First, we considered a global prediction problem where each sample corresponds to a triplet (chemical, time, replicate) and is represented by a vector of gene expression. The resulting dataset contains 800 samples (20 chemicals x 5 doses x 4 time points x 2 replicates) in 10 dimensions. There are three classes: neural, carcinogenic (genetic) and carcinogenic (non-genetic). The best cross-validation accuracy was 0.60, and was achieved with Gaussian SVM parameters C = 1000 and G = 0.001.

In figure 1, we used PARCA to visualize these samples. For visualization purpose, we only considered the highest dose.

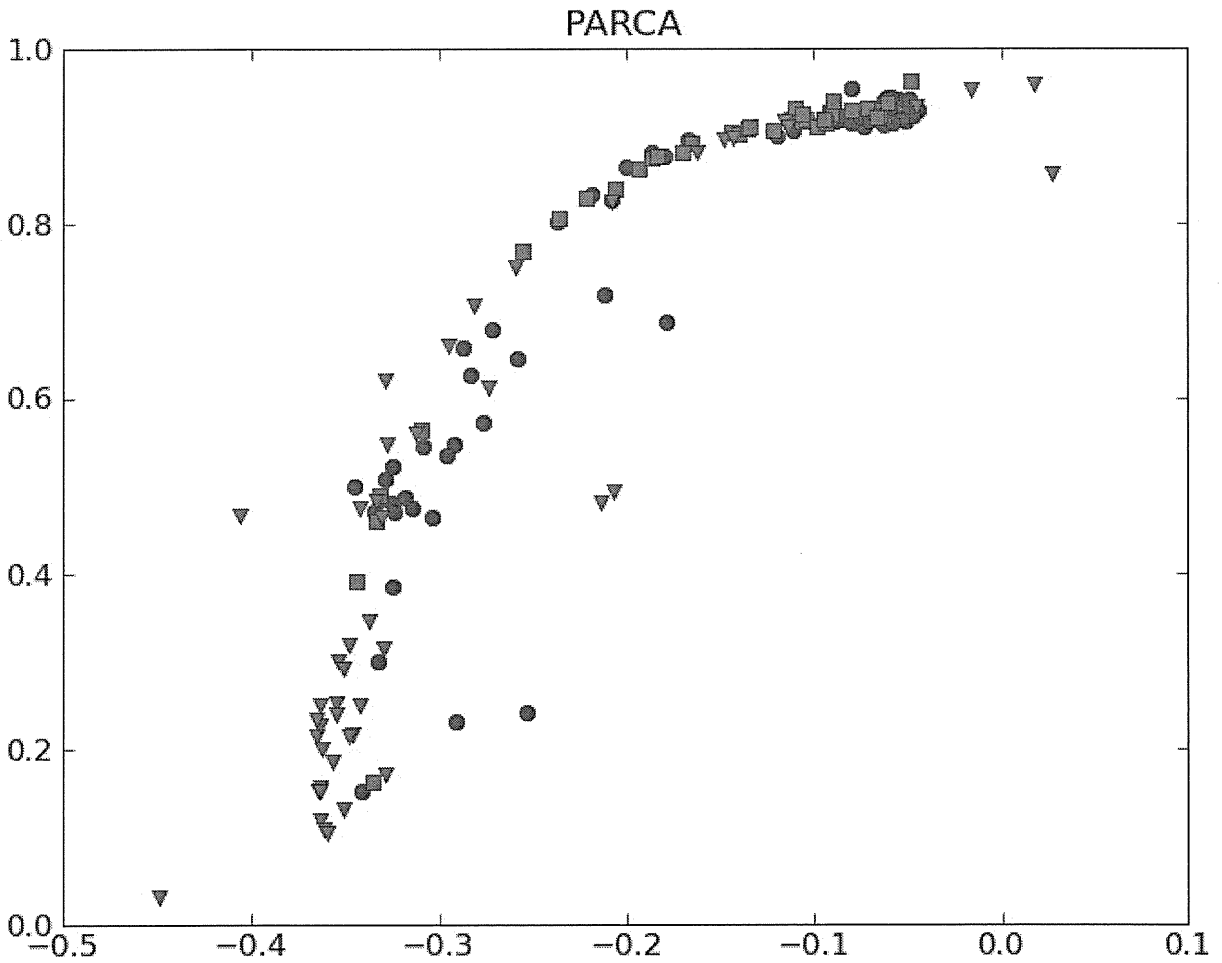

Figure 1. Chemical class structure for all time points. Blue circles: Neural class, green squares: Carcinogenic (genetic), red triangles: Carcinogenic (non-genetic).

We studied a second prediction problem, where the goal is to identify the chemical class for a given time point. As we have four different time points, this results in four datasets, each containing 200 samples (20 chemicals x 5 doses x 2 replicates) in 10 dimensions. We only considered the highest dose. There are three classes: neural, carcinogenic (genetic) and carcinogenic (non-genetic). The best cross-validation results, as well the parameters C and sigma of the gaussian SVM, are shown in table 1.

Table 1. Chemical classification results at different time points.

Time point	Best accuracy	C parameter	sigma
24H	0.64	1	1
48H	0.60	1000	0.001
72H	0.67	0.01	1000
96H	0.64	1	100

We now visualize these results by applying PARCA to the four time specific datasets. Figure 2 shows the results for T=72H, which corresponds to the highest accuracy. We only considered the highest dose.

