Applied Mathematics and Computation 320 (2018) 781–794

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Dependency and accuracy measures for directed graphs

G. Chiaselotti*, T. Gentile, F. Infusino, P.A. Oliverio

Department of Mathematics and Informatics, University of Calabria, Via Pietro Bucci, Cubo 30B, 87036 Arcavacata di Rende (CS), Italy

ARTICLE INFO

Keywords:
Digraphs
Rough set theory
Accuracy measure
Dependency measure
Dependency averages

ABSTRACT

In this paper we use finite directed graphs (digraphs) as mathematical models to study two basic notions widely analyzed in granular computing: the attribute dependency and the approximation accuracy. To be more specific, at first we interpret any digraph as a Boolean information table, next we study the approximation accuracy for three fundamentals digraph families: the directed path, the directed cycle and the transitive tournament. We also introduce a new global average for the attribute dependency in any information table and we determine such number for any directed path. For the transitive tournament we provide a lower bound.

1. Introduction

In database theory it is very frequent the need to study finite tables having a very large quantity of data, therefore many researches have been directed towards the purpose of reducing and simplifying the interpretation of these data. With such an aim, Pawlak [45] developed the so called *rough set theory* (abbreviated RST). RST is an elegant and powerful methodology in extracting and minimizing rules from data tables (*information systems* or, also, *information tables* in Pawlak terminology) which has been further developed and generalized in several recent works [35,38,43,44,54,63,64,79].

Next, RST has been considered as a part of the more general emerging methodological paradigm named *granular computing* (abbreviated GrC) [49,50,52,62]. GrC deals with representing and processing information in the form of some type of aggregates. These aggregates are generally called *information granules* or simply granules and they arise in the process of data abstraction and knowledge derivation from data. The scope of GrC covers various fields of study related to knowledge representation and data extraction. In 1979 the concept of *information granularity* was introduced by Zadeh [77] and it was related to the research on fuzzy sets. Next, the term *granular computing* was introduced again by Zadeh in 1997 (see [78]). Roughly speaking, information granules are collections of entities arranged together due to their similarity, functional or physical adjacency, indistinguishability, and so on. Since 1979, granular computing has become a very developed area of research in the scope of both applied and theoretical information science [50,62]. From a methodological perspective, GrC can be considered as an important attempt to investigate several research fields by means of the unifying granularity paradigm: rough set theory [16,46–48,65,75] and its generalizations [14,74,76], information tables [13,28,30,72], operative research [35], machine learning [73], interval analysis [40], formal concept analysis [39,67], database theory [36,53], data mining [37,41,42,69], fuzzy set theory [51,78], fuzzy-rough set theory [38], interactive computing [55,56]. The unifying perspective of GrC provides the useful interpretation tool which permit us to assign a same name for several notions used in different research fields. This is, for example, the case of two very important notions investigated (often with slightly

* Corresponding author.
E-mail addresses: giampiero.chiaselotti@unical.it, g.chiaselotti@gmail.com (G. Chiaselotti), gentile@mat.unical.it (T. Gentile), f.infusino@mat.unical.it (F. Infusino), paolo.oliverio@unical.it (P.A. Oliverio).

https://doi.org/10.1016/j.amc.2017.10.031
0096-3003/

different names) in data mining, database theory, rough set theory and fuzzy set theory: the *attribute approximation accuracy* [46–48,63,70,79,80] and the *attribute dependency* [68] for data tables. The utility of the above notions is so important in analyzing the data from a bi-dimensional table, that it goes beyond the scope of the specific disciplines that study them. Therefore, more generally, we prefer to speak of attribute approximation accuracy and attribute dependency in GrC by means of the unifying terminology derived from GrC itself.

