ISSN 1691-5402
ISBN 978-9984-44-071-2
Environment. Technology. Resources
Proceedings of the 8th International Scientific and Practical Conference. Volume I1

BRAIN CANCER ANTIBODY DISPLAY CLASSIFICATION

Madara Gasparovica, Ludmila Aleksejeva
Department of Modelling and Simulation, Institute of Information Technology
Riga Technical University
1 Kalku Str., Riga LV 1658, Latvia
E-mail: madara.gasparovica@rtu.lv, ludmila.aleksejeva@cs.rtu.lv

Abstract. *This article explores real data on brain cancer. This type of biological data has a few particularities like a great number of attributes – antibodies and genes. However the number of entries is rather small because the data have to be obtained from real patients. This process is time consuming and very costly. Due to that, this research provides detailed data description as well as analyzes their particularities, type and structure. Correspondingly, classification rules are also difficult to discover. This research is dedicated to finding applications of classification methods aimed at determining interconnections that could be used to classify brain cancer. Working exactly with such unique data has a great practical value, because the data obtained can be used in future to continue the research and in practical diagnostics with the possibility to offer the data to biologists for interpretation. To speed up the obtaining of interconnections, only important attributes were used. Various methods of interconnection determination were employed. Conclusions about this type of data analysis, obtaining classification rules and the precision of obtained rules are made and directions of future work are outlined.*

Keywords: *antibody display, classification, data mining, IF- THEN rules.*

Introduction

In the last few years the rapid development of computer systems has enabled performing even more complicated computing actions; thus data obtaining algorithms that can identify and classify various diseases have become very popular. One of the first works in this sphere is Golub et al. [1] research that is a basis for many other researches [2-4]. Different methods and the most popular gene expression data set description can be found in [5].

In the present work, the analyzed data are antibodies which are created as a response to some infectious disease microorganism, vaccine or another anti-gene and which react specifically to this particular anti-gene [6]. As a result, by creating antibodies, conclusions can be made whether a patient has been infected with particular disease. Such real data analysis and obtaining of important interconnections in classification is also done by biologists themselves [7], however the methods they are using differ a little from the methods used in the process of collecting and analyzing data.

Two methods of obtaining fuzzy rules were used in this paper - FURIA (An Algorithm For Fuzzy Rule Induction) and FLR (Fuzzy Lattice Reasoning) classifiers. The FURIA algorithm was proposed in 2009 by Hühn and Hüllermeier [8]. FURIA is a RIPPER algorithm modification, preserving all RIPPER [9] algorithm advantages, for example, a simple and well understood set of laws. In addition, it includes a number of modifications and extensions. FURIA obtains fuzzy rules instead of the usual strict rules, as well as an unordered rule set instead of the rule list. In addition, to address the problem of uncovered samples it uses an efficient method for stretching the rules. Combined with a sophisticated law induction method provided by the original RIPPER algorithm these improvements have led to a better rule induction algorithm for classification, which requires only a small increase in classification time. Authors have made extensive experiments that show that FURIA outperforms the original RIPPER algorithm, as well as other methods of obtaining fuzzy rules.

The FLR classifier was proposed in 2007 by Kaburlasos, Athanasiadis and Mitkas [10]. The FLR classifier is designed to obtain descriptive, decision-making knowledge (rules) in a

mathematical lattice data. Training takes place both gradually and rapidly, calculating the disjunctions of interval conjunctions. In this article, the authors study the problem of ozone concentration from both meteorological and air pollutant measurements. The FLR classifier induces rules from training examples, allowing a rise in the size of the diagonal of the rule to a maximum threshold. FLR is the Leader-Follower classifier, which learns quickly at a time, using the results of training. Data input order is vital. The total number of rules is not known a priori, but it is usually determined during the training period [10].

The first section of the paper describes the methods used and general principles underlying their work. The set of data used in the experiments is specified and compared to other publicly available databases. The second section discusses the experiments performed and their results. In conclusion, some observations about particular data set and directions of future research are provided.

Materials and methods

In the course of this study, experiments with real data of cancer research antibodies were performed. A short review of the data set is given in Table 1. Clearly biological data specifics can be seen – a great number of attributes and a rather small count of entries. There are also problems with data domination, which is a topical problem, because such biological data experiments are relatively expensive and complicated to perform. Important factor in cancer classification is the stage of the disease – the later the stage, the clearer it is to classify. However, in this research the stage has not been considered, as not all of the entries have this data provided. That is why 1229 attributes (genes) are used in experiments.

Table 1.

