An approach to finding reduced sets of information features describing discrete objects based on rough sets theory

D. Sitnikov[1], O. Titova[1], O. Romanenko[1] & O. Ryabov[2]
[1]*Kharkov State Academy of Culture, Ukraine*
[2]*National Institute of Advanced Industrial, Science and Technology, Japan*

Abstract

Modern Data Mining methods allow discovering non-trivial dependencies in large information arrays. Since these methods are used for processing and analysis of huge information volumes, reducing the number of features necessary for describing a discrete object is one of the most important problems.

One of the classical problems in intelligent data analysis is the problem of classifying new objects based on some a-priori information. This information might not allow us to exactly classify an object as one belonging to a certain set. In such cases using rough sets theory may be an effective solution as this theory operates with the concept of "indiscernible" elements and ambiguous information.

In this paper we introduce a concept of a local reduct as a reduced set of features allowing us to describe a particular subset of the original set with the same precision as with the help of the full set of features. A method has been suggested which allows finding reduced sets of features adequately describing a rough set without losing necessary information (so-called reducts), and also assessing the importance of each feature. The suggested method is based on the algebraic approach to finding rough set approximations developed by the authors earlier. The main idea of the developed approach is as follows: if the algebraic approximations of a rough set do not change substantially in the process of excluding features the resulting reduced set of features can be used instead of the original full set. Also the greater changes eliminating a particular feature causes in the approximations, the more important this feature is.
Keywords: data mining, rough sets, rough approximations, reduct.


WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080011

1 Introduction

Modern Data Mining methods allow discovering non-trivial dependencies in large information arrays. Since these methods are used for processing and analysis of huge information volumes, reducing the number of features necessary for describing a discrete object is one of the most important problems. Rough sets theory has turned out to be quite an effective mathematical tool for resolving classification problems associated with searching reduced sets of features without losing necessary information on the object under consideration.

One of the basic concepts in rough sets theory is the concept of "indiscernibility" (or indiscernibility relation) [1, 2]. It is assumed that the same information can be associated with different elements of a set, which makes it impossible to exactly determine whether or not an element belongs to the set (such a set is called rough). A rough set is characterized by its lower and upper approximations. The lower approximation defines elements that *must* belong to the given set, and the upper approximation defines elements that *may* belong to the set.

The difference between the upper and lower approximations is called "boundary region". If the boundary region does not contain any element the set is considered as "strict", in other cases it is considered as "rough".

For finding approximations of a rough set various approaches can be used, including the algebraic approach developed by the authors earlier [4]. The suggested approach uses only comparison and Boolean operations, which makes the process of searching approximations and building approximation-based logic rules quite quick from the computational viewpoint.

Sets of "indiscernible" elements with which the same information is associated are called "knowledge granules". In this context also the concept of a reduct has been introduced [3]. Reduct is a minimal set of features that allows distinguishing granules. Being minimal set means the impossibility of its further reduction without losing the ability of distinguishing different granules. Thus reduct defines a set of features adequately describing a rough set. Z. Pawlak calls finding reducts for an arbitrary set an interesting but complicated problem [3].

In this paper we introduce a concept of a local reduct as an irreducible subset of features allowing us to describe a particular subset of the original set with the same precision as with the help of the full set of features. We suggest a method for finding local reducts with the help of the algebraic approach to describing rough set approximations developed by the authors earlier. The suggested method also allows us to assess the importance of each feature participating in the rough set description.

2 Reducts

When talking about reducts Z. Pawlak [1] meant irreducible sets of features that allow describing *any* subset of the original set without changing the granule structure of information on objects. We suggest a concept of a *local reduct* as an irreducible set of features that allows describing a *particular* subset of the

original set without adding any ambiguity in comparison with the full set of features. Let us consider the concept of reducts with the help of a simple example used by Pawlak. Suppose we have a fragment of a medical database, which contains some information on 6 patients (Table 1) [1].

