal$. $BaseSupVal$ is the user-defined threshold value of support. According to Eqs. (1) and (2), support is defined by $Supp(X \cup Y)$ and confidence is defined by $Supp(X \rightarrow Y)/Supp(X) \geq BaseConfVal$. $BaseConfVal$ is the user-defined threshold value of confidence. A negative association rule $X \rightarrow \neg Y$ OR $\neg X \rightarrow Y$ OR $\neg X \rightarrow \neg Y$ is valid if $X \cap Y = \phi$, $Supp(X) > BaseSupVal$, $Supp(Y) > BaseSupVal$ and $Supp(X \cup Y) < BaseSupValNeg$, where $BaseSupValNeg$ is the user-defined threshold value of support for negative association rule.

Table 2: Variants of context variables with remedial statistical adjustments (COA = current value of context attribute, LLC = lower limit of the context variable, ULC = upper limit of the context variable, NSV = new support value, AS = actual support)

Positive rules	COA < LLC	Dissimilarity = (LLC − COA) ∗ 100/COA; NSV = AS + (AS ∗ Dissimilarity)/100
	COA > ULC	Dissimilarity = (LLC − COA) ∗ 100/COA; NSV = AS − (AS ∗ Dissimilarity)/100
Negative rules	COA > LLC	Dissimilarity = (LLC − COA) ∗ 100/COA; NSV = AS − (AS ∗ Dissimilarity)/100
	COA < ULC	Dissimilarity = (LLC − COA) ∗ 100/COA; NSV = AS + (AS ∗ Dissimilarity)/100

The aforementioned mathematical procedures generate a large number of positive and negative association rules. The measure of Interestingness measure proposed by [5] is used to apply first level pruning. The interestingness of the rules can be calculated using Eq. (4). After applying the first level pruning through an interestingness measure, the second-level pruning is applied to further reduce the number of rules. The second level pruning measure is the CPIR, which is defined in Eq. (5). All rules that are positively and negatively dependent are eligible to be included in the final rule set. In this level of pruning, only the rules in which antecedent and consequent are independent of one another are omitted. The values of the context variable are then taken

into account to evaluate the validity of the rules included in the final rule set. Four possible cases for the context variable as given in Tab. 2 are then applied. Rules that are wrongly added to the final list due to the out-of-range value of the context variable are omitted. Rules that are erroneously omitted on these grounds will be added to the final list.

The proposed algorithm for context-based association rule mining is given in the section below:

Algorithm: Context-based association rule mining

Name: CARM ( )

1: **Inputs:**
- a. SI: Database of Indicators for sustainable energy development given in Tab. 1.
- b. BaseValSupp: Threshold value for support variable
- c. BaseValSuppNeg: Threshold value for support variable of negative association
- d. BaseValConf: Threshold value for confidence variable
- e. BaseValInterest: Threshold value for interestingness variable
- f. ULC: Upper limit for context variable range
- g. LLC: Lower limit for context variable range

2: **Output:**
- a. List of association rules

3. **Begin**

/* The data of sustainability indicators have three dimensions (Year, Country and sustainability indicator). In this loop, it is converted to two-dimensional by averaging each sustainability indicators for all the years) */

4: **For** each country's SI in the list of countries and SIs in the list of SI

5: **For** each year in the list of years

6: $\text{Year-Sum}_{SI} = \text{Year-Sum}_{SI} + \text{year}$

7: $\text{Year-Avg}_{SI} = \text{Year-Sum}_{SI}/\text{total\_number\_of\_years}$

8: Store value of Year-AvgSI in the database for SIs of a country

9: Update database SI

10: **End For**

11: **End For**

12: **While** (No more frequent itemset in SI database)

/* Extract positive and negative frequent itemsets on the basis of frequency of each itemset in the database SI */

13: LPos = Find n-frequent positive itemsets

14: LNeg = Find n-frequent negative itemsets

/* The itemsets which do not qualify the criteria of minimum support are removed from the database */

15: While (No more frequent itemset in LPos and LNeg)

16: PFI = Positive_frequent_itemset U itemsets with Supp > BaseValSupp

17: NFI = Negative_frequent_itemset U itemsets with Supp < BaseValSuppNeg

18: End While

/* The itemsets which do not qualify the criteria of minimum confidence are removed from the database */

19: While (No more frequent itemset in PFI and NFI)

(Continued)

