Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 295–302

ISBN 978-83-60810-51-4

Utilization of Attribute Clustering Methods for Scalable Computation of Reducts from High-Dimensional Data

Andrzej Janusz* and Dominik Ślęzak*†
*Institute of Mathematics, University of Warsaw
ul. Banacha 2, 02-097 Warsaw, Poland
†Infobright Inc.
ul. Krzywickiego 34, lok. 219, 02-078 Warsaw, Poland
andrzejanusz@gmail.com, slezak@infobright.com

***Abstract*—We investigate methods for attribute clustering and their possible applications to a task of computation of decision reducts from information systems. We focus on high-dimensional data sets, for which the problem of selecting attributes that constitute a reduct can be extremely computationally intensive. We apply an attribute clustering method to facilitate construction of reducts from microarray data. Our experiments confirm that by proper grouping of similar, in some sense replaceable attributes it is possible to significantly decrease a computation time, as well as increase a quality of resulting reducts (i.e. decrease their average size).**

***Keywords*-attribute selection, attribute reduction; attribute clustering; high-dimensional data; microarray data; scalable reducts computation methods;**

I. Introduction

IN MANY applications, information about objects from a considered universe has to be reduced. It can be required in order to limit resources needed by algorithms analyzing the data or to prevent crippling their performance by noisy or irrelevant attributes [1], [2]. Computation of information and decision reducts has been widely discussed in literature related to data analysis and knowledge discovery, as a rough-set-based approach to attribute reduction. In particular, its practical significance for tasks such as attribute selection, rule induction and data visualization is unquestionable [3], [4].

There is plenty of literature showing how to apply rough-set-based methods to the microarray data analysis [5], [6], [7], [8]. In [9], [10], it was discussed that, given such a huge amount of attributes in microarray data sets, it is better to combine the standard computation mechanisms with some elements of attribute clustering. This paper aims at experimental verification of these ideas by combining the rough-set-based methods of attribute reduction with the rough-set-inspired methods for attribute clustering.

Although the paper focuses on analysis of microarray data, the discussed approach to dimensionality reduction can be applied to a much wider spectrum of high-dimensional data types. For example, similar techniques can be used to find concise conceptual representations of texts for the purpose of multi-label topical classification [11] or meaningful labelling of clustering results [12]. Both of those tasks require working on texts represented by a potentially large number of interdependent domain concepts. We go back to this topic in Section VII, while sketching the areas of our future research.

The paper is organized as follows: Section II recalls some basic notions of rough-set-based attribute reduction. Section III outlines our intuition behind combining attribute reduction with attribute clustering. Sections IV and V report our experimental framework for utilizing gene clustering methods for computation of reducts and the obtained results, respectively. Section VI overviews our preliminary research on reusing gene clustering results in analysis of different data sets. Finally, as already mentioned, Section VII concludes the paper with some future research directions.

II. Rough-Set-Based Attribute Reduction

In the rough set theory, by a reduct we usually mean a compact yet informative subset of available attributes. In this section, we outline some basic types of reducts.

Definition 1 (Decision reduct)**.**
Let $\mathbb{S}_d = (U, A, d)$ be a decision table with a decision attribute d indicating belongingness of objects to investigated concepts. A subset of attributes $DR \subseteq A$ will be called a decision reduct, iff the following conditions are met:

1) *For any pair $u, u' \in U$ of objects belonging to different decision classes (i.e. $d(u) \neq d(u')$), if u and u' are discerned by A (i.e. there is $a \in A$ such that $a(u) \neq a(u')$), then they are also discerned by DR.*
2) *There is no proper subset $DR' \subsetneq DR$, for which the first condition holds.*

A decision reduct can be interpreted as a set of attributes that are sufficient to discriminate all objects from different decision classes. At the same time this set has to be minimal, in a sense that no further attributes can be removed from DR without loosing the discernibility property. For example, $\{a_3, a_5\}$ and $\{a_3, a_6\}$ are decision reducts of the decision table $\mathbb{S}_d$ from Table I. The first condition in Definition 1 is often replaced

TABLE I
AN EXEMPLARY DECISION TABLE $\mathbb{S}_d$ WITH A BINARY DECISION.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	d
u_1	1	2	2	0	0	1	0	1	1
u_2	0	1	1	1	1	0	1	0	1
u_3	1	2	0	1	0	2	1	0	1
u_4	0	1	0	0	1	0	0	1	0
u_5	2	0	1	0	2	1	0	0	1
u_6	1	0	2	0	2	0	0	2	0
u_7	0	1	1	2	0	2	1	0	1
u_8	0	0	0	2	1	1	1	1	0
u_9	2	1	0	0	1	1	0	0	0

by some other requirements for preserving information about decision while reducing attributes [13], [14]. In this paper, for simplicity, we restrict ourselves to the above discernibility-based criterion, which is very well-documented in the rough set literature [15], [16].

