Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 331–338

ISBN 978-83-60810-51-4

Decision Bireducts and Approximate Decision Reducts: Comparison of Two Approaches to Attribute Subset Ensemble Construction

Sebastian Stawicki* and Sebastian Widz†‡
*Institute of Mathematics, University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
†Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
‡Xplus SA
ul. Puławska 435A, 02-801 Warsaw, Poland
sebastian.stawicki@gmail.com, sebastian.widz@xplus.pl

***Abstract*—We discuss the notion of a decision bireduct [1], which is an extension of the notion of a decision reduct developed within the theory of rough sets. We show relationships between the decision bireducts and some formulations of approximate decision reducts summarized in [2]. We investigate advantages of the decision bireducts and the approximate decision reducts within a rough-set-inspired framework for deriving attribute subset ensembles from data, wherein each of attribute subsets yields a single classifier, basically by generating its corresponding if-then decision rules from the training data. We also show how to use the above-mentioned relationships to build even more efficient rough-set-based ensembles in the future.**

***Keywords*-Attribute Subset Selection; Approximate Reducts; Bireducts; Classifier Ensembles; Randomized Search;**

I. Introduction

ATTRIBUTE subset selection plays an important role in knowledge discovery [3]. It establishes the basis for more efficient classification, prediction and approximation models. It also provides the users with a better insight into data dependencies. In this paper, we concentrate on attribute subset selection methods originating from the theory of rough sets [4]. There are numerous rough-set-based algorithms aimed at searching for so called *reducts* – irreducible subsets of attributes that satisfy predefined criteria for keeping enough information about decisions. Those criteria are verified on the training data and, usually, they encode the risk of misclassification by if-then decision rules with their antecedents referring to the values of investigated attribute subsets and their consequents referring to decisions.

Original definition of a reduct is quite restrictive, requiring that it should determine decisions or, if data inconsistencies do not allow full determinism, provide the same level of information about decisions as the complete set of attributes. There are a number of approaches to formulate and search for approximate or inexact reducts, which *almost* preserve the decision information [5]. Approximate reducts are usually smaller than standard ones, providing the basis for learning more efficient classifiers [6], [7].

In [1], the following issue concerning approximate reducts was outlined in relation to a popular idea of building classifier ensembles [8]. Combining classifiers is efficient especially if they are different from each other [9], [10]. In this way, one may increase stability of the classification and improve the ability to represent data dependencies. However, the original approximate reduct criteria do not allow controlling which parts of data are problematic for particular reducts. For example, when building an ensemble of reducts supposed to correctly classify at least 90% of the training objects, we may not anticipate that each of them will have problems with the same 10% of objects.

Given the above challenges, a new extension of the original rough-set-based notion of a decision reduct was proposed in [1]. Its interpretation seems to be simpler than in the case of most of the types of approximate decision reducts known from the literature. The emphasis here is on both, a subset of attributes that describes the decision classes and a subset of objects for which such a description is valid. Inspired by the methodology of biclustering [11], where it is crucial to work with objects and attributes simultaneously, the proposed notion is called a decision bireduct.

In this paper, we show that, although definitions and properties of decision bireducts and approximate reducts seem to be entirely different, there are some relationships between them. As already noticed in [1], the subsets of attributes being parts of the most representative decision bireducts turn out to be approximate decision reducts based on so called majority decision measure [12]. Moreover, it turns out that other measures utilized in [2] to define the approximate attribute reduction criteria can be modeled by the decision bireduct searching mechanisms as well.

The remainder of this paper is organized as follows. Sections II and III present approximate reducts and bireducts notions. In Section IV, we discuss how the two rough set

Table I
DECISION SYSTEM $\mathbb{A}$ WITH 14 OBJECTS IN U, FOUR ATTRIBUTES IN A, AND $d =$ SPORT?.

	Outlook	Temp.	Humid.	Wind	Sport?
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

Table II
EXAMPLES OF DECISION BIREDUCTS AND γ-BIREDUCTS

Bireduct	γ-Bireduct
({O H},{1..3 6..9 11..14})	({O H},{1..3 7..9 11..13})
({O H W},{1..14})	({O H W},{1..14})
({O T W},{1..14})	({O T W},{1..14})
({T W},{1..2 4..5 7 9..10 14}) ({T W},{2..6 9..13})	({T W},{2 5 9})
({T H},{1..2 6 8 10..11 13..14})	({T H},{10..11 13})
({O},{1..3 6..8 12..14}) ({O},{1..5,7..8,10,12..13})	({O},{3 7 12..13})
({O W},{2..7 9..10 12..14})	({O W},{3..7 10 12..14})
({O T},{1..4 6..10 12..13})	({O T},{1..3 7 9 12..13})
({H W},{1 5..6 8..10 12..13})	({H W},{5 9..10 13})

concepts differ and how they are related. The comparison of bireducts and approximate reducts notions is made. We show formal relationships expressed in terms of mathematical formulas. In Section V, we describe algorithms used for approximate reduct and bireduct calculation. Section VI describes ensembles of classifiers based on discussed reduct types. Finally, we conclude this paper in Section VII.

II. FOUNDATIONS OF DECISION BIREDUCTS

We use the standard notation of decision systems to represent data [4]. By a decision system we mean a tuple $\mathbb{A} = (U, A \cup \{d\})$, where U is a set of objects and A is a set of attributes and $d \notin A$ is a distinguished decision attribute. For simplicity, we refer to the elements of U using their ordinal numbers $i = 1, ..., |U|$, where $|U|$ denotes the cardinality of U. We treat attributes $a \in A$ as functions $a : U \to V_a$, V_a denoting a's domain. The values $v_d \in V_d$ correspond to decision classes that we want to describe using the values of attributes in A. Thus, decision systems can be employed within the standard supervised learning framework.