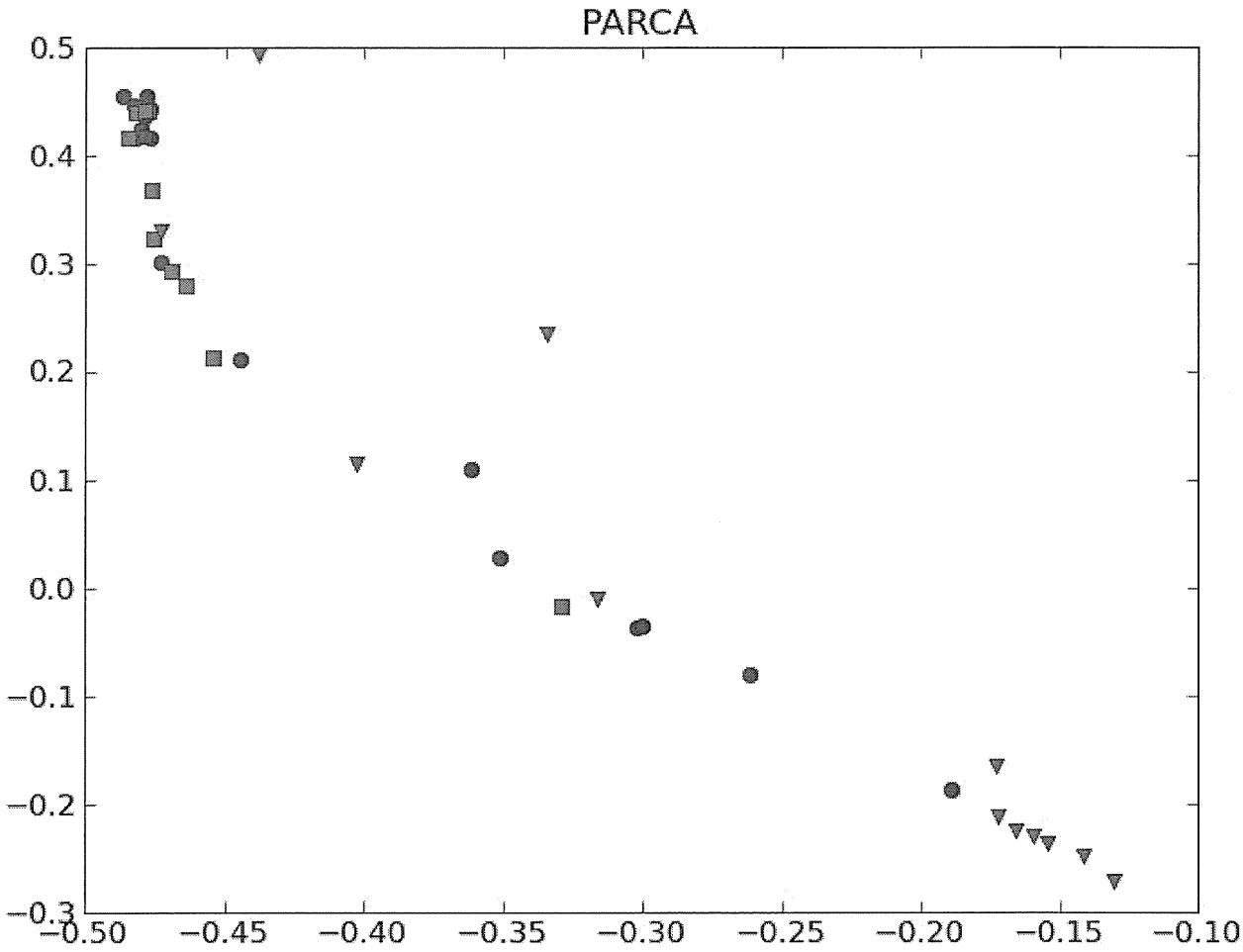

Figure 2. Chemical class structure at T=72H. Blue circles: Neural class, green squares: Carcinogenic (genetic), red triangles: Carcinogenic (non-genetic).

Dose versus control.

In this section, we are interested in studying the difference between two types of gene expression patterns: “dose” and “control”. Intuitively, we expect those two patterns to be different if the chemical induces a differential phenotype, and similar if the chemical has no effect.

We only considered the highest dose. The resulting dataset consisting of 184 samples in 10 dimensions, including 176 dose samples (22 chemicals x 4 time points x 2 replicates) and 8 control samples (4 time points x 2 replicates). The result of PARCA is shown in the figure 3.

