Technology for reduction feature space for classification immunosignature data

Koshechkin Alexander a, Andryushchenko Vladimir a, Romanovich Olga a, Stamate Daniel b,Zamyatin Alexander a

a Tomsk State University

b Department of Computing

Goldsmiths, University of London

**Abstract.** Random sequences of peptides in a microchip make it possible to generate specific immunosignatures that can diagnose various diseases. A large number of features does not allow for the quick and efficient analysis of such data. In this study, we propose technology to reduce feature space using various methods. The proposed technology makes it possible with minimal computational costs to ensure the accuracy and reliability of the classification of immunosignature data. The technology was tested on samples formed from a set of real data with the introduction of noise at various levels. The efficiency of the proposed technology on all test samples with various classifiers used for further data analysis is shown. .

**Keywords.** Immunosignature, informative peptide, feature selection, Early diagnostics diseases.

# Introduction

Recently, innovation in computer and information technology has generated huge amounts of information available for analysis. Often there is a problem of low representativeness of information as a result of missing values, noise, redundancy and unreliability of data. To build effective data analysis models, researchers need relevant, high-quality data. In this regard, the selection of informative features is one of the most important machine learning technologies aimed at reducing the dimension of the input space. By discarding useless or redundant attributes, it not only improves the performance of the model, but also facilitates its interpretation [1].

Currently, many types of cancer are potentially treatable if the early stage of the disease is diagnosed. For early detection of malignant tumors, simple, minimally invasive, but at the same time accurate diagnostic methods are needed [2]. The biomarker immunosignature technology is one of such methods [3].

The basis of this technology is a microchip representing a set of peptides with random amino acid sequences, which interacting with human serum; give a map of immune activity. Peptide microchips are diverse and contain from 10 000 to 350 000 peptides. Currently, the applicability of various classification methods based on supervised learning [4, 5, 6, 7, 8] is being actively investigated to analyze and interpret data obtained through biomarker immunosignature technology.

Due to specificity of peptide creation technology, not all of peptide are informative, which adversely affects machine learning. In addition, most classification models are not resistant to noise, and together with the high dimensionality of the data makes high-quality model training almost impossible. In this regard, there is the problem of optimal selection of informative features.

Thus, the purpose of this study is to evaluate existing methods and create an effective technology for selecting informative features for a rational time, as applied to high-dimensional immunosignature data, ensuring accuracy and reliability of classification at a high level. To achieve the goal it is necessary to solve the following tasks:

* Analyze the existing methods of selection of informative features;
* Estimate applicability of existing methods for selection informative features to immunosignature data;
* Based on obtained results , develop most appropriate technology for selection informative features applicable to immunosignature data.

# Materials and Methods

The materials obtained by analyzing biomarker immunosignatures represent a set of data for peptide fluorescence intensity values, where the names of the peptides are columns (features) and the class labels are rows (examples). The study used a normalized data set from a public biomedical data repository, obtained by digitizing data from immunosignatures [9, 10]. The data set has the following characteristics:

* number of copies - 1516;
* the number of signs - 10371;
* number of classes - 15;

The number of examples of each class is different, so the data is an unbalanced dataset. Table 1 shows the number of instances of each class.

**Table 1.** Dataset description

|  |  |
| --- | --- |
| **Class** | **Count** |
| Healthy control | 249 |
| Astrocytoma | 166 |
| Valley Fever | 142 |
| Breast cancer stages II, III | 141 |
| Pancreatic cancer | 136 |
| Multiple myeloma | 112 |
| Lung cancer | 107 |
| Mixed oligoastrocytoma | 97 |
| Ovarian cancer | 86 |
| Pancreatitis | 82 |
| Second BC | 61 |
| Oligodendroglioma | 48 |
| BC stage IV | 42 |
| GBM | 27 |
| Ewing sarcoma | 20 |

To date, there are a large number of different methods for selection of informative features. They can be divided into filter approaches, wrappers, and embedded methods.

