Using grey incidence to analyze the energy audit reports and rough set for rule extraction

Tooraj Karimi
Faculty of Farabi, University of Tehran, Tehran, Iran, and
Jeffrey Yi-Lin Forrest
Department of Mathematics,
Slippery Rock University, Rock University, Pennsylvania, USA

Abstract

Purpose – The purpose of this paper is to analyze the energy audit reports in order to define the most favorable factors affecting energy consumption of buildings. Since energy audit of buildings includes assessment of occupants comfort level in addition to the technical data of buildings so some rules are extracted to model the employees thermal comfort level in organization.

Design/methodology/approach – Some tools of RST and GIA are used in this research to analyze the energy consumption of official buildings. "Average energy consumption of building per year" is selected as a system characteristic in GIA and as a decision attribute in RST to show the behavior of buildings energy consumption. Ten technical sequences of buildings are chosen as relevant factors of behavior and conditional attributes in GIA and RST. In order to model the employees thermal comfort level in organization by RST, ten technical attributes of buildings are selected as condition attributes and thermal comfort level of employees is selected as decision attribute. Due to the different algorithms of data complement, discretization, reduction, and rule generation, four rule models are constructed. Cross-validation is used for evaluation of the model results and the best model is chosen with 62 rules and 99.8 percent of accuracy.

Findings – According to the results of GIA and RST, "Uncontrolled area of the building" has been diagnosed as the most important factor between other relevant factors/attributes and it has the greatest effect on energy consumption of building. Four rule models have been extracted from deferent decision tables in order to describe the thermal comfort level of employees in organization. The maximum number of rules relates to the conditional combination/GA model with 1263 rules and average accuracy of 99.7 percent and the minimum number relates to the conditional combination/Janson model with 62 rules and average accuracy of 99.8 percent.

Research limitations/implications – The total observations for rule extraction is 81 and the results can be improved by further samples.

Originality/value – It shows that "Uncontrolled area of the building" is the most important factor/attribute to define the consumption of buildings and thermal comfort level of employees in organization.

Keywords Rough set, Energy audit, Grey incidence, Iran Oil Ministry

Paper type Research paper

1. Introduction

Since the oil shocks of the1970s, there have been numerous studies of energy consumption behaviors from a wide range of disciplinary perspectives. These perspectives include microeconomics; technology adoption models; social and environmental psychology; and sociological theories (Marechal, 2008). Lutzenhiser, who has examined energy behavior since the early 1990s, states: "We are trying to change a very complex system, with lots of moving parts and it is not easily reduced to simple explanations

This research supported by Iran National Science Foundation (INSF: 92019066).

Kybernetes
Vol. 45 No. 7, 2016
pp. 1024-1035

0368-492X
DOI 10.1108/K-02-2016-0022

or simple policy approaches" (Lutzenhiser, 2008). Recently, "energy cultures framework" was introduced by Stephenson. This conceptual framework aims to assist in understanding the factors that influence energy consumption behavior. The energy cultures framework suggests that consumer energy behavior can be understood at its most fundamental level by examining the interactions between cognitive norms (e.g. beliefs, understandings), material culture (e.g. technologies, building form) and energy practices (e.g. activities, processes) (Stephenson *et al.*, 2010).

Although energy consumption per cubic meters in the buildings of Iran is much more than international standards, but thermal comfort level of occupants in buildings is not often satisfactory. Energy audit report of a building includes assessment of thermal comfort level of occupants in addition to the technical data of the building (Karimi and Forrest, 2014). In this study, we intend to generate a rule model based on the relationship between thermal comfort level of employees in organizations and the technical data of the organization buildings. Facing the complex uncertainty of human behavior, it is impractical to try to accurately characterize with a complete mathematical model; even though it is feasible, it is also very difficult to solve and analyze. To help people intelligently analyze uncertainty data, there has been a new generation of soft computing tools, such as rough set theory, grey set theory, and fuzzy set theory (Liu *et al.*, 2012). These soft computing technologies and the hybrid of their advantages are aimed to exploit inaccuracy, uncertainty, approximate reasoning, and partial correctness in the process of human behavior, so as to obtain processable, powerful, and low-cost solutions that are very similar to human decision-making. In the grey system theory, each sequences represents as a factor, a model, a program, an act, and so on. To understand the borders of grey system and make analysis of the primary and secondary factors, the recognition mode, the optimization program, disposal behavior, and so on, it is necessary to make modeling analysis on the relationship between sequences, and it is called the grey incidence analysis (Jian *et al.*, 2011). The rough set method is a series of logical reasoning procedures for analyzing an information system, a table composed of objects with values of conditional and decision attributes. Concepts such as indiscernibility relations, lower and upper approximations, and reducts are used to extract classifying rules (Liu and Qiao, 2014). In this research, at first we use reduct in rough set theory and GIA in grey set theory to find the most favorable technical factors affecting energy consumption of buildings and after that a decision table is organized. This table is composed of ten technical characteristics of buildings as condition attribute and one characteristic about the thermal comfort level of employees as decision attribute. We extract the rules from the reduced decision table and different rule models are generated. Finally, the validity of the models is tested.