There are described many algorithms for attribute reduction, utilizing various greedy or randomized search approaches [7], [17]. Most of them refer to the search of optimal (shortest, generating minimum number of rules, etc.) decision reducts or some larger ensembles of decision reducts that constitute efficient classification models. However, most of algorithms can be also adapted to search for other types of reducts, e.g., those aimed at the unsupervised analysis of data sets with no predefined decisions.

Definition 2 (Information reduct)**.**
Let $\mathbb{S} = (U, A)$ be an information system. A subset of attributes $IR \subseteq A$ will be called an information reduct, iff the following conditions are met:

1) *For any pair $u, u' \in U$ of objects, if u and u' are discerned by A, then they are also discerned by IR.*
2) *There is no proper subset $IR' \subsetneq IR$, for which the first condition holds.*

Definition 3 (Association reduct)**.**
Let $\mathbb{S} = (U, A)$ be an information system. A pair of subsets of attributes $AR = (L, R)$, where $L \cap R = \emptyset$, will be called an association reduct, iff the following conditions are met:

1) *For any pair $u, u' \in U$ of objects, if u and u' are discerned by R, then they are also discerned by L.*
2) *There is neither proper subset $L' \subsetneq L$ nor proper superset $R' \supsetneq R$, for which the first condition holds.*

Association reducts were studied in [10], [14] due to their usefulness for unsupervised learning over huge amounts of attributes, where there is a relatively low chance to obtain reasonable information reducts. In the case of all the above types of reducts, we can also consider their approximate versions, which are especially useful for noisy data sets [13], [17]. For instance, we can require that only some percentage of pairs of objects satisfies the first conditions in Definitions 1, 2, 3. Also, we may extend the discernibility notion towards the criteria of a sufficient dissimilarity or a discernibility in a degree, which are useful in the case of numeric data.

The above discussion shows that appropriately extended classical rough set notions can be successfully applied as an attribute selection and reduction framework for the analysis of large and complex data sets, such as microarray data (large amounts of attributes with potentially noisy, numeric values). However, there is yet another possibility of scaling with regard to the amounts of attributes. The basic idea is to search for clusters of attributes that can potentially replace each other while learning optimal reducts from data. In the next sections, we investigate this opportunity for the case of decision reducts (Definition 1). An analogous framework may be also considered for other types of reducts and other criteria of their construction.

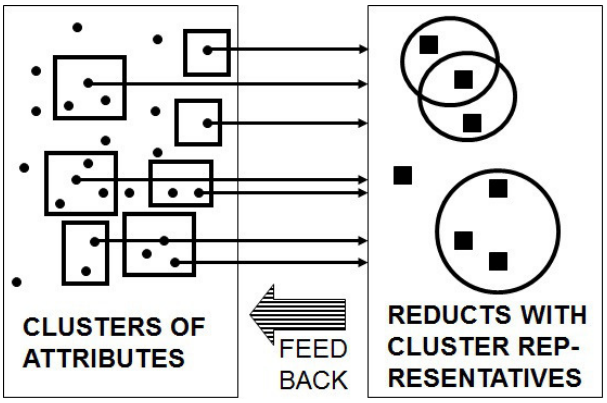

Fig. 1. A hybrid approach combining attribute clustering and reduction.

III. REDUCT-ORIENTED ATTRIBUTE CLUSTERING

From a perspective of the analysis of microarray data, the ideas presented in this section can be regarded as an example of gene clustering [18], [19]. In [9], we reported that the gene clustering outcomes may meet the domain experts' expectations to more extent when they are based on information-theoretic measures, rather than on standard numeric and rank-based correlations. In other words, interpreting genes as attributes with some approximate dependencies between them may bring better results than treating them simply as numeric vectors. In [10], we suggested that attribute clustering can be conducted also by means of dissimilarity functions based on discernibility between objects, utilized as a form of measuring degrees of functional dependencies between attributes. We also proposed a mechanism illustrated by Fig. 1, where reducts could be searched in a data table consisting only of the previously computed clusters' representatives, with their occurrence in reducts used as a feedback for splitting and merging the corresponding clusters. For instance, clusters with their representatives reoccurring more often in reducts could be considered for further split. On the other hand, clusters with their representatives not occurring in reducts could be eliminated or merged with the others.

In order to make sure that approaches such as the one proposed in Fig. 1 work efficiently, we need to guarantee that attributes falling more likely into the same clusters will

also be more frequently *replaceable* within reducts. From this perspective, it is reasonable to use analogous criteria for preserving information about decision while reducing attributes and measuring distances between them. As an example, let us compare attributes a_5 and a_6 in Table I. Note that in the case of most of pairs of objects, a_5 discerns them, iff a_6 does. This may indicate that there are relatively many pairs of reducts of the form $B \cup \{a_5\}$ and $B \cup \{a_6\}$, $B \subseteq A \setminus \{a_5, a_6\}$. Reducts $\{a_3, a_5\}$ and $\{a_3, a_6\}$ are an illustration of the replaceability understood in this way.