Let us now consider the notion of a reduct, which is one of the most important contributions of the theory of rough sets into knowledge discovery and data mining. We say that $B \subseteq A$ is a decision reduct for $\mathbb{A} = (U, A \cup \{d\})$, iff it is an irreducible subset of attributes such that each pair $i, j \in U$ satisfying inequality $d(i) \neq d(j)$ is discerned by B. For $\mathbb{A}$ in Table I, there are two decision reducts: $\{\text{Outlook}, \text{Temp.}, \text{Wind}\}$ and $\{\text{Outlook}, \text{Humid.}, \text{Wind}\}$ (or $\{\text{O}, \text{T}, \text{W}\}$ and $\{\text{O}, \text{H}, \text{W}\}$ for short).

Definition 1. *Let $\mathbb{A} = (U, A \cup \{d\})$ be a decision system. A pair (B, X), where $B \subseteq A$ and $X \subseteq U$, is called a decision bireduct, iff B discerns all pairs $i, j \in X$ where $d(i) \neq d(j)$, and the following properties hold:*

1) *There is no $C \subsetneq B$ such that C discerns all pairs $i, j \in X$ where $d(i) \neq d(j)$;*
2) *There is no $Y \supsetneq X$ such that B discerns all pairs $i, j \in Y$ where $d(i) \neq d(j)$.*

A decision bireduct (B, X) can be regarded as an inexact functional dependence linking the subset of attributes B with the decision d in a degree X, denoted by $B \Rrightarrow_X d$. The objects in $U \setminus X$ can be treated as the outliers. Remove of noisy data, outliers, redundant objects (generally known as the instance selection problem [13]) is often one of the first steps of data preparation and can significantly improve the results of the main data analysis algorithms. In the bireducts approach, the instance selection problem appears integrated in the whole methodology. The objects in X can be used to learn a classifier based on B from data. For instance, one can partition X with respect to its elements' values on B and treat the combinations of values labeling partition classes (called indiscernibility classes in the rough set literature [4]) as the antecedents of rules pointing at specific decision values, uniquely defined within X thanks to the properties of decision bireducts.

Table I shows a few decision bireducts and γ-bireducts for a well-known example of a decision system presented on Table I. The number of all decision bireducts for this data set is far higher than illustrated. One may notice that the same $B \subseteq A$ can occur as a component of many bireducts, with different subsets of objects.

Some interesting questions arise as to what is the best or optimal bireduct or when we can say that one bireduct is better than another. Following the idea from [1] it may be convenient to describe decision reducts in terms of their attributes and outliers. An implicit assumption is that bireducts shall minimize both those factors. Minimizing the size of the attribute subset is quite intuitive as to its great analogy of minimizing the size of reducts. The smaller the size the more general is the description of the decision system. The minimization of the size of the outliers set is also an intuitive tendency to minimize the number of objects that do not fit to (or disturb) the description of the decision system contained in a decision bireduct. The minimization task translates into a maximization of the size of the objects contained in a decision bireduct. In case of imbalanced data sets (with large disproportion between number of objects with certain decision value assigned) the simplest form of measuring the size based on the cardinality of the objects subset may be insufficient. In such cases one should pay more attention to a specified subset of objects, e.g. objects belonging to the minority decision classes.

We introduce a following modification of the bireduct Definition 1. The difference is the requirement that objects

belonging to the bireduct should now be discerned not only in context of X but of the entire U. One can notice the analogy of this attitude to a positive region sometimes denoted as POS from the theory of rough sets. In other words the object can be added to X if it belongs to positive region defined by attribute subset B in U. To distinguish between those two types of bireducts we call the latter γ-bireduct.

Definition 2. *Let $\mathbb{A} = (U, A \cup \{d\})$ be a decision system. A pair (B, X), where $B \subseteq A$ and $X \subseteq U$, is called a decision γ-bireduct, iff B discerns all pairs $i \in X, j \in U$ where $d(i) \neq d(j)$, and the following properties hold:*

1) *There is no $C \subsetneq B$ such that C discerns all pairs $i \in X, j \in U$ where $d(i) \neq d(j)$;*
2) *There is no $Y \supsetneq X$ such that B discerns all pairs $i \in Y, j \in U$ where $d(i) \neq d(j)$.*

A decision γ-bireduct (B, X) can be regarded as an inexact functional dependence linking the subset of attributes B with the decision d in a degree X on U, denoted by $B \Rrightarrow^g_X d$, where g can be regarded as an indicator for a global scope of the functional dependence. Indeed, the difference between the two mentioned bireduct types lies in the scope of considered objects. The latter type focuses on the discernibility on the whole set of object, while the former considers a local scope and ensures discernibility among the objects belonging to the bireduct.

III. Foundations of Approximate Reducts

There is a variety of methods of searching for approximate reducts in decision systems (eg.[14], [15]). The criteria usually include formulas for functions measuring degrees of decision information induced by subsets of attributes and thresholds for those functions' values specifying which subsets of attributes are *good enough*. The choice of functions may depend on the nature of particular data sets and methods of learning classifiers based on reduced sets of attributes.