Yu *et al.* (2011) have reported a new methodology for examining the influences of occupant behavior on building energy consumption based on a basic data mining technique. Grey relational grades were used as weighted coefficients of different attributes in their method. The results obtained could help prioritize efforts at modification of occupant behavior in order to reduce building energy consumption, and help improve modeling of occupant behavior in numerical simulation (Yu *et al.*, 2011).

Our research is delimited to buildings of the Oil Ministry in Tehran which their reports of energy auditing were perfect and available. Energy auditing of these buildings has been implemented by consultants since 2010 and all reports have the same structure.

Since the main object of this paper is analyzing the energy audit reports by using GIA and RST, so in Section 2 we briefly review the algorithm of the grey incidence analysis and essential concepts and definitions of RST. In Section 3, the most important technical factors affecting the energy consumption of building as a system characteristics are

determined with the two different approaches. After that, for modeling the thermal comfort level of employees, the process of using RST for rule generation, validation and implementation is explained in Section 4. Finally, conclusion is presented in Section 5.

2. Grey incidence analysis and rough set theory

In the grey relational analysis (GRA), the data that contain the same features are regarded as a series. The relationship between two series is determined by the difference of the two series, and the difference measure refers to a value of background for generating a grey relational grade. Compared with the usual distance measurement, the GRA combines with the concept of wholeness and can express the relationship of the two objects more exactly and objectively (Dafang and Qing-chun, 2009).

The basic idea of grey correlation analysis is to see whether the relation is close or not by the similarity among the geometrical shapes of sequence curves. The closer the sequence curves are, the greater the correlation is and vice versa (Jian *et al.*, 2011).

When we make system analysis and study the relationship between the behavior of the system characteristics and related factors, the main concern is usually the order of the size of degree, the system characteristic behavior sequence, and all relevant factors. The degree of grey incidence is used to demonstrate the relationship between two sequences, denoted $\gamma(x_0(k), x_i(k))$.

Grey incidence analysis includes such materials as grey incidence axioms, degree of grey incidence, generalized degree of grey incidence (absolute degree, relative degree, synthetic degree), the degrees of grey incidence based on either similar visual angles or nearness visual angles, grey incidence order, superiority analysis and others (Liu *et al.*, 2012).

The computing steps of grey degree of sequences are as follows.

- Step 1: find the average image of each sequence. Let:

$$X_i' = \frac{X_i}{\overline{X}} = \left(x_i'(1), x_i'(2), \ ..., \ x_i'(n)\right), \quad i = 0, 1, 2, \ ..., \ m \tag{1}$$

- Step 2: find difference sequences. Denote:

$$\Delta_i(k) = \left|x_0'(k) - x_i'(k)\right|, \tag{2}$$

$$\Delta_i = (\Delta_i(1), \Delta_i(2), \ ..., \ \Delta_i(n)), \quad i = 0, 1, 2, \ ..., \ m \tag{3}$$

- Step 3: find the maximum and minimum difference and write:

$$M = \max_i \max_k \Delta_i(k), \ m = \min_i \min_k \Delta_i(k) \tag{4}$$

- Step 4: find incidence coefficients:

$$\gamma_{0i}(k) = \frac{m + \xi M}{\Delta_i(k) + \xi M}, \xi \in (0, 1), \quad i = 0, 1, 2, \ ..., \ m, k = 1, 2, \ ..., \ n \tag{5}$$

- Step 5: compute the degree of incidences:

$$\gamma(k) = \tfrac{1}{n} \sum_{k=1}^{n} \gamma_{0i}(k) \tag{6}$$

Assume that X_0 is a sequence of a system's characteristic behaviors, that X_i and X_j are sequences of two relevant factors' behaviors, and that γ is the degree of grey incidence.