**Filtering methods** - do not depend on the data analysis algorithm that will be applied to the selected features after selection, since they evaluate their informativeness, looking only at the internal properties of the data. Filtering methods are easily scaled for high-dimensional data and do not have a high computational complexity, but most methods are one-dimensional and treat each attribute separately, ignoring the dependencies of attributes, which can lead to deterioration of subsequent data analysis [11, 12].

**Wrapper Methods** - use a data analysis algorithm to determine how good a feature set is selected. The search procedure in the space of possible subsets of features is determined, various subsets of features are generated and evaluated. In the framework of the classification task, the selected subset of features will be selected for a specific classifier, which acts as an evaluation function. A common disadvantage of wrapper methods is that they have a higher risk of retraining than filtering methods and have high computational complexity [13].

**Embedded methods** - in this approach, the ability to search for an optimal subset of features is built into the classifier design. Like the wrapper methods, the built-in methods are specific to a particular learning algorithm, but in terms of computational complexity, they are something between the methods of filtering and convolution [14, 15, 16, 17].

In connection with the absence of dependencies between features due to the biological specificity peptides, it is advisable to use filtration methods as the most computationally efficient. There are a large number of different methods for selecting informative features that are applicable within the framework of the classification task [18, 19]. Therefore, it is advisable to choose methods based on a fundamentally different mathematical approaches. Consider the selected methods in more detail.

Gain ratio — it is supervised, univariate, non-symmetrical and entropy based measure introduced to eliminate the bias of information gain. Enhances Information Gain as it offers a normalized score of a feature’s contribution to an optimal information gain based classification decision [20, 21, 22].

Where — information gain for a feature and the class label , — энтропия признака .

Relief-F — is an improved version of the popular and classic filtering algorithm Relief. It is an instance based feature selection which works by sampling an instance randomly from the data and then finds its nearest neighbor from the same and opposite class. The nearest neighbors are compared to the sampled instance and it is used to update relevance scores of each attribute [23].

M-statistic is a one-dimensional method for estimating the separability of two classes based on their probability distributions [25].

where μ\_1 — first class mean, μ\_2 — second class mean, σ\_1 — first class standard deviation, σ\_1 — second class standard deviation.

Jeffries–Matusita distance —one-dimensional method for assessing the separability of two classes is an improved version of the Bhattacharya distance..

где — mean values for classes and respectively, — covariance matrices and respectively,  — vector transposition.

**Fisher score** — it is a supervised and univariate feature weighting method which picks features that assigns similar value to the samples from the same class and different value to samples from different classes [26].

# Principle of filtering methods

Using of filtering methods consists in assigning an assessment to all features for the subsequent selection of a subset of informative features based on it. Most of the known filtering methods are applicable only under the conditions of the binary classification problem, but with the use of one-versus-one or one-versus-all strategy, it is possible to apply multiclass classification under conditions of problem. In this study, for all filtering methods, we consider decomposition of a multiclass problem into a set of binary problems using one-versus-one strategy [28].

Results of evaluating informativeness of features can be represented in the form of a matrix, where the element of the matrix is evaluation of filtering method, columns are various binary comparisons, and rows are numbers of attributes in original data set. A fragment of matrix with estimates of informativeness of signs using M-statistic algorithm is presented in Table 2.

**Table 2.** Fragment of feature score matrix for binary comparisons

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Comparison 1** | **Comparison 2** | **Comparison 3** |
| Feature 1 | 0.1814 | 0.7218 | 1.2067 |
| Feature 2 | 0.6448 | 0.0697 | 0.1479 |
| Feature 3 | 0.3237 | 0.6923 | 1.0535 |
| Feature 4 | 0.4239 | 0.8101 | 0.2108 |

The initial values ​​of estimates for each attribute are replaced by value of position number of attribute in sorted list. A fragment of the rating matrix for M-statistic algorithm is added in Table 3.