In order to follow the filter approach to feature subset selection, we need to design some measures that evaluate particular feature subsets in the selection process. From this point of view, the rough set literature may be regarded as a source of measures that draw correspondence between feature subsets and rule-based classifiers corresponding to those subsets. Let us consider the following three examples of such measures:

$$\begin{aligned} \gamma(B) &= \textstyle\sum_{E_B \in U/B : Pr(X^M_{E_B} | E_B) = 1} Pr(E_B) \\ M(B) &= \textstyle\sum_{E_B \in U/B} Pr(X^M_{E_B}, E_B) \\ R(B) &= \textstyle\sum_{E_B \in U/B} Pr(E_B | X^R_{E_B}) \end{aligned} \tag{1}$$

where "$U/\cdot$" denotes indiscernibility equivalence classes induced by a set of attributes. The above decision classes $X^M_{E_B}$ and $X^R_{E_B}$ are defined as follows:

$$\begin{aligned} X^M_{E_B} &= \arg\max_{E_d \in U/\{d\}} Pr(E_d | E_B) \\ X^R_{E_B} &= \arg\max_{E_d \in U/\{d\}} Pr(E_B | E_d) \end{aligned} \tag{2}$$

All above measures are inspired by the theory of rough sets: γ corresponds to the already mentioned rough set positive region [4], while M and R – to its probabilistic and Bayesian extensions, respectively [12], [16]. We refer to M na R as *decision* types Majority and Relative respectively. Majority type points at the most frequent decision for E_B. It is the most popular way of constructing rules in machine learning. Relative type points at decision that is most frequent within E_B relative to its prior probability. It may be particularly worth applying for imbalanced data sets, although it was used successfully also for data sets with larger number of uniformly supported decision classes (see [6] for references). It is also worth noting that, for the simplest case where feature subsets induce equivalence relations, we have $\gamma(C) \leq \gamma(B)$, $M(C) \leq M(B)$, and $R(C) \leq R(B)$, for any $C \subseteq B$. Such property is important for feature subset selection algorithms. In our studies, we search for so called (F, ε)-reducts, where F may mean γ, M, R, or any other measure that can represent decision information provided by feature subsets, and $\varepsilon \in [0, 1)$ decides how much of quality of determining d we agree to lose when operating with smaller subsets $B \subseteq A$ (and shorter rules as a consequence), according to the following constraint:

$$F(B) \geq (1 - \varepsilon) F(A) \tag{3}$$

In [2], we compared functions F by means of average/maximum classifiers' accuracies obtained for different settings of ε. We noted that γ, M, and R usually start leading towards empty feature sets (corresponding to trivial classifiers based on rules with empty left sides) for different values of ε. Therefore, instead of (3), we used the following re-scaled criterion to assure more accurate comparison:

$$F(B) - F(\emptyset) \geq (1 - \varepsilon)(F(A) - F(\emptyset)) \tag{4}$$

Equivalently:

$$F(B) \geq (1 - \varepsilon) F(A) + \varepsilon F(\emptyset) \tag{5}$$

IV. Approximate reducts and bireducts comparison

Although definitions and properties of decision bireducts and approximate reducts seem to be different, we may define relationships expressing a correspondence between them in terms of measures described in previous section. More precisely for a given attribute subset we will show that the size of X (expressed in terms of values of mathematical formulas) for the decision bireduct (intuitively the best or most meaningful) corresponds to the appropriate criterions of (F, ε)-approximate reduction where F is used to evaluate a quality for specified attribute subsets.

Proposition 1. *Let $\mathbb{A} = (U, A \cup \{d\})$ be a decision system and let M be the measure defined in (1). For $B \subseteq A$ let*

$$\mathbb{X}_B = \{X \subseteq U \mid (B, X) \text{ is a bireduct for } \mathbb{A}\}$$

be a family of object subsets which together with B create decision bireducts for $\mathbb{A}$. Then

$$\max_{X \in \mathbb{X}_\mathbb{B}} \frac{|X|}{|U|} = M(B)$$

Proof

Remark 1. *Let $\mathbb{A}$ be a decision system defined as above and (B, X) be a bireduct for $\mathbb{A}$. For each indiscernibility class $E_B \in U/B$ the following two statements are true:*

- *Objects from $X \cap E_B$ are of the same class (have the same value for the decision attribute).*
- *Moreover, all objects from E_B that belong to that decision class are contained in X.*

This is just a simple conclusion of the bireduct definition:

- *If $X \cap E_B$ contains objects from more than one decision class it will break the definition of bireduct because with the given attribute subset B we cannot discern at least one pair of objects with different decisions.*
- *If $X \cap E_B$ contains objects with the same decision but $X \cap E_B$ is not maximal in a sense that there is at least one object in E_B with the same decision which is not contained in X, it will break the second part of the definition of bireduct - maximality of object subset X.*

The left hand-side of the proposition's statement can be transformed as follows:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \frac{|X|}{|U|} \\ &= \max_{X \in \mathbb{X}_{\mathbb{B}}} Pr(X) \\ &= \max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} Pr(X, E_B) \end{aligned} \tag{6}$$

For $M(B)$ based on (1) and (2) we can write:

$$\begin{aligned} M(B) &= \sum_{E_B \in U/B} Pr(X^M_{E_B}, E_B) \\ &= \sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} Pr(E_d, E_B) \end{aligned} \tag{7}$$

Now we only need to show that:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} Pr(X, E_B) \\ &= \sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} Pr(E_d, E_B) \end{aligned} \tag{8}$$

