ISSN 1691-5402
ISBN 978-9984-44-071-2
Environment. Technology. Resources
Proceedings of the 8th International Scientific and Practical Conference. Volume I1
© Rēzeknes Augstskola, Rēzekne, RA Izdevniecība, 2011

FEATURE SELECTION APPROACHES IN ANTIBODY DISPLAY DATA ANALYSIS

Inese Polaka
Institute of Information Technology Riga Technical University
1 Kalku str, Riga, LV 1658, Latvia, e-mail: Inese.Polaka@rtu.lv

Abstract. *Molecular diagnostics tools provide specific data that have high dimensionality due to many factors analyzed in one experiment and few records due to high costs of the experiments. This study addresses the problem of dimensionality in melanoma patient antibody display data by applying data mining feature selection techniques. The article describes feature selection ranking and subset selection approaches and analyzes the performance of various methods evaluating selected feature subsets using classification algorithms C4.5, Random Forest, SVM and Naïve Bayes, which have to differentiate between cancer patient data and healthy donor data. The feature selection methods include correlation-based, consistency based and wrapper subset selection algorithms as well as statistical, information evaluation, prediction potential of rules and SVM feature selection evaluation of single features for ranking purposes.*

Keywords: *antibody display, classification, data mining, feature selection, ranking*

Introduction

In the last decade new technological advancements have made molecular diagnostics more accessible and it has become a popular and perspective field of research [1]. While gene, protein and antibody analysis and screening techniques are developed, the analysis techniques of the resulting data to extract new knowledge are less than satisfactory. The statistical approaches that are often used are demanding towards data and provide little useful information to help understand relationships between features and prognostic capabilities of features.

Antibody display data analysis is a relatively new approach and is less studied than other molecular diagnostics approaches but it has similar problems – high dimensionality (thousands of antibodies) and small numbers of instances due to high costs of experiments. Most classification methods are very sensitive to data dimensionality and the instance/feature ratio but the less sensitive ones are also shown to benefit from dimensionality reduction [2]. Therefore this study is dedicated to analyzing feature selection techniques known in data mining and investigating their performance in antibody display data.

Methods

The antibody selection can be performed using standard data mining techniques. All of the techniques can be divided into two major groups – subset selection and individual attribute ranking. Attribute ranking evaluates each attribute independently of others and does not consider dependencies between attributes. Subset selection in its turn searches for a set of attributes that together give the best result. The choice of the approach depends on the data features but subset selection has another advantage – it can provide more information about patterns in the data by explaining relationships between attributes. But subset selection methods and wrappers, in particular, have higher computation costs which can be an important matter in high-dimensional data.

Subset selection methods

Feature subset selection algorithms perform a search over the feature space to select the optimal subset. To perform the search they have to address four basic issues [3]:

- Starting point: starting with no features in the initial subset (*forward selection*) or starting with the full set of features (*backward elimination*);

- Search organization: consider each possible subset (*exhaustive search*) or locally changing the subset without returning to reconsider the change (*greedy search*); another possible approach is based on adding and removing a feature from the subset in each step to make the search more flexible (*stepwise selection*);
- Evaluation strategy: testing each feature of the subset individually (*filters*) against an evaluation merit or testing the whole subset (*wrappers*);
- Stopping criterion: lack of improvement on change, reaching the other end of the feature space or a particular subset size.

Correlation-based Feature Selector (CFS) is a filter algorithm that ranks feature subsets according to a correlation-based heuristic evaluation function that selects features highly correlated with the class feature and uncorrelated with each other [4]. It allows distinguishing features with a high predictive accuracy in the instance space that is not already covered by other selected features (the low inter-correlation of the selected features). The heuristic evaluation merit *M* for a subset *S* containing *k* features is calculated as shown in the Equation 1.

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \qquad (1)$$

where $r_{cf}$is the mean correlation between features and the class attribute,
r_{ff} is the average correlation between features.

Classifier Subset Evaluator uses classification algorithms applied to full data sets (or a division of the full data set into training and testing subsets for one run) to evaluate feature subsets. They are very similar to *Wrapper Subset Evaluators* but in this case the term Wrapper Subset Evaluators is used to address strategy that uses classification algorithms to evaluate feature subsets and cross-validation to estimate classification accuracy while fundamentally both, the Classifier Subset Evaluator and the Wrapper Subset Evaluator, are considered being wrappers. In both cases the classification process is treated as a black box giving evaluation values [5]. In this study all classification algorithms used to evaluate final subsets are employed to evaluate feature subsets while searching for the best combinations.

