Representative Prototype Sets for Data Characterization and Classification

Ludwig Lausser*, Christoph Müssel*, and Hans A. Kestler**

Research Group Bioinformatics and Systems Biology
Institute of Neural Information Processing, Ulm University, 89069 Ulm, Germany
hans.kestler@uni-ulm.de

Abstract. Common classifier models are designed to achieve high accuracies, while often neglecting the question of interpretability. In particular, most classifiers do not allow for drawing conclusions on the structure and quality of the underlying training data. By keeping the classifier model simple, an intuitive interpretation of the model and the corresponding training data is possible. A lack of accuracy of such simple models can be compensated by accumulating the decisions of several classifiers. We propose an approach that is particularly suitable for high-dimensional data sets of low cardinality, such as data gained from high-throughput biomolecular experiments. Here, simple base classifiers are obtained by choosing one data point of each class as a prototype for nearest neighbour classification. By enumerating all such classifiers for a specific data set, one can obtain a systematic description of the data structure in terms of class coherence. We also investigate the performance of the classifiers in cross-validation experiments by applying stand-alone prototype classifiers as well as ensembles of selected prototype classifiers.

1 Introduction

The rapid development of molecular high-throughput technologies has driven the need for computational approaches to mine and analyze the resulting data. These data sets usually comprise only a small set of probes, while being extremely high-dimensional. An intuitive understanding of such data is usually impossible. Classifiers can provide decision support to life scientists when judging new probes. However, researchers are often interested in the basic characteristics that distinguish probes of different types. Feature selection techniques try to identify those features (e.g. genes) that are relevant for the classification.

Instead of selecting relevant features, we propose an approach which identifies probes that characterize certain classes well. This method is based on simple prototype set classifiers that comprise one sample from each class. Due to their simple structure and their data dependency, it is possible to enumerate all such classifiers for low-cardinality data sets, such as microarray data. This allows for a systematic characterization of the data set. For instance, the classification

* L. Lausser and C. Müssel contributed equally.
** Corresponding author.

N. Mana, F. Schwenker, and E. Trentin (Eds.): ANNPR 2012, LNAI 7477, pp. 36–47, 2012.

performance of such simple base classifiers can give insight into the distribution and coherence of the classes. We describe several ways of extracting and visualizing information obtained from the universe of basic prototype set classifiers on a data set. Although single prototype set classifiers may be too simple to achieve a competitive classification accuracy, ensembles of such base classifiers that complement each other well can achieve a performance similar to state-of-the-art classifiers. We propose an ensemble method and evaluate it on several microarray data sets.

Prototype-based classification is used in many state-of-the-art classifiers, such as k-Nearest Neighbour (k-NN) classification [1], Learning Vector Quantization (LVQ) [2,3], Nearest Centroid classification, Nearest Medioid classification, or Nearest Shrunken Centroid classification [4]. With the exception of k-NN, these approaches generate prototypes based on the training set instead of directly drawing data points from the training set. Kuncheva and Bezdek [5] analyze whether prototypes should be selected or generated from the training set. They conclude that prototype selection should be preferred over prototype generation, as determining clusters in the data does not guarantee a good classification performance.

Our concept is related to approaches aiming at a reduction of the training set for the k-NN classifier. E.g., the Condensed k-NN approach tries to reduce the training subset in such a way that it maintains the performance of the full training set [6]. Many approaches make use of search heuristics, such as Genetic Algorithms (e.g. [7,8]). An overview of training subset selection for k-NN is given in [9] and [10].

Our approach differs from such neighbourhood condensation methods in several ways: Firstly, we focus on a very simple prototype representation using only a single prototype per class, while k-NN neighbourhood condensation can yield reduced training sets of arbitrary size. In this way, all possible classifiers can be examined, without the need of search heuristics to identify optimal sample subsets. Secondly, we consider these classifiers as base learners, i.e. they are not meant to achieve a high prediction accuracy on their own. Instead, they serve as data set descriptors and members of ensemble classifiers.