20: PFI = PFI U itemset with Conf > BaseConfVal
21: NFI = NFI U itemset with Conf > BaseConfVal
22: End While
/* The itemsets which do not qualify the criteria of minimum interestingness are removed from the database */
23: While (No more frequent itemset in PFI and NFI)
24: PFI = PFI U itemset with interest > BaseInterestVal
25: NFI = NFI U itemset with interest > BaseInterestVal
26: End While
/* The itemsets which do not qualify the criteria of CPIR are removed from the database */
27: While (No more frequent itemset in PFI and NFI)
28: if CPIR (antecedent, consequent in PFI) < 1
29: Omit the rule from PFI
30: End if
31: if CPIR (antecedent, consequent in NFI) >= 1
32: Omit the rule from NFI
33: End if
34: End While
/* Four possible cases of context variable, two for positive association rules and two for negative association rules are applied in this loop */
35: While (PFI and NFI are not empty)
For PFI
36: if (COA < LLC)
37: Dispos = (LLC − COA) ∗ 100/COA
38: NSVpos = S + (S ∗ Dispos/100)
39: Elseif (COA > ULC)
40: Dispos = (COA − ULC) ∗ 100/COA
41: NSVpos = S + (S ∗ Dispos/100)
42: End if
43: If (NSVpos) < BaseSuppVal
44: Omit the rule from PFI
45: else
46: Add the rule in PFI
47: End if
For Negative Itemset
48: if (COA < LLC)
49: Disneg = (LLC − COA) ∗ 100/COA
50: NSVneg = S − (S ∗ Disneg/100)
51: Elseif (COA > ULC)
52: Disneg = (COA − ULC) ∗ 100/COA
53: NSVneg = S − (S ∗ Disneg/100)
54: End if
55: If (NSVneg) > BaseSuppVal
56: Omit the rule from NFI
57: else
58: Add the rule in NFI
59: End if
60: End While
61: **End**

The time complexity of the proposed algorithm is $O(N^2)$ if one looks at the years and the number of countries. However, if the number of countries is set at its maximum, the time complexity is O(N), where N represents the number of years. The working of the proposed method is given in Fig. 1 below. In Fig. 1, the values of energy indicators are stored in a database that is then discretized to convert the data from conventional numeric format to ordinal format. A frequent itemset is obtained from the dataset based on support, confidence, interestingness and CPIR thresholds. The positive and negative association rules are then mined and evaluated using the values of context variables. The context variable in each dataset is selected by the user/ domain expert. Possibilities/cases in the context are also given in Fig. 1, the details of which appear in the algorithm above.

