Rough Sets Meet Statistics - A New View on Rough Set Reasoning about Numerical Data

Marko Palangetić[1][0000−0002−6366−0634], Chris Cornelis[1][0000−0002−7854−6025], Salvatore Greco[2,3][0000−0001−8293−8227], and Roman Słowiński[4,5][0000−0002−5200−7795]

[1] Ghent University, Ghent, Belgium
{marko.palangetic, chris.cornelis}@ugent.be
[2] University of Catania, Catania, Italy
[3] University of Portsmouth, Portsmouth, United Kingdom
salgreco@unict.it
[4] Poznań University of Technology, Poznań, Poland
[5] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
roman.slowinski@cs.put.poznan.pl

Abstract. In this paper, we present a new view on how the concept of rough sets may be interpreted in terms of statistics and used for reasoning about numerical data. We show that under specific assumptions, neighborhood based rough approximations may be seen as statistical estimations of certain and possible events. We propose a way of choosing the optimal neighborhood size inspired by statistical theory. We also discuss possible directions for future research on the integration of rough sets and statistics.

Keywords: Rough Sets · Statistical Learning · Neighborhood Based Rough Sets

1 Introduction

Zdzisław Pawlak introduced rough sets in 1982 to deal with inconsistencies within information tables [15]. His approach is applied to the representation of classes of objects in an information table using two new sets called lower and upper approximation. The lower approximation contains objects which certainly belong to the approximated class, while the objects which are possibly in the approximated class are included in the upper approximation. Formulated in another way, the approach identifies the objects which are certainly consistent with the available knowledge and the objects which are possibly consistent with it. The original method is designed to deal with categorical data or data with a finite domain.

The extension of the model to numerical data faces some difficulties. One possibility to deal with numerical data is to discretize the attributes in the information table and make them categorical [7]. However, such an approach may lead to a loss of information, since discretization considers a set of values as

one single value. The other option are neighborhood based rough sets where the equivalence class from Pawlak's approach is replaced with the neighborhood of an object in a high dimensional Euclidean space [9]. They are related to similarity based rough sets [21], and are part of the more general family of covering based rough sets [26]. The third approach are fuzzy rough sets which use fuzzy generalizations of equivalence relations suitable for application to numerical data [5]. In this paper, we use probability and statistics instead of fuzziness to model uncertainty in data.

From the very beginning, it was acknowledged that Pawlak's approach runs into limitations when it comes to problems which are more probabilistic than deterministic in nature [27]. In general, data consist of true values affected by some noise. Therefore, the first step in data analysis is to remove that noise in order to use the real values to solve the problem of interest. As a robust version of rough sets, the Variable Precision Rough Set (VPRS) approach was proposed by Ziarko [27]. It was also the first attempt to integrate the probabilistic approach and rough sets. Other probabilistic versions of rough sets were presented later, including decision theoretic rough sets [25] and parameterized rough sets [6]. Later on, Ziarko also introduced the assumption that the data are just a sample from an unknown space [28] into rough sets. That is a widely used assumption in statistics and machine learning: data are a realization of a random variable. With this assumption, we seek for a deeper integration of rough sets and statistics. In this paper, we propose a new view on the definition of rough sets, and provide a new definition independent of the type of data. It leads to a natural extension of the initial rough set approach to numerical data. We provide an example how to calculate rough sets for numerical data, elaborate on some of issues we are facing and present some ideas about how to direct the future research on integration of rough sets and statistics.

The paper is organized as follows. In the next section we recall basic concepts of rough set theory. In Section 3, statistical learning theory for Pawlak's rough sets is introduced. Section 4 presents rough approximations for numerical data. Section 5 identifies and discusses some potential pitfalls and drawbacks identified in Section 4 together with ideas for improvement. Conclusions are provided in Section 6.

2 Preliminaries

2.1 Rough sets

An information table is a 4-tuple $< U, Q \cup \{d\}, X \cup Y, f >$ where $U = \{u_1, \ldots, u_n\}$ is a finite set of objects or alternatives, $Q = \{q_1, \ldots, q_m\}$ is a finite set of condition attributes, d is a decision attribute; $X = \cup_{q \in Q} X_q$, where X_q is the domain of attribute $q \in Q$ while Y is the domain of d. The information function $f : U \times Q \cup \{d\} \to X \cup Y$ satisfies that $\forall u \in U, \forall q \in Q : f(u, q) \in X_q$ and that $f(u, d) \in Y$. Denote by $X_Q = \prod_{q \in Q} X_q$ the joint domain of condition attributes, while $f(u, Q) \in X_Q$ represents the $|Q|$-tuple of values $f(u, q)$ for $q \in Q$. If X_q is finite, we say that q is categorical, while if $X_q \subseteq \mathbb{R}$ we say that q is numerical.