The structure of the paper is as follows: Section 2 describes the basic Representative Prototype Set (RPS) classifiers, their use for the characterization of data sets, and ensembles of RPS classifiers. In Section 3, the results of an application of the new methods to six well-known microarray data sets are presented. Section 4 discusses the results and concludes the paper.

2 Representative Prototype Set Classification

2.1 Basic Prototype Classifiers and Their Properties

We define a basic Representative Prototype Set (RPS) classifier as a set of prototypes, one for each class. The prototypes $\mathcal{P}$ are chosen from the training set. The labels of unseen data points are predicted according to the label of the nearest prototype in the set.

Let $\mathcal{T} = \bigcup_{i=1,\dots,k} \mathcal{T}_i$ denote a labeled training set comprising k classes. A Representative Prototype Classifier consists of a set of prototypes

$$\mathcal{P} = \{p_i \in \mathcal{T}_i \mid i = 1, \dots, k\},$$

where each $p_i = (\mathbf{x}_i, i)$ is a feature vector $\mathbf{x}_i \in \mathbf{R}^n$ labeled with class label i.

The classifier predicts an unseen data point $\mathbf{v}$ by choosing the prototype p_i with the smallest distance $d(\mathbf{v}, \mathbf{x_i})$, i.e.

$$RPS_{\mathcal{P}}(\mathbf{v}) = \text{argmin}_{i=1,\dots,k}\ d(\mathbf{v}, \mathbf{x_i}).$$

In the following, we use the Euclidean distance.

The RPS classifier is a special case of general prototype classifiers PC which rely on a set of prototypes

$$\mathcal{Q} \subseteq \mathcal{T},$$

i.e, $\mathcal{Q}$ is not necessarily restricted to a single prototype per class.

The general prototype concept described above is data-dependent: A classifier c is called data-dependent if it can be determined entirely according to a relatively small set of training samples $\mathcal{T}' \subseteq \mathcal{T}$. That is, the training on both sets will result in the same classification model,

$$c_{\mathcal{T}'} = c_{\mathcal{T}}. \tag{1}$$

The set of samples $\mathcal{T}'$ is called the *compression set* of the data-dependent classifier. For the above concept, the compression set is equal to $\mathcal{Q}$. If the prototype set corresponds to the complete training set ($\mathcal{Q} = \mathcal{T}$), the data-dependent prototype classifier corresponds to the well-known 1-Nearest Neighbour classifier [1].

Data-dependent classifiers allow for the specification of sample compression bounds [11]. These bounds can be used to give an upper limit of the true classification error probability $\mathcal{R}$ of a data-dependent classifier c. The main component of a sample compression bound is an empirical error rate R calculated on the remaining set of samples $\mathcal{T} \setminus \mathcal{Q}$.

Theorem 1 (e.g. [12]). *For a random sample $\mathcal{T}$ of* iid *examples drawn from an arbitrary, but fixed distribution $\mathcal{D}$ and for all $\delta \in (0, 1]$,*

$$\Pr_{\mathcal{T} \sim \mathcal{D}^{|\mathcal{T}|}} \left(\forall \mathcal{Q} \subseteq \mathcal{T} \text{with } c = PC(x, \mathcal{Q}) \colon \mathcal{R}_{\mathcal{D}}(c) \leq \overline{\text{Bin}} \left(\text{R}(c, \mathcal{T} \setminus \mathcal{Q}), \frac{\delta}{|\mathcal{T}| \binom{|\mathcal{T}|}{|\mathcal{T} \setminus \mathcal{Q}|}} \right) \right) \geq 1 - \delta$$

Here, $\overline{\text{Bin}}$ denotes the binomial tail inversion.

Proof. The sample compression bound given in Theorem 1 is a direct application of the sample compression bound in [12]. □

If there is a set of data-dependent prototype-based classifiers that all achieve the same empirical error rate, it follows from Theorem 1 that the classifier with the smallest compression set is most reliable. This means that in case of several classifiers with different numbers of prototypes, but equal performance, the extreme case of choosing the smallest possible reference set $\mathcal{P}$ as in RPS is the most favorable option.

