BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 16 2007, pages 2147–2154
doi:10.1093/bioinformatics/btm312

Data and text mining

Visualization-based cancer microarray data classification analysis

Minca Mramor[1], Gregor Leban[1], Janez Demšar[1] and Blaž Zupan[1,2,*]
[1]Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia and [2]Department of Molecular and Human Genetics, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA

Received on February 12, 2007; revised on May 17, 2007; accepted on June 5, 2007
Advance Access publication June 22, 2007
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Methods for analyzing cancer microarray data often face two distinct challenges: the models they infer need to perform well when classifying new tissue samples while at the same time providing an insight into the patterns and gene interactions hidden in the data. State-of-the-art supervised data mining methods often cover well only one of these aspects, motivating the development of methods where predictive models with a solid classification performance would be easily communicated to the domain expert.

Results: Data visualization may provide for an excellent approach to knowledge discovery and analysis of class-labeled data. We have previously developed an approach called VizRank that can score and rank point-based visualizations according to degree of separation of data instances of different class. We here extend VizRank with techniques to uncover outliers, score features (genes) and perform classification, as well as to demonstrate that the proposed approach is well suited for cancer microarray analysis. Using VizRank and radviz visualization on a set of previously published cancer microarray data sets, we were able to find simple, interpretable data projections that include only a small subset of genes yet do clearly differentiate among different cancer types. We also report that our approach to classification through visualization achieves performance that is comparable to state-of-the-art supervised data mining techniques.

Availability: VizRank and radviz are implemented as part of the Orange data mining suite (http://www.ailab.si/orange).

Contact: blaz.zupan@fri.uni-lj.si

Supplementary information: Supplementary data are available from http://www.ailab.si/supp/bi-cancer.

1 INTRODUCTION

Although current methods used to distinguish and classify human malignancies rely on a variety of clinical, molecular and morphological parameters, precise cancer diagnosis remains a challenging task. Existing diagnostic classes are often heterogeneous and include diseases with different clinical courses, therapeutic response and metastatic potential. In just the last few recent years, DNA microarrays have become almost ubiquitous in biological research. Since cancer is a genetic disease, mostly resulting from acquired mutations and epigenetic changes that lead to changes in gene expression, cancer research and diagnosis is one of the most important emerging clinical applications of this technology. Genome-wide gene expression measurements can give an insight into carcinogenesis, oncogenic pathways and gene networks. They can point to new molecular markers that can be widely used in clinical diagnosis, and lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more reliable classification.

Golub *et al.* (1999) were some of the first researchers to show the superior diagnostic performance of gene expression signatures in cancer (acute leukemia) classification compared to the currently used diagnostic method. Many other studies in almost all cancer types followed (Bhattacharjee *et al.*, 2001; Khan *et al.*, 2001; Shipp *et al.*, 2002). At the same time, a wide range of statistical and data mining methods for microarray data analysis evolved (Allison *et al.*, 2006; Asyali *et al.*, 2006; Pham *et al.*, 2006), all faced two significant problems: the high number of variables (gene expression measurements) with respect to the number of data instances (patient tissue samples), and the substantial component of noise present in the data.

These approaches usually cover one or more components of microarray data analysis that include dimensionality reduction through a gene subset selection, the construction of new predictive features and model inference. While, especially in early research, analysts used unsupervised methods like clustering and principal components analysis, the inference of predictive models from cancer microarray data is essentially a supervised (predictive) data mining problem where the class information is used for the model induction (Simon *et al.*, 2003). Popular approaches in this respect include, e.g. support vector machines (SVMs) (Statnikov *et al.*, 2005), artificial neural networks (ANN) (Khan *et al.*, 2001), k-nearest neighbors (k-NN) and weighted voting of informative genes (Golub *et al.*, 1999) [also see (Asyali *et al.*, 2006) for a recent review].

The two most important aspects of predictive data mining are accuracy in predictions and a gain of insight. The inferred model should make an accurate prediction when classifying new data, i.e when considering data that was not used in the model's induction. To support knowledge discovery, models should also provide means to uncover patterns hidden in the data and allow domain experts to understand their interplay in

*To whom correspondence should be addressed.

