"Impact of cluster stability on class decomposition in antibody display data"

Inese Polaka (*Riga Technical University*), Arkady Borisov (*Riga Technical University*)

***Keywords* – analysis of data inner structure, clustering, cluster stability, data mining.**

I. Introduction

Bioinformatics data processing is a complicated process. This paper presents a data preprocessing step that analyzes the inner structures of a data set and that is called class decomposition. This step uses clustering (in this case applying hierarchical agglomerative clustering) to find high density areas in the classes present in data and re-labels them as subclasses. To implement class decomposition, the data is first analyzed using clustering then the clusters are evaluated and selected as subclasses. This article focuses on cluster stability evaluation to assess the characteristics of the data set and the found subclasses. The evaluation is an iterative process, making small changes to the data set in every step and reapplying cluster analysis. These small changes (removing one object from the data set 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).

II. Cluster evaluation

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 data set. In this case the clusters are believed to be stable. Otherwise these clusters do not represent the features of the whole group. This article analyzes the stability of clusters induced in bioinformatics data sets for the reason of class decomposition using hierarchical agglomerative clustering and Ward's linkage [5]. The minor changes mentioned above are considered to be subtraction of one object of the data set – after removing one random object of the set, the division of other objects into clusters should stay the same. Of course, 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).

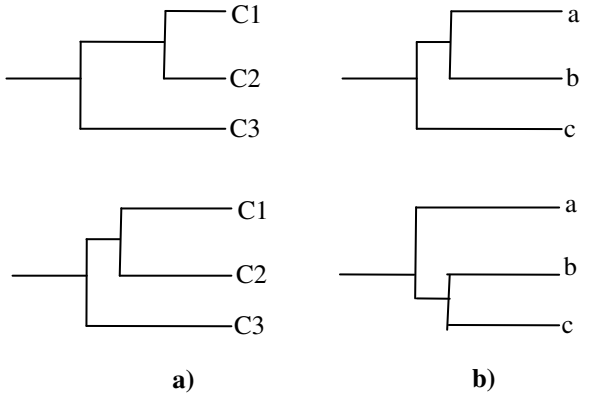

Figure 1. Changes in cluster and object allocation; top - before changes in record set, bottom - after changes

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

The experiments are carried out using eight bioinformatics data sets provided by Latvian Biomedical Research and Study Centre or available on the Internet.

III. Results and Discussion

Cluster stability was also tested using classification to find out if it had an impact on classification accuracy. The overall trend can be seen in Table 4 – many misplaced objects in the performed cluster stability test mean lower maximum gains in accuracy in the corresponding data sets. This means that the more stable and 'clean' clusters lead to better classification accuracy using class decomposition. If PrCa data set 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 other decreases, meaning that stability of clusters increases the maximum gain in classifier accuracy when clusters are used as subclasses.

TABLE IV

Gain in accuracy and cluster stability

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

IV. Conclusions

The experimental work in this study shows that cluster stability has a great impact on classifier accuracy. The more stable are the clusters, the higher is the quality of data division into clusters, which means that, when the clusters are used in classification as subclasses, the ability of a classification algorithm to discriminate between classes and subclasses grows.

V. References

[5] Ward, J. H., Jr., Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, Vol. 48, 1963, pp. 236–244.