Knowledge-Based Systems 24 (2011) 1159–1168

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

# Mining sustainability indicators to classify hydrocarbon development

Muhammad Shaheen [a,*], Muhammad Shahbaz [a,*], Aziz Guergachi [b], Zahoorur Rehman [a]

[a] Department of Computer Science & Engineering, University of Engineering & Technology, Lahore, Pakistan
[b] Information Technology Management Ryerson University, Toronto, ON, Canada

ARTICLE INFO

*Article history:*
Received 8 December 2010
Received in revised form 23 April 2011
Accepted 23 April 2011
Available online 29 April 2011

*Keywords:*
Sustainability indicators
Clustering
Decision Tree
Data mining
Hydrocarbon development
Energy development

ABSTRACT

The role of energy in economic, social and ecological development of a country defines its significance in sustainable development. We propose here a method to classify a nation's hydrocarbon development into one of five classes: (1) futuristic; (2) conforming; (3) sustainable; (4) unsustainable; or, (5) critical. *K* means clustering is a method of unsupervised classification in which the clusters cannot be labeled due to their lack of a class value. We propose a unique method to label unsupervised classes which is then used to divide the energy data of nations into five clusters. The labeled clusters are structured in an ID3 decision tree which provides a hierarchical structure to evaluate the hydrocarbon development in a given country. The results indicate some useful and interesting patterns in sustainability indicators.



## 1. Introduction

Sustainability is defined as, "The development that meets the needs of today without compromising the ability of future generations to meet their needs" [4]. Energy plays vital role in the socio-eco-economic development of a country. Energy is globally available in different forms; the most common energy is derived from hydrocarbons. In our energy based global economy, energy providers are desperately looking for ways to extract hydrocarbons to meet the needs of consumption. One third of the world's population relies on the use of animal power and non-commercial fuels and almost two billion people lack access to electricity [32]. Energy is precious as it leads to better living standards, health, environment and prosperity but two key questions remain: Why are the large reserves of energy in developing countries failing to bring remarkable change in the energy dependent socio-eco-economic dimensions? And, despite adequate energy planning, why do energy extraction and distribution practices still lead to wide-scale economic recession?

There is an absence of a procedure to assess energy development in any given country. In 2001, at a world summit on sustainable development in South Africa, 41 indicators for sustainable energy development were proposed [35], but only some of these variables are quantifiable. Many of the indicators are not directly related to energy development because energy is often framed only as a means to an end. The end is sustainable development for a nation's economy, ecology and social welfare. An annotated list of these indicators with descriptions is presented in Table 1.

The meaning of sustainable development varies in different contexts. In the case of energy, development is considered to be sustainable if the consumption rate conforms to the production rate. The first Enquête Commission formulated four rules of sustainability which emphasize the need for environmental protection and growth in natural resources. The commission also focused on potential effects of the above synergy on social and economic conditions [18]. One of the rules concerns conformity in consumption and production of energy resources. An imbalance between these factors may cause the depletion-midpoint of crude oil to be between 2010 and 2020 [30]. However, the natural gas market is younger and its depletion is not expected to occur in the next 60 years.

Data mining has increased the opportunities for decision makers to extract useful implicit knowledge from a large pool of collected data [29]. Data mining of substantial datasets results in the supervised/unsupervised classification of datasets or the prediction of an unknown real value [3,6]. Clustering is an unsupervised data mining technique that is used to classify patterns into clusters. A number of methods are proposed for clustering different types of datasets [8,26,16].

There is currently no method to assess a country's hydrocarbon development using the indicators put forward in the IAEA's sustainability indicators proposed in 2001. We propose a method to

* Corresponding authors. Tel.: +92 3314525045 (M. Shaheen), +92 3027424229 (M. Shahbaz).
*E-mail addresses:* shaheen@uet.edu.pk (M. Shaheen), m.shahbaz@uet.edu.pk (M. Shahbaz), a2guerga@ryerson.ca (A. Guergachi), xahoor@gmail.com (Z. Rehman).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.04.016

**Table 1**
Sustainability Indicators proposed by IAEA/EIA.

| S. no. | Indicator name | Description |
|---|---|---|
| 1. | Total Population | Total population |
| 2. | GDP Per Capita | Values of goods produced per person |
| 3. | Distance Traveled per Capita | Total distance traveled per capita |
| 4. | Agriculture Value Added | Contribution of agriculture |
| 5. | Industry Value Added | Contribution of industry to overall GDP |
| 6. | Services Value Added | Contribution of services to overall GDP |
| 7. | Energy Intensity | Energy efficiency of nation's economy |
| 8. | Energy Consumption per capita | Total energy consumed against each capita |
| 9. | Total Primary Energy Consumption | Total consumption of hydrocarbons |
| 10. | Energy Supply Efficiency | Supply of energy with respect to demand |
| 11. | Energy Use Per unit of GDP | Consumption of energy against each unit GDP |
| 12. | Expenditure on Energy Sector | What% of GDP is spent on Energy Sector |
| 13. | Ambient concentration of pollutants | Concentration of pollutants in atmosphere |
| 14. | Total Primary Energy Production | Total production of hydrocarbons |
| 15. | Net Energy Import Dependence | Need of quantity of energy to be imported |
| 16. | Ratio of Daily Disposable Income to Prices of Fuels | Ratio of income with prices of fuels |
| 17. | Daily Disposable Income | Income available for saving and spending |
| 18. | Quantities of Air Pollutant Emissions | Quantity of air pollutant in atmosphere |
| 19. | Urban Population | Total population in cities |
| 20. | Quantities of Green House Emissions | Greenhouse emissions in atmosphere |
| 21. | Lifetime of Proven Reserves | Expected life of hydrocarbon reserves |
| 22. | Proven Uranium Reserves | Expected life of uranium reserves |
| 23. | Net Nuclear Power Generation | Consumable nuclear power |
| 24. | Intensity of Use of Forest Resources | Utilization of forest resources |
| 25. | Rate of Deforestation | Rate of deforestation |
| 26. | Proven fossil fuel reserves | Quantity of hydrocarbon reserves |
| 27. | $CO_2$ Emissions | Quantity of $CO_2$ released to atmosphere |

approach the sustainability indicators by classifying countries into five clusters. As discussed earlier, clustering is an unsupervised classification technique. In unsupervised classification, the resulting classes do not have a class label. The clusters, if not labeled, will consist of few countries and will lack any assessment about those countries or how they are grouped and according to which dimensions of similitude. Hence, to identify the type of energy development in each of the five clusters, we devise a method of labeling unsupervised classes. Using this method, each cluster groups similar countries out of a total of 40, and each country has 27 indicator values. We have applied correlation analysis between (1) *sustainability indicators and total production*; and 2) *sustainability indicators and total consumption*.