Consistency Subset Evaluator (CSE) evaluates feature subsets by the degree of consistency in class values when the training instances are projected onto the set, i.e. the prevalence of one class in subsets that the data set is divided into by attribute values. This also means that feature values have to be discretized [6]. Consistency of a subset can never surpass that of the full set so the algorithm searches for the smallest subset which has the same consistency as the full set.

The consistency of a feature subset *S* in a data set with *N* instances is calculated using the equation presented by Liu [7]:

$$C_s = 1 - \frac{\sum_{i=0}^{J} |D_i| - |M_i|}{N} \qquad (2)$$

where *J* is the number of distinct attribute value combinations,
$|D_i|$ is the number of occurrences of the *i*-th attribute value combination,
$|M_i|$ is the cardinality of the majority class for the *i*-th attribute value combination.

Ranking methods

Ranking 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 based on 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* that is calculated with respect to the class [8]. It also works with discrete data types. The statistic for a problem with *k* classes and *N* instances is calculated as shown in Equation 3.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij}-E_{ij})^2}{E_{ij}} \qquad (3)$$

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 4.

$$E_{ij} = \frac{R_i \cdot C_j}{N} \qquad (4)$$

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* that is measured with respect to the class. Information Gain is used in decision tree induction and was introduced by J. R. Quinlan [9]. 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 would change (some information is given) and can be calculated by subtracting conditional entropy of the class from its entropy. Entropy of a feature C is calculated as shown in Equation 5. Conditional entropy of a feature C if the state of feature A is given is calculated as shown in Equation 6.

$$\mathrm{H(C)} = -\sum_{i=1}^{n} P(C=c_i) log_2(P(C=c_i)) \qquad (5)$$

$$\mathrm{H(C|A)} = -\sum_{j=1}^{k} P\left(A=a_j\right) H(C|A=a_j) \qquad (6)$$

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 is another metric used to evaluate features in decision tree induction [9]. 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 7.

$$GR(C,A) = \frac{\mathrm{H(C)-H(C|A)}}{H(A)} \qquad (7)$$

Another classification method that can be used as a basis for feature selection is the rule induction algorithm *OneR* [10]. 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 [11] evaluates a feature by randomly sampling instances and analyzing two neighboring instances of 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 [12]. It is adapted to work with multi-class problems by finding one or more (*k*) neighboring instance *M(C)* 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 8 [13].

$$W[A] := W[A] - \sum_{j=1}^{k} \frac{diff(A,R,H_j)}{m \cdot k} + \sum_{C \neq class(R)} \frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} \frac{diff\left(A,R,M_j(C)\right)}{m \cdot k} \qquad (8)$$

The number of the checked neighboring 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 9.

$$diff(A, I_1, I_2) = \frac{|value(A,I_1) - value(A,I_2)|}{\max(A) - \min(A)} \qquad (9)$$

Another approach that can be used in feature selection is *Support Vector Machines* (SVM) that was proposed by Guyon et al. [1]. In feature ranking the feature evaluation is done by using the square of the weight assigned by the SVM. SVMs have the deficiency that they only work with binary classes therefore feature evaluation for multi-class problems is conducted by ranking attributes for each class separately using the one-vs-all method. SVMs build decision functions *D(x)* whose weights w_i are a function of a small subset of the training examples called support vectors [1]. The squares of sums of these weights assigned to features by support vectors are considered evaluation metrics in feature ranking.

Classification methods

To evaluate feature subsets, various classification methods are used – decision function classification using SVM, 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 [14-17]. The best results have shown that SVM and Random Forests perform best on such specific data but C4.5 and other decision tree classifiers not only perform well but also allow extracting knowledge about feature relations; Naïve Bayes classification algorithm is a standard and best-performing probabilistic classification algorithm.

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 [18] that is used for training support vector classifiers. It was also improved by Keerthi and Shevade [18]. 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 [20].

