Electrical Power and Energy Systems 62 (2014) 727–734

Contents lists available at ScienceDirect

Electrical Power and Energy Systems

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

# Using the rough set theory to detect fraud committed by electricity customers

Josif V. Spirić [a,*], Slobodan S. Stanković [b,1], Miroslav B. Dočić [c,2], Tatjana D. Popović [d,3]

[a] I. Strele 6/27, 16000 Leskovac, Serbia
[b] Business Department, Company for electric power supply "Jugoistok" Niš, Branch "Elektrodistribucija – Leskovac", Stojana Ljubića 16, 16000 Leskovac, Serbia
[c] Company for electric power supply "Jugoistok" Niš, Branch "Elektrodistribucija – Leskovac", Stojana Ljubića 16, 16000 Leskovac, Serbia
[d] IT Department, Company for electric power supply "Jugoistok" Niš, Branch "Elektrodistribucija – Leskovac", Stojana Ljubića 16, 16000 Leskovac, Serbia

ARTICLE INFO

*Article history:*
Received 26 January 2014
Received in revised form 30 April 2014
Accepted 10 May 2014
Available online 12 June 2014

*Keywords:*
Fraud detection
Power utility
Total losses
Information system
Feature selection
Unsupervised discretization

ABSTRACT

High values of total percentage losses of electricity in Serbia's distribution network and a very small percentage of success in detecting electricity fraud indicate a need for a sophisticated means of detecting fraud perpetrators. In order for this task to be performed efficiently, the authors of the paper emphasize the need for a suitable and comprehensive use of the billing system or the database of the amounts of invoiced spent electricity in accounting periods and other relevant data regarding registered customers. For that purpose, they suggest using the rough set theory and give a general approach to its use. The point of the paper is forming a criterion for the estimation of accurate (suitable) discretization of original data. The criterion is based on the amount of lost not invoiced electricity due to electricity fraud. Based on consumption characteristics of detected fraud of customers whose measurement points were regularly (monthly) read, a list of the suspicious customers will be formed which will serve as the basis for sending expert teams to specified locations with the task to confirm or dismiss the fraud suspicion.



## Introduction

There are two predominant logics for estimating the size of the suspicious customers set based on the set of data recorded for years in the database formed in power utilities. The first logic is based on a significant change of characteristic parameters regarding a customer's electricity consumption in a certain time period with respect to: (a) their own average parameters until then and (b) with respect to the same, but average parameters of the substation TS MV/LV region or a series of measurement points in the region, which are periodically remotely or directly read. Basic or derived statistical techniques are used for realization [1–3]. A customer identified with one or more characteristic changes in accordance with the appropriate values of those changes, is identified as a suspicious customer.

The other logic is based on customers caught in fraud in an earlier period. In case of these customers, only a part of electricity goes through the meter which is regularly read. The remaining part of electricity is used off the meter. In other words, these customers steal electricity and will hereafter be referred to as thieves. Based on their characteristic consumption data, profiles of such customers are formed. Then, customers with profiles identical to the already registered fraud profiles are searched for. With this type of fraud, the set of customers expected to contain thieves, without excluding those who are not, is called the suspicious customers set. All customers whose profiles match thieves' profiles are declared suspicious customers. The rough set theory will be used for realization in this paper [4,5].

It is possible to use the fuzzy set theory for the same purpose [6,7]. First, it must be defined at least two criteria that express some customer specific consumption characteristics or the ratio of two customer consumption characteristics. The criteria can also be the ratio of the customer consumption characteristics and selected site characteristics where the customers belong. Based on the selected criteria, it is formed membership functions to their fuzzy sets, determined the membership functions of fuzzy sets to assess suspicions and putted a fuzzy rules under the "if-then". After the fuzzy reasoning procedure is done, defuzzification is performed which fuzzy conclusion turn into a real number that represents the suspicion evaluation. Values of suspicion evaluation (usually in%) make up the list of priorities for field testing. In

* Corresponding author. Tel.: +381 64 8367600; fax: +381 16 253196.
*E-mail addresses:* josif.vspiric@gmail.com (J.V. Spirić), slobodan.sstankovic@jugoistok.com (S.S. Stanković), miroslav.docic@jugoistok.com (M.B. Dočić), tatjana.popovic@jugoistok.com (T.D. Popović).
[1] Tel.: +381 64 8367635.
[2] Tel.: +381 64 8367619.
[3] Tel.: +381 64 8367634.

http://dx.doi.org/10.1016/j.ijepes.2014.05.004
0142-0615/

forming membership functions of fuzzy sets should be used the frauds that were already found.

Tariff systems for selling and buying electricity in other countries can generate a variety of consumption characteristic parameters which, in accordance with the selection made by the supervisor, can form different thief profiles. It is emphasized that the difference in tariff systems does not influence the general application of either statistical techniques or the rough and fuzzy set theory.

## The basics of the rough set theory

The rough set theory created by Z. Pawlak (1982) [8] is a mathematical tool for analyzing available data. Knowledge discovery in the database (KDD) is a process of handling valid, large groups of data with several important, interactive and iterative stages for identifying understandable patterns (samples). The rough set theory is used in medicine, economy, technology, pharmacy, the decision theory, software engineering, etc.

Rough set philosophy is based on the assumption that every object of the *universe* can be described using some characteristic information. Objects characterized by the same information are *indiscernible* (similar) in view of the available information about them. The indiscernibility relation formed in this manner is the mathematical basis of the rough set theory.