Brain cancer data set description

Number of attributes (genes)	1230	
Number of classes	2	Brain cancer (BrCa)
		Healthy donor (HD)
Number of instances	168	BrCa – 13
		HD – 155

The data set in the space of the two most important attributes determined (where 329 and 501 are two most relevant attributes) in the experiment is displayed in Fig. 1.

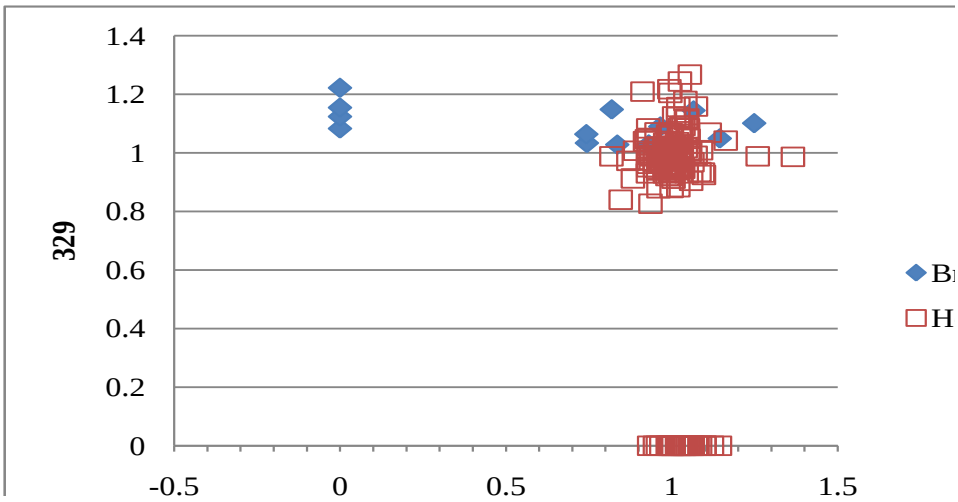

Fig. 1. **Brain cancer data set relevant attributes**

Let us describe the methods used in the experiments in more detail. Since six different methods were used in the classification and the first four of them - Ridor [11], PART[12], OneR [13] and JRIP (a version of RIPPER algorithm [15] that was created especially for WEKA) are rigorous training methods in Weka and only the last two - FLR and FURIA are based on fuzzy rules, they also will be given extra attention.

Let us describe the FURIA algorithm. The representation of rules is as follows. A fuzzy selector $A_i \in I_i^F$ covers an instance $x = (x_1 \ldots x_n)$ to the degree $I_i^F(x_i)$. A fuzzy rule r^F involving k selectors $(A_i \in I_i^F), i = 1,\ldots k,$ covers x to the degree [8]:

$$\mu_{r^F}(x) = \prod_{i=1\ldots k} I_i^F(x_i). \tag{1}$$

Rule fuzzification. For the fuzzification of a ($A_i \in I_i$) it is important to consider only the relevant training data D_T^i, i.e., to ignore those instances that are excluded by any other antecedent [8]:

$$D_T^i = \left\{x = (x_1 .. x_n) \in D_T \mid I_j^F(x_j) > 0 \, for \, all \; j \neq i\right\} \subseteq D_T \tag{2}$$

D_T^i is partitioned into the subset of positive instances, D_{T+}^i, and negative instances, D_{T-}^i. To measure the quality of a fuzzification, the rule purity will be used:

$$pur = \frac{p_i}{p_i + n_i} \tag{3}$$

where

$$p_i = \sum_{x \in D_{T+}^i} \mu_{Ai}(x)$$

$$n_i = \sum_{x \in D_{T-}^i} \mu_{Ai}(x).$$

Suppose that fuzzy rules $r_1^{(j)} \ldots \; r_k^{(j)}$ have been learned for class $\lambda_j . j$. For a new query instance x, the support of this class is defined by

$$s_j(x) = \sum_{i=1\ldots k} \mu_{r_i^{(j)}}(x) \cdot CF(r_i^{(j)}) \tag{4}$$

where $CF(r_i^{(j)})$ is the *certainty factor* of the rule $r_i^{(j)}$. It is defined as follows:

$$CF(r_i^{(j)}) = \frac{2\frac{\left|D_T^{(j)}\right|}{\left|D_T\right|} + \sum_{x \in D_T^{(j)}} \mu_{r_i^{(j)}}(x)}{\sum_{x \in D_T} \mu_{r_i^{(j)}}(x)}, \tag{5}$$

where $D_T^{(j)}$ denotes the subset of training instances with label λ_j [8].

