Journal of the Egyptian Mathematical Society 000 (2017) 1–5

Contents lists available at ScienceDirect

Journal of the Egyptian Mathematical Society

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

Topological approach to retrieve missing values in incomplete information systems

A.S. Salama*, O.G. El-Barbary

Department of Mathematics, Faculty of Science, Tanta University, Tanta, Egypt

ARTICLE INFO

Article history:
Received 1 April 2017
Revised 29 June 2017
Accepted 17 July 2017
Available online xxx

MSC:
54A05
54B05
54D35
03B70
68R01

Keywords:
Rough sets
Rough approximation
Near open sets
Incomplete information systems
Accuracy measure
Rule generation

ABSTRACT

We used some new results related to topology to retrieve the original values of the missing values in the incomplete information systems.

1. Introduction

The theory that Pawlak discovered in the early 1980s [1], that called "rough set theory" is an extension of the classical set theory for the study of incomplete information systems. Many researchers have been made proposals for the generalizations and interpretations of rough sets [2–20]. The Pawlak theory is considered as a new mathematical tool to computer applications using also fuzzy set theory [21,22]. Nowdays, a lot of researchers are interested to generalize Pawlak theory in other fields of applications [23–27].

A lot of attractive generalizations to equivalence relation have been proposed such as tolerance relations [28], similarity relations [29], topological bases and subbases [30–34] and others [35–38]. Many exertions from researchers to use other approaches of the universe of discourse for starting a new generalizations of rough sets by coverings [39]. Others [40–45] tried to combine fuzzy sets with rough sets in a fruitful way by defining rough fuzzy sets and fuzzy rough sets. Furthermore, another group of researchers tried to characterize a measure of roughness of a fuzzy set making use of the concept of rough fuzzy sets [46,47]. Some results of these generalizations are obtained about rough sets and fuzzy sets in [48–55]. Pawlak rough set theory generate a special type of topological space called the clopen space (the space have each open set is closed) and this topology is used in many fields of real life applications. The concept of topological rough set given by Wiweger [33] in 1989 is the first topological generalization of rough sets. In 1983 Abd El-Monsef et al. [56] introduced the concept of β-open sets that can used widely in the topological generalizations of rough sets. In 2006 Hatir and Noiri [57] introduced the concept of $\delta\beta$-open sets that we used in this paper to introduce a solution of the problem of the missing values in information systems.

Rokach in [58] considered the idea of ensemble methodology to build a predictive model by integrating multiple models. In [59] Hamid Parvin and others have been provided a good way to have a near-optimal classifying system for any problem depend on a technique called ensemble learning. This technique became one of the most challenging problems in classifier ensemble that introducing a suitable ensemble of base classifiers. Parvin with others in [60] have been proposed an ensemble based approach for feature selection. They aimed at overcoming the problem of pa-

* Corresponding author.
E-mail addresses: dr_salama75@yahoo.com, asalama@su.edu.sa (A.S. Salama), omniaelbarbary@yahoo.com (O.G. El-Barbary).

http://dx.doi.org/10.1016/j.joems.2017.07.004

rameter sensitivity of feature selection approaches. The effectiveness of non-weighing-based sampling technique, comparing the efficacy of sampling with and without replacement, in conjunction with several consensus algorithms have been invested in [61]. In addition, experimental results have shown improved stability and accuracy for clustering structures obtained via bootstrapping, subsampling, and boosting techniques. In [62] Ahmad and others have been presented a clustering algorithm based on k-mean paradigm that works well for data with mixed numeric and categorical features. They have been proposed new cost function and distance measure based on co-occurrence of values.

In this paper, we introduced a topological method to retrieve missing values in the incomplete informations. Short illustration on missing values are given in Section 2. In Section 3 we give the properties of rough set theory and their approximations. Section 4 gives the basic concepts of incomplete information systems. In Section 5 we introduce the basics of near topological approximations. The main aim of Section 6 is that how we make use of near topological approximations to retrieve missing values in incomplete information system. Section 7 shows the phsolophy that retrieve missing values. Conclusions of the work are given in Section 8.

2. Basics of missing values

In incomplete information systems, absent attribute values befall when no information value is stored for some characteristics. Lost data are a mutual occurrence and can have the noteworthy result of the assumptions that drained from the information.

For treatment, absent attribute values, we may be using one of the following:

1. Erasing cases with missing attribute values.
2. Replacing a lost attribute value of the greatest common value of that attribute.
3. Transmission all likely values for the missing attribute value.
4. Relieving a lost attribute value by the mean for numerical attributes.
5. Transfer to a lost attribute value the corresponding value taken from the closest case.
6. Relieving a lost attribute value by a new value computed from a new data set.
7. Seeing the original attribute as a decision.

We need some assumptions to find out the missing values some of them are:

- All result attribute values are well defined.
- The elements of the universe that have lost values have the same accidental in our classification.
- The field of every disorder attribute is completely definite.
- The dataset is reliable.

Noises of incomplete decision table represented objects U and columns represent attributes At. Qualities are dividing to main parts condition C and choice attributes. On this table, a function that maps the direct product of U and C into the set of all attribute values is called an information function and denoted by f.

