Generalized Parameterized Approximations

Jerzy W. Grzymala-Busse

[1] Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS 66045, USA
[2] Institute of Computer Science, Polish Academy of Sciences,
01–237 Warsaw, Poland
jerzy@ku.edu

Abstract. We study generalized parameterized approximations, defined using both rough set theory and probability theory. The main objective is to study, for a given subset of the universe U, all such parameterized approximations, i.e., for all parameter values. For an approximation space (U, R), where R is an equivalence relation, there is only one type of such parameterized approximations. For an approximation space (U, R), where R is an arbitrary binary relation, three types of parameterized approximations are introduced in this paper: singleton, subset and concept. We show that the number of parameterized approximations of given type is not greater than the cardinality of U. Additionally, we show that singleton parameterized approximations are not useful for data mining, since such approximations, in general, are not even locally definable.

1 Introduction

The entire rough set theory is based on ideas of the lower and upper approximations. Complete data sets, presented as decision tables, are well described by an indiscernibility relation, yet another fundamental idea of rough set theory. The indiscernibility relation is an equivalence relation. Standard lower and upper approximations were extended, using probability theory, to parameterized approximations. Such approximations were studied, among others, in [1–7]. The parameter, called a threshold and associated with the parameterized approximation, may be interpreted as a probability. The threshold is, in general, a real number.

So far parameterized approximations were usually defined as lower and upper approximations. As it was observed in [8], the only difference between so called lower and upper parameterized approximations is in the choice of the value of the threshold.

Due to the fact that we explore the set of all parameterized approximations of a given type, the distinction between lower and upper approximations is blurred. Therefore, we will define only one kind of parameterized approximations for an approximation space (U, R), where U is a finite set and R is an equivalence relation on U.

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 136–145, 2011.

This paper, for a given decision table and a subset of the universe explores the set of all parameterized approximations. It is shown that the number of all parameterized approximations is finite and quite limited.

Additionally, this paper generalizes the usual three types of approximations: singleton, subset and concept, used for approximation spaces (U, R), where R is an arbitrary binary relation. Similarly as for singleton standard approximations, a singleton parameterized approximation of a subset X of the universe U is, in general, not definable. There are two types of definability, local and global. If the set X is globally definable, it is locally definable, the converse is, in general, not true. Sets that is the singleton parameterized approximation of X are, in general, not even locally definable. The idea of parameterized approximations is applied to incomplete data sets. It is well known [9, 10] that incomplete data sets, i.e., data sets with missing attribute values, are described by characteristic relations, which are reflexive but, in general, neither symmetric nor transitive.

2 Equivalence Relations

In this section we will discuss data sets without missing attribute values, i.e., complete. Complete data sets are describable by equivalence relations. Then we will discuss all parameterized partitions defined over a space approximation (U, R), where U is a finite set and R is an equivalence relation.

2.1 Complete Data

Many real-life data sets have conflicting cases, characterized by identical values for all attributes but belonging to different concepts (classes). Data sets with conflicting cases are called inconsistent. An example of the inconsistent data set is presented in Table 1. The data set presented in Table 1 is inconsistent since it contains conflicting cases: the cases 2 and 4 are in conflict with the case 3 and the case 6 is in conflict with case 8.

In Table 1, the set A of all attributes consists of three variables *Temperature*, *Headache* and *Cough*. A *concept* is a set of all cases with the same decision value. There are two concepts in Table 1, the first one contains cases 1, 2, 4 and 6 and is characterized by the decision value *no* of decision *Flu*. The other concept contains cases 3, 5, 7 and 8 and is characterized be the decision value *yes*.

The fact that an attribute a has the value v for the case x will be denoted by $a(x) = v$. The set of all cases will be denoted by U. In Table 1, $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

For an attribute-value pair $(a,\ v) = t$, a *block* of t, denoted by $[t]$, is a set of all cases from U such that for attribute a have value v. An *indiscernibility relation* R on U is defined for all $x, y \in U$ by

$$xRy \text{ if and only if } a(x) = a(y) \text{ for all } a \in A.$$

Table 1. An inconsistent data set

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	normal	no	yes	no
2	normal	no	no	no
3	normal	no	no	yes
4	normal	no	no	no
5	high	yes	no	yes
6	high	yes	yes	no
7	high	no	yes	yes
8	high	yes	yes	yes