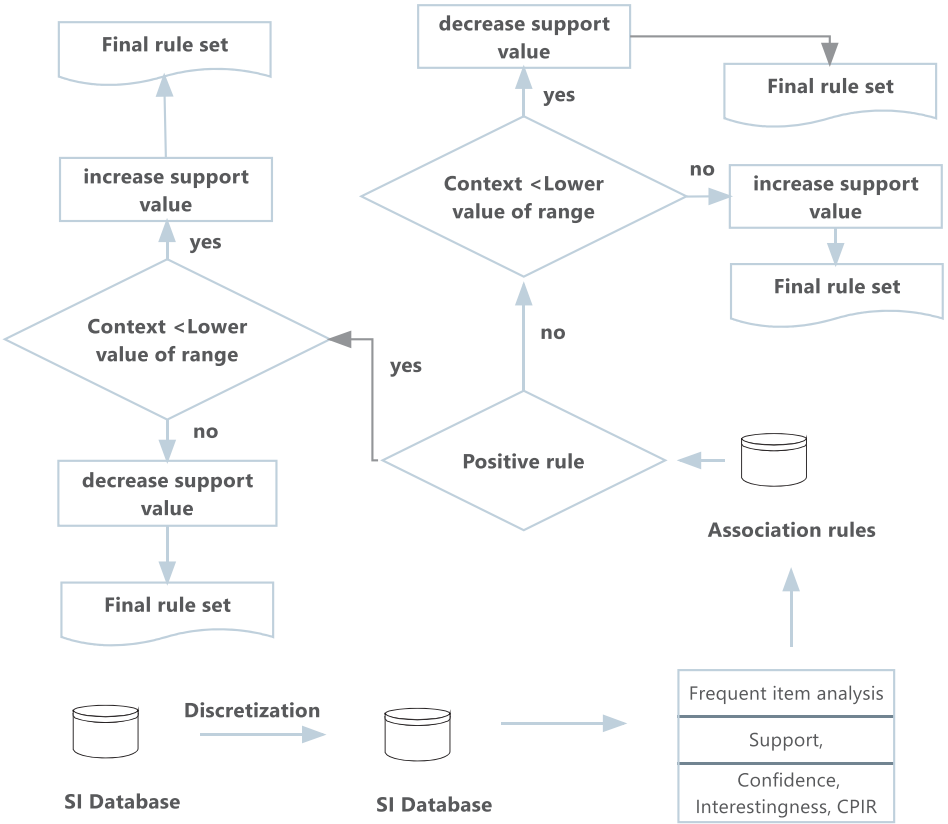


Figure 1: Proposed algorithm and its application in energy database

5 Experiments and Results

The algorithm proposed in the present document is encoded in python Jupyter notebook which is an open-source programming language. The experiment is performed on a machine with an i7-2.11 GHz Processor, 16 GB RAM and 500 GB hard disk installed with all necessary network conditions required for the Windows 10 operating system. Data for 23 sustainable energy development indicators are collected from 28 countries over 25 years from 1990 to 2015. All data is collected from the online energy data portals. Energy sustainability indicators contain quantifiable

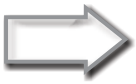
and unquantifiable attributes from which quantifiable attributes are used in this study. Data for the 30 attributes were not available in the online sources, and 23 of the 30 attributes are included in the final database. There were some attributes for which data were not available through online sources but they could be derived from the available attributes. The context variables taken into consideration for the study of sustainable energy development are presented in Tab. 3.

Table 3: Context variables and their normal value ranges

S. No.	Name of context variable	Data type	Category of indicators	Range
1.	Economic recession	Boolean	Economic domain	Normal-not normal
2.	No of power distributors	Numeric	Social domain	1–4 (for 35K people) = normal
3.	Index of pollution	Numeric	Ecological domain	0–0.4 = normal

The data from the first phase of the experiment are averaged and discretized to produce significant associations. As there were three dimensions of the data, the value of sustainability indicator, country and year, so for the discretization, it was necessary to convert the data into two dimensions. The values of each indicator were averaged over 25 years to obtain one value. The process of discretization was straightforward. Range values are determined for all data attributes on which data has been converted from values to ranges. An example of three indicators can be found in Fig. 2.

	SI1	SI2	SI3
C1	6	1.5%	4136.64
C2	0.4	4%	7726.9
C3	1	9%	8643.29
C4	0	7%	7680.57
C5	0.1	0.8%	986.583
C6	0.7	8%	15441.9



	SI1	SI2	SI3
C1	5-10	0-5	0-5000
C2	0-5	0-5	5000-10000
C3	0-5	5-10	5000-10000
C4	0-5	5-10	5000-10000
C5	0-5	0-5	0-5000
C6	0-5	5-10	15000-20000

Discretization of data is done on all the sustainability indicators by manually dividing data into ranges. e.g. the data in SI1 is divided among four ranges i.e. (0-5, 5-10, 10-15, 15-20). For SI2, the ranges are (0-5, 5-10, 10-15 and 15-20) and for SI3 (0-5000, 5000-10000, 10000-15000, 15000-20000)

Figure 2: Discretization of data values (C = country, SI = sustainability indicators)

In Fig. 2, an example of discretization of sustainability indicators for different countries from C1 to C6 is provided. The table on the left shows the non-discretized value that is converted in the table on the right illustrated in Fig. 2. For example, the SI2 value of C1 is converted in the interval 0–5 after discretization. The results shown in Fig. 2 show the relationship between the various energy SIs. The covariance of SI17 with SI19 shows that the greenhouse gas emissions caused by energy products have a strong association with the rate of deforestation caused by the energy products. Based on this pattern, energy decision-maker can build an optimal plan for sustainable energy development in the future. Another issue that the decision-maker can raise relates to the extent of covariance between SI17 and SI19. This can be calculated by using CPIR, interestingness, support and confidence measures.

A significant number of positive and negative association rules were extracted from the dataset using the CARM algorithm. It was nearly impossible to learn from these many rules. Different level of pruning's strategies as described in the proposed method is used. Some of the final rules

extracted after pruning are given in Fig. 3 and the detailed reduction in the number of rules after each pruning level is given in Tab. 4.