The rows of this table contain values of attributes (disease symptoms), the columns denote patients A1...A6. It can be seen from the table that if a patient has a very high temperature it can be deduced from the available information that he/she has a flue, if patient's temperature is normal he/she does not have a disease.

Table 1: Patients and symptoms.

Patient	A1	A2	A3	A4	A5	A6
Headache	No	Yes	Yes	No	Yes	No
Muscle pain	Yes	No	Yes	Yes	No	Yes
Temperature	High	High	Very high	Normal	High	Very high
Flue	Yes	Yes	Yes	No	No	Yes

Consider the attribute "Temperature", which does not carry unambiguous information on whether or not a patient has a flue. There are patients with a high temperature that have a flue (A1 and A2) and those that do not (A5). Note also that the objects A2 and A5 are "indiscernible" from the viewpoint of the available information ("Headache", "Muscle pain", "Temperature") and it is impossible for these patients to determine the presence or absence of the disease unambiguously. Therefore in this case we have a rough set of ill patients.

Let us eliminate the attribute "Headache" (Table 2).

Table 2: Table without "Headache".

Patient	A1	A2	A3	A4	A5	A6
Muscle pain	Yes	No	Yes	Yes	No	Yes
Temperature	High	High	Very high	Normal	High	Very high
Flue	Yes	Yes	Yes	No	No	Yes

Note that the set of attributes {("Headache", "Muscle pain"} adequately describes the rough set of ill patients and allows classifying them with the same precision as the full set of attributes {("Headache", "Muscle pain", "Temperature"}. We can say with certainty that the patients having a high temperature (A3 and A6) have a flue and those whose temperature is normal do not. We can see unambiguously that the patient A1 has a flue from the viewpoint of the available information as he has muscle pain and high temperature and nobody else who is ill has such a combination of symptoms. The patients A2 and A5 are still indiscernible, which has been noted before. It can be stated that the granule structure of this set has not been changed. On the other hand it can be shown that no row in Table 2 can be further eliminated without causing additional ambiguity. For example, if the row "Temperature" is deleted the

patients A1 and A4 become indiscernible from the view point of the available information as both have muscle pain. Therefore the set of features {"Muscle pain", "Temperature"} is a local reduct for the set of ill patients. It can be also shown that the set of features {"Headache", "Temperature"} is a local reduct as well. Note that these local reducts are the same as "global" reducts considered by Pawlak. Nevertheless in many cases it is not so. Consider for example Table 3 where the row flue is a bit different from that in Table 1.

Table 3: Row "Flue" modified.

Patient	A1	A2	A3	A4	A5	A6
Headache	No	Yes	Yes	No	Yes	No
Muscle pain	Yes	No	Yes	Yes	No	Yes
Temperature	High	High	Very high	Normal	High	Very high
Flue	Yes	Yes	Yes	No	Yes	Yes

In this case the set of features {Temperature} containing a single element becomes a local reduct for the set of ill patients as "Temperature" allows determining unambiguously whether or not a patient has a flue. Obviously if a patient has a high or very high temperature he has a flue, otherwise he does not in accordance with the above information.

3 A method for finding local reducts

Before describing the procedure of finding local reducts of a rough set consider the concepts of upper and lower approximations [1, 2]. The upper approximation (I^*) consists of the elements that *may* belong to the given set. From the above example (Table 1) we can not state unambiguously whether or not the patients A2 and A5 have a flue as they are indiscernible from the viewpoint of the available information represented by values of the attributes "Headache", "Muscle pain", "Temperature". Since A2 has "Yes" in the row "Flue" and A2 and A5 are indiscernible, A5 should also be included in the set of patients that may have a flue, i.e. in the upper approximation. Also patients A1, A3 and A6 will be included in the upper approximation.