2.2 Analyzing Data Set Characteristics

The total number of possible RPS classifiers for a training set $\mathcal{T}$ is $\prod_{i=1,\ldots,k} |\mathcal{T}_i|$. Consequently, it is often feasible to enumerate the complete set of RPS classifiers for a given data set, in particular regarding biomolecular data sets of low cardinality. This complete set can be used to characterize a data set according to the coherence of classes and representativeness of single training data points. For the following visualization and summarization approaches, we focus on two-class data sets.

For each RPS classifier c based on a prototype set $\mathcal{P}$, we can measure its empirical accuracy on the remaining training samples that are not included in the classifier:

$$\mathrm{A}(c, \mathcal{T} \setminus \mathcal{P}) = 1 - \mathrm{R}(c, \mathcal{T} \setminus \mathcal{P}) = \frac{|\{(\mathbf{v}, l) \in \mathcal{T} \setminus \mathcal{P} \mid c(\mathbf{v}) = l\}|}{|\mathcal{T} \setminus \mathcal{P}|}$$

For a two-class data set, this can be visualized in a heat map: the samples of the first class are plotted on the x axis, and the samples of the second class are plotted on the y axis. The greyscale color indicates the empirical accuracy of a combination, with a light color denoting a high accuracy and a dark color denoting a low accuracy. By applying complete-linkage hierarchical clustering, samples that exhibit a similar accuracy in combination with samples from the other class are grouped.

To get an impression of how well the data can be described by small sets of representative prototypes, we plot the distribution of the empirical accuracies A for all possible RPS classifiers c in form of a histogram. If many classifiers achieve high empirical accuracies, the histogram shows a right-skewed distribution, whereas data sets that are hard to separate by small prototype sets show a left-skewed distribution.

The empirical error rate of the classifiers is not the only performance measure that can be used to characterize a dataset. It is also of interest to know if all classifiers misclassify more or less the same set samples. A possible measure of this similarity of two classifiers c_a and c_b is Yule's Q statistic [13]:

$$Q_{i,j} = \frac{M^{11}M^{00} - M^{01}M^{10}}{M^{11}M^{00} + M^{01}M^{10}} \tag{2}$$

Here, M^{ij} denotes the number of co-occurences of correct predictions (1) or incorrect predictions (0) of c_a and c_b. E.g, M^{11} is the number of samples that are predicted correctly by both c_a and c_b, whereas M^{10} is the number of samples that are predicted correctly by c_a, but misclassified by c_b.

The range of Q is $[-1, 1]$. It is equal to 1 if both classifiers correctly predict exactly the same set of samples. It is equal to -1 if all samples correctly classified by c_a are misclassified by c_b and vice versa. If c_a and c_b are statistically independent, $\mathbb{E}[Q] = 0$.

Calculating Q for all pairs of possible RPS classifiers provides further information on a data set: If many pairs yield a Q close to 1, the data set is probably very coherent, and classes consist of a single cluster. By contrast, if many pairs

achieve a Q close to -1, a prototypic description of the data set is not easily available, and classes probably consist of several disjoint clusters. A histogram of the distribution of Yule's Q on a data set can reveal such information.

2.3 Ensemble RPS Classification

A single RPS classifier can be seen as a simple base learner, but may not achieve good accuracies if any of the classes is distributed over two or more separate clusters due to the fact that each class is represented by only one prototype. To achieve a higher robustness and accuracy, several basic RPS classifiers can be combined to an ensemble classifier (denoted as eRPS). In the following, the ensemble set of m base learners is written as

$$\mathcal{E}_m = \{c_1, \ldots, c_m\}$$

To train an eRPS ensemble on a training data set $\mathcal{T}$, we determine all possible base RPS classifiers c_j and order them by their empirical accuracies $A(c_j, \mathcal{T})$. $\mathcal{E}_m$ then consists of the m RPS classifiers with the highest empirical accuracies.

