Information Technology and Management Science

doi: 10.2478/v10313-012-0015-6

2012 /15

# The Impact of Cluster Stability on Class Decomposition in Antibody Display Data

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

***Abstract* – This article focuses on cluster stability evaluation to assess the characteristics of the dataset and the subclasses found in class decomposition. The evaluation is an iterative process, making small changes to the dataset in every step and reapplying the cluster analysis. These small changes (removing one object from the dataset is repeated for 20 iterations in this case) should not have any impact on clusters if they are stable (meaning that other objects that were not removed stay in the same clusters as in the full clustering).**

***Keywords* – class decomposition, clustering, cluster stability, data mining**

## I. Introduction

When working with biomedical microarray data, it is often useful to learn the inner structure of the data to improve the efficiency of classification models. This article deals with one of the methods that can be used to learn the inner structure of the data called the class decomposition. This method allows learning the structure of data inside a class by analyzing high density areas in the attribute space. These high density areas can be interpreted as disease subtypes in the analysis of this specific data. They can exist in the real diseases but can still be unidentified, and therefore the subclasses can provide also other significant information that is relevant to the medical experts that interpret the results of data analysis.

To use the class decomposition, the data belonging to each class should be clustered. Thus, this article focuses on the clustering quality by analyzing stability of separate clusters.

## II. Datasets

The datasets used in experiments are provided by the Latvian Biomedical Research and Study Centre (BMC) and obtained from the Broad Institute website [1]. The ones provided by BMC hold the antibody display data consisting of 1229 antibodies and the class label – a cancer patient (gastric cancer (GaCa), gastrointestinal inflammatory disease (GIS), prostate cancer (PrCa) or breast cancer (BrCa)) or healthy donor. The datasets obtained from the Broad Institute comprise cancer patient and healthy donor gene microarray data; cancers include: breast cancer (hereinafter called BC1 and BC2), carcinoma (carc) and prostate cancer (Pr).

## III. Cluster Analysis

Cluster analysis is performed in order to find subclasses (subtypes) of a disease. In this article, clustering is implemented by using the hierarchical agglomerative analysis and Ward’s distance [2], which measures the change in variance if two clusters (objects) are merged.

Hierarchical agglomerative clustering merges the closest objects/clusters (using the Ward’s distance metric) into clusters iteratively until all objects belong to one cluster. This forms the hierarchy of clusters that can be visualized as a tree (dendrogram), where the distance (Ward’s distance in this case) in each merge is shown by the distance between merges in the dendrogram. This can be used to find the most distant clusters (distance between merges in the dendrogram is longer than that of the others). Therefore, in this study the number of clusters is determined using dendrogram and the largest distance between two merges in it. The minimum number of clusters is set to three.

## IV. Cluster Testing

When objects are split into groups (clusters), this division is viewed as representing the characteristics of the whole set and should not show major changes if minor changes are made in the dataset. If there are no changes or adequately small changes are present, the clusters are believed to be stable; otherwise, these clusters are not stable and do not represent the features of the whole group. This article analyzes the stability of clusters induced in bioinformatics datasets for the reason of class decomposition using the hierarchical agglomerative clustering and Ward’s linkage. The minor changes mentioned above are considered to be subtraction of one object of the dataset – after removing one random object of the set, the division of other objects into clusters should remain the same, meaning that the other objects still belong to the same groups (clusters) as before the removal of the object. However, if there are small changes in the data, there will be changes in the clusters. These changes can be divided into two groups:

- changes in the distance at which the clusters are merged (Fig. 1a),
- changes of object allocation to clusters (Fig. 1b).