If $\gamma_{0i} \geqslant \gamma_{0j}$, then the factor X_i is said to be more favorable than the factor X_j, denoted as $X_i \succ X_j$. The relation "$\succ$" is called the grey incidence order. So in order to analyze a system, we can determine the most important system characteristic and also among under examined factors, the ones which have more effects on the future development of the systems can be recognized (Liu and Lin, 2006).

The data collected from the real world may contain all kinds of noise, and there are many uncertainties and incomplete information to be dealt with (Zou *et al.*, 2011). As an extension of conventional set theory, rough set theory is proposed to support approximations. Working as a math tool on fuzzy and uncertain knowledge, rough set theory plays an important role in machine learning, decision support, data mining, and process control. It has an irreplaceable advantage on handling the uncertainty problem (Liu and Qiao, 2014).

The core assumption of "rough set" is that the knowledge is embodied in the ability of classification (Tseng and Huang, 2007). In this theory, a data set is represented as a table, where each row represents a case, an event, or simply an object. Every column represents an attribute (a variable, an observation, a property, etc.) that can be measured for each object; the attribute may be also supplied by a human expert or user. This table is called an information system. More formally, it is a pair $S = (U, A)$, where U is a non-empty finite set of objects called the universe and A is a non-empty finite set of attributes such that a: $U \rightarrow V_a$ for every $a \in A$. The set V_a is called the value set of a.

In many applications there is an outcome of classification that is known. This is a posteriori knowledge expressed by one distinguished attribute called decision attribute. Information systems of this kind are called decision systems. A decision system is any information system of the form $S = (U, C \cup \{d\})$ where $d \notin C$ is the decision attribute. The decision attribute may take several values though binary outcomes are rather frequent. Let $S = (U, A)$ be an information system, then with any $B \subseteq A$ an equivalence relation $IND_S(B)$ is associated:

$$IND_S(B) = \left\{(x, x') \in U^2 \colon \forall a \in B, a(x) = a(x')\right\} \tag{7}$$

$IND_S(B)$ is called the B-indiscernibility relation. If $(x, x') \in IND_A(B)$ then objects x and x' are indiscernible from each other by attributes from B.

Given a knowledge representation system $S = (U, A)$, $P \subseteq A$, $X \subseteq U$, $x \in U$, the lower approximation and upper approximation of set X regarding $IND_S(B)$, respectively, are:

$$\underline{B}X = \underline{apr}_B(X) = \left\{x \colon IND_S(B) \subseteq X\right\} \tag{8}$$

$$\overline{B}X = \overline{apr}_B(X) = \left\{x \colon IND_S(B) \cap X \neq \phi\right\} \tag{9}$$

According to indiscernible relationship, it is convenient to define some important characteristics of the information system, among which the most important characteristic is the dependency of attributes. If the number of equivalent types (element sets) derived from attribute set A is the same as that derived from $(A - a_i)$, then attribute a_i is regarded as redundant; otherwise, the attribute a_i is indispensable in A.

Suppose $X = \{X_1, X_2, ..., X_n\}$ is a partition of universe U, where $X_i (i = 1, 2, ..., n)$ is one class of X, and $P \subseteq A$, then the quality of approximation of X is:

$$k = \frac{\left|POS_P(D)\right|}{|U|}, POS_P(D) = \bigcup_{X \in U/D} \underline{P}(X) \tag{10}$$

If the quality of approximation $k=1$, then the knowledge X is completely dependent on P. If $0<\mathrm{k}<1$, then we can say that the knowledge X is partly dependent on P, which reveals that only partial attributes in the P are available, or the data set has some initial defects.

Core and attribute reduct are two fundamental concepts while reduct is the smallest independent attribute subset that has the same data division with the overall attribute sets, and it is the essential part of the information system, which can be used to distinguish all the objects that can be discernible in the original information system. The core is the common part of all reducts.

Let S be an information system with n objects. The discernibility matrix of S is a symmetric $n \times n$ matrix with entries c_{ij} as given below:

$$C_{ij} = \{a \in A | a(x_i) \neq a(x_j)\}, \quad i,j = 1,2, \ldots, n \tag{11}$$

Each entry thus consists of the set of attributes upon which objects x_i, x_j differ. A discernibility function f_S for an information system S is a Boolean function of m Boolean variables $a_1^*, \ldots, a_m^*$ defined as follows (Jian *et al.*, 2011):

$$f_S(a_1^*, \ldots, a_m^*) = \wedge \{ \vee c_{ij}^* | 1 \leqslant j \leqslant i \leqslant n, c_{ij} \neq \phi \}, \quad c_{ij}^* = \{a^* | a \in c_{ij}\} \tag{12}$$

