Scientific Journal of Riga Technical University
Computer Science. Information Technology and Management Science

2011
Volume 49

Impact of Antibody Panel Size on Classification Accuracy

Inese Polaka[1], Arkady Borisov[2], [1-2]*Riga Technical University*

***Abstract* – This paper experimentally studies the influence of antibody panel size reduction on classification results. The presented study includes four classification methods and five feature evaluators that are applied to five different biomedical data sets with large dimensionality (1200 features). The behaviour of the classifiers in these data sets is examined to reveal overall trends of dimensionality reduction impact on classification accuracy.**

***Keywords* – bioinformatics, classification accuracy, dimensionality reduction, feature selection**

I. Introduction

Antibody phage display technology has made it possible to examine hundreds and thousands of antibodies that indicate immune responses of human body to different diseases. The display vector holds a lot of information about a patient's health but not all of this information is useful, especially if the focus is turned to some specific diseases that are hard to diagnose using the traditional symptomatic medicine. One of these specific diseases that is particularly interesting for researchers is cancer, because timely diagnosis often means higher survival chance and less invasive treatment with lighter side effects; but diagnosing cancer at the early stages is difficult because it does not manifest itself with implicit and obvious symptoms. One shortcoming of this possibility is that cancer indicating antibodies are not thoroughly studied and often not known. Also identifying and studying the auto-antibodies that human body produces as a response to cancer is very important in synthesizing antibodies for immunotherapy that could treat the disease by stimulating the immune system of a patient to attack the disease using the therapeutic antibodies.

To find these disease-related antibodies, researchers have used various statistical methods but also data mining techniques have been found to be very effective and less demanding towards data properties. One of the approaches is using classification algorithms to build classification models that represent the response of human body indicating possible marker antibodies. Another way is dimensionality reduction by feature selection finding the most important antibodies and the most efficient antibody panel (feature subset). However, these approaches are not mutually exclusive and can be used together to achieve better results. It has been previously proven that dimensionality reduction not only decreases the computing time but also increases the classification accuracy even for classification methods with good scalability. Therefore it helps finding the most important antibodies and the classification model illustrates the relationships and processes between antibodies. But it is also important to define whether the first selection of antibodies provides the best subset or it is only a working sample that is further decreased in size by building a classification model. Therefore this study investigates the impact of feature subset size on classification size. Although some classification models would successfully implement ten or twenty features, there are few that would use all hundred or two hundred features. Thus it shows the significance of classification in the further selection of the indicative antibody panel.

It is important to discard the uninformative antibodies not only because of the lack of correlation between a feature and the class variable but also because of existing correlation between two features that obviously hold redundant information and are not desired in the resulting panel of marker antibodies. Smaller feature subset (antibody panel) sizes also mean clearer data visualization and help to perceive the data.

This study introduces feature ranking methods and classification algorithms and later describes the use of both on antibody phage display data sets. The paper is organized as follows: Section 2 describes the used methods and the basics of their functioning; Section 3 introduces the experimental setup used in the study and Section 4 presents the results of the study and discussion.

II. Methods

This study uses feature selection methods based on ranking to successfully test various feature subset (antibody panel) sizes by selecting the top 10, 20, 50, 100 or 200 features. Also several feature evaluation approaches were used to dismiss the impact of the feature selection method on the classification process. Classification is performed using different classification approaches to reduce the preference of one method that would be more appropriate for one data set and perform badly in other data sets.

A. Ranking Methods

Ranking-based feature search methods evaluate single features using various metrics and assign a rank to each feature based on the performance of the feature. Ranking methods can filter the top features based on the metric and a predefined subset size. The evaluation metrics are usually based on statistical properties of features or the predictive potential of a feature.

One of the metrics used in ranking is *Chi-Square Statistic* (Chi in graphs and tables) that is calculated with respect to the

class [1]. It also works with discrete data types. The statistic for a problem with *k* classes and *N* instances is calculated as shown in Equation 1.

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}} \quad (1)$$

where A_{ij} is the number of instances in the *i*-th interval (with *i*-th value), *j*-th class,

E_{ij} is the expected frequency of A_{ij}, which is calculated as shown in Equation 2.

$$E_{ij} = \frac{R_i \cdot C_j}{N} \quad (2)$$

where R_i is the number of instances in the *i*-th interval,

C_j is the number of instances in the *j*-th class.