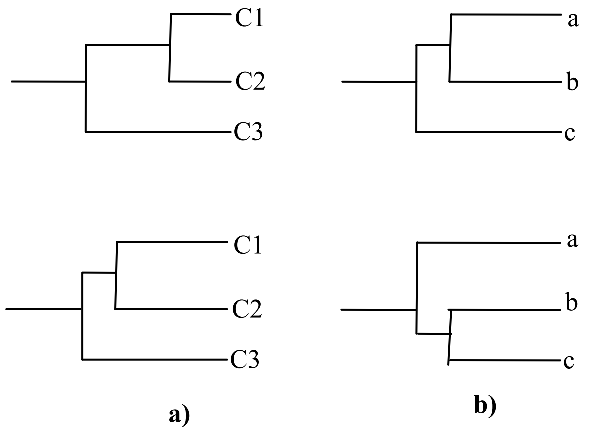


Fig. 1. Changes in cluster and object allocation; at the top – before changes in the record set, at the bottom – after changes

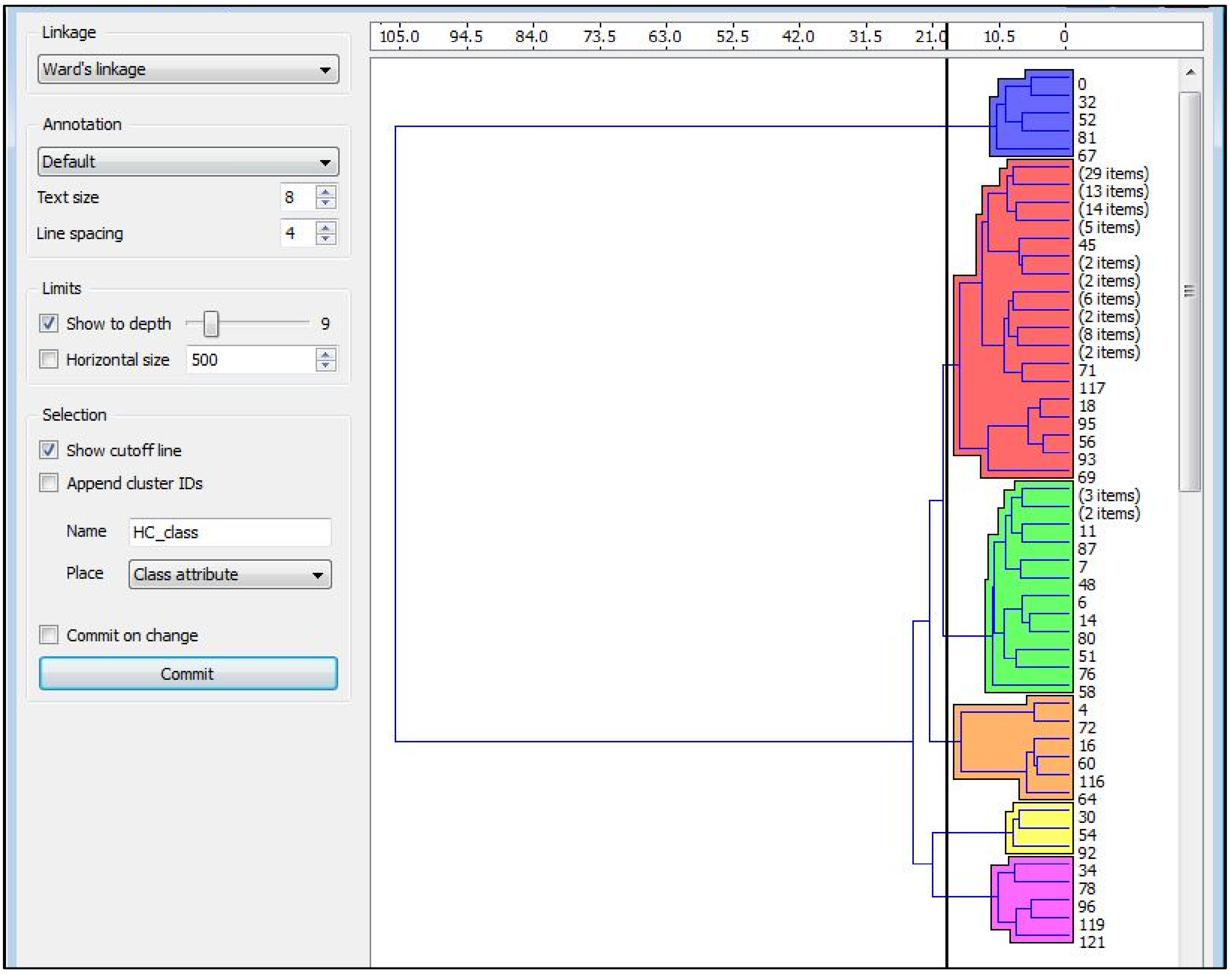
Although the sum of squares changes if one object is missing, the first type of changes is logical and not so important in this study. The second type of changes is crucial to determine cluster stability; in this case the objects are viewed as subclasses based on their membership to a cluster. Small changes in the dataset should not cause serious effect (object reallocation to different clusters) on the cluster structure.

