Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/100306

Correlation Coefficient Via Statistical and Rough Set Concepts

Manal E. Ali[1] *and Tamer Medhat*[2,*]

[1]Department of Physics and Engineering Mathematics, Faculty of Engineering, Kafrelsheikh University, 33516, Kafrelsheikh, Egypt
[2]Department of Electrical Engineering, Faculty of Engineering, Kafrelsheikh University, 33516, Kafrelsheikh, Egypt

Received: 11 Jun. 2021, Revised: 2 Jul. 2021, Accepted: 24 Jul. 2021
Published online: 1 Sep. 2021

Abstract: In this article, new definitions for a correlation coefficient using the rough sets technique are introduced. These definitions are more general than statistic definitions, which give the capability to handle all information system tables (qualitative, quantity, ordered, and unordered data). By using these definitions, dealing with all unordered data tables can be done, which can't deal with it by using a statistical definition, these definitions will be discussed in detail through some examples.

Keywords: Statistical correlation coefficient, Information systems, and Rough sets.

1 Introduction

In the past few years, rough sets theory [1,2] has attracted great attention in numerous domains, such as data mining, pattern identification, and machine learning. The classical rough sets (CRS) model [1,2,3,4] was effectively utilized as a mathematical method to handle uncertainty data, reduce the attributes (feature selection), extract the rules, and justify uncertainty. The classification analyses of data tables are discussed [5,6]. The classification research is discussed in the data tables. Measurements or human experts can be used for obtaining the data. The rough set analysis's main goal is to synthesize the approximation of the acquired data concepts and reduce the display of them to a minimum [7,8,9,10,11]. Several rough set modes were established in a rough set community in the recent decades including VPRS and GRS [1,3,4,5,11,12,13,14]. Some of them were implemented in industrial data mining projects, such as patient symptoms diagnostics, telemarket churner predictions, stock market predictions, and client attrition analyses for financial banks to resolve difficult business difficulties [8,9,10,15,16,17,18]. These rough-set models are aimed at extending Pawlak's original model [1,2] and trying to meet its limits, like statistical distribution or noisy data management. The semi-correlation factor [19] is used in information system tables for the reduction of attributes.

The approach we utilized depends on the positive region between the characteristics of the condition and decision or "attributes" which classify the objects with regard to all condition attributes as equivalence classes." In the positive region, we have a new definition of the correlation factor that applies to all data (quantity, qualitative, ordered, and unordered data). In this article, we start with section 2 as an introduction to the fundamental ideas of rough sets theory and coefficient of correlation. Section 3 presents the new definitions of correlation coefficient depending on the concepts of the rough sets, and we end this work in the concluding section.

2 Basic Concepts

2.1 Information System

A data set is portrayed as a table in which each row is represented by a case, patient, incident, or simply an object. Each column has an attribute for each item (a variable, property, inspection, and so on) and a human specialist or user may also give the attribute. This table is called information system (IS) table [2] that can be defined as IS=(U,A,ρ,V), where U is a non-empty finite set of objects called a universe and A is a non-empty

* Corresponding author e-mail: tmedhatm@eng.kfs.edu.eg

finite set of attributes.

$$IS = (U, A, \rho, V) \tag{1}$$

Any subset X ⊆U is termed a U category, each attribute $a \in$A is seen as mapping U elements into Va set, where Va set is called a value set of an attribute a.

$$\rho : UxA \rightarrow Va \tag{2}$$

2.2 Mathematical Model in Information System

There are mathematical models that may be utilized to extract knowledge by utilizing the provided data from information systems.

2.2.1 Indiscernibility Relation

For any Q ⊆ A there is an equivalence relation IND(Q) [2]:

$$IND(Q) = \{(x, y) \in U^2 : \forall a \in Q, a(x) = a(y)\} \tag{3}$$

The partition of U, generated by IND(Q) is denoted U/IND(Q) (or U/Q) and can be obtained as follows:

$$U/IND(Q) = \otimes\{a \in Q : U/IND(\{a\})\}, \tag{4}$$

where

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \emptyset\} \tag{5}$$

If (x,y)∈IND(Q), then x and y are indiscernible by attributes from Q. The equivalence classes of the Q-indiscernibility relation are denoted $[x]_Q$.