Equivalence classes of R are called *elementary sets* of R. An equivalence class of R containing x is denoted $[x]$. Any finite union of elementary sets is called a *definable set* [11]. Let X be a concept. In general, X is not a definable set. However, set X may be approximated by two definable sets, the first one is called a *lower approximation* of X, denoted by $\underline{appr}(X)$ and defined as follows

$$\{[x] \mid x \in U,\ [x] \subseteq X\},$$

The second set is called an *upper approximation* of X, denoted by $\overline{appr}(X)$ and defined as follows

$$\cup\ \{[x] \mid x \in U,\ [x] \cap X \neq \emptyset\}.$$

For example, for the concept $[(Flu,\ no)] = \{1,\ 2,\ 4,\ 6\}$,

$\underline{appr}(\{1, 2, 4, 6\}) = \{1\}$,

and

$\overline{appr}(\{1, 2, 4, 6\}) = \{1,\ 2,\ 3,\ 4,\ 6,\ 8\}$.

2.2 Parameterized Approximations

Let (U, R) be an approximation space, where R is an equivalence relation on U. A *p*arameterized approximation of the set X with the threshold α, $0 < \alpha \leq 1$, is denoted by $appr_{\alpha}(X)$ and defined as follows

$$\cup\{[x] \mid x \in U,\ Pr(X|[x]) \geq \alpha\},$$

where $[x]$ is an elementary set of R and $Pr(X \mid [x]) = \frac{|X \cap [x]|}{|[x]|}$ is the conditional probability of X given $[x]$ and $|X|$ denotes the cardinality of the set X.

Table 2. Conditional probabilities

$[x]$	{1}	{2, 3, 4}	{5}	{6, 8}	{7}
$Pr(\{1, 2, 4, 6\} \mid [x])$	1	0.667	0	0.5	0

Obviously, the equivalence relation R uniquely defines a partition on U defined as the family of all elementary sets of R. Such a partition will be denoted by R^*. For Table 1, $R^* = \{\{1\}, \{2, 3, 4\}, \{5\}, \{6, 8\}, \{7\}\}$.

For the set X and all equivalence classes from R^* we may compute the set of all distinct conditional probabilities $Pr(X \mid [x])$ and then sort these numbers in the ascending order. The number of all nonempty distinct parameterized approximations of X is equal to the number of distinct and positive conditional probabilities $Pr(X \mid [x])$.

Table 2 shows conditional probabilities for all members of R^*. In Table 2 there are three positive conditional probabilities: 0.5, 0.667 and 1. Therefore there are only three parameterized approximations:

$appr_{0.5}(\{1, 2, 4, 6\}) = \{1, 2, 3, 4, 6, 8\}$,

$appr_{0.667}(\{1, 2, 4, 6\}) = \{1, 2, 3, 4\}$,

and

$appr_{1}(\{1, 2, 4, 6\}) = \{1\}$.

Obviously, for the concept X, the parameterized approximation of X computed for the threshold equal to the smallest positive conditional probability $Pr(X \mid [x])$ is equal to the upper approximation of X. Additionally, the parameterized approximation of X computed for the threshold equal to 1 is equal to the lower approximation of X.

3 Arbitrary Binary Relations

In this section first we will study approximations defined on the approximations space $A = (U, R)$ where U is a finite nonempty set and R is an arbitrary binary relation. Then we will extend corresponding definitions to generalized parameterized approximations.

3.1 Nonparameterized Approximations

First we will quote some definitions from [12]. Let x be a member of U. The R-*successor* set of x, denoted by $R_s(x)$, is defined as follows

$$R_s(x) = \{y \mid xRy\}.$$

The R-*predecessor* set of x, denoted by $R_p(x)$, is defined as follows

$$R_p(x) = \{y \mid yRx\}.$$

For the rest of the paper we will discuss only R-successor sets and corresponding approximations.

Let X be a subset of U. The R-*singleton lower approximation* of X, denoted by $\underline{appr}^{singleton}(X)$, is defined as follows

$$\{x \mid x \in U, R_s(x) \subseteq X\}.$$

The singleton lower approximations were studied in many papers, see, e.g., [9, 10, 13–20].