## V. CLASSIFICATION

After assigning a subclass to all objects, the classification process is carried out as usual. The tested algorithms are C4.5 [3] (J48 implementation in Weka), Random Forest [4], Support Vector Machines [5] and Naïve Bayes [6] (also, they all are implemented in Weka library). Classification is performed using both – the initial classes (benchmark results for comparison) and the subclasses found in the cluster analysis. For all algorithms and datasets, 10-fold cross-validation is also applied to evaluate the results.

## VI. RESULTS AND DISCUSSION

After determining the number of clusters for each dataset using clustering in the full dataset, each set of clusters was tested for stability. The test included dismissing 20 records one by one and comparing the object memberships to clusters every time (see Fig. 2 – the gastrointestinal inflammatory disease dataset is clustered without record 84 and the clustering result using full dataset is given below; it is a good example of stable clustering where small changes do not affect the object membership allocation clusters).



Linkage

Ward's linkage

Annotation

Default

Text size 8

Line spacing 4

Limits

☑ Show to depth 9

□ Horizontal size 500

Selection

☑ Show cutoff line

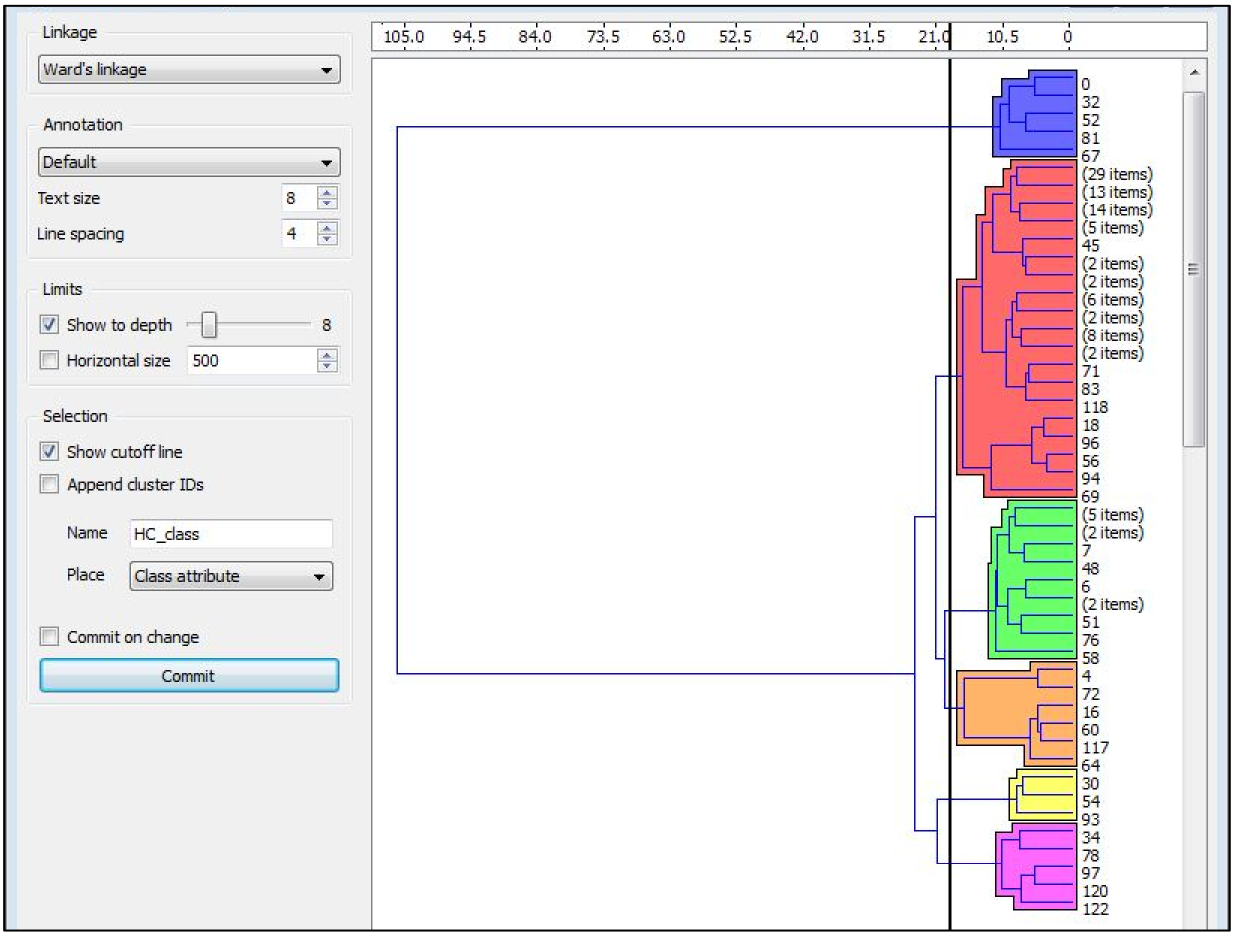
□ Append cluster IDs

Name HC_class

Place Class attribute

□ Commit on change

a) Clustering result of full GIS dataset



Linkage

Ward's linkage

Annotation

Default

Text size 8

Line spacing 4

Limits

☑ Show to depth 8

☐ Horizontal size 500

Selection

☑ Show cutoff line

☐ Append cluster IDs

Name HC_class

Place Class attribute

☐ Commit on change

b) Clustering of GIS dataset with one record removed from the full set

Fig. 2. GIS clustering with (a) and without (b) record 84

Then each number was divided by 20 showing the fraction of all records that were moved between clusters (see Table I). The average number then was chosen to describe the stability of each cluster group in each dataset. If the changes were greater than 5%, the clustering result was deemed as unstable. As can be seen from Table I, in 7 out of 8 cases the instability of the clusters is below this threshold.