The other dimension in reduction is to keep only those attributes that preserve the indiscernibility relation and, consequently, set approximation. The rejected attributes are redundant since their removal cannot worsen the classification. There is usually several such subsets of attributes and those which are minimal are called reducts. In practical application, it is not necessary to calculate all the reducts, but just some of them. Generally speaking, the reduct that contains the minimal attribute number is the most satisfying reduct (Zhong *et al.*, 2001). The common part of all reducts is called core so:

$$CORE(P) = \bigcap_{R_i \in RED(P)} R_i \; i = (1,2, \ldots, n) \tag{13}$$

In any learning system, rule generation is a very important task. The set of all condition elements in the universe is called the condition classes of S, denoted by $X_i (i = 1, 2, \ldots, n)$. The set of all decision elements in the universe is called the decision classes of S, denoted by $Y_j (j = 1, 2, \ldots, n)$ if $X_i \cap Y_j = \phi$ then:

$$r\text{:}\ Des_C(X_i) \Rightarrow Des_D(Y_j) \tag{14}$$

It is called the decision rules of (C, D), denoted as $\{r_{ij}\}$ for $\vee i, j, if\ X_i \subseteq Y_j$, then rule r_{ij} is decisive in S, otherwise it is indecisive. The syntax of a rule is as follows:

$$IF, f(x,q_1) = r_{q1} \wedge f(x,q_2) = r_{q2} \wedge \ldots \wedge f(x,q_p) = r_{qp} THEN\ x \in Y_{j1} \vee Y_{j2} \vee \ldots \vee Y_{jk}$$

$$\{q_1, q_2, \ldots, q_p\} \subseteq C; (r_{q1}, r_{q2}, \ldots, r_{qp}) \in V_{q1} \times V_{q2} \times \ldots V_{qp}.$$

The "if" part of decision rule is called the condition part, while the "then" part is called the decision part. If the consequences are univocal, that is, one object not only matches the condition part but also the decision part, then $k=1$, the rule is exact, or the rule is approximate. The number of objects that support the decision rules is called the supporting number (Jiang *et al.*, 2007).

3. Selecting the most important factor

The main purpose of this section is identifying the most important factors affect energy consumption of office buildings. For this purpose we have implemented two deferent methods. First, we calculate GI of each factor and after that, we extract the reduct sets and core attributes of data and finally the results are compared.

In order to analyze a system by GI, after choosing the quantity to reflect the characteristics of the system of concern, one needs to determine all factors that influence the behavior of the system. If a quantitative analysis is considered, one needs to process the chosen characteristic quantity and the effective factors using sequence operators so that the available data are converted to their relevant non-dimensional values of roughly equal magnitudes.

In this study, among the 95 audit reports relating to buildings of Iran's Oil Ministry in Tehran, 87 office buildings in the same climate zone are selected. Due to lack of measurements of some presented factors for these buildings, only 81 buildings that have complete data have been analyzed by GIA. System characteristic (X_0) and system factors ($X_1, X_2, \ldots, X_{10}$) of this study are shown in Table I. In this table, the meaning of "uncontrolled area of the building" is parts of buildings which have no heating and cooling systems such as parking and stairs. Moreover, load coefficient of a surface is the total heat transfer from the surface when the temperature difference between inside and outside is 1°C. For more detailed information about the definition and the method of calculating of each variable (see Wayne and Turnur, 2005).

According to the GIA, the system characteristic and ten-related factors are a sequence of 81 observations as follows:

$$X_j(k), j = 0, 1, 2, \ldots, 10, k = 1, 2, \ldots, 81 \tag{16}$$

After collecting the data for each variable, Grey System Modeling Software 6.0 is used to calculate γ_{0j} (Liu and Lin, 2010). The grade of grey incidences is as follows:

$$\gamma_{0j} = (0.86, 0.89, 0.82, 0.92, 0.90, 0.83, 0.89, 0.86, 0.87, 0.85) \tag{17}$$

$$X_4 \geqslant X_5 \geqslant X_2, X_7 \geqslant X_9 \geqslant X_1, X_8 \geqslant X_{10} \geqslant X_6 \geqslant X_3 \tag{18}$$

