Using Fuzzy clustering with bioinformatics data

Madara Gasparovica[1], Ludmila Aleksejeva[1], Vladislavs Nazaruks[2]

[1] Department of Modelling and Simulation, Institute of Information Technology, Riga Technical University, Kalku str.1, Rīga, LV-1658, Latvia
[2]Department of Applied Computer Science, Riga Technical University, Kalku str.1, Rīga, LV-1658, Latvia
madara.gasparovica@rtu.lv, ludmila.aleksejeva_1@rtu.lv, vladislavs.nazaruks@rtu.lv

Abstract: *The article describes a research about fuzzy clustering algorithms, their creation and classification with the goal to determine the possibilities to use them in bioinformatics data clustering to find the membership of each record to a class. The study uses sixteen data sets used in previous studies by the authors and other researchers. Experiments were carried out using fuzzy c-means clustering method. The first section of the article gives an overview of the historical development of fuzzy clustering algorithms, their classification as well as the hypothesis that fuzzy clustering algorithms can be used to construct membership functions. The second section gives the description of the applied algorithm and the sixteen data sets used in the experiments. The third section gives a summary of the performed experiments and their results. And finally conclusions are drawn about the use of the algorithms in the clustering of bioinformatics data. The fourth section gives the overall conclusions and describes the further research directions. It is proven that fuzzy clustering algorithms (including the most popular – fuzzy c-means) can be used in membership function construction. Therefore fuzzy c-means algorithm with slight modifications can be used to construct membership functions of separate record attributes.*

Keywords: Fuzzy clustering, fuzzy c-means, bioinformatics data.

Introduction

Fuzzy logic is used in various scopes for problem solving where regular (crisp) algorithms cannot show good results. Often algorithms are modified adding fuzzy logic to improve their performance. For example, Prism algorithm (Cendrowska, 1987) was modified into a fuzzy version of itself and called FuzzyPrim (Wang et al., 1999), Bexa algorithm (Theron et al., 1996) has a fuzzy versions FuzzyBexaI (van Zyl et al., 2004a) and FuzzyBexaII (van Zyl et al., 2004 b). Also similar trend is in clustering – algorithm k-means (MacQueen, 1967) was modified into its fuzzy version Fuzzy c-means (Bezdek, 1981; Klir et al., 1997).
Clustering differs from classification in that most often there is no information about the membership of a record to a particular class. This information can be obtained by performing analysis to find groups in the data. Cluster analysis is grouping of objects into clusters in a way that similarity between two objects in one cluster is larger than similarity of two objects belonging to different clusters. The first fuzzy clustering method was created in 1974 (Dunn, 1974) and soon thereafter new methods were created and improved, which is also happening until now (e.g., Li et al., 2006; Gadaras et al., 2009; Lu et al., 2012).
This study explores fuzzy c-means (Bezdek, 1981) clustering algorithm. The article studies clustering algorithm main working principles and presents the applied experiments with 16 real bioinformatics data sets. The main goal of this study is to determine if the use of fuzzy classification algorithms (fuzzy c-means in particular) could be perspective in construction of membership functions. It also gives conclusions about clustering results and proves that the use of clustering can be perspective. It is also confirmed by (Bilgic et al., 1995), who included it into their research about membership function construction methods and concluding that fuzzy clustering should be applied to the output data and then projected onto input data, then clusters should be generated and the variable data of input-output relationships should be chosen. Then the membership functions are formed for the chosen variables and then the cross-validation is applied to the input data to evaluate the model. The author also venture a guess that most fuzzy clustering methods use Euclidean norm and the direction for future research would be to examine the use of other distance metrics (Bilgic et al., 1995).
For an overview research of various fuzzy clustering methods see (Ali et al., 2008), which also offers a classification of fuzzy clustering methods (Fig.1).
The second section of the article gives a description of the implemented fuzzy clustering methods and the data sets used in the experiments.

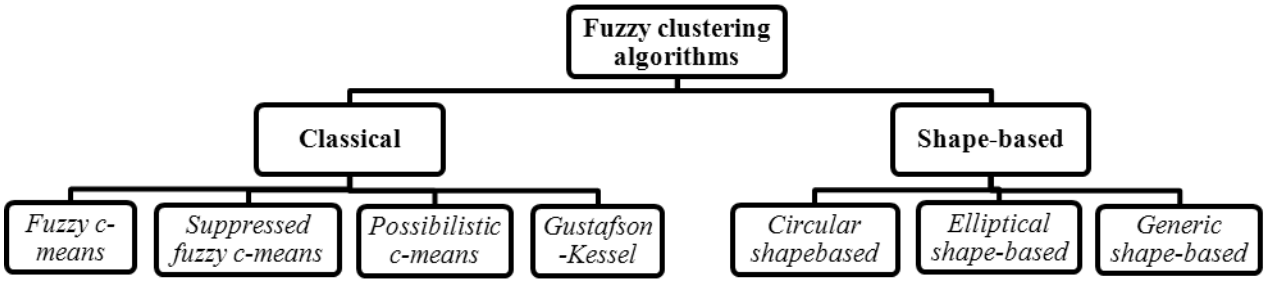

Fig. 1. **Fuzzy clustering algorithms; classification according to Ali.** (Ali et al., 2008)