TABLE I

FRACTIONS OF MISPLACED OBJECTS

| | | | Number of iterations | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Number of records | Average | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| BrCa | 13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| GaCa | 165 | 0.04 | 0.00 | 0.22 | 0.13 | 0.00 | 0.00 | 0.01 | 0.07 | 0.11 | 0.01 |
| GIS | 126 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| PrCa | 51 | 0.33 | 0.45 | 0.45 | 0.06 | 0.02 | 0.02 | 0.75 | 0.57 | 0.71 | 0.71 |
| BC1 | 17 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| BC2 | 41 | 0.02 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Carc | 18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 |
| Pr | 52 | 0.03 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 | 0.10 | 0.10 | 0.00 |

TABLE I (CONTINUED)
FRACTIONS OF MISPLACED OBJECTS

|  | Number of iterations |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| BrCa | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| GaCa | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.16 | 0.00 | 0.00 | 0.02 | 0.00 |
| GIS | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 |
| PrCa | 0.71 | 0.62 | 0.63 | 0.24 | 0.06 | 0.02 | 0.02 | 0.29 | 0.02 | 0.02 | 0.22 |
| BC1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 |
| BC2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.03 | 0.00 | 0.00 |
| Carc | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Pr | 0.08 | 0.08 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

The results show that the clustering results of prostate cancer dataset are clearly unstable – in some cases more than half of the records change the cluster they belonged to. Other datasets are below the acceptable margin of 5%. Nevertheless, all datasets are investigated further to see the impact of unstable clustering on the results of classification using subclasses.