Another popular metric to evaluate features is *Information Gain* (IG in graphs and tables) that is measured with respect to the class. Information Gain is used in decision tree induction and was introduced by J. R. Quinlan [2]. Prior to feature evaluation the numeric attribute values have to be discretized because this approach works with categorical data. This metric is based on the change of information entropy that would occur if the state of the information change (some information is given) and can be calculated by subtracting conditional entropy of the class from its entropy. The entropy of a feature C is calculated as shown in Equation 3. Conditional entropy of feature C if the state of feature A is given is calculated as shown in Equation 4.

$$\mathrm{H(C)} = -\sum_{i=1}^{n} P(C = c_i) log_2(P(C = c_i)) \quad (3)$$

$$\mathrm{H(C|A)} = -\sum_{j=1}^{k} P\left(A = a_j\right) H(C|A = a_j) \quad (4)$$

where $P(C=c_i)$ is relative appearance frequency of value c_i in feature *C* in the data set,

$H(C|A=a_j)$ is the entropy of feature *C* in the data subset where the value of attribute *A* is a_j.

Gain Ratio (GR in graphs and tables) is another metric used to evaluate features in decision tree induction [2]. It is based on Information Gain metric and eliminates its weakness that occurs in data sets that have features with large numbers of unique values which are given preference over other possibly better features with fewer values. Therefore Gain Ratio divides Information Gain by entropy of the considered feature as shown in Equation 5.

$$GR(C, A) = \frac{\mathrm{H(C)} - \mathrm{H(C|A)}}{H(A)} \quad (5)$$

Another classification method that can be used as a basis for feature selection is the rule induction algorithm *OneR* [3]. It also discretizes numeric features (using minimum bucket size as the criteria) and evaluates each feature using its error rate. *OneR* generates one rule for each feature and evaluates how this rule classifies the data. This classification error is also used to rank features in this feature selection approach.

Relief algorithm [4] evaluates a feature by randomly sampling instances and analyzing two neighbouring instances of the same and different classes. This algorithm was not able to work with missing data and data sets that included three or more classes; therefore it was improved resulting in *Relief-F* algorithm [4]. It is adapted to work with multi-class problems by finding one or more (*k*) neighbouring instances *MI* from each different class *C* and averages their contribution for upgrading estimates *W[A]* weighting it with the prior probability of each class. The estimation of weight *W* of feature *A* when the sampled instance is *R* (which is sampled m times) and the nearest instance of the same class is *H* is conducted as shown in Equation 6 [4].

$$W[A] := W[A] - \sum_{j=1}^{k}\frac{diff(A,R,H_j)}{m \cdot k} + \sum_{C \neq class(R)}\frac{P(C)}{1-P(class(R_i))}\sum_{j=1}^{k}\frac{diff\left(A,R,M_j(C)\right)}{m \cdot k} \quad (6)$$

The number of the checked neighbouring instances is determined by either predefining a number or the maximum distance. The difference *diff(A, I_1, I_2)* for discrete features is one if the values of instances are equal and 0 if the values are different. The difference of numeric features is calculated as shown in Equation 7.

$$diff(A, I_1, I_2) = \frac{|value(A,I_1) - value(A,I_2)|}{\max(A) - \min(A)} \quad (7)$$

B. Classification Methods

To evaluate feature subsets, different classification methods are used – decision function classification using support vector machines, probabilistic classification using Naïve Bayes method, decision tree induction algorithm C4.5 and tree ensemble Random Forest. The choice of classification algorithms is based on a number of studies on gene expression classification techniques that deal with similar problems [5]. The results have shown that support vector machines and Random Forests perform best on high-dimensional data but C4.5 and other decision tree classifiers not only perform well but also allow extracting knowledge about feature relations by constructing rules including several antibodies (or genes in gene expression studies); Naïve Bayes classification algorithm is a standard and best-performing probabilistic classification algorithm in similar tasks.

Support vector machine (SVM) builds a function of relevant features by assigning weights to them (irrelevant features are assigned weight 0) based on relevant instances (support vectors). The function is a hyperplane in the instance space that separates different classes with a maximum margin (distance from the hyperplane to the nearest instances). SVMs have various types and enhancements; this study employs an enhancement called *Sequential Minimal Optimization* (SMO) introduced by Platt [6] that is used for training support vector classifiers. It was also improved by Keerthi and Shevade [7]. This approach breaks training process into smaller, two-dimensional problems and reduces resource consumption comparing to large matrix computation needed for the classic

SVM training. SVMs also use kernels to transform feature spaces where they search for hyperplanes. In this study the Polynomial kernel was used to represent dot products.

