Simple and Effective Visual Models for Gene Expression Cancer Diagnostics

Gregor Leban, Minca Mramor, Ivan Bratko, Blaz Zupan
Faculty of Computer and Information Science
University of Ljubljana
Tržaška 25, Ljubljana, Slovenia
(gregor.leban, minca.mramor, ivan.bratko, blaz.zupan)@fri.uni-lj.si

ABSTRACT

In the paper we show that diagnostic classes in cancer gene expression data sets, which most often include thousands of features (genes), may be effectively separated with simple two-dimensional plots such as scatterplot and radviz graph. The principal innovation proposed in the paper is a method called VizRank, which is able to score and identify the best among possibly millions of candidate projections for visualizations. Compared to recently much applied techniques in the field of cancer genomics that include neural networks, support vector machines and various ensemble-based approaches, VizRank is fast and finds visualization models that can be easily examined and interpreted by domain experts. Our experiments on a number of gene expression data sets show that VizRank was always able to find data visualizations with a small number of (two to seven) genes and excellent class separation. In addition to providing grounds for gene expression cancer diagnosis, VizRank and its visualizations also identify small sets of relevant genes, uncover interesting gene interactions and point to outliers and potential misclassifications in cancer data sets.

Keywords

Gene Expression Analysis, Cancer Diagnosis, Machine Learning, Data Mining, Data Visualization

1. INTRODUCTION

DNA microarrays simultaneously measure the expression of thousand of genes in a biological sample to determine which genes are differently expressed in various cells and tissues. Gene expression measurement may provide for a powerful tool in uncovering the genetic mechanisms causing the loss of cell cycle control and the consecutive development of cancer. Several recent studies of different cancer types [1, 2, 10, 13, 17, 18, 19, 20] have demonstrated the superior performance of gene expression profiles for cancer classification when compared to standard morphological criteria. The ultimate goal of this approach is improvement and individualization of treatment and detection of pathogenomic biological markers of different tumors for earlier diagnosis and prognosis.

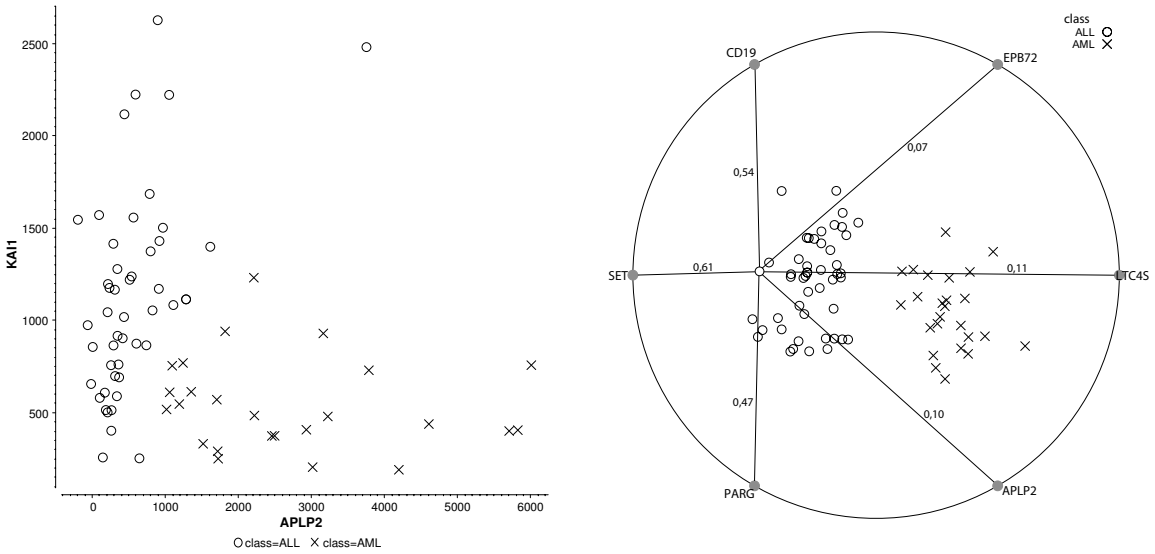
Gene expression data analysis is characterized by extremely high data dimensionality due to thousands of gene expression values measured for each sample on an array. At the same time, the number of samples (patients) is far smaller. The analysis of this peculiar and noisy data has been challenged by a number of different approaches in bioinformatics that include feature subset selection to focus only on genes that bear most information on the cancer type, unsupervised and supervised machine learning methods, and various visualization techniques.

In unsupervised analysis of cancer gene expression data, a standard procedure is to select a set of the most informative genes and then use them in principal component analysis (PCA) [13]. It has been shown that, under appropriate gene selection, visualization of data using the first two principal components may reveal separated clusters, each comprising the data of a prevailing diagnostic class. While such an approach can demonstrate that diagnostic classes may be separated by gene expression data, the clinical and genomic interpretation of the results is hard as each component may combine expression of tens or hundreds of genes.

To use the diagnostic class information in the learning process, a number of recent studies supervised machine learning techniques such as artificial neural networks [13], k-nearest neighbors with weighted voting of informative genes [10] and support vector machines (SVM) [9, 21]. While it was recently shown that SVM are a method to uncover models with most reliable classifications [21], such classification models often combine relatively weak contributions of up to thousands of genes and are therefore hard to understand and interpret by the domain experts.

The research reported in this paper aimed to investigate how "hard" are gene expression cancer data sets in terms of finding good and simple classifiers. Our working hypothesis was that visualizations that include only a few genes and use the untransformed gene expression data can provide for a clear separation of diagnostic classes. Differently from the related work, we address the problem of finding good visualizations directly with an algorithm that uses a powerful heuristic to search through a space of possible data projec-

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.



