VERSITA
Information Technology and Management Science
doi: 10.2478/itms-2013-0018
2013 / 16

The Application of Class Structure to Classification Tasks

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

***Abstract* – This article presents an approach in bioinformatics data analysis and exploration that improves classification accuracy by learning the inner structure of the data. The diseases studied in bioinformatics (diagnostic, prognostic etc. studies) often have the known or yet undiscovered subtypes that can be used while solving bioinformatics tasks providing more information and knowledge. This study deals with the problem above by studying inner class structures (probable disease subtypes) using a cluster analysis to find classification subclasses and applying it in classification tasks. The study also analyses possible cluster merges that would best describe classes. Evaluation is carried out using four classification methods that can be successfully used in bioinformatics: Naïve Bayes classifiers, C4.5, Random Forests and Support Vector Machines.**

***Keywords* – Bioinformatics, classification, class decomposition, data mining, data structure exploration**

I. Introduction

The advances in genomics and proteomics in the last decade have opened a way to the study of diseases from the systems biology point of view. Both gene expression microarrays and antibody display microarrays allow analysing thousands of genes/antibodies and their interactions in medical conditions covering several thousand genes/antibodies at a time. This also brings a new challenge in data analysis to effectively and accurately analyse the expressions. The methods used previously are not suitable for such high dimensionality; therefore, many bioinformatics researchers have turned to data mining methods that allow analysing data with so many features and comparably small sets of records. However, these methods, as well as data pre-processing, are also to be adapted for these new tasks.

This paper proposes class decomposition that helps deal with the data specifics by analysing their inner structures and allows discriminating between subclasses rather than existing classes in the diagnostic research. It is based on the natural properties of the diseases – many of them have several mechanisms of work, and many have subtypes that have not been distinguished yet. This way a classifier can build different models that suit all diseases rather than trying to adjust one model to several different diseases.

This article presents the decomposition approach using bioinformatics data (both the gene expression microarray data available from different sources on the Internet and the antibody display data provided by the Latvian Biomedical Research and Study Centre). Then it is validated using classification results provided by different classifiers that implement different approaches (the probability-based Naïve Bayes classifier, SVM that uses functional dependencies to discriminate between classes and has been proven to be one of the most successful classifiers working with this type of data; decision tree classification algorithm C4.5 and Random Forest that constructs an ensemble of decision tree classifiers).

II. Experimental Setup

All of the used data sets are either for genomic (BC1 – breast cancer, BC2 – inflammatory breast cancer, carcinoma, Pr – prostate cancer; all data sets are available at the website of Broad Institute Cancer Program [1]) or proteomic (GaCa – gastric cancer, GIS – gastrointestinal disease, Mel – melanoma, PrCa – prostate cancer; all antibody display data sets are provided by the Latvian Biomedical Research and Study Centre) task of diagnostics. Each proteomic data set holds information about 1229 antibodies expressed in sick and healthy donors. Each genomic set is composed of more than 10 000 gene expressions. Each dataset is pre-processed, imputing missing values (using average values of the attribute), and classes are decomposed.

Class decomposition is done by applying hierarchical agglomerative clustering and by analysing the cluster structures. First, each class is decomposed into 3 or less clusters based on the structure that has the largest distance among clusters; more clusters are not used because there are very few records comparing to the number of attributes. The distance between clusters is calculated using Ward's method [2].

The original and the decomposed data sets are evaluated using four different classification methods: Naïve Bayes classifier [3] (NB; NaiveBayes in *Weka* [4] software), C4.5 [5] (algorithm J48 in *Weka* software), Random Forests [6] (RF; RandomForest algorithm in *Weka* software) and SVM (SMO algorithm [7] in *Weka* software).

III. Results and Discussion

This section will provide the analysis of the results of single data sets gradually shifting attention to more abstract results that can help to draw conclusions.

Table I shows the results of classification using all three methods for breast cancer antibody data set. The table row named 'Bench' gives the benchmark results that are classification accuracy percentage without using class decomposition. Next row gives accuracy results for all clusters without any merges. The next three rows show the result when clusters are merged – the first and second in the row named '1 and 2' and so on. The results show that the classification accuracies of Naïve Bayes and C4.5 (J48) are significantly lower than those of Random Forest and SVM. The accuracy of Naïve Bayes classifier increases in three out of four cases

when class decomposition is applied but the accuracy of C4.5 drops significantly. While Random Forest and SVM show identical classification accuracies on the initial data sets, the performance on data with decomposed class structure differs significantly, reaching 88.33%, which is the best accuracy for this data set and is a 5% improvement over the next best result (Random Forests with decomposed data).