While SVMs only work with binary classes, the multi-class problem is solved using pairwise classification 1-vs-1 (*pairwise coupling method*) proposed by Hastie and Tibshirani [8].

The *Naïve Bayes* classifier (NB in graphs and tables) uses probabilistic knowledge to assign class values [9]. It assumes that features are conditionally independent (hence the naïve approach) and predicts the most probable class according to class probabilities that are calculated for class set C with value c and feature value vector X with values x as shown in Equation 8.

$$P(C = c|X = x) = \frac{p(C=c)p(X=x|C=c)}{p(X=x)} \quad (8)$$

C4.5 is a decision tree induction algorithm proposed by Quinlan [2]. The trees are constructed from a data set by dividing the training set into subsets until a class value can be assigned to each subset. The tree construction starts with choosing a root node representing a feature that splits the initial data set into subsets according to its values. Then nodes are selected for the second level split and so on. The features are chosen based on evaluation using *Gain Ratio* (described previously). *Random Forest* is an ensemble of random trees [10]. Random trees are constructed considering a predefined number of randomly chosen features. In these experiments the Forest consists of ten trees each considering 11 features (this number *k* is determined based on the number of instances *N* in the data set using Equation 9).

$$k = \log_2 N + 1 \quad (9)$$

Then the class to assign to a new instance in classification process is chosen using the most frequent tree output.

III. Experimental Setup

The experiments were carried out using five different data sets that hold antibody display data of patients and healthy donors and that were provided by Latvian Biomedical Research and Study Center. The patients whose data were featured in the data sets were affected by breast cancer (data set marked as Br in further text), gastric cancer (data set Ga), gastrointestinal series (data set GIS), melanoma (data set Mel) or prostate cancer (data set Pr). Each data set held information about 1230 antibodies and consisted of few hundred records (Breast cancer data set had 168 records, Gastric cancer data set had 328 records, Gastrointestinal series data set had 281 records and Melanoma data set had 343 records and Pr had 207 records).

Each of the data sets was used to select the most important features and then to test these subsets by classifying data. Feature selection ranged the features by their importance (various metrics described previously) and created a list; then the top 10, 20, 50, 100 and 200 features were chosen to form data subsets and evaluate them using the previously described classification algorithms and 10-fold cross-validation.

To eliminate the influence of feature selection methods and classification algorithms, the average classification accuracies across different feature subsets of the same size and across different classifiers were used to estimate the most suitable antibody panel size.

IV. Results and Discussion

The experiments were carried out using five different data sets and the results varied a lot depending on the features of each single data set; but the trends were similar for all data sets, so the following generalizations can be made:

- the accuracy of single experiments with reduced data sets improves significantly when the dimensionality is reduced;
- the average accuracy improves when the feature subset size increases but the improvement becomes ever less significant and the results worsen with the full feature subset;
- the average accuracy when using 100-200 features is very similar but higher than the results obtained using full data sets.

Figure 1 depicts the average accuracies of classifiers using reduced data sets (the top 10, 20, 50, 100 and 200 features) and the full data set (1200).

