23rd Annual

1995 ACM COMPUTER SCIENCE CONFERENCE

The Shrinking Footprint and Growing Impact

Nashville, Tennessee
February 28 - March 2, 1995

PROCEEDINGS

Editors
C. Jinshong (C.J.) Hwang
Southwest Texas State University
Betty W. Hwang
Victina Systems International

Program Chair
Richard Brice
MCC Corporation

ASSOCIATION FOR COMPUTING MACHINERY
1515 Broadway
New York, NY 10036-9998

The Association for Computing Machinery, Inc.
1515 Broadway
New York, NY 10036-9998

ISBN: 0-89791-737-5

Additional copies may be ordered prepaid from:

ACM Order Department
P.O. Box 12114
Church Street Station
New York, NY 10257
1-800-342-6626 (USA and Canada)
1-212-626-0500 (All other countries)
FAX: 1-212-944-1318
E-Mail: acmpubs@acm.org

ACM Order Number: 404950

WORKSHOP ON ROUGH SETS AND DATABASE MINING

T.Y. Lin

The theory of rough sets, originated by Zdzislaw Pawlak, has been develop rapidly in the past decade, and has evolved into a technology. This emerging new technology concerns the classificatory analysis of imprecise, uncertain or incomplete information. Database mining can be defined as the process of mining for implicit, previously unknown, and potentially useful information from very large databases by efficient knowledge discovery techniques. It is one of the most promising research topics in the fields of database systems and machine learning. In the past few years, it has been demonstrated that rough set theory is a very effective methodology for data analysis in the attribute-value based domains. It is an efficient technique for investigating data mining in relational databases.

The main objective of the Workshop on CSC'95 is to provide a forum for researchers from rough set and database mining communities to discuss their results, their viewpoints, and to identify the future directions of the development and research in rough sets and database mining.

WORKSHOP PARTICIPANTS:
T. Y. Lin (Chair), San Jose State University, USA,
Nick Cercone (Co-Chair), University of Regina, Canada,
Zdzislaw Pawlak (Honorary Chair), Warsaw University of Technology, Poland,
Jerzy Grzymala-Busse, University of Kansas, USA,
Vijay Raghavan, University of Southwestern Louisiana,
Zbigniew Ras, University of North Carolina, USA,
Andrzej Skowron, Warsaw University, Poland,
Wojciech Ziarko, University of Regina, Canada

ROUGH REAL FUNCTIONS AND ROUGH CONTROLLERS*

Zdzislaw Pawlak
Institute of Computer Science
Warsaw University of Technology
ul. Nowowiejska 15/19, 00 665 Warsaw, Poland

and

Institute of Theoretical and Applied Informatics
Polish Academy of Sciences
ul. Baltycka 5, 44 000 Gliwice, Poland

1 Introduction

This paper is an extension of articles Pawlak (1987, 1994), where some ideas concerning rough functions were outlined. The concept of the rough function is based on the rough set theory (cf. Pawlak, 1991) and is needed in many applications, where experimental data are processes, in particular as a theoretical basis for rough controllers (cf. Czogala et al., 1994, Mrozek and Plonka, 1994).

The presented approach is somehow related to nonstandard analysis (Robinson, 1970), measurement theory (cf. Orlowska and Pawlak, 1984) and cell-to-cell mapping (cf. Hsu, 1980) but these aspects of rough functions will be not considered here.

In recent years we witness rapid grow of development and applications of fuzzy controllers. The philosophy behind fuzzy control is that instead of describing, as in the case of classical control theory, the process being controlled in terms of mathematical equations - we describe the behavior of human controller in terms of fuzzy decision rules, i.e. rules that involve rather qualitative then quantitative variables and can be seen as a common-sense model of the controlled process, similarly as in qualitative physics physical phenomena are described in terms of qualitative variables instead of mathematical equations.

The idea of rough (approximate) control steams yet from another philosophical background. It is based on the assumption that the controlled process is observed and data about the process are registered. The data are then used to generate the control algorithms, which can be afterwards optimized. Both, the generation of the control algorithm from observation, as well the optimization of the algorithm can be based on the rough set theory, which seems to be very well suited for this kind of tasks. The control algorithms

*This work was supported by grant No. 8 S503 021 06 from State Committee for Scientific Research

obtained in this way are objective and can be viewed as an intermediate approach between classical and fuzzy approach to control systems.

In some cases the observation can be postponed and control algorithm can be obtained directly from the knowledgeable expert, similarly as in the fuzzy set approach. In this case the control algorithm can be also simplified using the rough set theory technic.