2.2.2 Lower and upper approximations

Let Y⊆U. Y can be approximated using only the information contained within Q by constructing the Q-lower and Q-upper approximations of Y [2]:

$$\underline{QY} = \{y : [y]_Q \subseteq Y\} \tag{6}$$

$$\overline{QY} = \{y : [y]_Q \cap Y \neq \emptyset\} \tag{7}$$

2.2.3 Positive, negative, and boundary regions

If C and D are equivalence relations over U, then the positive, negative, and boundary regions are defined as follows [2]:

$$POS(C, D) = \bigcup_{X \in U} \underline{Q}X \tag{8}$$

$$NEG(C, D) = U - \bigcup_{X \in U} \overline{Q}X \tag{9}$$

$$BND(C, D) = U - \{POS(C, D) \cup NEG(C, D)\} \tag{10}$$

The positive region includes all objects U categorized by information in characteristics C in U/D classes. BND(C,D) is a collection of items that could be categorized in this way, but not surely. NEG(C,D) is the negative region of items that cannot be categorized into U/D classes.

2.3 Statistical correlation coefficient

In essence, the Pearson product-moment coefficient [20] is only a particular example where the data are translated into ranking before the coefficient is calculated. However, for the sake of *r* calculations, a simpler technique is typically employed. The raw scores are transformed to rankings, and the differences d between ranks are calculated on each of the two variables.

Correlation Coefficient of Spearman Rank [20]. Spearman proposes non-parametric (distribution-free) rank statistics as a measure for the force of the linkages between two variables in 1904. The coefficient of the Spearman grade correlation might be used to produce an R-estimate and is a measure of the monotone relationship used to make the Pearson correlation coefficient unwanted or deceptive by the distribution of data.

The correlation coefficient of Spearman rank is defined by

$$r = 1 - \frac{6\sum {d_i}^2}{n(n^2 - 1)} \tag{11}$$

where d is the difference between each rank of matching x and y values, and n is the number of pairs of values.

The above formulation is a close approximation to the precise correlation coefficient.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{12}$$

derived from the original data.The Spearman rank correlation coefficient is significantly easier to compute since it utilizes ranks.

Example 1.

Let U={1,2,3,4,5} be persons, A={a1,a2,a3} be subjects which are: a1=Mathematics, a2=English, a3=Chemistry, and the values E=excellent, VG=Very Good, G=Good, P=Pass are grade of students as shown in Table 1.

Table 1: Grade of students in some subjects

U/A	a1	a2	a3
1	VG	VG	E
2	VG	VG	E
3	G	G	VG
4	G	G	VG
5	P	P	G

Table 2: Rank between a_1 and a_2

U/A	a1	a2	R(a1)	R(a2)	d	d^2
1	VG	VG	4.5	4.5	0	0
2	VG	VG	4.5	4.5	0	0
3	G	G	2.5	2.5	0	0
4	G	G	2.5	2.5	0	0
5	P	P	1	1	0	0
Sum	-	-	-	-	0	0

Table 3: Rank between a_2 and a_3

U/A	a2	a3	R(a1)	R(a2)	d	d^2
1	VG	E	4.5	4.5	0	0
2	VG	E	4.5	4.5	0	0
3	G	VG	2.5	2.5	0	0
4	G	VG	2.5	2.5	0	0
5	P	G	1	1	0	0
Sum	-	-	-	-	0	0

By using the statistical definition of the correlation coefficient (Spearman coefficient), see the following rank tables Table 2 and Table 3:
From Table 2 and Table 3, we find that:
The statistical correlation coefficient between a1 and a2 is 1 or 100%, see Table 2:

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{0}{5(25-1)} = 1$$

The statistical correlation coefficient between a2 and a3 is 1 or 100%, see Table 3:

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{0}{5(25-1)} = 1$$

also, the statistical correlation coefficient between a1 and a3 is 1 or 100%.

3 Rough Set Correlation Coefficient

Using rough sets theory approaches, this section provides a new definition for the correlation coefficient. This concept applies to quantitative, qualitative, ordered, and unordered data. Statistical correlation definitions cannot work with unordered data.