TABLE I
ACCURACY (%) RESULTS FOR BREAST CANCER DATA

	NB	C4.5	RF	SVM
Bench	75.00	73.33	81.67	81.67
All	76.67	66.67	78.33	85.00
1and2	76.67	65.00	83.33	85.00
2and3	75.00	68.33	83.33	85.00
1and3	78.33	65.00	76.67	88.33

Table II shows the results of classification using all three methods for gastric cancer data set. It is obvious that C4.5 has gained most from any decomposition structure. Naïve Bayes classifier, Random Forest and SVM methods have also shown improvements, but the class decomposition combinations with the best results do not match among methods, meaning that the best structure here can be dependent on classification algorithm specifics. The C4.5 algorithm has also shown the worst accuracy in benchmark results but improved a lot using the proper class decomposition, showing even better results than the Random Forest method. SVM has shown the best benchmark accuracy and also the best result that can be reached in the gastric cancer data set after class decomposition.

TABLE II
ACCURACY (%) RESULTS FOR GASTRIC CANCER DATA

	NB	C4.5	RF	SVM
Bench	60.63	55.63	59.06	64.38
All	60.63	59.06	56.25	65.63
1and2	60.94	59.69	59.38	63.75
2and3	60.31	62.19	56.88	65.31
1and3	61.56	61.25	60.00	62.81

Table III shows the classification accuracies for the gastro-intestinal disease data set. C4.5 algorithm again benefits in all but one case, whereas Random Forests and Naïve Bayes classifier do not benefit at all. SVM that has the highest benchmark accuracy also benefits in all but one case, but this one case is not for the same structure as it is for C4.5. Once again this shows that the same class inner structure descriptions can lead to completely different trends in classification accuracies in classification methods with different approaches. The best accuracy is again achieved by SVM method.

TABLE III
ACCURACY (%) RESULTS FOR GASTRO-INTESTINAL DISEASE DATA

	NB	C4.5	RF	SVM
Bench	56.07	54.29	58.57	61.07
All	53.93	55.71	57.14	62.86
1and2	55.36	55.00	56.07	62.86
2and3	50.36	51.07	56.07	63.57
1and3	52.86	55.71	51.07	60.36

Table IV shows the accuracies for the melanoma data set. Here the best classification accuracy for the initial data set is achieved by C4.5, but the performance of this method does not improve when class decomposition is applied (in fact, it decreases significantly). Also the accuracy of Naïve Bayes classifier shows no improvements when the information about class structure is introduced. Random Forest benefits in all but one combination, which is also one of the worst decomposed class structure combinations for other algorithms. This method also shows the overall best result for this data set – 85.42%.

TABLE IV
ACCURACY (%) RESULTS FOR MELANOMA DATA

	NB	C4.5	RF	SVM
Bench	74.64	83.09	82.22	79.59
All	72.59	75.22	85.42	80.76
1and2	74.05	77.55	84.84	78.13
2and3	74.05	76.97	84.55	81.34
1and3	73.76	75.22	80.17	78.43

Table V gives the results for prostate cancer antibody data. Here the two methods that benefit from class decomposition are Naïve Bayes classifier and SVM. Although C4.5 and Random Forests show comparatively good results for the initial data set, their performance does not improve when the class structure is introduced to the training data. Naïve Bayes classifiers perform worse than other methods on initial data, but the increase in accuracy from class decomposition is small (not more than 1%); therefore, the best result is again achieved using SVM method. Although the increase in accuracy for this method also does not exceed 1%, the best overall result is reached by SVM using class decomposition.

TABLE V
ACCURACY (%) RESULTS FOR PROSTATE CANCER DATA

	NB	C4.5	RF	SVM
Bench	83.5	85.5	87.5	88.5
All	83.5	74.5	86.0	89.5
1and2	82.5	68.5	84.0	89.0
2and3	84.5	70.5	83.5	89.0
1and3	84.0	73.5	80.0	87.0

Table VI shows results for breast cancer (BC1) gene expression data set. Here again Random Forests do not benefit from class decomposition, the results for the other algorithms are very different and do not lead to any more abstract conclusions. In fact, each algorithm shows the best accuracy in a data set with different subclass combinations (Naïve Bayes excels in the data set, where subclasses one and two are merged; C4.5 shows the best accuracy in the data set with three subclasses; SVM achieves its best result in the data set, where subclasses two and three are merged). The overall best result (69,05% accuracy) is shown by Random Forest in the initial data set and by C4.5 in the data set with original class decomposed into three subclasses.