The lower approximation (I_*) consists of the elements about which it can be stated unambiguously (from the viewpoint of the available information) that they belong to the rough set of ill patients. In this example these are patients A1, A3 and A6. The patient A2 does not belong to the lower approximation as he/she is indiscernible with A5 who has "No" in the row "Flue".

Thus for this example:

$$I^*=\{1,2,3,5,6\},\ I_*=\{1,3,6\}.$$

A method for quick finding the upper and lower approximations of a rough set with the help of manipulating Boolean strings has been suggested in [4]. It allows finding approximations for a given set using only Boolean operations. Let us consider a brief description of this method.

Consider a non-empty set of objects $U=\{a_1, a_2,\ldots,a_n\}$ called universe and some subset X represented in the form of a Boolean row (Table 4).

Table 4: Elements and binary features.

Features \ Elements	a_1	a_2	…	a_n
P_1	δ_{11}	δ_{12}	…	δ_{1n}
P_2	δ_{21}	δ_{22}	…	δ_{2n}
…	…	…	…	…
P_k	δ_{k1}	δ_{k1}	…	δ_{kn}
X	λ_1	λ_2	…	λ_n

where:
$\delta_{ij}=1$, if an element a_i has a property (feature) P_j represented by a predicate P_j;
$\delta_{ij}=0$, if an element a_i does not have a property P_j;
$\lambda_i=1$, if an element a_i belongs to the set X;
$\lambda_i=0$, if an element a_i does not belong to X.

In the general form formulae for the approximations can be represented as follows [4]:

$$\begin{aligned} I^* &= (\lambda_1 \wedge P_1 * \delta_{11} \wedge P_2 * \delta_{21} \wedge \ldots \wedge P_k * \delta_{k1}) \vee \\ &\vee (\lambda_2 \wedge P_1 * \delta_{12} \wedge P_2 * \delta_{22} \wedge \ldots \wedge P_k * \delta_{k2}) \vee \ldots \\ &\ldots \vee (\lambda_n \wedge P_1 * \delta_{1n} \wedge P_2 * \delta_{2n} \wedge \ldots \wedge P_k * \delta_{kn}); \end{aligned} \quad (1)$$

$$\begin{aligned} I_* &= (\lambda_1 \vee P_1 * (1-\delta_{11}) \vee P_2 * (1-\delta_{21}) \vee \ldots \\ &\ldots \vee P_k * (1-\delta_{k1})) \wedge (\lambda_2 \vee P_1 * (1-\delta_{12}) \vee \\ &\vee P_2 * (1-\delta_{22}) \vee \ldots \vee P_k * (1-\delta_{k2})) \vee \ldots \\ &\ldots \vee (\lambda_n \vee P_1 * (1-\delta_{1n}) \vee P_2 * (1-\delta_{2n}) \vee \ldots \\ &\ldots \vee P_k * (1-\delta_{kn})), \end{aligned} \quad (2)$$

where $P*\delta=P$ if $\delta=1$, and $P*\delta=\overline{P}$ if $\delta=0$ for any predicate P.

Consider an example of finding approximations. Elements a_1, a_2, a_3, a_4, and a_5 are described with the help of features P_1, P_2, P_3. The set X consists of elements a_2, a_4, a_5 (Table 5).

Table 5: Example of elements and predicates.

Features \ Elements	a_1	a_2	a_3	a_4	a_5
P_1	1	0	0	1	0
P_2	0	0	1	1	1
P_3	0	1	0	0	0
X	0	1	0	1	1

The upper and lower approximations for X found in accordance with (1) and (2) are as follows:

$$I^* = \left(\overline{P_1} \wedge \overline{P_2} \wedge P_3\right) \vee \left(P_1 \wedge P_2 \wedge \overline{P_3}\right) \vee \left(\overline{P_1} \wedge P_2 \wedge \overline{P_3}\right),$$

$$I_* = \left(\overline{P_1} \vee P_2 \vee P_3\right) \wedge \left(P_1 \vee \overline{P_2} \vee P_3\right).$$

The calculated values of approximation vectors are represented in Table 6.