Definition 1. Let U be a universe and A be a set of condition attributes, $a_i, a_j \in A$, then $r_{i,j}$is a rough set correlation coefficient between attributes a_i *and* a_j, and it may be computed as follows:

$$r_{i,j} = \frac{||Pos(\{a_i\},\{a_j\})||}{||U||}, \tag{13}$$

where

$$POS(\{a_i\},\{a_j\}) = \cup_{Yi \in U/IND(\{a_j\})}\{X \in U/IND(a_i), X \subseteq Y_i\} \tag{14}$$

where $a_i, a_j \in A$, and i,j=1,2,3,....,$\|C\|$

Definition 2. In a decision information system tables, Let U be a universe and A={C,D} be a set of attributes of condition C and attribute of decision D, $c_i \in C$, then r_i is a rough set correlation coefficient between condition attribute c_i and decision attribute D, and it may be computed as follows:

$$r_i = \frac{||Pos(\{c_i\},D)||}{||U||}, \tag{15}$$

where

$$POS(\{c_i\},D) = \cup_{Yi \in U/IND(D)}\{X \in U/IND(c_i), X \subseteq Y_i\} \tag{16}$$

where $c_i \in C$, and i=1,2,3,....,$\|C\|$
See the following examples.

Example 2.
Continue from Table 1 in Example 1, and by using the concepts of rough sets theory, and rough set correlation coefficient, we get the following:

U/IND({a1}) = {{1,2}, {3,4}, {5}}
U/IND({a2}) = {{1,2}, {3,4}, {5}}
U/IND({a3}) = {{1,2}, {3,4}, {5}}
Then, U/IND({a1}) = U/IND({a2}) = U/IND({a3})

$$r_{a1,a2} = \frac{||Pos(\{a1\},\{a2\})||}{||U||} = \frac{||\{1,2,3,4,5\}||}{||\{1,2,3,4,5\}||} = \frac{5}{5} = 1$$
$$also:$$
$$r_{a1,a3} = 1,\ r_{a2,a3} = 1$$

This means that: The correlation coefficients between a1, a2 and a3 are equivalent and equal to 1.

Example 3.

Let U={x_1,x_2,x_3,x_4,x_5,x_6} be the universe, A={C,D} be the attributes, where C={c_1,c_2,c_3} is the condition attributes, and D is the decision attribute. The values of these condition attributes are symbols which can't be ordered as α, β, γ, S, R, H, M, and L as shown in Table 4.
Using rough sets techniques, we get;
U/IND({c_1})={{x1,x3,x6},{x2,x5},{x4}}
U/IND({c_2})={{x1,x2,x4},{x3,x5,x6}}
U/IND({c_3})={{x1,x2,x5},{x3,x6},{x4}}
and

Table 4: Qualitative information system

U/A	c_1	c_2	c_3	D
x_1	α	S	H	H
x_2	β	S	H	H
x_3	α	R	L	L
x_4	γ	S	M	L
x_5	β	R	H	H
x_6	α	R	L	L

U/IND(D)={{x1,x2,x5},{x3,x4,x6}}
Then, the rough sets correlation coefficient between c_1 and D is

$$r_1 = \frac{||Pos(\{c_1\},D)||}{||U||} = \frac{||\{x2,x4,x5||}{||\{x1,x2,x3,x4,x5,x6\}||} = \frac{3}{6} = 0.5$$

The rough sets correlation coefficient between c_2 and D is

$$r_2 = \frac{||Pos(\{c_2\},D)||}{||U||} = \frac{0}{6} = 0$$

The rough sets correlation coefficient between c_3 and D is

$$r_3 = \frac{||Pos(\{c_3\},D)||}{||U||} = \frac{6}{6} = 1$$

4 Conclusion

The statistical correlation coefficients can't deal with all data such as unordered data. So, we can use the definitions of rough set correlation coefficient. These definitions are used for all data (qualitative, quantity, ordered and unordered data).