**Table 3.** Fragment of matrix of ratings of each feature in binary comparisons rating

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Comparison 1** | **Comparison 2** | **Comparison 3** |
| Feature 1 | 6841 | 3002 | 522 |
| Feature 2 | 1009 | 9706 | 8936 |
| Feature 3 | 4386 | 3324 | 1042 |
| Feature 4 | 2973 | 2195 | 8324 |

The next step is to calculate relative ranking matrix by the formula:

where – relative ranking of feature , – feature rating value, matching binary comparison , – number of features in source data. Thus, all initial values are normalized to the interval [0, 1]. A fragment of the rank matrix for M-statistic algorithm is presented in Table 4.

**Table 3.** Fragment of the matrix of ranks of attributes

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Comparison 1** | **Comparison 2** | **Comparison 3** |
| Feature 1 | 0.3403 | 0.7105 | 0.9496 |
| Feature 2 | 0.9027 | 0.0641 | 0.1383 |
| Feature 3 | 0.5770 | 0.6794 | 0.8995 |
| Feature 4 | 0.7133 | 0.7883 | 0.1973 |

For formation of final data set, a strategy is used to select the N best features for each binary comparison.

# Technology

Filtering algorithms have various disadvantages that do not allow to find optimal set of informative features, as a result of which the efficiency of various classifiers varies considerably. There are many possible ways to improve feature filtering. In this study, the construction of an ensemble of filtering algorithms is considered.

When building an ensemble of filtering algorithms, we selected: Gain ratio, Relief-F, and M-statistic. The values ​​of rank matrix for each binary comparison are determined by following formula:

где , , – feature ranks on list, appropriate binary comparison , calculated using selected algorithms, filter weighting factor, – filter weighting factor , – filter weighting factor . Filter weights are calculated as:

где – Accuracy value when classifying a data set formed from N best features selected by -th filter, – sum of Accuracy values when classifying a dataset formed from N best attributes selected by each filter independently.

Further, based on values of ranks, N best features are selected for each binary comparison and a final data set is formed.

# Experimental studies

The purpose of the experimental studies is to assess the applicability and performance of selected filtering algorithms. To effectively apply the strategy of selecting the N best attributes, it is necessary to determine the value of the parameter N. To do this, we will conduct an experiment as follows.

* Selection of informative features, varying the number of N best features.
* Conduct a classification for each generated data set.
* Evaluation results.

In early studies for the classification problem, effectiveness of algorithms Random forest (RF) and Support vector machine (SVM) was revealed, therefore their use in this study as a classifier is quite reasonable [28].

When choosing a classification quality metric, we take into account the peculiarities of the processed data. The task of classifying data on biomarker immunosignatures is multiclass classification. In addition, the data set is an unbalanced sample. Therefore, in this experiment, it is advisable to use the Cohen's kappa classification metric [28].

To assess effectiveness of selection of informative features, it is necessary to form samples with various parameters that vary widely. It is advisable to implement this by distorting informativeness of features by entering noise into the data set. As noise, this experiment uses additive and multiplicative Gaussian white noise with constant mean and varying standard deviation. To analyze effectiveness of selection informative features, we will conduct an experiment using the following algorithm:

* Distort informative features in a Gaussian white noise dataset with different standard deviation values.
* Selection of informative features by various methods in each data set.
* Perform classification for each generated data set.
* Repeat steps 1-3 for both additive and multiplicative noise.
* Evaluation of the results.

For experiment, the R programming language and the implementation of data analysis methods from CRAN public repository were used. Table 5 lists the sources (packages) and settings for each method.

**Table 5.** Used packages and their parameters

|  |  |
| --- | --- |
| **Method** | **Package and Settings** |
| Healthy control | 249 |
| Gain ratio | “CORElearn” |
| Relief-F | “CORElearn” |
| M-statistic | “spatialEco” |
| Jeffries–Matusita distance | “spatialEco” |
| Fisher score | “PredPsych” |
| Support vector machine | “e1071”, kernel = "linear" |
| Random forest | “randomForestSRC”, ntree =100 |

# Results

In fig. 1, 2 shows the graphs of the quality of classification of the RF and SVM algorithm on the number of selected best features. It can be seen that the selection of more than 10 best features for the SVM and RF classifier does not improve the quality of classification for all the considered methods. In this regard, in the next experiment 10 best features will be used for each pairwise comparison.



