# Applications of Approximate Reducts to the Feature Selection Problem*

Andrzej Janusz and Sebastian Stawicki

Faculty of Mathematics, Informatics, and Mechanics, The University of Warsaw,
Banacha 2, 02-097 Warszawa, Poland
{andrzejanusz,sebastian.stawicki}@gmail.com

**Abstract.** In this paper we overview two feature rankings methods that utilize basic notions from the rough set theory, such as the idea of the decision reducts. We also propose a new algorithm, called Rough Attribute Ranker. In our approach, the usefulness of features is measured by their impact on quality of the reducts that contain them. We experimentally compare the reduct-based methods with several classic attribute rankers using synthetic, as well as real-life high dimensional datasets.

**Keywords:** attribute filtering, feature selection, approximate reducts.

## 1 Introduction

Contemporary applications of data mining often involve working on data described by extremely high number of attributes. On one hand such a detailed representation may help to capture some important aspects of the phenomena under scope but on the other, abundance of information makes it difficult, even for an expert, to distinguish between relevant and irrelevant features in a given context. Unnecessary attributes can not only cripple performance of predictive models but also increase their construction cost and hinder their interpretability. For this reason, selecting informative features is one of the key steps during construction of any classification model for high dimensional data ([1], [2]).

Numerous researchers have investigated the problem of feature selection for predictive models. In a general case, attribute selection algorithms can be divided into two separate groups, i.e. wrapper and filter methods[1] ([3], [4]). The problems of feature selection and discovery of dependencies between features have been closely related to the rough set theory from its very beginning ([5], [6]). In a standard rough set approach, subsets of attributes take a form of reducts. In [7] it is shown that a decision reduct can consist only of strongly and weakly relevant features (cf. [4]), if the available data sufficiently cover the universe.

* The authors are supported by the grant N N516 077837 from the Ministry of Science and Higher Education of the Republic of Poland and by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

[1] Some researchers distinguish embedded methods as a third, hybrid group ([2]).

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 45–50, 2011.

Approximate reducts (cf. [8]) extend this idea by considering reducts that not necessarily discern all objects, but are less sensitive to random data disturbances.

Several feature ranking techniques that make use of the concept of reducts have been used by the rough set community. In this paper we propose a new one, called Rough Attribute Ranker (RAR). It utilizes the notions of approximate reducts and discernibility measure to rank features according to their influence on classification. We compare its performance with two other attribute ranking algorithms that derive from the rough sets theory. In those approaches, the ranking is based on frequencies of the attribute occurrence in the reducts. We also compare effectiveness of the reduct-based rankers to several classic methods. Apart from tests on synthetic data, we apply the algorithms to the gene selection problem for three microarray datasets from different medical domains.

## 2 Preliminaries

Feature selection techniques aim at finding an optimal set of attributes to represent objects for a purpose of a given task. Depending on a specification of the task, the attributes which compose the optimal set may be different. For example, to visualize general dependencies in data one may prefer to select a small set of the most relevant and diversified attributes, whereas for a classification task a larger set that contains also less relevant features is usually preferred [2].

The rough set theory is an extension within the classical set theory, suitable for describing concepts in presence of inaccurate information. A basic information unit for the rough sets is an indiscernibility class. Some indiscernibility classes of objects from the same decision class can be aggregated to form information granules. For numeric data, this aggregation can be done using a discretization heuristic that is based on a discernibility measure (see [9]).

Once data is discretized, objects can be compactly represented by features forming a *decision reduct*. We use a definition of the reduct which is adapted to the case when all the original attributes are numeric and all objects are discerned.

**Definition 1.** *Let $T = (U, A, d)$ be a decision system and $A'$ denote a set of symbolic attributes which were obtained from numeric data by discretization. A decision reduct $RED \subseteq A'$ is a set of attributes which is sufficient to discriminate among all objects from different decision classes and there is no $a \in RED$ such that $RED \setminus \{a\}$ would still discern all pairs of $u, u' \in U$, $d(u) \neq d(u')$.*

It has been showed in [7] that, assuming completeness of a decision system[2], any decision reduct can consist only of relevant attributes and that all strongly relevant attributes are shared by all decision reducts. Unfortunately, this assumption is rarely met in practice. Usually the reducts also contain irrelevant attributes and are vulnerable to disturbances in data. To overcome this issue, several generalizations of the decision reducts have been proposed, such as the *dynamic reducts* and the *approximate reducts*.