The replaceability of attributes in the context of the discernibility can be easily noticed by studying a dendrogram generated by a hierarchical clustering algorithm. An example of such a tree generated for the decision table from Table I is presented in Fig. 2. As expected, the attributes a_5 and a_6 are merged into a single cluster as the second pair.

The methods of attribute reduction and attribute grouping can be put together in many different ways. As an example, in [20] it is noted that so called signatures (irreducible subsets of genes providing enough information about probabilities of specific types of cancer – the reader may notice an interesting correspondence of this notion to a probabilistic version of a decision reduct [13]) can contain genes (attributes) that are interchangeable with the others because of data correlations or multiple explanations of some biomedical phenomena. Moreover, such an interchangeability can be observed not only for single elements – there may be a larger subset of genes / attributes replaceable with another one, leading to another signature / reduct.

In the next sections, we operate with relatively straightforward dissimilarity functions based on the comparison of attributes' abilities to discern important pairs of objects. It would be possible to extend these functions in order to measure dissimilarities between subsets of attributes. However, let us focus on individual attributes.

The first function we considered, called a *direct* discernibility function, is a ratio between a number of pairs of objects from different decision classes that are discerned by *exactly one* attribute to a number of such objects discerned by *at least one* of the compared attributes. It can be written down in a way that emphasizes its analogy to some standard measures used in data clustering [21], [22]. Namely, $direct(a,b) =$

$$= 1 - \frac{|\{(u,u'):d(u)\neq d(u')\wedge a(u)\neq a(u')\wedge b(u)\neq b(u')\}|}{|\{(u,u'):d(u)\neq d(u')\wedge (a(u)\neq a(u')\vee b(u)\neq b(u'))\}|}$$

The second of the considered functions, called a *relative* discernibility function, takes into account the fact that some pairs of objects belonging to different decision classes can be harder to discern than the others. The following formula should be regarded as one of many possible mathematical formulations of this basic intuition. Namely, $relative(a,b) =$

$$= 1 - \frac{\sum_{u,u':d(u)\neq d(u')\wedge a(u)\neq a(u')\wedge b(u)\neq b(u')} \alpha(u,u')}{\sum_{u,u':d(u)\neq d(u')\wedge (a(u)\neq a(u')\vee b(u)\neq b(u'))} \alpha(u,u')}$$

where $\alpha(u,u') = |\{c \in A : c(u) = c(u')\}|/|A|$ plays a role of a weighting factor. The pairs of objects, which are generally more difficult to discern by the attributes in A, are considered to be relatively more important when computing dissimilarity between a and b. For instance, if a is the only attribute able to distinguish between u and u' from different decision classes, then its corresponding $\alpha(u,u')$ occurs only in the nominator of the above formula and causes relatively high decrease of the value of $relative(a,b)$. On the other hand, if some pair of objects is discerned by all attributes, it has no impact on the value of $relative(a,b)$.

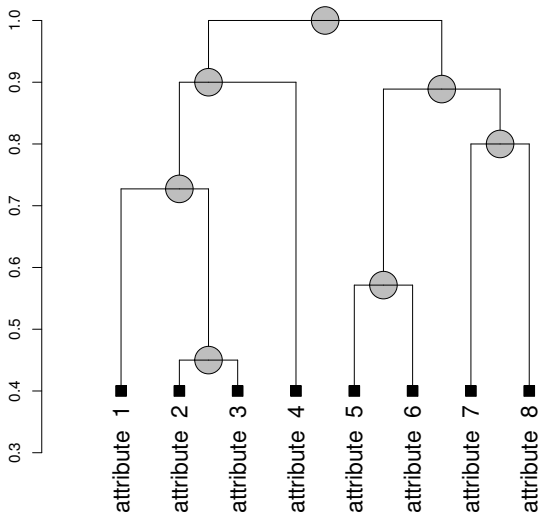

Fig. 2. An attribute clustering tree for the decision table from Table I obtained by applying the agglomerative nesting algorithm in combination with the direct discernibility dissimilarity function.

The direct and relative discernibility functions will be used in experiments described in Sections IV and V. We will verify usefulness of the above intuitions in a slightly different scenario than the one originally proposed in [10] and depicted by Fig. 1. However, the fundamental idea will remain the same – by using appropriate attribute clusters, we can avoid unnecessary attempts to co-locate replaceable or interchangeable attributes in the same reducts.

IV. Framework for Experimental Validation

We conducted a series of experiments to verify usefulness of the discernibility-based attribute clustering for scalable computation of decision reducts. We wanted to find answers to two main questions. The first one was whether the attribute grouping can speed up searching for reducts. The second question was related to a quality of reducts generated using discernibility-based clustering – we wanted to check if such reducts are more concise. The minimal number of attributes is not the only possible optimization criterion for decision reducts [13], [17]. However, it is indeed the most straightforward idea to rely on minimal reducts in order to clearly visualize the data dependencies.