The R-*singleton upper approximation* of X, denoted by $\overline{appr}^{singleton}(X)$, is defined as follows

$$\{x \mid x \in U, R_s(x) \cap X \neq \emptyset\}.$$

The singleton upper approximations, like singleton lower approximations, were also studied in many papers, e.g., [9, 10, 13, 14, 17–20].

The R-*subset lower approximation* of X, denoted by $\underline{appr}^{subset}(X)$, is defined as follows

$$\cup \{R_s(x) \mid x \in U, R_s(x) \subseteq X\}.$$

The subset lower approximations were introduced in [9, 10].

The R-*subset upper approximation* of X, denoted by $\overline{appr}^{subset}(X)$, is defined as follows

$$\cup \{R_s(x) \mid x \in U, R_s(x) \cap X \neq \emptyset\}.$$

The subset upper approximations were introduced in [9, 10].

The R-*concept lower approximation* of X, denoted by $\underline{appr}^{concept}(X)$, is defined as follows

$$\cup \{R_s(x) \mid x \in X, R_s(x) \subseteq X\}.$$

The concept lower approximations were introduced in [9, 10].

The R-*concept successor upper approximation* of X, denoted by $\overline{appr}^{concept}(X)$, is defined as follows

$$\cup \{R_s(x) \mid x \in X, R_s(x) \cap X \neq \emptyset\}$$

The concept upper approximations were studied in [9, 10, 16].

3.2 Parameterized Approximations

By analogy with standard approximations defined for arbitrary binary relations, we will introduce three kinds of parameterized approximations for such relations: singleton, subset and concept.

A *singleton parameterized approximation* of X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr_{\alpha}^{singleton}(X)$, is defined as follows

$$\{x \mid x \in U,\ Pr(X \mid R_s(x)) \geq \alpha\},$$

where $Pr(X|R_s(x)) = \frac{|X \cap R_s(x)|}{|R_s(x)|}$ is the conditional probability of X given $R_s(x)$.

A *subset parameterized approximation* of the set X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr_{\alpha}^{subset}(X)$, is defined as follows

$$\cup\{R_s(x) \mid x \in U,\ Pr(X \mid R_s(x)) \geq \alpha\},$$

where $Pr(X|[x]) = \frac{|X \cap R_s(x)|}{|R_s(x)|}$ is the conditional probability of X given $R_s(x)$.

A *concept parameterized approximation* of the set X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr_{\alpha}^{concept}(X)$, is defined as follows

$$\cup\{R_s(x) \mid x \in X,\ Pr(X \mid R_s(x)) \geq \alpha\}.$$

The number of different R-successor sets $R_s(x)$, where $x \in U$, is obviously not greater than n, where n is the cardinality of U. Therefore, for a given concept X, there is at most n different conditional probabilities $Pr(X \mid R_s(x))$. Thus, the number of different parameterized approximations of given type (singleton, subset or concept) is also not greater than n.

Obviously, for the concept X, the parameterized approximation of a given type (singleton, subset or concept) of X computed for the threshold equal to the smallest positive conditional probability $Pr(X \mid [x])$ is equal to the standard upper approximation of X of the same type. Additionally, the parameterized approximation of a given type of X computed for the threshold equal to 1 is equal to the standard lower approximation of X of the same type.

3.3 Incomplete Data Sets

It is well-known that any incomplete data set is described by a *characteristic relation* R, a generalization of the indiscernibility relation. The characteristic relation is reflexive but, in general, is neither symmetric nor transitive. For incomplete data sets R-definable sets are called *characteristic sets*, a generalization of elementary sets.

We distinguish between two types of missing attribute values: *lost* (e.g., the value was erased) and *"do not care" conditions* (such a value may be any value of the attribute), see [9, 10].

An example of incomplete data set is presented in Table 3.