First we assume that all condition attributes are categorical. We define the equivalence relation $\equiv$ on objects u and v as $u \equiv v \Leftrightarrow \forall q \in Q, f(u,q) = f(v,q)$. This means that two objects are related (indiscernible) if they are equally evaluated on all attributes. Let $[u]_\equiv$ denote the equivalence class of object u, and $A \subseteq U$. We recall Pawlak's lower and upper approximations on U:

$$\underline{\text{apr}}_\equiv(A) = \{u \in U | [u]_\equiv \subseteq A\}, \quad \overline{\text{apr}}_\equiv(A) = \{u \in U | [u]_\equiv \cap A \neq \emptyset\}.$$

In the lower approximation of A, we include objects u for which all identically evaluated objects are also in A. Therefore, we may conclude that u for sure belongs to A based on available knowledge, since all the instances with the same values are also in A. We include object u in the upper approximation of A if there is an instance in A identically evaluated as u. Hence, we may say that u is possibly in A if some instances, identically evaluated as u, are in A. In this way, we distinguish certain and possible knowledge. Below, we list the important properties of inclusion and duality [15]:

- (inclusion) $\underline{\text{apr}}_\equiv(A) \subseteq \overline{\text{apr}}_\equiv(A)$,
- (duality) $\underline{\text{apr}}_\equiv(A^c) = (\overline{\text{apr}}_\equiv(A))^c$, $\overline{\text{apr}}_\equiv(A^c) = (\underline{\text{apr}}_\equiv(A))^c$.

A question arises: how to apply a similar reasoning when we have numerical data? If we apply the reasoning presented above, the equivalence classes will mostly consist of only one object since it is almost impossible that two objects with numerical characteristics will be identically evaluated on all attributes. This means that all objects from A belong to the lower approximations of A, i.e., all objects from A certainly belong to A. However, in this way we ignore the fact that the noise present in data affects the certainty of objects belonging to a set. The noise is related to imprecision of numerical attributes and, even if the measurement of numerical attributes is precise, to human perception of these precise values.

A way to handle this problem is the neighborhood based rough set approach. Assume now that condition attributes are taking real values and let d be Euclidean distance on $X_Q \subseteq \mathbb{R}^m$. Here, any distance metrics can be used, but Euclidean distance corresponds with the later statistical approach we will use. For object $u \in U$ we define its ϵ-neighborhood $n_\epsilon(u) = \{v \in U; d(f(u,Q), f(v,Q)) < \epsilon\}$. We define the approximations in the following way [9]:

$$\underline{\text{apr}}_\epsilon(A) = \{u \in U; n_\epsilon(u) \subseteq A\}, \quad \overline{\text{apr}}_\epsilon(A) = \{u \in U; n_\epsilon(u) \cap A \neq \emptyset\}.$$

Here, object u certainly belongs to A if its close neighborhood only contains objects from A. Object u possibly belongs to A if its close neighborhood contains at least one object from A. Equivalent properties of inclusion and duality also hold in this case [9].

From the definition we may see that the approximations heavily depend on the parameter ϵ. The question is, what is the optimal neighborhood size which will identify certain and possible knowledge. Later on we will see that statistical techniques may be useful for this purpose.

2.2 Value-based definitions and inconclusive regions

Pawlak defines the approximations as sets of objects (SO). The main goal of these definitions is to distinguish possible knowledge from certain knowledge and for this we do not need to refer exactly to the set of objects. We can define the approximations as sets of values (SV), i.e., the sets which will only contain values from the domain of condition attributes. Let $x \in X_Q$. Similarly as in [8] we define sets $[x] = \{u \in U; f(u, Q) = x\}$. The SV approximations are

$$\underline{\text{apr}}^{\text{SV}}_{\equiv}(A) = \{x; [x] \neq \emptyset \wedge [x] \subseteq A\}, \quad \overline{\text{apr}}^{\text{SV}}_{\equiv}(A) = \{x; [x] \cap A \neq \emptyset\}.$$

We refer to this definition as SV definition while the original one will be called SO definition. We note that the SV definition keeps the same knowledge as the SO definition. The SO approximations can be obtained from the SV definition by collecting all objects with condition values belonging to the SV approximations (lower or upper). The SV approximations can be obtained from the SO definition as a set of unique condition values $f(u, Q)$ of the objects from the SO approximations. Therefore, in terms of Pawlak's environment of categorical data, SO and SV definitions are equivalent.

We notice that there are values from the domain which cannot be assigned to any approximation. In particular, the condition $|[x]| > 0$ is necessary in the definitions. Otherwise a value x for which $|[x]| = 0$ would belong to the lower approximations of A and A^c at the same time, i.e., it would certainly belong to two opposite classes. Of course, that is not possible and such values from the domain are called inconclusive. We denote the set $I \subseteq X_Q$ of inconclusive values by

$$I = \{x; x \in X_Q \wedge [x] = \emptyset\}$$

The inclusion property is clearly preserved while duality still holds if the complement operator on X_Q excludes inconclusive values i.e., if it is defined as: $S^c = X_Q - I - S$ for $S \subseteq X_Q$.

On the other hand, for the SV extension in the neighborhood based approximations, neighborhood may be defined for any value from the domain X_Q. If $X_Q \subseteq \mathbb{R}^m$ and $x \in X_Q$ we define $n_\epsilon(x) = \{u \in U; d(x, f(u, Q)) < \epsilon\}$. The SV approximations are:

$$\underline{\text{apr}}^{\text{SV}}_{\epsilon}(A) = \{x; n_\epsilon(x) \neq \emptyset \wedge n_\epsilon(x) \subseteq A\}$$

$$\overline{\text{apr}}^{\text{SV}}_{\epsilon}(A) = \{x; n_\epsilon(x) \cap A \neq \emptyset\}.$$

An arbitrary value $x \in X_Q$ is in the lower approximation of A if its ϵ-neighborhood contains only objects from A while it is in the upper approximation if it contains at least one object from A. Here again we consider the inconclusive areas, i.e., values in which neighborhood there are no objects from U. As for the SV definitions for Pawlak's rough sets, the inclusion property is preserved while duality holds with exclusion of the inconclusive areas. The SO and SV definitions are not equivalent in this case since SV is more general, and SO can be obtained from it, but not vice versa. For example, there can exist a value $x \in X_Q$ such that

its neighborhood contains exactly one object $u \in A$ and no elements from A^c, and such that u is not in the SO lower approximation of A. The latter holds in particular if there exists some $v \in A^c$ such that $d(f(u,Q), f(v,Q)) < \epsilon$. However, x belongs to the SV lower approximation, and such x cannot be reconstructed from the SO lower approximation.