Let us prove it by a contradiction. Suppose that inequality holds. Let $\overline{X} \in \mathbb{X}_{\mathbb{B}}$ be an object subset which maximizes the left-hand side expression:

$$\overline{X} = \arg \max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} Pr(X, E_B)$$

There are two cases. The first one is as follows:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} Pr(X, E_B) \\ &= \sum_{E_B \in U/B} Pr(\overline{X}, E_B) \\ &< \sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} Pr(E_d, E_B). \end{aligned} \tag{9}$$

It is easy to notice that there exists $E^{\overline{X}}_B \in U/B$ such that

$$Pr(\overline{X}, E^{\overline{X}}_B) < \max_{E_d \in U/\{d\}} Pr(E_d, E^{\overline{X}}_B)$$

This means that for $(B, \overline{X})$ bireduct we can show an object subset

$$\overline{\overline{X}} = \overline{X} \setminus E^{\overline{X}}_B \cup ((\arg \max_{E_d \in U/\{d\}} Pr(E_d, E^{\overline{X}}_B)) \cap E^{\overline{X}}_B)$$

such that $|\overline{\overline{X}}| > |\overline{X}|$. According to the Remark 1 it is impossible (it breaks the bireduct's object subset maximality property), contradiction.

Now the second case:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} Pr(X, E_B) \\ &= \sum_{E_B \in U/B} Pr(\overline{X}, E_B) \\ &> \sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} Pr(E_d, E_B) \end{aligned} \tag{10}$$

There exists $E^{\overline{X}}_B \in U/B$ such that

$$Pr(\overline{X}, E^{\overline{X}}_B) > \max_{E_d \in U/\{d\}} Pr(E_d, E^{\overline{X}}_B)$$

This means that the number of objects in $\overline{X}$ which are contained in $E^{\overline{X}}_B$ indiscernibility class is greater than the largest decision class within that equivalence class. This implies that $\overline{X}$ contains objects from different decision classes. According to Remark 1 it is impossible, contradiction.

The assumption that

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} Pr(X, E_B) \\ &\neq \sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} Pr(E_d, E_B) \end{aligned} \tag{11}$$

leads to a contradiction. This shows that

$$\max_{X \in \mathbb{X}_{\mathbb{B}}} \frac{|X|}{|U|} = M(B)$$

and completes the proof.

Proposition 2. *Let $\mathbb{A} = (U, A \cup \{d\})$ be a decision system, let R be the measure defined in (1) and for $u \in U$ let*

$$\mu(u) = \frac{1}{Pr(d(u))}$$

be a weight function defined for objects from U. For $B \subseteq A$ let

$$\mathbb{X}_B = \{X \subseteq U \mid (B, X) \text{ is a bireduct for } \mathbb{A}\}$$

be a family of object subsets which together with B create decision bireducts for $\mathbb{A}$. Then

$$\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{u \in X} \frac{\mu(u)}{|U|} = R(B)$$

Proof The proof can be conducted in a similar fashion that for Proposition 1. The left hand-side of the proposition's statement can be transformed as follows:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{u \in X} \frac{\mu(u)}{|U|} \\ &= \max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} \sum_{u \in E_B \cap X} \frac{\mu(u)}{|U|} \end{aligned} \tag{12}$$

For $R(B)$ based on (1) and (2) we can write:

$$\begin{aligned} R(B) &= \sum_{E_B \in U/B} Pr(E_B | X^R_{E_B}) \\ &= \sum_{E_B \in U/B} Pr(E_B | \arg\max_{E_d \in U/\{d\}} Pr(E_B | E_d)) \\ &= \sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} \frac{Pr(E_d, E_B)}{Pr(E_d)} \end{aligned} \tag{13}$$

We only need to show that:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} \sum_{u \in E_B \cap X} \frac{\mu(u)}{|U|} \\ &= \sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} \frac{Pr(E_d, E_B)}{Pr(E_d)} \end{aligned} \tag{14}$$

Let us prove it by a contradiction. Suppose that inequality holds. Let $\overline{X} \in \mathbb{X}_{\mathbb{B}}$ be an object subset which maximizes the left-hand side expression:

$$\overline{X} = \arg\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{E_B \in U/B} \sum_{u \in E_B \cap X} \frac{\mu(u)}{|U|}$$

There are two cases. The first one is as follows:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \textstyle\sum_{E_B \in U/B} \sum_{u \in E_B \cap X} \frac{\mu(u)}{|U|} \\ &= \textstyle\sum_{E_B \in U/B} \sum_{u \in E_B \cap \overline{X}} \frac{\mu(u)}{|U|} \\ &< \textstyle\sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} \frac{Pr(E_d, E_B)}{Pr(E_d)} \end{aligned} \tag{15}$$

It is easy to notice that there exists $E_B^{\overline{X}} \in U/B$ such that

$$\sum_{u \in E_B^{\overline{X}} \cap \overline{X}} \frac{\mu(u)}{|U|} < \max_{E_d \in U/\{d\}} \frac{Pr(E_d, E_B^{\overline{X}})}{Pr(E_d)}$$

$(B, \overline{X})$ is a bireduct and all objects from $E_B^{\overline{X}} \cap \overline{X}$ are from the same decision class $\overline{E_d} \in U/\{d\}$. We can write:

$$\sum_{u \in E_B^{\overline{X}} \cap \overline{X}} \frac{\mu(u)}{|U|} = |E_B^{\overline{X}} \cap \overline{X}| \cdot \frac{\frac{1}{Pr(\overline{E_d})}}{|U|} = \frac{Pr(\overline{E_d}, E_B^{\overline{X}})}{Pr(\overline{E_d})}$$