In each cluster, the average correlation value of a sustainability indicator with production and consumption rate is multiplied by the actual value of the indicator. All indicators are added then to assign a single value to a cluster. The clusters can be sorted in order of precedence and five class values result: (1) futuristic; 2) conforming; 3) sustainable; 4) unsustainable; or 5) critical. These form the labeled clusters. We then applied a ID3 decision tree classification algorithm to the clusters to simulate sustainability indicators on a hierarchical structure for ease of its tracking and use when assessing a country's energy development. The decision tree will also help decision makers to identify weak dimensions. Thus, using this technique, a decision maker will be able to address the question, "What dimensions, with respect to sustainability indicators, should be ameliorated to make energy development sustainable?"

The paper is organized as follows. Section 2 reviews the key literature on clustering and decision tree classification. We develop a method for clustering world countries with respect to energy development in Section 3. In Section 4, we perform experiments and finally the work is concluded in Section 5.

## 2. Clustering and decision tree classification

### 2.1. Clustering

The classification techniques in data mining are divided into two groups: (1) supervised classification and (2) unsupervised classification. In supervised classification, the user is provided with inputs bearing class labels while unsupervised classification does not apply class labels to input data. Unlabeled patterns of data are grouped by using clustering techniques for which it is said to be unsupervised classification technique.

A typical clustering activity would involve the following:

1. pattern representation;
2. definition of pattern proximity measure; and
3. grouping and data abstraction.

A number of approaches have been adopted to cluster larger datasets. For example, Gao and Hitchcock [9] offer taxonomy of these approaches [9]. In our study, we use squared error partitioned *K*-mean clustering.

#### 2.1.1. K-mean clustering

Consider a dataset with multiple points in Cartesian space, $DS = A, B, C, \ldots, Z$ where each point within a dataset is represented as $A(xa_1, xa_2, xa_3, \ldots, xa_n)$, $B(xb_1, xb_2, xb_3, \ldots, xb_n)$, $C(xc_1, xc_2, xc_3, \ldots, xc_n), \ldots, Z(xz_1, xz_2, xz_3, \ldots, xz_n)$. These datasets are to be allocated to $k$ number of clusters. Let $pts_b$ be the number of points assigned to cluster b. Following are the steps of *K*-mean clustering algorithm taken from Al-Sultan and Khan [2].

*Step 1*: At the first step, select cluster centers randomly

$$\text{ClusterCenters} = CS_p \text{ where } p = 1, 2, \ldots, n$$

and

$$CS_p \epsilon [A(xa_1, xa_2, xa_3, \ldots, xa_n), B(xb_1, xb_2, xb_3, \ldots, xb_n), C(xc_1, xc_2, xc_3, \ldots, xc_n), \ldots, Z(xz_1, xz_2, xz_3, \ldots, xz_n)]$$

*Step 2*: Find the Euclidean distance (ED) of each data point from each cluster center and allocate the dataset to the closet cluster center

$$ED(A(xa_1, xa_2, xa_3, \ldots, xa_n), B(xb_1, xb_2, xb3, \ldots, xb_n), C(xc_1, xc_2, xc_3, \ldots, xc_n), \ldots, Z(xz_1, xz_2, xz_3, \ldots, xz_n), CSp) = |(xa_1 - xp_1)^2 + (xa_2 - xp_2)2 + \ldots + (xa_n - xp_n)^2|$$

*Step 3*: Calculate the new cluster centers [2].

$$CS_p^* = \left(\frac{1}{pts}\right) \sum_{index(pts)=i=1}^{n} X_{a1} \quad \text{if } pts > 0$$

*Step 4*: Calculate the criterion function

$$CF = \sum_{i=1}^{c} \cdot \sum_{index(pts)=i=1}^{n} \left\| X_{a_i} - CS_p^* \right\|^2$$

*Step 5*: Go to Step 3 to compute the new assignments.

The *K*-means algorithm divides the datasets into the desired number of clusters and converges to a local minimum [28,9]. The algorithm does not specify anything about the labels of clusters. There are certain applications where labels of clusters are required in order to make results useful in decision making. Detailed information on the *K*-means algorithm is available in the existing literature [20,15].

### 2.2. Decision tree classification

A decision tree is a hierarchical structure that corresponds to the sequence of decision rules [11]. The tree is built by subdividing the training set on the basis of a criterion. DT is one of the popular methods of learning and reasoning from feature based examples [5]. These trees are constructed to help actors to make decisions. It has also been considered a predictive model. The leaves of the tree represent classification and branches represent classes of attributes that lead to classification [36]. Different algorithms ID3, C4.5 are used for decision tree classification. Decision tree also produced some improvement in performance of industrial applications by using it in combination with other learning techniques [19,21,13,37].

In order to construct a decision tree using ID3, the dataset should be labeled with class outputs which lead the decision tree to be a supervised classification method. Entropy is a measure used to gauge the level of disorder in the dataset. In decision tree construction, entropy is used as an attribute selection measure [36]. The step wise procedure of the decision tree construction is as stated below [14].

Let $X$ be the subset dataset containing n samples. $X = \{x_1, x_2, \ldots, x_n\}$. Suppose that the dataset $X$ is labeled by $k$ distinct class labels $(1, 2, \ldots, k)$. Let xi be the number of samples of $X$ having class label $i$.

*Step 1*: Calculate the information gain (IG) of each attribute in dataset $X$. Select test attribute at each node of the tree

$$IG = -\sum_{i=1}^{n} P_i \log_2(p_1)$$

Where Pi is the probability of a sample belonging to a particular class and Pi = $x_i/x$

$$IG = -\sum_{i=1}^{n} \frac{x_i}{x} \log_2\left(\frac{x_1}{x}\right)$$

*Step 2*: Calculate the entropy (E) based on dataset partitioning

$$E = -\sum_{i=1}^{n} \frac{x_{1i} + \cdots + x_{mi}}{x} * IG$$

*Step 3*: Calculate net gain (NG) of the attribute

$$NG = (IG - E)$$

*Step 4*: The attribute with the largest gain value will be selected on the node.

The details of the same algorithm can be found in Han and Kamber [14]. The algorithm will produce a logical decision tree from the dataset to support assessing any decision on to the branches of the tree.

