Rough Numbers and Rough Regression

Marcin Michalak
Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland
Marcin.Michalak@polsl.pl

Abstract. In this article a new model of regression is defined. On the basis of the rough sets theory a notion of rough number is defined. Typical real numbers calculations do not keep the additional information like the uncertainty or the error of input data. Rough numbers remove this limitation. It causes that rough numbers seem to be interested as the basis of the new way of regression: rough regression.

Keywords: rough sets, rough numbers, machine learning, nonparametric regression, rough regression.

For over forty years people have been trying to specify the inexactitude with notions like the fuzzyness [7] or roughness [3]. This human need of problems simplification by generalization is associated with the price of loosing the accuracy. In this article a new approach of inexact data representation is described that extends the limited Pawlak definition of a rough number [4]. With the wider definition of the approximation space this definition makes it possible to define basic arithmetical operation on rough numbers based on the whole \( \mathbb{R} \) set. It finds the application in the nonparametric regression.

1 Classical and Extended Models of Rough Numbers

Rough numbers [4] are based on the several following notions. For the set \( \mathbb{R}^+ \) a sequence of nonnegative reals \( S=\{x_i\}_{i=1}^n \) such that \( x_i<x_j, i<j \) is called the categorization. The approximation space is the ordered pair \( A=(\mathbb{R}^+, S) \). Every categorization \( S \) of \( \mathbb{R}^+ \) induces partition \( \pi(s) \) on \( \mathbb{R}^+ \) defined as \( \pi(S)=\{0,(0,x_1),x_1,\ldots,x_i,(x_i,x_{i+1}),\ldots\} \) where \( (x_i,x_{i+1}) \) denotes the open interval. \( S(x) \) denotes the block of partition \( \pi(S): x \in \pi(S) \). Let \( x\in(x_i, x_{i+i}) \). The closed interval \( \overline{S}(x)=\langle x_i, x_{i+1} \rangle \) is called the closure \( S(x) \). \( Q(x) \) is the closed interval \( \langle 0, x \rangle \). For a given approximation space \( A(\mathbb{R}^+, S) \) for every \( Q(x) \) its lower and upper approximation may be defined, denoted as \( S_*(Q(x)) \) and \( S^*(Q(x)) \) respectively. \( S_*(Q(x))=\{y\in\mathbb{R}^+: S(y)\subseteq Q(x)\}; S^*(Q(x))=\{y\in\mathbb{R}^+: S(y)\cap Q(x)\neq\emptyset\} \). As the S-lower and S-upper approximation of the nonnegative real number \( x \) the following values are considered: \( S_*(x) = \text{Sup}\{y \in S : y \leq x\}; S^*(x) = \text{Inf}\{y \in S : y \geq x\} \). All notions above lead to the definition of an approximation of a real number \( x \) on the basis of a set of real numbers \( S: S(x) = (S_*(x), S^*(x)) \). The number is exact in \( A = (\mathbb{R}^+, S) \) iff \( S_*(x) = S^*(x) \), otherwise it is inexact (rough) in \( A \). Every inexact number may be represented as a pair of exact numbers.
or as the interval. Now, let \( S \) is the categorization of \( \mathbb{R} \): \( S \subseteq \mathbb{R} \). An approximation space \( A = (\mathbb{R}, S) \) will be denoted as \( A_S \). For a given \( S = \{x_i\}_{i=1}^n \) the generalized projection \( P \) is defined: \( P(S) = \{(-\infty, x_1), x_1, (x_1, x_2), x_2, \ldots\} \). Notions \( S(x) \) and \( \overline{S}(x) \) remain as above. As \( A_S \) is based on \( \mathbb{R} \), \( Q(x) \) should be defined as \( Q(x) = (-\infty, x) \) what causes that definitions of its \( S \)-lower and \( S \)-upper approximations remain unchanged. Also definitions of real number \( x \) remain the same as in [4]. This extension of the Pawlak definitions leads us to the notion of rough real number. Every \( A_S \) gives \( \overline{S} \) exact numbers and \( \overline{S} + 1 \) rough numbers. As the exact number \( x \) is also the rough number \( \langle x, x \rangle \) we say that every \( A_S \) introduces \( 2\overline{S} + 1 \) rough real numbers. Two of them has infinite limits: \( (-\infty, x_1), (x_{\text{max}}, \infty) \) and will be called nonfinite rough numbers. All other will be called finite rough numbers. Exact rough number will be denoted as \( x : \langle x, x \rangle = \mathbb{R} \); inexact finite rough number \( x : \langle S_0(x), S^*(x) \rangle \); inexact nonfinite rough number \( x : \langle -\infty, x_1 \rangle \) and \( \langle x_{\text{max}}, \infty \rangle \); set of finite rough numbers: \( \mathcal{R} \); set of rough numbers (finite and nonfinite): \( \mathcal{R}_\infty \).