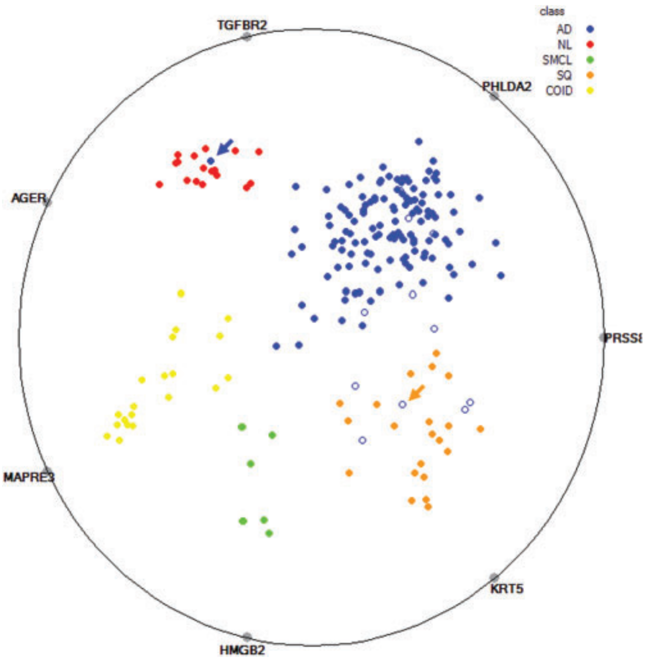


Fig. 1. Radviz data visualization for the lung cancer data set that uses gene expression information on six genes. Points represent tissue samples and are colored with respect to diagnostic class (adenocarcinomas-blue, squamous cell carcinomas-brown, carcinoids-yellow, small cell lung carcinomas-green and normal lung samples-red). The adenocarcinomas with histologically determined squamous features are shown as empty blue circles. Two arrows (red and brown) point to cases which are specifically referred to in the text.

the classification. Not all methods cover these two aspects equally well; SVM with non-linear kernels often outperform most other approaches but, due to the high dimensional transformations they use, they cannot convey the classification patterns. In microarray classification, SVM seems superior to classification trees (Statnikov *et al*., 2005) which, on the other hand, have a clear and communicable structure that is often easy to interpret.

Among popular approaches to data analysis, data visualization offers means for graphically exposing interesting patterns. Let us start by demonstrating this with an example. Figure 1 shows the radviz visualization (Hoffman *et al*., 1997) of a lung cancer data set (Bhattacharjee *et al*., 2001) which contains five diagnostic classes and 203 tissue samples represented with points in the plot. In radviz the data features (e.g. genes) are represented as equidistant anchors on the circle. For now, it suffices to say that the position of each point depends on the values of its features (e.g. gene expressions); the higher the expression of a gene, the closer the placement of the data point to the gene's anchor. Importantly, using the information on only seven genes the visualization in Figure 1 almost perfectly separates the lung cancer diagnostic classes. Normal lung cells (NL, red points), for instance, are characterized by a higher expression of the genes AGER and TGFBR2 compared to the expression of the other five genes shown. The only problematic, i.e overlapping classes are the adenocarcinomas (AD, blue points) and squamous cell carcinomas (SQ, brown points). Reviewing the supplementary information on adenocarcinoma tissue samples, we noticed that some of them have been histologically diagnosed as adenocarcinomas with squamous features (ADSQ, blue circles). These data instances are shown as empty blue circles in Figure 1. As one can observe, most of them lie in the overlapping region. Moreover, the three histologically diagnosed ADSQ samples that do not lie in this region have also been clustered in the 'not squamous carcinoma' or 'weak squamous carcinoma' clusters in the original research of this data set (Bhattacharjee *et al*., 2001). The visualization also includes one clear outlier (adenocarcinoma within the normal lung cells cluster, e.g. the blue point among the red points marked with a blue arrow); interestingly, in the original publication, this sample (AD363) and two other adenocarcinomas (AD4 and AD262) were clustered together with normal lung samples. In our visualization, AD4 and AD262 are clustered within other adenocarcinomas (the blue group).

It is not trivial to find such a clear, class-separating projection like that from Figure 1. The lung cancer data set includes expression measurements on 12 600 genes and there are myriads of possible radviz projections with small sets of features. The search for projections with a clear class separation thus needs to be automated.

In our previous work, we developed an approach called VizRank (Leban *et al*., 2006) that can score the visualizations according to the degree of class separation and that can work through projection candidates to find those with the highest scores. Here, we apply this method to find interesting projections from cancer gene expression data sets and extend it with approaches for feature ranking, outlier detection and classification. Our working hypothesis was that the best-ranked radviz visualizations using the expressions from a small subset of genes can well separate the diagnostic classes in cancer microarray data sets, thus offering a simple means to gain an insight, identify any outliers and classify new cases. The expected benefit of such visualizations is the simplicity of the model, which can easily be communicated to domain experts.

Data visualization played an important role already in early reports in cancer microarray studies. For instance, (Khan, *et al*., 2001) summarized their analysis results in a planar visualization that shows a clear separation of diagnostic cases. Differently to the approach proposed in this article, information in their visualization could not be traced back to the original genes as the plot was obtained by multi-dimensional scaling and used features crafted in several data preprocessing steps (feature selection through neural network learning and feature construction by principal component analysis). Their visualization was therefore not a result of an explicit search. McCarthy *et al*. (2004) were the first to show how radviz can be applied to the analysis of class-labeled data sets from biomedicine. They focused on feature subset selection to reduce the number of genes in visualization and feature grouping (clustering anchors of correlated features), and showed that in such an arrangement the visualizations can provide for a clear separation of instances of different class.