## 3. Related work

Sustainable development of energy is vital for eradicating poverty, improving living standards and human welfare [32]. The vitality of sustainable development is understood at a time when unsustainable development causes much anguish for people all over the world. Energy is a key contributor in the scenario because most of the patterns in energy demand and supply are unsustainable. There are some listed variables which help us to assess the development either sustainable or unsustainable.

The Commission on Sustainable Development (CSD) at the World Summit in South Africa [33] in 2001 provided a list of indicators for sustainable development [35]. UNDESA's indicators [34] were also considered in the discussion. UNDESA produced 58 indicators for sustainable development out of which only three are related to energy resources. The International Atomic Energy Agency (IAEA) produced a set of 41 indicators for sustainable energy development [35]. The original set of indicators were developed by justly considering four, not three, dimensions of sustainable development: (1) economic; 2) social; (3) environmental; and (4) institutional. The development of energy indicators was a collaborative effort by UNDESA, IEA, Eurostate and the European Environment Agency (EEA) [17]. The indicators are analyzed using statistical tools to explore future developments in the context of energy policy and national priorities [17].

The existing literature lacks suitable guidance on the application of data mining techniques on sustainability indicators in order to assess a particular country's hydrocarbon development. Further, there are no cited reports of an application of any type of statistical or intelligent techniques on sustainability indicators to evaluate the current standings of a country's energy sector. However, in our study, we have reviewed some related literature in which the application of data mining techniques to assist energy sector is discussed broadly and cursorily in a specific country or at fine-grain level in an industry.

Mostafa et al. used multilayer perception neural network, regression neural network and probabilistic neural network to classify the ecological footprint of 140 nations. The classification accuracy of three methodologies is compared by using accuracy indices. The study showed suitability of neuro-computational techniques over traditional statistics [24]. Rodriguez-Ortiz et al. classify energy consumption using a data mining technique based on an adaptive resonance theory (ART) algorithm modified with a Euclidean distance measure. The results of using data mining were quick retrieval and better visualization of results [25]. Graphet Inc. has developed algorithms for clustering, regression, classification and association rule mining for analyzing energy usage pattern, key indicators to predict energy usage, critical events in system and equipment stable modes and of operations [10].

Didem and Gulgun proposed a technique for scenario analysis by using Bayesian networks. A forecasting decision model based on scenario analysis of renewable energy resources is proposed and various scenarios of Turkey's energy sector are tested to aim sustainability [7]. Maricar et al. utilize data mining techniques in an energy audit to optimize the use of energy, hence ensuring comfort [23]. TSO and Yau [31] detected patterns in domestic energy usage in Hong Kong by using data mining. [1] used general and stepping regression techniques for forecasting electrical energy consumption in Saudi Arabia.

The contribution in this paper is multifold. The application of data mining techniques on sustainability indicators is unique. Two techniques, one for clustering time series data and the other

for labeling unsupervised classes are also proposed here. In the literature, the time series classification technique is proposed by Gregorio and Lacus [12]. In the proposed technique the stochastic differential equations are observed at discrete times and dissimilarity measures among objects are based upon distance between markov operators. The technique is applied on real financial data and will specify the distances to create final clusters. A brief overview of extant techniques is also provided in the same paper.

## 4. Proposed methodology

As mentioned, the IAEA proposed 41 sustainability indicators were proposed in 2002. All of these indicators are not quantifiable. We have selected 27 indicators (as shown in Table 1) which are quantifiable and available.

Let $SI_n$ represents the value of sustainability indicators where $n = 1, 2, \ldots, 27$. Let $WC_n$ represents countries where $n = 1, 2, \ldots, 40$. Let $C_1$, $C_2$, $C_3$, $C_4$, $C_5$ represent five clusters. $WC_n$ are classified into five clusters for every value of $SI_n$ by using $K$-mean clustering. The dataset of sustainability indicators do not contain any class value on the basis of which supervised classification may be preceded. To group the data into unlabeled classes, $K$-mean clustering is the simplest, fastest and most reliable technique and it is the most commonly used technique for clustering. The technique is not favored when difficult to define the mean, but this is not the case is with provided datasets [22]. Each $WC$ in $WC_n$ can be allocated to different clusters in different years based on the values of sustainability indicators. To select unitary belonging of each WC to one of the clusters, we used the frequent membership rule. The $WC$ is allocated to the cluster to which it belongs more frequently for different sustainability indicators. The same is illustrated in Eq. (1)

$$\forall WC \in WC_n \text{ and } \forall C_i \in C_n \therefore WC \in C_i \text{ iff}[membership(WC \to C_i)] > [membership(WC \to C_n) \text{ where } n = 1,2,3,4,5 \text{ and} i \text{ is the index of allocated cluster.} \quad (1)$$

### 4.1. Labeling clusters

The membership procedure devised above is used to place a country in its appropriate cluster. These clusters do not have an identification label because unsupervised classification does not have a class output. In this work, the energy development of each country is to be classified in one of the five clusters: (1) futuristic; (2) conforming; (3) sustainable; 4() unsustainable; and (5) critical. In order to allocate these labels we devised a methodology in which the correlation value of each of the sustainability indicators with total production and consumption is calculated. The correlation value can be calculated by using Eqs. (2) and (3)

$$Corr(Production, SI) = \frac{\sum_{i=1}^{n} p(Production_i, SI)}{n} \quad (2)$$

$$P(production, SI) = \frac{V(Production, SI) - V(Production)V(SI))}{V(Production^2) - V^2(Production)\sqrt{V(SI^2) - V^2(SI)}} \quad (3)$$

The class output can have five values in which three—futuristic, conforming and sustainable—point to sustainable development and two—unsustainable and critical—point to unsustainable development. As mentioned earlier, the balance between energy production and consumption will ensure sustainability. Hence each correlation value of sustainability indicator with energy production and consumption bears weight value $W$ for that indicator. Each of the clusters contains only a few countries where 27 indicator values are associated with each. The weight value $W$ of each indicator is multiplied with the actual value of that indicator. All the resulting indicator values are summed up and allocated to $Weighted\_Sum_c$. Let $New\_SI_n$ is the new value of sustainability indicator such that;

$$New\_SI_n = W * SI_n \quad (4)$$

For each WC

$$Weighted\_sum_{WC} = \sum_{i=1}^{n} new\_SI_i \quad (5)$$

where $n = 1, \ldots, 27$.