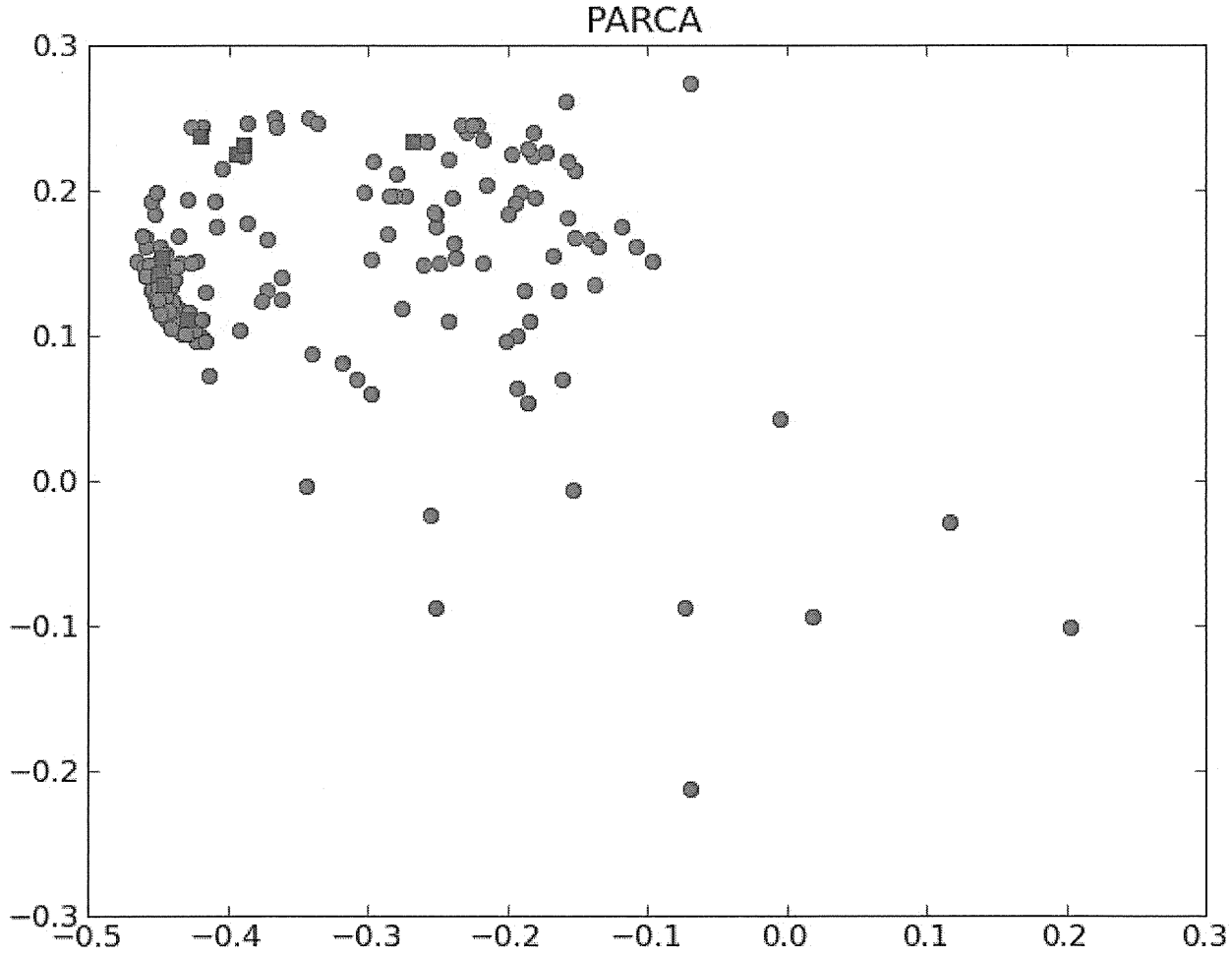

Figure 3. Dose versus control for all chemicals. Blue squares: control samples, red circles: dose samples.

Then, we performed the same analysis for each chemical separately. We prepared 22 datasets for the 22 chemicals. Each dataset contained 8 dose samples (2 replicates x 4 time points) and 8 control samples. We ran PARCA on each dataset and plotted the result. Figure 4 shows the PARCA results for bisphenol A, and figure 5 shows the results for Diethylbestrol.

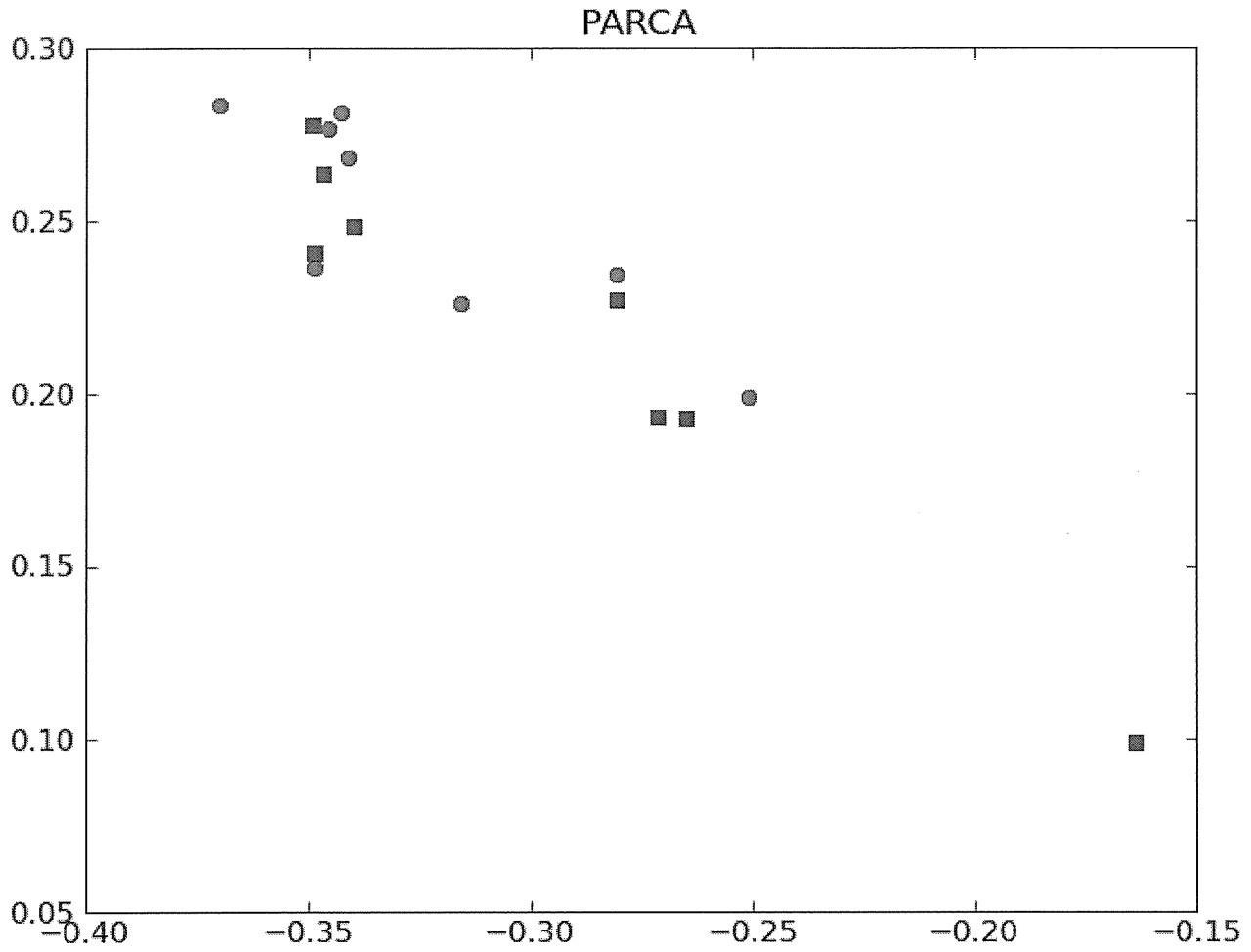

Figure 4. Dose versus control for Bisphenol A. Blue squares: control samples, red circles: dose samples.