[2] By a complete decision system we mean a decision system which contains enough information to be representative for objects from the given universe.

**Definition 2.** *Let $RED(T)$ denote a set of all decision reducts of a decision table $T = (U, A, d)$ and let $disc(A, T)$ be a number of pairs of objects from different decision classes of $T$ that are discerned by an attribute set $A$. A set of attributes $AR$ will be called an $\epsilon$–approximate reduct iff there exists $RED \in RED(T)$ such that $AR \subset RED$ and $disc(AR, T) \geq (1 - \epsilon) * disc(RED, T)$.*

Intuitively, an $\epsilon$–approximate reduct is a subset of some decision reduct that is able to discern sufficiently many objects from different decision classes of $T$.

## 3 Approximate Reduct-Based Feature Selection Methods

This section presents attribute ranking methods that utilizes a rough set tool, i.e. a concept of approximate reducts, to measure relevance of individual features. All of those methods are multivariate as they consider attributes in a context of others and are able to detect dependencies between them.

In our research we needed to adjust the existing methods to better fit to the problem of selecting relevant attributes from high dimensional decision systems ($|A| \geq 1000$). To generate approximate reducts we use an algorithm described in [10] with a modified stopping criteria. This method makes use of random sampling of the attribute set in order to discover reducts that capture diverse characteristics of data and reduce the computation cost. It constructs reducts from numeric data using the maximum-discernibility discretization heuristic ([9]).

Commonly used rough set feature ranking methods exploit the fact that the informative attributes are usually able to discern more objects from different decision classes, and thus are more likely to be present in a reduct. We examine two frequency-based approaches to attribute ranking. We test them in combination with approximate reducts and compare to our attribute ranking method.

Let us denote a finite set of approximate reducts of the decision system $T$ by $ARED(T)$, $|ARED(T)| = m$ and assume that each $AR_i \in ARED(T)$ was computed using a subset of attributes $B_i \subseteq A$, $i = 1, ..., m$. The first and the simplest of the compared methods ranks attributes by counting how many times they appear in a given set of decision reducts:

$$Score_1(a) = |\{AR_i \in ARED(T) : a \in AR_i\}| \ . \tag{1}$$

The second method considers a predictive potential of the reducts. In this approach, each $AR_i$ is assigned with a score $Scr(AR_i)$ which expresses its quality. This value is used as a weight during computation of the frequencies:

$$Score_2(a) = \sum_{i=1}^{m} \Big( Scr(AR_i) * \chi_{AR_i}(a) \Big) \ / \ \sum_{i=1}^{m} Scr(AR_i), \tag{2}$$

where $\chi_{AR_i}$ is a characteristic function of the attribute set $AR_i$.

The value of $Scr(AR_i)$ could be computed using one of many heuristics. In this paper, we propose to assess a quality of the reduct by a direct application of the discernibility measure to objects, which ware removed from the training set during approximate reduct assembling.

The ranking algorithm, which we propose in this paper, is an extension to the methods described above. It uses a discernibility-based scoring function $Scr_{AR}$ to examine how particular attributes from a reduct influence the reduct's quality. The impact of a single attribute is estimated as an average difference in the score assigned to the reduct after exchanging this attribute with a random probe, which is generated by a random permutation of the attribute values. Value of $Scr_{AR}(a)$ can be expressed as $Scr(AR) - \sum_i^K Scr(AR'_i)/K$, where $AR$ is an approximate reduct, $a \in AR$, $K$ is a number of random probes used for the estimation and $AR'_i$, $i = 1, ..., K$ are sets of attributes that were constructed from $AR$ by changing $a$ with a random probe. The total score assigned to an attribute $a$ is equal to its mean impact on quality of the reducts:

$$Score_{RAR}(a) = \sum_{i=1}^{m'} Scr_{AR_i}(a) \,/\, m', \tag{3}$$

where $m' \leq m$ is a number of the approximate reducts containing the attribute $a$. This method may be seen as an analogy to the Breiman's relevance measure for the random forest, which assesses the importance by examining how randomization of particular attributes influence the error rate of trees.

## 4 Evaluation of the Reduct-Based Ranking Methods

The three attribute ranking methods described in the previous section were empirically evaluated and compared to results of several commonly used statistical feature rankers, i.e. a correlation-based ranker, a Wilcoxon test-based ranker, the information gain and the relief algorithm. This comparison has been performed using two different evaluation methods on synthetic and real-life data.