An unseen sample $\mathbf{v}$ is classified according to a majority vote of the base learners, i.e.

$$eRPS_{\mathcal{E}_m}(\mathbf{v}) = \text{argmax}_{1,\ldots,k} \mid \{c_j \in \mathcal{E}_m \mid c_j(\mathbf{v}) = i\} \mid$$

3 Experiments

We applied our data set characterization as well as the ensemble classifier to several well-known microarray data sets:

The *Bittner data set* [14] contains expression profiles of 31 melanomas and 7 controls in 8067 features. The initial analysis of this data showed a stable cluster of 19 of the melanomas. In this analysis, the samples from this cluster (ML1) and the 19 remaining samples (melanomas and controls, ML2) were treated as distinct classes.

The *Golub data set* [15] contains data from a microarray experiment of acute Leukemia. The data set contains examples for two disease subtypes: ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia). The 47 ALL and 25 AML examples consist of 3571 expression measurements. The probes were selected and normalized according to a procedure proposed by Dudoit et al. [16].

The *Notterman data set* [17] contains 18 paired samples of colon adenocarcenomas (CA) and normal tissues (N). The expression profiles comprise 7457 features.

The *Pomeroy data set* [18] data set contains examples of two different kinds of embryonal tumors of the central nervous system, 25 classic medulloblastomas (CMD) and 9 desmoplastic medulloblastomas (DMD). The dataset contains 7129 unspecific genes.

The *Shipp data set*[19] consists of 77 samples of single B-cell lineage. 58 of these samples are classified as diffuse large B-cell lymphoma (DLBCL); the other

19 samples are follicular lymphoma (FL). The expression profiles of 7129 genes were collected using an Affymetrix HU6800 chip.

The *West data set* [20] comprises different breast cancer types. It contains 49 samples, which can be distinguished according to their estrogen receptor status (25 ER+ and 24 ER-). The expression profiles of 7129 features were measured using the HuGeneFL Chip.

For these six data sets, we include the heatmaps and accuracy histograms for single RPS classifiers as well as the histograms of Yule's Q statistic on pairs of RPS classifiers in Figures 1 and 2. The left columns of the figures show heatmaps of the accuracies of possible prototype sets in the data set. The greyscale color indicates the empirical accuracy of a combination, with a light color denoting a high accuracy and a dark color denoting a low accuracy. By applying a hierarchical clustering algorithm, samples that exhibit a similar accuracy in combination with samples from the other class are grouped. The columns in the middle show histograms of the same accuracies. The right columns depict histograms of the distribution of Yule's Q statistic for all pairs of prototype sets. This statistic measures how similarly these prototype sets predict the labels of samples.

To assess the classification performance of the RPS classifier, we conducted stratified cross-validation experiments, splitting the data into 10 random subsets, each of which was once used as a test set to measure the error, while the remaining samples were used for training. The cross-validation error was summed up over the 10 subsets, and the whole procedure was repeated 10 times. This yields a mean cross-validation error over the 10 runs.

Table 1 lists the mean cross-validation error percentage of the eRPS classifier, the k-Nearest Neighbour classifier, and the SVM with linear and RBF kernel for different configurations:

For the eRPS classifier, the number of prototype sets m in the ensemble was varied from 1 to 15. For k-NN, the number of neighbours k was varied, and for the linear SVM, different values of the cost parameter were applied. For the SVM with RBF kernel, we set the cost parameter to 100 (which appeared to yield the best results) and varied the γ parameter.

On the Bittner data set, eRPS clearly outperforms all other classifiers with only a single prototype set. The performance decreases when adding more sets to the majority vote, but remains clearly better than the results of the other classifiers. The accuracy histogram in Figure 1 can give a possible explanation for the good performance with few prototype sets: Most such base classifiers achieve a bad accuracy of 0.5 or less. However, there is a small number of prototype sets with an accuracy of more than 90%. As their performance is drastically better than nearly all other prototype sets, these good prototype sets are likely to be the top-ranked sets even in resampling settings, so that they are also included in the ensembles derived from the cross-validation subsets.This can also be seen in the heatmap: Many of the samples seem to be completely unsuitable as prototypes, as they yield a bad performance in any combination. Only a small set of configurations achieves a considerably higher performance.