We will use the SV definition to derive a statistical extension of rough sets to numerical data.

3 A statistical view of Pawlak's rough sets

One widely used assumption in statistics and machine learning (ML) is that data are realizations of a joint random variable. Let objects be outcomes of the joint random variable $\mathcal{U} = (\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ is a random variable corresponding to the condition attributes, while $\mathcal{Y}$ corresponds to the decision attribute. Since we are dealing with classification problems, we know that $\mathcal{Y}$ is always discrete, while $\mathcal{X}$ is discrete if we work with categorical data, or $\mathcal{X}$ takes values from $\mathbb{R}^m$ if we have numerical data. Those random variables are unknown in practice, so using data as their realizations, we explain the relations between $\mathcal{X}$ and $\mathcal{Y}$.

The idea here is to redefine the approximations in terms of random variables instead of data. The SV approximations were defined on the domain w.r.t. neighborhood operators, while here the approximations are defined on the domain w.r.t. a random variable. In terms of statistics these are the "true" approximations dependent on unknown random variables. The SV approximations on data will play the role of estimators of such approximations.

Since $\mathcal{Y}$ is discrete, assume that its domain is the set $\{0, 1, \ldots, K\}$ for some K. Classification tasks in machine learning often refer to calculation of the conditional probabilities of the particular classes. More formally, for class $k \in \{0, 1, \ldots, K\}$ we want to model the expression $P(\mathcal{Y} = k|\mathcal{X} = x)$ as a function of x for all x from the domain space (either a space of categories or $\mathbb{R}^m$). Assume now that the domain X_Q of $\mathcal{X}$ is finite i.e., $\mathcal{X}$ is discrete. If certainty is modeled in a probabilistic environment, we say that an event is certain if its probability is 1 while an event is possible if its probability is greater than 0. We want to know if value $x \in X_Q$ certainly belongs to class k, i.e., if $P(\mathcal{Y} = k|\mathcal{X} = x) = 1$. In practice, we do not have exact knowledge about the conditional distribution of $\mathcal{Y}$ on $\mathcal{X}$, so we need to estimate it. We recall the set of objects $U = \{u_i = (x_i, y_i)|i = 1 \ldots n\}$ which is now a set of realizations of random variable $\mathcal{U}$, known as a sample. The empirical estimation of the above mentioned conditional probability is

$$\hat{P}(\mathcal{Y} = k|\mathcal{X} = x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{\{y_i=k, x_i=x\}}}{\mathbf{1}_{\{x_i=x\}}} = \frac{|\{\hat{y} = k\}| \cap |\{\hat{x} = x\}|}{|\{\hat{x} = x\}|},$$

where $\mathbf{1}_A$ is the indicator function, $|\{\hat{y} = k\}|$ is the number of objects y_i equal to k, while $|\{\hat{x} = x\}|$ is the number of objects x_i equal to x. To estimate the set of values x for which $P(\mathcal{Y} = k|\mathcal{X} = x) = 1$, we use the estimated probability

instead of the true one. We have that:

$$\frac{|\{\hat{y}=k\}|\cap|\{\hat{x}=x\}|}{|\{\hat{x}=x\}|}=1\Leftrightarrow|\{\hat{y}=k\}|\cap|\{\hat{x}=x\}|=|\{\hat{x}=x\}|\wedge|\{\hat{x}=x\}|>0$$

$$\Leftrightarrow\{\hat{x}=x\}\subseteq\{\hat{y}=k\}\wedge|\{\hat{x}=x\}|>0.$$

We obtain

$$\{x\in X_Q;\hat{P}(\mathcal{Y}=k|\mathcal{X}=1)\}=\{x\in X_Q;|\{\hat{x}=x\}|>0\wedge\{\hat{x}=x\}\subseteq\{\hat{y}=k\}\}.$$

The right side of the latter equality is identical to the SV definition of Pawlak's rough sets, where $[x]$ is replaced by $\{\hat{x}=x\}$ while A is replaced with $\{\hat{y}=k\}$. Here, it can be noticed that the SV lower approximation may be seen as an estimation of the unknown lower approximation dependent on random variables. A similar procedure may be used for the upper approximation. This leads to the definition of the lower and upper approximations of the class k with respect to random variable $\mathcal{X}$:

$$\begin{aligned}\underline{\text{apr}}_{\mathcal{X}}^{RV}(\mathcal{Y}=k)&=\{x;P(\mathcal{Y}=k|\mathcal{X}=x)=1\},\\ \overline{\text{apr}}_{\mathcal{X}}^{RV}(\mathcal{Y}=k)&=\{x;P(\mathcal{Y}=k|\mathcal{X}=x)>0\}.\end{aligned}\tag{1}$$