(a) scatterplot (b) radviz

Figure 1: An example of scatterplot (a) and radviz (b) visualization of the leukemia data set (see Section 4.1 for details) that both exhibit good separation of diagnostic classes. For radviz and a selected data instance we also show the corresponding attachments to gene anchors ("springs") and related values of normalized gene expressions.

tions and a machine learning approach that evaluates and assigns a score for each projection. The visualization scoring and search algorithm called VizRank and exploration of its utility in the analysis of cancer expression data sets are two principal contributions of this paper. Using two planar geometric visualizations, namely scatterplot when visualizing two genes and radviz for visualizations with three genes or more, we show that we are always able to find simple visualizations which include only a handful of genes and provide for a clear split of diagnostic classes.

Two examples of such graphs visualizing gene expression data on leukemia are shown in Figure 1. Besides being simple, allowing to depict both individual role of visualized genes and their interactive effects, these visualizations also identify relevant genes for cancer prediction and can provide grounds for identification of potential outliers.

In the paper, we first introduce the two geometric visualization methods used, scatterplot and radviz. Components of VizRank, which include search algorithm, a method for scoring of visualizations, and a search heuristic are presented next. We then use VizRank on a set of eight cancer gene expression data sets, where we report on the role of the heuristic, experimentally assess how likely it is to find a good and simple visualization, and present the best visualization for each of the data sets. We show that these simple visualizations provide clear separation of diagnostic classes and include biologically relevant genes.

2. TWO-DIMENSIONAL GEOMETRIC VISUALIZATION METHODS

Many techniques exist that can be used to visualize multidimensional data. VizRank method proposed in this paper can be applied to any geometric visualization method, that is, any method where data instances are visualized as points in a two-dimensional space and the values of visualized features only influence the position of the data point in the graph, and not its size, shape or color. In this paper we present results using two such methods – a scatterplot, a method for visualizing data using two features, and radviz, which can concurrently visualize a larger number of features. Since scatterplot is perhaps the simplest and well-known multi-dimensional visualization method, we here in detail describe only radviz.

In radviz [12], m visualized features are represented as anchor points equally spaced around the perimeter of a unit circle (see Figure 1.b for example). Data instances are shown as points inside the circle, and the position of the instance within a graph can be explained using a physical analogy with multiple springs. There are m springs attached to the instance point, one for each of the feature anchors. The stiffness of each spring in terms of Hooke's law is determined by the corresponding feature value – the greater the value, the greater the corresponding stiffness. The point representing the data instance is then placed at the position where the sum of all spring forces equals 0 (see [4] for mathematical details). To bring values of features on the same scale and to allow for a simpler interpretation of radviz graphs, features are standardized to the interval between 0 and 1.

Figure 1.b shows an example of a radviz graph; the graph shows "springs" for a selected data instance, this being close to the anchor with gene SET bearing the largest standardized expression (0.61) and somewhere in-between anchors for genes CD19 and PARG with a similar expression (0.54 vs. 0.47, respectively). The "attractive force" of other three genes in this particular graph is smaller.

In radviz the points that have approximately equal values of features that lie on the opposite sides of the circle will lie close to the center of the circle. On the other hand, the points with one of the features having a much larger value

than others will lie close to the anchor of that feature. The particular visualization using a selected set of features will therefore largely depend on the position/order of feature anchors. Placing two highly correlated features that are good at discriminating between classes on the opposite side of the circle will make them useless in the visualization, since their joint effect will be cancelled out. On the other hand, they might generate a projection with well separated classes if their anchors are placed adjacently. The "correct" placement of feature anchors was for instance crucial for a nice separation of classes in Figure 1.b, where the three anchors (genes) on the left side of the circle attract data points from the ALL class and anchors (genes) on the right side of the circle attract points with AML class value.

3. VIZRANK: FINDING INFORMATIVE DATA VISUALIZATIONS

The two visualizations in Figure 1 can be considered interesting because the data instances that belong to different class are well-separated. They show data projections for which we can visually infer rules for discriminating between different class values. When features are in abundance, the main question is how to find the most informative data projections as the manual search through the projection space is not feasible.

3.1 VizRank Algorithm

We have developed a method called VizRank to enable an automatic, algorithm-based search for the most informative data visualizations. For a given data set and a visualization method, VizRank returns a ranked list of most informative data projections along with the numerical assessments of their "interestingness". In this way the data analyst is relieved of the unguided search through numerous possible projections, and can focus only on the top-rated visualizations that can provide the best insight into the data.

VizRank evaluates each by generating a corresponding visualization using a selected visualization method and computes its score based on how well the visualization separates the instances of different class. In other words, the visualization score is related to how likely it is for an analyst to spot a visual pattern in the projection that reveals some regularity in the domain.

VizRank solves the problem of projection assessment by applying a machine learning method on the graphically represented data and estimates the accuracy of the induced classifier. Input to machine learning are x and y coordinates of the points representing data instances in the assessed graph and their corresponding class labels. The estimated predictive accuracy of the classifier on this data set is then used as a score for the particular projection. Notice that if the data instances in the projection are clearly separated, the predictive accuracy for some machine learning algorithms is expected to be high. High visualization scores computed in this way are therefore indicators of usefulness of a visualization.

We use k-nearest neighbor (k-NN) with Euclidian distance metrics as a supervised machine learning method for visualization scoring. When applied to our 2-dimensional visualizations, Euclidian distance well matches the intuitive distance used by human observer when viewing the graph [5]. Using this distance metrics, k-NN predicts class value of an instance by observing the class distribution of its k nearest instances in the evaluated projection. If the prediction matches the true observation, this would mean that an instance is surrounded by instances with the matching prevailing class. Since k-NN does not impose any constraints on decision boundaries that separate instances from different classes, we believe this method may be the most suitable for our visualization scoring and may well match the interestingness as seen from the perspective of human analysts.

