Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial
ISSN: 1137-3601
revista@aepia.org
Asociación Española para la Inteligencia Artificial
España

Strickert, M.; Schleif, F. M.; Seiffert, U.; Villmann, T.
Derivatives of Pearson Correlation for Gradient-based Analysis of Biomedical Data
Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, vol. 12, núm. 37, 2008, pp. 37-44
Asociación Española para la Inteligencia Artificial
Valencia, España

ARTÍCULO

Derivatives of Pearson Correlation for Gradient-based Analysis of Biomedical Data

M. Strickert[a*], F.-M. Schleif[b], U. Seiffert[a], T. Villmann[b]

[a] Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben (IPK)
[b] University of Leipzig, Medical Department
[*] Corresponding author: stricker@ipk-gatersleben.de

Abstract

In biomedical analytics one of the major criteria for the characterization of similarities between measured data items is correlation. We demonstrate the use of the formal derivative of Pearson correlation for gradient-based optimization of data models. Firstly, individual data attributes can be rated according to their impact on pairwise data relationships, analogous to the variance measure in Euclidean space. Secondly, a versatile method is presented to perform faithful multi-dimensional scaling from a high-dimensional space of source data to a low-dimensional target space, a method driven by maximization of correlation between distances of static source data and adaptive target vectors. As shown for mass spectroscopy data, a combination of attribute rating and data visualization helps revealing interesting data properties.

Keywords: Feature rating, data visualization, correlation.

1 Introduction

Massive data sets with a high number of samples and/or attributes create challenges in *de novo* data analysis. Particularly, high-throughput biomedical devices like mass spectrometers or gene expression arrays generate thousands of data points in parallel for which accurate computational screening technologies are required. Initial data screening usually involves the calculation of data-specific characteristic moments, such as mean value and variance; moreover, principal component analysis (PCA) is used to reduce the complexity of high-dimensional data by projecting data to potentially interesting axes of maximum variance [4]. Further data abstraction can be obtained by vector quantization, such as realized by self-organizing maps (SOM) which allows to obtain centroid-based low-dimensional data visualizations [5].

Best flexibility in data inspection is provided by methods that allow application of generic, particularly task-specific, data similarity measures. PCA, for example, is designed to map Euclidean data spaces from high to low dimensions. Its optimization targets are principal data directions of maximum variance.

Custom data similarity measures beyond Euclidean distance are available in other methods such as self-organizing maps (SOM) or k-means [2]. Yet, for example the SOM update rule 'make close centroids more similar to the data' is traditionally implemented as identity relationship: centroids are moved on straight *Euclidean* lines, in portions depending on their closeness, towards presented data points. The model update is thus optimum for Euclidean spaces, but not necessarily for the chosen similarity measure. Such discrepancy can be circumvented by coupling the update with analytic properties of the selected similarity measure.

In cost function frameworks, commonly used in artificial intelligence methods, the model parameters can be adapted by optimization of the desired similarity relationships.

Here, foundations are laid for gradient-based optimization of Pearson correlation similarity. Correlation is often used in biomedical analysis tasks, and its formal derivative can be plugged into existing gradient methods such as the Heskes variant of SOM [3], neural gas [6], and generalized learning vector quantization [7, 8], to name just a few.

In the following, the application of correlation derivative is attribute characterization and outlier detection in correlation contexts. For visualization, the derivative is further used in the cost function of a very powerful dimension reduction technique called high-throughput multidimensional scaling (HiT-MDS). The methods are applied to a large mass spectroscopy data set.

2 Methods

Variance is one of the most essential concept in statistics. It is not only a statistical index to measure attribute variability, but it is also required in more complex operations ranging from data normalization, like the z-score transform, to PCA for dimension reduction and visualization. In this section, an alternative point of view is presented to assess the contribution of data attributes involved in calculations of metrics and similarity measures for data vectors. Thereby, Pearson correlation is the focus of choice for enhancing pattern-based mass spectrum analysis. The fundamental concept of gradients of Pearson correlation will be used for a revised version of high-throughput multidimensional scaling (HiT-MDS).