Data visualization may also be performed by dimension reduction, finding (linear) combinations of subsets of genes that would best predict the tumor type. Partial least squares (Boulesteix and Strimmer, 2007), for instance, can find sets of uncorrelated components, where those that best explain the between-class variance are used in the data visualization. The technique was successfully applied to microarray data analysis (e.g. Boulesteix, 2004), and compared favorably (Dai *et al.*, 2006) to other similar dimension-reduction approaches, such as sliced inverse and sliced average variance regression (Bura and Pfeiffer, 2003). The major difference between these approaches and the method proposed in this article is that instead of a single visualization VizRank proposes a ranked set of visualizations, which directly use the (minimal set of) original features instead of their computed combinations.

Perhaps closest to our idea of projection search that explicitly takes into account class information is the work of Grate (2005), who exhaustively scanned among all gene pairs and triplets to find those that defined a scatterplot which included a linear-hyperplane that perfectly discriminated instances of a different class in two-class problem data sets.

We continue with the description of a method for projection scoring and ranking which we then extend with approaches for visualization-based classification, outlier detection and feature ranking. We then evaluate different aspects of the approach on six different cancer microarray data sets. We include the best-rated projections in the article and show that, despite their simplicity, they provide a clear separation of diagnostic classes, expose biologically relevant genes, reveal relations between classes, uncover outliers and provide grounds for a reliable prediction of new cases. An experimental evaluation on 18 other data sets is included in the Supplementary Material.

2 METHODS AND IMPLEMENTATION

2.1 Radviz visualization

Radviz (Hoffman *et al.*, 1997) is a non-linear visualization method which presents visualized features as anchor points equally spaced around the perimeter of a unit circle. Data instances are shown as points inside the circle, with their positions determined by a metaphor from physics [see (Brunsdon *et al.*, 1998) for a formal definition]: each point is held in place with springs that are attached at the other end to the feature anchors. The stiffness of each spring is proportional to the value of the corresponding feature and the point ends up at the position where the spring forces are in equilibrium. Prior to visualization, feature values are scaled to lie between 0 and 1. Data instances that are close to a set of feature anchors have higher values for these features than for the others. For example, data points representing normal lung samples in Figure 1 have higher expression of genes AGER and TGFBR2 as compared to other genes in this visualization. Samples with similar values of all features appear close to the center. Finding the right placement of feature anchors is essential since, for instance, a pair of correlated features anchored at different sides of the unit circle even out each other, in this way concealing their potential importance.

2.2 VizRank visual projection scoring and ranking

VizRank (Leban *et al.*, 2006) is a method for ordering visual projections of class-labeled data by their potential interestingness. In this way, the data analyst does not have to randomly search through myriads of possible projections in order to find those with clear classification patterns, but can instead focus only on a usually small subset of most promising visualizations that are likely to provide the best insight into the data.

VizRank defines the interestingness of the projection by estimating how well data instances of the same class are clustered together and separated from instances from other classes. For example, visualizations from Figure 2a and b show the same data set, but in Figure 2a the class separation is better and, as expected, its VizRank score is higher than that of the visualization from Figure 2b (99.7 > 59.6%).

Radviz projections are planar so every point is described with the horizontal (x) and vertical (y) coordinate. To compute the score of a particular projection, VizRank uses these (x, y) positional features together with their class label, and presents the so-constructed positional data set to a k-NN machine-learning algorithm. The VizRank projection score is then the accuracy of k-NN on a positional data set as estimated through leave-one-out process where each data is classified using the k-NN classifier obtained from all other data points.

The nearest neighbors classification method uses the Euclidean metric to find the closest data instances. We set parameter k to $k = N/c$, where N is the number of data instances and c is the number of classes. k-NN is in our implementation used as a probabilistic classifier, where predicted probabilities are estimated as a relative class frequency of kNN. Clearly, classifications that assign a high probability to the correct class are favored; using leave-one-out, accuracy is for this reason computed as the average k-NN probability assigned to the correct class.