In our experiments, we use leave-one-out evaluation schema to obtain the predictive accuracy of a k-NN classifier. For a parameter k, we have followed the recommendation by Dasarathy [6] which, to classify each instance, uses a neighborhood of $k = \sqrt{N}$ instances, where N is the number of instances in the data set. To make the method less sensitive to the choice of k we also use weighted voting, where contribution of each instance in the neighborhood decreases with the distance, so that close neighbors have greater influence on the prediction than those farther away from the instance being classified.

3.2 Measure for Prediction Accuracy

There are several measures (scoring functions) we could use to evaluate the performance of k-NN classifier. One of the most often used measures is classification accuracy that is defined as the proportion of correctly classified instances. For the purpose of projection evaluation we found that classification accuracy is too crisp as it considers only if the example was correctly or incorrectly classified and ignores the prediction uncertainty. As an example, Figure 2 shows two radviz projections of the MLL data set that both have 100% classification accuracy. However, the visualization on the left has a much nicer separation of the classes and should therefore be favored over the projection on the right.

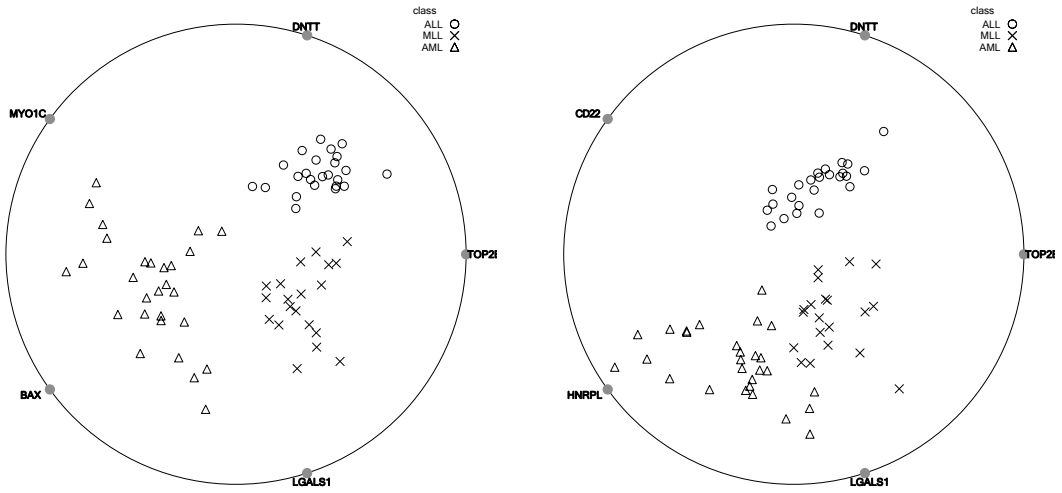
When prediction is based on the weighted proportion of the neighboring instances belonging to each class, k-NN can be regarded as a probabilistic classifier. To appropriately consider the predicted class probabilities, we measure the projection interestingness as an average probability $\overline{P}$ that the classifier assigns to the correct class:

$$\overline{P} = E(P_C(\mathbf{y}|\mathbf{x})) = \frac{1}{N}\sum_{i=1}^{N} P_C(y_i|x_i) \quad (1)$$

Here, N is the number of instances in the data set, and $P_C(y_i|x_i)$ is the probability assigned to the correct class value y_i for example x_i by the classifier C. Using this measure we can take into account the prediction uncertainty for examples in Figure 2.b that lie on the boundary between MLL and AML group and lower the projection value accordingly. Average probabilities assigned to the correct classes $\overline{P}$ for these two projections are 99.63% and 97.98%, respectively, favoring the visualization with a clearer class separation.

3.3 Search Heuristic

In the data sets with thousands of features, such as those on cancer gene expressions, number of possible possible data projections is extremely high. Using a data set with m features, there are $m(m-1)/2$ different scatterplot projections. For SRBCT, the smallest data set in terms of number of the features considered in this paper, the number of different scatterplot visualizations is 2,656,508. For radviz



(a) (b)

Figure 2: Two projections of the MLL data set (see Section 4.1 for details). Classification accuracy of k-NN classifier on both projections is 100%, while the average probability of correct classification $\overline{P}$ is 99.63% (a), and 97.98% (b).

method, the number of different projections is even considerably higher, since the method can concurrently visualize a larger number of features and has to consider different placement of these in the graph. Even for a computerized method it is therefore impossible to exhaustively search over all possible projections and search heuristic must be used instead.

The search heuristic we developed for VizRank starts by estimating the predictive quality of each single feature using the ReliefF measure [15]. Other feature scoring function may be used instead, but we choose to use ReliefF since it can detect feature interactions and could possibly assign higher scores to features that could be overlooked using some other, univariate analysis measure. Next, for each projection, a rough estimate of its usefulness is computed as the sum of ReliefF's values for features that are present in the projection. VizRank then assesses projections starting with those with those most promising according to sum-of-ReliefF value. As we show in the next section, such heuristic is successful and allows VizRank to evaluate only a small subset of projections to find those with best class discrimination.

4. EXPERIMENTS AND CASE STUDIES

We have used eight cancer gene expression data sets to evaluate the proposed approach. Experiments on these data sets were not only an academic exercise to assess our algorithms: the data sets come from recent clinical studies for which the problem of finding most informative genes and gene interactions is still open and highly relevant.

4.1 Data sets

In our experimental study we have considered eight publicly available cancer gene expression data sets with two to five distinct diagnostic categories, 40 to 203 samples (patients) and 2308 to 12625 features (gene expressions). The basic information on these is summarized in Table 1.

Three data sets, leukemia [10], diffuse large B-cell lymphoma (DLBCL) [19] and prostate tumor [20] include two diagnostic categories. The leukemia data consists of 73 tissue samples, including 48 with acute lymphoblastic leukemia (ALL) samples and 25 with acute myeloid leukemia (AML), each with 7074 gene expression values. The DLBCL data set includes expressions of 7070 genes for 77 patients, 59 with DLBCL and 19 with follicular lymphoma (FL). The prostate tumor data set includes 12533 genes measured for 52 prostate tumor and 50 normal tissue samples.

