2006 International Conference on Hybrid Information Technology
November 9th ~ 11th, 2006

Rough Discretization of Gene Expression Data

Dominik Ślęzak[1,3] and **Jakub Wróblewski**[2,1]
[1]Infobright Inc.
e-mail : {slezak,jakubw}@infobright.com
[2]Polish-Japanese Institute of Information Technology
e-mail : jakubw@pjwstk.edu.pl
[3]Department of Computer Science, University of Regina
e-mail : slezak@uregina.ca

Abstract We adapt the rough set-based approach to deal with the gene expression data, where the problem is a huge amount of genes (attributes) $a \in A$ versus small amount of experiments (objects) $u \in U$. We perform the gene reduction using standard rough set methodology based on approximate decision reducts applied against specially prepared data. We use rough discretization – Every pair of objects $(x,y) \in UxU$ yields a new object, which takes values "$\geq a(x)$" if and only if $a(y) \geq a(x)$; and "$<a(x)$" otherwise; over original genes-attributes $a \in A$. In this way: 1) We work with desired, larger number of objects improving credibility of the obtained reducts; 2) We produce more decision rules, which vote during classification of new observations; 3) We avoid an issue of discretization of real-valued attributes, difficult and leading to unpredictable results in case of any data sets having much more attributes than objects. We illustrate our method by analysis of the gene expression data related to breast cancer.

Keyword: Rough Sets, Feature Selection, Discretization, Gene Expression Data

1. Introduction

DNA microarrays provide a huge quantity of information about genetically conditioned susceptibility to diseases [1,3]. However, their analysis is uneasy because of large number of genes. A typical gene expression data set has a few cases, while the number of attributes is counted in thousands. This yields a problem for methods assuming data to be representative enough.

In this paper, we deal with the above challenge using rough discretization, applied before to gene clustering [5]. We apply the rough-set-based method for finding optimal approximate reducts and rules [2,8,12]. We proceed with roughly discretized data, to express explicit comparisons of objects' values for particular attributes. Resulting rules have clear interpretation in terms of inequality conditions for gene expressions. Obtained reducts (minimal subsets of attributes, which provide approximately same levels of information about decisions, as the full sets of attributes) are more credible, because of calculation over far larger universe of cases. Finally, experimental results related to the breast cancer data confirm that our approach is quite promising, though it surely requires further research.

2. Gene Expression Data

The *DNA microarray* technology [1] enables simultaneous analysis of thousands of genes in biological samples. It is automated, much quicker, and less complicated than the previous methods of molecular biology, allowing scientists to study no more than a few genes at a time. Microarray production starts with preparing two samples of mRNA. Actual sample of interest is paired with a healthy control sample. Fluorescent labels are applied to both the control (green) and actual (red) samples. The procedure of mixing two labeled samples is repeated for each of genes. Then the slide is washed and the color intensities of every gene-spot are scanned. Fluorescence of red and green colors indicates to what extent particular genes are expressed.

3. Rough Set Approach

In the rough set theory [6], a sample of data takes a form of *information system* $\mathbf{A}=(U,A)$, where each attribute $a \in A$ is a function $a:U \rightarrow V_a$ from universe U into a value set V_a. (In this paper, A corresponds to the set of genes and U is the set of observations.) In case of classification problem, we have a distinguished decision $d \notin A$, with values to be determined using A. (Here, d labels behaviors related to breast cancer.) Then, a data set takes a form of *decision system* $\mathbf{A}=(U,A \cup \{d\})$. Rough set models balance between accuracy and complexity using decision reducts: minimal subsets $B \subseteq A$ that determine d. Obtained reducts produce decision rules. Smaller reducts induce shorter and more general rules. Quite often, it is even better to delete more attributes, to get shorter rules, at a cost of a slight loss of decision determination. To measure determination level for $B \subseteq A$, the rough set theory introduces *positive region*

POS(B) = {$u \in U$: all objects that have the same values as u on attributes $a \in B$, have also to the same decision as u}

By a *decision reduct* for $\mathbf{A}=(U,A \cup \{d\})$ we mean $B \subseteq A$ such that POS(B)=POS(A), and there is no proper subset of B, which holds analogous equality [6]. Further, in this paper, by an *ε-reduct* for $\mathbf{A}=(U,A \cup \{d\})$ we mean $B \subseteq A$ such that:

$$|POS(B)| \geq (1-\varepsilon)\,|POS(A)|$$

and there is no proper subset of B, which holds analogous inequality. |POS(B)| denotes cardinality of POS(B), i.e. the number of objects $u \in U$ which can be correctly classified by exact rules induced using **A**. Approximation threshold $\varepsilon \in [0,1)$ expresses willingness to reduce more attributes (and simplify rules) on the cost of losing positive region. The above inequality is an example of the approximate reduction criterion introduced in [8], further considered in rough set literature in many different variants (cf. [12]).