We call this the RV definition of rough sets. Such defined "true" approximations do not require any assumptions on $\mathcal{X}$ ($\mathcal{X}$ being discrete or continuous) as long as the conditional probability is defined. This version of the approximations provides a natural extension of rough sets to numerical data (and all other types of data). In practice, approximation estimates for categorical and numerical data are different since the probability estimation is different in the discrete and the continuous case. We have already seen the estimation of the lower approximation for categorical data. Later on it will be shown how to estimate the approximations in the numerical case. The RV rough set definitions can be taken out of the context of classification and they can be extended to arbitrary events. Let A be an event and $\mathcal{X}$ be a random variable. The lower and upper approximations of A w.r.t. $\mathcal{X}$ are defined as:

$$\underline{\text{apr}}_{\mathcal{X}}^{RV}(A)=\{x;P(A|\mathcal{X}=x)=1\},\quad\overline{\text{apr}}_{\mathcal{X}}^{RV}(A)=\{x;P(A|\mathcal{X}=x)>0\}.$$

However, such general definition will not play an important role for our goal, but it may find some other applications in data analysis.

4 Rough approximations for numerical data

In the previous section we have seen how the approximations may be estimated in practice when we deal with categorical data, and that such estimation coincides with Pawlak's approach. Since the approximations do not depend on the type of data, the question is how to estimate them for numerical data. To make things simpler, we assume that classification is binary, i.e., $K=1$, and we only have two

values for the variable $\mathcal{Y}$, 0 and 1. Assume also that the domain of $\mathcal{X}$ is $X_Q \subseteq \mathbb{R}^m$ i.e., $\mathcal{X}$ is a continuous random variable. By $f_{\mathcal{X}}$ we denote the probability density function (PDF) of $\mathcal{X}$, while by $f_{\mathcal{Y}}(k) = P(\mathcal{Y} = k)$ we denote the PDF of the binary random variable $\mathcal{Y}$. The joint PDF of $\mathcal{Y}$ and $\mathcal{X}$ is denoted as $f_{\mathcal{Y},\mathcal{X}}$. From probability theory it holds that $f_{\mathcal{Y}}(0) + f_{\mathcal{Y}}(1) = 1$, $f_{\mathcal{X}}(x) > 0$ for $x \in X_Q$ and $\int_{X_Q} f_{\mathcal{X}}(x)dx = 1$. We calculate the approximations of class 1. Probability theory tells us that:

$$P(\mathcal{Y} = 1|\mathcal{X} = x) = \frac{f_{\mathcal{Y},\mathcal{X}}(1,x)}{f_{\mathcal{X}}(x)} = 1 - \frac{f_{\mathcal{X}}(x) - f_{\mathcal{Y},\mathcal{X}}(1,x)}{f_{\mathcal{X}}(x)} = 1 - \frac{f_{\mathcal{Y},\mathcal{X}}(0,x)}{f_{\mathcal{X}}(x)}.$$

For the lower approximation we have that

$$P(\mathcal{Y} = 1|\mathcal{X} = x) = 1 \Leftrightarrow 1 - \frac{f_{\mathcal{Y},\mathcal{X}}(0,x)}{f_{\mathcal{X}}(x)} = 1 \Leftrightarrow \frac{f_{\mathcal{Y},\mathcal{X}}(0,x)}{f_{\mathcal{X}}(x)} = 0 \Leftrightarrow f_{\mathcal{Y},\mathcal{X}}(0,x) = 0.$$

The last equality can be divided by $f_{\mathcal{Y}}(0)$ and we get the condition $f_{\mathcal{X}|\mathcal{Y}=0}(x) = 0$. Here $f_{\mathcal{X}|\mathcal{Y}=0}$ stands for the conditional PDF of $\mathcal{X}$ on event $\{\mathcal{Y} = 0\}$. For the upper approximation we have:

$$P(\mathcal{Y} = 1|\mathcal{X} = x) > 0 \Leftrightarrow \frac{f_{\mathcal{Y},\mathcal{X}}(1,x)}{f_{\mathcal{X}}(x)} > 0 \Leftrightarrow f_{\mathcal{Y},\mathcal{X}}(1,x) > 0.$$

The last equality can be divided by $f_{\mathcal{Y}}(1)$ and we get the condition $f_{\mathcal{X}|\mathcal{Y}=1}(x) > 0$.

The conclusion we may derive from the calculations is that x certainly belongs to class 1 if the conditional PDF of $\mathcal{X}$ on $\{\mathcal{Y} = 0\}$ evaluated in x is 0. We have that x possibly belongs to class 1 if the conditional PDF of $\mathcal{X}$ on $\{\mathcal{Y} = 0\}$ evaluated in x is greater than 0. These conditions depend on conditional PDFs which are unknown in practice and have to be estimated. More precisely, we need to estimate the so-called level sets, i.e., areas on which the PDF is smaller or greater than some value [2]. In our case, the thresholds we consider for the PDFs are when they are equal to 0 and greater than 0 (lower and upper approximation).

The estimation of level sets is an emerging field in statistics and ML [2, 3, 20]. Such estimations are essentially different from estimating the PDF itself since we are searching for good estimators for a particular area of the PDF, not for the whole PDF.