In the first test a dataset containing 10000 objects described by 1000 numeric attributes was generated from normal distribution. Three different decision attributes were constructed using the first 20 features so that each of the selected features had the same impact on the decision value:

$$\begin{aligned}
&Decision_1(u) = 1 \Leftrightarrow \textstyle\sum_{i=1}^{20} a_i(u) \geq 0,\\
&Decision_2(u) = 1 \Leftrightarrow a_1(u)a_{20}(u) + \textstyle\sum_{i=1}^{19} a_i(u)a_{i+1}(u) \geq 0,\\
&Decision_3(u) = 1 \Leftrightarrow \Big(a_1(u) \in [-\delta, \delta] \wedge a_{20}(u) \in (-\infty, -\delta) \cup (\delta, \infty)\Big) \vee\\
&\qquad\qquad \bigvee_{i=1,\ldots,19} \Big(a_i(u) \in [-\delta, \delta] \wedge a_{i+1}(u) \in (-\infty, -\delta) \cup (\delta, \infty)\Big).
\end{aligned}$$

The first decision depended linearly on the attribute values and as such might be seen as the easiest one, whereas the dependency of the second and the third decision attributes was not linear. In order to better mimic a real-life situation additional noise was introduced to data. For each of the decision attributes 20% of randomly selected values were rearranged by a random permutation.

The compared ranking algorithms were used to select relevant features for each decision vector. The number of attributes that each of the algorithms should choose was estimated using the random probes test described in [2]. Quality of

**Table 1.** Results of the attribute ranking methods on the synthetic data

| | $Decision_1$ | | | | $Decision_2$ | | | | $Decision_3$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ranker: | $N$ | $Prec.$ | $Recall$ | $F_{score}$ | $N$ | $Prec.$ | $Recall$ | $F_{score}$ | $N$ | $Prec.$ | $Recall$ | $F_{score}$ |
| CorrRank | 32 | 0.63 | 1.0 | 0.77 | 14 | 0.07 | 0.05 | 0.06 | 20 | 0.05 | 0.05 | 0.05 |
| Wilcoxon | 28 | 0.71 | 1.0 | 0.83 | 11 | 0.09 | 0.05 | 0.06 | 19 | 0.05 | 0.05 | 0.05 |
| InfoGain | 27 | 0.74 | 1.0 | 0.85 | 8 | 0.13 | 0.05 | 0.07 | 16 | 0.56 | 0.45 | 0.5 |
| Relief | 39 | 0.51 | 1.0 | 0.68 | 19 | 0.16 | 0.15 | 0.15 | 22 | 0.09 | 0.1 | 0.09 |
| $AR_{Score1}$ | 28 | 0.71 | 1.0 | 0.83 | 10 | 0.0 | 0.0 | 0.0 | 12 | 0.25 | 0.15 | 0.19 |
| $AR_{Score2}$ | 26 | 0.77 | 1.0 | 0.87 | 10 | 0.0 | 0.0 | 0.0 | 12 | 0.33 | 0.2 | 0.25 |
| $RAR$ | 29 | 0.69 | 1.0 | 0.82 | 26 | 0.62 | 0.8 | 0.7 | 21 | 0.48 | 0.5 | 0.49 |

**Table 2.** Results of the attribute ranking methods on the three microarray data. The mean number of selected genes (N) and balanced accuracies are given.

| | $acuteLymphLeukemia$ | | | $hepatitisC$ | | | $skinPsoriatic$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Ranker: | N | BAC:kNN | BAC:RF | N | BAC:kNN | BAC:RF | N | BAC:kNN | BAC:RF |
| CorrRank | 6880 | 0.91 | 0.75 | $> 13K$ | 0.86 | 0.79 | $> 30K$ | 0.76 | 0.81 |
| Wilcoxon | 3538 | 0.91 | 0.75 | 6348 | 0.87 | 0.77 | $> 22K$ | 0.76 | 0.82 |
| InfoGain | 5733 | 0.91 | 0.75 | $> 12K$ | 0.86 | 0.78 | $> 24K$ | 0.76 | 0.82 |
| Relief | 12054 | 0.92 | 0.82 | 2551 | 0.89 | 0.79 | 4904 | 0.76 | 0.81 |
| $AR_{Score1}$ | 1512 | 0.91 | .82 | 1471 | 0.90 | 0.80 | 1492 | 0.78 | 0.83 |
| $AR_{Score2}$ | 2021 | 0.91 | 0.81 | 1696 | 0.90 | 0.80 | 1993 | 0.77 | 0.84 |
| $RAR$ | 1953 | 0.92 | 0.81 | 2068 | 0.91 | 0.80 | 3186 | 0.77 | 0.84 |

the selected feature sets was assessed using classic measures from the information retrieval domain – precision, recall and F-score (Table 1).