In the experiments, we use a microarray data set from the RSCTC'2010 data mining competition [23]. Microarrays are

usually described by many thousands of attributes whose values correspond to expression levels of genes. The considered data set is related to the investigation of a chronic hepatitis C virus role in the pathogenesis of HCV-associated hepatocellular carcinoma. It contains data on 124 tissue samples described by $22,277$ numeric attributes (genes). It was obtained from the ArrayExpress repository [24] (data set accession number: E-GEOD-14323). The gene expression levels in this data set were obtained using Affymetrix GeneChip Human Genome U133A 2.0 microarrays.

We preprocessed the data by discretizing attributes using an unsupervised method. Every expression level value of a given gene was replaced by one of the three labels: *over_expressed*, *normal* or *under_expressed*. A label for an attribute a and a sample u is decided as follows:

$$\overline{a}(u) = \begin{cases} \textit{over\_expressed} & \text{if } a(u) > mean_a + sd_a, \\ \textit{under\_expressed} & \text{if } a(u) < mean_a - sd_a, \\ \textit{normal} & \text{otherwise.} \end{cases}$$

where $mean_a$ and sd_a denote the mean and the standard deviation of expression level values of the gene a in the whole data set. One might also apply more intelligent discretization techniques [15], [7] or utilize some rough-set-based approaches that do not require discretization at all [25], [26]. However, we proceed with the above-proposed discretization strategy for the sake of simplicity.

In order to assess an impact of different attribute clustering methods on computation of reducts we clustered the genes using several techniques. We combined the discernibility-based functions described in the previous section with the agglomerative nesting (*agnes*), which is a hierarchical grouping algorithm [21], [22]. We compared it with *kmeans* and *agnes* algorithms working on dissimilarities computed using Euclidean distance on non-discretized data. As a reference, we took results obtained for a random clustering, which is equivalent to no clustering at all. We also checked the worst-case scenario, in which the attributes are grouped so that the most dissimilar genes (according to the *direct* discernibility function) are in the same clusters.

In each experiment, we generated 100 decision reducts for all the compared clustering methods. For the reduct computation we used Algorithm 1, which is a modified version of the permutation-based attribute reduction framework considered in [14], [17]. The permutations for each run of the algorithm were generated based on the clusterings corresponding to the tested grouping methods. Algorithm 2 explains the permutation construction process. In practice, there is no need to pre-generate a permutation for the reduct computation since it might be an integral part of the algorithm. However, we explicitly generated the permutations for the sake of reproducibility of the results. In the next section, we also mention some relatively easier greedy techniques. All the algorithms were implemented in R System [27].

Algorithm 1: A permutation-based attribute reduction.

Input: a decision table $\mathbb{S}_d$, an ordered list of attributes $perm_A = [a_1, ..., a_n]$
Output: a decision reduct DR of $\mathbb{S}_d$

```
DR = ∅;
i = 1;
while DR does not meet condition 1 in Definition 1 do
    a_i = perm_A[i];
    DR = DR ∪ a_i;
    i = i + 1;
end
for j = i to 1 do
    if DR \ a_j discerns all objects with different decisions from S_d then
        DR = DR \ a_j;
    end
end
return DR;
```

Algorithm 2: A cluster-based permutation generator.

Input: a clustering of attributes $CL_A = (C_1, ..., C_k)$
Output: an ordered list $perm_A = [a_1, ..., a_n]$

```
perm_A = []                (an empty list);
perm_k = [p_1, ..., p_k]   (a permutation of numbers 1 to k);
i = 1;
while length(perm_A) ≠ n do
    if |C_{p_i}| > 0 then
        p_i = perm_k[i];
        select at random an attribute a from C_{p_i};
        perm_A = [perm_A, a];
        C_{p_i} = C_{p_i} \ a;
    end
    i = i + 1;
    if i > k then
        i = 1;
    end
end
return perm_A;
```

V. Obtained Experimental Results

Table II summarizes computation times. For each clustering method the mean and standard deviation of 20 independent repetitions of the experiment are given. The results show the advantage of using the *direct* discernibility function in combination with a hierarchical clustering algorithm to speed up generation of decision reducts. Times obtained by this method are significantly lower than those of all other approaches. The significance was measured using t-test [28] and the p-values obtained at 0.95 confidence level were all lower than 10^{-10}. For instance, the times obtained by this method when grouping was made into 1000 clusters are on average lower by 34% than the corresponding times for the random method. Moreover, robustness of results is confirmed by their stability with regard to a number of considered clusters.

The results obtained for the *relative* discernibility func-

tion may be regarded as disappointing. The tested weighting schema seems to degrade performance of the reduct computation algorithm, especially when a low number of gene clusters is considered. Explanation of this behavior will be definitely in a scope of our future research. The experiments show that distinguishing between the cases that are easier and more difficult to discern may not be necessary. On the other hand, a better-adjusted mathematical formula for such distinguishing may lead to more promising outcomes.

The results from Table II obtained for the two Euclidean distance-based clusterings also show a clear advantage of using hierarchical methods for grouping genes in microarray data. The times for the combination of *kmeans* clustering with Euclidean metric can not be regarded as statistically different from results of random clusterings, whereas the grouping into 10 clusters, made using the *agnes* algorithm resulted in significantly lower reduct construction times.