The clusters are then sorted with respect to $Weighted\_Sum_{WC}$ in descending order. The cluster with the highest value of $x$ is labeled as Conforming and the remaining clusters are labeled accordingly.

Once the clusters are labeled, the association of a particular country to one of the five classes can then be evaluated using the membership function defined in Eq. (1).

### 4.2. Decision tree

The labeled clusters are again stored in the database with corresponding output labels extracted from the above method. The organized data is clustered and is not easy to visualize because of one or both of the following reasons:

- The decision maker is not aware of clustering methodology and interpretation. Visualizing the data to prepare an energy development plan is not feasible.
- The data, organized into clusters, do not suggest an improvement in an energy development plan as the indicators cannot be tracked individually in order to assess overall energy development.

It is better to plot clustering results onto a decision tree thus making all sustainability indicators explicit to the decision maker(s). The decision tree will enable one to address individual sustainability indicators to improve energy development in a given country. The clustering results of Section 4.1 classify each country into one of the predefined classes. After assessing the energy development of the country, the decision maker will wish know the reasons for unsustainable or critical development. A decision tree will map all sustainability indicator rules to assist the decision maker(s) in preparing a better energy development plan. The data of 27 sustainability indicators for 40 countries over a period of 30 years is clustered in the previous section. The data of 30 years against each of the sustainability indicators is averaged to find an aggregate value for each. After labeling the clusters, the same data is labeled with class values and is stored in the database. The labeled data is mapped onto a decision tree using the ID3 algorithm [14]. Decision tree representation helps to develop an approximation discrete function that produces some useful expressions. An ID3 algorithm is used in this study because it generates shorter and more compact trees. The attributes with lower entropies can be found near the root of the tree. The algorithm is ideal for numeric attributes [14,27]. The proposed approach is shown in Fig. 1.

## 5. Results and discussion

The data of 27 energy sustainability indicators of 40 countries is collected for a period spanning 30 years. Twenty-seven out of 41 sustainability indicators are selected because the remaining indicators were either not quantifiable or the data related to those sustainability indicators was not available. The data is collected from the Energy Information Administration, USA and the World Bank web server. The data is stored in the database as shown in Table 2.

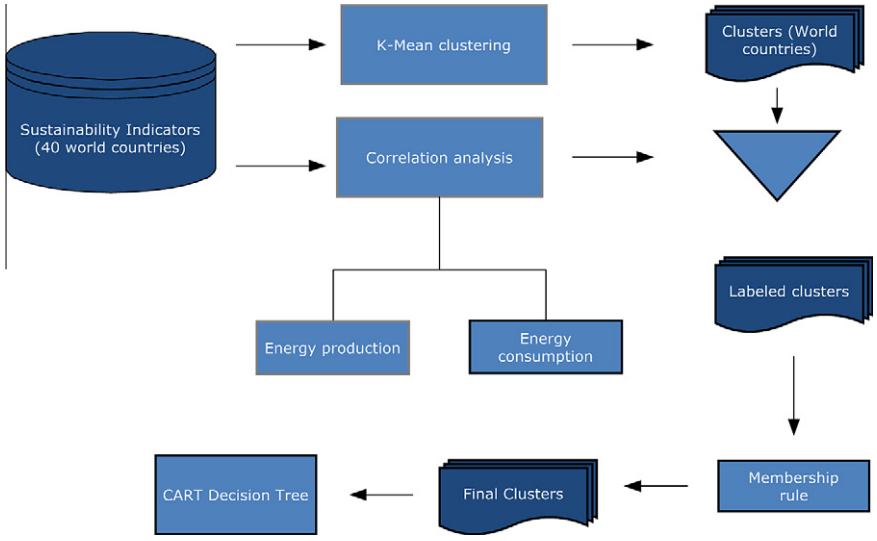


**Fig. 1.** The proposed approach.

**Table 2**
Orientation of sustainability indicators database.

| 1980 ,. . . , 2008 | | | | |
|---|---|---|---|---|
| | $SI_1$ | $SI_2$ | . . . | $SI_n$ |
| $WC_1$ | $SI_1WC_1$ | $SI_2WC_1$ | . . . | $SI_nWC_1$ |
| $WC_2$ | $SI_1WC_2$ | $SI_2WC_2$ | . . . | $SI_nWC_2$ |
| . . . | $SI_iWC_i$ | | | |
| $WC_n$ | $SI_1WC_n$ | $SI_2WC_n$ | . . . | $SI_nWC_n$ |

**Table 3**
Correlation values of sustainability indicators with energy production and consumption.

| S# | Indicator name | Production | Consumption | Average |
|---|---|---|---|---|
| 1. | Total Population | 0.978 | 0.980 | 0.979 |
| 2. | GDP Per Capita | 0.913 | 0.901 | 0.907 |
| 3. | Distance Traveled per Capita | 0.975 | 0.866 | 0.920 |
| 4. | Agriculture Value Added | 0.654 | 0.750 | 0.702 |
| 5. | Industry Value Added | 0.772 | 0.648 | 0.710 |
| 6. | Services Value Added | 0.942 | 0.910 | 0.926 |
| 7. | Energy Intensity | 0.998 | 0.971 | 0.984 |
| 8. | Energy Consumption per capita | 0.224 | 0.230 | 0.227 |
| 9. | Total Primary Energy Consumption | 0.882 | 1.000 | 0.941 |
| 10. | Energy Supply Efficiency | 0.407 | 0.474 | 0.440 |
| 11. | Energy Use Per unit of GDP | 0.224 | 0.230 | 0.227 |
| 12. | Expenditure on Energy Sector | 0.386 | 0.554 | 0.470 |
| 13. | Ambient concentration of pollutants | 0.148 | 0.226 | 0.187 |
| 14. | Total Primary Energy Production | 1 | 0.736 | 0.868 |
| 15. | Net Energy Import Dependence | 0.519 | 0.585 | 0.552 |
| 16. | Ratio of Daily Disposable Income to Prices of Fuels | 0.180 | 0.404 | 0.292 |
| 17. | Daily Disposable Income | 0.658 | 0.702 | 0.680 |
| 18. | Quantities of Air Pollutant Emissions | 0.984 | 0.983 | 0.983 |
| 19. | Urban Population | 0.774 | 0.645 | 0.709 |
| 20. | Quantities of Green House Emissions | 0.919 | 0.910 | 0.914 |
| 21. | Lifetime of Proven Reserves | 0.994 | 0.924 | 0.959 |
| 22. | Proven Uranium Reserves | 0.422 | 0.411 | 0.416 |
| 23. | Net Nuclear Power Generation | 0.650 | 0.650 | 0.650 |
| 24. | Intensity of Use of Forest Resources | 0.550 | 0.486 | 0.518 |
| 25. | Rate of Deforestation | 0.786 | 0.656 | 0.721 |
| 26. | Proven fossil fuel reserves | 0.954 | 0.944 | 0.949 |
| 27. | CO2 Emissions | 0.472 | 0.388 | 0.430 |

Countries are grouped into five clusters by using *K*-mean clustering. Initially, the countries are clustered on the basis of sustainability indicators for every year from 1980 to 2008. The country is represented as follows:

$X\ (SI_1, SI_2, SI_3, \ldots, SI_{27})$ where $X$ is the name of country and $SI$ represents sustainability indicator.