The first classification is applied to the initial datasets without class decomposition. The results are shown in Table II. In the original data, the most accurate classification algorithm is SVM, which has the highest results in seven datasets out of eight (surprisingly it shows the worst accuracy in breast cancer gene expression (BC1) dataset. Both C 4.5 and RF show both the worst results and the best results depending on the dataset.

In general, the results point to the most and less complex datasets to classify – there are datasets where all classifiers show good results (like carcinoma (carc) and prostate cancer auto-antibody display (PrCa) datasets) and there are datasets with obviously more intricate structure, where all classifiers show mediocre classification accuracy (like gastro-intestinal inflammatory disease auto-antibody display (GIS) dataset, where all methods show results close to 50% accuracy).

TABLE II
BENCHMARK CLASSIFICATION ACCURACY

| Dataset | Classification algorithm |  |  |
|---|---|---|---|
|  | J48 | RF | SVM |
| BrCa | 57.69 | 84.62 | 88.46 |
| GaCa | 59.38 | 58.75 | 66.88 |
| GIS | 49.64 | 55.00 | 58.57 |
| PrCa | 83.50 | 85.00 | 90.00 |
| BC1 | 64.29 | 69.05 | 59.52 |
| BC2 | 64.58 | 67.71 | 79.17 |
| Carc | 91.67 | 91.67 | 97.22 |
| Pr | 85.29 | 79.41 | 91.18 |

Then the same classification algorithms were applied to the datasets with decomposed classes. The results are given in Table III. The shaded cells show results that are the same or higher than those without using class decomposition. The results for gastro-intestinal inflammatory disease auto-antibody display (GIS) dataset shows that class decomposition can give a great boost to classification accuracy in complex datasets because it has the increase in accuracy of 14% for C4.5 classification algorithm. Also data sets classified very well in their original form can gain from class decomposition as it can be seen in the case of carcinoma gene expression (carc) dataset – the Random Forest classification algorithm achieves perfect classification (100% accuracy).

As can be seen from Table III, the unstable clustering subclasses provide in the increase in accuracy but it is just one percent for SVM algorithm and no increase for other methods. In fact, the accuracy of C4.5 and of Random Forest algorithms decreased by 8.5% and 3.5%, respectively.

TABLE III

CLASSIFICATION ACCURACY USING CLASS DECOMPOSITION

| | Classification algorithm | | |
|---|---|---|---|
| Data set | J48 | RF | SVM |
| BrCa | 73.08 | 57.69 | 88.46 |
| GaCa | 61.88 | 55.94 | 67.19 |
| GIS | 63.57 | 55.87 | 63.93 |
| PrCa | 75.00 | 81.50 | 91.00 |
| BC1 | 69.05 | 66.67 | 64.29 |
| BC2 | 66.67 | 70.83 | 75.00 |
| Carc | 91.67 | 100 | 97.22 |
| Pr | 72.57 | 75.49 | 90.20 |

In other cases, J48 algorithm has a more significant gain in accuracy (up to 16% increase), when class decomposition is applied. Random Forest has the smallest gain in accuracy from class decomposition – the results improved only in three cases out of eight. SVM has also a significant gain in one of the datasets (GIS) and smaller gains (within one percent) in other datasets. Table IV provides a summary of maximum classification gain in a dataset and the average number of misplaced objects over 20 iterations (i.e. our measure of instability).