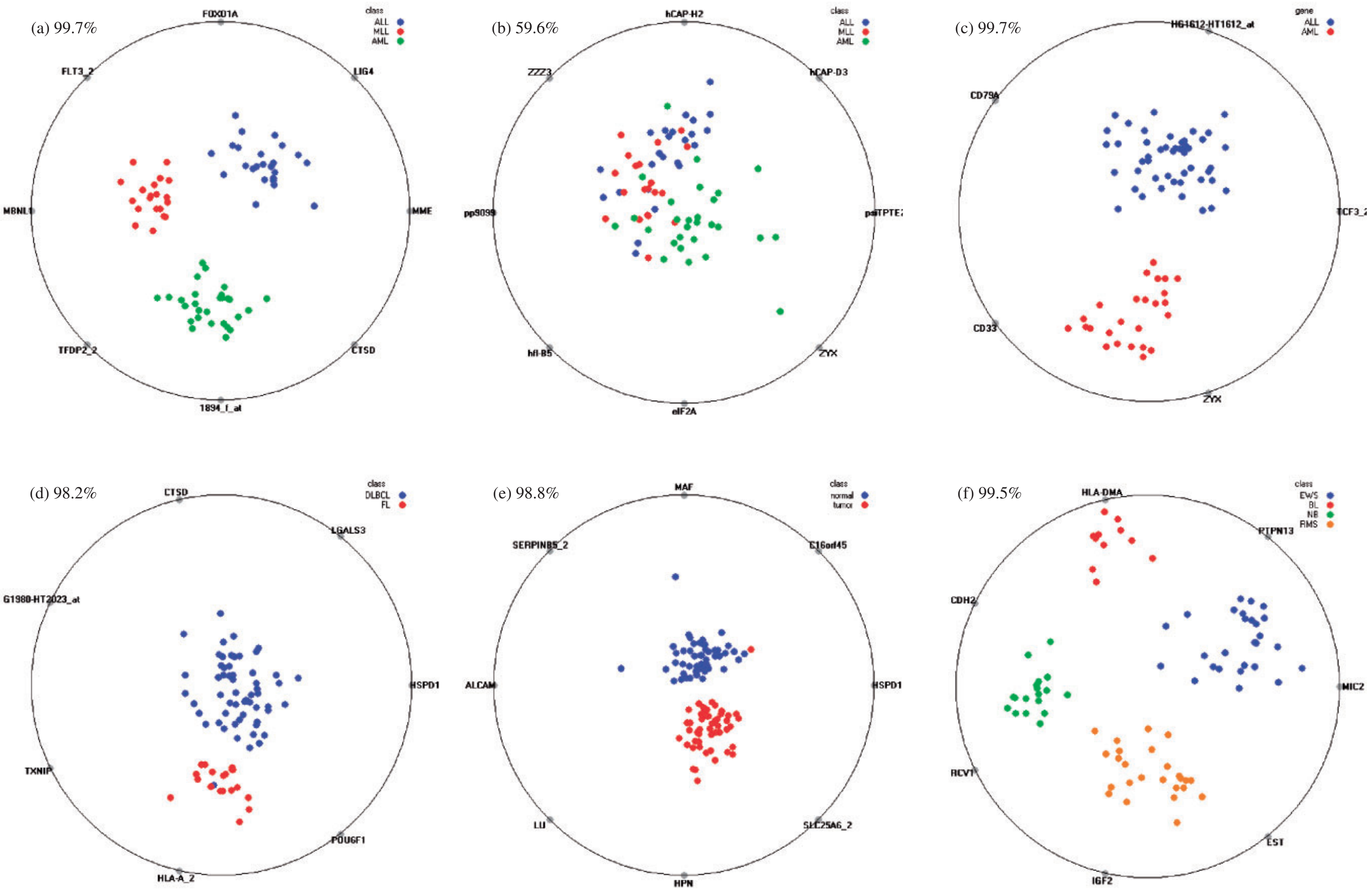
The estimated accuracy of k-NN will be higher for projections with well-separated points with different class labels, making it a good candidate for a measure of projection's interestingness. Other machine-learning methods could be used to assess the quality of projections, but a particular advantage of k-NN using Euclidian distance in the projected plane is that it is close to the human perception of distances and class assignment based on proximal data points (Leban *et al.*, 2006).

2.3 VizRank projection search

Although VizRank can score several thousand projections per minute of computer runtime, this is not fast enough for a reasonably timed search over all possible projections. For instance, for the lung cancer data set from Figure 1 there are $\sim 4 \times 10^{31}$ different radviz projections that include from 3 to 8 genes. VizRank uses a simple search heuristic to remedy this limitation (Leban *et al.*, 2006). Genes are first ranked using signal-to-noise ratio (Golub *et al.*, 1999) and a subset of genes to be considered in visualization is then chosen randomly favoring genes with higher ranks (the probability that a gene is chosen is proportional to the gamma probability density function with the gene's rank as an argument). This ensures that genes with more information about the given classification problem are more likely to be included in a radviz projection to be scored by VizRank. Given a gene subset, VizRank then exhaustively evaluates all possible radviz projections defined by different permutations of gene anchors on a unit circle. Our implementation (see Section 2.7) lets the user set the number of projections to be screened or to interactively observe the evolving list of top-rated projections and stop the search process at will.

2.4 Classification

The projections found by VizRank can be used to classify new samples—cases. The sample's position in the projection is determined by its expression of the genes used in the projection. The sample is then classified to the prevailing class of k-closest samples from the original visualization. The classification algorithm is thus the same as that used in the ranking of the projections. Besides classification to a single class, the procedure can also predict class probability distributions which are estimated from the relative class frequencies of the k-closest instances.



(a) 99.7%

(b) 59.6%

(c) 99.7%

(d) 98.2%

(e) 98.8%

(f) 99.5%

Fig. 2. The best and medium quality radviz projections for mixed lineage leukemia data set (**a** and **b**), and best projections for leukemia (**c**), DLBCL (**d**), prostate cancer (**e**) and SRBCT (**f**) data sets. Associated to each plot is a projection score computed by VizRank.