TABLE VI
ACCURACY (%) RESULTS FOR BREAST CANCER (BC1) DATA

	NB	C4.5	RF	SVM
Bench	60.63	64.29	69.05	59.52
All	61.90	69.05	66.67	64.29
1and2	64.29	61.90	57.14	59.52
2and3	59.52	66.67	52.38	66.67
1and3	61.90	45.24	54.76	64.29

Table VII gives the classification accuracies for inflammatory breast cancer gene expression (BC2) data set. In this set, SVM and Naïve Bayes classifier are the algorithms that do not benefit from class decomposition, which contradicts results from previous data sets. J48 again benefits with all combinations and so do Random Forests; this also contradicts the conclusions drawn from the previous data sets. The overall best result is achieved by Naïve Bayes classifier without using the class structure information. Other methods could not reach this benchmark accuracy even after introducing class structure information.

TABLE VII
ACCURACY (%) RESULTS FOR INFLAMMATORY BREAST CANCER (BC2) DATA

	NB	C4.5	RF	SVM
Bench	86.46	64.58	67.71	79.17
All	70.83	79.17	75.00	76.04
1and2	80.21	65.63	71.88	77.08
2and3	76.04	73.96	68.75	76.04
1and3	79.17	81.25	68.75	69.79

Table VIII shows classification accuracies for carcinoma gene expression data set. The carcinoma data set is rather small (only 18 records in each of the two classes) and the best clustering (class splitting) result was at two clusters; therefore, the data set after class decomposition holds only two positive classes and the only combinations of subclass merging use all found subclasses or the initial data set (the case when two found subclasses are merged). Here none of the methods improves its performance after introducing the class inner structure information.

TABLE VIII
ACCURACY (%) RESULTS FOR CARCINOMA DATA

	NB	C4.5	RF	SVM
Bench	91.67	91.67	91.67	97.22
All	69.44	91.67	83.33	94.59

Table IX shows classification results for prostate cancer gene expression data set. Here the results improve after the class decomposition in one case for each method, and it is a different combination of subclasses for each method except Naïve Bayes classifier, which does not show increase in classification accuracy from the initial data set result. SVM clearly shows the best accuracies in both the initial data set and the datasets, where class inner structure information has been used.

TABLE IX
ACCURACY (%) RESULTS FOR PROSTATE CANCER GENE EXPRESSION DATA

	NB	J48	RF	SVM
Bench	62.75	79.41	79.41	91.18
All	56.86	68.63	71.57	94.12
1and2	55.88	76.47	79.41	91.18
2and3	62.75	83.33	75.49	91.18
1and3	61.76	67.65	81.37	91.18

Figure 1 shows the performance of Naïve Bayes classifier across all data sets with and without class decomposition (the result of the best cluster combination). In three out of five cases with antibody data, the accuracy of Naïve Bayes classifier benefits from class decomposition and the loss in accuracy in other cases is very small. In gene expression data sets, where the number of attributes reaches 10 000 and 15 000, the performance of the algorithm is significantly worse when class decomposition is applied. Gene expression data sets hold few records and have a high dimensionality that significantly increases the complexity of Naïve Bayes classification models that use all attributes to differentiate between classes. If additional information is added increasing the complexity of class description, the classification models become even more complex in order to explain more complex class structures.