Name of variables RST	GST	Description	Unit	Value domain	Attributes type
D	X_0	Average energy consumption of building per year	Mega Jul (MJ)	[242.8,13013.2]	Float
C_1	X_1	Number of floors including basement	Number	[1,21]	Integer
C_2	X_2	Building area	Cubic meters (m^3)	[3250,238000]	Integer
C_3	X_3	Approximate age of building	Year	[2,54]	Integer
C_4	X_4	Uncontrolled area of the building	Cubic meters (m^3)	[141.7,53482.8]	Float
C_5	X_5	Number of employees	Number	[4,1300]	Integer
C_6	X_6	Walls load coefficient	Watts per Kelvin (W/K)	[0.47,4.32]	Float
C_7	X_7	Windows load coefficient	Watts per Kelvin (W/K)	[3,7.8]	Float
C_8	X_8	Doors load coefficient	Watts per Kelvin (W/K)	[2,5.97]	Float
C_9	X_9	Roof load coefficient	Watts per Kelvin (W/K)	[0.5,2]	Float
C_{10}	X_{10}	Floor load coefficient	Watts per Kelvin (W/K)	[0.2,2.42]	Float

Table I. Research variables

Equation (18) shows that the fourth factor, "uncontrolled area of the building," is a more favorable factor and has more effect on the system behavior. In contrary, the third factor, "approximate age of building," has the least effect on the behavior of the system. In short, among the factors affecting the behavior of system which is energy consumption, "uncontrolled area of the building" is the most effective factor.

In order to use rough set, ten condition attributes and one decision attribute are selected as Table I. Each row of decision table relates to one office building and each column relates to an attribute. The domain of V_a for $a \in A$ is shown in Table I.

Rough set theory analysis requires miss value data to be completed. There are a variety of techniques to complete the miss value of observations. Two most common ones are "conditional mean/mode fill" and "conditional combinatorial completion" which are used in this research. After using the latter, 81 observations convert to 1908 data. For more information about these methods (see Komorowski *et al.*, 2002).

Since rough set theory does not include digital data discretization, we must make use of a proper way to convert the digital data into discrete interval before the application of the rough set method, accordingly, we need to transform the continuous number into a series of natural numbers (Ruiz *et al.*, 2008). In the discretization of a decision table $S = (U, A)$, where $V_a = (v_a, w_a)$ is an interval of real, we search for a partition P_a of V_a for any $a \in A$. Any partition of V_a is defined by a sequence of the so-called cuts $v_1 \prec v_2 \prec \ldots \prec v_k$ from V_a.

In this research, all data of the energy auditing are numeral; therefore, we must make use of a proper way to convert the digital data into discrete interval before the application of the rough set methods. Common data discrete methods are such as "expert discrete method" and "entropy method." The latter has been used in this research by ROSSETTA software. The results of discretization are shown in Table II. To read more about this and other algorithms of discretizing (see Ludl and Widmer, 2000; Clarke and Barton, 2000).

Reduct is defined as the minimum data content including input and output features necessary to represent an object (Smith and Bull, 2003). More common algorithms for reduct extraction are "genetic algorithm" and "Johnson's algorithm" which are used in this research. The minimum reduct set of each algorithm for each completed data table are shown in Table III. For more information about these methods (see Liang, 2009; Starzyk *et al.*, 2000).

Since the Johnson's algorithm is not an optimization programming for reducing attributes and gives one quick reduct set; so the members of reduct which is obtained by this algorithm is equal or less than genetic algorithm.

Table II.
Number of intervals of attributes

	Number of intervals			Number of intervals	
Attributes	Conditional combinatorial completion	Conditional mean/mode fill	Attributes	Conditional combinatorial completion	Conditional mean/mode fill
D	18	12	C6	12	8
C1	7	2	C7	9	3
C2	16	16	C8	7	3
C3	9	3	C9	7	5
C4	14	9	C10	9	10
C5	17	9			

According to the members of reduct sets, C4 is core of all data sets, so the most important attribute to define system is "uncontrolled area of the building." In other word, this attribute can be used alone to discern energy consumption of buildings.

By comparing the results of RST and GIA, it can be seen that "uncontrolled area of the building" is selected as the most important attribute/factor to define and control energy consumption of buildings.