The other five data sets analyzed in this work are multicategory. The mixed lineage leukemia (MLL) [1] data set includes 12533 gene expression values for 72 samples obtained from the peripheral blood or bone marrow samples of affected individuals. The ALL samples with a chromosomal translocation involving the mixed lineage gene were diagnosed as MLL, so three different leukemia classes were obtained (AML, ALL and MLL). The small round blue cell tumors (SRBCT) data set [13] consists of four types of tumors in childhood, including Ewing's sarcoma (EWS), rhabdomyosarcoma (RB), neuroblastoma (NB) and Burkitt's lymphoma (BL). It includes 83 samples derived from both tumor biopsy and cell lines and 2308 genes. For the analysis of the brain tumor gene expression data, we used the A1 data set [18] that includes 40 embryonal tumor samples of the central nervous system (10 medulloblastomas (MD), 10 malignant gliomas (MG), 5 rhabdoid tumors (Rh), 6 primitive neuroectodermal tumors (PN) and 4 normal cerebella (Nc)) and 7129 genes. The glioblastoma data set [17] consists of 50 brain tumor samples, including 28 glioblastomas and 22 anaplastic oligodendrogliomas that are additionally classified as classic (CG, CO) or non-classic (NG, NO). The last data set is the lung cancer data set [2] that contains 12600 gene expression values for 203 lung tumor samples (139 adenocarcinomas (AD), 21 squamous cell lung carcinomas (SQ), 20 pulmonary carcinoids (COID), 6 small cell lung cancers (SMLC) and 17 normal lung samples (NL)).

All data sets except the SRBCT were obtained from Affy-

Table 1: Cancer-related gene expression data sets used in our study. Basic statistics of the data sets include the number of examples, diagnostic classes and genes included in a data set, and proportion of examples in the majority diagnostic class. Last two columns show the average probability of correct classification ($\overline{P}$) for the top-ranked scatterplot and radviz projection.

	Number of			Major	Score for top projection	
Data set	Samples	Classes	Genes	class	Scatterplot	Radviz
Leukemia	73	2	7074	52.8%	98.0	100.0
MLL	72	3	12533	38.9%	94.8	99.9
SRBCT	83	4	2308	34.9%	87.6	100.0
Prostate	102	2	12533	51.0%	91.7	97.7
DLBCL	77	2	7070	75.3%	96.8	100.0
Glioblastoma	50	4	12625	30.0%	80.4	94.6
Brain tumor	40	5	7129	25.0%	78.5	92.6
Lung cancer	203	5	12600	68.5%	93.4	96.5

metrix gene chips and are available at http://www.broad.-mit.edu/cancer/. The SRBCT gene expression data set was obtained from cDNA microarrays and is available at http://research.nhgri.nih.gov/microarray/Supplement/.

4.2 Distribution of Visualization Scores

In the study, we let VizRank evaluate a large number of projections and were interested in the distribution of evaluation scores. It turns out that only a relatively small proportion of visualizations bear high interestingness score as assigned by VizRank. For instance, Figure 3 shows that among 5,000 top-ranked scatterplots identified by our heuristic in the MLL data set, there only a few projections scored above 90%. The fast drop in the projection scores is welcomed and indicates that there are only a few projections that are most relevant and are needed to be considered for a detailed inspection by an analyst. We have observed similar projection score distributions on other data sets, also when increasing the total number of evaluated projections.

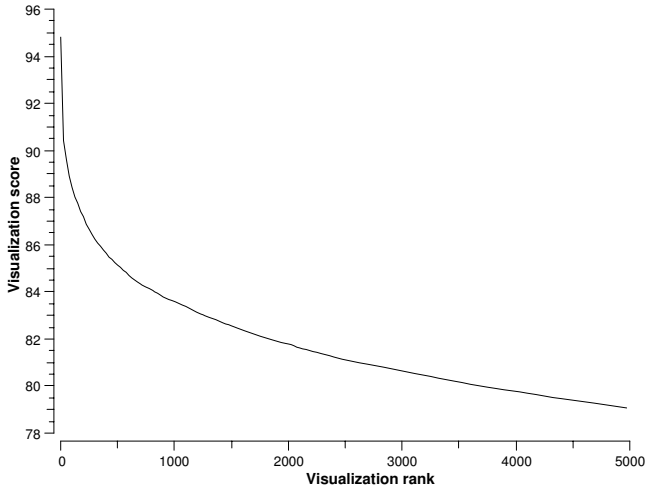


Figure 3: Visualization scores for the best 5,000 scatterplot projections of the MLL data set. Visualizations were ranked based on their score, starting with best-scored visualization.

4.3 Utility of Search Heuristic

Our hypothesis is that the utility of search heuristic would allow VizRank to reduce the number of projections it needs to evaluate, and search over projections with consequently highest projection scores first. The later is important in terms of user's interface, and would allow to present projections with good class separation even within the first few seconds of the search time.

We have compared a heuristic search and a search with a random selection of projections in terms of the best projection found. Figure 4 shows the results on SRBCT data set: the plot shows the score of the best-found projection as a function of a number of projections being evaluated. The value of heuristic is clear, as when compared to a random search allows to find much better projections even if only a few have been considered. For the reasons of brevity the results on other data sets are not shown here, but are qualitatively very similar to those shown in Figure 4.