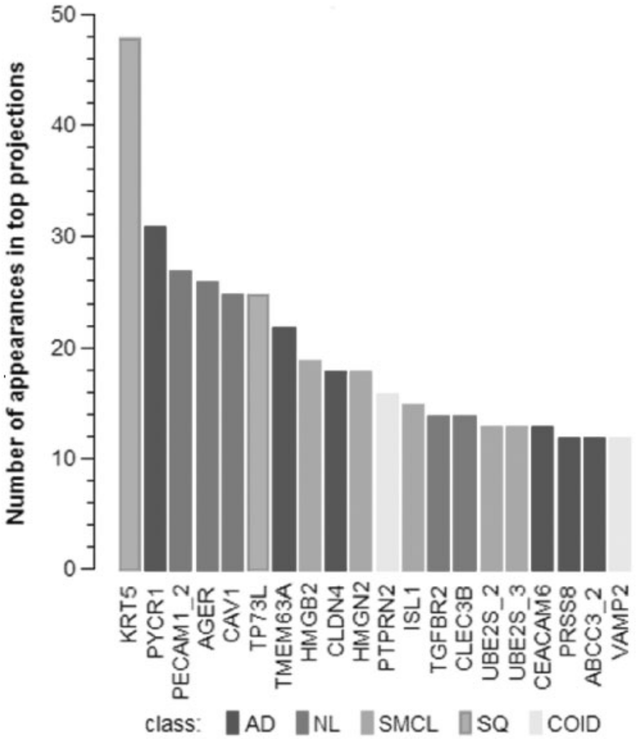


Fig. 3. A histogram of the 20 genes most often used in the 100 best radviz visualizations of the lung cancer data set.

The classification of a new sample can be based on the best-ranked projection as determined from the training data set or it could use a set of best-ranked projections and combine their prediction through some voting schema. The latter could be advantageous as different projections use different features and thus give rise to different visual models. In general, classification ensembles are known to have advantages in predictive accuracy (Witten and Frank, 2005), but can also hinder the simplicity of an original model as it requires considering many models instead of a single optimal one. Another disadvantage of ensembles is that their performance depends on the user-defined number of models in the ensemble. For these reasons, in this report we decided to use only the single best-ranked projection for the classification.

2.5 Gene ranking

Genes that appear in the top-ranked projections are expected to be those that hold the most information for class discrimination. The co-location of genes in individual projections and their interaction in each projection can be explored manually (e.g. through explorative data analysis). To assess the overall utility of genes across a subset of most informative projections, we define a so-called gene utility score which is the number of appearances of the gene in P best-ranked projections. In this article, we use $P = 100$. For example, Figure 3 includes scores of the 20 top-ranked genes from the lung cancer data set (Bhattacharjee *et al.*, 2001), with the colors of the bars indicating which diagnostic class has the highest average expression at that gene. For instance, gene PYCR1 was included in the 31 top-ranked projections and is highly expressed in cases with adenocarcinoma.

The proposed gene ranking is designed for explorative data analysis related to the visualizations suggested in this article. But while graphs such as that in Figure 3 in principle answer questions related to the utility of genes in top-ranked visualizations, the proposed utility score can also be used as a multivariate gene scoring technique (Lai *et al.*, 2006) which, in contrast to univariate analysis, examines the role of the gene taking into account its interaction with other genes in the data set. A comparison of our scoring technique is beyond the scope of this article, where we primarily use gene ranking in the context of explorative data analysis.

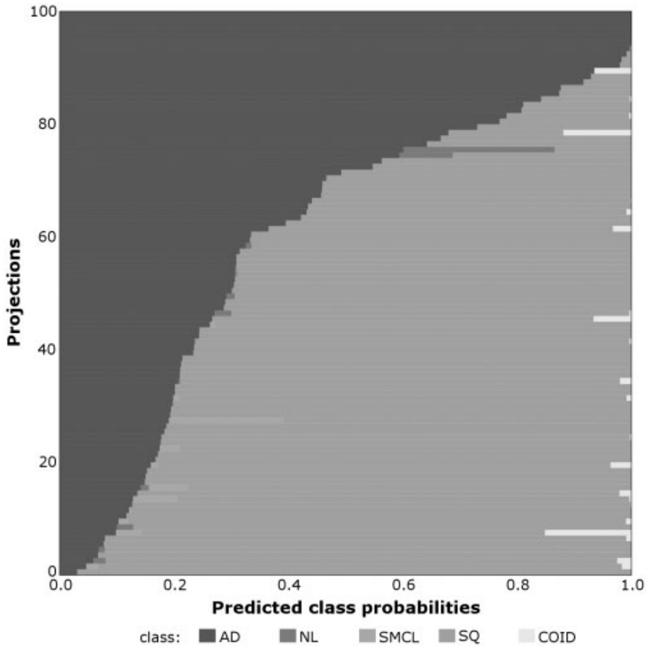


Fig. 4. Class predictions for the selected data instance (blue arrow, Fig. 1) from the lung cancer data set. Rows show the class probability predictions in each of 100 best-ranked projections and are ordered by decreasing probability of the original class label (AD, light gray, bars on the left).