Below we present a naive approach of estimating level sets using the estimation of the PDF. Density estimation is a well studied area of statistics [18, 19, 23]. The main methods are histogram density estimation, kernel density estimation (KDE) and nearest neighbour density estimation. Histograms are known for performing badly in high dimensions [18], while the nearest neighbour methods do not assume that there are areas where the PDF is equal to 0 [14]. For these reasons, KDE appears the most appropriate choice to calculate level sets. We refer the reader to [19] for an overview of density estimation methods.

4.1 Rough sets and KDE

A kernel $K : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is a positive and symmetric mapping for which it holds that $\forall t \in \mathbb{R}^M, \int_{\mathbb{R}^m} K(t,s)ds = 1$ [24]. It may be seen as a measure of

similarity between points from $\mathbb{R}^m$. The kernel density estimator is defined as:

$$\hat{f}^K(t) = \frac{1}{n}\sum_{i=1}^{n} K(t, t_i),$$

where $\{t_1, t_2, \ldots, t_n\}$ is a given sample from the unknown PDF f. The motivation behind this definition is that if x has more points in its proximity, then value $\hat{f}^K(x)$ will be larger, which indicates an area of higher density.

Similarity measures are usually based on distances between points since, intuitively, the closer points are, the more similar they are to each other. Therefore, we use kernels based on Euclidean distance, called radial kernels [12]:

$$K(x, y) = \frac{1}{h}k\left(\frac{\|x - y\|}{h}\right).$$

The notation $\|\cdot\|$ stands for the standard norm on $\mathbb{R}^m$, h is a positive real parameter called bandwidth while k is a univariate positive function. Using radial kernels, the PDF estimator becomes:

$$\hat{f}^{k,h}(x) = \frac{1}{nh^m}\sum_{i=1}^{n} k\left(\frac{\|x - x_i\|}{h}\right). \tag{2}$$

From before we have that the lower approximation can be formulated as:

$$\underline{\text{apr}}^{RV}_{\mathcal{X}}(\mathcal{Y} = 1) = \{x; f_{\mathcal{X}|\mathcal{Y}=0}(x) = 0\}.$$

Therefore, using (2) we get the estimator of the lower approximation:

$$\underline{\text{apr}}^{RV}_{\hat{\mathcal{X}}}(\mathcal{Y} = 1) = \{x; \hat{f}^{k,h}_{\mathcal{X}|\mathcal{Y}=0}(x) = 0\}.$$

Although it is not possible that $f_{\mathcal{X}|\mathcal{Y}=0}(x) = 0$ and $f_{\mathcal{X}|\mathcal{Y}=1}(x) = 0$ at the same time, it may happen that $\hat{f}^{k,h}_{\mathcal{X}|\mathcal{Y}=0}(x) = 0$ and $\hat{f}^{k,h}_{\mathcal{X}|\mathcal{Y}=1}(x) = 0$ for some x. Such values we will denote as inconclusive and we will exclude them from the approximations, as before. Following this, we redefine the estimation of the lower approximation:

$$\underline{\text{apr}}^{RV}_{\hat{\mathcal{X}}}(\mathcal{Y} = 1) = \{x; \hat{f}^{k,h}_{\mathcal{X}|\mathcal{Y}=0}(x) = 0 \wedge \hat{f}^{k,h}_{\mathcal{X}|\mathcal{Y}=1}(x) > 0\}. \tag{3}$$

Henceforth we will focus on the lower approximation. A very similar procedure can be used to estimate the upper approximation.

We have to decide which area satisfies the condition from (3). To estimate $f_{\mathcal{X}|\mathcal{Y}=0}$ we use objects from class 0 and to estimate $f_{\mathcal{X}|\mathcal{Y}=1}$ we use objects from class 1. Recall $U = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ as the set of objects or the sample. Set U is split into two subsets; objects which belong to class 0, and objects which belong to class 1. We denote those sets $U^0 = \{(x^0_1, 0), (x^0_2, 0), \ldots, (x^0_{n_0}, 0)\}$ and $\{U^1 = (x^1_1, 1), (x^1_2, 1), \ldots, (x^1_{n_1}, 1)\}$. To estimate the conditional PDFs $f_{\mathcal{X}|\mathcal{Y}=0}$

and $f_{\mathcal{X}|\mathcal{Y}=1}$ we use the objects from U^0 and U^1 respectively. To estimate the level set $f_{\mathcal{X}|\mathcal{Y}=0}(x) = 0$ we have to find values of x for which $\hat{f}^{k,h}_{\mathcal{X}|\mathcal{Y}=0}(x) = 0$ and to estimate $f_{\mathcal{X}|\mathcal{Y}=1}(x) > 0$ we are searching for x where $\hat{f}^{k,h}_{\mathcal{X}|\mathcal{Y}=1}(x) > 0$. It follows that:

$$\frac{1}{nh}\sum_{i=1}^{n_0} k\left(\frac{\|x - x_i^0\|}{h}\right) = 0 \Leftrightarrow \forall i \in \{1, \dots n_0\};\, k\left(\frac{\|x - x_i^0\|}{h}\right) = 0.$$

$$\frac{1}{nh}\sum_{i=1}^{n_1} k\left(\frac{\|x - x_i^1\|}{h}\right) > 0 \Leftrightarrow \exists i \in \{1, \dots n_1\};\, k\left(\frac{\|x - x_i^0\|}{h}\right) > 0.$$

The derivation up to now is general and holds for all functions k and bandwidths h. The question is, which kernel best suits the last condition. The most used kernel in practice is the Gaussian kernel which is also radial: $k(x) = \frac{1}{\sqrt{(2\pi)^m}} e^{-\frac{1}{2}x^2}$. Its main drawback is that it is nowhere equal to 0. It is used under the assumption that there are no impossible or certain events which is not the case here. Therefore, a better choice would be a kernel with different assumptions. In particular, we require a kernel for which k is bigger than 0 on a bounded set i.e., a kernel with bounded support.