This means that $\overline{X}$ is not maximal, contradiction.
The second case:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \textstyle\sum_{E_B \in U/B} \sum_{u \in E_B \cap X} \frac{\mu(u)}{|U|} \\ &= \textstyle\sum_{E_B \in U/B} \sum_{u \in E_B \cap \overline{X}} \frac{\mu(u)}{|U|} \\ &> \textstyle\sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} \frac{Pr(E_d, E_B)}{Pr(E_d)} \end{aligned} \tag{16}$$

There exists $E_B^{\overline{X}} \in U/B$ such that

$$\sum_{u \in E_B^{\overline{X}} \cap \overline{X}} \frac{\mu(u)}{|U|} > \max_{E_d \in U/\{d\}} \frac{Pr(E_d, E_B^{\overline{X}})}{Pr(E_d)}$$

By analogy to the previous case, we have:

$$\frac{Pr(\overline{E_d}, E_B^{\overline{X}})}{Pr(\overline{E_d})} > \max_{E_d \in U/\{d\}} \frac{Pr(E_d, E_B^{\overline{X}})}{Pr(E_d)}$$

which is impossible, contradiction.
The assumption that

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \textstyle\sum_{E_B \in U/B} \sum_{u \in E_B \cap X} \frac{\mu(u)}{|U|} \\ &\neq \textstyle\sum_{E_B \in U/B} \max_{E_d \in U/\{d\}} \frac{Pr(E_d, E_B)}{Pr(E_d)} \end{aligned} \tag{17}$$

leads to a contradiction. This shows that

$$\max_{X \in \mathbb{X}_{\mathbb{B}}} \sum_{u \in X} \frac{\mu(u)}{|U|} = R(B)$$

and completes the proof.

Proposition 3. *Let $\mathbb{A} = (U, A \cup \{d\})$ be a decision system, and let γ be the measure defined in (1). For $B \subseteq A$ let*

$$\mathbb{X}_B^{\gamma} = \{X \subseteq U \mid (B, X) \text{ is a } \gamma\text{-bireduct for } \mathbb{A}\}$$

be a family of attribute subsets which together with B create decision γ-bireducts for $\mathbb{A}$. Then

$$\max_{X \in \mathbb{X}_{\mathbb{B}}^{\gamma}} \frac{|X|}{|U|} = \gamma(B)$$

Remark 2. *Let $\mathbb{A}$ be a decision system defined as above and (B, X) be a γ-bireduct for $\mathbb{A}$. For each indiscernibility class $E_B \in U/B$ the following two statements are true:*

- *If all objects from E_B have the same value for the decision attribute then the whole E_B is included in X. Otherwise no object from X_B belongs to X.*

This is just a simple conclusion of the γ-bireduct definition:

- *If all objects from E_B have the same value for the decision attribute then all of them must be in X, otherwise breaking the definition in a sense that X would not be maximal. If E_B contains objects from more than one decision class then no object from E_B can belong to X because with the given attribute subset B we cannot discern at least one pair of objects from U with different decision, thus violating the definition of γ-bireduct.*

A simple observation can be made on base of Remark 2. For each indiscernibility class $E_B \in U/B$ based on its content it is predetermined if it is contained in γ-bireducts or it is excluded. Hence, $\mathbb{X}_{\mathbb{B}}^{\gamma}$ has only one element.
The left hand-side of the proposition's statement can be transformed as follows:

$$\begin{aligned} &\max_{X \in \mathbb{X}_{\mathbb{B}}} \tfrac{|X|}{|U|} \\ &= \max_{X \in \mathbb{X}_{\mathbb{B}}} Pr(X) \\ &= \max_{X \in \mathbb{X}_{\mathbb{B}}} \textstyle\sum_{E_B \in U/B} Pr(X, E_B) \\ &\textit{using the Remark 2 we can write} \\ &= \max_{X \in \mathbb{X}_{\mathbb{B}}} \textstyle\sum_{E_B \in U/B : Pr(X|E_B)=1} Pr(E_B) \\ &= \tfrac{POS(B)}{|U|} \end{aligned} \tag{18}$$

For $\gamma(B)$ we can write:

$$\begin{aligned} &\gamma(B) = \textstyle\sum_{E_B \in U/B : Pr(X^M_{E_B}|E_B)=1} Pr(E_B) \\ &= \textstyle\sum_{E_B \in U/B : \max_{E_d \in U/\{d\}} Pr(E_d|E_B)=1} Pr(E_B) \\ &\textit{we can easily notice that we consider those } E_B \\ &\textit{which contain objects from only one decision class} \\ &= \tfrac{|POS(B)|}{|U|} \end{aligned} \tag{19}$$

This shows that

$$\max_{X \in \mathbb{X}_{\mathbb{B}}^{\gamma}} \frac{|X|}{|U|} = \gamma(B)$$

and completes the proof.