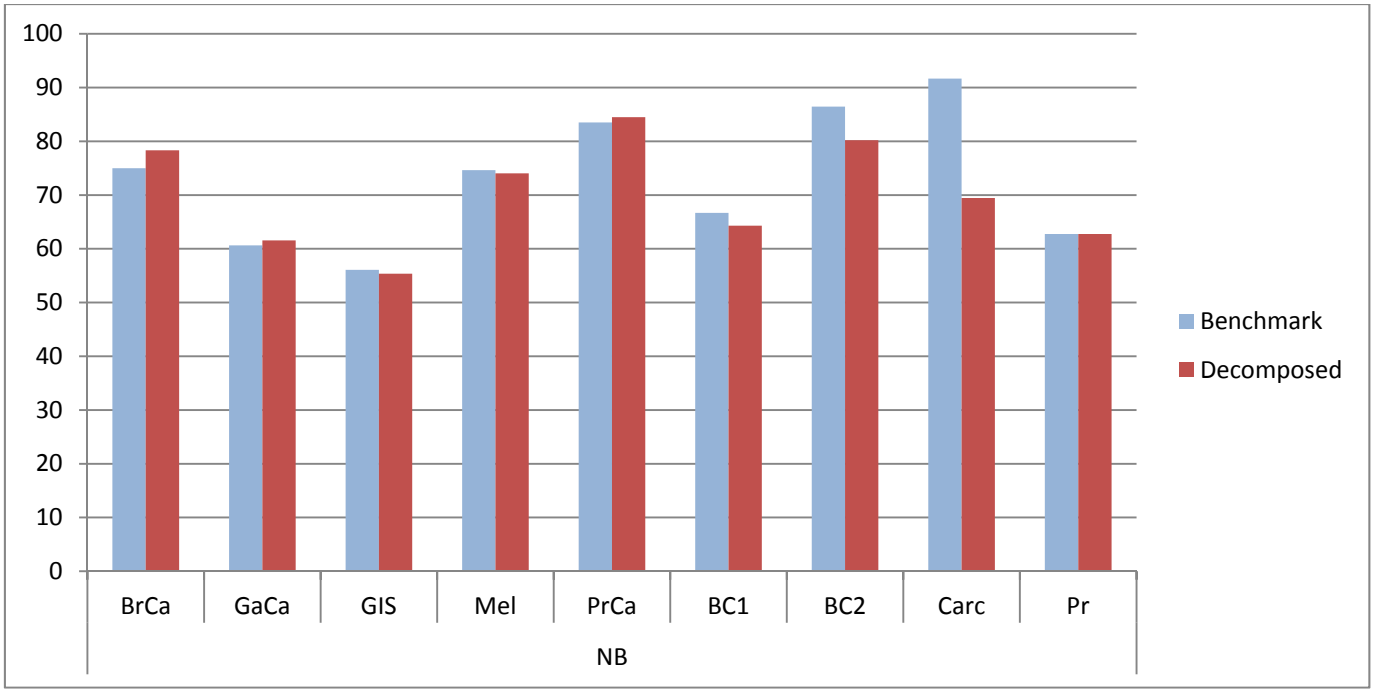

Fig. 1. Results of Naïve Bayes algorithm across all data sets

Figure 2 depicts the classification accuracy of C4.5 method (J48 implementation algorithm) across all data sets. C4.5 method builds simpler classification models and only uses the most informative attributes due to the built-in attribute selection mechanism. For complex data sets, such as biomedical sets (like the antibody display and gene expression microarray data used in this study), the simple models can be too small to describe all the necessary knowledge to discriminate between classes, whereas more complex, larger decision trees can be overfitted to the training data. Therefore, using additional information for class description can either help the classification trees describe the classes or make them more prone to overfitting because, while searching for more complex descriptions, models can incorporate unnecessary information and be overfitted. That is also visible in the results – in most cases the accuracy either drops or rises significantly. Due to the built-in attribute selection, the trend is not influenced by the dimensionality.