4. Rough Discretization

Gene expression data is an information system with real-valued attributes (gene expressions). When enriched with clinical information, it becomes decision system with well-specified decision classes (e.g. indicating disease types). Standard rough set methods are not applicable unless we use discretization [2] or apply more advanced techniques, based on similarities and/or hierarchical models (cf. [7]). Still, according to our experience, both the rough-set-like techniques and other tools using explicitly real values do not deal properly with disproportions between attributes and objects. Hence, we suggest another, non-invasive form of discretization, already applied to gene clustering [5], which enables to increase the number of objects and provides relevant results in terms of reducts and rules.

Table 1. Decision system with four objects u1,u2,u3,u4, two conditional attributes a,b, and three decision classes d=0,1,2.

	a	b	d
u1	3	7	0
u2	2	1	1
u3	4	0	1
u4	0	5	2

Let the decision system **A**=(U,A∪{d}) with real-valued conditional attributes be given. (Table 1 illustrates the system, where U={u1,u2,u3,u4} and A={a,b}). *Rough discretization* is the procedure of creation from **A** the decision system **A***=(UxU,A*∪{d*}), where, for every x,y∈UxU, we put d*(x,y)=d(y), and, for every a∈A,

a*(x,y)="≥a(x)" if and only if a(y)≥a(x)
a*(x,y)="<a(x)" if and only if a(y)<a(x)

Table 2. Decision system obtained using rough discretization.

	a*	b*	d*
(u1,u1)	≥ 3	≥ 7	0
(u1,u2)	< 3	< 7	1
(u1,u3)	≥ 3	< 7	1
(u1,u4)	< 3	< 7	2
(u2,u1)	≥ 2	≥ 1	0
(u2,u2)	≥ 2	≥ 1	1
(u2,u3)	≥ 2	< 1	1
(u2,u4)	< 2	≥ 1	2
(u3,u1)	< 4	≥ 0	0
(u3,u2)	< 4	≥ 0	1
(u3,u3)	≥ 4	≥ 0	1
(u3,u4)	< 4	≥ 0	2
(u4,u1)	≥ 0	≥ 5	0
(u4,u2)	≥ 0	< 5	1
(u4,u3)	≥ 0	< 5	1
(u4,u4)	≥ 0	≥ 5	2

Example: Table 2 illustrates a decision system obtained from Table 1 via rough discretization. Region POS(a,b)= {(u1,u1),(u1,u3),(u2,u3),(u2,u4),(u3,u3),(u4,u2),(u4,u3)} yields the following decision rules:

IF a≥3 AND b≥7 THEN d=0
IF a≥3 AND b<7 THEN d=1
IF a≥2 AND b<1 THEN d=1
IF a<2 AND b≥1 THEN d=2
IF a≥4 AND b≥0 THEN d=1
IF a≥0 AND b<5 THEN d=1

{b} is ε-reduct for ε=3/7, POS(b)={(u2,u3),(u4,u2),(u4,u3)}, while {a} is ε-reduct for ε=2/7, POS(a)={(u2,u4),(u3,u3)}.

The rules generated from reducts of **A*** are of similar form as those obtained by discretization described in [2], where each object's value can yield a cut over the given attribute. However, [2] and many other discretization tools (see e.g. [4]) correspond to creating new attributes for every (a,u), a∈A, u∈U, and keeping universe U unchanged. Clearly, it does not help but, instead, further deepens the problem of disproportion between numbers of attributes and objects. Our method, on the other hand, keeps unchanged the set of attributes treating them, actually, as source of rankings for comparing the objects (cf. [5]). In this way, it should be rather compared to the rough set approaches to decision system with the ranking attributes (cf. [10]). Actually, the usage of attribute quality functions for roughly discretized data to deal with inexact dependencies between attribute-rankings is one of interesting future research directions.

5. Classification Results

We study the breast cancer data downloaded from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=360), anlayzed in [2,11]. It contains 24 core biopsies taken from patients, who are resistant (first decision, 14 objects) or sensitive (second decision, 10 objects) to the docetaxel treatment. There are 12,625 genes-attributes. Figure 1 shows results using leave-one-out method. For every epsilon, we take 5 best reducts and classify by voting [2,12]. For ε close to 0, accuracy and coverage are practically 100%. (Compare to k-NN providing accuracy below 85% and standard rough set methods, without rough discretization, that are even worse.) We can also see that reducts are extremely small against the whole set of genes-attributes. For higher ε we get worse accuracy but better reduction, with ε-reducts of average size equal to 6 and accuracy at the level of k-NN.

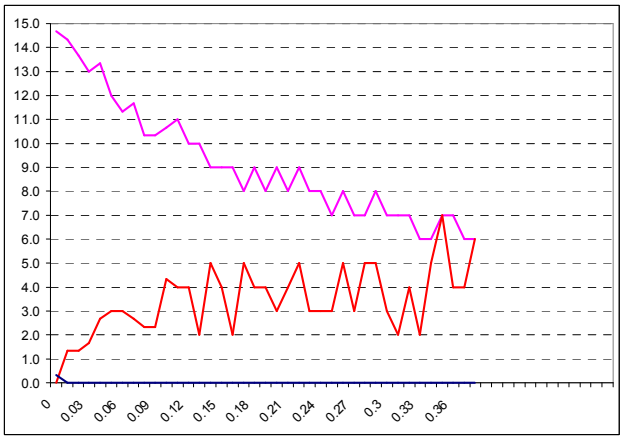

Fig. 1. Results for the breast cancer data. For particular epsilons (x axis), purple line denotes average number of genes in reducts, red line – number of wrongly recognized objects, and blue line (at the very bottom, almost invisible) – not recognized objects.