4. Employees thermal comfort-level modeling

The purpose of this section is analyzing the employees thermal comfort level in organization using the RS modeling. For this purpose, ten technical aspects of the building (Table I) are considered as condition attributes and thermal comfort level of employees in organization which is calculated and presented in energy audit reports are considered as decision attribute. The type of decision attribute is string and its value set is:

$$V_a = \{Comfortable,\ Normal,\ Uncomfortable\}$$

The conceptual framework of RST to elicit decision rules consists of the following steps (Gaojun and Yan, 2006):

(1) problem definition;

(2) data preprocessing including completion of miss value and discretization of numerical attributes;

(3) performing a standard rough set based analysis of data;

(4) search of a core and reducts of attributes permitting data reduction;

(5) inducing sets of decision rules from rough approximations of decision classes;

(6) evaluating sets of rules in classification experiments; and

(7) using sets of decision rules as classifiers.

Conditional mean/mode and conditional combinatorial completion have been used to complete the data, entropy method to discrete the data and Janson and Genetic algorithm to extract the reduct sets. The result has been shown in Table IV.

The rules have been extracted for each complete decision table using reducts which is generated in two ways (Janson and GA). Table V shows some rules which is related to Janson algorithm reduct generated from mod/mean completed decision table.

Accuracy and coverage are indices of the approximation quality. Accuracy measures the probability that an object belonging to the approximation belongs also to the approximated set. Coverage measures the percent of objects in a set that are included in its approximation. When the approximation accuracy is equal to 1 and the

Complete method	Discretizing method	Reduct algorithm	Reduct set (minimum set)	Core set
Conditional mean/mode fill	Entropy	Johnson	{C4, C5, C9}	{C4}
Conditional mean/mode fill	Entropy	Genetic	{C4, C5}	{C4}
Conditional combinatorial completion	Entropy	Johnson	{C4}	{C4}
Conditional combinatorial completion	Entropy	Genetic	{C4}	{C4}

Table III. Reduct sets of different tables

	Reduct Set		Reduct Set	Reduct Set
Janson algorithm and conditional mean/mode	{C4, C10}	Genetic algorithm and conditional mean/mode fill	{C4}	{C3, C8, C11}
			{C7, C10}	{C5, C6, C7}
Janson algorithm and conditional combinatorial	{C4}		{C2, C6}	{C1, C9, C10}
			{C2, C10}	{C2, C5, C9}
			{C5, C10}	{C2, C7, C11}
Genetic algorithm and conditional combinatorial completion	{C4, C10}		{C3, C10}	{C3, C6, C8}
	{C2, C10, C11}		{C1, C8, C10}	{C2, C3, C11}
	{C2, C9, C10}		{C6, C7, C8}	{C3, C5, C6}
	{C5, C10, C11}		{C5, C6, C9}	{C1, C5, C8, C9}
	{C2, C6, C10}		{C5, C7, C11}	{C1, C6, C7, C11}
	{C5, C6, C10}		{C5, C7, C9}	{C2, C7, C8, C9}
	{C2, C6, C9, C11}		{C1, C6, C10}	{C1, C6, C9, C11}
	{C2, C3, C9, C11}		{C3, C8, C9}	{C3, C7, C9, C11}
	{C4, C6, C9, C11}		{C2, C3, C9}	{C6, C7, C9, C11}

Table IV.
Reduct sets of deferent decision tables

The first five rules	Accuracy	Coverage
C4([362.59, 370.96)) AND C10([1.57, 1.71)) → D(comfortable)	1	0.5
C4([4434.50, *)) AND C10([*, 1.10)) → D(uncomfortable)	1	0.4
C4([*, 362.59)) AND C10([*, 1.10)) → D (normal)	1	0.23
C4([*, 362.59)) AND C10([1.28, 1.57)) → D (normal)	1	0.23
C4([517.25, 750.15)) AND C10([1.28, 1.57)) → D (normal)	1	0.23

Table V.
Some rules of mean/mode completion and Janson

coverage is maximized the approximation may be considered as lower and when the approximation coverage is equal to 1 and the accuracy is maximized the approximation may be considered as upper (Chan *et al.*, 2008).

Table VI shows the number of rules that have been obtained from each of the four models. In order to predict the decision attribute of new objects using the rules, it is necessary to compare the validation of four rule models.

To obtain good estimates of the true classification performance it is important to use a test set that is representative for the observations that the classifier is likely to encounter in the future. In practice, it is common to divide the available labeled observations randomly into a training set and a test set. The training set is used to induce a classifier and the test set is used for estimating the classification performance (Liu and Qiao, 2014).