For each clustering method we also measured an average length of the generated reducts. This statistic reflects a quality of reducts, both by means of data-based knowledge representation and ability to construct efficient classification models. These results are displayed in Table III.

The standard deviations given in the above table are not computed directly from the lengths of reducts, but from the average lengths of 100 reducts in each of the 20 experiment runs. This explains such low values of this statistic.

Our method significantly outperformed other approaches also in terms of the reduct length. As before, the significance was checked using t-test. On average, decision reducts generated by using the hierarchical clustering based on the direct discernibility function are shorter than those computed from the random clusterings by nearly 1.5 gene. They were also shorter than the reducts computed for the *agnes* algorithm and working on the Euclidean distances by over 0.5 gene. It confirms that a proper attribute clustering increases efficiency of the reduct computation methods.

In order to further assess usefulness of the considered attribute clustering methods for computation of decision reducts, we compared the permutation-based reducts obtained by utilization of groupings of genes with a reduct computed using a simple deterministic greedy heuristic, which starts with an empty set and iteratively adds the most promising attributes until the decision determination criterion is satisfied [25], [29]. Out of many possibilities, we selected the GiniGain measure to evaluate quality of attributes during such an iterative construction of the solution. At the first stage of this experiment, we did not use any form of attribute clustering. For the investigated data, a reduct was constructed in 544 seconds and it contained 6 attributes. The experiment was conducted on the same machine as before. The applied heuristic was implemented in R as well.

The construction of a greedy reduct is a couple of orders of magnitude more time-consuming than the construction of a single reduct by the permutation-based algorithm. This is because at each iterative step we need to examine (almost) all remaining attributes against the data, which is a huge overhead in the case of microarray data sets. On the other hand, for the case of the GiniGain measure and this particular data set, the greedy approach leads to a significantly shorter solution. In order to confirm it, we checked lengths of the shortest reducts computed in each of the 20 repetitions of the previous experiment. The average results for each clustering method are presented in Table IV.

The above observation motivated us to measure an impact of attribute grouping on a computation time of greedy reducts. We introduced constraints to the greedy algorithm which allow to select only a single attribute from each cluster (the selection

TABLE II
AVERAGE COMPUTATION TIMES OF 100 DECISION REDUCTS BASED ON PERMUTATIONS GENERATED FROM DIFFERENT CLUSTERINGS.

clustering method	10 clusters	100 clusters	1000 clusters
agens & direct	3.536 ± 0.112	3.151 ± 0.097	3.015 ± 0.117
agens & relative	4.680 ± 0.156	4.164 ± 0.161	3.705 ± 0.134
agens & Euclidean	3.965 ± 0.158	4.430 ± 0.251	4.839 ± 0.199
kmeans & Euclidean	4.872 ± 0.239	4.434 ± 0.229	4.545 ± 0.148
random	4.597 ± 0.155	4.665 ± 0.190	4.543 ± 0.147
worst	5.485 ± 0.219	9.901 ± 0.753	11.929 ± 0.628

TABLE III
AVERAGE LENGTHS OF 100 REDUCTS COMPUTED FOR DIFFERENT CLUSTERING METHODS.

clustering method	10 clusters	100 clusters	1000 clusters
agens & direct	11.209 ± 0.099	11.095 ± 0.087	11.103 ± 0.093
agens & relative	12.102 ± 0.132	11.790 ± 0.134	11.638 ± 0.114
agens & Euclidean	11.709 ± 0.123	11.860 ± 0.118	12.198 ± 0.114
kmeans & Euclidean	12.590 ± 0.089	12.228 ± 0.069	12.283 ± 0.130
random	12.519 ± 0.127	12.470 ± 0.092	12.471 ± 0.128
worst	12.731 ± 0.133	14.800 ± 0.159	15.624 ± 0.180

TABLE IV
AVERAGE MINIMAL LENGTHS AMONG 100 REDUCTS COMPUTED FOR DIFFERENT CLUSTERING METHODS.

clustering method	10 clusters	100 clusters	1000 clusters
agens & direct	8.900 ± 0.307	8.950 ± 0.223	9.200 ± 0.410
agens & relative	9.600 ± 0.502	9.250 ± 0.444	9.550 ± 0.510
agens & Euclidean	9.500 ± 0.512	9.250 ± 0.444	9.600 ± 0.502
kmeans & Euclidean	10.000 ± 0.458	9.650 ± 0.489	9.600 ± 0.502
random	9.85 ± 0.489	9.900 ± 0.447	10.000 ± 0.324
worst	9.950 ± 0.394	10.900 ± 0.640	11.550 ± 0.604

itself was still done in the greedy fashion). This modification resulted in a significant decrease of time needed for computation of a single greedy reduct – it took 392 seconds when the grouping was done into 10 clusters. The size of a reduct obtained in this way was 6, which is equal to the classical case. However, those two greedy reducts differed on 4 out of 6 attributes. In particular, this shows that searching for a single decision reduct provides highly incomplete knowledge about dependencies in the data, especially for such huge amounts of attributes. Hence, the approaches aimed at extraction of larger families of reducts should be preferred [17], [30].