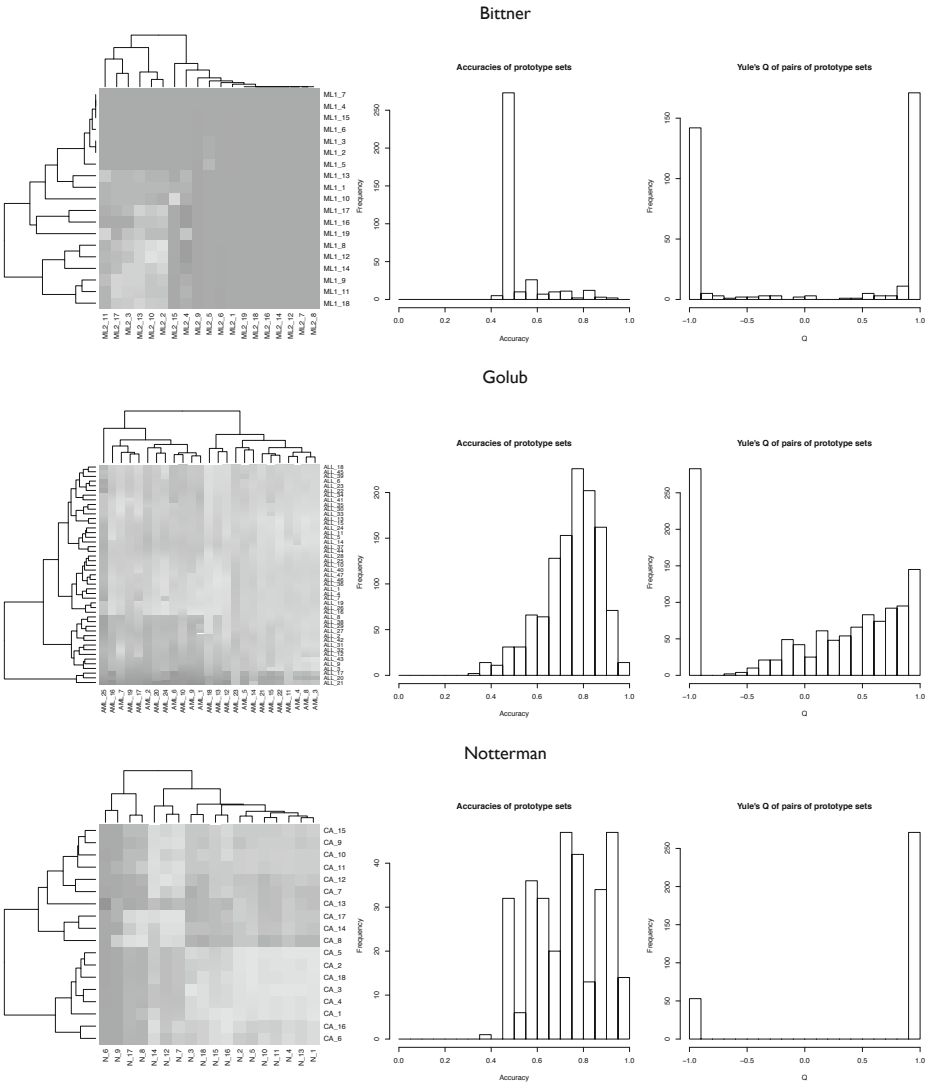

Fig. 1. Visualization of data set properties for the Bittner data set, the Golub data set and the Notterman data set