Cross-validation is a technique that provides several test sets while fully utilizing all the available data for training and testing. The data are divided into k approximately equally sized subsets. Each fold is then consecutively used as a test set while the remaining k-1 folds are used as a training set. Thus each object appears in the test set once and in the training set k-1 times. The prediction performance is recorded for each test set and variance is computed. The ROSETTA system includes an algorithm for doing cross-validation. In this study, k is considered to be 5. In each iteration, a confusion matrix is presented. The confusion matrix shows the overall accuracy, as well as the sensitivity and accuracy for each class. Table VII shows an example of 20 confusion matrix of cross-validation. This table is related to the first iteration of Janson algorithm and conditional combinatorial decision table. At the end of the cross-validation of each model, the max, min and average accuracy and its standard deviation is given. The results of cross-validation of four rule models have been shown in Table VI.

Table VI shows that the highest validation relates to the rules extracted from conditional combinatorial decision table and Janson algorithm. In other words, although the number of rules generated in this model is less than other models but cross-validation results show that the accuracy of this rule model is more than the other models. Therefore based on this model, we can predict the employees thermal comfort level in organization according to the "uncontrolled area of the building" and the accuracy of prediction will be more than 99 percent.

5. Conclusion

Since GST and RST techniques can overcome weaknesses of statistical methods, some techniques of them were used to analyze the results of energy audit of Iranian Oil Ministry buildings in this research. At first step, "average energy consumption of building per year" was selected as a system characteristic in GIA and as a decision attribute in RST to show the behavior of buildings' energy consumption. In GIA, ten factors included in the audit reports were considered as the factors affecting this behavior and in RST as conditional attributes. Based on calculated GI vector,

Table VI. Cross-validation results of four models

Data completion method	Reduct algorithm	Number of rules	Min accuracy	Max accuracy	Average accuracy
Conditional mean/mode	Janson	70	0.375	0.625	0.464
Conditional mean/mode	Genetic	282	0.666	0.875	0.810
Conditional combination	Janson	62	0.994	1	0.998
Conditional combination	Genetic	1,263	0.992	1	0.997

Table VII. First confusion matrix of Janson algorithm and Conditional Combinatorial decision table

	Predicted				
	Comfortable	Normal	Uncomfortable	Undefined	Accuracy
Actual					
Comfortable	0	0	0	2	0
Normal	0	12	0	0	100
Uncomfortable	0	0	2	0	100
Undefined	0	0	0	0	–
Accuracy	–	100	100	0	85.7

"uncontrolled area of the building" was the most favorable factor. According to the core of reduct sets, "uncontrolled area of the building" was the most important attribute in information system of energy auditing of buildings. Comparison of the results of both methods showed that "uncontrolled area of the building" is the most influential factor. So it has more effects on the system behavior and it is necessary to control it.

In the next step, RST was used to model the employees thermal comfort level in organization. Ten technical attributes of buildings were selected as condition attributes and thermal comfort level of employees was selected as decision attribute. The miss value of data were completed by conditional mean/mode and conditional combination method. Two completed decision tables were discretized by entropy algorithm. Janson and Genetic algorithms were used to generate the reduct sets. Finally, four rule models were extracted from deferent reducts and decision tables. Maximum number of the rules related to the conditional combination/GA model with 1263 rules and average accuracy of 99.7 percent and minimum number of the rules related to the conditional combination/Janson model with 62 rules and average accuracy of 99.8 percent. Despite the lowest number of rules in latter model, the validity and accuracy of this model was higher than others. Since only one reduct set was generated by Johnson algorithm and the reduct set in conditional combination/Janson model had only one member so it can be said that with the knowledge of "uncontrolled area of the building" as condition attribute, thermal comfort level of employees in organization can be predicted with the accuracy of 99.8 percent. So it can be said that "uncontrolled area of the building" is the most important characteristic of decision system. If this attribute accurately be calculated in the energy audit of office buildings there is no need to assess the level of employees comfort in the form of time-consuming and costly projects and it can be accurately predicted using rule model of this research. This analysis helps the improvement of future audits, and assists in making energy conservation policies.

References

Chan, C.C., Jerzy, W. and Ziarko, W.P. (2008), "Rough sets and current trends in computing", 6th International Conference, *RSCTC, Akron, OH, October 23-25*.

Clarke, E.J. and Barton, B.A. (2000), "Entropy and MDL discretization of continuous variables for bayesian belief networks", *International Journal of Intelligent Systems*, Vol. 15 No. 1, pp. 61-92.

Dafang, L. and Qing-chun, W. (2009), "Grey relational analysis for local government public service evaluation", *Proceedings of 2009 IEEE International Conference on Grey Systems and Intelligent Services, Nanjing, November 10-12*.