In general we assume that a rough controller can be seen as an implementation of rough (approximate) function, i.e. function obtained as a result of physical measurements with predetermined accuracy, depending on assumed scale.

The aim of this paper is to give basic ideas concerning rough functions, which are meant to be used as a theoretical basis for rough controllers synthesis and analysis. The presented ideas can be also applied to other problems – in general to discrete dynamic systems, and will be discussed in further papers.

2 Basic of the Rough Set Concept

Basic ideas of the rough set theory can be found in Pawlak (1991). In this section we will give only those notions which are necessary to define concepts used in this paper.

Let U be a finite, nonempty set called the *universe*, and let I be an equivalence relation on U, called an *indiscernibility relation*. By $I(x)$ we mean the set of all y such that xIy, i.e. $I(x) = [x]_I$, i.e.- is an equivalence class of the relation I containing element x. The indiscernibility relation is meant to capture the fact that often we have limited information about elements of the universe and consequently we are unable to discern them in view of the available information. Thus I represents our lack of knowledge about U.

We will define now two basic operations on sets in the rough set theory, called the *I-lower* and the *I-upper approximation*, and defined respectively as follows:

$$I_*(X) = \{x \in U : I(x) \subseteq X\},$$

$$I^*(X) = \{x \in U : I(x) \cap X \neq \emptyset\}.$$

The difference between the upper and the lower approximation will be called the *I-boundary* of X and will be denoted by $BN_I(X)$, i.e.

$$BN_I(X) = I^*(X) - I_*(X).$$

If $I^*(X) = I_*(X)$ we say the the set is *I-exact* otherwise the set X is *I-rough*. Thus rough sets are sets with unsharp boundaries.

Usually in order to define a set we use the membership function. The membership function for rough sets is defined by employing the equivalence relation I as follows:

$$\mu_X^I = \frac{card(X \cap I(x))}{card I(x)}.$$

Obviously

$$\mu_X^I(x) \in [0, 1].$$

The value of the membership function expresses the degree to which the element x belongs to the set X in view of the indiscernibility relation I.

The above assumed membership function, can be used to define the two previously defined approximations of sets, as shown below:

$$I_*(X) = \{x \in U : \mu_X^I(x) = 1\},$$

$$I^*(X) = \{x \in U : \mu_X^I(x) > 0\}.$$

3 Rough Sets on the Real Line

In this section we reformulate the concepts of approximations and the rough membership function referring to the set of reals, which will be needed to formulate basic properties of rough real functions.

Let R be the set of reals and let (a,b) be an open interval. By a *discretization* of the interval (a,b) we mean a sequence $S = \{x_0, x_1, \ldots, x_n\}$ of reals such that $a = x_0 < x_1 < \ldots < x_n = b$. Besides, we assume that $0 \in S$. The ordered pair $A = (R, S)$ will be referred to as the *approximation space* generated by S or simple as S-*approximation space*. Every discretization S induces the partition $\pi(S) = \{\{x_0\}, (x_0 x_1), \{x_1\}, (x_1, x_2), \{x_2\}, (x_2, x_3), \{x_3\}, \ldots, \{x_{n-1}\}, (x_{n-1}, x_n), \{x_n\}\}on(a, b)$. By $S(x)$ (or $[x]_S$) we will denote block of the partition $\pi(S)$ containing x. In particular, if $x \in S$ then $S(x) = \{x\}$. The closed interval $[a,b]$ will be denoted by $S'(x)$, and will be referred to as the *closure* of $S(x)$.

In what follows we will be interested in approximating intervals $(0, x) = Q(x)$ for any $x \in [a, b]$.

Suppose we are given an approximation space $A = (R, S)$. By the S-*lower* and the S-*upper* approximation of $Q(x)$, denoted by $S_*(Q(x))$ and $S^*(Q(x))$ respectively, we mean sets defined below:

$$S_*(Q(x)) = \{y \in R : S(y) \subseteq Q'(x)\}$$

$$S^*(Q(x)) = \{y \in R : S(y) \cap Q'(x) \neq \emptyset\}.$$

The above definitions of approximations of the interval $< 0, x >$ can be understood as approximations of the real number x which are simple the ends of the interval $S(x)$, therefore we will use the following abbreviations: $S_*(Q(x)) = S_*(x)$ and $S^*(Q(x)) = S^*(x)$. If $X \subseteq R$, then $\Delta(X) = Sup|x - y|, x, y \in X$. In particular $\Delta(S(x))$ will be denoted by $\Delta_S(x)$ and will be called the *length* of X with respect to scale S.

