Characterizing Pawlak’s Approximation Operators

Victor W. Marek

Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046, USA

To the memory of Zdzisław Pawlak, in recognition of his friendship and guidance

Abstract. We investigate the operators associated with approximations in the rough set theory introduced by Pawlak in his [14,11] and extensively studied by the Rough Set community [16]. We use universal algebra techniques to establish a natural characterization of operators associated with rough sets.

1 Introduction

The concept of rough set determined by an equivalence relation $R$ has been introduced by Pawlak [14,11] in his studies of data mining. It is a natural extension of a model of database introduced in [10] that treats records as objects which may be indiscernible in the language (i.e. the tables are bags, not sets, of records). Rough sets and a set of associated numerical measures allow for capturing various degrees of similarity of objects such as records, documents, or other data units.

Rough sets and a set of associated numerical measures allow for capturing various degrees of similarity of objects such as records, documents, or other data units. The point of departure of Pawlak was the realization that the descriptive languages are often inadequate to correctly describe concepts (i.e. – in set-theoretic terms – subsets of the domain). The express goal of rough sets was to operate in the following situation: we have a collection of objects $X$ and some description language $L$. We have some collection of objects $Y \subseteq X$. We would like to describe $Y$ in the language $L$. That is we would like to find a formula $\varphi$ of $L$ so that

$$ Y = \{ x \in X : \varphi[x] \}. $$

We call such sets $Y$ definable. While usually the number of available definitions is infinite, even in the situation when $X$ is finite, not every subset has to be definable. Yet another point, made in [12], is that a set $Y$ may be definable in the language $L$ but all the definitions are prohibitively large. In such circumstances we may want to find a smaller language $L'$ where $Y$ is not definable $L'$, but the approximations are definable in $L'$ by short formulas. This is certainly the case in various medical applications.

In his analysis, Pawlak observed that in the case of finite set $X$, there is a largest subset of $Y$ that is definable, and a least superset of $Y$ that is definable.
There is a way to compute these largest and least definable subsets of $X$ that are associated with $Y$. Specifically, this is done with the *indiscernibility relation* associated with the language $L$. In finite case, there is always a formula defining a *least definable set* containing a given object $x$. Let us call these sets *monads* (for the lack of better name, and for the fact that they resemble Leibniz monads). Then, it turns out that the largest definable set included in $Y$ is the union of monads that are entirely included in $Y$, while the least definable set containing $Y$ consists of those monads that have a nonempty intersection with $Y$. Abstracting from the existence of a specific language and its logical operations, Pawlak introduced the notion of *indiscernibility relation* in the set $X$. This is the equivalence relation $R$ so that the monads are its cosets.

We believe that the guiding examples motivating Pawlak were standard medical terminologies such as SNOMED ([19]) or ICD-9 ([5]) and their inadequacies for description of classes of medical cases. It is worth mentioning that for many years Pawlak collaborated with physicians interested in Medical Informatics (needless to say, this started long before the term *Medical Informatics* were even coined). Pawlak was concerned with the fact that medical reasoning approximates the available data, often disregarding values of some attributes. As a result, it is often difficult, for a variety of reasons, to classify medical cases. If one treats a medical condition as an ideal set of cases and attempts to describe it within a concrete language of some terminology then all a physician can do is to produce a differential diagnosis. This leads, naturally, to lower and upper approximations of the classes of medical cases. While Pawlak’s intuitions were motivated by his collaborations with practicing physicians, it turned out that the methodology of approximations and indiscernibility relations are a common phenomenon. We refer the reader to monographs and journals devoted to rough set theory ([16]) for further motivations.

Let us assume that the underlying set $X$ is finite. Denoting by $\overline{R}(Y)$ and $\overline{R}(Y)$, respectively, the largest definable subset of $Y$ and the least definable superset of $Y$, we get the desired approximation relationships

$$\overline{R}(Y) \subseteq Y \subseteq R(Y).$$

The sets $\overline{R}(Y)$ and $\overline{R}(Y)$ provide collectively measure of adequacy of the underlying language to the task of describing $Y$. Moreover, by various statistical operations on those sets, and on other sets derived by set-theoretic means, we can analyze the properties of the set $Y$ itself. For that reason we would like to know more about the sets $\overline{R}(Y)$ and $R(Y)$. We would like to know what are possible operators of the form $\overline{R}(\cdot)$ and $R(\cdot)$, and how those behave when $R$ vary (i.e. when the language changes). These issues, to some extent were addressed in recent [9], but the review of the literature indicates that the Rough Sets community investigated a number of possible explanations for the rough set formalism by immersing it into various well-known mathematical areas. Those areas are all related to a variety of ways in which one can describe databases. We will list several different areas which were explored, although more could be mentioned. The references are, by necessity, incomplete. The first one is the