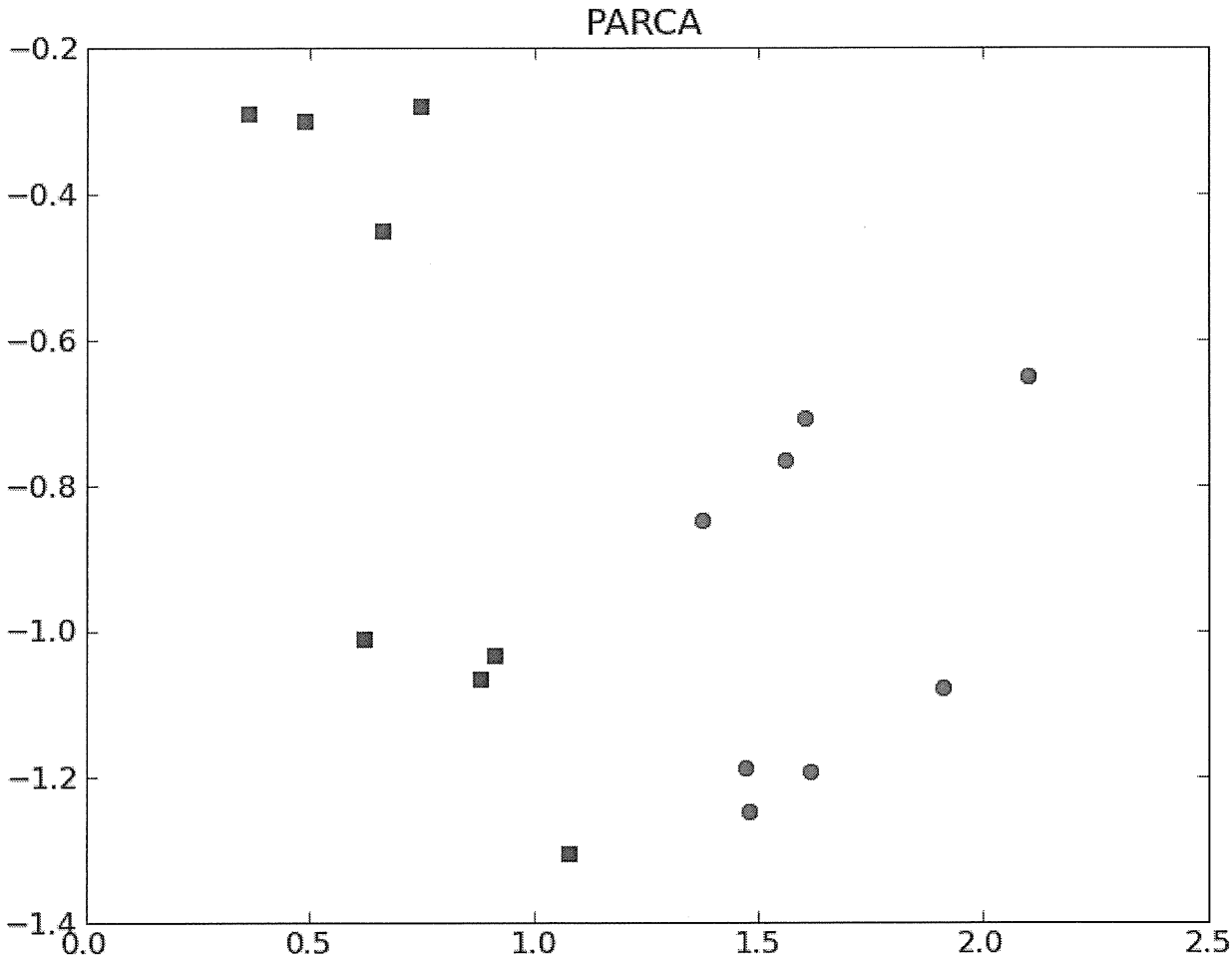

Figure 5. Dose versus control for Diethylbestrol. Blue squares: control samples, red circles: dose samples.

Visualization of chemicals separability.

In this section, we wanted to study if the chemicals are separable from each other, when considering all time points. We only considered the highest dose. We prepared a dataset containing 176 samples divided into 22 classes, and represented in 10 dimensions. Each class represents a chemical, and contains 8 samples (4 time points x 2 replicates). If the time only has limited influence on the gene expression patters, then the samples corresponding to a given chemical should be close to each other and possibly away from other samples. Figure 6 shows the result of PARCA.