In the future we would like to investigate a possibility of combining the greedy and permutation-based heuristics to facilitate fast computation of representative ensembles of short decision reducts. Certainly, the choice of parameters responsible for generation of permutations, the greedy heuristic measures, or the attribute reduction criteria may have a significant influence on the results. However, in any of such scenarios it is expected that attribute clustering can improve computations and interpretation of the results. In particular, in the case of microarray data sets it may be far more intuitive to work with reducts understood as subsets of genes representing some larger clusters.

VI. Reuse of Gene Clustering Results for Other Microarray Data Sets

In our research on applications of gene clustering for scalable analysis of microarray data we were also interested in a possibility of reusing knowledge obtained from one data set for analysis of different data. This problem is especially important in the context of the microarray research, since due to expensiveness of a single microarray experiment the number of cases available in a typical microarray data set is small. This characteristic data feature often makes it difficult to perform meaningful analysis on the microarray data.

On the other hand, researchers attempt to come up with standards that aim at unification of acquisition, preprocessing and description of microarray data. The ultimate goal of this effort is to aid the reuse of knowledge and to facilitate creation of larger microarray data sets by hybridization of multiple smaller set that correspond to experiments conducted in different research centres. One example of such an initiative is the MIAME (Minimum Information About a Microarray Experiment) standard [31].

Many microarray experiments for different medical problems are conducted using the same types of microarray chips and with the same sets of reagents. As a consequence, data samples corresponding to unrelated medical issues can be represented by attributes with exactly the same semantics. A question arises whether it is possible to facilitate computation of decision reducts of one microarray data set using a clustering of genes obtained from a different data. Unfortunately, the reuse of knowledge for data related to different problems is usually not that simple because of differences in the applied acquisition procedures and data preprocessing methods.

TABLE V
Average computation times of 100 decision reducts for acute lymphoblastic leukemia data using different clusterings obtained for the hepatitis C data.

clustering method	10 clusters	100 clusters	1000 clusters
agens & direct	3.751 ± 0.086	3.726 ± 0.118	3.786 ± 0.094
agens & relative	3.642 ± 0.114	3.746 ± 0.086	3.766 ± 0.074
agens & Euclidean	3.745 ± 0.108	3.730 ± 0.091	3.699 ± 0.099
kmeans & Euclidean	3.775 ± 0.096	3.699 ± 0.077	3.658 ± 0.072
random	3.661 ± 0.104	3.700 ± 0.088	3.7155 ± 0.070
worst	3.845 ± 0.106	3.668 ± 0.098	3.671 ± 0.081

TABLE VI
Average lengths of 100 reducts computed for acute lymphoblastic leukemia data using different clusterings obtained for the hepatitis C data set.

clustering method	10 clusters	100 clusters	1000 clusters
agens & direct	12.534 ± 0.096	12.479 ± 0.100	12.501 ± 0.093
agens & relative	12.420 ± 0.097	12.467 ± 0.087	12.513 ± 0.081
agens & Euclidean	12.389 ± 0.093	12.410 ± 0.118	12.368 ± 0.084
kmeans & Euclidean	12.477 ± 0.108	12.377 ± 0.074	12.314 ± 0.064
random	12.402 ± 0.083	12.428 ± 0.108	12.391 ± 0.104
worst	12.744 ± 0.084	12.281 ± 0.088	12.203 ± 0.093

We designed the following experiment along the lines of the above discussion. First, we searched the ArrayExpress repository looking for a microarray data set in which gene expressions were measured using the same microarray chips as in the hepatitis C data set discussed in Section IV. We selected a data set related to recognition of acute lymphoblastic leukemia (ALL) genetic subtypes (experiment accession number E-GEOD-13425). This data set has been used in the already-mentioned RSCTC'2010 data mining competition as well. It contains data on 190 blood samples annotated with expression levels of the same $22,277$ genes as before. We discretized the data and computed decision reducts using the same methodology as described in Section IV. However, to generate the permutations we utilized the clusterings previously computed for the hepatitis C data set. The results in terms of a reduct average length and computation speed is shown in Tables V and VI.

Computation times and reduct lengths obtained by application of clusterings from the hepatitis C data set are not statistically different from random outcomes. Apart from highlighting the need for unified preparation methods for microarray data, this result confirms the importance of considering a specific decision problem as a context when forming groups of genes. It explains why the attribute dissimilarity functions not refer-

ring to a given decision task may perform worse than those taking decision attributes into account.

In the future, we will investigate whether such a reuse of knowledge can bring benefit for computation of association reducts, where the decision attribute is not fixed. We would also like to verify if the gene clustering can be improved by utilization of a decision table hybridized from several different microarray data sets. However, one needs to remember that even though different microarray data sets may be created using the same microarray chips, the already mentioned differences in a processes of acquiring their data rows may cause significant differences in statistical characteristics of the corresponding attributes. For example, in two series of microarray experiments, samples may be collected from different kinds of tissues and as a result, the same genes may have different neutral expression levels.