Let there be a finite set of objects (instances, cases or observations) $U$ which is called the *universe* and let there be a finite set of attributes (qualities, characteristics, variables) $A$. Each attribute $a \in A$ is joined with a set $V_a$ whose values are called the domain of $a$. The pair $S = (U, A)$ denotes an *information system*. The set of attributes $A$ can be divided into the set of condition attributes $C$ and the set of decision attributes $D$, with a condition that they are disjoint or $C \cap D = \emptyset$. Any subset $B \subseteq A$ determines $I_B$, the binary relation from $U$, which is called the *indiscernibility relation* and is marked with:

$$I_B = \{(x, y) \in U \times U : a(x) = a(y),\ \forall a \in B\}, \tag{1}$$

where $a(x)$ and $a(y)$ denote the values of attribute $a$ for objects $x$ and $y$, respectively.

In other words, relation (1) states that two different objects $x$ and $y$ are described by the same attributes and that such relation between them is expressed by the indiscernibility relation.

Taking into consideration that relation $I_B$ is reflexive, symmetrical and transitive, it represents the relation of equivalence. Equivalence relation $I_B$ divides set $B$ into mutually disjoint non-empty subsets (equivalence classes) whose union is set $B$.

The equivalence classes of the indiscernibility relation with respect to $B$ are denoted $[x]_B$ or $[x]_{I_B}$.

The family of all equivalence classes is marked $U/I_B$ or simply $U/B$.

The family of all classes, already marked $U/B$ is called $B$ *elementary notions* or $B$ *granules*.

The point of the rough set theory is forming approximate sets for the given information system $IS = (U, A)$ and let $B \subseteq A$ and $X \subseteq U$.

Using the rough set theory, a set of electricity customers can be determined in the category of households for which there are assumptions regarding electricity fraud.

In accordance with established marks, let: $X$ be the set of customers who steal electricity (hereafter thieves), $U$ the finite set of customers (in our case $N$ customers), $C$ the finite set of condition attributes, $B(x)$ the equivalence class of $x$ with respect to the attribute set $B$ and $IS$ the information system.

$\underline{B}$**- *Lower approximation*** of set $X$ is defined as the union of all elementary sets contained in $X$ or

$$\underline{B}(X) = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\}, \tag{2}$$

and denotes customers which certainly belong to the observed set $X$.

$\overline{B}$**-*Upper approximation*** of set $X$ is defined as the union of those elementary sets which have non-empty intersection with set $X$ or

$$\overline{B}(X) = \bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \emptyset\}, \tag{3}$$

and denotes customers which possibly belong to the observed set.

The difference between the upper and lower approximation is defined as $B$-***boundary region*** of set $X$:

$$BN_B(X) = \overline{B}(X) - \underline{B}(X), \tag{4}$$

and denotes customers which may, but do not necessarily belong to set $X$ based on the available knowledge in $B$.

If the boundary region of set $X$ is an empty set, i.e. $BN_B(X) = \emptyset$, then set $X$ is crisp (exact) with respect to $B$. In case of $BN_B(X) \neq \emptyset$, set $X$ is rough (inexact) with respect to $B$ (see Fig. 1).

We can also define the ***positive region*** of set $X$ in which the elements of the set are certainly members of the set, as

$$posB(X) = \underline{B}(X), \tag{5}$$

and the ***negative region*** of set $X$ in which there are elements that are certainly not members of set $X$ as

$$negB(X) = U - \overline{B}(X). \tag{6}$$

In order to estimate how well approximation to $A$ has been performed, the accuracy of approximation is defined:

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|}, \tag{7}$$

as a relation between the cardinality of lower approximation set $\underline{B}(X)$ and the cardinality of upper approximation set $\overline{B}(X)$.

It is obvious that $0 \leqslant \alpha_B(X) \leqslant 1$. If $\alpha_B(X) = 1$, then $X$ is exact with respect to $B$ and if $\alpha_B(X) < 1$, then $X$ is rough with respect to $B$.

The *quality of the lower approximation* can also be significant, defined as:

$$\underline{\gamma}_B(X) = \frac{|\underline{B}(X)|}{|U|}, \tag{8}$$

with intention to perceive the degree of knowledge totality (completeness) of set $X$ with respect to $B$ as well as *the quality of the upper approximation* defined as:

$$\overline{\gamma}_B(X) = \frac{|\overline{B}(X)|}{|U|}, \tag{9}$$

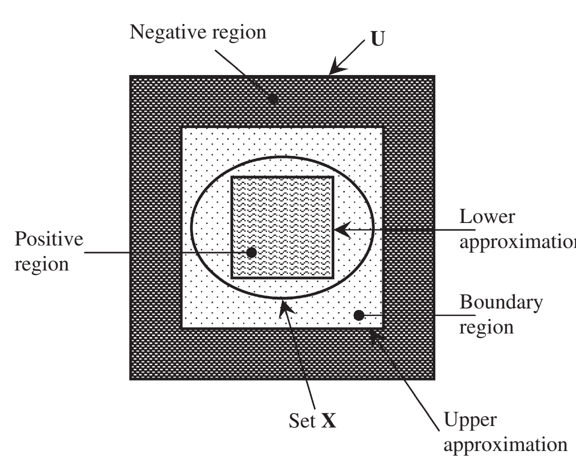


**Fig. 1.** Graphic interpretation of approximate sets.