In other words given any real number x and a discretization S, by the S-lower and the S-upper approximation of x we mean the numbers $S_*(x)$ and $S^*(x)$, which can be defined as

$$S_*(x) = Sup\{y \in S : y \leq x\}$$

$$S^*(x) = Inf\{y \in S : y \geq x\}.$$

for $x \geq 0$ and

$$S_*(x) = Inf\{y \in S : y \geq x\}$$

$$S^*(x) = Sup\{y \in S : y \leq x\}.$$

for $x \leq 0$.
Thus $S(x) = (S_*(x), S^*(x))$.

We will say that the number x is exact in $A = (R, S)$ if $S_*(x) = S^*(x)$, otherwise the number x is *inexact* (*rough*) in $A = (R, S)$. Of course x is exact iff $x \in S$. Thus to every inexact number x we can associate pair of exact numbers $S_*(x)$ and $S^*(x)$ (the lower and the upper approximations) and the interval $S(x)$).

Any discretization S can be interpreted as a scale (e.g. km, in, etc), by means of which reals from R are measured with some approximation due to the scale S.

Remark

We can also assume that the discretizatin S induces partition $\pi(S) = \{(-INF, x_0), \{x_0\}, (x_0 x_1), \{x_1\}, (x_1, x_2), \{x_2\}, (x_2, x_3), \{x_3\}, \ldots, \{x_{n-1}\}, (x_{n-1}, x_n), \{x_n\}, (x_n, +INF)\}$ on $\mathbf{R}$. In this case for $x > b$ the upper approximation of x is $S^*(x) = +INF$, and similarly for $x < a$, we have $S^*(x) = -INF$. However for the sake of simplicity we will not consider this case here. □

The introduced idea of the rough set on the real line corresponds exactly to those defined for arbitrary sets and can be seen as a special case of the general definition.

Now we give the definition of the next basic notion in the rough set approach - the rough membership function – referring to the real line (cf. Pawlak and Skowron, 1993).

The rough membership function for set on the real line have the form

$$\mu_{Q(x)}(y) = \frac{\Delta(Q(x) \cap S(y))}{\Delta(S(y))},$$

where $\Delta(X) = Sup|x - y|, x, y \in X$.

Assuming that $x = y$, we get

$$\mu_{Q(x)}(y) = \mu(y),$$

which can be understood as an error of measurement of x in scale S.

4 Rough Sequencies and Rough Functions

Let $A = (R, S)$ be an approximation space and let $\{a_n\}$ be an infinite sequence of real numbers.

A sequence $\{a_n\}$ is *roughly convergent* in $A = (\mathbf{R}, S)$, (*S-convergent*), if there exists i such that for every $j > i$ $S(a_j) = S(a_i); S_*(a_i)$ and $S^*(a_i)$ are referred to as the *rough lower* and the *rough upper limit* (*S-upper, S-lower limit*) of the sequence $\{a_n\}$. Any roughly convergent sequence will be called rough *Cauchy sequence.*

A sequence $\{a_n\}$ is *roughly monotonically increasing* (*decreasing*) in $A = (\mathbf{R}, S)$, (*S-increasing* (*S-decreasing*)), if $S(a_n) = S(a_{n+1})$ or $a_n < a_{n+1}(a_n > a_{n+1})$ and $S(a_n) \neq S(a_{n+1})$.

Obviously, $\{a\}$ is the Cauchy sequence iff $\{a\}$ is roughly monotonically increasing or decreasing.

A sequence $\{a_n\}$ is *roughly periodic* in $A = (\mathbf{R}, S)$ (*S-periodic*), if there exists k such that $S(a_n) = S(a_{n+k})$. The number k is called the period of $\{a_n\}$.

A sequence $\{a_n\}$ is *roughly constant* in $A = (\mathbf{R}, S)$ (*S-constant*), if $S(a_n) = S(a_{n+1})$.

Suppose we are given a real function $f : X \to Y$ with discretizations S and P on X and Y respectively. With every function f we associate the function $F_f : \pi(S) \to \pi(P)$ such that

$$F_f(S(x)) = P(f(x)).$$

Thus the function F_f assigns unequally to each block of the partition $\pi(S)$ one block of the partition $\pi(P)$. We can enumerate blocks of partitions $\pi(S)$ and $\pi(P)$ by integers in the following way:

$N(S(x)) = i, 0leqi \leq n$, if $S_*(x)) = x_1$, where $S = \{x_0, x_1, \ldots, x_n\}$. Now instead of function F_f we can use the function $f_S : \{n\} \to \{n\}$, from integers to integers defined as follows:

$$f_S(i) = N(P(f(x_i))).$$

The function f_S will be called the *discretization* of f.