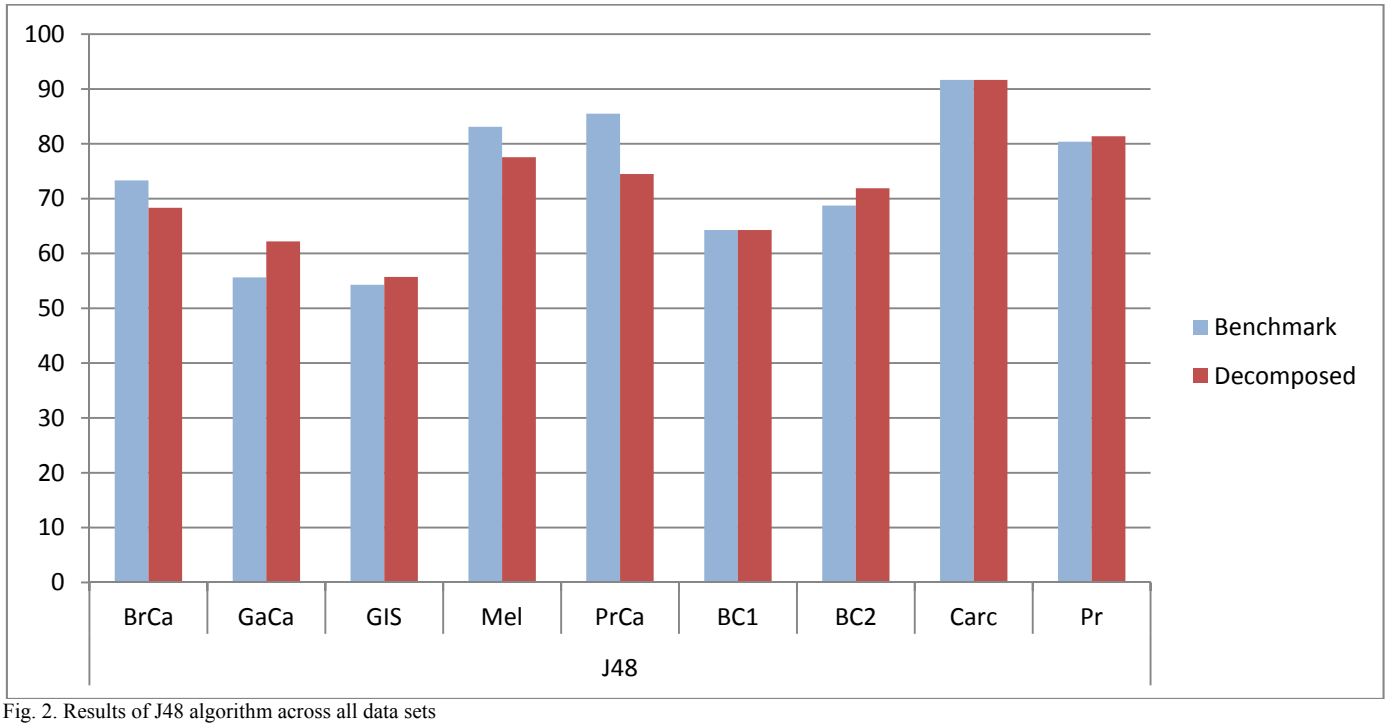

Fig. 2. Results of J48 algorithm across all data sets

Figure 3 sums up the performance of Random Forests (percentage accuracy). It can be seen that in most of the cases the algorithm benefits from class decomposition, but the amplitude of the changes is rather narrow. The only exception is in the data sets BC1 (breast cancer gene expression data set), where the accuracy rises by almost 10%. It is the smallest data set holding only 42 records, while its dimensionality reaches 16 382 attributes. This is a good example of how additional information about the class structure can improve the discriminating power of a method even in such complex data sets with a small number of records and high dimensionality.