**Figure 1.** The results of the classification of SVM with a different number of N best features.



**Figure 2.** RF grading results with various numbers of N best attributes.

In fig. 3.4 shows results of work SVM and RF classifiers on data previously distorted by multiplicative noise. The diagrams show quality of classifier, depending on selection algorithm informative features. It can be seen that the best of the studied algorithms for the selection of informative features for classification using SVM are M-statistic, Fisher score and Stacking, and using the RF - Gain ratio and Ensemble.



**Figure 3.** Results of selection informative features for classification of SVM at different levels of multiplicative noise.



**Figure 4.** Results of selection of informative features for classification of RF at different levels of multiplicative noise.

In fig. 5.6 shows results of work SVM and RF classifiers on the data previously distorted by additive noise. The diagrams show quality of classifier, depending on selection algorithm informative features. It can be seen that for classification using SVM, advantage of algorithms M-statistic, Fisher score and Ensemble for the selection of informative features is preserved. At the same time, for the RF algorithm there are practically no differences in the work of the selection methods of informative features.



**Figure 5.** Results of selection of informative features of methods for classification of SVM at different levels of additive noise.



**Figure 6.** Results of selection informative features of methods for the classification of RF at different levels of additive noise.

# Conclusion

High dimension of immunosignatures combined with a high content of non-informative features creates significant difficulties for effective use of existing methods of machine learning. To improve efficiency of classification on such data, a technology for reducing attribute space has been proposed, which makes it possible to ensure accuracy and reliability of the classification of immunosignature data with minimal computational costs.

The technology was tested on samples formed from a set of real data with introduction of noise at various levels. The efficiency of the proposed technology on all test samples with various classifiers used for further data analysis is shown.