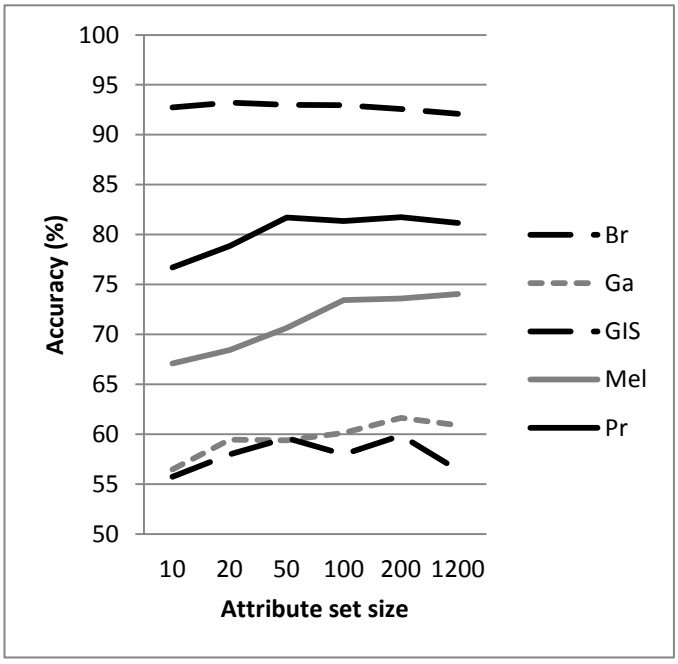

Fig. 1. Average classification accuracy of all classifiers across different data sets and numbers of attributes in a subset (Br – breast cancer data; Ga – gastric cancer data; GIS – gastro-intestinal series data; Mel – melanoma data; Pr – prostate cancer data).

Table I shows the loss (negative numbers) or gain (positive numbers) in accuracy comparing the classification results (classifier accuracy) using the full data set and the reduced data sets.

TABLE I
CHANGES IN CLASSIFICATION ACCURACY

	10	20	50	100	200
Br-NB	1.79	2.50	2.38	2.26	2.62
Br-SVM	-0.71	0.48	0.36	1.43	0.83
Br-C45	1.67	1.19	0.83	-0.71	-1.67
Br-RF	-0.24	0.24	0.00	0.48	0.12
Br average	0.63	1.10	0.89	0.86	0.48
Ga-NB	-4.70	-3.66	-3.66	0.98	0.12
Ga-SVM	-11.65	-6.65	-6.77	-6.22	-3.48
Ga-C45	-4.76	-0.30	-0.61	-2.13	-0.30
Ga-RF	3.35	4.88	5.06	4.27	6.65
Ga average	-4.44	-1.43	-1.49	-0.78	0.75
GIS-NB	0.50	2.06	5.20	2.35	3.77
GIS-SVM	-3.84	0.00	2.49	2.92	2.35
GIS-C45	0.43	1.71	0.64	0.14	1.21
GIS-RF	0.28	2.56	4.56	0.85	6.55
GIS average	-0.66	1.58	3.22	1.57	3.47
Mel-NB	-7.11	-7.99	-7.52	-4.49	-1.17
Mel-SVM	-16.03	-13.18	-6.65	-2.68	-2.62
Mel-C45	1.98	3.21	1.69	2.86	0.00
Mel-RF	-6.65	-4.55	-1.17	1.87	1.92
Mel average	-6.95	-5.63	-3.41	-0.61	-0.47
Pr-NB	-7.54	-2.90	-1.64	-2.42	0.00
Pr-SVM	-11.40	-8.70	-5.99	-5.41	-1.84
Pr-C45	4.93	2.90	4.73	5.12	1.35
Pr-RF	-3.77	-0.58	5.12	3.48	2.80
Pr average	-4.44	-2.32	0.56	0.19	0.58
Average	**-3.17**	**-1.34**	**-0.05**	**0.25**	**0.96**

The average loss in accuracy decreases as the data subset size increases but this trend is not definite in all data sets. Although the accuracy is lower with very small number of features (10 feature subsets) in almost all data sets, it gets better with subset sizes of 50 features. The results also vary a lot depending on classification methods and the specifics of the data sets. SVM method seems to benefit the least from dimensionality reduction which is understandable because this method is scalable and uses all available information in the creation of classification patterns. It is very similar to Naïve Bayes method that uses information available in all attributes to construct the best classifier and therefore loses in accuracy with very small data sets although the results with Breast cancer and Gastrointestinal series data sets were better after feature space reduction, which could be due to specifics of the features – there is a small group of features that hold the most information responsible for correlation with the class variable. The two methods based on decision tree classifiers have similar constructions that take part in classification process but the influence of dimensionality reduction by feature selection is very different on both of these methods and often has opposite effect – if the accuracy of one method increases, the result of the other method improves. Another exception is the Melanoma data set that has all of the accuracy changes as negative numbers and the only method that gains accuracy from dimensionality reduction is C4.5.

The results of classification experiments on all data sets using all classification methods and feature subsets chosen by all methods are shown in Table II.

The results demonstrate that the single best results for each data set are obtained using different classification algorithms, feature subset sizes and feature evaluation approaches. This stresses one of the main problems of classification in data mining, which is the problem of choosing the most suitable pre-processing and classification techniques and methods for a certain data set to achieve the best results.