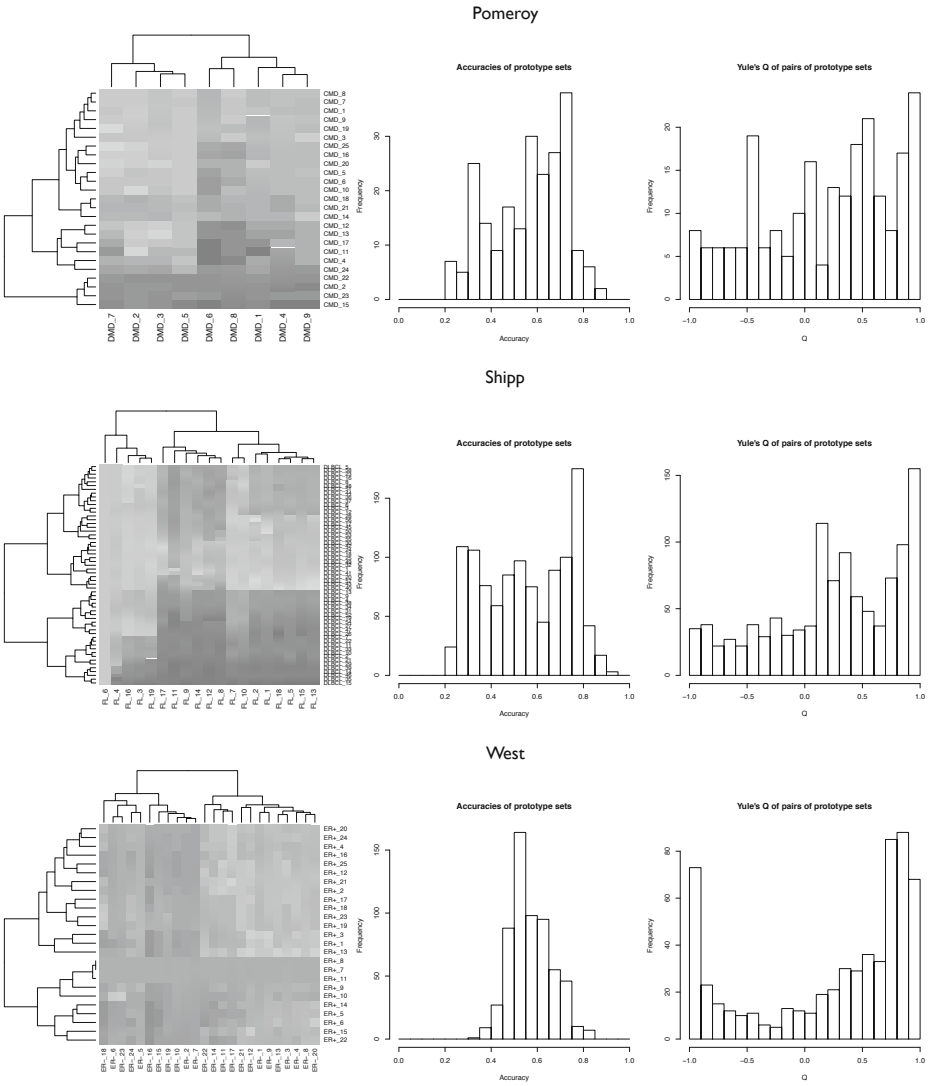

Fig. 2. Visualization of data set properties for the Pomeroy data set, the Shipp data set and the West data set

Table 1. Mean cross-validation prediction error percentage of Representative Prototype Set ensembles, k-Nearest Neighbour, and Support Vector Machines on six microarray data sets

	Bittner	Golub	Notterman	Pomeroy	Shipp	West
eRPS, $m = 1$	**7.63**	10.69	4.72	27.06	19.48	34.69
eRPS, $m = 3$	7.89	4.58	3.61	19.12	12.08	25.71
eRPS, $m = 5$	8.16	3.33	3.06	17.35	8.70	21.02
eRPS, $m = 7$	10.00	2.50	3.33	19.41	6.49	18.57
eRPS, $m = 9$	10.00	1.67	**2.78**	22.94	5.84	18.16
eRPS, $m = 11$	10.00	1.94	**2.78**	24.12	5.71	18.16
eRPS, $m = 13$	10.53	2.36	**2.78**	23.53	5.71	18.16
eRPS, $m = 15$	10.79	2.92	**2.78**	25.29	5.71	18.78
k-NN, $k = 1$	27.89	2.92	**2.78**	21.18	13.12	10.41
k-NN, $k = 3$	22.89	1.94	3.33	21.76	12.47	23.47
k-NN, $k = 5$	19.21	3.33	3.06	19.41	11.43	25.10
k-NN, $k = 7$	32.11	3.06	5.83	24.41	10.13	26.94
k-NN, $k = 9$	37.63	2.50	5.83	25.00	7.66	27.96
k-NN, k = 11	41.05	3.33	6.67	26.47	7.40	27.14
linear SVM, cost = 0.001	22.89	**1.39**	**2.78**	18.53	**3.77**	**8.98**
linear SVM, cost = 0.01	22.89	1.94	**2.78**	18.53	**3.77**	**8.98**
linear SVM, cost = 0.1	22.89	1.94	**2.78**	18.53	**3.77**	**8.98**
linear SVM, cost = 1	22.89	1.94	**2.78**	18.53	**3.77**	**8.98**
linear SVM, cost = 10	22.89	1.94	**2.78**	18.53	**3.77**	**8.98**
RBF SVM, cost = 100, $\gamma = 10^{-07}$	32.89	34.72	24.72	26.47	24.68	57.14
RBF SVM, cost = 100, $\gamma = 10^{-06}$	22.89	25.97	**2.78**	20.88	**3.77**	11.63
RBF SVM, cost = 100, $\gamma = 10^{-05}$	21.32	1.53	5.56	**16.47**	5.32	12.65
RBF SVM, cost = 100, $\gamma = 10^{-04}$	28.16	2.78	12.50	17.65	7.92	16.94
RBF SVM, cost = 100, $\gamma = 10^{-03}$	46.58	10.42	46.67	26.47	24.68	40.61