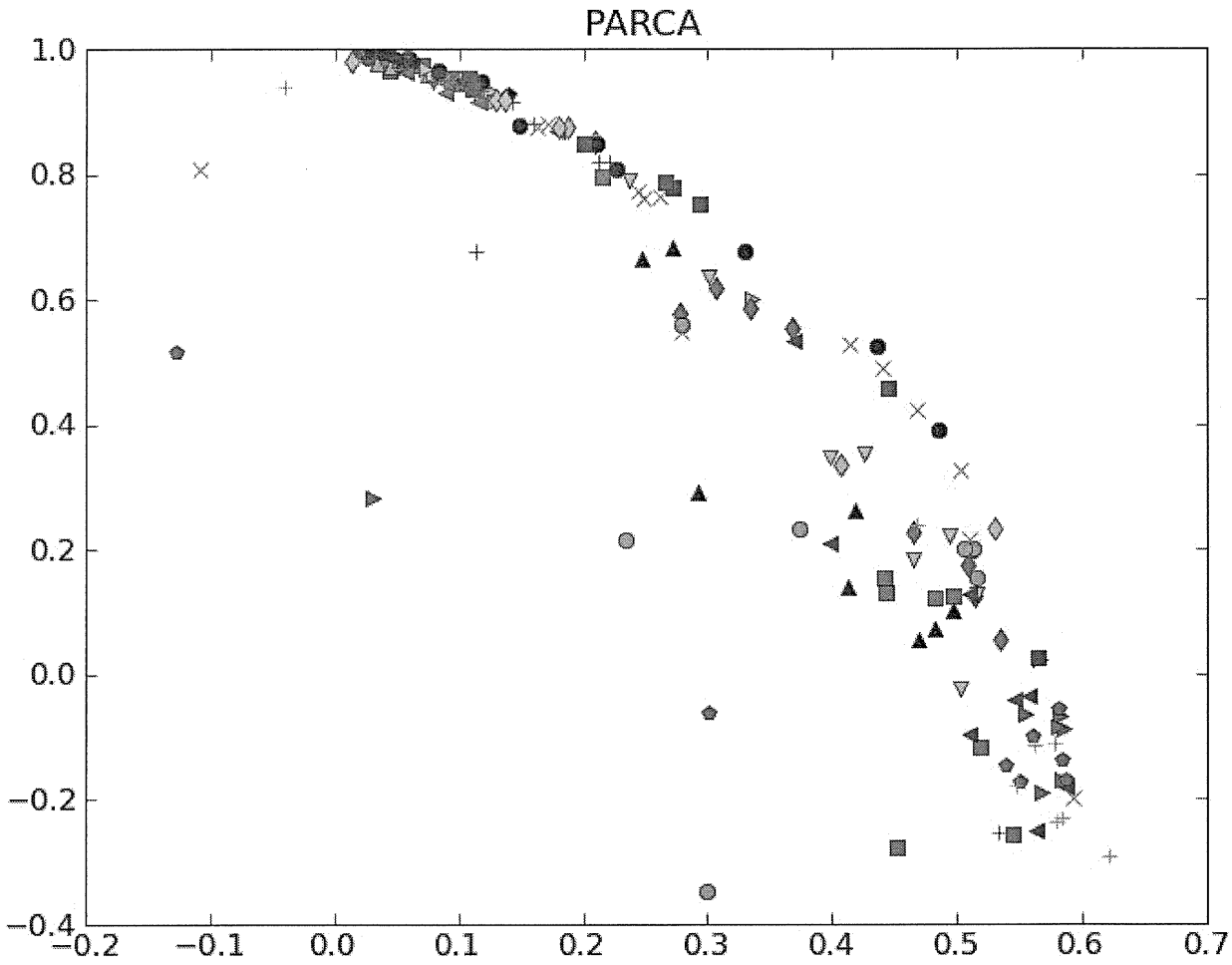

Figure 6. Visualization of chemicals separability. Each colored symbol represents a chemical.

Low dose versus High dose.

In this section, we studied the difference between low doses (1/16 and 1/8) and high doses (1/2 and 1). We wanted to know if gene expression patterns corresponding to low doses are different from gene expression patterns for high doses. If they are similar, it would suggest that the dose level does not have a strong influence on the gene expression patterns of a chemical.

We generated one dataset for each of the 22 chemicals. Each chemical contains 16 high dose samples (2 doses x 4 time points x 2 replicates) and 16 low dose samples (2 doses x 4 time points x 2 replicates) in 10 dimensions. We ran PARCA on each of the 22 datasets. Figure 7 shows the PARCA results for Bisphenol A and figure 8 shows the results for Diethylbestrol.

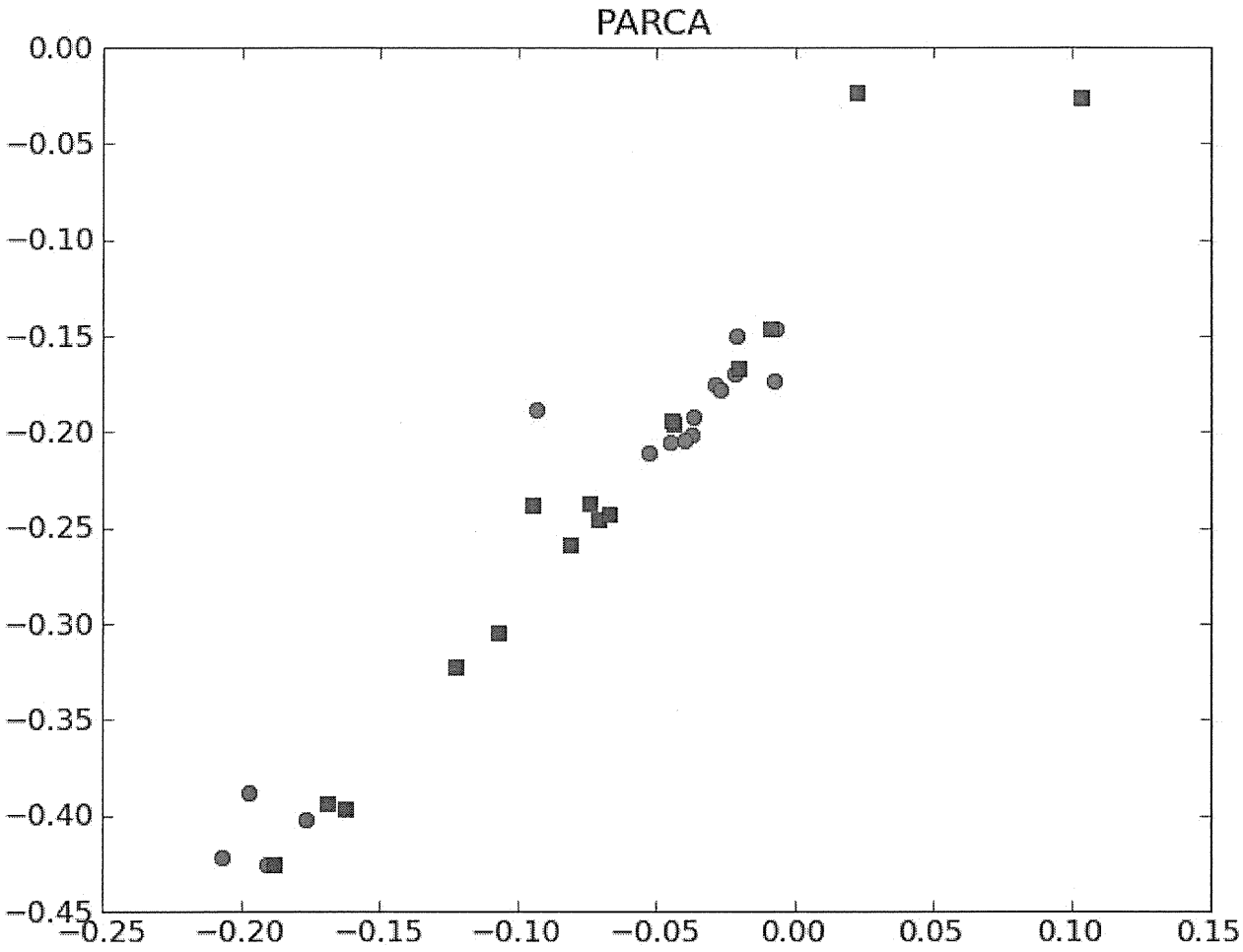

Figure 7. Low dose versus high dose for Bisphenol A. Red circles: low dose samples, blue squares: high dose samples.