Forty countries are represented by 40 data points where each data point is represented in the given format. These countries are grouped into five clusters for each sustainability indicator thus producing $27 \times 5$ clusters. The association of a country to a particular cluster is observed to be reassigned across different clusters for different indicators.

The resulting clusters do not have any label. What this means is that the countries have been grouped into five classes but we cannot identify to which group a particular country might belong. If we have labels on these clusters, then it would mean that we identified the type of energy development of a country. A unique method of labeling clusters is presented in Section 4.1 according

**Table 4**
Classification of world countries in five clusters (w.r.t. sustainability indicators).

| Conforming | Futuristic | Critical | Unsustainable | Sustainable |
|---|---|---|---|---|
| Canada (7) | Nigeria (5) | Azerbaijan (6) | Trinidad & Tobago (8) | Saudi Arabia (8) |
| Kuwait (6) | Brazil (8) | | Russia (7) | Venezuela (8) |
| UAE (5) | Algeria (6) | | Kazakhstan (7) | Oman (7) |
| USA (8) | Mexico (8) | | Iran (6) | China (7) |
| Norway (5) | Angola (8) | | | Qatar (8) |
| United Kingdom (6) | Sudan (8) | | | Iraq (7) |
| Australia (8) | Ecuador (7) | | | Libya (7) |
| Brunei (6) | Yemen (7) | | | Syria (7) |
| Denmark (7) | Argentina (5) | | | Congo (7) |
| | Gabon (6) | | | India (6) |
| | Colombia (8) | | | Malaysia (7) |
| | | | | Egypt (7) |
| | | | | Indonesia (6) |
| | | | | Chad (7) |
| | | | | Pakistan (8) |

to which we should initially find the correlation value of each sustainability indicator with production and consumption of hydrocarbons. The correlation analysis is performed and the results are presented in Table 3.

The average correlation value of sustainability indicators with production and consumption is considered to be a weight value for that indicator. For example, the weight value for End use Energy Prices is 0.182. This value shows the cumulative effect of each sustainability indicator on sustainable development hence it is selected as the weight value. To label the clusters, these weight values are multiplied with the actual value of the sustainability indicator. As a result, 27 new sustainability indicator values for each country are produced. These values are averaged to define one value for each country (i.e., overall weight). The clusters are then sorted according to the values of overall weight. The clusters, in order of precedence, are labeled: (1) futuristic, (2) conforming, (3) sustainable, (4) unsustainable; and (5) critical.

Cluster 1 ↔ conforming
Cluster 2 ↔ futuristic
Cluster 3 ↔ critical
Cluster 4 ↔ unsustainable
Cluster 5 ↔ sustainable

On the basis of membership function defined in Eq. (1), five clusters from 27 × 5 clusters are selected. These clusters are shown in Table 4 where each country is sufficed by its membership value to the cluster in parentheses.

The membership value of the UAE (5), for example, means that the UAE is put into the conforming cluster for five years out of 30, which causes the UAE to ultimately fall into cluster 1.

In Fig. 2, the clusters are mapped on a Cartesian system in which classes of energy development are at the *x*-axis and the period of sustainability is plotted at the *y*-axis. It is observed that the patterns reflected some mismatches from the actual condition of energy development in a given country. For example, Iran is clustered in the unsustainable class and Malaysia is in sustainable class. The mechanism for a calculation of error is proposed as an extension to this work.

The clusters do not reflect the criticality level of the individual indicators. An ID3 decision tree is plotted to evaluate the energy