In this paper, we study the two above notions in the digraph context. To be more specific, we interpret any finite digraph D (without loops and parallel arcs) as a particular type of Boolean information system. We use the more natural way to interpret a digraph D in terms of information system: its adjacency matrix $Adj(D)$. We apply the classical notion of *indiscernibility relation* [45] for any information system, in order to compute the basic rough approximation functions and the positive regions for three basic digraph families: the n-directed cycle $\vec{C}_n$, the n-directed path $\vec{P}_n$ and the n-transitive tournament $\vec{T}_n$. We associate to any information system two new data tables: the *approximation accuracy table* and the *attribute dependency table*, which contain, respectively, the approximation accuracies and the attribute dependencies with respect to any possible pair of attribute subset of the information system. We can note that if one user needs to know *all* the approximation accuracies or *all* the attribute dependencies of one only specific information system, he must build the corresponding approximation accuracy table or the corresponding attribute dependency table of its information system. Nevertheless, if the examined information system has n attributes, the dimension of the above tables is $2^n \times 2^n$, therefore in practice it is very difficult to build information system families with increasing attribute number in order to evaluate how these tables change when n becomes very large. From this perspective, some basic types of digraphs are excellent mathematical models to study the variability of the attribute dependency and the approximation accuracy when the number of attributes grows indefinitely. In the case of the attribute dependency, for any information system we also introduce an average which globally measures a type of dependency degree associated to the examined information system. We completely determine this new numerical parameter for any finite directed path: in this case we provide both a closed formula and an exact asymptotic estimate. For the transitive tournament we also provide a lower estimate for the aforementioned dependency average.

This work can be considered a continuation of a research project started in [18,22,27,29,31] and further developed in [23–26], where the incidence matrix of a hypergraph and the adjacency matrix of a simple undirected graph have been studied as particular types of information tables (for investigations concerning graphs and hypergraphs with RST and related methods in the scope of the mathematical morphology and spatial information systems see [57–61]). As a matter of fact, in [22,27,31] has been showed how a detailed study of finite undirected simple graphs by means of the investigation tools derived from GrC, leads to discover new and unexplored mathematical properties of these discrete structures. Let us note that in any finite discrete structure which is formalizable by means of the notion of information system, we obtain two basic advantages in studying such a structure from the standpoint of the information systems: (C1) to isolate *only* the specific information of our interest, neglecting all not relevant notions; (C2) to provide a temporal dynamic interpretation to an apparently static situation. In reference to (C1), to many discrete structure (graphs, digraphs, groups, posets) we can associate several types of information systems, depending on our study necessity. When we choose a specific type of information system for our structure, this means that we are selecting a way to reduce the structure study complexity towards a more particular point of view. Then, after making this choice, we can use all RST investigation tools to study our structure from the simplified observatory induced from our decision. In reference to (C2), the papers [8,9,15,33] can be methodologically considered some of the more recents and interesting tempts of studying dynamically an information table. In this perspective, the main advantage of interpreting a discrete structure as an information system is that to provide a global and hierarchical classification for the knowledge obtained from *all* possible choices of attribute subsets. The key tool to visualize dynamically the study of the indiscernibility relations in an information system is the *granular partition lattice* (see [71]). This order structure of all the indiscernibility partitions is the formal tool allowing us to visualize hierarchically and globally the knowledge evolution in any information system. We can consider the study of the granular partition lattice of any discrete structure having a corresponding associated information system as a natural way to provide a dynamic interpretation to an apparently static situation. In fact, we can imagine an initial indistinct knowledge, obtained by choosing an attribute empty subset, which next evolves through various temporal ramifications whenever one adjoins further attributes to its available knowledge. In this temporal process, the complete knowledge is reached when one can obtain the availability of *all* attributes in its information system. For example, the graphs can be studied by means of sequential or parallel dynamics (see [1,2,4–7]), or also for their analogies with both sequential and parallel dynamics on order structures (see [10–12,17,19–21]).

We conclude this introductory section with a brief description of the content of each section of this paper. In Section 2 we give the necessary definitions and introduces the notations which we will use in the sequel. In Section 3 we introduce the basic granular computation on digraphs. In particular, we provide a geometrical characterization of the indiscernibility relation in digraph context. Section 4 is devoted to the study of the Pawlak exactness in our context. To illustrate our analysis we use three simple families of digraphs: the n-path $\vec{P}_n$, the n-cycle $\vec{C}_n$ and the n-transitive tournament $\vec{T}_n$. In Section 5 we introduce the concept of attribute dependency average for an information system $\mathcal{I}$. This statistical concept is introduced in order to measure the average capacity to transmit dependency of a generic attribute subset of $\mathcal{I}$. In Section 6 we determine the attribute dependency average for the information system associates to the digraph $\vec{P}_n$ and we provide a lower bound for the digraph $\vec{T}_n$.