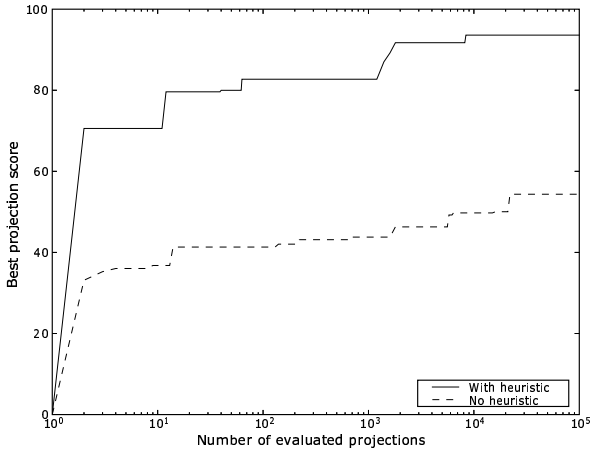


Figure 4: The importance of the heuristic for fast identification of top-ranked projections. The figure shows projection score for the best found projections on the SRBCT data set using radviz method with and without the use of heuristic.

4.4 Best-Scored Visualizations

For each of the cancer gene expression data sets we used VizRank to find top-ranked scatterplot and radviz visualizations. Due to extremely large number of possible projections, for each data set VizRank was constrained to evaluate only 200,000 projections as selected by the search heuristic. The typical performance of VizRank on such data sets using a Pentium 4 PC with 2.4GHz processor is about 30 projections per second, so the typical evaluation time for 200,000 projections was about two hours.

Although the radviz method can in principle visualize an arbitrary number of features, this has a significant influence on interpretation value of the visualization. Also, if the method is able to find visualizations with clear separation of diagnostic classes using only a small number of features, these should be preferred over visualizations with many features. For this reason, we have only investigated radviz projections with 3 to 7 features.

Visualizations with highest VizRank scores for MLL, prostate, DLBCL, glioblastoma, brain tumor, and lung cancer data sets are shown in Figure 6, for leukemia in Figure 1.b, and for SRBCT in Figure 5.a. The best radviz visualizations scored higher than the best scatterplots. The best scored radviz visualizations included anywhere from five (MLL) up to seven features (glioblastoma and brain tumor). Most of the visualizations show a clear separation of instances from different diagnostic classes, with the only exception of brain tumor (two outliers) and lung cancer, where instances of the AD class group together but are placed within the group with prevailing SQ class. Interestingly, the adenocarcinomas in the lung cancer data set are also histologically not a unique class. It was reported by Bhattacharjee *et. al* [2] that seven adenocarcinomas express high levels of squamous-associated genes and also display histological evidence of squamous features. In addition to these seven mixed AD-SQ tumor samples, 12 other adenocarcinomas were suspected to be extrapulmonary metastases, thus adding to histological diversity of the AD class.

4.5 Considerations on Overfitting

One could claim that in the space of a very high number of projections there is always a chance to find a projection with good or even excellent class separation. Even for a random data set, if using enough features, VizRank could then find an excellent projection and would, in this sense, overfit the data.

Our theoretical analysis, though, points to a quite different conclusion. The probability that a random planar geometric visualization will offer a clear separation of the instances with different class is:

$$p = c! \cdot \left(\frac{1}{c}\right)^N$$

where N denotes a number of instances in the data set and c is the number of different class labels. The above formula was derived by computing the probability that in such visualization instances are grouped within c clusters, each containing only instances of the same class.

If, for example, we compute the chance that a random visualization of SRBCT data set offers a clear class separation, we obtain the probability of 2.57×10^{-49} (SRBCT data set has 83 instances and 4 class values). Notice that this probability is low, and even with high number of projections the chances that we will find one with clear separation, if the data would be random, are slim.

To address this issue in an experiment, we have randomly permuted the class values in the SRBCT data set and used VizRank to rank the projections. We evaluated 500,000 most promising radviz projections as identified by our heuristic. The best found projection is shown in Figure 5.b. Notice that the resulting visualization is almost completely uninformative as the classes overlap. We performed a similar experiments on all other data sets and observed similar results, *i.e.*, none of the visualizations found separated the classes well.

Using this results together with the biological interpretation of results in our case studies (see next section), we can conclude that the clear separation of classes in the shown projections could not be attributed to chance but are rather a demonstration of a true regularity in the data.

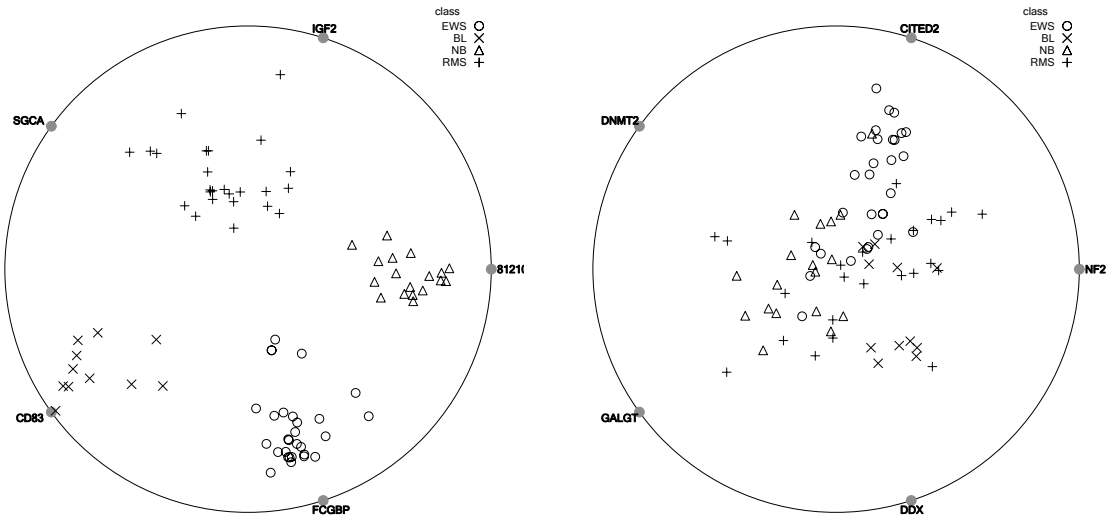
4.6 On Biological Relevance and Gene Selection

We studied the biological relevance of genes appearing in the best visual projections. We assumed that most useful genes in discriminating different tumor types would mostly be markers of different tissue or cell origin and not be necessary related to cancer pathogenesis. However, many of the genes appearing in the best radviz projections are annotated as cancer or cancer-related genes according to the atlas of genetics and cytogenetics in oncology and haematology (www.infobiogen.fr/services/chromcancer/index.html). For example, BAX, DNTT, CD22 and TOP2B genes shown in the two projections of the MLL data set (Figure 2) are cancer related.