The third section of this work gives a summary of the performed experiments and their results. It also gives conclusions about the use of algorithms in bioinformatics data clustering.
The fourth section gives overall conclusions and how the research can develop in the future studies.

Materials and methods

Fuzzy c-means clustering method

This study approaches practical analysis of clustering methods using bioinformatics data to evaluate the use of classical fuzzy clustering methods in the classification of bioinformatics data. Fuzzy c-means (fuzzy version of k-means) clustering was chosen as one of the top 10 data mining algorithms in the world (Wu et al., 2008). Therefore one can assume that Fuzzy c-means is one of the most used Top 10 algorithms.

Fuzzy c-means clustering algorithm tries to divide a finite element set $X = \{x_1, \dots, x_n\}$ into c cluster sets based on a particular criterion. If a finite data seti s given the algorithm returns a list of cluster canters $V = \{v_1; \dots; v_c\}$ and a conformity matrix $U = \{u_i(x) | i = 1 \dots c; x \in X\}$, where each element $u_i(x)$ determines level at which an element x belongs to cluster c_i (Klir et al., 1995; Bezdek, 1981). Usually for each data element its sum of membership levels to different clusters equals one, meaning that it is normalized:

$$\forall x \left(\sum_i u_i(x) = 1 \right)$$

The centroid v_k of each cluster k is calculated as follows:

$$v_k = \frac{\sum_x u_k(x)^m x}{\sum_x u_k(x)^m}$$

When centroids of all clusters have been calculated the coefficients $u_k(x)$ that show the membership levels of a point for each point x can be calculated according to the following formula:

$$u_k(x) = \frac{1}{\sum_j \left(\frac{d(v_k; x)}{d(v_j; x)} \right)^{2/m-1}}$$

where $d(a; b)$ is the distance between data points a and b. Often the Euclidean distance is chosen for this:

$$d(a; b) \equiv \sqrt{\sum_{i=1}^{\dim a} (a_i - b_i)^2}$$

m is the parameter of the real number that describes the influence on membership levels (Klir et al., 1995).
Algorithm fuzzy *c*-means itself is very similar to the rough *k*-means algorithm and can be described as follows:

1. Choose the number of clusters c.
2. Randomly assign membership coefficients $u_i(x)$ for each data point.
3. Repeat until the necessary accuracy is achieved:
    a. Calculate centroids v_i of all clusters.
    b. For each point calculate its membership coefficients $u_i(x)$.

To determine if the algorithm stops because the necessary accuracy ε is achieved, in the third step: calculate the absolute values $|\Delta u_k(x)|$ of changes of all coefficients $u_k(x)$ in the last iteration; find the maximum $\max_{k,x} |\Delta u_k(x)|$ of the absolute values and then compare to the accuracy ε that was chosen in the beginning of

the algorithm – if the maximum is not larger than ε, then the desired accuracy has been reached and the algorithm stops (Klir et al., 1995; Bezdek, 1981).

Used data sets

The study uses 16 popular data sets (Gasparovica et al,. 2012a; Gasparovica et al., 2012b) mentioned in literature (see Table 1): DLBCL data set (Shipp et al., 2002) - diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas (FL); GSE2191 - acute myeloid leukemia prognosis after treatment (Yagi et al., 2003); GSE349_350 - breast cancer treatment response (Chang et al., 2003); GSE3726 - breast & colon cancer gene expressions (Chowdary et al., 2006); GSE89 - bladder cancer gene expressions, (Dyrskjøt et al., 2003). GSE2535 - chronic myeloid leukemia treatment response (Crossman et al., 2005); GSE967- childhood tumors - Ewing's sarcoma (EWS), embryonal and alveolar rhabdomyosarcoma (eRMS and aRMS) gene expression (Baer et al., 2004); GSE2685 - diffuse and intestinal gastric cancer gene expressions (Hippo et al., 2002); GSE1577 - lymphoma & leukemia (T-ALL, T-LL and B-ALL) gene expressions (Raetz et al., 2006); GSE1987 - lung cancer (Dehan et al., 2007).