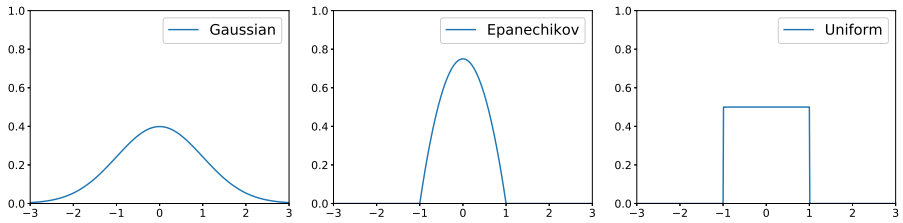

Fig. 1. Kernel examples in univariate case

The theory developed in [13] states that the smallest estimation error under certain conditions is achieved for the Epanechikov kernel. The Epanechikov kernel is radial with

$$k(x) = \max\left\{0, \frac{m+2}{2c_m}(1 - x^2)\right\},$$

where c_m is the volume of the m-dimensional unit ball. According to the definition, its support is the unit hypersphere, which implies that it is bounded. Another kernel with bounded support is the spherical uniform kernel, i.e., the constant radial kernel for which

$$k(x) = \begin{cases} \frac{1}{c_m} & \text{if } x \in (0, 1) \\ 0 & \text{otherwise.} \end{cases}$$

Let h_e and h_u be the bandwidths corresponding to the Epanechikov kernel and spherical uniform kernel, respectively. For the Epanechikov kernel, we have that:

$$k\left(\frac{\|x-x_i^0\|}{h_e}\right)=0 \Leftrightarrow \frac{m+2}{2c_m}\left(1-\frac{\|x-x_i^0\|^2}{h_e^2}\right)\leq 0 \Leftrightarrow \|x-x_i^0\|\geq h_e,$$

$$k\left(\frac{\|x-x_i^1\|}{h_e}\right)>0 \Leftrightarrow \frac{m+2}{2c_m}\left(1-\frac{\|x-x_i^1\|^2}{h_e^2}\right)>0 \Leftrightarrow \|x-x_i^1\|< h_e,$$

while for the spherical uniform kernel it holds that:

$$k\left(\frac{\|x-x_i^0\|}{h_u}\right)=0 \Leftrightarrow \|x-x_i^0\|\geq h_u, \quad k\left(\frac{\|x-x_i^1\|}{h_u}\right)>0 \Leftrightarrow \|x-x_i^1\|< h_u,$$

In both cases, value x certainly belongs to class 1 if in the neighborhood there are no objects from the opposite class and there are some objects from the same class. Hence, by using kernels with bounded support, we obtain simple conditions for estimating the lower approximations.

4.2 Relationship to neighborhood based rough sets

We summarize the results obtained so far: we defined the lower approximation of class $\{\mathcal{Y}=1\}$ as : $\underline{\text{apr}}_{\mathcal{X}}^{RV}(\mathcal{Y}=1)=\{x; f_{\mathcal{X}|\mathcal{Y}=0}(x)=0\}$ for continuous random variable $\mathcal{X}$. We estimated the approximation by estimating the PDF from the expression using kernel density estimators as:

$$\underline{\text{apr}}_{\hat{\mathcal{X}}}^{RV}(\mathcal{Y}=1)=\{x; \hat{f}^K_{\mathcal{X}|\mathcal{Y}=0}(x)=0 \wedge \hat{f}^K_{\mathcal{X}|\mathcal{Y}=1}(x)>0\}.$$

We have shown that the estimators for certain radial kernels with bounded support lead to the expression:

$$\underline{\text{apr}}_{\hat{\mathcal{X}}}^{RV}(\mathcal{Y}=1)=\{x; \forall i : \|x-x_i^0\|\geq h \wedge \exists i : \|x-x_i^1\|< h\},$$

for some h. Let us write the neighborhood definition replacing ϵ with h: $n_h(x)=\{x_i\in U; d(x,x_i)<h\}$, where d is the Euclidean distance. Condition $\exists i : \|x-x_i^1\|<h$ means that there is at least one object from U^1 in $n_h(x)$, i.e., $n_h(x)\neq\emptyset$, while $\forall i : \|x-x_i^0\|\geq h$ means that there are no objects from U^0 in $n_h(x)$, i.e., $n_h(x)\subseteq U^1$. It follows that the approximation estimator can be written as:

$$\underline{\text{apr}}_{\hat{\mathcal{X}}}^{RV}(\mathcal{Y}=1)=\{x; n_h(x)\neq\emptyset \wedge n_h(x)\subseteq U^1\}.$$

The latter expression is exactly the SV (set of values) definition of the neighborhood based rough sets. We can conclude that the estimators of the RV approximations coincide with the SV definition of the neighborhood based rough sets. The advantage of this representation of the neighborhood based rough sets is that we have proper mathematical tools to calculate the neighborhood size in order to get better results. We are now able to use statistical methods to obtain a proper bandwidth which plays the role of the neighborhood size.