In Fig. 3, a snapshot of the extracted rules is given. SI in the figure represents the sustainability indicator and C represents one country. SI3 $\Rightarrow$ SI4 indicates that SI4 varies with SI3 and C1 $\Rightarrow$ C17 indicates an association between two countries represented by C1 and C17. Examples of negative rules from the dataset are also shown in Fig. 3. Tab. 4 summarizes the total number of positive and negative association rules in different scenarios. The results of our algorithm are also compared to some of the existing association rule mining algorithms including Apriori, PNARM, CBPNARM with normal context and CBPNARM with out-of-range context. The results of the algorithm are compared to the number of rules, average confidence of the rules, average dependence and execution time of the algorithms. Two plots Figs. 4 and 5 show the number of rules extracted by different algorithms. Many rules retrieved without applying a pruning measure are shown in Fig. 4. In Fig. 5, the number of rules extracted after the pruning measure is applied are reported. The number of rules extracted in these two plots is at the largest in support of 0.2. The rules extracted by CARM are the minimum although both positive and negative rules are extracted by CARM. The reason there are fewer rules is to include different pruning measures. Including the context variable in this dataset further reduced the number of rules in this case. Some rules were pruned by other pruning measures but the context variable included them in the final list. The numbers of rules extracted by CARM after pruning exceeds that extracted before pruning, which is interesting. The reason for an increase in the number of rules after pruning is to include the context variable. The context variable added few rules to the final set that were not included when the value of the context variable was in the normal range. This is the only case where the number of rules increased after applying the context variable. As mentioned earlier, the context variable can also increase the number of rules by adding those rules in the final rule set that were previously ignored because of the out-of-range value of the context variable. The number of rules applying the pruning measures is shown in Fig. 5. CARM is at the lowest level, which is evident to CARM as it extracts more meaningful rules, which helps to reach the decision.

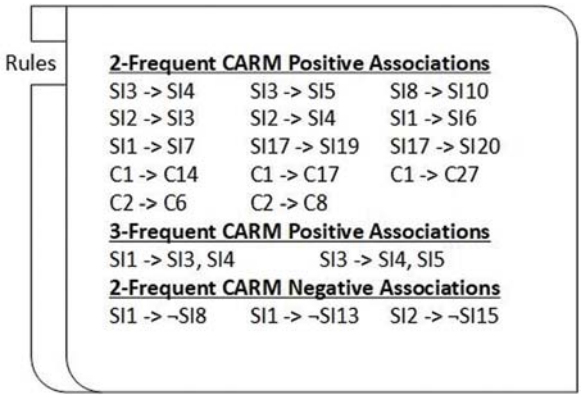


Figure 3: A snapshot of final set of association rules obtained through CARM based on three context variables

Table 4: Number of association rules extracted after applying different levels of pruning (C = confidence, I = interestingness, CP = CPIR, CN = context, N = nil)

Algo	Apriori	Apriori	PNARM	PNARM	PNARM	PNARM	CBPNARM	CBPNARM
Pruning	N	C	N	I	I, C	I, C, CP	CN	I, C, CP, CN
Rules	204	106	438	312	266	186	298	101

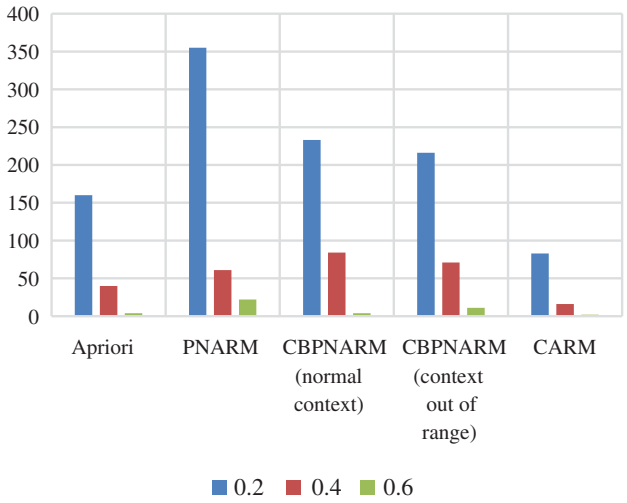


Figure 4: Plot of number of rules for Apriori, PNARM, CBPNARM (normal context), CBPNARM (context out of range), and CARM