Reusing knowledge discovered by attribute clustering seems appealing also in a context of text mining and information retrieval [32]. Intuitively, when analyzing large document corpora related to the same domain (e.g. articles from two different biomedical journals), it is expected that the same terms would have the same or similar sets of possible meanings, and that occurrence frequencies of particular meanings would be similar as well. Knowing this, one may think of reusing term clustering results obtained for a single corpus, in order to facilitate analysis of other corpora from that domain. In the next section, we refer to one of our research projects aiming at development of a text analysis system, where such a reuse of the grouping results would be applicable [33], [34].

VII. Concluding Remarks

We proposed a new approach to attribute clustering and its application to a task of computation of optimal decision reducts from data sets with a large number of attributes. We showed that by utilization of clustering results it is possible to significantly speed up the search for decision reducts and that the resulting reducts tend to be shorter than those obtained without the clustering. We also proposed a discernibility-based attribute similarity measure, which is useful for identifying groups of attributes that are likely to be interchangeable in many decision reducts.

In the future, we intend to combine our methods with other knowledge discovery approaches that involve attribute grouping and selection [9], [20]. One may also consider an idea of full integration of the algorithms for attribute clustering and selection, so they can provide feedback to each other within the same learning process. Such a new process may be performed separately for particular microarray data sets or, as suggested in Section VI, over their larger unions.

Also, as already noted, the integration of the attribute clustering and selection procedures may not only bring significant performance improvements but also provide a new meaning with regard to the attribute selection outcomes. Namely, instead of operating with subsets of individual attributes chosen from thousands of genes, it may be truly better to deal with subsets of representatives selected from much more robust clusters of pair-wise replaceable attributes.

Although our experiments referred primarily to microarray data sets, we plan to use our methods also for the analysis of a broader class of biomedical data sources, such as medical texts and clinical data [6], [35], where there is a common need for a scalability of the attribute selection techniques.

In particular, we intend to utilize the developed methods in the system already mentioned in the end of Section VI, in which they can be applied to different data types, such as corpora of textual documents. In this context, both attribute clustering and reduction, can play a significant role for a search engine, whose aim is to facilitate retrieval, synthesis and visualization of semantically meaningful information from scientific document repositories. The observations on a possibility of reusing knowledge related to textual data was actually among the factors, which motivated us in designing architecture of such an engine.

Acknowledgments

This work was partially supported by the Polish National Science Centre grant 2011/01/B/ST6/03867 and by the National Centre for Research and Development (NCBiR) grant SP/I/1/77065/10, the strategic scientific research and experimental development program: “Interdisciplinary System for Interactive Scientific and Scientific-Technical Information”.

References

[1] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, Eds., *Feature Extraction: Foundations and Applications*, ser. Studies in Fuzziness and Soft Computing. Springer, 2006, vol. 207.

[2] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[3] Z. Pawlak, *Rough Sets – Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, 1991.

[4] R. W. Świniarski and A. Skowron, “Rough Set Methods in Feature Selection and Recognition,” *Pattern Recognition Letters*, vol. 24, no. 6, pp. 833–849, 2003.

[5] J. Fang and J. W. Grzymała-Busse, “Leukemia Prediction from Gene Expression Data – A Rough Set Approach,” in *ICAISC*, ser. LNCS, vol. 4029. Springer, 2006, pp. 899–908.

[6] H. Midelfart, H. J. Komorowski, K. Nørsett, F. Yadetie, A. K. Sandvik, and A. Lægreid, “Learning Rough Set Classifiers from Gene Expressions and Clinical Data,” *Fundamenta Informaticae*, vol. 53, no. 2, pp. 155–183, 2002.

[7] A. Janusz and S. Stawicki, “Applications of Approximate Reducts to the Feature Selection Problem,” in *RSKT*, ser. LNCS, vol. 6954. Springer, 2011, pp. 45–50.

[8] D. Ślęzak and J. Wróblewski, “Roughfication of Numeric Decision Tables: The Case Study of Gene Expression Data,” in *Rough Sets and Knowledge Technology, Second International Conference, RSKT 2007, Toronto, Canada, May 14-16, 2007, Proceedings*, ser. Lecture Notes in Computer Science, J. Yao, P. Lingras, W.-Z. Wu, M. S. Szczuka, N. Cercone, and D. Ślęzak, Eds., vol. 4481. Springer, 2007, pp. 316–323.

[9] A. Gruźdź, A. Ihnatowicz, and D. Ślęzak, “Interactive Gene Clustering – A Case Study of Breast Cancer Microarray Data,” *Information Systems Frontiers*, vol. 8, no. 1, pp. 21–27, 2006.

[10] D. Ślęzak, “Rough Sets and Few-Objects-Many-Attributes Problem: The Case Study of Analysis of Gene Expression Data Sets,” in *FBIT*. IEEE, 2007, pp. 437–442.

