PROGRESS: Projection-Based Gene Expression Classification

Kristóf Marussy[1,2] and Krisztián Buza[2]

[1]Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Hungary, marussy@cs.bme.hu

[2]Institute of Genomic Medicine and Rare Disorders, Semmelweis University, Budapest, Hungary, chrisbuza@yahoo.com

Background

Gene expression profiles were found to be highly relevant for safety assessment, diagnostics and prognostics applications [1, 5]. Recent advancements in high-throughput sequencing technology lead to growing interest in predictive classification (see Figure 1) models for gene expression data. For example, Ion AmpliSeq™ technology delivers simple and fast library construction for affordable targeted sequencing of specific human genes [3].

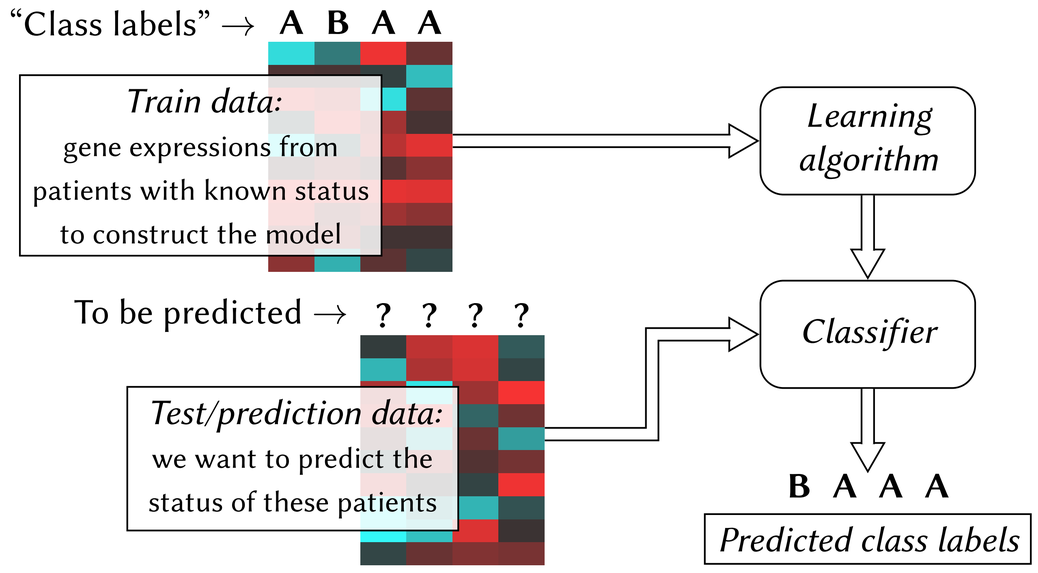

Figure 1: Machine learning for gene expression data.

Hubness in Gene Expression Data Sets

- A gene expression *instance* may contain expression values of thousands of genes. Therefore, instances are represented as vectors in very high-dimensional Euclidean space.
- *Hubness* is a phenomenon in high-dimensional data sets, such as gene expression data, that challenges classification algorithms [4].
- *Hubs* are instances that are similar to a suprisingly large number of other instances according to some measure of similarity, e.g. *Euclidean distance* $d(x_i, x_j) = \sqrt{\sum_\ell (x_{i,\ell} - x_{j,\ell})^2}$, where $x_{i,j}$ is the expression value of the ℓth gene in the ith sample.
- Hubness-aware classifiers (HW-KNN, HFNN, NHBNN, HIKNN) [6] are one of the most promising research directions aiming to enchance classificitation in high-dimensional spaces. To compare our approach to hubness-aware methods, we run HIKNN with parameter $k = 5$ as a baseline.

Our Contribution

- We attempt to mitigate hubness artefacts with dimensionality reduction via instance *projection* using *base points* (see Figure 2). A *logistic regression* classifier is trained on the resulting representation.
- Our results show that a single projection classifier has suboptimal accuracy (see Figure 3 and Table 1). However, it is very simple to construct an ensemble of such learners, which increases accuracy substantially.
- Each member of the ensemble performs projection using a different random base point set. Prediction output is decided by majority vote. We call this method PROGRESS: Projection-Based Gene Expression Classification.