V. Attribute Reduction Algorithms

Let us go back to decision bireducts and consider the algorithm proposed in [1] in order to search for optimal bireducts in data. It takes as an input a permutation $\sigma : \{1, ..., n+m\} \to \{1, ..., n+m\}$ mixing the ordinal numbers of attributes counted from 1 to n, $n = |A|$, together with objects represented by numbers from $n+1$ to $n+m$, $m = |U|$. The algorithm is initiated with the pair (B, X), where $B = A$ and $X = \emptyset$. Then, it examines the values of $\sigma(i)$, for

$i = 1, ..., n + m$. Depending on whether $\sigma(i)$ corresponds to an attribute (the case of $\sigma(i) \leq n$) or an object (the case of $\sigma(i) > n$; the corresponding object is then retrieved as $\sigma(i) - n$), it attempts to remove it from B or to add it to X, respectively. The removal/addition conditions are defined for the decision bireducts calculation as $B \setminus \{a_{\sigma(i)}\} \Rrightarrow_X d$ and $B \Rrightarrow_{X \cup \{\sigma(i)-n\}} d$, using the inexact functional dependence notation introduced earlier. In case of γ-bireduct calculation the conditions are $B \setminus \{a_{\sigma(i)}\} \Rrightarrow^g_X d$ and $B \Rrightarrow^g_{X \cup \{\sigma(i)-n\}} d$ respectively.

Algorithm 1 presents details of bireduct calculation. In case of γ-bireduct the algorithm is almost the same. It is enough to change the functional dependence $\Rrightarrow_X$ with its equivalent from the description of the decision $gamma$-bireduct $\Rrightarrow^g_X$. For comparison we also put the procedure for (F, ε)-approximate decision reduct calculation in Algorithm 2.

Algorithm 1 Decision bireduct calculation for a decision system $\mathbb{A} = (U, A \cup \{d\})$

```
B ← A, X ← ∅
for i = 1 → n + m do
    if σ(i) ≤ n then
        if B \ {a_σ(i)} ⇛_X d then
            B ← B \ {a_σ(i)}
        end if
    else
        if B ⇛_{X∪{σ(i)−n}} d then
            X ← X ∪ {σ(i) − n}
        end if
    end if
end for
```

In [1] it is also noted that for each permutation $\sigma : \{1, ..., n + m\} \rightarrow \{1, ..., n + m\}$, where $n = |A|$ and $m = |U|$, the output (B, X) of the above algorithm is a decision bireduct. In fact, this algorithm is a generalization of a permutation-based method developed for searching for standard rough-set-based reducts (see [17]), which was also adapted for the approximate reducts [5]. The difference is that in case of the standard reducts the algorithm works with permutations $\sigma : \{1, ..., n\} \rightarrow \{1, ..., n\}$, $n = |A|$, and needs to assure irreducibility of the generated subsets of attributes. In the case of the bireduct search we need to assure both, irreducibility of the subset of attributes and non-extendability of the subset of objects.

Examples of bireducts and γ-bireducts are shown on Table I and V respectively.

It is worth notifying that for (F, ε)-reducts the value of ε can be understood as a threshold for allowed decrease of classifier determination and can address the balance between simplicity and accuracy. For larger ε, calculated reducts contain less attributes and generated decision rules become shorter. Shorter decision rules are more applicable. By the cost of slight inconsistencies we gain higher simplicity and applicability for unseen cases. On the other hand for smaller ε generated reducts contain more attributes and decision rules generated based on attributes belonging to a reduct are potentially less applicable but more accurate.

Algorithm 2 Permutation-based (F, ε)-REDORD (a bit modified comparing to [5])

Input: $\varepsilon \in [0, 1)$, $\mathbb{A} = (U, A \cup \{d\})$,
$\sigma : \{1, ..., n\} \rightarrow \{1, ..., n\}$, $n = |A|$
Output: $B \subseteq A$

```
B ← A
for i = 1 → n do
    if F(B \ {a_σ(i)}) ≥ (1 − ε)F(A) + εF(∅) then
        B ← B \ {a_σ(i)}
    end if
end for
return B
```

Table III
EXAMPLES OF DECISION BIREDUCTS WITH PERMUTATIONS USED TO GENERATE THEM.

Permutation	Bireduct
O 8 W 1 4 7 2 14 10 12 9 T 6 3 13 5 11 H	({H},{1..2 5 7..11 13..14})
H 13 T 8 W 6 11 3 14 10 O 5 7 9 2 1 4 12	({O},{1..3 6..8 12..14})
3 8 T 1 W 11 9 O 14 12 6 4 7 H 10 13 2 5	({O H},{1..3 6..9 11..14})
2 13 5 14 11 7 12 4 3 1 9 6 8 10 H O W T	({O T W},{1..14})
9 4 12 14 1 8 7 3 10 13 6 11 2 5 W T H O	({O H W},{1..14})
11 O 2 H 1 10 5 7 9 8 3 13 T 6 14 12 4 W	({T},{1..2 4..5 7 9..12})
T 2 5 H 10 11 W 14 1 12 7 9 13 6 4 8 3 O	({O},{1..5,7..8,10,12..13})
W 6 H O 5 8 4 7 3 2 10 9 12 11 13 14 1 T	({T},{3 6 8 13..14})
O 2 3 13 1 H 4 T W 6 12 14 5 8 9 10 11 7	({W},{2..6 9..10 13..14})
O H 14 1 10 7 4 3 12 13 5 W 9 T 11 8 2 6	({T W},{1..2 4..5 7 9..10 14})
6 5 10 9 H O 12 T 8 W 4 2 13 3 7 1 14 11	({T W},{2..6 9..13})
11 14 9 13 3 7 8 2 5 1 12 W 6 4 10 H O T	({O H},{1..3 5 7..14})
13 8 6 H 7 W 9 T 5 3 4 12 O 2 10 14 11 1	({O T},{1..4 6..10 12..13})
9 H 2 4 6 13 14 7 T 11 10 O W 3 5 1 8 12	({O W},{2..7 9..10 12..14})

In case of bireducts the approximation threshold is not defined directly but it is somehow expressed in the way permutations are generated [1]. We can define a parameter that can control probability of selecting an attribute in first place rather than an object while building the permutation σ. One can notice that when the permutation σ contains more attributes at its beginning, this results in a bireduct having smaller number of attributes but also higher number of outliers. In case more attributes are put at the end of the permutation, a resulting bireduct will have larger number of objects (less outliers) but also contain potentially more attributes.