[11] A. Janusz, H. S. Nguyen, D. Ślęzak, S. Stawicki, and A. Krasuski, “JRS’2012 Data Mining Competition: Topical Classification of Biomedical Research Papers,” in *Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2012), Chengdu, China, August 17-20, 2012*, ser. LNAI, J.T. Yao et al., Ed., vol. 7413. Springer, Heidelberg, 2012, pp. 156–165.

[12] S. Nguyen, W. Świeboda, and G. Jaśkiewicz, "Extended Document Representation for Search Result Clustering," in *Intelligent Tools for Building a Scientific Information Platform*, ser. Studies in Computational Intelligence, R. Bembenik, L. Skonieczny, H. Rybinski, and M. Niezgodka, Eds. Springer Berlin / Heidelberg, 2012, vol. 390, pp. 77–95.
[13] D. Ślęzak, "Approximate Entropy Reducts," *Fundamenta Informaticae*, vol. 53, no. 3-4, pp. 365–390, 2002.
[14] D. Ślęzak, "Rough Sets and Functional Dependencies in Data: Foundations of Association Reducts," *LNCS Transactions on Computational Science V*, vol. 5540, pp. 182–205, 2009.
[15] J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, and J. Wróblewski, "Rough Set Algorithms in Classification Problem," in *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems*, ser. Studies in Fuzziness and Soft Computing, L. Polkowski, S. Tsumoto, and T. Y. Lin, Eds. Physica-Verlag, 2000, vol. 56, pp. 49–88.
[16] H. S. Nguyen, "Approximate Boolean Reasoning: Foundations and Applications in Data Mining," *LNCS Transactions on Rough Sets V*, vol. 4100, pp. 334–506, 2006.
[17] J. Wróblewski, "Ensembles of Classifiers Based on Approximate Reducts," *Fundamenta Informaticae*, vol. 47, no. 3-4, pp. 351–360, 2001.
[18] P. Baldi and G. W. Hatfield, *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. Cambridge University Press, 2002.
[19] B. A. Mckinney, D. M. Reif, M. D. Ritchie, and J. H. Moore, "Machine Learning for Detecting Gene-Gene Interactions: A Review," *Applied Bioinformatics*, vol. 5, no. 2, pp. 77–88, 2006.
[20] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys, "Robust Biomarker Identification for Cancer Diagnosis with Ensemble Feature Selection Methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.
[21] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.
[22] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Interscience, 1990.
[23] M. Wojnarski, A. Janusz, H. S. Nguyen, J. G. Bazan, C. Luo, Z. Chen, F. Hu, G. Wang, L. Guan, and H. Luo, "RSCTC'2010 Discovery Challenge: Mining DNA Microarray Data for Medical Diagnosis and Treatment," in *RSCTC*, ser. LNCS, vol. 6086. Springer, 2010, pp. 4–19.
[24] H. E. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. I. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma, "ArrayExpress Update - From an Archive of Functional Genomics Experiments to the Atlas of Gene Expression," *Nucleic Acids Research*, vol. 37, no. Database-Issue, pp. 868–872, 2009.
[25] R. Jensen and Q. Shen, "New Approaches to Fuzzy-Rough Feature Selection," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 824–838, 2009.
[26] R. Słowiński, S. Greco, and B. Matarazzo, "Dominance-Based Rough Set Approach to Reasoning About Ordinal Data," in *RSEISP*, ser. LNCS, vol. 4585. Springer, 2007, pp. 5–11.
[27] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. [Online]. Available: http://www.R-project.org
[28] J. Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
[29] A. Janusz, "Dynamic Rule-based Similarity Model for DNA Microarray Data," *LNCS Transactions on Rough Sets XV*, vol. 7255, pp. 1–25, 2012.
[30] 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.
[31] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum Information About a Microarray Experiment (MIAME) - Toward Standards for Microarray Data," *Nature Genetics*, vol. 29, no. 4, pp. 365–371, 2001.
[32] R. Feldman and J. Sanger, Eds., *The Text Mining Handbook*. Cambridge University Press, 2007.
[33] M. Grzegorowski, P. W. Pardel, S. Stawicki, and K. Stencel, "SONCA: Scalable Semantic Processing of Rapidly Growing Document Stores," in *New Trends in Databases and Information Systems*, ser. Advances in Intelligent Systems and Computing, M. Pechenizkiy and M. Wojciechowski, Eds. Springer Berlin Heidelberg, 2013, vol. 185, pp. 89–98.
[34] D. Ślęzak, K. Stencel, and H. S. Nguyen, "(No)SQL Platform for Scalable Semantic Processing of Fast Growing Document Repositories," *ERCIM News*, no. 90, 2012.
[35] D. Ślęzak, A. Janusz, W. Świeboda, H. S. Nguyen, J. G. Bazan, and A. Skowron, "Semantic Analytics of PubMed Content," in *USAB*, ser. LNCS, vol. 7058. Springer, 2011, pp. 63–74.