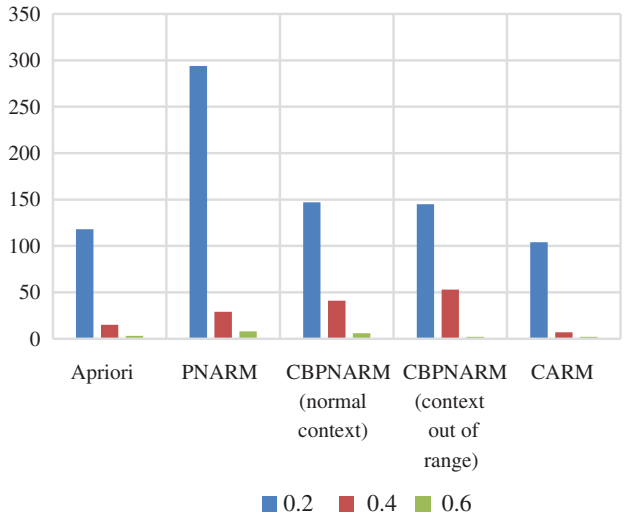


Figure 5: Plot of number of rules after pruning for Apriori, PNARM, CBPNARM (normal context), CBPNARM (context out of range), and CARM

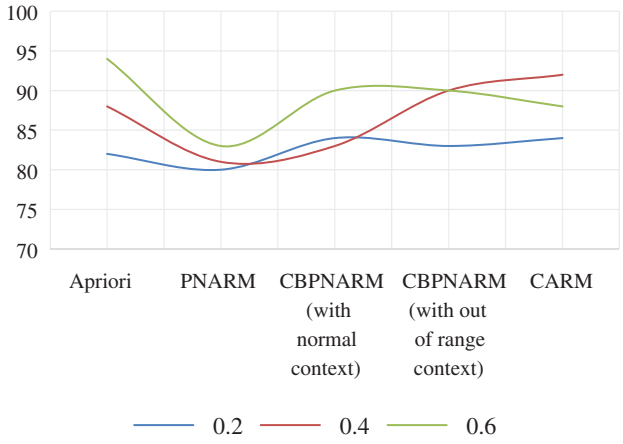


Figure 6: Plot of average confidence for Apriori, PNARM, CBPNARM (normal context), CBPNARM (context out of range), and CARM

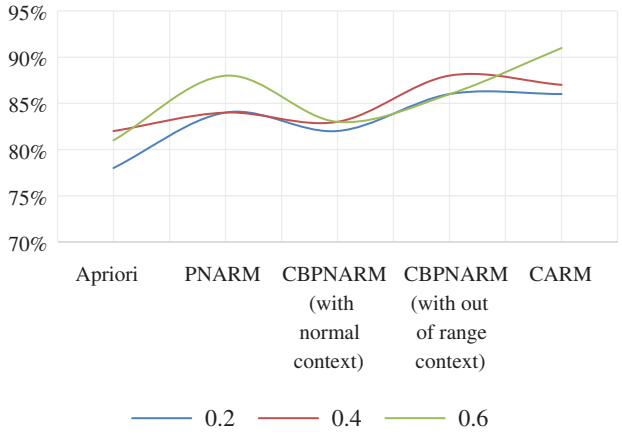


Figure 7: Plot of average confidence after pruning for Apriori, PNARM, CBPNARM (normal context), CBPNARM (context out of range), and CARM

The average confidence graphs for the unpruned and pruned rules extracted through all algorithms are given in Figs. 6 and 7 respectively. The average confidence of the proposed algorithm is greater in most cases. For support values 0.2 and 0.4, it is practically equal to CBPNARM. This is because the context variable value for the sustainability indicators dataset was normal in most cases for which the net impact on the final association rule set was too low. Rules confidence is a predictor of the CARM algorithm producing rules with higher certainty. The higher average confidence value for CARM indicates that the rules extracted are not unfamiliar. There is a higher co-occurrence of antecedent and consequent and the consequent is less escorted by any other antecedent. This proves the certainty about the rules extracted from the database. The average dependence plots are almost the same with pruning and without pruning for which a single plot is given in Fig. 8. PNARM performed best over the given dataset because the PNARM algorithm is intended to maintain rule dependence by interestingness, dependence and CPIR. The proposed algorithm applied interestingness and CPIR for the curve are in the lower

domain of PNARM but higher in comparison with the rest. Fig. 9 illustrates the execution time of all algorithms. After integrating all pruning techniques, the execution time of the proposed algorithm is inferior to PNARM and CBPNARM and superior to the Apriori algorithm. The reason is obvious because the Apriori algorithm mine frequent items only and level-1 pruning is only done in Apriori. PNARM, CBPNARM and CARM are all considering various pruning measures where their execution time is expected to be at the top end. The execution time of CARM is less than CBPNARM because CBPNARM was developed for spatial data that is a pseudo form of image data.