The best results for Breast cancer data set were achieved by SVM classification method that was applied to the feature subset with 100 best features as evaluated by Chi-square statistic and Gain Ratio (both feature subsets had the same classifier accuracy). Classification accuracy was 96.43%, which is almost by 3% higher than SVM accuracy on the full data set (93.45%). The best accuracy using the full Breast cancer data set was also achieved by SVM (see Table I).

Support vector machines also had the highest classification accuracy for Gastric cancer data set but in this case the best feature subset size was 200 best features and the best evaluation metrics were OneR and ReliefF (both had the same accuracy 69.51%). The best result using the full data set was obtained using SVM and was 67.68%.

Gastrointestinal system data set was the most difficult to classify and had the lowest classification accuracies. The best result was obtained using Random Forest with the 20 best features as evaluated by OneR metric. This result was 69.40% that was 15% improvement from 54.45% using the full data set for classification using Random Forest. The best result using the full data was obtained by again using SVM (59.43%).

The best result for Melanoma data set was reached by SVM classifier using 100 features with the highest Gain Ratio (83.80%). SVM was also the most precise classifier using the whole data set with 81.34%. The situation was very similar with Prostate cancer data set where the best result was achieved by SVM for both the full data set (89.37%) and the reduced data set with 200 features selected by Chi-square statistic and Gain Ratio (both feature subsets had the same classifier accuracy – 91.30%).

All of the single best results were significantly higher than the average results because there were results that were worse than the initial classification accuracies obtained using the full data sets. This shows that it is still a hard task to choose the best classification approach and the right data pre-processing technique. But dimensionality reduction by feature selection can improve the classification accuracy even for the methods that are scalable if the right method is chosen. Dimensionality reduction also helps to decrease the time required for classification by eliminating uninformative features from analysis; it also reduces the sizes of classifiers whereas the many uninformative features are replaced by fewer more

TABLE II. CLASSIFICATION RESULTS ACROSS DIFFERENT DATA SETS

		Breast cancer					*Gastric cancer*					*Gastro-intestinal series*					*Melanoma*					*Prostate cancer*				
		10	20	50	100	200	10	20	50	100	200	10	20	50	100	200	10	20	50	100	200	10	20	50	100	200
Naive Bayes	Chi	93	94	94	94	95	58	55	56	62	60	58	59	61	58	59	74	73	66	69	71	79	85	86	83	85
	GR	93	95	95	93	95	55	55	56	62	60	57	59	61	58	59	51	52	63	67	72	82	86	86	83	86
	IG	95	95	94	94	95	55	55	56	61	60	58	59	61	58	59	68	68	66	68	71	79	84	86	83	85
	OneR	91	92	92	93	92	53	58	56	59	60	61	63	63	63	62	72	71	64	73	76	75	81	82	87	88
	ReliefF	92	92	92	92	92	53	58	56	59	60	51	52	62	58	62	63	60	67	65	69	80	83	85	85	88
	Average	93	94	93	93	94	55	56	56	61	60	57	59	62	59	60	66	65	65	68	72	79	84	85	84	86
SVM	Chi	91	95	95	96	96	61	63	59	59	60	55	60	63	63	64	71	71	77	79	80	79	83	87	89	91
	GR	92	95	95	96	95	59	60	61	58	61	55	61	63	63	64	55	57	75	83	79	79	84	87	86	91
	IG	96	95	95	96	96	59	61	59	59	60	56	61	62	64	63	71	71	74	79	81	79	85	88	89	91
	OneR	92	93	92	93	93	50	61	63	66	70	57	60	64	64	65	64	71	73	78	78	75	75	76	77	80
	ReliefF	92	92	92	92	91	50	61	63	66	70	55	55	57	58	53	65	70	74	75	75	77	77	78	80	85
	Average	93	94	94	95	94	56	61	61	61	64	56	59	62	62	62	65	68	75	79	79	78	81	83	84	88
C4.5	Chi	93	92	91	89	88	60	60	59	59	60	57	56	56	53	54	70	65	68	70	69	79	77	78	79	74
	GR	93	92	91	89	88	57	62	60	58	60	57	56	56	53	59	66	73	70	71	69	81	79	78	77	74
	IG	92	92	90	88	88	57	59	59	59	60	56	56	56	53	54	71	67	68	72	69	79	76	77	79	74
	OneR	93	93	95	95	95	51	58	59	56	58	54	64	62	59	60	72	78	69	68	63	75	72	78	76	75
	ReliefF	92	92	92	91	88	51	58	59	56	58	54	54	51	58	55	66	68	69	69	65	73	72	75	76	72
	Average	93	92	92	90	89	55	59	59	58	59	56	57	56	55	56	69	70	69	70	67	77	75	77	78	74
RandomForest	Chi	93	94	92	93	93	59	64	58	59	64	52	57	57	50	60	77	76	78	78	79	81	85	88	84	81
	GR	94	95	93	95	93	62	65	61	61	59	54	53	57	54	60	66	74	78	80	77	84	85	87	80	81
	IG	93	93	94	94	94	62	61	64	59	64	51	53	57	54	62	76	75	78	80	79	84	85	85	84	78
	OneR	93	94	92	92	92	58	58	62	62	64	62	69	63	58	62	56	57	62	69	74	45	52	72	76	80
	ReliefF	90	89	92	92	92	58	58	62	62	64	54	53	60	60	60	67	70	73	76	75	70	72	75	75	75
	Average	93	93	93	93	93	60	61	61	61	63	55	57	59	55	61	68	70	74	77	77	73	76	81	80	79