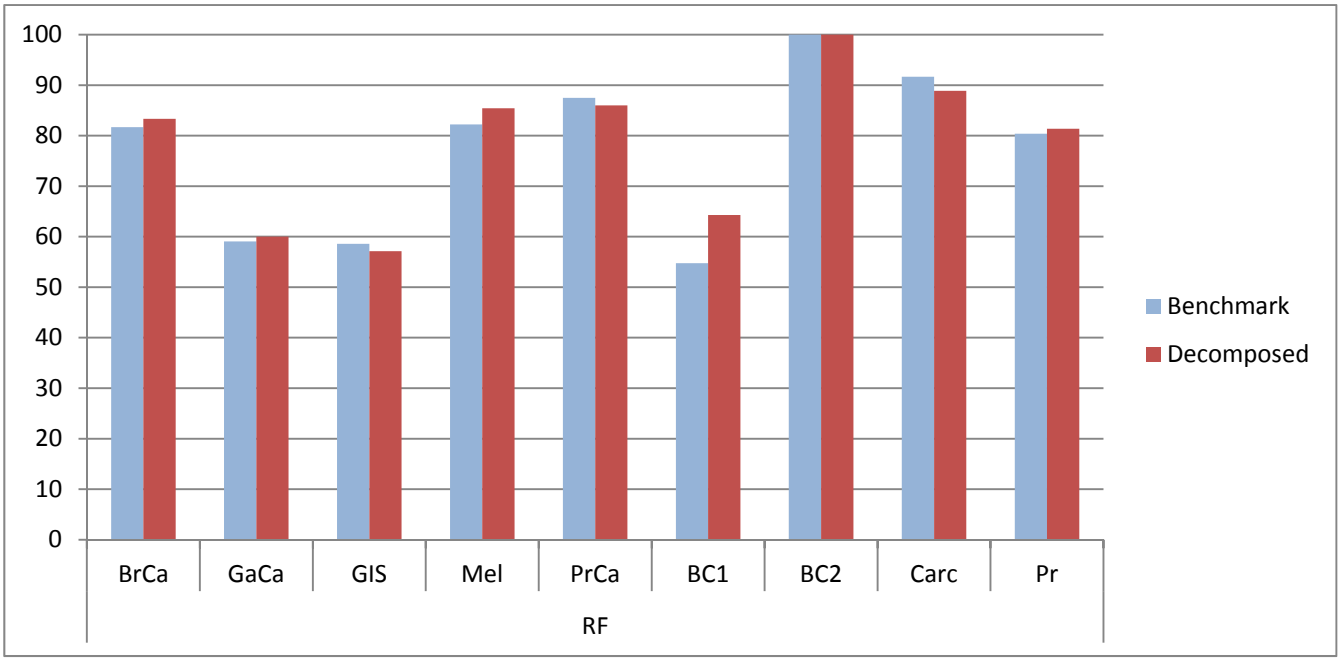

Fig. 3. Results of Random Forests algorithm across all data sets

Figure 4 demonstrates the results of SVM classifier (percentage accuracy) that performed better than the other methods in the most of the data sets without using the class decomposition. The graph also shows that the SVM method (SMO algorithm, that works with more than two classes) in almost all cases performed better when class decomposition was applied. The other two cases show that the classification accuracy did not change. This method can handle very complex data and build appropriate models, which can explain why its accuracy does not drop when additional information about the internal class structure is added. It can utilize this information very well, which allows improving the performance significantly.

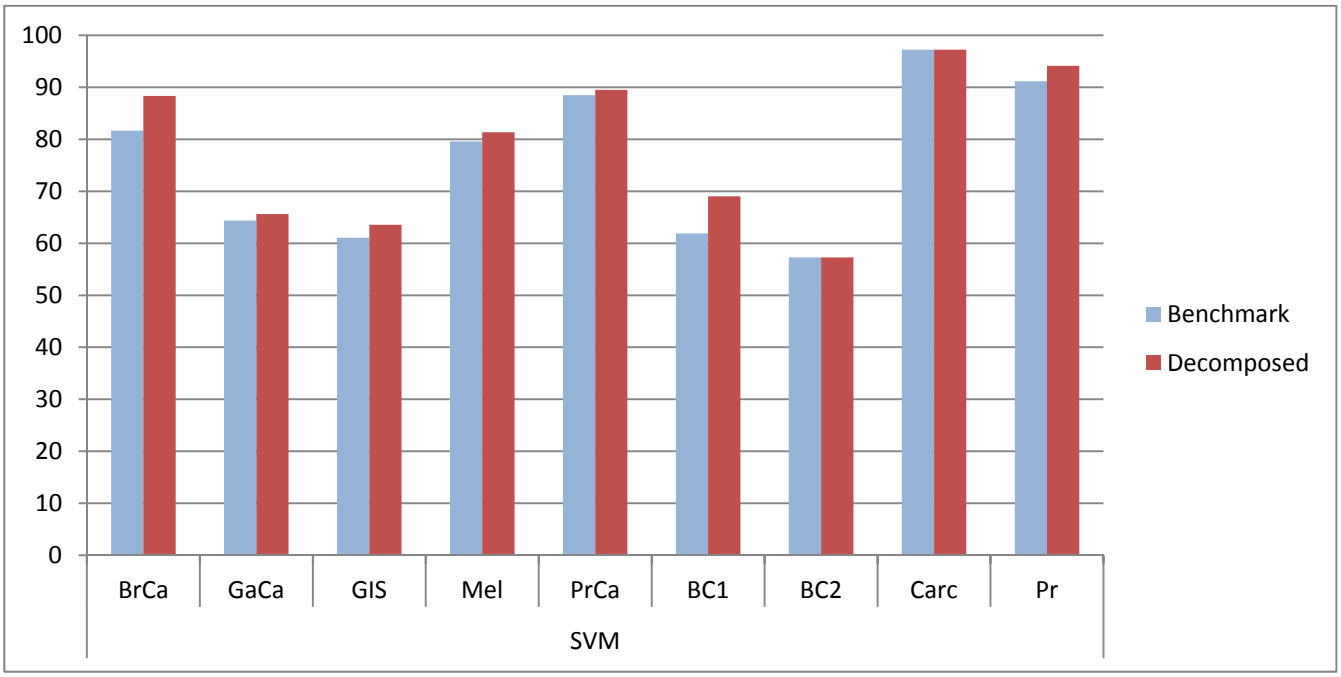

Fig. 4. Results of SVM (SMO algorithm) across all data sets

If we sum up the best results across all data sets, both benchmark best and best after decomposition, we get the graph in Fig. 5. It shows that class decomposition gives higher accuracy in all data sets but BC2, where the best benchmark result and the best result after class decomposition are equal and it is a perfect classification. The most significant increases are in the data sets with antibody data that hold 1229 attributes. Data sets with more than 10 000 attributes (BC1, BC2, Carc and Pr) have more similar results with and without class decomposition, except for prostate cancer gene expression data set where the accuracy after class decomposition is significantly higher than the accuracy in the initial data set.