Table 1

Used data sets

Name	Diagnostic classes	Number of genes	Genes after FCBFS	Number of samples
DLBCL	Diffuse large B-cell lymphoma (DLBCL): 58 ex.	7070	74	77
	Follicular lymphoma (FL): 19 ex.			
GSE2191 (AML prognosis)	Remission (remission): 28 ex.	12625	54	54
	Relapse (relapse): 26 ex.			
GSE349_350 (Breast cancer)	Resistant to docetaxel treatment (resistant): 14 ex.	12625	2	24
	Sensitive to docetaxel treatment (sensitive): 10 ex.			
GSE3726 (breast & colon cancer)	Breast cancer (breast): 31 ex.	22283	37	52
	Colon cancer (colon): 21 ex.			
GSE89 (bladder cancer)	Tumor stage T2-T4 (T2+): 10 ex.	5724	48	40
	Tumor stage Ta (Ta): 19 ex.			
	Tumor stage T1 (T1): 11 ex.			
GSE2535 (CML treatment)	Non-responder to imatinib treatment (Non-Responder): 12 ex.	12625	33	28
	Responder to imatinib treatment (Responder): 16 ex.			
GSE967 (childhood tumors)	Ewing's sarcoma (EWS): 11 ex.	9945	2	23
	Rhabdomyosarcoma (RMS): 12 ex.			
GSE967 (childhood tumors)	Ewing's sarcoma (EWS): 11 ex.	9945	28	23
	Embryonal rhabdomyosarcoma (eRMS): 3 ex.			
	Alveolar rhabdomyosarcoma (aRMS): 9 ex.			
GSE2685 (gastric cancer)	Normal gastric tissue (Normal): 8 ex.	4522	2	30
	Diffuse gastric tumor (Diffuse): 5 ex.			
	Intestinal gastric tumor (Intestinal): 17 ex.			
GSE2685 (gastric cancer)	Normal gastric tissue (Normal): 8 ex.	4522	25	30
	Advanced gastric cancer tissue (Tumor): 22 ex.			
GSE1577 (lymphoma & leukemia)	T-cell lymphoblastic lymphoma (T-LL): 9 ex.	15434	2	19
	T-cell acute lymphoblastic leukemia (T-ALL): 10 ex.			
GSE1577 (lymphoma & leukemia)	T-cell lymphoblastic lymphoma (T-LL): 9 ex.	15434	53	29
	T-cell acute lymphoblastic leukemia (T-ALL): 10			
	B-cell acute lymphoblastic leukemia (B-ALL): 10			
GSE1987 (lung cancer)	Squamous cell carcinoma (Squamous): 17 ex.	10541	41	34
	Adenocarcinoma (Adenocarcinoma): 8 ex.			
	Normal lung tissue (Normal): 9 ex.			

Name	Diagnostic classes	Number of genes	Genes after FCBFS	Number of samples
GSE468 (medulloblastoma)	Metastatic medulloblastoma (Met): 10 ex.	1465	20	23
	Non-metastatic medulloblastoma (NonMet): 13 ex.			
GSE2443 (prostate cancer)	Androgen dependent tumor (dependent): 10 ex.	12627	2	20
	Androgen - independent tumor (independent): 10			
Brain tumor	Medulloblastoma (medulloblastoma): 10 ex.	7129	67	40
	Malignant glioma (glioma): 10 ex.			
	Rhabdoid tumor (RhabdoidTu): 10 ex.			
	Normal cerebellum (Normal): 4 ex.			
	Primitive neuroectodermal tumor (PNET): 6 ex.			

GSE468 - metastatic (Met) or non-metastatic (NonMet) medulloblastoma gene expressions (MacDonald et al., 2001); GSE2443 - prostate cancer gene expressions (Best et al., 2005); Braintumor (Pomeroy et al., 2002) - distinguishes between different embryonal tumors of the central nervous system on the basis of DNA expression signatures. The popular data sets are provided by University of Ljubljana, Faculty of Computer and Information Science Bioinformatics Laboratory on their home page (University of Ljubljana, 2012). All popular data sets characterize gene expression data. The column 'Genes after FCBFS' shows the number of genes that were used in the second series of experiments after applying attribute selection method Fast correlation based filter solution (Yu at al., 2003; Gasparovica 2012a) to the initial data set.

Fuzzy clustering results and discussion

Difference between clustering and real data

The experiments were carried out in two sessions – the first used data sets reduced with Fast-correlation baser filter solution attribute selection method (Gasparovica, et al. 2012a), the second used full data sets shown in the table. Fuzzy c-means experiments were carried out with the following parameters: number of clusters from 2 to 10, ε=0.0001 for reduced sets (with FCBFS), and ε=0.001 for the full data sets.