Table 6: Calculated approximations.

Features \ Elements	a_1	a_2	a_3	a_4	a_5
P_1	1	0	0	1	0
P_2	0	0	1	1	1
P_3	0	1	0	0	0
X	0	1	0	1	1
I^*	0	1	1	1	1
I_*	0	1	0	1	0

Searching for local reducts can be carried out by iterative forming reduced sets of features that can be obtained by eliminating one or several features from the full set of features $P_1, P_2, ..., P_k$.

When eliminating a feature we should be able to quickly conclude whether or not the descriptive strength of the resulting reduced set has changed. In this paper we suggest a new approach to assessing the reduced set of features based on evaluating changes in the boundary region (difference between the approximations) after excluding a feature.

It can be concluded from the above considerations that when a feature is eliminated from the set of features the upper approximation can only increase (in the sense that the true values in its Boolean vector remain intact and may be some false values will turn into true ones) and the lower approximation can only decrease (in the sense that the false values in its Boolean vector remain intact and may be some true values will turn into false ones). Thus the boundary region can only become greater when any feature is excluded.

Before searching for local reducts let us define a maximum possible change in the boundary region ($\Delta(BN_I)$) for which we can say that it does not influence substantially the descriptive strength of the resulting set of features. This value depends on the problem being solved and should be set by the analyst. Suppose that the maximum allowable change in the boundary region is 1 element. Let us exclude the feature P_1 from Table 5 and calculate new approximations. The results can be obtained from the following formulae and are represented in Table 7.

$$I_1^* = \left(\overline{P_2} \wedge P_3\right) \vee \left(P_2 \wedge \overline{P_3}\right),$$

$$I_{1*} = \left(P_2 \vee P_3\right) \wedge \left(\overline{P_2} \vee P_3\right).$$

Table 7: Approximations after excluding P_1.

Features \ Elements	a_1	a_2	a_3	a_4	a_5
P_2	0	0	1	1	1
P_3	0	1	0	0	0
X	0	1	0	1	1
I_1^*	0	1	1	1	1
I_{1*}	0	1	0	0	0

It can be seen from Table 7 that in this case the power of the upper approximation has not changed, but the power of the lower approximation has decreased by 1 element. Thus the boundary region has become 1 element larger. Since the allowable deviation $(\Delta(BN_I) = 1)$ has not been exceeded, the feature P_1 can be excluded and we obtain a reduced set P_2 , P_3.

Further we exclude the feature P_2 and calculate new approximations. The results can be calculated with the help of the following formulae and they are represented in Table 8.

$$I_2^* = \left(\overline{P_1} \wedge P_3\right) \vee \left(P_1 \wedge \overline{P_3}\right) \vee \left(\overline{P_1} \wedge \overline{P_3}\right);$$
$$I_{2*} = \left(\overline{P_1} \vee P_3\right) \wedge \left(P_1 \vee P_3\right).$$

Table 8: Approximations after excluding P_2.

Features \ Elements	a_1	a_2	a_3	a_4	a_5
P_1	1	0	0	1	0
P_3	0	1	0	0	0
X	0	1	0	1	1
I_2^*	1	1	1	1	1
I_{2*}	0	0	0	0	0

It can be seen from Table 8 that the upper and lower approximations have changed by 1 element each. Thus the change in the boundary region is now 2 elements $(\Delta(BN_I) = 2)$. This exceeds the maximum allowable deviation in the boundary region therefore the reduced set P_2, P_3 can not be a local reduct. It is obviously seen from Table 8 that the upper approximation includes all objects and the lower approximation does not contain any therefore all objects are indiscernible and the set X can be described only in a trivial way.