Naïve Bayes classifier uses probabilistic knowledge to assign class values [21]. 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 10.

$$P(C=c|X=x)=\frac{p(C=c)p(X=x|C=c)}{p(X=x)} \qquad (10)$$

C4.5 is a decision tree induction algorithm proposed by Quinlan [9]. The trees are constructed from a data set by dividing the training set into subset 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 [22]. Random trees are constructed considering a predefined number of randomly chosen features. In these experiments the Forest consists of ten trees each considering eleven features (this number *k* is determined based on the number of instances *N* in the data set using Equation 11).

$$k=\log_2 N+1 \qquad (11)$$

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

Design of experiments

The experiments were carried out using a data set describing patient antibody displays that held 1230 attributes and 343 instances divided into classes 'melanoma patients' (188 instances) and 'healthy donors' (155 instances); the data were provided by Latvian BioMedical Research & Study Center. To determine the baseline error, a set of experiments was conducted using the full data set and performing classification with all methods (C4.5, Random Forest, Naïve Bayes and SVM).

The second step involves attribute selection by all of the methods described using 10 fold cross-validation that divides the data set into 10 subsets and performs attribute selection on data withholding one subset each time. The feature subset selection methods use greedy forward stepwise selection using both filter and wrapper strategies (as described earlier) and stop when the evaluation metric starts to decrease. Attribute subsets selected by subset selection methods are used complete (including attributes used in more than one fold) to assure robustness and avoid overfitting to specific data subset, whereas ranker methods rank all of the attributes and only a subset can be used in the experiments to reduce dimensionality. According to Golub et al. [23], the attribute subset size differences of data sets that contain 50, 100 and 200 attributes have a minimal impact if data sets hold microarray data with several thousand attributes and a much smaller number of instances. Considering similarities of gene microarray data and antibody display data the number of selected ranked attributes is set to 50 best attributes.

The selected antibody subsets are then evaluated using classification algorithms C4.5, Random Forest, Naïve Bayes and SVM and 10 fold cross-validation and the results are compared to baseline results.

Results and discussion

Overall experimental results show that the most appropriate classification algorithms for this data set are SVM and Random Forest, which had the lowest error rate in the most attribute subsets. The error rate in data subsets that were created using ranking approach had less deviation and mostly were around 20% the only outfitter being SVM classifier when applied to data set with dimensions reduced by SVM approach (the classification error being a little over 5%), which is a logical result. The classifier precision in data sets where subset selection was used varied more corresponding to classifier sensitivity to inter-feature relationships in the selected subset.

The obtained results evaluated by classification error (percentage of the incorrectly classified instances) in 10 fold cross-validation are given in

Table 1. It shows classification errors of all classifiers in all data sets used in the experiments; the shaded cells show the best result for the data subset used in the experiments.

In almost all data subsets where wrapper technique was used to reduce dimensionality the best results were shown by the methods which were used in the wrapper indicating that choosing the right dimensionality reduction approach is very important because it can significantly improve the results but also decrease results if used improperly.

The performance of Classifier and Wrapper methods is similar because they both use classification algorithms when evaluating feature subsets. Wrapper-based feature subset selection showed the largest increase in classification accuracy for tree based classification methods (improvement in C4.5 being 18% and almost 5% in Random Forest. The best results for Naïve Bayes classification method were in the data subset selected by Classifier method. It can be explained by the fact that it did not use cross-validation leaving more data for the method to be trained and it is crucial for Naïve Bayes to show good results.

Table 1.

Classification errors for data subsets

	Attri butes	Error (%)			
		C4.5	RF	SVM	NB
Baseline	1230	32,9446	23,9067	18,0758	25,3644
CFS	123	31,7784	19,8251	17,7843	21,8659
Classifier J48	6	17,2012	41,1079	45,1895	45,1895
Classifier RF	2	46,9388	47,2303	45,1895	44,898
ClassifierSVM	13	37,3178	35,8601	18,9504	32,07
ClassifierNB	7	29,7376	36,7347	41,3994	16,6181
Consistency	16	34,6939	31,7784	31,4869	38,1924
Wrapper J48	6	14,8688	41,1079	45,1895	38,484
Wrapper RF	1	42,8571	18,0758	44,0233	18,0758
Wrapper SVM	12	39,3586	34,6939	20,6997	40,2332
Wrapper NB	5	24,1983	31,4869	45,1895	19,242
Chi	50	32,3615	22,1574	23,0321	33,5277
GR	50	30,0292	21,5743	25,3644	37,3178
IG	50	32,07	22,449	25,656	33,5277
OneR	50	31,1953	38,484	26,8222	35,8601
ReliefF	50	32,9446	26,5306	25,656	32,9446
SVM	50	36,4431	23,3236	**5,2478**	27,1137

The best results using CFS and Consistency feature subset selection methods have been shown by SVM because the algorithm benefits from correlation reduction and it showed the best overall and baseline results. Correlation reduction in the features (CFS method) benefited all tested classification results showing better results than baseline although the feature set was reduced to 10% of the initial set meaning that the information in the data was preserved.