The function f_S can be used to define some properties of real functions.

A function f is roughly *monotonically increasing* (*decreasing*) if $f_S(i+1) = f(i) + \alpha$, where α is a non-negative integer, (α is non-positive integer), for every $i = 0, 1, 2, \dots n-1$.

A function f is *roughly periodic* if there exist k such that $f_S(i) = f_S(i+k)$ for every $i = 0, 1, \dots, n-1$.

A function f is *roughly constant* if $f_S(i) = f_S(i+1)$, for every $i = 0, 1, \dots, n-1$.

Now we give a definition of a very important concept, the rough continuity of real function.

Suppose we are given a real function $f : X \to Y$, where both X and Y are sets of reals and S, P are discretizations of X and Y respectively.

A function f is (*S,P*)-*continuous* (*roughly continuous*) in x if

$$f(S(x)) \subseteq P(f(x)).$$

In other words a function f is roughly continuous in x iff for every $y \in S(x) f(y) \in P(f(x))$.

If f is roughly continuous in x for every $x \in \delta(S)$, where $\delta(S) = (x_0, x_n)$, we say that f is (*S,P*)-*roughly continuous.*

The intuitive meaning of this definition is obvious. Whether the function is roughly continuous or not depends on the information we have about the function, i.e. it depends on how exactly we "see" the function through the available information (the indiscernibility relation).

Obviously a function f is roughly continuous iff $F_f(i+1) \in \{-1, 0, +1\}$ for every $i = 0, 1, \dots, n-1\}$.

Particularly interesting is the relationship between dependency of attributes in information systems and the rough continuity of functions

Let $\mathbf{S} = (U, A)$, be an *information system*, (cf. Pawlak, 1991), where U is a finite set of *objects*, called the *universe* and A is a finite set of attributes. With every attribute $a \in A$ a set of *values* of attribute a, called *domain* of a, is associated and is denoted by V_a . Every attribute $a \in A$ can be seen as a function $a : U \to V$, which to every object $x \in U$ assigns a value of the attribute a. Any subset of attributes $B \subseteq A$ determines the equivalence relation $IND(B) = \{x, y \in U : a(x) = a(y) \text{for every} a \in A\}$. Let $B, C \subseteq A$. We will say that the set of attributes C *depends* on the set of attributes B, in symbols $B \to C$, iff $IND(B) \subseteq IND(C)$. If $B \to C$ then there exists a *dependency function* $f_{B,C} : V_{b_1} x V_{b_2} x \dots x V_{b_n} \to V_{c_1} x V_{c_2} x \dots x V_{c_m}$, such that $f_{B,C}(v_1, v_2, \dots, v_n) = (w_1, w_2, \dots, w_m)$, iff $\sigma(v_1) \cap \sigma(v_2) \cap, \dots, \cap \sigma(v_n) \subseteq \sigma(w_1) \cap \sigma(w_2) \cap, \dots, \cap \sigma(w_m)$, where $v_1 \in V_b, w_j \in V_{c_j}, \sigma(v) = \{x \in U : a(v) = x\}$ and $v \in V_a$. The dependency function $B \to C$, where $B = \{b_1, b_2, \dots, b_n\}$ and $C = \{c_1, c_2, \dots, c_m\}$ assigns uniquely to every n-tuple of values of attributes from B the m-tuple of values of attributes from C.

There exists the following important relationship. $B \to C$ iff $f_{B,C}$ is (*B,C*)-roughly continuous.

Many other basic concepts concerning functions can be expressed also in the rough function setting.

By the (*P*)-*lower approximation* of f we understand the function $f_* : X \to Y$ such that

$$f_*(x) = P_*(f(x)) \text{for every} x \in X.$$

Similarly the (*P*)-*upper approximation* of f is defined as

$$f^*(x) = P^*(f(x)) \text{for every} x \in X.$$

We say that a function f is *exact* in x iff $f_*(x) = f^*(x)$; otherwise the function f is *inexact* (*rough*) in x. The number $f^*(x) - f_*(x)$ is the *error of approximation* of f in x.

If f is a (S,P)-continuous function, then F_f is exact for every $x \in S$.

Finally in many applications we need the fix point properties of functions.

We say that x is a rough fix-point (*rough equilibrium point*) of a real function f if $F_f(N(S_*(x)) = N(S_*(x)$.

5 Optimal Discretization of Rough Functions

The function $f_S : S \to Y$ such that $f_S(x) = f(x)$ for any $x \in S$ will be called a *S-discrete representation* of f or in short *S-discretization* of f.