Following the above procedure of eliminating features, exclude P_3 from the original set and find the following approximations (Table 9).

$$I_3^* = \left(\overline{P_1} \wedge \overline{P_2}\right) \vee \left(P_1 \wedge P_2\right) \vee \left(\overline{P_1} \wedge P_2\right),$$
$$I_{3*} = \left(\overline{P_1} \vee P_2\right) \wedge \left(P_1 \vee \overline{P_2}\right).$$

Table 9: Approximations after excluding P_3.

Features \ Elements	a_1	a_2	a_3	a_4	a_5
P_1	1	0	0	1	0
P_2	0	0	1	1	1
X	0	1	0	1	1
I_3^*	0	1	1	1	1
I_{3*}	0	1	0	1	0

In this case the boundary region has not changed therefore no descriptive strength has been lost after eliminating feature P_3. Therefore the reduced set of features P_1, P_2 satisfies the required precision. It can be easily shown that after eliminating one of the features in the reduced sets P_1, P_2 and P_2, P_3 the remaining feature induces a boundary region differing from the original one in more than 1 element. Thus we can conclude that these reduced sets of features are local reducts for the set X.

It should be noted that the less elements the boundary region contains, the more "strict" (versus rough) the set being described is. The size of the boundary region can be used for measuring the quality of data from the viewpoint of the available information. The above procedure of eliminating features and recalculating boundary regions should be stopped when further elimination of variables produces unsatisfactory boundary regions differing from the original one in too many elements.

4 Assessing the importance of features

The suggested method allows assessing the importance (or classification strength) of the features participating in the description of a rough set and select the most important ones.

The *importance* $V(P_i)$ of a feature P_i can be calculated with the help of the following formula:

$$V(P_i) = \frac{\Delta(BN_I)}{M(X)} * 100\% ,$$

where M(X) is the number of true values in the Boolean vector X.

A special threshold minDeterioration can be defined by the analyst for determining whether or not a feature is important. If $V(P_i) \geq$ minDeterioration, the feature P_i is important, if $V(P_i) <$ minDeterioration, this feature is non-salient. Let us calculate the importance of the features from our example:

1. $V(P_1) = \frac{1}{3} * 100\% = 33.3\%$;
2. $V(P_2) = \frac{2}{3} * 100\% = 66.6\%$;
3. $V(P_3) = \frac{0}{3} * 100\% = 0\%$.

5 Conclusions and discussion

In this paper we suggest a method for finding reduced sets of features that allow describing a given set of objects with the same precision as with the help of the full set of features. In this connection we introduce a concept of local reducts that depend on particular sets being considered. For the purpose of finding such reducts we exclude features one by one and measure the changes in the resulting boundary regions in comparison with the original region. If after excluding a feature the boundary region does not change substantially we conclude that this feature can be eliminated and consider the remaining ones. Using our approach it is also possible to measure the importance of separate features from the viewpoint of the available information. The suggested method uses only Boolean operations, which makes it efficient from the computational viewpoint.

Of course the order in which we exclude features can play an important part for efficient extracting local reducts, therefore it would be nice to have a good criterion as to which feature should be excluded at a particular step. In this connection there may exist a variety of improvements to the procedure of random elimination. We would suggest that the least important features are eliminated first at each step but this problem requires more research and computer modeling.

References

[1] Pawlak Z.; (1995) *Rough set approach to knowledge-based decision support* / Proc. of the 14 European Conference on Operational Research Jerusalem, Israel.

[2] Pawlak, Z.; (1995) *Vagueness and uncertainty: a rough set perspective*, Computational Intelligence, vol. 11 (issue 2), pp. 227–232.

[3] Pawlak Z.; (1997) *Rough set approach to knowledge-based decision support*. European Journal of Operational Research, 99, pp. 420–432.

[4] D. Sitnikov, O. Ryabov.; (2004), *An algebraic approach to defining rough set approximations and generating logic rules,* in Zanasi, A.; Ebecken, N.; Brebbia, C.; (eds), Data Mining V, Malaga, Spain, pp. 179–188.