On the Golub data set, one configuration of the linear SVM achieves a slightly smaller error than the best eRPS classifiers, but both achieve an error of less than 2%. Here, more than one RPS classifier is required in the ensemble, with the best performance achieved for 9 base classifiers. This is also visible in the plots in Figure 1: The accuracy histogram shows that many RPS classifiers achieve good accuracies of 80% or more, while the distribution of the Q statistic indicates that there are many prototype sets that predict the samples differently. Hence, combining several good prototype sets that complement each other well can increase the performance. The heatmap indicates that the last three ALL samples are unsuitable as prototypes, while some other samples (e.g. AML_12, AML_13, AML_18, ALL_16, ALL_19, ALL_26) achieve good accuracies in almost any combination.

On the Notterman data set, the performance of all classifiers is similar, with the best error of 2.8% achieved by many configurations. The distribution of Yule's Q in Figure 1 reveals a possible reason: Most of the prototype classifiers seem to predict the same labels, while a small fraction of prototype sets behaves entirely

different. Thus, ensembles comprising prototype classifiers of both categories will be able to improve the accuracy compared to single prototype classifiers, but the predictions of such ensembles will all be similar.

On the Pomeroy data set, the classification error is mostly similar for all three classifiers. The best accuracy is achieved by a configuration of the SVM with RBF kernel, followed by a configuration of 5 prototype sets of eRPS. The data set shows a broad variety of possible prototype sets whose classifications partly overlap, but are different for other samples (see the Q statistic histogram in Figure 2). Four samples in class CMD seem to be unsuitable prototypes (see heatmap), which indicates that this class is possibly incoherent.

On the Shipp data set, the SVM shows the best performance with an error of 3.8%. Some configurations of eRPS still achieve a very low error of 5.7%. Here, many basic prototype classifiers yield the same performance, but few achieve an accuracy of more than 90% (see Figure 2). Many of the classifiers predict similar labels. Interestingly, the heatmap shows that sample FL_6 yields a high accuracy in combination with any data point from the DLBCL class, which means that it is an excellent representative for his class.

On the West data set, eRPS is clearly outperformed by 1-NN and the SVM. Figure 2 shows that most prototype set classifiers achieve accuracies of 0.6 or less, such that this data set is probably unsuitable for prototype-based classification. At the same time, many of the base classifiers behave similarly, i.e. they misclassify the same samples. As a consequence, ensembles cannot benefit from a diverse set of base learners.

4 Discussion and Conclusion

The high dimensionality of current biomolecular data sets often makes an intuitive understanding impossible. Data mining approaches can provide decision support for such data. However, such models are usually not designed for easy interpretation.