Table 2

Difference between the original and the clustering prognosis

Name	Number of classes	Original data set	Data set with FCBFS
DLBCL	2	0.69	0.57
GSE2191 (AML prognosis)	2	0.59	0.57
GSE349_350 (Breast cancer)	2	0.83	0.79
GSE3726 (breast & colon cancer)	2	0.65	0.65
GSE89 (bladder cancer)	3	0.60	0.83
GSE2535 (CML treatment)	2	0.54	0.79
GSE967 (childhood tumors – 2cl.)	2	0.70	0.70
GSE967 (childhood tumors – 3cl.)	3	0.52	0.48
GSE2685 (gastric cancer -3cl.)	3	*	0.67
GSE2685 (gastric cancer – 2cl.)	2	0.93	0.90
GSE1577 (lymphoma & leukemia – 2cl.)	2	0.84	0.89
GSE1577 (lymphoma & leukemia – 3cl.)	3	0.72	0.97
GSE1987 (lung cancer)	3	**	0.85
GSE468 (medulloblastoma)	2	0.42	0.74
GSE2443 (prostate cancer)	2	0.70	0.75
Brain tumor	5	**	0.73

(*) The healthy class is well separated but both cancer classes have the same membership; (**) All classes have the same membership.

Table 2 shows the difference between original class and prognosis for each data set, respectively it shows how much the defined membership of each record to a specific class corresponds to the real class. Analysis of clustering results and comparison of C-means algorithm applied to real data sets and to FCBFS data sets, it can be seen that better results are acquired in the data sets reduced using FCBFS. It can be explained by the fact that the original data set holds many attributes that require additional computations and consideration of so many criteria that they are hard to balance.

Clustering results

To evaluate the clustering results obtained by the fuzzy *c*-means algorithm, the internal evaluation metric S, which is described further, is used.

For degrees of belonging $u_k(x)$ of each data item x, the standard deviation σ is calculated (Weisstein):

$$\sigma(x) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(u_k(x) - \overline{u_k(x)}\right)^2}$$

where K is a number of clusters. As the values of $u_k(x)$ are normalized, the arithmetic mean $\overline{u_k(x)}$ equals to $\frac{1}{K}$; therefore, $\sigma(x) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(u_k(x) - \frac{1}{K}\right)^2}$. The sense of the metric σ is the following: for a data item x it equals to 0 if (and only if) the degrees of belonging of this data item to all clusters are equal ($u_1(x) = \cdots = u_K(x)$); and it takes the maximal value, if the data item belongs to one cluster with the degree of 1: $u_l(x) = 1$, $\forall k \neq l(u_k(x) = 0)$.

To compare the values of σ of the same data items independently of different numbers of clusters K, σ needs to be normalized in order to take values in the interval $[0; 1]$. The maximal value of σ is $\sigma_{max} = \sqrt{\frac{1}{K}\left(\left(1 - \frac{1}{K}\right)^2 + (K-1)\left(0 - \frac{1}{K}\right)^2\right)} = \frac{\sqrt{K-1}}{K}$. Therefore, the normalized value of σ is $\sigma_n = \frac{\sigma(x)}{\sigma_{max}}$.

For all data items x, the metric S is defined as the arithmetic average of $\sigma(x)$: $S = \overline{\sigma(x)}$. It takes the value of 0 (or near it), if the clustering is poor (the data item is not associated with any cluster with a significant certainty), and the value of 1 (or near it), if the clustering is authoritative (however, not obligatory true); the higher value of the metric is the better one.

The results of applying the metric S to the clustering results obtained by the fuzzy *c*-means algorithm are showed on the following two figures (two figures were used in order to improve the appearance of the graphs):