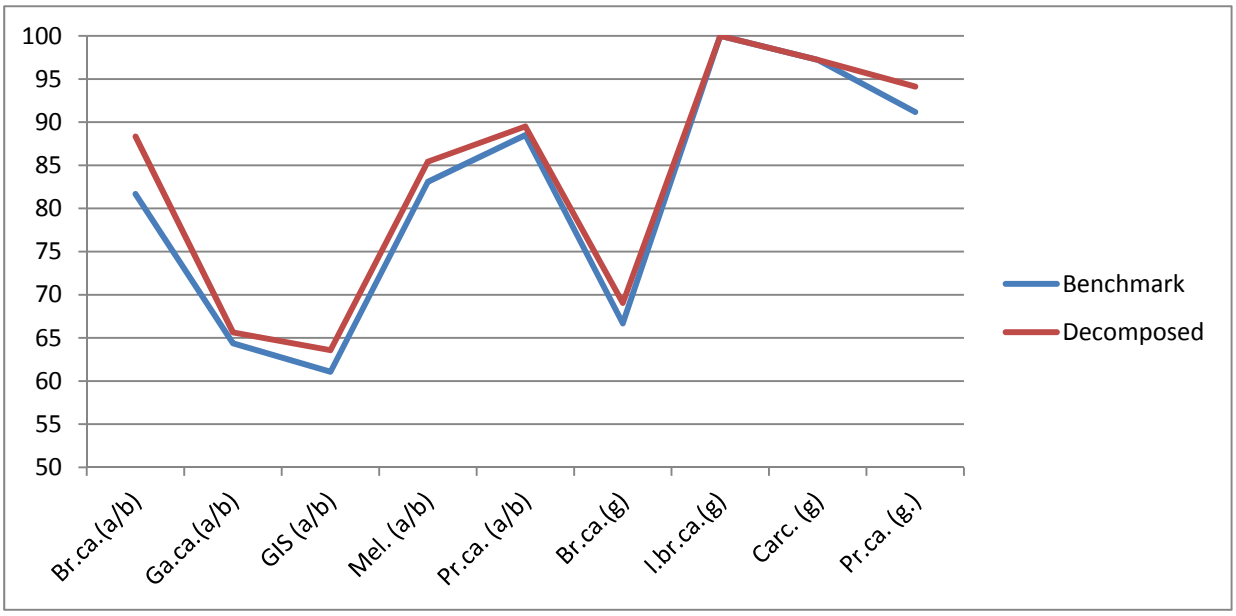

Fig. 5. The best results using decomposition compared to the best benchmark results

IV. Conclusion

The article shows that the use of proper class decomposition can increase the accuracy of almost any classification algorithm. The method that showed the best results the most was SVM (the accuracy of the *Weka* implementation of SMO algorithm increased in all but two data sets and remained the same in the other two data sets when compared to benchmark results (without class decomposition)), whereas Random Forests improved its accuracy in only five out of nine cases. C4.5 implementation J48 in *Weka* improved its accuracy in four cases out of nine and suffered some significant decreases in three cases due to overfitting while searching for more complex classification models that would incorporate the additional information about the class structure.

The overall results show that when the best benchmark (without using class decomposition or any other additional information about the class structure) results (out of all classification algorithms) are compared to the best result where class decomposition was applied, the best results were achieved using class decomposition.

The proposed approach gives a better overall description of the classes but it still leaves room for future research to describe the classes even better and represent the information in a way that would not make classifiers like decision tree-based methods prone to overfitting. Other clustering algorithms and distance metrics can be studied to give even more precise information about class structures using the approach presented in this article.

Acknowledgements

The 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 13, 2013]

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

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

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

[5] Quinlan, R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA., 1993.

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

[7] 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, 1998, pp. 185-208.