For incomplete decision tables the definition of a block of an attribute-value pair must be modified in the following way:

- If for an attribute a there exists a case x such that $a(x) =?$, i.e., the corresponding value is lost, then the case x should not be included in any blocks $[(a, v)]$ for all values v of attribute a,

Table 3. An incomplete data set

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	normal	no	*	no
2	?	no	no	no
3	normal	*	no	yes
4	normal	no	?	no
5	high	yes	*	yes
6	high	yes	yes	no
7	high	?	yes	yes
8	high	yes	yes	yes

- If for an attribute a there exists a case x such that the corresponding value is a "do not care" condition, i.e., $a(x) = *$, then the case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

For a case $x \in U$ the *characteristic set* $K_B(x)$ is defined as the intersection of the sets $K(x, a)$, for all $a \in B$, where the set $K(x, a)$ is defined in the following way:

- If $a(x)$ is specified, then $K(x, a)$ is the block $[(a, a(x))]$ of attribute a and its value $a(x)$,
- If $a(x) =?$ or $a(x) = *$ then the set $K(x, a) = U$.

The characteristic set $K_B(x)$ may be interpreted as the set of cases that are indistinguishable from x using all attributes from B and using a given interpretation of missing attribute values.

For the data set from Table 3, the set of blocks of attribute-value pairs is
$[(Temperature, normal)] = \{1, 3, 4\}$,
$[(Temperature, high)] = \{5, 6, 7, 8\}$,
$[(Headache, no)] = \{1, 2, 3, 4\}$,
$[(Headache, yes)] = \{3, 5, 6, 8\}$,
$[(Cough, no)] = \{1, 2, 3, 5\}$,
$[(Cough, yes)] = \{1, 5, 6, 7, 8\}$.

The corresponding characteristic sets are

$K_A(1) = K_A(4) = \{1, 3, 4\}$,
$K_A(2) = \{1, 2, 3\}$,
$K_A(3) = \{1, 3\}$,
$K_A(5) = K_A(6) = K_A(8) = \{5, 6, 8\}$,
$K_A(7) = \{5, 6, 7, 8\}$.

Table 4. Conditional probabilities

$R_s(x)$	{1, 3, 4}	{1, 2, 3}	{1, 3}	{5, 6, 8}	{5, 6, 7, 8}
$Pr(\{1,2,4,6\} \mid R_s(x))$	0.667	0.667	0.5	0.333	0.25

Conditional probabilities of the concept {1, 2, 4, 6} given a characteristic set $K_A(x)$ are presented in Table 4.

For Table 3, all parameterized approximations (singleton, subset and concept) are

$appr_{0.25}^{singleton}(\{1,2,4,6\}) = U,$

$appr_{0.333}^{singleton}(\{1,2,4,6\}) = \{1,2,3,4,5,6,8\},$

$appr_{0.5}^{singleton}(\{1,2,4,6\}) = \{1,2,3,4\},$

$appr_{0.667}^{singleton}(\{1,2,4,6\}) = \{1,2,4\},$

$appr_{1}^{singleton}(\{1,2,4,6\}) = \emptyset,$

$appr_{0.25}^{subset}(\{1,2,4,6\}) = U,$

$appr_{0.333}^{subset}(\{1,2,4,6\}) = \{1,2,3,4,5,6,8\},$

$appr_{0.5}^{subset}(\{1,2,4,6\}) = \{1,2,3,4\},$

$appr_{0.667}^{subset}(\{1,2,4,6\}) = \{1,2,3,4\},$

$appr_{1}^{subset}(\{1,2,4,6\}) = \emptyset,$

$appr_{0.25}^{concept}(\{1,2,4,6\}) = \{1,2,3,4,5,6,8\},$

$appr_{0.333}^{concept}(\{1,2,4,6\}) = \{1,2,3,4,5,6,8\},$

$appr_{0.5}^{concept}(\{1,2,4,6\}) = \{1,2,3,4\},$

$appr_{0.667}^{concept}(\{1,2,4,6\}) = \{1,2,3,4\},$

$appr_{1}^{concept}(\{1,2,4,6\}) = \emptyset.$

3.4 Definability

Definability for completely specified decision tables should be modified to fit into incomplete decision tables. For incomplete decision tables, a union of some intersections of attribute-value pair blocks, where such attributes are members of B and are distinct, will be called B-*locally definable* sets. A union of characteristic sets $K_B(x)$, where $x \in X \subseteq U$ will be called a B-*globally definable* set. Any set X that is B-globally definable is B-locally definable, the converse is not true. For example, the set $\{1\}$ is A-locally definable since $\{1\} = [(Temperature, normal)] \cap [(Cough, yes)]$. However, the set $\{1\}$ is not A-globally definable. On the other hand, the set {1, 2, 4} $= appr_{0.667}^{singleton}(\{1, 2, 4, 6\})$ is not even locally definable since in all blocks of attribute-value pairs containing the case 4 contain also the case 3 as well. Obviously, if a set is not B-locally definable then it cannot be expressed by rule sets using attributes from B. This is why it is so important to distinguish between B-locally definable sets and those that are not B-locally definable. In general, subset and concept parameterized approximations are globally definable while singleton parameterized approximations are not even locally definable.