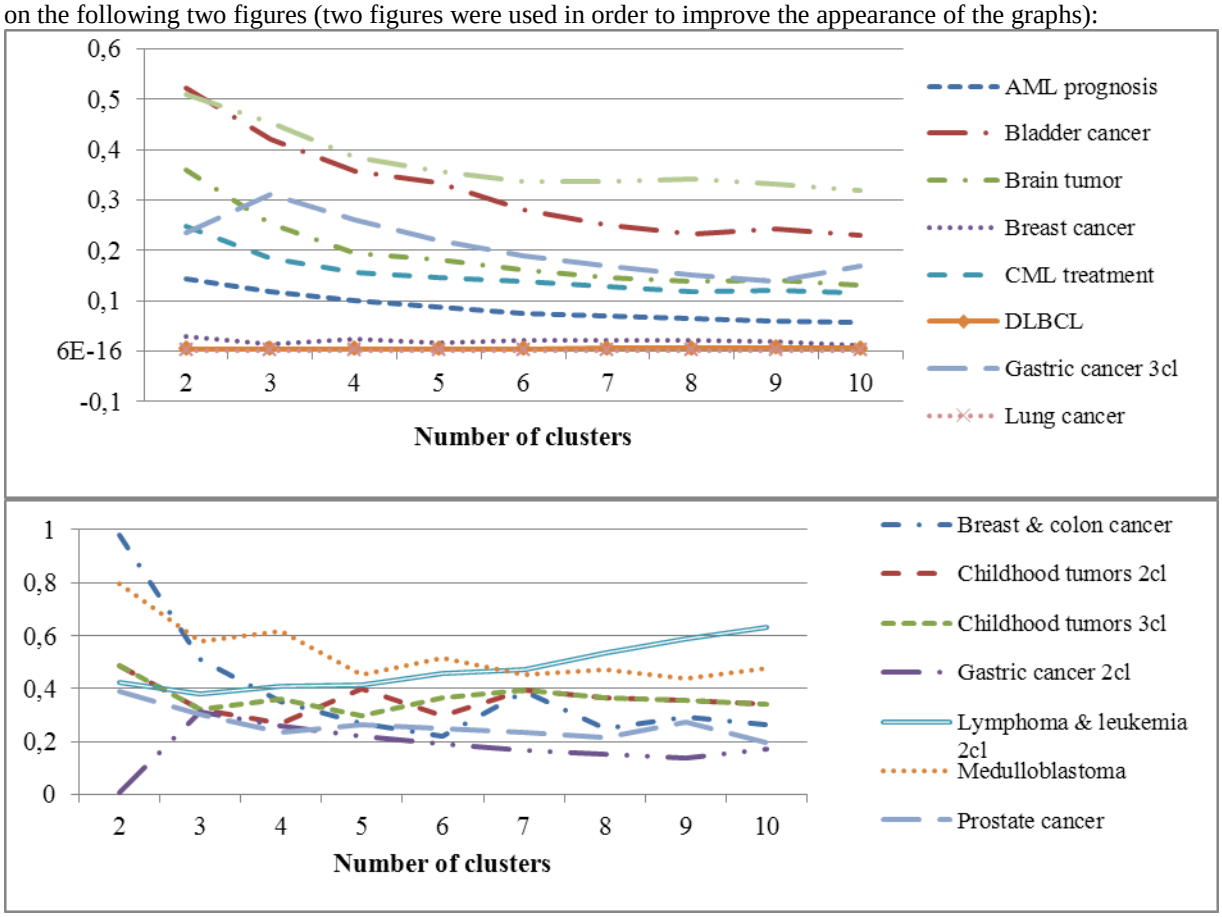

Fig. 2. **Fuzzy c-means results of applying the metric S.**

As one can see, the most common tendency of S depending on the number of clusters is to decrease. (The exception is the *Lymphoma leukemia 2cl* data set; however, it can be explained with a low (19) number of data items in the data set.) The maximums on the graphs show the numbers of clusters where the fuzzy *c*-means

algorithm can potentially give best results. For example, for the majority of data sets, the potentially best number of clusters is 2 (except for the *Gastric cancer 3cl* and *Gastric cancer 2cl* data sets).
Speaking about the absolute values of the metric, one can conclude that for the *Lung cancer*, *Breast cancer* and *CML treatment* data sets the fuzzy *c*-means clustering algorithm cannot give good results; however, for the *Breast & colon cancer*, *Medulloblastoma*, *Bladder cancer*, *Lymphoma leukemia 3cl* and *Childhood tumor 3kl* data sets this algorithm can potentially give good results. (The word 'potentially' here shows that we used the internal evaluation (not the external one), where known cluster labels were not compared with the clustering results.)

Iteration and Cluster number experiments

While conducting the experiments it was decided to also study the interactions of number of iterations and the number of clusters. As it can be seen in Fig.3 precise conclusions cannot be drawn but the data set behaviours have at least two trends. As it is shown in the upper graph in Fig. 3, the numbers of iterations display sharp fluctuations when the number of clusters is increased and then the number of iterations returns to the initial state, and beginning from three clusters the number of iterations for each number of clusters grows. This lets concluding that there is a number of clusters when the number of iterations increases very sharply. The lower graph in Fig. 3 shows a different trend – there are several numbers of clusters that have high numbers of iterations.