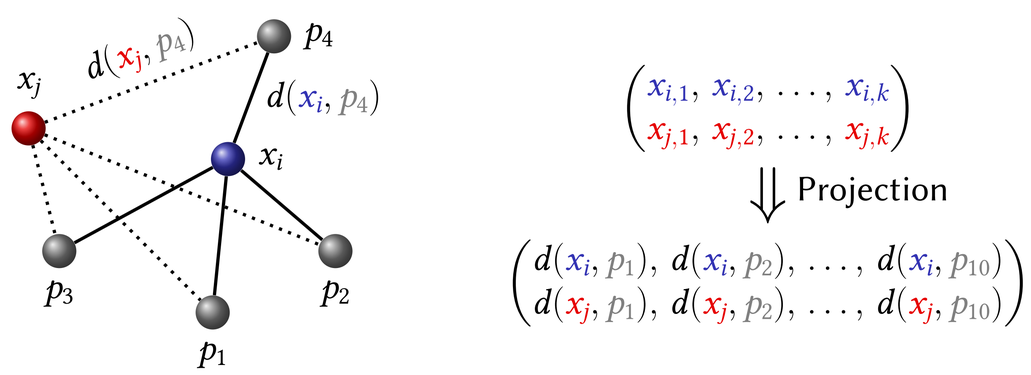

Figure 2: The distance of the instances is measured from the randomly selected projection base points $p_1, p_2, \ldots, p_{10}$, which are instances themselves. This projection represents the instances as vectors of distances instead of vectors of gene expressions.

Experimental Evaluation

In our ongoing research, we ran classification experiments on two publicly available data sets:

- The Breast Cancer data set consists of 32 ER− and 65 ER+ specimens from breast cancer patients with 7650 genes [5].
- The Colon Cancer data set consists of 40 colon tumor tissue samples and 22 normal colon tissue samples with 2000 genes [1].

We evaluated the following classifiers:

- Support Vector Machines with linear, polynomial and RBF kernels [2],
- HIKNN with $k = 5$ and Euclidean distance [6],
- Classification with logistic regression after projection to 10 randomly selected base points (Figure 2),
- Homogenous PROGRESS ensemble of 1000 projection classifiers.

We report the accuracy of the classifiers averaged over 10×10-fold cross-validation in Figure 3 and Table 1.

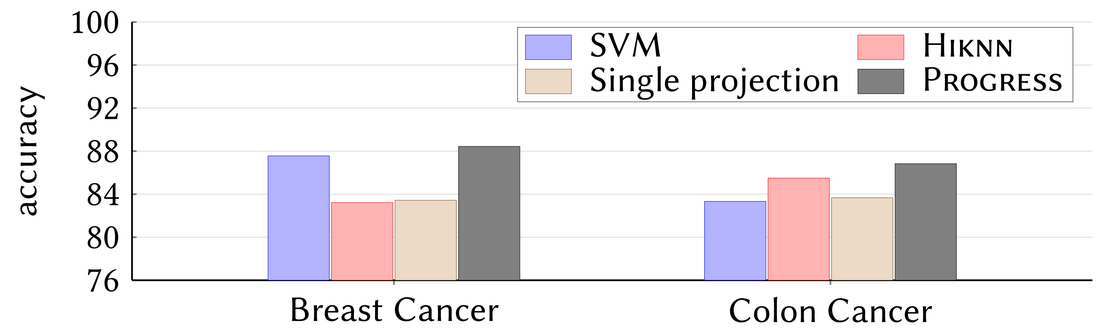

Figure 3: Average accuracy of classifiers over 10×10-fold cross-validation.

	Breast Cancer		Colon Cancer	
Best SVM	linear	87.56%	polynomial	83.33%
HIKNN	—	83.22% •	—	85.50%
Single projection	—	83.44% •	—	83.67%
PROGRESS ensemble	—	**88.44%**	—	**86.83%** ○

Table 1: Best-performing classifiers and their accuracies. Significantly better accuracy than SVM is denoted by ○, while significantly worse accuracy is denoted by •. Statistical significance was evaluated by two-tailed permutation test at $p < 0.05$.

Conclusions and Outlook

- Our preliminary results show that the PROGRESS projection ensemle can outperform Support Vector Machines and hubness-aware HIKNN on gene expression data sets.
- On the Breast Cancer data set, PROGRESS delivered the highest accuracy.
- On the Colon Cancer data set, both HIKNN and PROGRESS outperformed SVMs, but only PROGRESS outperformed them significantly.
- As future work, we plan to explore data pre-processing for classification and learning of distance functions.

Acknowledgement: This work was performed within the framework of the Hungarian Scientific Research Fund (grant no. OTKA 111710). This research was partially funded by the National Brain Research Program (project no. KTIA_13_NAP-A-III/6, project id KTIA_NAP_13-1-2013-0001).

References

[1] U. Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.*, 96(12):6745–6750, 1999.
Data set publicly available at http://genomics-pubs.princeton.edu/oncology/.

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines, 2001.

[3] Life Technologies. Targeted RNA sequencing by Ion Torrent next-generation sequencing, 2013. http://www.lifetechnologies.com, accessed Sept. 11, 2014.

[4] M. Radovanovic et al. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

[5] C. Sotiriou et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U.S.A.*, 100(18):10393–10398, 2003. Data set publicly available at http://www.pnas.org/.

[6] N. Tomasev and D. Mladenic. Nearest neighbor voting in high dimensional data: Learning from past occurrences. *Comput. Sci. Inf. Syst.*, 9(2):691–712, 2012.