In the following subsection, we will outline a procedure to select the bandwidths in theory, that is: we provide some insights on how the bandwidths can be calculated independently from data, using only the chosen kernel and the original PDF.

4.3 Bandwidth selection - an example

This subsection relies on the work presented in [19]. Using the KDE theory, we are able to construct the proper bandwidths for different kernels in order to obtain the best possible estimator of PDFs (or at least close to the best). The bandwidths are chosen to minimize the error of the PDF estimation. A widely used error function is Mean Integrated Square Error (MISE):

$$MISE(\hat{f}^{k,h}) = \int_{X_Q} E((\hat{f}^{k,h}(x) - f(x))^2)dx$$

where E stands for the expected value. When n is significantly larger than the number of attributes m, the MISE of radial kernels can be approximated as:

$$MISE(\hat{f}^{k,h}) \approx C_1 h^4 + \frac{C_2}{nh^m}.$$

The latter expression is also called AMISE or Asymptotic MISE. By minimizing the expression above, we get the optimal bandwidth:

$$h^{opt} = C_3 n^{-\frac{1}{m+4}}.$$

Constants C_1, C_2 and C_3 are dependent on the kernel and on the actual probability density function f. Assuming that our data are normally distributed (or something close to normal with bounded support), we are able to calculate the optimal bandwidths. Under normality assumption, the optimal bandwidths for the Epanechikov and spherical uniform kernels are:

$$h_e^{opt} = [8(d+4)c_m^{-1}(2\sqrt{\pi})^d n^{-1}]^{\frac{1}{m+4}}, \quad h_u^{opt} = [4(d+2)c_m^{-1}(2\sqrt{\pi})^d n^{-1}]^{\frac{1}{m+4}}.$$

From the AMISE expression, we may see that the rate of convergence is not dependent on constant C_3. Therefore, in order to avoid the assumptions and to achieve better results one can try to tune constant C_3 using data. Under h^{opt} for some kernel we also ensure that:

$$\lim_{n\to\infty} MISE(\hat{f}^{k,h^{opt}}) = 0.$$

That ensures that for a sufficiently large sample size n, the inconclusive areas will become negligible. That is also intuitive since with more data we acquire more knowledge which leaves less space for uncertainty.

5 Discussion

We have presented a new way to calculate the neighborhood size in neighborhood based rough sets. A question arises: does it provide satisfactory results in practice?

It is well known that rough sets are widely used in attribute selection [4, 10]. The attribute selection in rough sets focuses on preservation of certain knowledge; we delete attributes as long as the lower approximations of all classes remain unchanged.

We have run a series of experiments applying the attribute selection using neighborhood based rough sets together with the calculated bandwidths. Unfortunately, the results were not satisfactory. First, we simulated data with normal distribution to fulfill the assumption from the previous subsection. We have noticed that for lower dimensions, both h_e^{opt}-neighborhood and h_u^{opt}-neighborhood are too wide, meaning that they cover a large amount of data. Consequently, the lower approximations obtained with them consist of a low percentage of data which is unrealistic. With higher dimensions, we observed the opposite problem; the neighborhoods are too narrow which leads to the lower approximation containing almost all data, which is also unrealistic. We can conclude that the naive approach of estimating PDF and searching for the optimal bandwidth is not the best idea. The reason for the failure, even under the normality assumption, may lie in the fact that the optimal bandwidths are mainly useful in the following cases.

- The number of objects in the sample is significantly larger than the number of attributes since the bandwidth optimality is asymptotic.
- The MISE error is calculated using l_2 norm (the integral of the squared difference). Our interest is to get the optimal bandwidth for the level set where PDF is equal to 0. The l_2 convergence does not guarantee that the estimator also uniformly converges to the actual PDF [17]. Thus, we may have that h^{opt} is suitable for the higher density regions where the PDF is significantly larger than 0 and that it may have poor performance for the regions where the PDF is close to 0.

We have also applied the procedure on real data for which the normality assumption does not hold. As soon as the assumption is not fulfilled, the results are getting worse. For example, we considered binary classification in mammographic data from UCI [1] for which $n = 830$ and $m = 5$. In all cases, the lower approximations contained less than 7 % of data, meaning that only 7 % of data can be certainly classified. Keeping in mind that the classification accuracy we obtained with SVM on this dataset is around 85%, 7 % of certainty is unrealistic.

To overcome the limitations of the theoretical bandwidth selection, we identify the following options for future integration of rough sets, KDE and statistics in general.

- **Data driven estimation.** The calculation of bandwidths may be data driven. There is also a statistical theory on how to calculate bandwidths based on data (again [19]). Data driven bandwidths will help us to overcome any a priori assumptions on the distribution of data.
- **Robust approaches.** Having 0 probability regions is a strong assumption which usually does not coincide with reality. Mostly, numerical data exhibit

rare events, which may occur in the training data and/or during the prediction process. Having the assumption that data lie in a bounded region may be misleading in many cases and it can produce bad results. The 0 probability regions can be eliminated by applying robust approaches similar to Variable Precision Rough Sets (VPRS).