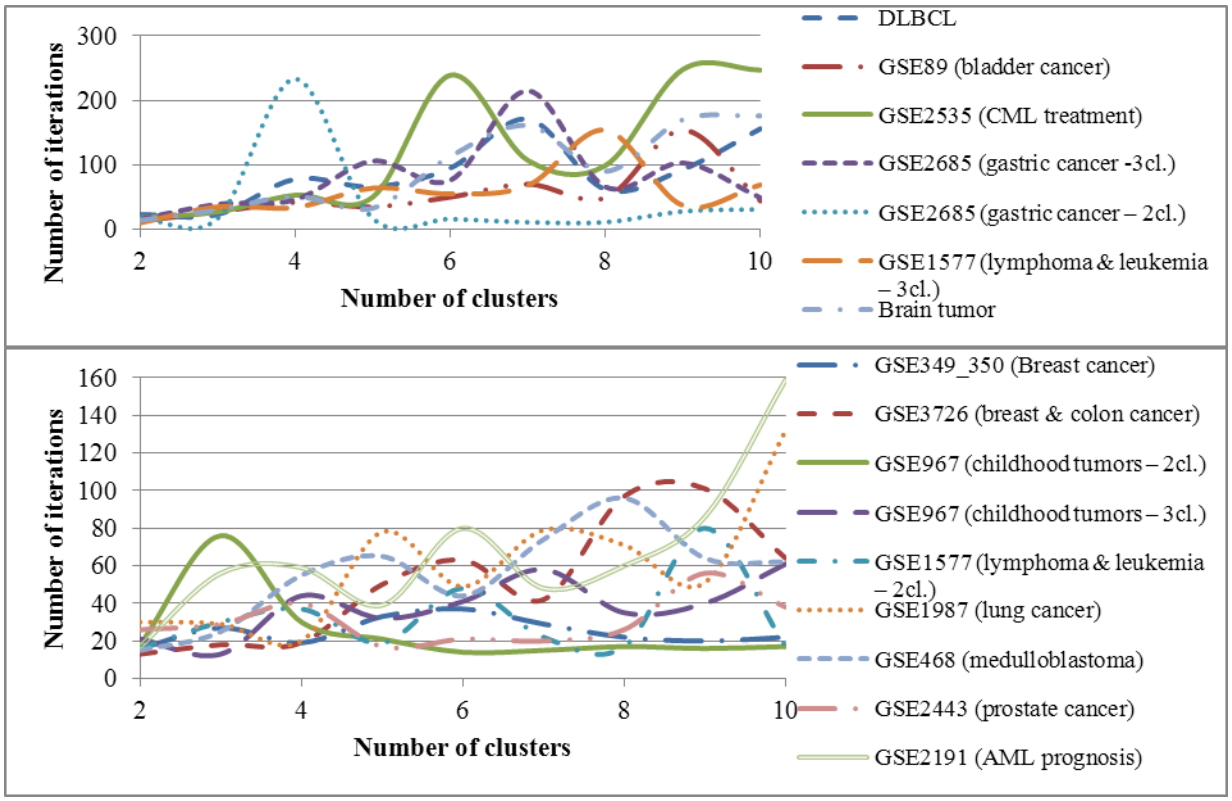

Fig. 3. **Iteration and cluster number experiments data sets with FCBFS.**

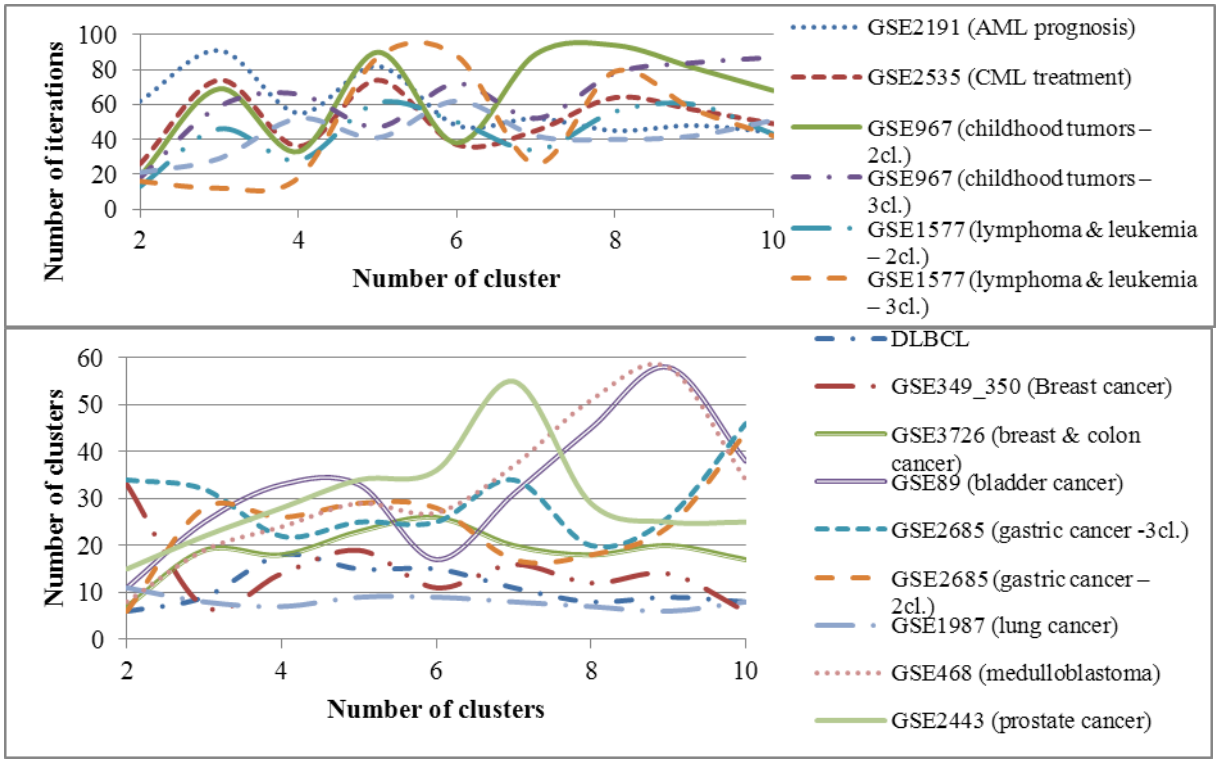

Fig. 4. **Iteration and cluster number experiments original data sets.**

If we look at graphs shown in Fig.4 that depict the trends in the original data sets with a slightly lower accuracy, it can be seen that the number of iterations to reach this accuracy is smaller. But one can see analogical trends in these data – some data sets have one number of clusters that requires much more iterations, but other show fluctuations in a small interval.

Conclusion

The use of fuzzy clustering algorithms while determining class membership is perspective because the acquired membership functions can be used in further research. A question that should be addressed in future is how to transfer information about the class weights in the process of constructing membership functions for the initial training data.

The research about the relations between the number of clusters and the number of iterations shows that there are certain relations but strong conclusions cannot be drawn. To do that there should be additional experimental research carried out.

Comparison of full and reduced data set clustering results shows that in future studies it is more perspective to use data sets reduced with FCBFS because the comparative parameters are better than in the full data sets.

The research about other membership function construction methods and algorithms should be continued in order to improve the shortcomings of fuzzy c- means algorithm.

Acknowledgements

Thanks to Dr.habil.sc.comp. Professor Arkady Borisov (Riga Technical University) for help and support.