We describe very simple base classifiers that predict the labels of unseen data points according to a set of single prototypes. Due to the simple nature and the data dependency of the Representative Prototype Set classifier, it is possible to describe data sets systematically by enumerating all possible classifiers. Furthermore, ensembles of such basic classifiers yield a prediction accuracy similar to state-of-the-art approaches. We have shown how the two key components of this paper, data set analysis and ensemble classification, complement each other well and can give insights into the data structure. In principle, the proposed methods work for any other type of concept class that allows for enumerating all classifiers.

As for any other classification approach, there is "no free lunch" [21]: the RPS approach is particularly suitable for certain data distributions, but inappropriate for others. The data set analysis also provides a way of judging the ensemble classifier's suitability for a specific type of data. For example, both the Bittner data set – where eRPS performs excellently – and the West data set – where

eRPS is inferior to other approaches – show characteristic profiles in the data set analysis.

Future work will include different distance and correlation measures as well as other ways of aggregating the votes of the base classifiers in the ensembles. Furthermore, the selection of ensembles could be modified in such a way that the correct and incorrect predictions of the base classifiers on the training set complement each other.

Acknowledgement. This work is supported by the Graduate School of Mathematical Analysis of Evolution, Information and Complexity at the University of Ulm (CM, HAK) and by the German federal ministry of education and research (BMBF) within the framework of the program of medical genome research (PaCa-Net; project ID PKB-01GS08) and GerontoSys (Forschungskern SySTaR).

References

1. Fix, E., Hodges, J.: Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties. Technical Report Project 21-49-004, Report Number 4, USAF School of Aviation Medicine, Randolf Field, Texas (1951)
2. Kohonen, T.: Learning vector quantization. Neural Networks 1, 303 (1988)
3. Kohonen, T.: Learning vector quantization. In: Arbib, M. (ed.) The Handbook of Brain Theory and Neural Networks, pp. 537–540. MIT Press, Cambridge (1995)
4. Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99(10), 6567–6572 (2002)
5. Kuncheva, L., Bezdek, J.: Nearest prototype classification: Clustering, genetic algorithms or random search? IEEE Transactions on Systems, Man, and Cybernetics C28(1), 160–164 (1998)
6. Hart, P.E.: The condensed nearest neighbor rule. IEEE Transactions on Information Theory 14, 515–516 (1968)
7. Kuncheva, L.: Fitness functions in editing k-nn reference set by genetic algorithms. Pattern Recognition 30(6), 1041–1049 (1997)
8. Gil-Pita, R., Yao, X.: Using a Genetic Algorithm for Editing k-Nearest Neighbor Classifiers. In: Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds.) IDEAL 2007. LNCS, vol. 4881, pp. 1141–1150. Springer, Heidelberg (2007)
9. Brighton, H., Mellish, C.: Advances in instance selection for instance-based learning algorithms. Data Mining and Knowledge Discovery 6(2), 153–172 (2002)
10. Dasarathy, B.: Nearest neighbor (NN) norms: NN pattern classification techniques. IEEE Computer Society Press (1991)
11. Littlestone, N., Warmuth, M.: Relating data compression and learnability (1986) (unpublished manuscript)
12. Langford, J.: Tutorial on practical prediction theory for classification. Journal of Machine Learning Research 6, 273–306 (2005)
13. Yule, G.: On the association of attributes in statistics: With illustrations from the material of the childhood society. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 194, 257–319 (1900)

14. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., Sondak, V.: Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406(6795), 536–540 (2000)
15. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286(5439), 531–537 (1999)
16. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97(457), 77–87 (2002)
17. Notterman, D., Alon, U., Sierk, A.J., Levine, A.: Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Research 61(7), 3124–3130 (2001)
18. Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E., Golub, T.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415(6870), 436–442 (2002)
19. Shipp, M., Ross, K., Tamayo, P., Weng, A., Kutok, J., Aguiar, R., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G., Ray, T., Koval, M., Last, K., Norton, A., Lister, T., Mesirov, J., Neuberg, D., Lander, E., Aster, J., Golub, T.: Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature Medicine 8(1), 68–74 (2002)
20. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.J., Marks, J.R., Nevins, J.R.: Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS 98(20), 11462–11467 (2001)
21. Wolpert, D.: The lack of a priori distinctions between learning algorithms. Neural Computation 8(7), 1341–1390 (1996)