informative ones. Overall the best results were obtained using SVM classifiers (all of the best accuracies using full data sets were achieved using this method) and Random Forests.

Another conclusion can be made about the optimal feature subset size – using all features does not mean obtaining the most accurate classifier; the best results in this study were achieved using from one hundred to a few hundred features. Although this number is too big for a diagnostic anti-body panel, it helps eliminating the uninformative anti-bodies and the classifiers then can be used to reduce this number even more.

ACKNOWLEDGMENTS

This work has been supported by the European Social Fund within the project «Support for the implementation of doctoral studies at Riga Technical University». This work has been developed in 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.

REFERENCES

[1] Witten, I. H., Frank, E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann series in data management systems. San Mateo: Morgan Kaufmann Pub., 2005. 560 p.

[2] Quinlan J. R. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Pub., 1993. 302 p.

[3] Holte, R. C. Very simple classification rules perform well on most commonly used datasets. Machine Learning 11-1, 1993, p. 63–90.

[4] Kononenko, I. Estimating attributes: analysis and extensions of RELIEF. *Proceedings of the European Conference on Machine Learning* (ECML-94), Catania, Italy, April 6-8, 1994. Secaucus: Springer-Verlag New York, Inc., 1994, p. 171-182.

[5] Poļaka I., Tom I., Borisovs A. Decision Tree Classifiers in Bioinformatics. Scientific Journal of RTU. 5. Series, Computer Science, Information Technology and Management Science 44, 2010, p. 118-123.

[6] Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In: B. Schoelkopf and C. Burges and A. Smola (eds), Advances in Kernel Methods - Support Vector Learning. Cambridge, MA, USA: The MIT Press, 1998, 386 p.

[7] Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., Murthy, K. R. K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13-3, 2001, p. 637-649.

[8] Hastie, T., Tibshirani, R. Classification by Pairwise Coupling. Annals of Statistics 26-2, 1998, p. 451-471.

[9] John, G. H., Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995. San Mateo: Morgan Kaufmann Pub., 1995, p. 338-345.

[10] Breiman, L. Random Forests. Machine Learning 45-1, 2001, p. 5-32.

Inese Polaka is a second year Doctoral student at Riga Technical University. She finished her Master studies at Riga Technical University majoring in Information Technology in 2010.

Her research interests include machine learning methods and classification tasks in bioinformatics, decision tree classifiers, classifier efficiency improvement methods, use of ontology in machine learning, ontology based classifier design, descriptive statistics and exploratory data analysis.

Arkady Borisov holds a Doctoral degree in Technical Sciences in the field of Control in Technical Systems from Taganrog State Radio-Engineering University and the Dr.habil.sci.comp. degree from the Latvian Council of Science.

He is Professor of Computer Science in the Faculty of Computer Science and Information Technology at Riga Technical University (Latvia). His research interests include fuzzy sets, fuzzy logic and computational intelligence. He has 205 publications in the field.