References

Ali, A., Karmakar M., Gour C., Dooley, L.S., 2008. Review on Fuzzy Clustering Algorithms. *Journal of Advanced Computations*, Vol.2, No.3, pp. 169–181.

Baer, C., Nees, M., Breit, S., Selle, B. et al., 2004 Jul 10. Profiling and functional annotation of mRNA gene expression in pediatric rhabdomyosarcoma and Ewing's sarcoma. Int J Cancer;110(5), pp.687-694.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers Norwell, MA, USA, p. 256.

Bilgic, T., Türksen, I.B., 2005. Measurement Of Membership Functions: Theoretical And Empirical Work. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.6691, Accessed 05.01.2013.

Cendrowska J., 1987, PRISM: An Algorithm for Inducing Modular Rules. *International Journal of Human-computer Studies / International Journal of Man-machine Studies - IJMMS*, Vol. 27, no. 4, pp. 349-370.

Chang, JC., Wooten, EC., Tsimelzon, A., Hilsenbeck, SG. et al., 2003 Aug 2. Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. Lancet. 362(9381), pp.362 -369.

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K. et al., 2006 Feb, Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. J Mol Diagn;8(1), pp.31-39.

Crossman, L.C., Mori, M., Hsieh, Y.C., Lange, T. et al., 2005 Apr. In chronic myeloid leukemia white cells from cytogenetic responders and non-responders to imatinib have very similar gene expression signatures. Haematologica;90(4), pp.459-464.

Dehan, E., Ben-Dor, A., Liao, W., Lipson, D. et al., 2007 May. Chromosomal aberrations and gene expression profiles in non-small cell lung cancer. Lung Cancer;56(2), pp.175-184

Dunn, J.C., 1974. Well-Separated Clusters and Optimal Fuzzy Partitions, *Journal of Cybernetics and Systems*, vol. 4, pp.95-104.