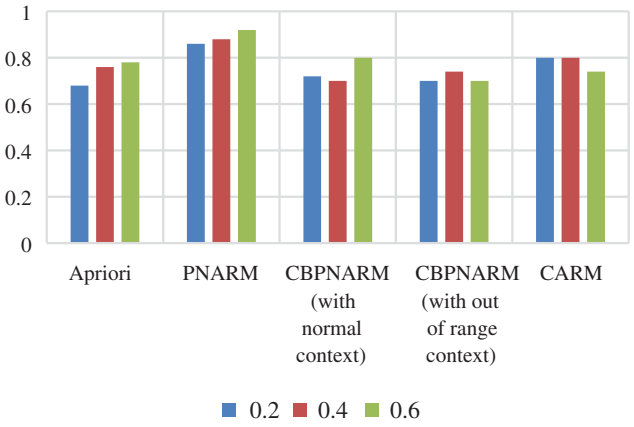


Figure 8: Plot of average dependence for Apriori, PNARM, CBPNARM (normal context), CBP-NARM (context out of range), and CARM

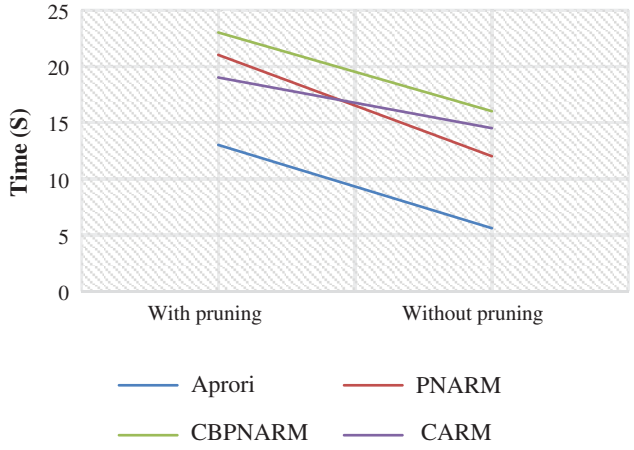


Figure 9: Plot of execution time for Apriori, PNARM, CBPNARM and CARM

CARM with additional pruning measure took lesser execution time than CBPNARM. The execution time of CARM is at last but one position if no pruning technique is used for asso-ciation rule mining. The algorithm was designed to improve the quality of association rules extracted from the datasets for which comparing algorithms based on precision, recall and F-measure depicted a clearer picture. The comparison of the algorithms based on average values

on multiple energy datasets is shown in Tab. 5. The rules extracted from the dataset is divided into true positives, false positives, true negatives and false negatives according to the measures above. Higher precision, recall, and F-measure for the CARM algorithm indicate that the algorithm has extracted more useful rules. The values given in Tab. 5 are calculated by comparing the result of the algorithms for extracting association rules with the real rules that are used in the energy sector and validated by the expert of the domain.

Table 5: Precision, recall, F-measure for evaluation of the extracted rules

	TP rate	FP rate	Precision	Recall	F-Measure	ROC area	PRC area
Apriori	0.52	0.16	0.764	0.912	0.831	0.433	0.837
PNARM	0.54	0.3	0.642	0.984	0.777	0.337	0.652
CBPNARM	0.74	0.014	0.981	0.922	0.950	0.925	0.998
CARM	**0.78**	**0.006**	**0.992**	**0.938**	**0.964**	**0.998**	**0.999**

TP rate: rate of true positives, FP rate: rate of false positives (instances falsely extracted as a rule), ROC/PRC area: trade-off between true positive and false positive rates. Whereas;

$$Precision = \frac{relevant\ retrived\ rules}{retrieved\ rules}$$

$$Recall = \frac{relevant\ retrived\ rule}{relevant\ rules}$$

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

6 Conclusion and Future Work

The CARM algorithm for mining context-based association rules is proposed in this paper as an extension of the CBPNARM algorithm. A few association rule pruning techniques are incorporated into the CARM algorithm including confidence, interestingness and CPIR to improve insights by decreasing the number of rules extracted. The context is used in the algorithm to eliminate certain rules and/or add those excluded from the final rule set defined based on the out-of-range-value of the context variable. The algorithm is applied to sustainable energy indicators to find co-varying sustainability indicators and countries for sustainable energy development. The rules produced by CARM are more robust, relevant and insightful in terms of average confidence, dependence and relevance.