Let us describe the FLR classifier [10]. The main rules used in this algorithm are as follows.

Definition 1. A fuzzy lattice is a pair $\langle L, \mu \rangle$, where L is a crisp lattice and $(L \times L, \mu)$ is a fuzzy set with membership function $\mu : L \times L \to [0,1]$ such that $\mu(x, y) = 1$ if and only if $x \leq y$.

Definition 2. An inclusion measure σ in a complete lattice L is a real function $\sigma : L \times L \to [0,1]$ such that for $u, w, x, y \in L$ the following conditions are satisfied:

(C0) $\sigma(X, O) = 0, x \neq O$

(C1) $\sigma(x, x) = 1, \forall x \in L$

(C2) $u \leq w \Rightarrow \sigma(x, u) \leq \sigma(x, w)$ – The Consistency Property

(C3) $x \wedge y < x \Rightarrow \sigma(x, y) < 1..$

Proposition 3. If $\sigma : L \times L \to [0,1]$ is an inclusion measure on lattice L, then $\langle L, \sigma \rangle$ is a fuzzy lattice.

Proposition 4. If L is a (complete) lattice and $v: L \to R$ is a positive valuation (with $v(O) = 0$) then (1) $k(x,u) = \frac{v(u)}{v(x \vee u)}$ and (2) $s(x,u) = \frac{v(x \wedge u)}{v(x)}$ are inclusion measures.

Proposition 5. Let L_i be a totally-ordered lattice, let $v: L \to R$ be a positive valuation, and let $\theta: L_i^{\partial} \to L_i$ be an isomorphic function in L_i. Then a positive valuation function $v: \tau(L_i) \to R$ is given by $v([a,b]) = v(\theta(a)) + v(b)$.

Definition 6. Consider a product lattice $L = L_1 \times ... \times L_n$. Let $v_i : L_i \to R$ be a positive valuation function in the constituent lattice $L_i, i = 1,... N.$. Then the diagonal of an interval $[a,b] \in \tau(L)$, with $a \le b$, is defined as a non-negative real function $diag_p : \tau(L) \to R_0^+$ given by $diag_p([a,b]) = d_p(a,b),\ p = 1,2...$

Proposition 7. For $p = 1,2,\ldots$ we have $diag_p([a,b]) = \max_{x,y \in [a,b]} d_p(x,y)$ [10].

Results and discussion

In the course of this work several experiments with various attribute importance methods available in software WEKA [14] were made to determine the number of important attributes that can be used in further experiments. As can clearly be seen, eight different combinations of methods were used to reach the goal – to obtain most important attributes, to narrow the data capacity that can be used to successfully perform classification several times. In every series of experiments 10 fold cross validation was used to get more accurate results with different methods and to make them less affected by coincidence. As a result, 75 attributes were found to be most important in this data set and all other experiments were performed with already narrowed data set.

To perform classification, classifiers based on interconnections that can be accessed in WEKA software were used. The results obtained are summarized in Table 2. In the first column the name of an algorithm is given; in the second – the number of correctly classified examples; in the third – the number of incorrectly classified examples; in the fourth – the accuracy of classification and in the fifth – the summarized number of obtained interconnections.

The last two methods - FLR and FURIA use fuzzy set theory to obtain the rules. As can be seen from the results, FLR reached the highest results - the classification accuracy of 95%.

Table 2.

Result of classification

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Accuracy	Number of rules
Ridor	149	19	88.69%	1
PART decision	159	9	94.64%	2
OneR	160	8	95.24%	2
JRIP	152	16	90.48%	2
Fuzzy lattice Reasoning (FLR)	161	7	95.83%	7
FURIA	156	12	92.86%	5

Since in such real life problem one class has a significantly greater (155 against 13) number of entries, it is important to clarify which examples have been classified incorrectly. In Table 3 the results of classification are summarized by the value, how precise every attribute classifies entries of each class. We can see that the best score for the “Brain cancer” class is

shown by Fuzzy FLR classificatory, then PART decision algorithm, which by overall accuracy score takes the second place a little behind OneR algorithm. The results of other algorithms are not so good. Of course, the practical result is important here – if a patient who is perfectly healthy is classified as a cancer case, it definitely is no good, though the patient is not in danger. However, if the case is opposite: a cancer patient is classified as healthy, it is extremely dangerous, because often timely diagnosis of this disease provides recovery possibilities.

Table 3.