Disclosure: This study was performed as part of the employment of the authors: University of Kafrelsheikh, Egypt.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] Z. Pawlak, "Rough Sets", International Journal of Information and computer Science, 11(5):341-356, (1982).
[2] Z. Pawlak, "Rough Sets - Theoretical Aspects of Reasoning about data.", Kluwer Academic Publishers, Dordrecht, Boston, London, (1991).
[3] L. Polkowski, A. Skowron, "Rough mereology: A new paradigm for approximate reasoning", J. of Approximate Reasoning, 15(4), 333-365, (1996).
[4] W. Ziarko, "Variable Precision Rough Set Model", Journal of Computer and System Sciences, 46(1), 39-59, (1993).
[5] X. Hu, N. Cercone, J. Han, W. Ziarko, "GRS: A Generalized Rough Sets Model", in Data Mining, Data Mining, Rough Sets and Granular Computing, T.Y. Lin, Y.Y.Yao and L. Zadeh (eds), Physica-Verlag, 447- 460, (2002).
[6] J.Y. Liang, J.H. Wang, Y.H. Qian, "A new measure of uncertainty based on knowledge granulation for rough sets", Inform. Sci. 179(4), 458–470, (2009).
[7] M. Sammany and T. Medhat, "Dimensionality Reduction Using Rough Set Approach for Two Neural Networks-Based Applications[A]" in Rough Sets and Intelligent Systems Paradigms[C], Heidelberg:Springer Berlin, pp. 639-647, (2007).
[8] T.Y. Lin, "From rough sets and neighborhood systems to information granulation and computing in words", Proceedings of European Congress on Intelligent Techniques and Soft Computing, 1602-1607, (1997).
[9] T.Y. Lin, "Granular computing on binary relations I: data mining and neighborhood systems, II: rough set representations and belief functions", In Rough Sets in Knowledge Discovery, Lin T.Y., Polkowski L., Skowron A., (Eds.). Physica-Verlag, Heidelberg,107-140, (1998).
[10] T.Y. Lin, Y.Y. Yao, L.A. Zadeh, (Eds.) " Rough Sets, Granular Computing and Data Mining", Physica-Verlag, Heidelberg, (2002).
[11] T. Medhat, "Topological applications on information analysis by rough sets", Master thesis, Egypt, Tanta University, Faculty of Engineering, (2004).
[12] A. Skowron, J. Stepaniuk, "Tolerance approximation spaces", Fundamenta Informaticae, 27(2-3), 245-253, (1996).
[13] H.M. Chen, T.R. Li, R. Da, J.H. Lin, C.X. Hu, "A rough-set based incremental approach for updating approximations under dynamic maintenance environments", IEEE Trans. Knowl. Data Eng. 25(2), 274–284, (2013).
[14] L.J. Dong, D.G. Chen, N.L. Wang, Z.H. Lu, "Key energy-consumption feature selection of thermal power systems based on robust attribute reduction with rough sets", Inform. Sci., 532, 61–71, (2020).
[15] P.F. Zhang, T.R. Li, G.Q. Wang, C. Luo, H.M. Chen, J.B. Zhang, D.X. Wang, Z. Yu," Multi-source information fusion based on rough set theory: A review", Inf. Fusion 68, 85–117, (2021).
[16] Q.H. Hu, D.R. Yu, W. Pedrycz, D.G. Chen, "Kernelized fuzzy rough sets and their applications", IEEE Trans. Knowl. Data Eng. 23(11), 1649–1667, (2011).
[17] T.P. Hong, T.T. Wang, S.L. Wang, B.C. Chien," Learning a coverage set of maximally general fuzzy rules by rough sets", Expert Syst. Appl. 19(2), 97–103, (2000).
[18] X.Y. Zhang, H. Yao, Z.Y. Lv, D.Q. Miao, "Class-specific information measures and attribute reducts for hierarchy and systematicness", Inform. Sci., 563, 196–225, (2021).
[19] M. E. Ali, A. A. Abo Khadra, A. M. Kozae, "Reduction using semi correlation factor", International Journal of Engineering Science and Technology, Vol. 2(12), 7500-7509, (2010).
[20] R. V. Hogg, A. T. Craig, "Introduction to Mathematical Statistics", 5th ed. New York: Macmillan, (1995).