Our main task is to give interpolation algorithm for discrete representation f_S giving the best approximation of f.

Let us first consider the *linear interpolation* formula. The linear interpolation of f will be denoted by f_αFs and is defined as follows:

$$f_\alpha(x) = f(S_*(x)) + \mu_{Q(S_*(x))}(x).\Delta f(S(x)),$$

where $\Delta f(S(x)) = f(S^*(x)) - f(S_*(x))$.
The number

$$\frac{|f\alpha(x) - f(x)|}{f(x)}$$

will be called the relative error of the interpolation of f in x. The maximal error of interpolation will be called the error of interpolation of f.

If $f(S_*(x))$ and $f(S^*(x))$ are unknown we can use another interpolation formulas shown below.

1) Lower interpolation

$$f^1_\alpha(x) = P_*(f(S_*(x))) + \mu_{Q(S_*(x))}(x).\Delta^1 f(S(x)),$$

where $\Delta^1 f(S(x)) = P_*(f(S^*(x))) - P_*(f(S_*(x)))$;

2) Upper interpolation

$$f^2_\alpha(x) = P^*(f(S_*(x))) + \mu_{Q(S_*(x))}(x).\Delta^2 f(S(x)),$$

where $\Delta^2 f(S(x)) = P^*(f(S^*(x))) - P^*(f(S_*(x)))$;

3) Lower cross interpolation

$$f^3_\alpha(x) = P_*(f(S*(x))) + \mu_{Q(S_*(x))}(x).\Delta^3 f(S(x)),$$

where $\Delta f(S(x)) = P^*(f(S^*(x))) - P_*(f(S_*(x)))$;

4) Upper cross interpolation

$$f^4_\alpha(x) = P^*(f(S_*(x))) + \mu_{Q(S_*(x))}(x).\Delta^4 f(S(x)),$$

where $\Delta^4 f(S(x)) = P_*(f(S^*(x))) - P^*(f(S_*(x)))$;

The meaning of the above interpolation formulas is obvious.

We will be also interested in the following problem. Given a function $f : X \to Y$ and a number $0 \le \epsilon \le 1$. Find categorizations S and P such the the error of interpolation of f is less than ϵ.

The following algorithm solves the problem.

$$x_0 = z \in X$$

$$x_{i+1} = Sup\{x > x_1 : |f(y) - p_{x_i x}(y)| \leq \epsilon\} \text{for any} y \in (x_i, x >,$$

where $p_{x_i,x}(y)$ denotes the straight line determined by points x_i, x.

6 Conlusions

Rough function concept is meant to be used as a theoretical basis for rough controllers. Basic definitions concerning rough functions were given and some basic properties of these functions investigated.

Applications of the above discussed ideas will be presented in the forthcoming papers.

References

[1] **Czogala, E., Mrozek, A. and Pawlak, Z. (1994).** Rough-Fuzzy Controllers. *ICS WUT Reports*, 32/94.

[2] **Hsu, C.S. (1980).** A Theory of Cell-To-Cell Mapping Dynamical Systems.*ASME Journal of Applied Mechanics,* Vol. 47, pp. 931-939.

[3] **Mrozek, A. and Plonka, L. (1994).** Rough Sets for Controller Synthesis. *ICS WUT Report* 51/94.

[4] **Orlowska, E., and Pawlak, Z. (1984)**. Measurement and Indiscernibility. *Bull. PAS, Math. Ser.* Vol. 32, No. 9-10, pp. 617-624.

[5] **Pawlak, Z. (1991).** Rough Sets - TheoreticaL Aspects of Reasoning about Data. KLUWER ACADEMIC PUBLISHERS.

[6] **Pawlak, Z. and Skowron, A. (1993).** Rough Membership Functions. In: Yaeger, R.R., Fedrizzi, M., and Kacprzyk, J. (Eds.), *Advances in the Dempster Shafer Theory of Evidence,* John Wiley and Sons, pp. 251-271.

[7] **Pawlak, Z. (1987).** Rough Functions, *Bull. PAS, Tech. Ser.Vol.* 35, No.5-6, pp. 249-251.

[8] **Pawlak, Z. (1994).** Rough Sets, Rough Real Functions, *ICS WUT Repart,* 50/94

[9] **Robinson, A. (1970).** Non-Standard Analysis. NORTH- HOLLNAD PUBLISHING COMPANY.

[10] **Slowinski, R. (Ed.) (1992).** Intelligent Decision Support - Handbook of Advances and Applications of the Rough Set Theory. KLUWER ACADEMIC PUBLISHERS.