He has supervised several national research grants and participated in the European research project ECLIPS.

Inese Poļaka, Arkadijs Borisovs. Antivielu paneļa lieluma ietekme uz klasifikācijas precizitāti

Rakstā apskatīta datu augstas dimensionalitātes problēma un tās ietekme uz datu ieguves klasifikācijas uzdevumu bioinformātikas sfērā. Tika veikta eksperimentālā analīze, izmantojot reālus datus, kas satur cilvēka antivielu ekspresijas līmeņu nolasījumus par 1230 antivielām, kas izanalizētas vēža (melanoma, krūts, kuņģa un prostatas vēzis) un kuņģa-zarnu trakta slimību pacientiem un veseliem indivīdiem. Lai noteiktu dimensionalitātes ietekmi uz klasifikācijas rezultātiem, tika veikta atribūtu (antivielu) atlase, samazinot atribūtu kopu līdz 10, 20, 50, 100 un 200 atribūtiem, izmantojot dažādas atribūtu novērtēšanas metodes (tādi mēri kā Hī kvadrāta metodes novērtējums, *Gain Ratio*, *Information Gain*, kā arī *OneR* precizitāte un metode *ReliefF*). Tad šajās samazinātās dimensionalitātes datu kopās tika veikta klasifikācija, izmantojot metodes, kas implementē dažādas pieejas – uz varbūtību balstītais Naivais Baijesa klasifikators, koku klasifikatori C4.5 un Random Forest, kā arī atbalsta vektoru mašīnas SVM. Klasifikācijas rezultāti salīdzināti ar pilno datu kopu rezultātiem, kas ļauj izdarīt secinājumus par antivielu paneļa lieluma ietekmi uz klasifikācijas precizitāti, kā arī par atribūtu atlases izmantošanas iespējām diagnostiska antivielu paneļa izveidē. Rakstā sniegts arī izmantoto metožu apraksts. Rezultāti parāda, ka arī metodes, kas labi darbojas ar augstas dimensionalitātes datiem, precizitātes ziņā iegūst, ja dimensionalitāti samazina, jo tiek atmesti atribūti, kas veicina pārapmācību un samazina izveidoto klasifikācijas modeļu spēju pareizi atpazīt līdz tam neredzētus ierakstus. Tāpat pēc eksperimentu datiem var secināt, ka atkarībā no datu rakstura atsevišķas atribūtu novērtēšanas metodes darbojas labāk par citām, taču nav izteikti labākas metodes, tāpēc rezultāti analizēti pēc vidējām vērtībām visu metožu rezultātiem.

Инесе Поляка, Аркадий Борисов. Влияние размерности пространства признаков антител на точность классификации

В статье рассматривается проблема влияния уровня размерности начальных данных на результаты интеллектуального анализа и классификации в области биоинформатики. Произведён экспериментальный анализ реальных данных пациентов, проанализированы 1230 антител на признаки рака (меланомы, молочной железы, желудка и предстательной железы) и заболеваний желудочно-кишечной системы как пациентов, так и здоровых доноров. Для получения данных о влиянии размерности на результаты классификации всё пространство признаков антител было уменьшено до 10, 20, 50, 100 и 200 признаков с использованием различных вычислительных методов (метрика хи-квадрат, Gain Ratio, Information Gain, а также - проверка точности OneR и метод ReliefF). Полученные наборы данных были использованы для классификации с использованием различных методов — вероятностный классификатор Naive Bayes, деревья решений с использованием алгоритмов C4.5 и Random Forest, а также - метод опорных векторов (SVM). Результаты классификации были сравнены с результатами классификации по полным наборам данных, что позволило получить оценку влияния размерности пространства признаков антител на точность классификации, а также оценить потенциал использования выборок атрибутов антител для диагностики. В статье также представлены описания использованных методов и проведённых экспериментов. Результаты показывают, что даже методы, которые хорошо себя показали с данными с высокой размерностью, работают ещё лучше при сокращении размерности из-за удаления атрибутов, которые способствуют переобучению и снижают способность модели классификации правильно распознавать ранее неизвестные записи. Кроме того, экспериментальные данные показывают, что, в зависимости от характера данных, некоторые методы оценки атрибутов работают лучше, чем другие; но не существует метода, исключительно превосходящего другие, поэтому результаты были проанализированы, используя средние значения результатов всех методов.