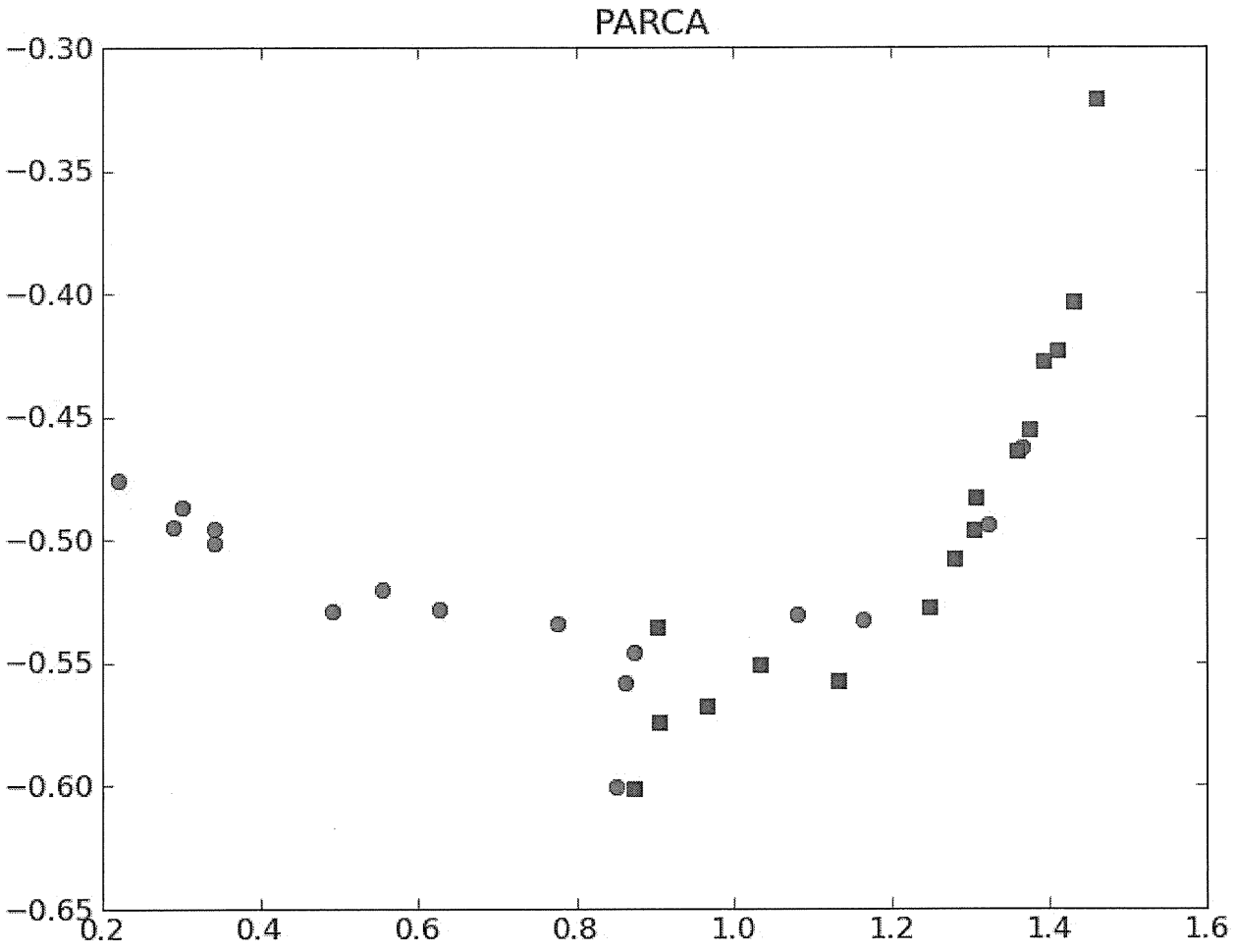

Figure 8. Low dose versus high dose for Diethylbestrol. Red circles: low dose samples, blue squares: high dose samples.

D. DISCUSSION

Chemical class identification.

The test accuracy of the global classification problem result is 0.60, which is only slightly better than random prediction (accuracy 0.50). This shows that the identification of the chemical class is a difficult prediction problem. One way to improve the prediction performance is to change the initial feature space, i.e., to add more genes to the initial gene list.

In the lower part of figure 1, we can see that the neural and carcinogenic (non-genetic) are somewhat separated. In the upper part, the three classes tend to be mixed in a single dense cluster, which explains the poor performance of the Gaussian SVM.

One of the difficulties might come from the fact that different types of chemicals might behave similarly at specific time points. For example, if the timing is too early to observe the chemical influence, or if the phenotype is back to normal, then no difference should be observed between two different chemicals.

The accuracy results for the four time-wise classification problems (table 1) show that the identification of chemical class becomes slightly easier when we consider specific time points. The lowest prediction accuracy 0.60 corresponds to T=48H, which is similar to the prediction performance when using all time points. The best prediction accuracy 0.67 corresponds to T=72H, and suggests that the specific phenotype differences of each chemical class are stronger at T=72H. Overall, the accuracies remain relatively low, and show that the differences between the three chemical classes neural, carcinogenic (genetic) and carcinogenic (non-genetic) are not strong.

In figure 2, we observe two different regions. In the right part of the figure, there is fairly good separation between the neural class (blue circles) and the carcinogenic (non-genetic) class (red triangles). This suggests that at T=72H, those two types of chemicals exhibit specific phenotypic differences. In the left part of the figure, the neural class and the carcinogenic (genetic) class (green squares) tend to be mixed together, suggesting that those two classes exhibit similar gene expression patterns at T=72H.

This confirms the class structure that we observed previously, and suggests that the time scale only plays a minor role in the specificity of gene expression patterns between the three chemical classes. It also shows that although the class prediction of a chemical is a difficult problem, it is also class specific. The visualization results suggest that predicting the neural class from the carcinogenic (non genetic) might be easier than predicting the neural class from the carcinogenic (genetic).

Dose versus control.

In figure 3, we cannot see any obvious separation between dose samples (red circles) and control samples (blues squares). A possible explanation is that for some chemicals, the phenotypes

corresponding to dose and control are very similar. For the remaining chemicals, a long distance between the dose samples and the control samples suggest that the corresponding chemical induces a strong differential phenotype. When the dose samples and the control samples are close, the corresponding chemical has no observable effect on the gene expression. However, this does not necessarily imply that those chemicals do not have any biological effect.