The proposed method outperformed the previous methods in terms of the number of rules generated, confidence and dependency. The inclusion of the context variable and CPIR reduced the number of rules and increased the robustness and usability of the rules. Confidence and dependency values show that fewer rules do not suggest a loss of useful patterns. The execution time of the algorithm is higher than a few other algorithms, which is expected due to additional functions added for the context variable and CPIR. The complexity of the algorithm can be improved in future by using object-oriented approaches for context variable and CPIR.

The results obtained in terms of the application domain of sustainable energy development are also insightful and reported interesting covariances in the indicators and underlined the criticality of some countries for their energy development. The energy sector in a country can use associations derived from the proposed method to construct an optimal plan to ensure sustainable energy development. The associations among sustainability indicators can lead the

energy sector to devise a plan according to the individual deficiencies of energy development and its relation with other developmental factors. Thus, the study can lead an energy sector to achieve optimal energy development without compromising the economy, ecology and social justice that are essential ingredients for sustainability. The work can be extended to automate the selection of context variable because manually selecting context variables can add some bias to the results. An automated mechanism interpreting negative association rules can also be added to the algorithm in future work. Different classification algorithms and learning approaches can be added to the system to reduce the complexity arising from the data structure.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Shaheen, M. Shahbaz, Z. Rehman and A. Guergachi, "Data mining applications in hydrocarbon exploration," *Artificial Intelligence Review*, vol. 35, no. 1, pp. 1–18, 2010.

[2] M. Shaheen and M. Z. Khan, "A method of data mining for selection of site for wind turbines," *Renewable and Sustainable Energy Reviews*, vol. 55, pp. 1225–1233, 2017.

[3] S. H. Liao, P. H. Chu and P. Y. Hsio, "Data mining techniques and applications—A decade review from 2000 to 2011," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11303–11311, 2017.

[4] A. Azevedo, "Data mining and knowledge discovery in databases," in *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics*. Pennsylvania, United States: IGI Global, pp. 502–514, 2019.

[5] X. Wu, C. Zhang and S. Zhang, "Efficient mining of both positive and negative association rules," *ACM Transactions on Information Systems*, vol. 22, no. 3, pp. 381–405, 2004.

[6] K. Vadim, "Overview of different approaches to solving problems of data mining," *Procedia Computer Science*, vol. 123, pp. 234–239, 2018.

[7] M. Shaheen, M. Shahbaz and A. Guergachi, "Context based positive and negative spatio-temporal association rule mining," *Knowledge-Based Systems*, vol. 37, pp. 261–273, 2013.

[8] M. Shaheen, M. Shahbaz, A. Guergachi and Z. Rehman, "Mining sustainability indicators to classify hydrocarbon development," *Knowledge-Based Systems*, vol. 24, no. 8, pp. 1159–1168, 2011.

[9] S. Ray and M. Saeed, "Applications of educational data mining and learning analytics tools in handling big data in higher education," *Applications of Big Data Analytics*, pp. 135–160, 2018. https://www.springerprofessional.de/en/applications-of-big-data-analytics/15972258.

[10] C. Romero and S. Ventura, "Data mining in education," *WIREs Data Mining and Knowledge Discovery*, vol. 3, pp. 12–27, 2013.

[11] M. Shahbaz, A. Guergachi, A. Noreen and M. Shaheen, "A data mining approach to recognize objects in satellite images to predict natural resources," In: G. C. Yang, Ao S., Gelman L. (Eds.) *IAENG Transactions on Engineering Technologies*, Lecture Notes in Electrical Engineering, vol. 229. Dordrecht: Springer, pp. 215–230, 2013. https://doi.org/10.1007/978-94-007-6190-2_17.

[12] M. Shaheen and M. Shahbaz, "An algorithm of association rule mining for microbial energy prospection," *Scientific Reports*, vol. 7, no. 1, pp. 68, 2017.

[13] S. G. Alonso, I. D. I. Torre-Diez, S. Hamrioui, M. Lopez-Coronado, D. C. Barreno *et al.,* "Data mining algorithms and techniques in mental health: A systematic review," *Journal of Medical Systems*, vol. 42, no. 9, pp. 161, 2018.