The use of Chi-square statistic only slightly improved the performance of the decision tree based classification methods that had the best increase in accuracy among ranker selected subsets when GainRatio and Information Gain metrics were applied. This is also understandable because these metrics are used in tree construction.

The evaluation of single feature predictive capabilities did not show any notable results the only accuracy increase being for C4.5 method. The ReliefF method also did not show any significant results in this data decreasing the classification accuracies. Notably dimensionality reduction using SVM feature evaluation method only showed an increase in classification accuracy for SVM classifier.

The methods that are scalable perform well on full data sets but they also benefit from the right feature selection methods (the accuracy of C4.5 improved by 18%, Random Forest by almost 6% and SVM by almost 13%). Another important aspect in favor of feature selection also with scalable methods is the reduction in computational resources. On average, the computation time decreased by half.

Most frequently chosen features are shown in Fig. 1; the antibodies are coded by their ID used in the study.

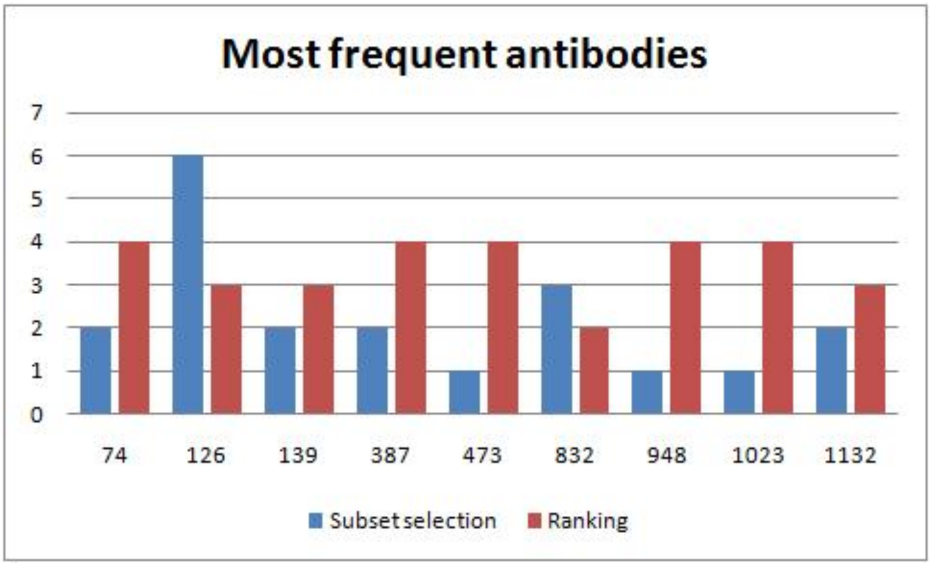

Fig. 1. **Frequency diagram of the most popular attributes**

The lighter columns show the occurrence frequency of the antibody in feature subsets chosen by subset selection methods; the darker columns show the number of occurrences in the top 50 antibodies of ranked lists.

Conclusions

Although accuracy fluctuations are greater in feature subset selection methods, the best results of all classification methods were shown in data subsets selected by these methods. Ranker methods show more stable results across all methods that would ease the selection of the right method, they do not show the best results.

The data subsets that were acquired using feature subset selection methods held less features than the selected threshold for rankers (50 best features) showing that the size of the feature subsets does not have to be large to build effective classifiers.

Overall experimental results show that data mining methods can be used to reduce antibody display data dimensionality for data analysis keeping the significant information intact and the accuracy does not suffer; on the contrary – the results even show some increase in accuracy and the computation resource consumption decreases.