We observed good separability for most chemicals. This includes Diethylbestrol, for which the PARCA results are shown in Figure 5. For benzo[a]pyrene, bisphenol A, methapyrilene, and Phenobarbital, there was a poor separability between dose and control. Figure 4 shows the PARCA results for bisphenol A.

For those chemicals, it is possible that the effects are weak on individual genes, and need additional biomarkers (i.e., genes) to be accurately detected.

Visualization of chemicals separability.

Figure 6 shows the influence of time on each chemical's expression patterns. Each colored symbol represents a chemical. We observe a data structure where most samples are clustered together in the upper part of the figure. This suggests that the corresponding chemicals will be difficult to identify from each other. This result is different from our analysis using the Gaussian SVM, which showed that the class prediction was difficult. In figure 6, the remaining samples are scattered though the figure, but most chemicals tend to overlap each other, and show poor separability. As we previously noted, this might suggest that the available gene expression are not sufficient to distinguish the different chemicals, and additional genes could be useful.

Low dose versus high dose.

We observed that for most chemicals, the low dose samples and the high dose samples were clustered together. Figure 7 shows the PARCA results for Bisphenol A. For Diethylbestrol, Methapyrilene, phenobarbital and TCDD, there was a better separation between low dose samples and high dose samples . Figure 8 shows the results for Diethylbestrol.

For most chemicals, the dose does not seem to have a strong influence on the gene expression patterns. This suggests that for some chemicals, the resulting changes in phenotype become stronger as the dose increases. For some other chemicals, the changes do not depend much on the dose, and increases of the dose don't seem to further modify the changes in gene expressions. In the case of four chemicals (Diethylbestrol, Methapyrilene, phenobarbital and TCDD), we observed a good separation between low doses and high doses. This suggests that those chemicals are more sensitive to the dose levels, and show progressive effects in the way they affect the samples.

E. CONCLUSION

In this work, we analyzed toxicogenomic data in the framework of machine learning. First, we use a classification approach to predict the chemical class: Neural, Carcinogenic (genetic), or

Carcinogenic (non genetic). The prediction accuracy using the whole dataset was fairly low, and suggests that the current biomarkers (genes) do not contain sufficient information to predict the chemical class. We also showed that the chemical classes are not equivalent, and some classes such as the Neural class seem easier to predict. This was confirmed using PARCA, our supervised dimensionality reduction methods based on ranking the ranking approach. The visualization provided by PARCA clearly showed that the Neural class and the Carcinogenic (non-genetic) are well separated.

We also used PARCA to check if dose samples are well separated from control samples. For most chemicals, the gene expression patterns differed between dose and control. However, few chemicals didn't show any difference, which suggest that the current genes are not optimal biomarkers for these chemicals.

Next we used PARCA to visualize the separability of the chemicals, when considering all time points. The chemicals were not clearly separated from each other. Therefore, the initial gene space in which the chemicals are represented may not contain enough information to properly distinguish the chemical specificities.

Last, we used PARCA to study the differences between low doses and high doses. For most chemicals, the dose does not seem to have a strong influence on the gene expression patterns.

F. LIST OF PRESENTATIONS

1. JOURNAL PAPERS

Pessiot Jean-Francois, Kim Hyeryung, Fujibuchi Wataru. Pairwise Ranking Component Analysis. Submitted to Knowledge and Information Systems.

2. CONFERENCES

Conference papers with oral presentations.

Pessiot Jean-Francois, Kim Hyeryung, Fujibuchi Wataru. Learning similarity functions for multi-platform gene expression data. SIG-BIO (2011), 1-15.

Pessiot Jean-Francois, Fujibuchi Wataru. A ranking-based alternative to the discriminant analysis framework. Behaviormetric Society of Japan (2011), 1-4.

Poster presentations.

Pessiot Jean-Francois, Kim Hyeryung, Fujibuchi Wataru. Learning similarity functions for gene expression data. LS-BT (2011).

Pessiot Jean-Francois, Fujibuchi Wataru. Reverse engineering of genetic regulatory networks. Bioinformatics Week in Odaiba (2011).

G. INTELLECTUAL PROPERTIES

1. PATENTS
2. NEW IDEAS
3. OTHERS

H. REFERENCES

1. Amini MR, Truong TV, Goutte C (2008) A boosting algorithm for learning bipartite ranking

functions with partially labeled data. SIGIR' 08: Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval 99–106. doi:10.1145/1390334.1390354

2. Baccini A, Dejean S, Lafage L, Mothe J (2011) How many performance measures to evaluate information retrieval systems? Knowledge and Information Systems 1–21. Doi: 10.1007/s10115-011-0391-7

3. Baker LD, McCallum AK (1998) Distributional clustering of words for text classification. SIGIR ' 98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval 96–103. doi:10.1145/290941.290970

4. Bekkerman R, El-Yaniv R, Tishby N, Winter Y (2003) Distributional word clusters vs. words for text categorization. Journal of Machine Learning Research 3:1183–1208. doi:10.1.1.10.4861

5. Burges S, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. ICML' 05: Proceedings of the 22nd International Conference on Machine Learning 89–96. doi:10.1145/1102351.1102363

6. Burges CJC, Ragno R, Le QV (2007) Learning to Rank with Nonsmooth Cost Functions. Machine Learning 19:193–200. doi:10.1007/s10994-010-5185-8

7. Chapelle O, Shivaswamy P, Vadrevu S, Weinberger K, Zhang Y, TsengB (2010) Multi-task learning for boosting with application to web search ranking. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining 1189–1198. doi: 10.1145/1835804.1835953

8. Chen Y, Rege M, Dong M, Hua J (2008) Non-negative matrix factorization for semi-supervised data clustering. Knowledge and Information Systems 17:355–379. Doi: 10.1007/s10115-008-0134-6

9. Cohen WW, Schapire RE, Singer Y (1999) Learning to order things. Journal of Artificial Intelligence Research 10:243–270. doi:10.1613/jair.587

10. Cover TM, Thomas JA (1991) Elements of Information Theory. Wiley-Interscience. DOI:10.1002/047174882X.fmatter

11. Davis JV, Kulis B, Jain P, Sra S, Dhillon IS (2007) Information-theoretic metric learning. ICML ' 07: Proceedings of the 24th International Conference on Machine Learning 209–216. doi:10.1145/1273496.1273523