On the other hand, for the prostate data set, where we try to differentiate tumor and normal tissue samples based on gene expression profile, one would expect the "marker" genes to be cancer related. We support our hypothesis by ascertaining that all six genes used in the best radviz projection (Figure 6.b) are cancer related according to the cancer gene atlas.

We here present a biological interpretation of the genes used in the best visualizations of the MLL data set. One can observe in Figure 2.b that instances with ALL class label lie closer to the anchor points of the DNTT and CD22 gene than instances either in MLL or AML diagnostic class. This finding is consistent with the work of Armstrong *et al.* [1], in which they report on genes DNTT and CD22 to be specifically expressed in ALL. There is also a well-founded biological explanation for the appearance of the CD22 and DNTT genes in some of the best projections separating different classes of the MLL data set. It was proven that the presence of cytoplasmic CD22 protein, a human B-lymphocyte-restricted antigen, is a useful marker for B-cell precursor acute lymphocytic leukemia [3]. There is evidence also that terminal deoxynucleotidyl transferase (DNTT) is a unique DNA polymerase expressed in the lymphoid precursors of B- and T-cell lineage at the earliest recognizable stages of lymphopoiesis. DNTT is also expressed on their malignant counterparts, making it an important marker in distinguishing lymphoblastic leukemia from other haematologic neoplasms [16].

Instead of considering only the best rated visualization, we can examine several top ranked projections to find genes



(a) (b)

Figure 5: (a) The best radviz plot from the SRBCT data set. (b) The best radviz projection from the SRBCT data set, where the class labels were randomly permuted.

that are relevant in the differentiation of different cancer types. Therefore it is valuable to know if a particular gene is present only in one good projection or if it appears in several top ranked projections. In Figure 7 we show a plot that lists the first 20 genes present in the top-rated scatterplot projections of the MLL data set. For each pair of genes (one from the x and one from the y axis), a black box indicates if their scatterplot projection is ranked among the best 500. The figure shows that the three genes – MME, POU2AF1 and LGALS1 – stand out in the number of their appearances in the top-ranked projections. Interestingly, all three genes are among the specifically expressed genes in MLL, ALL or AML leukemic samples reported by Armstrong*et al.* [1]. In their work, MME and POU2AF1 are the first and tenth gene, respectively, most highly correlated with ALL, while LGALS1 is the eight most highly correlated gene with MLL compared with the remaining two classes.

We found similar biological relevance of genes that participated in the best visualizations of other data sets. It turns out that VizRank does not only find good projections which can well separate the diagnostic classes, but at the same time also finds genes that were already experimentally proven to be relevant in the diagnosis of different cancer types. Most of our visualizations included in this paper point to non-linear gene interactions, giving VizRank an advantage over univariate feature selection algorithms prevailingly used in the current related work in the area.

Tumorigenesis in humans is a multi-step process, where the steps reflect four to seven stochastic genetic alterations that drive the progressive transformation of normal human cells into highly malignant derivatives [11]. During the process of this transformation gene regulatory networks are disrupted causing alteration in the expression of many genes. We can not, in general, assert that the genes shown in the best ranked projections are those that are responsible for the cancer transformation. However, these genes can clearly be used to differentiate between different cancer types and

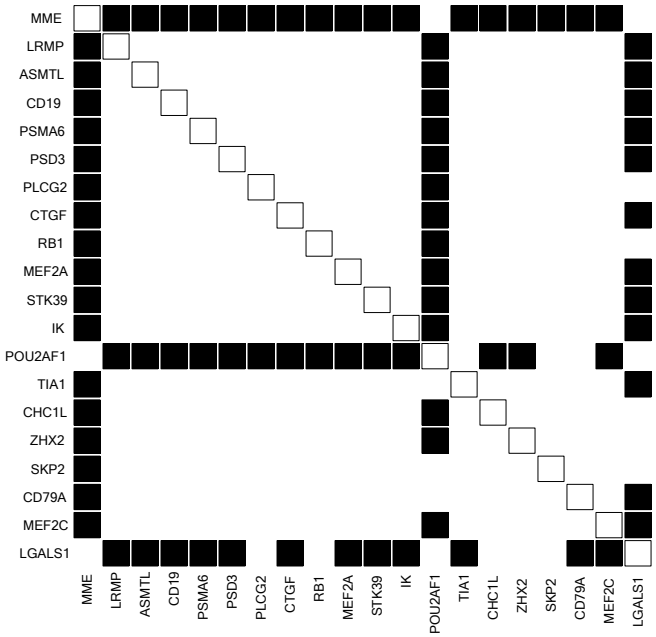
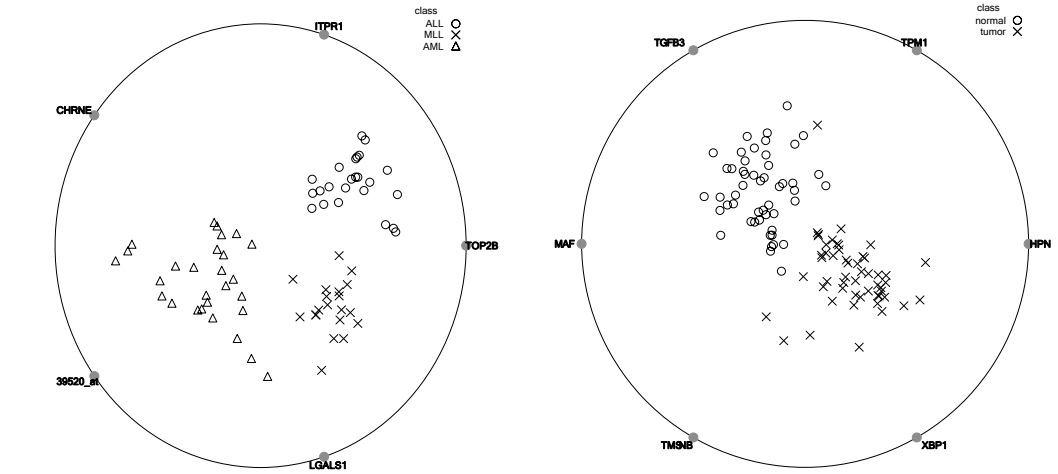