References

1. Sundaresh, S. et al. From protein microarrays to diagnostic antigen discovery. *Bioinformatics* 23-13, 2007, p. i508-i518.
2. Guyon, I., Weston, J., Barnhill, S., Vapnik, V. Gene selection for cancer classification using support vector machines. Machine Learning. 46, 2002, p. 389-422.
3. Langley, P. Selection of relevant features in machine learning. *Proceedings of the AAAI Fall Symposium on Relevance*. New Orleans, Louisiana, USA, November 4-6, 1994. New Orleans: AAAI Press, 1994, p. 140-144.
4. Hall, M. A. Correlation-based Feature Subset Selection for Machine Learning. Dissertation at University of Waikato (Hamilton, New Zealand), 1998. 198 p.
5. Kohavi, R., John, G. H. Wrappers for feature subset selection. *Artificial Intelligence* 1-2, 1997, p. 273-324.
6. Tan, C.P., Lim, K.S., Lai, W.K.. Multi-Dimensional Features Reduction of Consistency Subset Evaluator on Unsupervised Expectation Maximization Classifier for Imaging Surveillance Application. International Journal of Image Processing, 2-1, 2008, p. 18-26.
7. Liu, H., Setiono, R. A probabilistic approach to feature selection - a filter solution. Proceedings of the 13th International Conference on Machine Learning (ICML'96), Bari, Italy, July 3-6, 1996. San Mateo: Morgan Kaufmann Pub., 1996, p. 319-327.
8. 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.
9. Quinlan J. R. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Pub., 1993. 302 p.

10. Holte, R. C. Very simple classification rules perform well on most commonly used datasets. Machine Learning 11-1, 1993, p. 63–90.
11. Kira, K., Rendell, L. A. A Practical Approach to Feature Selection. Ninth International Workshop on Machine Learning, Aberdeen, Scotland, UK, July 1-3, 1992. San Mateo: Morgan Kaufmann Pub., 1992, p. 249-256.
12. Kononenko, I. Estimating attributes: analysis and extensions of RELIEF. *Proceedings of the European conference on machine learning on Machine Learning* (ECML-94), Catania, Italy, April 6-8, 1994. Secaucus: Springer-Verlag New York, Inc., 1994, p. 171-182.
13. Robnik-Šikonja, M., Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning* 53, 1-2, 2003, p. 23-69.
14. Dudoit, S., Fridlyand, J., Speed, T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97-457, 2002, p. 77-87.
15. Lu, Y. Han, J. Cancer classification using gene expression data. Information Systems 28-4, 2003, p. 243-268.
16. Lee, J. W., Lee, J. B., Park, M., Song, S. H. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 48-4, 2005, p. 869-885.
17. 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.
18. 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.
19. 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.
20. Hastie, T., Tibshirani, R. Classification by Pairwise Coupling. Annals of Statistics 26-2, 1998, p. 451-471.
21. 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.
22. Breiman, L. Random Forests. Machine Learning 45-1, 2001, p. 5-32.
23. Golub, T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science (New York, N.Y.) 286(5439), 1999, p. 531–537.

Anotācija. *Ar molekulārās diagnostikas rīkiem tiek iegūti specifiski dati, kuriem ir augsta dimensionalitāte, kas saistīta ar lielo apskatīto faktoru skaitu vienā eksperimentā, un neliels ierakstu skaits, kas saistīts ar augstajām eksperimentu izmaksām. Rakstā apskatīta dimensionalitātes problēma melanomas pacientu antivielu analīžu datos, šim mērķim izmantojot datu ieguves atribūtu atlases metodes. Tiek apskatītas atribūtu ranžēšanas atlases un atribūtu apakškopu izvēles pieejas, kā arī dažādu metožu veiktspēja, novērtējot izvēlētās atribūtu apakškopas ar klasifikācijas algoritmu C4.5, Random Forest, SVM un Naivā Baijesa palīdzību. Klasifikatoriem jāspēj maksimāli labi atšķirt vēža pacientu dati no veselo donoru datiem. Atribūtu atlases metodes iekļauj uz korelāciju un konsekvenci balstītās metodes un wrapper tipa apakškopu atlases metodes, kā arī atribūtu novērtēšanu, izmantojot statistiskās īpašības, informatīvuma novērtējumu, likumu prognozētspēju un SVM atribūtu atlases novērtējumu, ranžēšanas vajadzībām*.