2 Rough Calculations and Rough Regression

Let \( \rho_1, \rho_2 \) are finite rough numbers: \( \rho_1 = \langle a_1, a_2 \rangle, \rho_2 = \langle b_1, b_2 \rangle \). Consider: addition: \( \rho_1 + \rho_2 = \langle a_1 + b_1, a_2 + b_2 \rangle \); multiplication: \( \rho_1 \cdot \rho_2 = \langle a_1 \cdot b_1, a_2 \cdot b_2 \rangle \); mirroring: \( \rho_1 \) and \( \rho_2 \) will be called mirrored iff \( a_1 = b_2 \) and \( a_2 = b_1 \). The function \( \mu \) is called a mirror function: \( \mu : \mathcal{R} \rightarrow \mathcal{R}; \mu(\langle a, b \rangle) = \langle b, a \rangle \). It is easy to prove that \( \rho + \mu(\rho) = \mathbb{R} \). Addition and multiplication are commutative, have neutral and opposite elements, and satisfy distributivity of the multiplication over the addition. For every kernel function \( K : \mathcal{R} \rightarrow \mathcal{R} \) a rough kernel function \( K : \mathcal{R} \rightarrow \mathcal{R} \) may be defined as follows: \( K(\langle \rho_1, \rho_2 \rangle) = (\rho_2 - \rho_1)^{-1} \int_{\rho_1}^{\rho_2} K(x)dx \). Rough kernel function should take real values for real arguments. It is easy to prove that \( K(\langle \rho, \rho \rangle) = K(\rho) \) and that \( K(\rho) = K(\mu(\rho)) \). One of the most popular kernel estimators of the regression function is Nadraya-Watson estimator [2][6]: \( \hat{f}(x) = \frac{1}{\sum_{i=1}^{n} y_i(x - x_i)/h} \). K is a kernel function and \( h \) is the smoothing parameter. In this section the extension of NW estimator for rough numbers is presented. The process of roughing the real data is as follows: for the set of real valued observations \( U = \{(x_i, y_i)\}_{i=1}^{n} \) the set of \( 2n - 1 \) rough valued observations is defined in the following way: \( \mathcal{R}(U) = \{(x_1, x_1), (y_1, y_1), (x_1, x_2), (y_1, y_2), (x_2, x_2), (y_2, y_2), \ldots\} \). As the set \( X \) is a categorization of \( \mathbb{R} \) we may define the \( X \)-lower and \( X \)-upper function regression. If \( \hat{f}(\rho) \) is the rough kernel regressor then the rough function \( \mathcal{F}_X \) may be defined as follows \( \mathcal{F}_X = \{(u, \hat{f}(u)) : u \in \mathcal{R}(X) \} \) or: \( \mathcal{F}_X = \{(u_{i1}, u_{i2}), (\hat{f}(u_{i1}), \hat{f}(u_{i2})) : u_i = u_{i1} \oplus u_{i2} \in \mathcal{R}(X), \hat{f}(u_{i1}) = \langle \hat{f}(u_{i1}), \hat{f}(u_{i2}) \rangle, i = 1, \ldots, 2\overline{S} - 1 \} \). From the definition above we may define the \( \mathcal{X} \)-lower and \( \mathcal{X} \)-upper rough estimators of the function:

\[
\mathcal{F}_X = \{(u_{i1}, u_{i2}), (\hat{f}(u_{i1}), \hat{f}(u_{i2})): (u_{i1}, u_{i2}), (\hat{f}(u_{i1}), \hat{f}(u_{i2})) \in \mathcal{F}_X \}
\]

\[
\overline{\mathcal{F}}_X = \{(u_{i1}, u_{i2}), (\hat{f}(u_{i1}), \hat{f}(u_{i2})): (u_{i1}, u_{i2}), (\hat{f}(u_{i1}), \hat{f}(u_{i2})) \in \mathcal{F}_X \}
\]