In case of unbalanced data sets it might be desirable to focus on objects from the minority decision classes. To draw more classifier attention to those objects one can assign higher weights to them, that would increase the chance of placing them at the beginning of permutation thus increase the chances of covering them by calculated bireduct. We can express this relation as follows - the more important object is, the closer to the beginning of the permutation it should appear.

VI. ENSEMBLES

Classifier ensembles perform usually better than their components used independently. Combining several classifiers is efficient especially if they are substantially different from each

Table IV
EXAMPLES OF DECISION γ-BIREDUCTS WITH PERMUTATIONS USED TO GENERATE THEM.

Permutation	γ-Bireduct
11 2 T 9 W 6 O H 13 8 7 12 5 3 10 4 14 1	({O H},{1..3 7..9 11..13})
7 14 3 5 10 12 9 6 11 W O 13 1 4 T H 8 2	({O H W},{1..14})
6 O W 9 4 5 11 1 H T 14 12 13 7 3 2 8 10	({O T W},{1..14})
O H 14 1 10 9 13 W 6 7 12 T 8 2 5 4 3 11	({T W},{2 5 9})
W 10 H O T 12 3 14 2 7 5 13 8 4 11 6 1 9	({T H},{10..11 13})
W H 4 T 7 6 O 11 10 9 1 12 3 13 8 14 2 5	({O},{3 7 12..13})
4 O T 10 H 7 1 14 W 11 9 12 5 3 6 8 13 2	({O W},{3..7 10 12..14})
H 7 W 5 4 8 13 1 10 O T 14 6 12 9 2 11 3	({O T},{1..3 7 9 12..13})
O 9 W T H 7 1 4 6 8 5 13 11 2 3 14 12 10	({H W},{5 9..10 13})

other. One of the approaches is to construct many different classifiers based on possibly least overlapping subsets. Permutation based algorithms described in the previous section can be successfully utilized in construction of reducts ensembles following such criteria. As mentioned earlier, when generating permutations one can control the order of attributes (also order of objects for the bireduct version).

It is already known that in the case of the original permutation-based algorithm, the standard (or approximate) reducts that contain less common attributes are more likely to be obtained by employing a randomly chosen $\sigma : \{1, ..., n\} \rightarrow \{1, ..., n\}$ [5]. This property has an impact on the search for the most meaningful reducts, as well as the reduct ensembles – the sets of reducts with minimal intersections [2], [18]. The analogous behavior can be seen the algorithms proposed in [1], where randomly chosen permutations $\sigma : \{1, ..., n + m\} \rightarrow \{1, ..., n + m\}$ are more likely to lead to the bireducts with more diverse attributes and outliers. Such an ability to control the ensembles of bireducts with respect to the areas of objects that they cover is especially important for robustness of the resulting classification systems and completeness of data representation.

In bireduct version each subset in the ensemble is tuned to recognize particular set of objects. Assigning greater weights to objects not included in any of bireducts would cause placing them at the beginning of the permutation thus increasing chances of to be recognized by the classifier. As we described in previous sections bireduct object weights are used to construct permutations that are more likely to begin with objects having less frequent decision classes. The weights can be also assigned dynamically during the iterative process of bireducts ensemble creation. In consecutive iteration steps one may analyze the factor of object coverage by bireducts already selected for the ensemble. This process is somehow similar to the mechanisms applied in boosting type algorithms [19] where objects that are misclassified gain more weight during classifier training process.

Each reduct in the ensemble can be referred as a single classifier producing a set of decision rules. When different rules apply for the same object one must perform aggregate their outcomes using some simple voting mechanisms. In [2] we analyzed six different strategies of voting in decision rule based classifiers. Table V illustrated parameters that were taken into

Table V
SIX OPTIONS OF WEIGHTING DECISIONS BY RULES, CORRESPONDING TO THE *right-side* WEIGHT TYPES PLAIN, CONFIDENCE AND COVERAGE, AND *left-side* WEIGHT TYPES SINGLE AND SUPPORT. E_B DENOTES THE SUPPORT OF A RULE'S LEFT SIDE. $X^*_{E_B}$ DENOTES $X^M_{E_B}$ OR $X^R_{E_B}$ DEPENDING ON *decision* TYPE.

	Single	Support
Plain	1	$Pr(E_B)$
Confidence	$Pr(X^*_{E_B}\|E_B)$	$Pr(X^*_{E_B}, E_B)$
Coverage	$Pr(X^*_{E_B}\|E_B)/Pr(X^*_{E_B})$	$Pr(E_B\|X^*_{E_B})$

account. We considered three *right-side* weight possibilities to assign a value to a given decision class $X^*_{E_B}$ pointed by a rule supported by $E_B \in U/B$: 1 (Plain), $Pr(X^*_{E_B}|E_B)$ (Confidence), and $Pr(X^*_{E_B}|E_B)/Pr(X^*_{E_B})$ (Coverage). The rule's vote can be also strengthen by its left side's support $Pr(E_B)$ (*left-side* weight type Support) or not (Single).