TABLE IV

GAIN IN CLASSIFICATION ACCURACY AND THE CORRESPONDING

CLUSTER INSTABILITY (AVERAGE FRACTION OF MISPLACED OBJECTS)

| Data set | Max gain in accuracy | Average number of misplaced objects |
|---|---|---|
| BrCa | 15.39 | 0.00 |
| GaCa | 2.5 | 0.04 |
| GIS | 13.93 | 0.01 |
| PrCa | 1.0 | 0.33 |
| BC1 | 4.77 | 0.01 |
| BC2 | 3.12 | 0.02 |
| Carc | 8.33 | 0.00 |
| Pr | - | 0.03 |

The overall trend can be seen in Table IV – many misplaced objects in the performed cluster stability test mean lower maximum gains in accuracy in the corresponding datasets. This means that more stable and ‘clean’ clusters lead to better classification accuracy using class decomposition and the found clusters. If PrCa dataset is removed (it has significantly higher average number of object misplacement), the correlation is -0.76 at p<0,05, which is statistically significant negative correlation – one of the variables grows, while the other decreases. This means that there is a statistically significant connection between the stability of a cluster group and the efficiency of classification using these clusters as subclusters (sub-diseases).

## VII. CONCLUSIONS

The experimental work has shown that cluster stability has a great impact on classifier accuracy when working with subclasses found by clustering the data. Therefore, more stable clusters lead to better data division into clusters, i.e., when the clusters are used in classification as subclasses, the ability of a classification algorithm to discriminate between classes and subclasses grows. The most significant cluster instability results in the smallest gain in accuracy among the datasets that had accuracy growth after class decomposition (in 7 cases out of 8 – prostate cancer dataset consisting of gene microarray readings did not show any increase in classification accuracy after applying class decomposition). Also the most significant gains in accuracy have been observed in the datasets with the clustering instability being equal to 0.01 and a few misplaced objects in 20 iterations.

## ACKNOWLEDGEMENTS

This research has been supported by the European Social Fund within the project “Support for the Implementation of Doctoral Studies at Riga Technical University”.

## REFERENCES

[1] Cancer Program Data Sets. [Online.] Available: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi [Accessed September 14, 2012].

[2] J. H. Ward, Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association Vol. 58, Issue 301, 1963, pp. 236–244.

[3] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers, 1993. 302 p.

[4] L. Breiman, Random Forests, Machine Learning Vol. 45, Issue 1, 2001, pp. 5-32.

[5] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 1995.188 p.

[6] G. H. John, P. Langley, Estimating Continuous Distributions in Bayesian Classifiers, Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995, pp. 338-345.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations, Vol. 11, Issue 1, 2009, pp. 10-18.

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

Contact information: Riga Technical University, 1 Kalku Street, Riga LV-1658. Phone: +371 67089530, e-mail: inese.polaka@rtu.lv.

**Arkady Borisov** holds a degree of Doctor of Technical Sciences in Control of Technical Systems and a habilitation degree in Computer Science.

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

He has supervised a number of national research grants and participated in the European research project ECLIPS.

Contact information: Riga Technical University, 1 Kalku Street, Riga LV-1658. Phone: +371 67089530, e-mail: arkadijs.borisovs@cs.rtu.lv.