Classical rough set theory used an equivalence relation as a tool to deal with the uncertain data given in an incomplete information system. This relation for any subset of the conditional attributes is defined as follows:

$$R = \{(x, y) \in U \times U : \forall a \in X \ f(x, a) = f(y, a)\}, X \subseteq C.$$

Rough elementary sets are generated using the partition U/R.

Two rough approximations are defined using the relation R for any subset of objects $A \subseteq U$ as follows: The first one is the lower approximation $\underline{R}(A) = \cup\{G \in R/U : G \subseteq A\}$. The second one is the upper approximation $\overline{R}(A) = \cup\{G \in R/U : G \cap A \neq \varphi\}$.

The boundary region of A, $A^b = \overline{R}(A) - \underline{R}(A)$is the set of fundamentals for which there is a hesitation about their transmission to A. Currently, by means of new topological strategies the boundary region rotten into many districts that well known and smaller.

Reductions of information system are all subsets X of attributes such that if it is a minimal subset $X' \subset X \subseteq C$, then it must keep the quality of classification unmoved.

We make use of the decision rules generated from the incomplete information system to construct a knowledge base for this system. A decision rule is defined by the equation:

$$\underset{a \in X}{\wedge}(f(a(x), v)) \rightarrow (D(x), k) \ \text{ where } v \in V_a \text{ and } k \in V_D$$

The straightforward condition is usually resembled to as an attribute value pair $(a(x), v)$ and decision value pair $(D(x), k)$. Each decision rule r has two main parts: r_s and r_t the condition and decision parts congruently.

3. Topological spaces and their approximations

Usually a topological space is defined to be a set U and a family τ of subsets of U satisfying the following conditions:

(1) $\varphi, U \in \tau$
(2) Random unions of open subsets of U is being a member in τ.
(3) Limited intersections of open subsets of U is being a member in τ.

$cl_\tau(A) = \{F \subseteq U : A \subseteq F, F^c \in \tau\}$ is the closure of a subset $A \subset U$. $\text{int}_\tau(A) = \bigcup\{G \subseteq X : G \subseteq A, G \in \tau\}$ is the interior of a subset $A \subseteq U$. $b(A) = cl_\tau(A) - int_\tau(A)$ is the boundary of a subset $A \subseteq U$.

Now we will make use of the relation R to define a topology τ_R on the universe of dicourse. τ_R has the neighborhood $R_x = \{y : (x, y) \in R, x, y \in U\}$. We define two topological notion as follows:

$cl_{\tau_R}(A) = \{x \in U : R_x \cap A \neq \varphi\}$ and $\text{int}_{\tau_R}(A) = \{x \in U : R_x \subseteq A\}$, called the closure approximation and interior approximation of $A \subseteq U$ respectively. $P_{\tau_R}(A) = \text{int}_{\tau_R}(A)$ the positive region, $N_{\tau_R}(A) = U - cl_{\tau_R}(A)$ the negative region and $b_{\tau_R}(A) = cl_{\tau_R}(A) - \text{int}_{\tau_R}(A)$ the boundary region of A, respectively.

The degree of wholeness can also be characterized by the accurateness measure, in which $|R|$ represents the cardinality of a subset $A \subseteq U$ as follows:

$$\alpha_{\tau_R}(A) = \frac{|\text{int}_{\tau_R}(A)|}{|cl_{\tau_R}(A)|}, \ A \neq \varphi.$$

4. Generalizations of approximations

The δ-closure of a subset $A \subseteq U$ of the topology τ, is defined by $cl^\delta_\tau(A) = \{x \in X : A \cap \text{int}_\tau(cl_\tau(G)) \neq \varphi, G \in \tau, x \in G\}$. A is δ-closed set if $A = cl^\delta_\tau(A)$. The complement of a δ-closed set is δ-open, such that $\text{int}^\delta_\tau(A) = U \backslash cl^\delta_\tau(U \backslash A)$.

A subset A is Ω-open if $A \subseteq cl_\tau(int_\tau(cl^\delta_\tau(A)))$. The family of all Ω-open sets of U is denoted by $\Omega O(U)$. The complement of Ω-open set is called Ω-closed set. The family of Ω-closed sets is denoted by $\Omega C(U)$.

The pair (U, R_Ω) is called a Ω-approximation space where R_Ω is a general relation used to get a subbase for a topology τ_{R_Ω} on U which generates the class $\Omega O(U)$. Ω-interior and Ω-closure of any subset $A \subseteq U$ is defined as:

$$\text{int}_{R_\Omega}(A) = \cup\{G \in \Omega O(U) : G \subseteq A\}, \ cl_{R_\Omega}(A) = \cap\{F \in \Omega C(U) : F \supseteq A\}.$$

We call $P_{\tau_{R_\Omega}}(A) = \text{int}_{\tau_{R_\Omega}}(A)$ the confident region of A, $N_{\tau_{R_\Omega}}(A) = U - cl_{\tau_{R_\Omega}}(A)$ denote the undesirable region of A and the set $b_{\tau_{R_\Omega}}(A) = cl_{\tau_{R_\Omega}}(A) - \text{int}_{\tau_{R_\Omega}}(A)$ denote the frontier region of A.

Please cite this article as: A.S. Salama, O.G. El-Barbary, Topological approach to retrieve missing values in incomplete information systems, Journal of the Egyptian Mathematical Society (2017), http://dx.doi.org/10.1016/j.joems.2017.07.004