In case of bireducts based classifiers only Plain and Coverage right-side weighting types can be used. For bireduct ensembles voting weights are calculated on local X where confidence is always equal to 1 (rather than on global U as in case of approximate reducts where rule confidence matters).

VII. CONCLUSIONS

We have compared the notion of a decision bireduct originally presented in [1] and extend its definition to so called γ-bireduct. Although their concept is different from the concept of approximate decision reducts we have shown that certain analogies and relationships exist. We can compare (R, ε)-reducts based on so called Relative decision measure $R(B)$ to bireducts generated from permutation where objects having minor decision are put at fist place. Analogously similar relations were shown between majority decision measure based (M, ε)-reducts and bireducts where objects can be added to X only when their decision is most frequent in certain equivalence classes induced by attribute subset B. γ-bireducts where objects can be added to X only when they belong to positive region induced by B have strong analogy to (γ, ε)-reducts.

We have addressed several questions about the quality of bireducts. As it was shown in section IV there should be strong relation between the quality of X for certain bireduct types and measures types used in their approximate reducts counterparts.

Bireduct properties and their relationship with approximate reducts still need closer investigation. For example we think that there exists a relation between level of approximation ε used with approximate reducts and the parameter α used to control the order of objects and attributes within a permutation used for calculating bireducts. We also look forward experimenting with various types of bireducts and ways of constructing ensembles and testing such classifiers on a benchmark data.

ACKNOWLEDGMENTS

This work was partially supported by the Polish National Science Centre grant 2011/01/B/ST6/03867.

REFERENCES

[1] D. Ślęzak and A. Janusz, "Ensembles of Bireducts: Towards Robust Classification and Simple Representation," in *Proc. of FGIT 2011*, ser. LNCS, vol. 7105, 2011, pp. 64–77.

[2] D. Ślęzak and S. Widz, "Rough-Set-Inspired Feature Subset Selection, Classifier Construction, and Rule Aggregation," in *Rough Sets and Knowledge Technology (RSKT) - 6th International Conference, RSKT 2011, Banff, Canada, October 9-12, 2011. Proceedings*, ser. Lecture Notes in Computer Science, vol. 6954. Springer, 2011, pp. 81–88.

[3] H. Liu and H. Motoda, Eds., *Computational Methods of Feature Selection*. Chapman & Hall/CRC, 2008.

[4] Z. Pawlak and A. Skowron, "Rudiments of Rough Sets," *Information Sciences*, vol. 177, no. 1, pp. 3–27, 2007.

[5] D. Ślęzak, "Rough Sets and Functional Dependencies in Data: Foundations of Association Reducts," *LNCS Transactions on Computational Science*, vol. V, pp. 182–205, 2009.

[6] S. Widz and D. Ślęzak, "Approximation Degrees in Decision Reduct-based MRI Segmentation," in *FBIT*. IEEE, 2007, pp. 431–436.

[7] A. Janusz and S. Stawicki, "Applications of Approximate Reducts to the Feature Selection Problem," in *RSKT*. Springer, 2011, pp. 45–50.

[8] T. G. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[9] L. I. Kuncheva and C. J. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," *Mach. Learn.*, vol. 51, no. 2, pp. 181–207, May 2003. [Online]. Available: http://dx.doi.org/10.1023/A:1022859003006

[10] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *Journal of Information Fusion*, vol. 6, pp. 5–20, 2005.

[11] B. Mirkin, *Mathematical Classification and Clustering*. Kluwer, 1996.

[12] D. Ślęzak, "Normalized Decision Functions and Measures for Inconsistent Decision Tables Analysis," *Fundamenta Informaticae*, vol. 44, no. 3, pp. 291–319, 2000.

[13] J. Derrac, C. Cornelis, S. García, and F. Herrera, "Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection," *Inf. Sci.*, vol. 186, no. 1, pp. 73–92, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2011.09.027

[14] M. J. Moshkov, M. Piliszczuk, and B. Zielosko, *Partial Covers, Reducts and Decision Rules in Rough Sets - Theory and Applications*, ser. Studies in Computational Intelligence. Springer, 2008, vol. 145.

[15] H. Banka and S. Mitra, "Feature selection, classification and rule generation using rough sets," in *Rough Sets: Selected Methods and Applications in Management and Engineering*, ser. Advanced Information and Knowledge Processing, G. Peters, P. Lingras, D. Ślęzak, and Y. Yao, Eds. Springer London, 2012, pp. 51–76.

[16] D. Ślęzak and W. Ziarko, "The Investigation of the Bayesian Rough Set Model," *International Journal of Approximate Reasoning*, vol. 40, no. 1-2, pp. 81–91, 2005.

[17] J. Bazan, H. Nguyen, S. Nguyen, P. Synak, and J. Wróblewski, "Rough Set Algorithms in Classification Problem," in *Rough Set Methods and Applications*, ser. Studies in Fuzziness and Soft Computing 56, L. Polkowski, S. Tsumoto, and T. Lin, Eds. Physica Verlag, 2000, pp. 49–88.

[18] S. Widz and D. Ślęzak, "Rough Set Based Decision Support – Models Easy to Interpret," in *Selected Methods and Applications of Rough Sets in Management and Engineering*, ser. Advanced Information and Knowledge Processing, G. Peters, P. Lingras, D. Ślęzak, and Y. Yao, Eds. Springer, 2012, pp. 95–112.

[19] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *ICML*, L. Saitta, Ed. Morgan Kaufmann, 1996, pp. 148–156.