4 Conclusions

In this paper we study a set of all parameterized approximations, first for the approximation space (U, R), where U is a nonempty finite set and R is an equivalence relation, and then for the approximation space (U, R), where R is an arbitrary binary relation. For an arbitrary binary relation R standard definitions of singleton, subset and concept approximations are generalized to parameterized approximations. It is shown that the set of such parameterized approximations, even if R is an arbitrary binary relation, is finite and quite limited. Moreover, singleton parameterized approximations of a subset X of the universe U is, in general, not even locally definable, so X is not expressible by a rule set. Therefore, singleton parameterized approximations should not be used for data mining.

References

1. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. International Journal of Man-Machine Studies 29, 81–95 (1988)
2. Tsumoto, S., Tanaka, H.: PRIMEROSE: probabilistic rule induction method based on rough sets and resampling methods. Computational Intelligence 11, 389–405 (1995)
3. Yao, Y.Y.: Decision-theoretic rough set models. In: Yao, J., Lingras, P., Wu, W.-Z., Szczuka, M.S., Cercone, N.J., Ślęzak, D. (eds.) RSKT 2007. LNCS (LNAI), vol. 4481, pp. 1–12. Springer, Heidelberg (2007)
4. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. International Journal of Man-Machine Studies 37, 793–809 (1992)
5. Yao, Y.Y., Wong, S.K.M., Lingras, P.: A decision-theoretic rough set model. In: Proceedings of the 5th International Symposium on Methodologies for Intelligent Systems, pp. 388–395 (1990)

6. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences 46(1), 39–59 (1993)
7. Ziarko, W.: Probabilistic approach to rough sets. International Journal of Approximate Reasoning 49, 272–284 (2008)
8. Grzymala-Busse, J.W., Marepally, S.R., Yao, Y.: An empirical comparison of rule sets induced by LERS and probabilistic rough classification. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 590–599. Springer, Heidelberg (2010)
9. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Workshop Notes, Foundations and New Directions of Data Mining, in Conjunction with the 3-rd International Conference on Data Mining, pp. 56–63 (2003)
10. Grzymala-Busse, J.W.: Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Transactions on Rough Sets 1, 78–95 (2004)
11. Pawlak, Z.: Rough Sets. In: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht (1991)
12. Grzymala-Busse, J.W., Rzasa, W.: Definability and other properties of approximations for generalized indiscernibility relations. Transactions on Rough Sets 11, 14–39 (2010)
13. Kryszkiewicz, M.: Rough set approach to incomplete information systems. In: Proceedings of the Second Annual Joint Conference on Information Sciences, pp. 194–197 (1995)
14. Kryszkiewicz, M.: Rules in incomplete information systems. Information Sciences 113(3-4), 271–292 (1999)
15. Lin, T.Y.: Neighborhood systems and approximation in database and knowledge base systems. In: Proceedings of the ISMIS-1989, the Fourth International Symposium on Methodologies of Intelligent Systems, pp. 75–86 (1989)
16. Lin, T.Y.: Topological and fuzzy rough sets. In: Slowinski, R. (ed.) Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory, pp. 287–304. Kluwer Academic Publishers, Dordrecht (1992)
17. Slowinski, R., Vanderpooten, D.: A generalized definition of rough approximations based on similarity. IEEE Transactions on Knowledge and Data Engineering 12, 331–336 (2000)
18. Stefanowski, J., Tsoukiàs, A.: On the extension of rough sets under incomplete information. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 73–82. Springer, Heidelberg (1999)
19. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Computational Intelligence 17(3), 545–566 (2001)
20. Yao, Y.Y.: Relational interpretations of neighborhood operators and rough set approximation operators. Information Sciences 111, 239–259 (1998)