12. Dela Rosa K, Metsis V, Athitsos V (2011) Boosted ranking models: a unifying framework for ranking predictions. Knowledge and Information Systems 1–26. Doi: 10.1007/s10115-011- 0390-8

13. Dhillon IS, Modha DS (2001) Concept Decompositions for Large Sparse Text Data Using Clustering. Machine Learning 42:143–175. doi: 10.1023/A:1007612920971

14. Duda RO, Hart PE, Stork DG (2000) Pattern Classification. Wiley-Interscience.

doi:10.1007/s00357-007-0015-9

15. Forman G (2003) An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research 3:1289-1305. doi: 10.1162/153244303322753670

16. Freund Y, Schapire RE (1996) Experiments with a New Boosting Algorithm. ICML' 96: Proceedings of the 13th International Conference on Machine Learning 148-156. doi:10.1.1.133.1040

17. Freund Y, Iyer R, Schapire RE, Singer Y (2003) An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research 4:933-969. doi:10.1.1.30.8189

18. Globerson A, Roweis S (2006) Metric learning by collapsing classes. Advances in Neural Information Processing Systems 451-458. doi:10.1.1.61.7998

19. Goldberger J, Roweis S, Hinton G, Salakhutdinov R (2004) Neighbourhood components analysis. Advances in Neural Information Processing Systems 513-520.doi:10.1.1.108.7841

20. Har-peled S, Roth D, Zimak D (2003) Constraint classification for multiclass classification and ranking. Advances in Neural Information Processing Systems 785-792. doi:10.1.1.71.5954

21. Huang K, Ying Y, Campbell C, (2011) Generalized sparse metric learning with relative comparisons. Knowledge and Information Systems 28:25-45. Doi:10.1007/s10115-010-0313-0

22. Jain P, Kulis B, Dhillon IS, Grauman K (2008) Online Metric Learning and Fast Similarity Search. Advances in Neural Information Processing Systems. doi:10.1.1.140.1609

23. Jollife I (1986) Principal Component Analysis. Springer, New York, NY, USA. doi:10.1007/b98835

24. Kulis B, Sustik M, Dhillon IS (2006) Learning low-rank kernel matrices. ICML' 06: Proceedings of the 23rd International Conference on Machine Learning 505-512. doi:10.1145/1143844.1143908

25. Lee DD, Seung HS (1999) Learning the parts of ob jects by non-negative matrix factorization. Nature 401:788-791. doi:10.1038/44565

26. Lee DD, Seung HS (2001) Algorithms for Non-negative Matrix Factorization. Advances in Neural Information Processing Systems 556-562. doi:10.1.1.19.8636

27. Liu TY (2009) Learning to Rank for Information Retrieval. Foundations and Trends in Information Retrieval 3:225-331. doi: 10.1561/1500000016

28. Manning CD, Raghavan P, Schutze H (2008) Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA. doi:10.1007/s10791-009-9115-y

29. Martınez AM, Kak AC (2001) PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence 23:228-233. doi:10.1109/34.908974

30. Pereira F, Tishby N, Lee L (1993) Distributional clustering of English words. Proceedings of the

31st Annual Meeting on Association for Computational Linguistics 183-190. doi: 10.3115/981574.981598

31. Salton G, McGill MJ (1986) Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA

32. Schultz M , Joachims T (2004), Learning a Distance Metric from Relative Comparisons. Advances in Neural Information Processing Systems. doi:10.1.1.142.2314

33. Shalev-Shwartz S, Singer Y, Ng AY (2004) Online and batch learning of pseudo-metrics. ICML '04: Proceedings of the 21st International Conference on Machine Learning 743-750. doi:10.1145/1015330.1015376

33-b. Schölkopf,B. *et al.* (2004) *Kernel Methods in Computational Biology.* MIT Press, Cambridge, MA.

34. Shental N, Hertz T, Weinshall D, Pavel M (2002) Adjustment Learning and Relevant Component Analysis. ECCV '02: Proceedings of the 7th European Conference on Computer Vision 776-792. doi:10.1.1.19.2871

35. Slonim N, Tishby N (2000) Document clustering using word clusters via the information bottleneck method. SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval 208-215. doi:10.1145/345508.345578

36. Sugiyama M (2006) Local Fisher discriminant analysis for supervised dimensionality reduction. ICML '06: Proceedings of the 23rd International Conference on Machine Learning 905-912. doi:10.1145/1143844.1143958

37. Thurau C, Kersting K, Wahabzada M, Bauckhage C (2010) Convex non-negative matrix factorization for massive datasets. Knowledge and Information Systems 1-22. Doi: 10.1007/s10115-010-0352-6

38. Usunier N, Amini MR, Gallinari P (2005) Generalisation Error Bounds for Classifiers Trained with Interdependent Data. Advances in Neural Information Processing Systems 1369-1376. doi:10.1.1.69.9435

39. Usunier N, Buffoni D, Gallinari P (2009) Ranking with ordered weighted pairwise classification. ICML'09: Proceedings of the 26th International Conference on Machine Learning 1057-1064. doi:10.1145/1553374.1553509

39-b. Vapnik,V.N. (1998) *Statistical Learning Theory.* Wiley, New York.

40. Wang D, Li T, Ding C (2010), Weighted Feature Subset Non-negative Matrix Factorization and Its Applications to Document Understanding. Proceedings of the 2010 IEEE International Conference on Data Mining 541-550. doi: 10.1109/ICDM.2010.47

41. Weinberger KQ, Blitzer J, Saul LK (2006) Distance Metric Learning for Large Margin Nearest Neighbor Classification. Advances in Neural Information Processing Systems 1473- 1480. doi:10.1.1.117.5831

42. Xing EP, Ng AY, Jordan MI, Russell S (2002), Distance Metric Learning, with Application to

Clustering with Side-information. Advances in Neural Information Processing Systems 505-512. doi:10.1.1.58.3667

43. Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval 267-273. doi:10.1145/860435.860485

44. Yang Y, Pedersen JO (1997) A Comparative Study on Feature Selection in Text Categorization. ICML '97: Proceedings of the 14th International Conference on Machine Learning 412-420. doi:10.1.1.32.9956