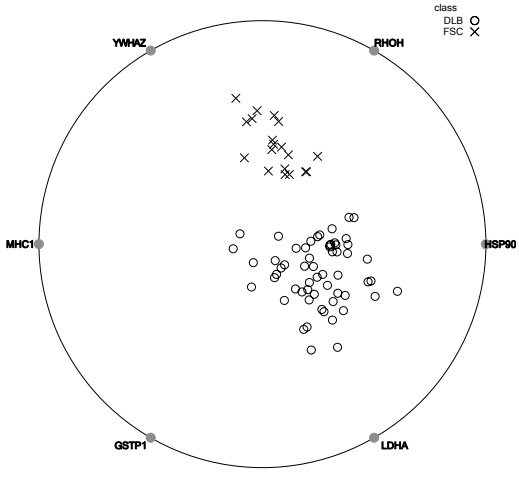
Figure 7: Genes on the x and y axis are the first 20 genes from the list of top-ranked scatterplot projections of the leukemia data set. Each black box indicates that the corresponding genes on the x and y axis form a scatterplot that is ranked as one of the best 500 scatterplots.

moreover, some of them are known pathognomonic markers of special cancer types, while others might turn out to be so in the future.

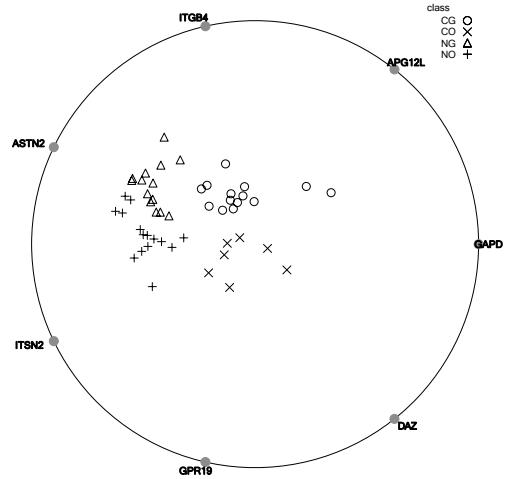


(a) MLL

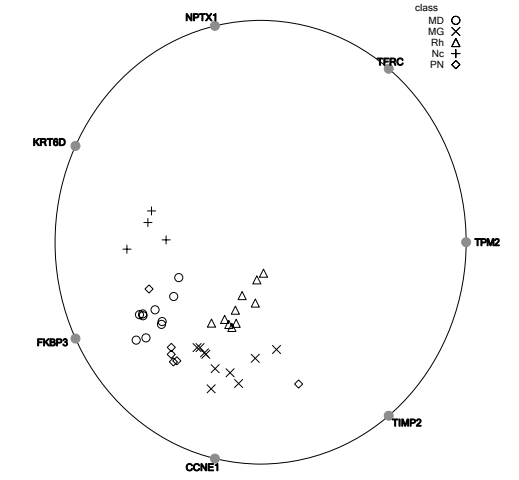
(b) prostata



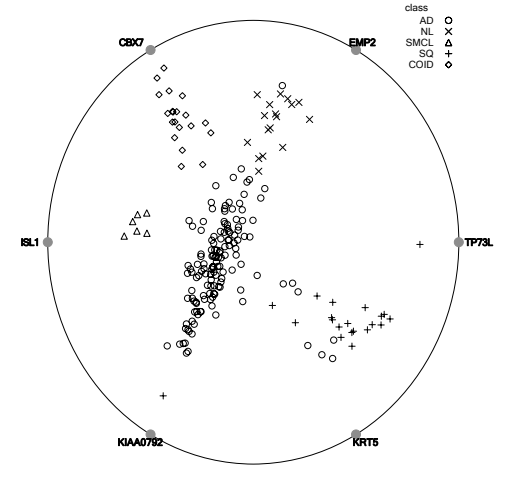
(c) DLBCL



(d) glioblastoma



(e) brain tumor



(f) lung cancer

Figure 6: Top-ranked radviz projections from the MLL, prostata, DLBCL, glioblastoma, brain tumor, and lung cancer data sets.