The second experiment was conducted on microarray datasets related to different medical domains. The data was downloaded from ArrayExpress[3]. The *acuteLymphoblasticLeukemia* dataset (190 samples, 22276 genes) describes five genetic subtypes of acute lymphoblastic leukemia, *hepatitisC* (124 samples, 22276 genes) regards a role of chronic hepatitis C virus in the pathogenesis of HCV-associated hepatocellular carcinoma and *skinPsoriatic* (180 samples, 54675 genes) contains profiles of genetic changes related to the skin psoriasis.

In this test, the compared methods were used to selected gene sets in a repeated 5-fold cross-validation schema. A quality of each gene set was evaluated based on classification results achieved by two prediction models which are commonly used in the microarray experiments, i.e. $k$-NN and random forest. Due to uneven sizes of the decision classes, the prediction accuracy was measured using the balanced accuracy score. As in the experiment on synthetic data, the number of genes selected in each fold of the cross-validation cycle was determined using the random probes method. Table 2 summarizes the mean results achieved by each ranking algorithm after 10 executions of 5-fold cross-validation tests.

The reduct-based rankers outperformed the classic algorithms in both tests. For the synthetic data only RAR method was able to reasonably select attributes which are relevant for the second decision vector. Interestingly, in case when the relation between relevant features and the decision is not very complex, the simplest, frequency based attribute rankers may yield better results. It is also noticeable, especially for microarray data, that the frequency based approaches tend to select less attributes than the RAR. In terms of classification accuracy, the RAR on average performed slightly better than other algorithms but in cases of other reduct-based rankers the difference was not statistically significant.

[3] www.ebi.ac.uk/arrayexpress

## 5 Conclusions

We described three attribute ranking methods that make use of reducts. The first two are based on the intuition that relevant features more occur in reducts than the irrelevant ones. Our Rough Attribute Ranking algorithm extends this approach with introduction of the attribute impact measure, by an analogy to the Breiman's ranking method for the random forest. Performance of those algorithms was empirically evaluated and compared to several classic rankers on synthetic as well as real-life high dimensional data. The results of conducted experiments confirm that RAR is able to detect relevant attributes even in situations when the target decision is not linearly dependent on the features.

The problem of selecting relevant attributes is extremely important in many fields related to data analysis and should be regarded as one of the main research directions for the rough set community. For this reason, in the future we plan to further investigate attribute selection and feature extraction methods that derive from the rough set theory. For example, we are interested how different heuristics for generating decision reducts would influence quality of the proposed ranking algorithms. We would also like to investigate a geometry of reducts from high dimensional datasets by examining co-occurrence of particular attributes.

## References

1. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)
2. Guyon, I., et al.: Feature Extraction: Foundations and Applications. Studies in Fuzziness and Soft Computing. Springer, Heidelberg (2006)
3. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant Features and the Subset Selection Problem. In: International Conference on Machine Learning, pp. 121–129 (1994)
4. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. 97, 273–324 (1997)
5. Pawlak, Z.: Rough sets: present state and the future. Foundations of Computing and Decision Sciences 18(3-4), 157–166 (1993)
6. Modrzejewski, M.: Feature selection using rough sets theory. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 213–226. Springer, Heidelberg (1993)
7. Pawlak, Z.: Rough sets - Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
8. Ślęzak, D.: Rough sets and functional dependencies in data: Foundations of association reducts. In: Gavrilova, M., Tan, C., Wang, Y., Chan, K. (eds.) Transactions on Computational Science V. LNCS, vol. 5540, pp. 182–205. Springer, Heidelberg (2009)
9. Nguyen, H.S.: On efficient handling of continuous attributes in large data bases. Fundamenta Informaticae 48(1), 61–81 (2001)
10. Janusz, A.: Utilization of dynamic reducts to improve performance of the rule-based similarity model for highly-dimensional data. In: Proceedings of the WI/IAT Workshops. IEEE Computer Society, Los Alamitos (2010)