**Inese Poļaka, Arkādijs Borisovs. Klasteru stabilitātes ietekme uz klašu dekompozīciju antivielu mikrorežģa datos**

Rakstā apskatīta klasterizācijas rezultātu novērtēšanas problēma. Par klasterizācijas kvalitātes mēru šeit tiek pieņemts klasterizācijas robustums jeb stabilitāte – noturība pret izmaiņām. Tas ir, nedaudz izmainot klasterizējamo datu kopu, klasterizācijas rezultātos zīmīgām izmaiņām nevajadzētu būt. Šajā gadījumā nenozīmīgas izmaiņas ir viena ieraksta likvidēšana pamatdatu kopā. Veicot klasterizāciju 20 reizes ar šādām izmaiņām, tiek iegūts vidējais izmaiņu apjoms (cik ieraksti procentuāli ir mainījuši piederību sākotnējiem klasteriem), kas arī ir stabilitātes novērtējums. Klasterizācija apskatīta un stabilitātes mērs pārbaudīts bioinformātikas datu kopās – gēnu ekspresijas vai antivielu mikrorežģu datos. Tiem ir īpatnība, ka dimensionalitāte ir ļoti augsta – tūkstošos atribūtu, bet ierakstu skaits ir salīdzinoši neliels – daži simti. Klasterizācija izmantota, veicot klašu dekompozīciju, tas ir, meklējot iespējamos slimības paveidus (kā, piemēram, leikēmijā pastāv leikoze, mieloleikoze un limfoleikoze; bet, iespējams, citām slimībām, kas apskatītas, apakštipi vēl nav atklāti) ar atšķirīgām biomedicīniskām izpausmēm. Lai noteiktu slimību paveidus, tiek veikta klasterizācija slimo pacientu datos, pieņemot klasterus par slimību paveidiem. Tad tiek veikta klasifikācija, nosakot veselos un slimos indivīdus, procesā ņemot vērā klases iekšējo blīvuma struktūru jeb slimību paveidus. Dati, kas izmantoti eksperimentos, iekļauj kuņģa vēža, gremošanas sistēmas iekaisuma slimību un melanomas antivielu datus, krūts vēža un prostatas vēža gēnu ekspresijas un autoantivielu datus, kā arī vispārīgus karcinomas gēnu ekspresijas datus. Rezultāti uzrāda, ka pastāv negatīva korelācija starp klasteru stabilitāti un klasifikācijas precizitātes pieaugumu klašu dekompozīcijas rezultātā, tātad, lai izmantotu klašu iekšējās blīvuma struktūras, klasterizācijas rezultātiem ir jābūt stabiliem.

**Инесе Поляка, Аркадий Борисов. Влияние стабильности кластеров на декомпозицию классов в данных микрочипов антител**

В статье рассмотрена проблема оценки результатов кластеризации. Мерой качества кластеризации здесь принята устойчивость или стабильность кластеризации – сопротивление изменениям. То есть, при небольшом изменении кластеризуемого набора данных в результатах кластеризации не должно быть значительных изменений. В этом случае несущественным изменением является устранение одной записи в основной выборке данных. Повторив процесс кластеризации 20 раз со следующими изменениями, получаются средние изменения (сколько записей процентуально изменили свою принадлежность исходному кластеру), которые также являются оценкой стабильности. Рассмотрен процесс кластеризации; мера стабильности проверена на выборках данных биоинформатики - микрочипов экспрессий генов или антител, особенность которых состоит в том, что их размерность очень высока - тысячи атрибутов, а количество записей относительно невелико - несколько сотен. Кластеризация используется, чтобы выполнить декомпозицию классов, то есть при поиске возможных вариантов заболевания (таких как лейкоз, миелоидный лейкоз и лимфолейкоз в лейкемии, и, возможно, в других рассмотренных заболеваниях подтипы до сих пор не обнаружены) с разными биомедицинскими проявлениями. Для определения подтипа заболевания кластеризация проводится в данных больных, предполагая, что кластеры - это типы заболеваний. Затем осуществляется классификация, различая здоровых и больных индивидов, с учетом внутренней структуры плотности класса, которая определяется подтипами заболеваний. Данные, использованные в экспериментах, включают данные антител рака желудка, кишечных воспалительных заболеваний и меланомы, данные экспрессии генов и антител рака молочной железы и рака простаты, и общие данные экспрессии генов карциномы. Результаты показывают, что существует отрицательная корреляция между стабильностью кластеров и увеличением точности классификации в результате декомпозиции классов, то есть, чтобы использовать внутреннюю структуру плотности классов, результаты кластеризации должны быть стабильны.