2.6 Outlier detection and analysis

The identification and analysis of the outliers in the best projections can reveal interesting characteristics of the data. These instances might be special cases of a specific disease or perhaps even misdiagnosed samples. Visualizations such as that from Figure 1 can clearly identify the outliers, i.e cases whose class label does not match the prevailing label of the surrounding cases. An outlier is then a data instance whose class is different from the class predicted by a classification algorithm (such as that from Section 2.4) developed from all other instances. While using single data projections the identification of outliers is trivial and can be left to human detection and judgment, we also provide a supplementary technique that uses a range of best visualizations. The technique supports explorative data analysis and examines a single selected case. It reports on its predicted class probabilities using a set of top-ranked visualization. An example of such a report is given in Figure 4; also see Section 3.5 for a discussion.

2.7 Implementation

Vizrank and other methods described in this article are integrated within Orange, an open-source data mining suite featuring Python scripting and a visual programming graphical interface (Curk *et al.*, 2005; Demsar *et al.*, 2004). Details on the installation and user's graphical interface are provided in the Supplementary Material.

Table 1. Cancer microarray gene expression data sets used in the experimental analysis

Data set	Samples (instances)	Genes (features)	Diagnostic classes	Best projection Score
Leukemia	72	7074	2	99.7%
DLBCL	77	7070	2	98.2%
Prostate	102	12 533	2	98.8%
MLL	72	12 533	3	99.8%
SRBCT	83	2308	4	99.5%
Lung cancer	203	12 600	5	93.5%

3 EXPERIMENTAL ANALYSIS AND DISCUSSION

3.1 Data sets

The experimental analysis was performed on a set of cancer gene expression data sets, where each contained five or less diagnostic classes and each class was sufficiently well represented (coverage by at least five data instances was requested, whereas in most data sets the classes include more than 10 instances). While not imposing technical difficulties, the constraint on the number of diagnostic classes was due to radviz, whose visualizations involving too many classes are harder to interpret. Class coverage constraint was due to statistical reasons. To compare the performance of our algorithms to multi-category support vector machines (MC-SVMs), reportedly the most accurate supervised data mining method for this class of problems, we here considered six data sets used from Statnikov *et al*. (2005) that also match the above constraints. These include a data set on leukemia (Golub *et al*., 1999), diffuse large B-cell lymphomas (DLBCL) (Shipp *et al*., 2002), prostate tumor (Singh *et al*., 2002), mixed lineage leukemia (MLL) (Armstrong *et al*., 2002), small round blue cell tumors (SRBCT) (Khan *et al*., 2001) and lung cancer (Bhattacharjee *et al*., 2001) (Table 1). The Supplementary Material includes additional information on these data sets, and the analysis of 18 other publicly available cancer gene expression data sets.

3.2 Top-ranked visualizations

For each data set, VizRank evaluated 100 000 projections that included from three to eight genes and were proposed by search heuristics. The runtime for the largest of the data sets in terms of the number of instances was half an hour on Pentium PC 2.6 GHz with 1 GB of RAM. The particular upper bound for the number of genes was chosen to focus on simpler visualizations that are easier to interpret, and also came as a practical result from the experiments as eight genes were in most cases sufficient to find visualizations with a clear class-separation. The best-ranked projections for all six data sets are shown in Figures 1 and 2. As most of our data sets include more than two diagnostic classes, a direct comparison with the search for two-feature scatterplot projections as proposed by Grate (2005) was impossible. We did, however, instead use VizRank to search through the space of such projections and found that the scatterplots scored from ~1% (leukemia) to 12% (SRBCT) lower than the radviz projections reported above.

Examining the best-ranked projections in Figures 1 and 2 supports our principal finding that VizRank can identify simple planar visualizations of microarray cancer data sets which show a clear separation of diagnostic classes.

3.3 Comments on the biological relevance of genes in top-ranked projections

We further examined the biological relevance of the genes appearing in the best visual projections. For example, in the top-ranked projection of the lung cancer data set (Fig. 1) the raised expression of the gene KRT5 with respect to other genes in the projection subset was found to be an indicator of the squamous carcinoma class. Lung squamous cell carcinoma is a malignant tumor that arises from the bronchial epithelium. The epithelium is lined by squamous cells in which proteins named keratins compose the 8 nm intermediate filaments. The other lung tumors do not arise from epithelial cells and, therefore, have very low levels of expression of KRT5. The usefulness of keratins as tumor markers for sqamous cell carcinomas has already been shown (Fleischhacker *et al*., 1999). Also, in the original report of this data set (Bhattacharjee *et al*., 2001) KRT5 was recognized as one of the most differentially expressed genes, distinguishing the SQ instances from other classes. Moreover, the genes TGFBR2 and AGER that are separating the normal lung instances from tumor samples in Figure 1, were found among the most important markers of normal lung class by Bhattarachjee *et al*.