Algorithm confusion matrix

Classifier		Classified as		
		a= BR	b=HD	Classification accuracy for each class
RIpple DOwn Rule Learner(Ridor)	a=BR	2	11	15.38%
	b=HD	8	147	94.94%
PART decision	a=BR	7	6	53.85%
	b=HD	3	152	98.06%
OneR	a=BR	6	7	46.15%
	b=HD	1	154	99.35%
JRIP	a=BR	3	10	23.08%
	b=HD	6	149	96.13%
FLR	a=BR	9	4	69.23%
	b=HD	3	152	98.06%
FURIA	a=BR	2	11	15.38%
	b=HD	1	154	99.35%

Interconnections obtained as a result of classification are shown in Table 4. As displayed, various classifiers use different classification methods. That is why the resulted interconnections are with different attributes; however one of them, namely 329, dominates significantly. Prominently displayed are the differences of each algorithm in the rule induction process and the contents of the rules (see Table 4).

Table 4.

Rules

Classif.	**Rule**
Ridor	IF 329 <= 0.852165 THEN Category = Br
PART	IF 329 > 0.835414 AND 1142 <= 1.132913 THEN Category=HD
	IF 568 <= 0 AND 223 <= 0.851652 THEN Category=Br
ONE R	IF 329< 0.78095255 THEN Category=Br
	IF 329>= 0.78095255 THEN Category=HD
JRIP	IF 958 <= 0.837699 and 115 <= 0.845161 THEN Category=Br
	IF 329 <= 0 THEN Category=Br
FLR	IF 329 [0.835414, 0.8458] THEN Category=Br (CF = 0.74)
	IF 997 [0, 0.815175] and 115[0.845161, 0.853948] THEN Category=Br (CF = 0.77)
	IF 501 [1.024421, 1.028017]) THEN Category=HD (CF = 1.0)
	IF 87 [0, 0.803008] and 104 [1.259394, 1.283743] THEN Category=HD (CF = 1.0)
	IF 997 [1.021784, 1.026568] THEN Category=HD (CF = 1.0)

Classif.	Rule
FURIA	IF data in interval [0.0 1.7643473] [0.0 1.1474993] [0.0 1.2891427] [0.0 1.0846912] [0.744659 1.3009045] [0.0 1.0004329] [0.0 1.2837429] [0.6871632 1.108951] [0.0 1.1201585] [0.0 1.4158141] [0.0 1.6925139] [0.0 1.0963819] [0.0 1.2071105] [0.0 0.8506509] [0.9561629 1.1801563] [0.0 1.2243818] [0.9354909 1.3019569] [0.0 1.5064373] [0.0 1.0559155] [0.0 1.1105751] [0.0 1.2031143] [0.0 2.3868061] [0.0 1.2479679] [0.0 2.1671135] [0.8866751 1.0796594] [0.0 1.260543] [0.0 0.8510461] [0.0 1.3208536] [0.8985972 1.2919535] [0.9869045 1.840857] [0.0 1.2867797] ….. [0.9319126 1.2546412] [0.7294844 1.1581615] [0.0 1.1322578] THEN Class=Br, et al.

Conclusions and future work

As expected, the results are difficult to evaluate, because the number of cancer patients in the training data set is very small; even after performing 10 fold cross-validation the results are still not satisfactory, because the best result in classifying cancer patients is 69%, which means that only in a bit more than half of the cases in this classification has been correct. Due to that, it is necessary to use other methods for data classification.

However, it should be emphasized that the use of fuzzy classification methods produces higher-quality results and comparably the best result of the crisp methods of Brain Cancer classification is 53% as compared with 69% obtained by fuzzy method (FLR).

In general, it can be concluded that the FURIA algorithm shows worse results than FLR; however, to objectively assess capabilities of this algorithm it should be possible to make comparisons with other publicly available data sets whose classification results are publicly available. Of course, the main advantage of the fuzzy rule-based technique is the decision making process. Each person can easily and intuitively perceive the classification process, as it operates with IF -THEN rules, which are closer to the real, everyday language. In this situation rules inducted by FURIA are better understandable. So the main issue is classification accuracy or well understandable rules.

The directions of further research include studying other types of cancer separately and also researching all of the available types of cancer with data of healthy donors. It is necessary to continue research on different methods of important attributes selection to optimize the required work.

Acknowledgments

This work has been developed within LATVIA – BELORUS Co-operation programme in Science and Engineering within the project «Development of a complex of intelligent methods and medical and biological data processing algorithms for oncology disease diagnostics improvement», Scientific Cooperation Project No. L7631.

Thanks to Dr.habil.sc.comp. Professor Arkady Borisov (Riga Technical University) for help and support.