Inese Polaka is the fourth-year postgraduate student at Riga Technical University. She finished her Master studies at Riga Technical University majoring in Information Technology in 2010 obtaining the Mg.sc.ing. degree. 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.
E-mail: inese.polaka@gmail.com

Arkady Borisov received his Doctoral Degree in Technical Cybernetics from Riga Polytechnic Institute in 1970 and Dr.habil.sc.comp. degree in Technical Cybernetics from Taganrog State Radio Engineering University in 1986.

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

He is a member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany).

E-mail: arkadijs.borisovs@cs.rtu.lv

Inese Poļaka, Arkādijs Borisovs. Klašu struktūras izmantošana klasifikācijas uzdevumos
Pētījumā tiek risināta bioinformātikas problēma – datu klasifikācijā tiek izmantotas datu ieguves metodes, lai noteiktu diagnostikai svarīgu informāciju, kas ir šajos datos. Tā kā tipiskās klasifikācijas metodes reti ir pietiekami precīzas, tiek veikts papildus datu priekšapstrādes solis, kurā ar klasteru analīzes palīdzību tiek izzināta un aprakstīta klašu iekšējā struktūra, atrodot viegli nošķiramus blīvuma apgabalus, kas tālākajā darbā tiek uzskatīti par apakšklasēm. Šī pieeja balstās uz medicīnisko hipotēzi, kas daudzkārt apstiprinājusies citām slimībām, ka vienu un to pašu slimību var izraisīt atšķirīgi gēni (atšķirīgi slimības fenotipi) vai antigēni, pret kuriem darbojas humorālā imūnsistēma. Apakšklašu atrašanai tiek izmantota hierarhiskā aglomeratīvā klasterizācija ar vidējo attālumu un Varda attālumu. Datu kopas, kurās veikta klašu dekompozīcija, tika tālāk analizētas, izmantojot klasifikācijas metodes. Klašu iekšējās struktūras izmantošanas (klašu dekompozīcijas) novērtēšana tika veikta, par metriku izmantojot klasifikācijas precizitāti. Ja klašu iekšējās struktūras izmantošana palīdz atklāt papildus zināšanas, kas saistītas ar slimību, klasifikācijas precizitātei pēc klašu dekompozīcijas būtu jāpieaug, tāpēc klasifikācija tiek veikta datu kopās pirms un pēc klašu dekompozīcijas, izmantojot bioinformātikā populāras klasifikācijas metodes – Naivo Baijesa klasifikatoru, atbalsta vektoru mašīnas, lēmumu koku un to ansambļu klasifikatorus. Rezultāti uzrāda klasifikācijas precizitātes paaugstināšanos, izmantojot klašu dekompozīciju, bet lielākoties tā uzlabo to metožu darbību, kuras spēj veidot pietiekami sarežģītus klasifikatorus, lai aprakstītu ne vien klases, bet arī apakšklases.

Инесе Поляка, Аркадий Борисов. Применение структуры классов в задачах классификации
В исследовании решается задача в области биоинформатики – в классификации данных для определения существенной информации используются методы интеллектуального анализа данных. Поскольку типичные методы классификации редко бывают достаточно точны, выполняется дополнительный шаг предварительной обработки, на котором исследуется и описывается внутренняя структура классов, используя кластерный анализ. На этом этапе находятся легко отделяемые области плотности, которые в дальнейшей работе будут использованы как подклассы. Этот подход основан на медицинской гипотезе, которая неоднократно была доказана для других заболеваний, - одно и то же заболевание может быть вызвано разными генами (различные фенотипы болезни) или антигенами, против которых действует гуморальная иммунная система. Для нахождения подклассов была использована иерархическая агломеративная кластеризация, среднее расстояние и расстояние Уорда. Наборы данных, в которых проводилась декомпозиция классов, были далее проанализированы с помощью методов классификации. Оценка использования внутренней структуры классов (декомпозиции классов) проводилась с помощью точности классификации как метрики. Если использование внутренней структуры класса помогает обнаружить дополнительные знания о болезни, то точность классификации после декомпозиции классов должна быть улучшена, поэтому классификация проводится в наборах данных до и после декомпозиции классов с помощью методов, которые популярны в биоинформатике – Naive Bayes classifier, Support Vector Machines, C4.5 и Random Forests. Результаты показывают улучшение точности классификации после декомпозиции классов, но в большинстве случаев повышается точность методов, которые могут строить сложные классификаторы, способные описать не только классы, но и подклассы.