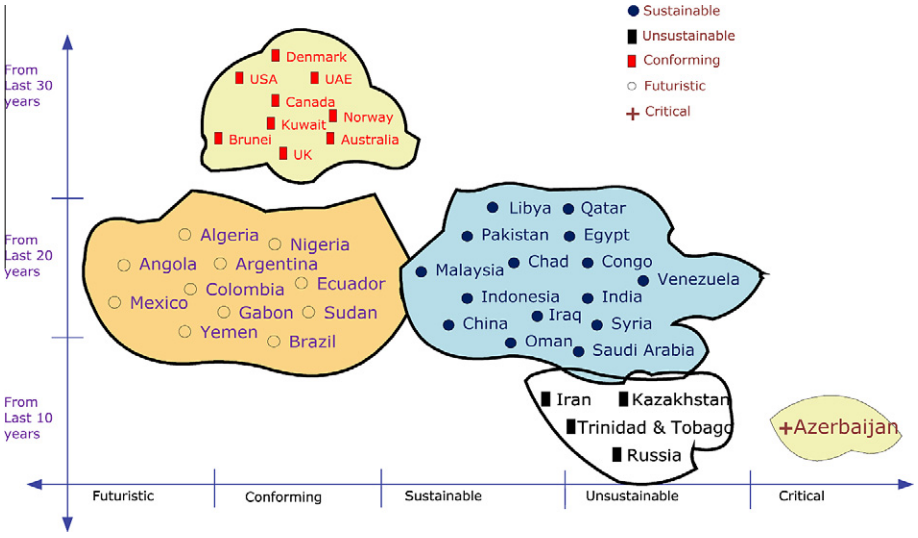


**Fig. 2.** Labeled clusters of the countries of the world w.r.t. energy development.

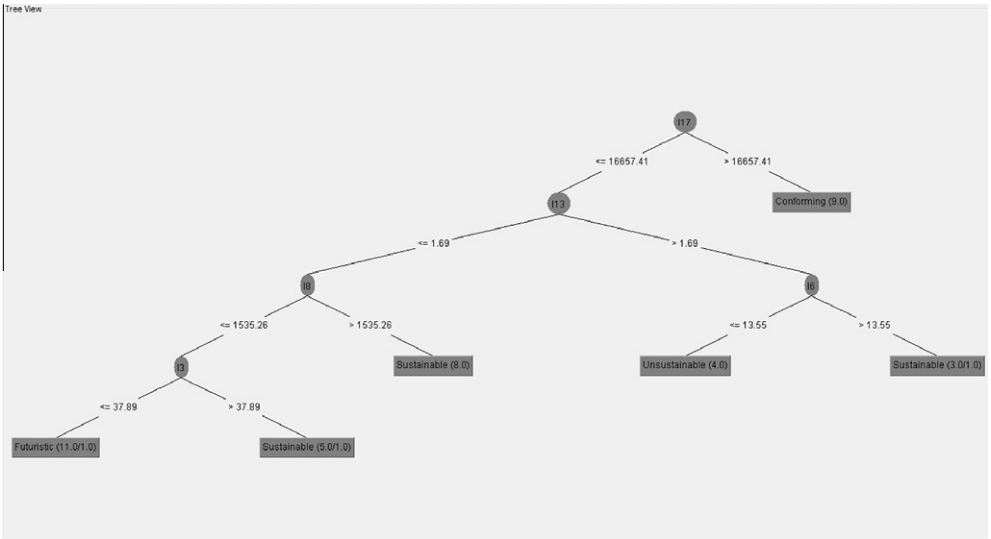


**Fig. 3.** J48 decision tree drawn (WEKA 3.6.2).

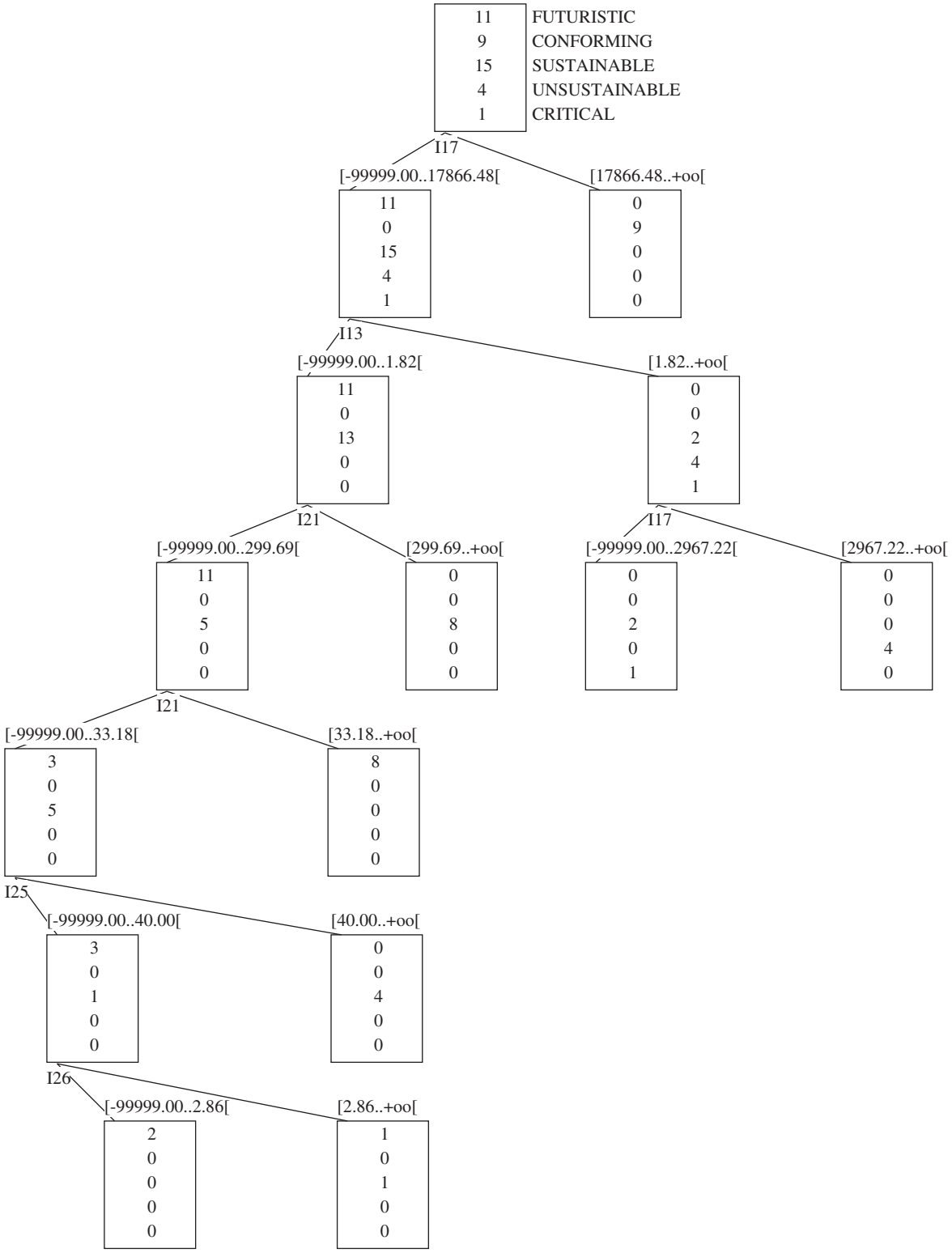


**Fig. 4.** ID3 decision tree for energy development classification (Spinia Software).

development of a country. The decision tree also enables the decision maker to identify the factors for improvement in energy development plan. The decision maker can now address the question, "What indicators should be ameliorated to improve overall energy development in the country?"

In order to draw the decision tree, the data is stored again in the database with pertaining class values which are extracted in the above step. After allocating class values, supervised classification can now be used. An ID3 decision tree is drawn on the above data by using Weka (Fig. 3) and Spinia software (Fig. 4).