Gaojun, L. and Yan, Z. (2006), "Credit assessment of contractors: a rough set method", *Tsinghua Science and Technology*, Vol. 11 No. 3, pp. 357-3621.

Jian, L., Liu, S. and Lin, Y. (2011), *Hybrid Rough Sets and Applications in Uncertain Decision-Making*, Taylor and Francis Group, LLC, New York, NY.

Jiang, W., Zhong, X., Qi, J. and Zhu, C. (2007), "Grey rough sets hybrid scheme for intelligent fault diagnosis", *IEEE International Conference on Grey Systems and Intelligent Services, Nanjing, November 18-20*.

Karimi, T. and Forrest, J. (2014), "Analyzing the results of buildings energy audit by using grey incidence analysis", *Grey Systems: Theory and Application*, Vol. 4 No. 3, pp. 386-399.

Komorowski, K., Øhrn, A. and Skowron, A. (2002), "The ROSETTA rough set software system", *Handbook of Data Mining and Knowledge Discovery*, ISBN 0-19-511831-6, Oxford University Press, pp. 1554-1559.

Liang, W.Y. (2009), "Apply rough set theory into the information extraction: the application of the clustering", *Fifth International Joint Conference on INC, IMS and IDC, Seoul.*

Liu, J. and Qiao, J. (2014), "A grey rough set model for evaluation and selection of software cost estimation methods", *Grey Systems: Theory and Application*, Vol. 4 No. 1, pp. 3-12.

Liu, S. and Lin, Y. (2006), *Grey Information Theory and Practical Applications*, Springer-Verlag Limited, London.

Liu, S. and Lin, Y. (2010), *Grey Systems Theory and Applications,* Springer-Verlag, Berlin and Heidelberg.

Liu, S., Forrest, J. and Yang, Y. (2012), "A brief introduction to grey systems theory", *Grey Systems:Theory and Application*, Vol. 2 No. 2, pp. 89-104.

Ludl, M.-C. and Widmer, G. (2000), "Relative unsupervised discretization for association rule mining" *Proceeding of the Fourth European Conf. Principles of Data Mining and Knowledge Discovery* (PKDD), pp. 148-158.

Lutzenhiser, L. (2008), "Setting the stage: why behavior is important", *Overview of Address Given to the Behavior, Energy and Climate Change Conference*, delivered to California Senate June, Sacramento CA, November 7-9.

Marechal, K. (2008), *An Evolutionary Perspective on the Economics of Energy Consumption: the Crucial Role of Habits*, Sovalay Business School, Brussels.

Ruiz, F.J., Angulo, C. and Agell, N. (2008), "IDD: supervised interval distance-based method for discretization", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20 No. 9, pp. 1230-1238.

Smith, M.G. and Bull, L. (2003), "Feature construction and selection using genetic programming and a genetic algorithm", *Proceedings of the 6th European Conference on Genetic Programming*, pp. 229-237.

Starzyk, J.A., Nelson, D.E. and Sturtz, K. (2000), "A mathematical foundation for improved reduct generation in information systems", *Journal of Knowledge and Information Systems*, Vol. 2 No. 2, pp. 131-146.

Stephenson, J., Barton, B., Carrington, G., Gnoth, D., Lawson, R. and Thorsnes, P. (2010), "Energy cultures: a framework for understanding energy behaviours", *Energy Policy*, Vol. 38 No. 10, pp. 6120-6129, doi: 10.1016/j.enpol.2010.05.069.

Tseng, T.L. and Huang, C.C. (2007), "Rough set-based approach to feature selection in customer Relationship management", *Omega*, Vol. 35 No. 4, pp. 365-383.

Wayne, C. and Turnur, B. (2005), *Energy Management Handbook*, 5th ed., Fairmont press.

Yu, Z., Fung, C.M., Haghighat, F., Yoshino, H. and Morofsky, E. (2011), "A systematic procedure to study the influence of occupant behavior on building energy consumption", *Energy and Buildings*, Vol. 43, pp. 1409-1417.

Zhong, N., Dong, J. and Ohsuga, S. (2001), "Using rough sets with heuristics for feature selection", *Journal of Intelligent Information Systems*, Vol. 16 No. 3, pp. 199-214.

Zou, Z., Tseng, T.L., Sohn, H., Song, G. and Gutierrez, R. (2011), "A rough set based approach to distributor selection in supply chain management", *Expert Systems with Applications*, Vol. 38 No. 1, pp. 106-115.

Corresponding author
Tooraj Karimi can be contacted at: tkarimi@ut.ac.ir