[14] D. AI, H. Pan, X. Li, Y. Gao and D. He, "Association rule mining algorithms on high-dimensional datasets," *Artificial Life and Robotics*, vol. 23, pp. 420–427, 2018.

[15] S. Cokpinar and T. I. Gundem, "Positive and negative association rule mining on XML data streams in database as a service concept," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7503–7511, 2012.

[16] S. Mahmood, M. Shahbaz and A. Guergachi, "Negative and positive association rules mining from text using frequent and infrequent itemsets," *Scientific World Journal*, vol. 973750, no. 2, pp. 1–11, 2014.

[17] I. Batyrshin, D. S. Pamucar, P. Crippa and F. Liu, "Select actionable positive or negative sequential patterns," *Journal of Intelligent and Fuzzy Systems*, vol. 29, no. 6, pp. 2759–2767, 2015.

[18] R. Anuradha and N. Rajkumar, "Mining generalized positive and negative inter-cross fuzzy multiple-level coherent rules," *Journal of Intelligent and Fuzzy Systems*, vol. 32, no. 3, pp. 2269–2280, 2017.

[19] A. Aggarwal and D. Toshniwal, "Frequent pattern mining on time and location aware air quality data," *IEEE Access*, vol. 7, pp. 98921–98933, 2019.

[20] Cowen and J. David, "Geospatial metadata," in *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, Boston, MA: Springer, pp. 1–6, 2016.

[21] R. Agarwal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. of the 20th VLDB Conf.*, Santiago, Chile, vol. 487, pp. 499, 1994.

[22] D. J. Prajapati, S. Garg and N. C. Chauhan, "Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment," *Future Computing and Informatics Journal*, vol. 2, no. 1, pp. 19–30, 2017.

[23] United Nations Department of Economics and Social Affairs, "International energy agency, Eurostate, and European environment agency," in *Energy Indicators for Sustainable Development: Guidelines and Methodologies*. Vienna: International Atomic Energy Agency, 2005.

[24] United Nations (UN), *Report of the World Summit on Sustainable Development, A/CONF. 199(20)*. New York: United Nations, 2002.

[25] B. K. Sovacool and I. Mukherjee, "Conceptualizing and measuring energy security: A synthesized approach," *Energy*, vol. 36, no. 8, pp. 5343–5355, 2011.

[26] U. Al-Mulali, B. Saboori and Ozturk, "Investigating the environmental Kuznets curve hypothesis in Vietnam," *Energy Policy*, vol. 76, no. 3, pp. 123–131, 2015.

[27] S. Pachauri and D. Spreng, "Measuring and monitoring energy poverty," *Energy Policy*, vol. 39, no. 12, pp. 7497–7504, 2011.

[28] M. Salahuddin, K. Alam, I. Ozturk and K. Sohag, "The effects of electricity consumption, economic growth, financial development and foreign direct investment on CO_2 emissions in Kuwait," *Renewable and Sustainable Energy Reviews*, vol. 81, no. 2, pp. 2002–2010, 2018.

[29] J. W. Moon and J. Ahn, "Improving sustainability of ever-changing building spaces affected by users' fickle taste: A focus on human comfort and energy use," *Energy and Buildings*, vol. 208, pp. 109662, 2020.

[30] Y. Dong and M. Z. Hauschild, "Indicators for environmental sustainability," *Procedia CIRP*, vol. 61, pp. 697–702, 2017.

[31] Y. Liu, P. Xie, Q. He, X. Zhao, X. Wei *et al.,* "A new method based on association rules mining and geo-filter for mining spatial association knowledge," *Chinese Geographical Science*, vol. 27, no. 3, pp. 389–401, 2017.

[32] F. Yoseph and M. Heikkila, "A new approach for association rules mining using computational and artificial intelligence," *Journal of Intelligent and Fuzzy Systems*, Prepress, pp. 1–14, 2020.

[33] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier, Online, pp. 243–278, 2012.

[34] L. K. Sharma, O. P. Vyas, U. S. Tiwary and R. Vyas, "A novel approach of multilevel positive and negative association rule mining for spatial databases," In: P. Perner, A. Imiya (Eds.) *Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science*, vol. 3587. Berlin, Heidelberg: Springer, pp. 620–629, 2005.

[35] W. Yang, Q. Liao and C. Zhang, "An association rule mining algorithm on context-factors and users preference," in *Proc. 2013 5th Int. Conf. on Intelligent Human–Machine Systems and Cybernetics*, China, pp. 190–195, 2013.