The used data set is from Latvian BioMedical Research & Study Center.

References

1. T.R. Golub, D.K. Slonim et.al. Huerta, Molecular Classification of Cancer: Class Discovery and Class prediction by gene expression Monitoring, Science, vol. 286, pp. 531-537, 1999.
2. S.A. Vinterbo, E.-Y. Kim, L. Ohno – Machado, Small, fuzzy and interpretable gene expression based classifiers, Bioinformatics, vol. 21, no. 9, pp. 1964-1970, 2005.

3. S.-Y. Ho, C.-H. Hsieg, H.-M. Chen, H.-L. Huang, Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis, BioSystems, vol. 85, pp.165-176, 2006.
4. G. Schaefer, Fuzzy Rule-Based Classification Systems and Their Application in the Medical Domain: 16th International Conference on Soft Computing MENDEL 2010, June 23-25, 2010, Brno, Czech Republic. Brno University of Technology, pp. 229-235.
5. Gasparoviča M., Novoselova N., Aleksejeva L. Using Fuzzy Logic to Solve Bioinformatics Tasks // Scientific Journal of Riga Technical University. Issue 5, Computer Science. Information Technology and Management Science, vol.44, pp.99-105, 2010.
6. Popular Medical Encyclopedia. –Rīga : Galvenā enciklopēdiju redakcija. pp. 623, 1984 (in Latvian).
7. Kalniņa Z, Siliņa K, Meistere I, Zayakin P, Rivosh A, Ābols A, Leja M, Minenkova O, Schadendorf D and Linē A. Evaluation of T7 and Lambda phage display systems for survey of autoantibody profiles in cancer patients. J. Immunol. Methods, vol. 334(1-2) pp.37-50, 2008.
8. Hühn J., Hüllermeier E. FURIA: an algorithm for unordered fuzzy rule induction// Data Mining and Knowledge Discovery, Springer Netherlands, Computer Science, Volume: 19, Issue: 3, pp. 293-319, 2009.
9. Cohen W. Fast effective rule induction // Proceedings of the 12th International Conference on Machine Learning, ICML, pp. 115 – 123, 1995.
10. Kaburlasos V. G., Athanasiadis I. N., Mitkas P. A. Fuzzy lattice reasoning (FLR) classifier and its application for ambient ozone estimation, International Journal of Approximate Reasoning, Volume 45, Issue 1, pp. 152-188, 2007.
11. Gaines B.R., Compton P. Induction of Ripple-Down Rules Applied to Modeling Large Databases. J. Intell. Inf. Syst., vol. 5, issue 3, pp. 211-228, 1995.
12. Frank E., Witten I.H. Generating Accurate Rule Sets Without Global Optimization // Fifteenth International Conference on Machine Learning, pp. 144-151, 1998.
13. Holte R.C. Very simple classification rules perform well on most commonly used datasets. Machine Learning. Vol. 11, pp. 63-91.
14. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P, Witten I.H. The WEKA Data Mining Software: An Update. SIGKDD Explorations, vol.11, issue 1, 2009.

Anotācija. *Šajā rakstā pētīti reāli smadzeņu vēža dati. Šāda veida bioloģiskajiem datiem ir savas īpatnības – liels atribūtu – gēnu, antivielu skaits. Toties ierakstu skaits ir neliels, jo datus nepieciešams iegūt no reāliem pacientiem, šāds process ir lēns un ar lielām materiālajām izmaksām. Tāpēc darbā dots izmantoto datu sīks apraksts, analizētas to īpatnības, veids un struktūra. Attiecīgi arī klasifikācijas likumu atklāšana šādos datos ir sarežģīta. Šis pētījums veltīts klasifikācijas metožu pielietošanai ar mērķi atrast likumsakarības, kuras būtu iespējams izmantot smadzeņu vēža klasifikācijā. Tieši darbam ar šādiem unikāliem datiem ir ārkārtīgi liela praktiska vērtība, jo tos iespējams nākotnē izmantot turpmāko pētījumu veikšanai, kā arī praktiskam pielietojumam diagnosticēšanā ar iespēju nākotnē tos piedāvāt arī interpretēšanai biologiem. Lai paātrinātu likumsakarību iegūšanu, tika izmantoti tikai nozīmīgie atribūti. Tika pielietotas vairākas metodes nozīmīgo atribūtu iegūšanā. Izdarīti secinājumi par šāda veida datu apstrādi, klasifikācijas likumu iegūšanu, iegūto likumu precizitāti, kā arī aprakstīti nākotnē plānotie darbi.*