- **Direct level set estimation.** The bandwidth calculation needs to be more adjusted to the problem of the level set estimation, rather than to the PDF estimation. After we identify the regions of interest, we have to set up the optimization problem to get the best possible (or close to the best) bandwidth for that particular case.
- **Different estimators than KDE.** We can try to use other estimators for level sets, besides KDE. The nearest neighbor based estimator can give interesting results [14].
- **Integration with SVM.** Do we have to use densities to estimate the approximations defined in (1)? We showed that the estimation of the RV approximations (1) boils down to the estimation of level sets. We may explore the relation between SVM and level set estimation as has been done in [11, 16, 22]. On the other hand, there is a direct correspondence between principles of rough sets and SVM. The applications of rough sets in binary classification divide the domain into three sets, two certain regions for each class and one boundary region. SVM is doing something similar where it trains two margins which divide the space similarly as the rough sets: one boundary region and two regions for two classes. Thus, using the similarities between rough sets and SVM, we can try to integrate them in order to achieve better results.

6 Conclusion

We presented a new view on the definition of rough sets for the case when data are not necessarily categorical. From the statistical point of view, the calculation of rough set approximations is basically the estimation of the unknown RV (random value) approximations dependent on random variables that generate data. Such estimation under certain conditions (i.e., using radial kernels with bounded support) is equivalent to the definition of neighborhood based rough sets. We also showed a simple way how to calculate the neighborhood size using statistics. Moreover, we discussed several options for future research on the integration of rough sets and statistics. Of course, for each of the proposals it should be studied if it can be tailored to the main applications of rough sets: rule induction and attribute selection.

Acknowledgements

This work was supported by the Odysseus program of the Research Foundation-Flanders.

References

1. Asuncion, A., Newman, D.: Uci machine learning repository (2007)
2. Cadre, B.: Kernel estimation of density level sets. Journal of multivariate analysis **97**(4), 999–1023 (2006)
3. Chen, Y.C., Genovese, C.R., Wasserman, L.: Density level sets: Asymptotics, inference, and visualization. Journal of the American Statistical Association **112**(520), 1684–1696 (2017)
4. Choubey, S.K., Deogun, J.S., Raghavan, V.V., Sever, H.: A comparison of feature selection algorithms in the context of rough classifiers. In: Proceedings of IEEE 5th International Fuzzy Systems. vol. 2, pp. 1122–1128. IEEE (1996)
5. Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. International Journal of General System **17**(2-3), 191–209 (1990)
6. Greco, S., Matarazzo, B., Słowiński, R.: Rough membership and bayesian confirmation measures for parameterized rough sets. In: International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. pp. 314–324. Springer (2005)
7. Grzymala-Busse, J.W., Stefanowski, J.: Three discretization methods for rule induction. International Journal of Intelligent Systems **16**(1), 29–38 (2001)
8. Grzymala-Busse, J.W., Werbrouck, P.: On the best search method in the lem1 and lem2 algorithms. In: Incomplete Information: Rough Set Analysis, pp. 75–91. Springer (1998)
9. Hu, Q., Yu, D., Liu, J., Wu, C.: Neighborhood rough set based heterogeneous feature subset selection. Information sciences **178**(18), 3577–3594 (2008)
10. Jensen, R.: Rough set-based feature selection: A review. In: Rough computing: theories, technologies and applications, pp. 70–107. IGI Global (2008)
11. Kloft, M., Nakajima, S., Brefeld, U.: Feature selection for density level-sets. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 692–704. Springer (2009)
12. Kulczycki, P.: Kernel estimators in industrial applications. In: Soft Computing Applications in Industry, pp. 69–91. Springer (2008)
13. Muller, H.G., et al.: Smooth optimum kernel estimators of densities, regression curves and modes. The Annals of Statistics **12**(2), 766–774 (1984)
14. Orava, J.: K-nearest neighbour kernel density estimation, the choice of optimal k. Tatra Mountains Mathematical Publications **50**(1), 39–50 (2011)
15. Pawlak, Z.: Rough sets. International journal of computer & information sciences **11**(5), 341–356 (1982)
16. Rakotomamonjy, A., Davy, M.: One-class svm regularization path and comparison with alpha seeding. In: ESANN. pp. 271–276. Citeseer (2007)
17. Rudin, W.: Real and complex analysis. Tata McGraw-hill education (2006)
18. Scott, D.W.: Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons (2015)
19. Silverman, B.W.: Density estimation for statistics and data analysis. Routledge (2018)
20. Singh, A., Scott, C., Nowak, R., et al.: Adaptive hausdorff estimation of density level sets. The Annals of Statistics **37**(5B), 2760–2782 (2009)
21. Slowinski, R., Vanderpooten, D.: A generalized definition of rough approximations based on similarity. IEEE Transactions on knowledge and Data Engineering **12**(2), 331–336 (2000)

22. Steinwart, I., Hush, D., Scovel, C.: Density level detection is classification. In: Advances in neural information processing systems. pp. 1337–1344 (2005)
23. Wand, M.P., Jones, M.C.: Kernel smoothing. Chapman and Hall/CRC (1994)
24. Węglarczyk, S.: Kernel density estimation and its application. In: ITM Web of Conferences. vol. 23, p. 00037. EDP Sciences (2018)
25. Yao, Y.: Decision-theoretic rough set models. In: International conference on rough sets and knowledge technology. pp. 1–12. Springer (2007)
26. Yao, Y., Yao, B.: Covering based rough set approximations. Information Sciences **200**, 91–107 (2012)
27. Ziarko, W.: Variable precision rough set model. Journal of computer and system sciences **46**(1), 39–59 (1993)
28. Ziarko, W.: Probabilistic rough sets. In: International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. pp. 283–293. Springer (2005)