References

1. Ashraf M., Chetty G., Tran D. (2013). Feature selection techniques on thyroid, hepatitis, and breast cancer datasets. International Journal on Data Mining and Intelligent Information Technology Applications (IJMIA), 3(1), 1-8, DOI: https://doi.org/10.4156/ijmia.vol3.issue1.1
2. Osipova T. V., Ryabykh T. P., Baryshnikov A. Yu. (2006). Diagnostic microchips: Application in oncology. Russian journal of biotherapy, 5(3), 72-81
3. O’Donnell B., Maurer A., Papandreou-Suppappola A., Stafford P. (2015). Time frequency analysis of peptide microarray data: Application to brain cancer immunosignatures. Cancer informatics, 14(2), 219-233, DOI: https://doi.org/10.4137/cin.s17285
4. Richer J., Johnston S.A., Stafford P. (2014). Epitope identification from fixed complexity random sequence peptide microarrays. Molecular & cellular proteomics, 14(1), 136-147, DOI: <https://doi.org/10.1074/mcp.m114.043513>
5. Kukreja M., Johnston S. A., Stafford P. (2013). Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases. Journal of proteomics & bioinformatics, 1(S6), 1-5, DOI: https://doi.org/10.4172/jpb.s6-001
6. Lake B. B., Rossmeisl J. H., Cecere J., Stafford P., Zimmerman K. L. (2018). Immunosignature differentiation of non-infectious meningoencephalomyelitis and intracranial neoplasia in dogs. Frontiers in veterinary science, 5(97), DOI: https://doi.org/10.3389/fvets.2018.00097
7. Rowe M., Melnick J., Gerwien R., Legutki J.B., Pfeilsticker J., Tarasow T.M., Sykes K.F. (2017). An ImmunoSignature test distinguishes trypanosoma cruzi, hepatitis B, hepatitis C and West Nile virus seropositivity among asymptomatic blood donors. PLOS neglected tropical diseases, 11(9), DOI: https://doi.org/10.1371/journal.pntd.0005882
8. Vard A., Firouzabadi F., Sehhati M., Mohebian M. (2018). An optimized framework for cancer prediction using immunosignature. Journal of medical signals and sensors, 8(3), 161-169
9. Stafford P., Halperin R., Legutki J.B., Magee D.M., Galgiani J., Johnston S.A. (2012). Physical characterization of the “immunosignaturing effect”. Molecular & cellular proteomics, 11(4), 1-14, DOI: https://doi.org/10.1074/mcp.m111.011593
10. GSE52580.URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52580.
11. Duch W. (2006). Filter Methods. In: Guyon I., Nikravesh M., Gunn S., Zadeh L.A. (eds) Feature Extraction. Studies in Fuzziness and Soft Computing, vol. 207, Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-540-35488-8\_4
12. Liu H., Motoda H. (2007). Computational methods of feature selection. New York: Chapman & Hall/CRC Data Mining and knowledge discovery series. DOI: <https://doi.org/10.1201/9781584888796>
13. Kohavi R., John G.H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273–324, DOI: https://doi.org/10.1016/s0004-3702(97)00043-x
14. Guyon I., Elisseeff A. (2006). An introduction to feature extraction. In: Guyon I., Nikravesh M., Gunn S., Zadeh L.A. (eds) Feature Extraction. Studies in Fuzziness and Soft Computing, vol 207. Springer, Berlin, Heidelberg, DOI: https://doi.org/10.1007/978-3-540-35488-8\_1
15. Liu H., Yu L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering, 17(4), 491-502, DOI: https://doi.org/10.1109/tkde.2005.66
16. Ma S., Huang J. (2008). Penalized feature selection and classification in bioinformatics. Briefings in bioinformatics, 9(5), 392–403, DOI: <https://doi.org/10.1093/bib/bbn027>
17. Saeys Y., Inza I., Larranaga P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507–2517, DOI: <https://doi.org/10.1093/bioinformatics/btm344>
18. Mani K., Kalpana P. (2016). A Review on Filter Based Feature Selection. International Journal of Innovative Research in Computer and Communication Engineering, 4(5), 9146-9156
19. Tang, J., Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review. Data Classification: Algorithms and Applications, CRC Press, 37-64, DOI: https://doi.org/10.1201/b17320
20. Koller D., Sahami M. (1996). Toward Optimal Feature Selection. International Conference on Machine Learning (ICML).
21. Yu L. Liu H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In International Conference on Machine Learning, 8(2), 856-863
22. Peng H., Long F., Ding C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226 1238, DOI: https://doi.org/10.1109/tpami.2005.159
23. Robnik-Sikonja M., Kononenko I. Theoretical and empirical analys is of relieff and rrelieff. Machine learning, 53(1-2), 23–69, 2003, DOI: https://doi.org/10.1023/A:1025667309714
24. Kaufman Y. J., Remer L. A. (1994). Detection of forests using mid-IR reflectance: an application for aerosol studies. IEEE Transactions on Geoscience and Remote Sensing, 32(3), 672-683, DOI: <https://doi.org/10.1109/36.297984>
25. Kumar P., Prasad R., Choudhary A., Mishra V. N., Gupta D. K., Srivastava P. K. (2016). A statistical significance of differences inclassification accuracy of crop types usingdifferent classification algorithms. Geocarto International, 1-19, DOI: https://doi.org/10.1080/10106049.2015.1132483
26. Gu Q., Li Z., Han J. (2012). Generalized fisher score for feature selection. arXiv: <https://arxiv.org/abs/1202.3725>
27. Izetta J., Verdes P. F., Granitto P. M. (2017). Improved multiclass feature selection via list combination. Expert Systems with Applications, 88, 205-216, DOI: <https://doi.org/10.1016/j.eswa.2017.06.043>
28. Andryushchenko V.S., Uglov A.S., Zamyatin A.V. (2018). Statistical classification of immunosignatures under significant reduction of the feature space dimensions for early diagnosis of diseases. Modern technologies in medicine, 10(3), 14 20, DOI: <http://dx.doi.org/10.17691/stm2018.10.3.2>
29. Arlot S., Celisse A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4(0), 40 79, DOI: https://doi.org/10.1214/09-ss054