Dyrskjøt L., Thykjaer T., Kruhøffer M., Jensen JL., et al., 2003 Jan, Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet;33(1), pp.90-96.

Gadaras, I., Mikhailov, L. 2009. An interpretable fuzzy rule-based classification methodology for medical diagnosis. *Artificial Intelligence in Medicine*, Vol. 47, pp. 25 – 41.

Gasparovica, M., Aleksejeva, L., 2012. Feature Selection for Bioinformatics Data Sets – Is It Recommended? *Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012)*, Latvia, Jelgava, pp. 325-335.

Gasparovica, M., Aleksejeva, L., Gersons, V. 2012. The Use of BEXA Family Algorithms in Bioinformatics Data Classification. *Scientific Journal of RTU*. 5. series., Information Technology and Management Science, vol. 15, pp 120-126.

Hippo, Y., Taniguchi, H., Tsutsumi, S., Machida, N. et al., 2002 Jan 1. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res;62(1), pp.233-240.

http://www.biolab.si/supp/bi-cancer/projections/index.htm, Accessed: 02.01.2012.

Klir, J.G., Yuan B., 1997. Fuzzy Sets and Fuzzy Logis. Theory and Applications. Prentice Hall of India. Private Limited, p. 592.

Klir, J.G., Yuan, B., 1995. Fuzzy Sets and Fuzzy Logis. Theory and Applications. Prentice Hall PTR, Upper Saddle River, NJ, p. 592.

Li, X.L., Tan, Y.C., Ng, S.K., 2006. Systematic Gene Function Prediction Using a Fuzzy Nearest-Cluster Method on Gene Expression Data. Proceedings of the First International Multi- Symposiums on Computer and Computational Sciences (IMSCCS|06), pp.171 - 178

Lu, W., Rankin, J.G., Bondra, A., Trader, C., Heeren, A., Harrington, P. de B. 2012. Ignitable liquid identification using gas chromatography/mass spectrometry data by projected difference resolution mapping and fuzzy rule-building expert system classification. Forensic science international, Vol.220, pp. 210-218.

MacDonald, T.J., Brown, K.M., LaFleur, B., Peterson, K. et al., 2001 Oct. Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. Nat Genet;29(2), pp.143-152.

MacQueen, J. 1967. Some methods for classification and analysis ofvmultivariate observations. *Proceedings of the Fifth Berkeley Symp. On Math. Stat. and Prob.*, 1, pp. 281-296.

Pomeroy, S.L, Tamayo, P., Gaasenbeek, M., et al., 2002 Jan 24. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature.;415(6870), pp.436-442.

Raetz, E.A., Perkins, S.L., Bhojwani, D., Smock, K. et al., 2006 Aug. Gene expression profiling reveals intrinsic differences between T-cell acute lymphoblastic leukemia and T-cell lymphoblastic lymphoma. Pediatr Blood Cancer;47(2), pp.130-40.

Shipp, M.A., et al., January 2002. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning. Nat Med., Vol.8, No.1, pp. 68-74.

Theron, H., Cloete, I., 1996. BEXA: A covering algorithm for learning propositional concept descriptions. *Machine Learning*, Vol. 24, No. 1, pp 5-40.

University of Ljubljana, Faculty of computer and information science Bioinformatics Laboratory home page:

van Zyl, J., Cloete, I., 2004. FuzzConRI - A Fuzzy Conjunctive Rule Inducer. *Proceedings of the ECML/PKDD-04 Workshop on Advances in Inductive Rule Learning*, pp.194-203.

van Zyl, J., Cloete, I., 2004. Simultaneous Concept Learning of Fuzzy Rules. *Proceedings of the Fifteenth European Conference on Machine Learning,* pp. 548-559.

Wang, C. H., Liu, J. F., Hong, T. P., Tseng, S.S., 1999. A fuzzy inductive learning strategy for modular rules, *Fuzzy sets and Systems*, 103, pp. 91–105.

Wu, X.,·Kumar, V.,·Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, S.H., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. *Knowl Inf Syst* ,14, pp.1–37.

Weisstein, E. W. Standard Deviation. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/StandardDeviation.html, Accessed: 05.01.2013.

Yagi, T., Morimoto, A., Eguchi, M., Hibi, S. et al., 2003 Sep 1. Identification of a gene expression signature associated with pediatric AML prognosis. Blood;102(5), pp.1849-1856.

Yu, L., and Liu, H., 2003. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution: Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), August 21-24, 2003, Washington DC. AAAI Press, Menlo Park, California.

.