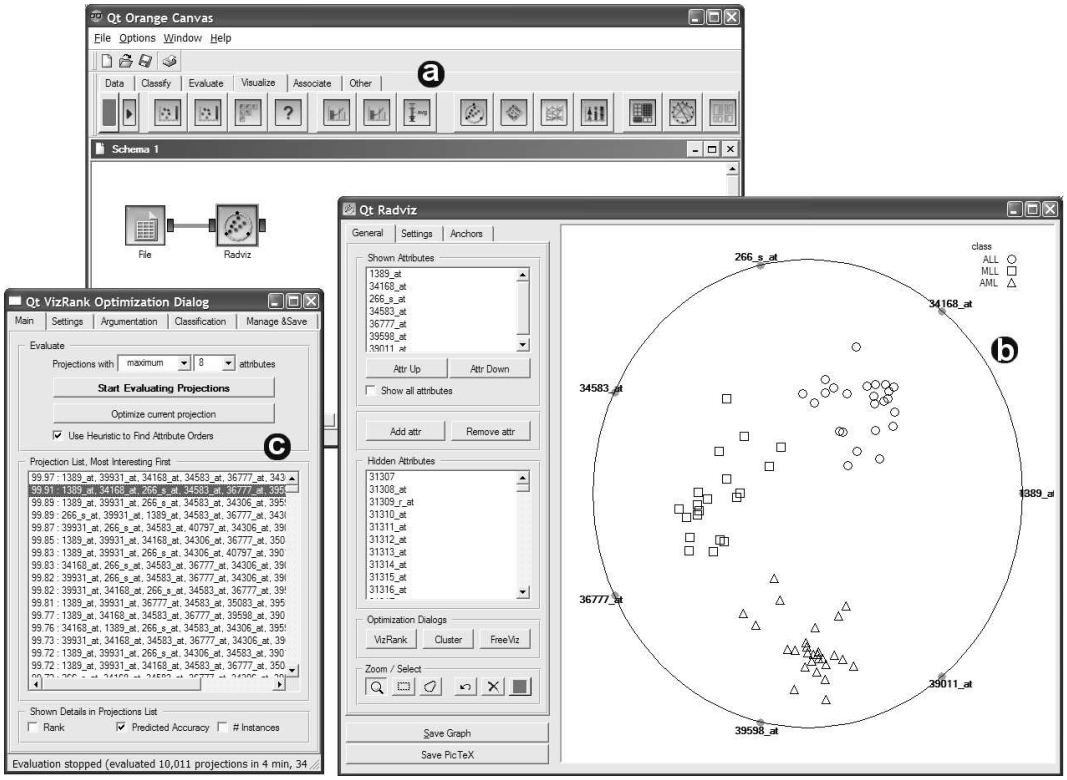


Figure 8: Screenshot of the Orange data mining suite (a) with radviz visualization widget (b) and VizRank dialog (c) that shows a list of best-rated projections for the MLL data set. Second-best projection is selected and shown in radviz.

5. CONCLUSION

Perhaps most striking and to a good degree unexpected result from experiments reported in this work is that we found a simple geometric visualizations that clearly visually differentiate among cancer types for all cancer gene expression data sets investigated. This finding complements recent related work in the area that demonstrates that gene expression cancer data can provide ground for reliable classification models. However, our "visual" classification models are much simpler and comprise much smaller number of genes when compared to those of, say, recently published artificial neural networks and support vector machines models that most often use anywhere from 50 features (genes) and encrypt their relation with the diagnostic class in at best hard-to-interpret model.

VizRank, a method we propose to find the most informative visualizations, is relatively fast: good visualizations with clear class separations are often provided to an experimentalists within minutes, with subsequent small improvements in the score of best rated visualization by letting the program run further.

The approach presented here is of course not limited to cancer gene expression analysis, and can be applied to search for good geometric visualizations on any class-labeled data set that includes continuous or nominal features. VizRank is freely available within a Scatterplot and Radviz widget in Orange open-source data mining suite [7, 8] (see Figure 8).

6. REFERENCES

[1] S. A. Armstrong, J. E. Staunton, L. B. Silverman, *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, 2001.

[2] A. Bhattacharjee, W. G. Richards, J. Staunton, *et al.* Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 98(24):13790–13795, 2001.

[3] D. Boue and T. LeBien. Expression and structure of cd22 in acute leukemia. *Blood*, 71(5):1480–1486, 1988.

[4] C. Brunsdon, A. S. Fotheringham, and M. Charlton. An investigation of methods for visualising highly multivariate datasets. *Case Studies of Visualization in the Social Sciences*, pages 55–80, 1998.

[5] J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Handbook of perception and cognition*, pages 69–117. Academic Press, San Diego, CA, 1995.

[6] B. W. Dasarathy. *Nearest neighbor (NN) norms: NN pattern classification techniques*. IEEE Computer Society Press, 1991.

[7] J. Demšar and B. Zupan. From experimental machine learning to interactive data mining, a white paper. *AI Lab, Faculty of Computer and Information Science, Ljubljana*, 2004.

[8] J. Demšar, B. Zupan, G. Leban, and T. Curk. Orange: From experimental machine learning to interactive

data mining. In *PKDD*, pages 537–539, 2004.

[9] L. M. Fu and C. S. Fu-Liu. Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. *FEBS Letters*, 561(1-3):186–190, 2004. TY - ABST.

[10] T. R. Golub, D. K. Slonim, P. Tamayo, *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[11] D. Hanahan and R. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[12] P. E. Hoffman, G. G. Grinstein, K. Marx, *et al.* DNA visual and analytic data mining. *IEEE Visualization 1997*, 1:437–441, 1997.

[13] J. Khan, J. S. Wei, M. Ringnr, *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *7*, 6(1):673–679, 2001.

[14] D. Komura, H. Nakamura, S. Tsutsumi, *et al.* Multidimensional support vector machines for visualization of gene expression data. *Bioinformatics*, 21(4):439–444, 2005.

[15] I. Kononenko and E. Simec. Induction of decision trees using relieff. In *Mathematical and statistical methods in artificial intelligence*. Springer Verlag, 1995.

[16] L. Liu, L. McGavran, M. A. Lovell, *et al.* Nonpositive terminal deoxynucleotidyl transferase in pediatric precursor b-lymphoblastic leukemia. *American Journal of Clinical Pathology*, 121(6):810–815, 2004.

[17] C. L. Nutt, D. R. Mani, R. A. Betensky, *et al.* Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, 63(7):1602–1607, 2003.

[18] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[19] M. A. Shipp, K. N. Ross, P. Tamayo, *et al.* Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

[20] D. Singh, P. G. Febbo, K. Ross, *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[21] A. Statnikov, C. F. Aliferis, I. Tsamardinos, *et al.* A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5): 631–643, 2005.