Similarly, the inclusion of some of the genes in the best projection of the MLL data set (Fig. 2a) is justified by their biological functions. Cases with acute lymphocytic leukemia (ALL, blue points) lie closer to the anchors of the MME and CD24 genes than the cases of the other two diagnostic classes. On the contrary, the anchor point of gene FLT3 most strongly attracts the points representing cases with mixed lineage leukemia (MLL, red point), and anchors for the TFDP2 and ANPEP genes those cases with acute myeloid leukemia (AML, green points). These genes were recognized as being among the most important in distinguishing the three leukemia classes also in the work of Armstrong *et al*. (2001). Additionally, these findings can also be explained biologically. For example, the gene FLT3 is a cancer gene expressed in myeloid and B-lymphoid progenitor cells (Birg *et al*., 1992). According to Armstrong *et al*. (2001), FLT3 is important in distinguishing the unique leukemic entity of MLL and represents an attractive target for rationale drug development. On the other hand, the genes MME and CD24 are lymphoid-specific genes and important cell surface markers in the diagnosis of human ALL (Hardy and Hayakawa, 2001). Similar observations of biological relevance for genes included in top-rated projections were also found for the other data sets included in our study.

3.4 Gene ranking

Figure 3 shows a histogram of the 20 genes most often used in the 100 top-rated radviz visualizations of the lung cancer data set, colored according to the diagnostic class in which they are

Table 2. Bootstrap-estimated classification accuracies of best-scored visualization (VizRank) compared to four standard machine-learning algorithms

	Classification accuracy				
Data set	VizRank	SVM	*k*-NN	Naïve Bayes	Decision trees
Leukemia	96.40%	97.57%	92.72%	84.34%	90.46%
DLBCL	93.03%	97.85%	88.60%	83.76%	85.46%
Prostate	94.00%	93.47%	84.51%	81.10%	85.47%
MLL	95.00%	97.32%	89.65%	75.20%	88.31%
SRBCT	96.39%	99.42%	86.29%	75.31%	87.32%
Lung	92.72%	94.67%	90.35%	75.28%	91.21%
Average rank	1.83	1.17	3.50	5.00	3.50

most highly expressed. According to this graph, KRT5 is the most important gene present in almost half of the top-rated projections with the role of separating squamous cell carcinomas (SQ) from other lung tumors. In the previous section, we have already presented the possible biological background of this result. Following in the order of appearance in the top-rated projections are PYCR3 for adenocarcinomas and PECAM1 for normal lung. Many genes that appear in the top-rated visualizations [e.g. PECAM1, AGER, and CAV1 for normal lung ISL1 for pulmonary carcinoids (COID), etc.) were also recognized as important in separating a specific class in the original work on this data set (Bhattacharjee *et al*., 2001).

Similar to the above, we also found correspondence between the list of genes defining the best ranked projections and earlier reports on their important role in the cancer classification from other data sets in our study. This confirmed our expectations since visualizations with a good class separation are bound to include the most informative genes. Similar visualizations for other data sets that we have studied are included in the supplementary Material.

3.5 Outlier analysis

We present the utility of outlier analysis through an example. Suppose we are interested in the instance of adenocarcinoma marked with a brown arrow in Figure 1 (blue circle) that was in this visualization placed within a group of squamous cell carcinomas (brown points). Figure 4 shows how this instance is classified in the 100 top-ranked radviz projections. In, roughly half of these this tumor sample is classified as a squamous cell carcinoma. Analyzing the supplemental information of the original publication on this data set (Bhattacharjee *et al*., 2001), we noticed that this adenocarcinoma (and the other seven shown as empty blue circles in Fig. 1) have been histologically diagnosed as adenocarcinomas with squamous features. The instance marked with a brown arrow in Figure 1 is therefore a mixture of AD and SQ classes, as can also be seen from the visualization shown in Figure 4.

3.6 Predictive accuracy

To estimate the predictive accuracy, we followed the suggestion of Braga-Neto and Dougherty (2004) and used the bootstrap resampling technique, where we repeated the sampling 100 times. The classification performance was obtained using the 0.632 bootstrap estimator where the classification accuracy (CA, proportion of correctly classified instances) is computed based on the error obtained on the training data set and the average test error of all bootstrap repetitions. To compare VizRank's accuracy to that of the state-of-the-art machine-learning approaches, the study included SVMs with a linear kernel, a *k*-NN learner [*k* equal to square root of number of data instances in the learning set (Dasarathy, 1991)], a naive Bayesian classifier and a decision tree [C4.5 (Quinlan, 1993) with default settings]. CAs of all the algorithms for the six data sets are reported in Table 2. The methods are ranked based on their performance on each data set (ranks from 1 to 5, where 1 is used for the best-scoring and 5 for the worst-scoring method). Table 2 (last row) reports on the methods' average rank across all the data sets. The SDs of the CAs are reported on the supplementary Material, which also includes the scoring of learners using an area under ROC curve (AUC), with very similar rankings as those based on CA.