The tree shown in above figure is drawn by using WEKA 3.6.2. The attribute with the maximum value of information gain (i.e., I17) is selected at the root. If the value of I17 becomes less than 16657.41, I13 (the attribute with second highest value of information gain), it will be tested; otherwise, the energy development is considered to be conforming. The whole tree is traversed

```
=== Run information ===
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
I17 <= 16657.41
|
|  I13 <= 1.69
|  |
|  |  I8 <= 1535.26
|  |  |
|  |  |  I3 <= 37.89: Futuristic (11.0/1.0)
|  |  |
|  |  |  I3 > 37.89: Sustainable (5.0/1.0)
|  |  |
|  |  .
|  |  I8 > 1535.26: Sustainable (8.0)
|  |
|  I13 > 1.69
|
|     I6 <= 13.55: Unsustainable (4.0)
|
|     I6 > 13.55: Sustainable (3.0/1.0)
I17 > 16657.41: Conforming (9.0)

Number of leaves:                       6

Size of the tree:                       11

Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly classified instances          26          65%
Incorrectly classified instances        14          35%
Kappa statistic                         0.5009
Mean absolute error                     0.151
Root mean squared error                 0.3601
Relative absolute error                 51.25%
Root relative squared error             94.0192%
Total Number of Instances               40
=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|---|
| 1. | 0.545 | 0.138 | 0.6 | 0.545 | 0.571 | 0.763 | Futuristic |
| 2. | 1 | 0.032 | 0.9 | 1 | 0.947 | 0.984 | Conforming |
| 3. | 0.667 | 0.32 | 0.556 | 0.667 | 0.606 | 0.683 | Sustainable |
| 4. | 0.25 | 0.028 | 0.5 | 0.25 | 0.333 | 0.611 | Unsustainable |
| 5. | 0 | 0 | 0 | 0 | 0 | 0.462 | Critical |
| **W.avg** | **0.65** | **0.168** | **0.626** | **0.65** | **0.631** | **0.76** | |

=== Confusion Matrix ===

| a | b | c | d | e | ← classified as |
|---|---|---|---|---|---|
| 6 | 0 | 5 | 0 | 0 | a= Futuristic |
| 0 | 9 | 0 | 0 | 0 | b= Conforming |
| 4 | 1 | 10 | 0 | 0 | c= Sustainable |
| 0 | 0 | 3 | 1 | 0 | d= Unsustainable |
| 0 | 0 | 0 | 1 | 0 | e= Critical |

in a similar manner. The run information taken from WEKA is given below.The decision tree drawn in Spinia depicts the effects of variations in individual sustainability indicators. I1, I2, . . . , I27 in the figure is representing 27 sustainability indicator values whose descriptions are detailed in Table 1. The root node shows that a total of 11 of 40 countries fit into the futuristic class, 9 in conforming, 15 in sustainable, 4 in unsustainable, and 1 in the critical class. The indicator I17 has the largest information content; hence it is selected for evaluation at the root node. The values against I17 indicators are split into two partitions based on their entropy value. In the preceding node, the class values are assigned according to the group range selected for that particular node. For example, there are 9 values of I17 which are greater than 17866.48 and these belong to the conforming class and none of the values under 17866.48 belongs to the conforming class. The decision maker is now able to see the optimal value for the I17 indicator. A similar procedure is adopted to draw the whole hierarchy.

Spinia also provides the function to compute decision tree rules extracted from the database. A few of them are displayed in Fig. 5.

The rule shown in the above figure gives the range of values for particular types of energy development. For example, if the value of I13 is less than 1.82, the value of I17 is less than 17866.48, and I21 is less than 299.69, etc., then the EDEV (energy development) is considered to be Conforming.

## 6. Conclusion and future work

Sustainability management in various sectors including the energy sector tends to be one of the widely studied areas after an economic recession or crisis. Unfortunately, data mining is rarely used to extract the patterns necessary for sustainable development. In this study, a method for assessing energy development of a country by exploiting data mining techniques and sustainability indicators

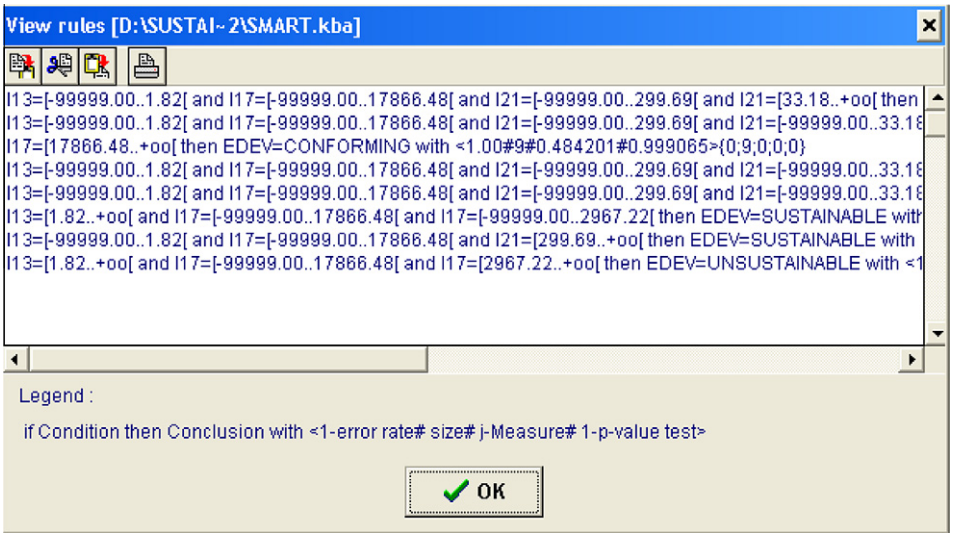


**Fig. 5.** Rules extracted from decision trees.

is presented. Both supervised and unsupervised classification techniques are used on a sustainability indicators database to find the current standings of countries in energy development and to extract generic rules for sustaining or improving energy development. Clustering is used for unsupervised classification and a unique method for labeling the clusters is proposed here. The proposed method of labeling can be generalized to be used on any dataset. Most of the countries listed here appear to be appropriately classified and show likeness with the energy testimony. Despite this, there are still some countries which seem to be inappropriately classified but analyzing the specific contexts of those countries leads us to some interesting clues about those countries' energy development practices over time. The procedure is adequate for producing accurate decision rules, need-based classes and an energy assessment grid. The extracted patterns are clear, specific and decision oriented.

The following extensions of the work are possible in future.

1. The proposed method of labeling clusters is generalized but its design for this application seems to be specific. The procedure can be formalized and generalized to be used in any type of dataset.
2. The indicators considered in this work are specifically designed for energy assessment. There may be some other social, political and economic factors which affect energy development in a country. The work can be extended to include those sustainability indicators for assessment of energy development.
3. The time series analysis of sustainability indicators can draw a temporal sketch of the energy development of a country. An energy development plan can be devised on the basis of these temporal scenarios.

## Acknowledgment

We are indebted to Ms. Eva Woyzbun from Ryerson University, Canada for her valuable error/omission rectification of this manuscript.