Obviously, further experimental study is needed here, both regarding comparison to other methods [1,3,5,11], as well as tuning parameters of relevant rough set tools [2,12].

6. Conclusions and Further Research

A new way of processing numeric data with large number of attributes versus low number of objects turns out to be well-suited to the gene expression data. Interpretation of decision rules while voting nicely merges advantages of rough set reduction and the proposed method of rough discretization. Among further research directions, there is hybridization of rough set reduction framework with gene clustering. In [5] we based self organizing maps on the entropy distance calculated for roughly discretized data. Figure 2 shows our system's interface, where one is able to drag&drop genes between clusters interactively, while running the learning process. In [10] it was proposed that gene clustering can be used for reduction of attributes and considering clusters' representatives for further analysis. Methodology described in [10] is, however, quite complex and we think that comparable performance is achievable by only simple gene clustering and approximate reducts, given rough discretization and well-adjusted functions.

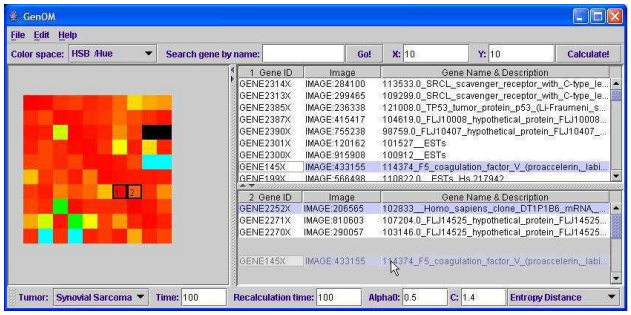

Fig. 2. GenOM system [5].

It is also worth remembering that gene expression data is a rich information source for unsupervised learning about gene dependencies [1,5]. Hence, the above methodology could be combined with other rough set models, like, e.g., association reducts [9], to reflect multi-gene relationships.

Acknowledgements

This paper is supported by Research Grant from Natural Sciences and Engineering Research Council of Canada, awarded to the first author, as well as by Research Centre of Polish-Japanese Institute of Information Technology.

References

[1] P. Baldi and W.G. Hatfield "DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modelling" Cambridge University Press, 2002.

[2] J. Bazan et al "Rough Set Algorithms in Classification Problem" Rough Set Methods and Applications, Physica-Verlag, 2000.

[3] J.C. Chang et al "Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer" The Lancet, Vol. 362, 2003.

[4] B. Ganter and R. Wille "Formal Concept Analysis: Mathematical Foundations" Springer-Verlag, New York, 1997.

[5] A. Grużdź, A. Ihnatowicz, and D. Ślęzak „Interactive gene clustering: A case study of breast cancer microarray data" Information Systems Frontiers, Vol. 8, 2006.

[6] Z. Pawlak "Rough sets: Theoretical aspects of reasoning about data" Kluwer Academic Publishers, Dordrecht, Netherlands, 1991.

[7] A. Skowron et al "A Hierarchical Approach to Multimodal Classification" Proc. of RSFDGrC'05, 2005.

[8] D. Ślęzak "Approximate reducts in decision tables" Proc. of IPMU'96, 1996, Vol. 3.

[9] D. Ślęzak "Association Reducts: Complexity and Heuristics" Proc. of RSCTC'06, 2006.

[10] R. Słowiński, S. Greco, and B. Matarazzo „Rough Set Based Decision Support" Introductory Tutorials on Optimization, Search and Decision Support Methodologies, Springer-Verlag, Boston, 2005.

[11] J.J. Valdes and A.J. Barton "Relevant Attribute Discovery in High Dimensional Data: Application to Breast Cancer Gene Expressions" Proc. of RSKT'06, 2006.

[12] J. Wróblewski "Ensembles of classifiers based on approximate reducts" Fundamenta Informaticae, Vol. 47, No. 3-4, 2001.

Authors

Dominik Ślęzak

Received MSc in Mathematics (1996) and PhD in Computer Science (2002), Warsaw University, Poland. Cooperating with Group of Logic, Warsaw University, Robotics Laboratory, PJIIT, Poland, and Rough Set Laboratory, Regina, Canada. Chief Scientist at Infobright Inc. Executive Member of IRSS. Member of 2 IEEE technical committees. Editor-In-Chief of Online Int. Journal on Rough Set Methods. Guest Co-Editor of several journal special issues. Co-Chair of several int. conferences. About 60 papers.

Jakub Wróblewski

Received MSc in Mathematics (1996) and PhD in Computer Science (2002), Warsaw University, Poland. Assistant Professor in PJIIT, Poland. Cooperating with Infobright Inc. and Group of Logic, Warsaw University. About 40 papers.