The approach proposed in this article performs comparably to state-of-the-art classification algorithms. In our study, this classification approach is second to SVM and performed better than the other algorithms studied. The performance of these could be improved through preprocessing by feature-subset selection (Lai *et al*., 2006) and ensemble approaches. Such a comparison exceeds the topic of our report as our aim is not to replace the already established techniques that were proved to yield a high classification accuracy, such as SVM. Instead, the comparable performance of our technique speaks of the approach's stability; VizRank seems able to choose a relevant subset of genes without overfitting the data. The principle gain, especially in comparison with SVM, are the much simpler, interpretable graphical models based on expressions of only eight or less features without a significant loss in performance. Finally, VizRank often identifies the best projections within minutes of runtime, which is significantly less than the hours of required SVM runtime reported in a recent study (Statnikov *et al*., 2005).

4 CONCLUSION

The main contribution of our work is a method for analyzing gene expression data that as a result provides a reliable classification model and gives a valuable insight into the data in the form of informative visualizations. The proposed method of projection scoring and ranking can find simple visualizations of cancer gene expression data sets that use a very small subset of genes yet provide a clear visual differentiation between cancer types. Especially due to its potential role in explorative data analysis, short runtimes and interactive interface, we propose that data visualization supported with efficient projection search techniques should complement other established techniques in cancer microarray analysis and become part of the standard analysis toolbox.

Conflict of Interest: none declared.

ACKNOWLEDGEMENTS

This work was supported in part by Slovenian Research Agency's grants J2-9699 and P2-0209 and by a P01 HD39691 grant from the National Institute of Child Health and Human Development.

REFERENCES

Allison,D.B. *et al*. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet*., 55–65.

Armstrong,S.A. *et al*. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet*., **30**, 41–47.

Asyali,M.H. *et al*. (2006) Gene expression profile classification: a review. *Curr. Bioinformatics*, **1**, 55–73.

Bhattacharjee,A. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Birg,F. *et al*. (1992) Expression of the FMS/KIT-like gene FLT3 in human acute leukemias of the myeloid and lymphoid lineages. *Blood*, **80**, 2584–2593.

Boulesteix,A.L. (2004) PLS dimension reduction for classification with microarray data. *Stat. Appl. Genet. Mol. Biol*., **3**, , Article 33.

Boulesteix,A.L. and Strimmer,K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinformatics*, **8**, 32–44.

Braga-Neto,U.M. and Dougherty,E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.

Brunsdon,C. *et al*. (1998) An investigation of methods for visualising highly multivariate datasets. In: Unwin,D. and Fisher,P. (eds) *Case Studies of Visualization in the Social Sciences*. pp. 55–80.

Bura,E. and Pfeiffer,R.M. (2003) Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, **19**, 1252–1258.

Curk,T. *et al*. (2005) Microarray data mining with visual programming. *Bioinformatics*, **21**, 396–398.

Dai,J.J. *et al*. (2006) Dimension reduction for classification with gene expression microarray data. *Stat. Appl. Genet. Mol. Biol*., **5**, Article 6.

Dasarathy,B.W. (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA.

Demsar,J. *et al*. (2004) Orange: from experimental machine learning to interactive data mining. White Paper, Faculty of Computer and Information Science, University of Ljubljana.

Fleischhacker,M. *et al*. (1999) Molecular genetic characteristics of lung cancer–useful as real' tumor markers? *Lung Cancer*, **25**, 7–24.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Grate,L.R. (2005) Many accurate small-discriminatory feature subsets exist in microarray transcript data: biomarker discovery. *BMC Bioinformatics*, **6**, 97.

Hardy,R.R. and Hayakawa,K. (2001) B cell development pathways. *Annu. Rev. Immunol*., **19**, 595–621.

Hoffman,P.E. *et al*. (1997) DNA visual and analytic data mining. *In the Proceedings of the IEEE Visualization*. Phoenix, AZ, pp. 437–441.

Khan,J. *et al*. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med*., **7**, 673–679.

Lai,C. *et al*. (2006) A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, **7**, 235.

Leban,G. *et al*. (2006) VizRank: data visualization guided by machine learning. *Data Mining Knowl. Discov*., **13**, 119–136.

McCarthy,J.F. *et al*. (2004) Application of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. *Ann. N.Y. Acad. Sci*., **1020**, 239–262.

Pham,T.D. *et al*. (2006) Analysis of microarray gene expression data. *Curr. bioinformatics*, **1**, 37–53.

Quinlan,J.R. (1993) C4.5: *Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California.

Shipp,M.A. *et al*. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med*., **8**, 68–74.

Simon,R. *et al*. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*., **95**, 14–18.

Singh,D. *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Statnikov,A. *et al*. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631–643.

Witten,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. 2nd edn. Morgan Kaufmann, San Francisco, California.