## References

[1] A.Z. Al-Garni, S.M. Zubair, J.S. Nizami, A regression model for electric energy consumption forecasting in Eastern Saudi Arabia, Energy 19 (2005) 1043–1049.
[2] K.S. Al-Sultan, M.M. Khan, Computational experience on four algorithms for the hard clustering problem, Pattern Recognition Letters 17 (1996) 295–308.
[3] S. Amreshi, C. Conati, Combining unsupervised and supervised classification to build user models for exploratory learning environments, Journal of Educational Data Mining 1 (1) (2009) 1–54.
[4] G.H. Bruntland, Our common future. World Commission on Environment and Development, University Press Oxford, 1987.
[5] Y.L. Chen, T. Wang, B.S. Wang, Z.J. Li, A survey of fuzzy decision tree classifier, Springer Fuzzy Information and Engineering 1 (2) (2009) 149–159.
[6] S.H. Constantinos, A.M. Paris, An application of supervised and unsupervised learning approaches to telecommunications fraud detection, Knowledge-Based Systems 21 (7) (2008) 721–726.
[7] C. Didem, K. Gulgun, Scenario analysis using bayesian networks: a case study in energy sector, Knowledge-Based Systems 23 (3) (2010) 267–276.
[8] G. Gan, C. Ma, J. Wu, Data clustering: theory, algorithms, and applications. SIAM-ASA Series on Statistics and Applied Probability, Alexandria, VA, 2007.
[9] J. Gao, D.B. Hitchcock, James-Stein shrinkage to improve *K*-means cluster analysis, Computational Statistics and Data Analysis 54 (2010) 2113–2127.
[10] Graphet Inc., Energy data mining and analysis toolset, 2000. <http://www.graphet.com/index.php?id=24>.
[11] J. Brain Gray, F. Guangzhe, Classification tree analysis using TARGET, Computational Statistics and Data Analysis 52 (2008) 1362–1372.
[12] A.D. Gregorio, S.M. Lacus, Clustering of discretely observed diffusion processes, Computational Statistics and Data Analysis 54 (2010) 598–606.
[13] M. Hall, A decision tree-based attribute weighing filter for Naïve Bayes, Knowledge-Based Systems 20 (2) (2007) 120–126.
[14] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, New York, 2001. pp. 284–291.
[15] J.A. Hartigan, M.A. Wong, A *k*-means clustering algorithm, Journal of the Royal Statistical Society Series C – Applied Statistics 28 (1979) 100–108.
[16] K.Y. Huang, A hybrid particle swarm optimization approach for clustering and classification of datasets, Knowledge-Based Systems 24 (3) (2011) 420–426.
[17] International Atomic Energy Agency (IAEA), United Nations Department of Economic and Social Affairs (UNDESA), International Energy Agency (IEA), Eurostat, European Environment Agency (EEA), Energy indicators for sustainable development: methodologies and guidelines, Vienna, 2005.
[18] M. Janicke, H. Jorgens, K. Jorgensen, R. Nordbeck, Governance for sustainable development in Germany: institutions and policy making, Enquette Commission Report, 2001.
[19] W. Jin-Mao, W. Shu-Qin, W. Ming-Yang, Y. Jun-Ping, L. Da-You, Rough set based approach for inducing decision trees, Knowledge-Based Systems 20 (8) (2007) 695–702.
[20] T. Kanungo, D.M. Mount, N. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, An efficient *K*-means clustering algorithm: analysis and implementation, IEEE Transactions and Pattern Analysis and Machine Intelligence 24 (7) (2002) 881–892.
[21] W. Li-Min, L. Xiao-Lin, Y. Sen-Miao, Combining decision tree and Naïve Bayes for classification, Knowledge-Based Systems 19 (7) (2006) 511–515.
[22] J.B. Mac Queen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability Berkeley, vol. 1, University of California Press, 1967, pp. 281–297.
[23] N.M. Maricar, G.C. Kim, N. Jamal, Data mining application in industrial energy audit for lightning, in: Proceedings of the European Power and Energy Systems, 2005.

[24] M.M. Mostafa, R. Natarjaan, A neuro-computational intelligence analysis of ecological footprint of nations, Computational Statistics and Data Analysis 53 (2009) 3516–3531.
[25] G. Rodriguez-Ortiz, V. Fernandez-Espinosa, G. Ramos-Niembro, M. Mejia-Lavalle, Using data mining technique to classify energy demand, International Journal of Power and Energy Systems 19 (2) (1994) 168–172.
[26] Xu. Rui, C. Donald, Wunsch, Recent advances in cluster analysis, International Journal of Intelligent Computing and Cybernetics 1 (4) (2008) 484–508.
[27] S.R. Safavin, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Transactions on Systems, Man and Cybernetics 21 (3) (1991) 660–674.
[28] S.Z. Selim, M.A. Ismail, K-means-type algorithm: generalized convergence theorem and characterization of local optimality, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 (1984) 81–87.
[29] M. Shahbaz, N. Rahman, Data mining for engineering sector in Pakistan: issues and implications, in: Proceedings of the World Congress on Engineering and Computer Science, 2008, pp. 787–792.
[30] M. Shaheen, M. Shahbaz, Z. Rehman, A. Guergachi, Mining sustainability indicators to predict optimal hydrocarbon exploration rate, in: Proceedings of the Tenth IASTED International Conference on Artificial Intelligence and Applications Austria, 2010, pp. 394–400.
[31] G.K.F. Tso, K.K.W. Yau, A study of domestic energy usage pattern in Hong Kong, Energy 28 (2003) 1671–1682.
[32] US Energy Information Administration (EIA), Annual energy review 2009, 2010, pp. 227–271.
[33] United Nations (UN), Report of the World Summit on Sustainable Development, A/CONF.199(20), United Nations, New York, 2002.
[34] United Nations Department of Economic and Social Affairs (UNDESA), Indicators of Sustainable Development: Guidelines and Methodologies, second ed., New York, 2001.
[35] I. Vera, L. Langlois, H. Rogner, Energy Indicators for Sustainable Development. International Atomic Energy Agency IAEA, Austria, 2005, pp. 6–20.
[36] N. Yang, T. Li, J. Song, Construction of decision trees based entropy and rough sets under tolerance relation, in: Advances in Intelligent System Research ISKE-2007 Proceedings, 2007.
[37] Z. Zhi-Hua, C. Zhao-Qian, Hybrid decision tree, Knowledge-Based Systems 15 (8) (2002) 515–528.