2.1 Statistical variance transfered to Pearson correlation

Variance is tightly bound to data spaces with Euclidean distance, as will be discussed the following. The main consideration is to analytically evaluate the influence of individual vector components on the underlying similarity measure (or metric) used for pairwise vector comparisons.

Variance – the standard case. For two data vectors $\mathbf{x}, \boldsymbol{w} \in \mathbb{R}^d$ the squared Euclidean distance d^2 and component derivatives are calculated by

$$\begin{aligned} \mathrm{d}^2(\mathbf{x}, \boldsymbol{w}) &= \sum_{i=1}^{d} (x_i - w_i)^2 \quad , \\ \frac{\partial \mathrm{d}^2(\mathbf{x}, \boldsymbol{w})}{\partial w_k} &= -2 \cdot (x_k - w_k). \end{aligned} \tag{1}$$

The derivative indicates the contribution of the k-th vector component of $\boldsymbol{w}$ to the distance. Since the negative gradient refers to the direction from $\boldsymbol{w}$ towards minimizing its distance to $\mathbf{x}$, the largest absolute component of the gradient represents the most influential attribute. Because of symmetry in Euclidean space, same contributions are obtained for the gradient from $\mathbf{x}$ to $\boldsymbol{w}$.

For an arbitrary number n of data vectors $\mathbf{x}^i \in \mathbf{X}$ and a general similarity measure d we define the overall contribution of the k-th attribute for all vector pairs by the quantity

$$V_k(\mathbf{X}) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\partial \mathrm{d}(\mathbf{x}^i, \mathbf{x}^j)}{\partial x_k} \right)^2 . \tag{2}$$

In case of the squared Euclidean distance d^2 the term V_k can be transformed to

$$\begin{aligned} V_k(\mathbf{X}) &= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_k^i - x_k^j)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_{i,k} - \mu_{\mathbf{X}_{*,k}})^2 \\ &= \mathrm{var}(\mathbf{X}_{*,k}) \, . \end{aligned}$$

Thus, the variance $\mathrm{var}(\mathbf{X}_{*,k})$ of the k-th data attribute $\mathbf{X}_{*,k}$, i.e. k-th column of the data matrix $\mathbf{X}$, can be turned from a double sum expression into the usual mean-centered sum of squares using the pre-computed average $\mu_{\mathbf{X}_{*,k}}$. Since the derivative of squared Euclidean distance Eqn. 1 yields a decomposition into individual attributes, variance is computed independently for each attribute. As demonstrated below, however, such independence does not hold for all similarity measures. Moreover, changing back to $\mathbf{x}$ and $\boldsymbol{w}$, the symmetry of $\partial \mathrm{d}(\mathbf{x}, \boldsymbol{w}) / \partial w_k = -\partial \mathrm{d}(\mathbf{x}, \boldsymbol{w}) / \partial x_k$, valid in Euclidean case, is not always guaranteed.

Variance analogon for Pearson correlation. The derivation of sensitive data components for Pearson correlation comparisons is analogous to the previous one for the squared Euclidean distance. Pearson correlation is given by

$$\mathsf{r}(\mathbf{x}, \boldsymbol{w}) = \frac{\sum_{i=1}^{d} (x_i - \mu_{\mathbf{x}}) \cdot (w_i - \mu_{\boldsymbol{w}})}{\sqrt{\left(\sum_{i=1}^{d} (x_i - \mu_{\mathbf{x}})^2\right) \left(\sum_{i=1}^{d} (w_i - \mu_{\boldsymbol{w}})^2\right)}} \, . \tag{3}$$

With abbreviations $\mathsf{r}(\mathbf{x}, \boldsymbol{w}) =: \frac{\mathscr{B}}{\sqrt{\mathscr{C} \cdot \mathscr{D}}}$ the required derivative w.r.t. component w_k is given by the formula

$$\frac{\partial \mathsf{r}(\mathbf{x}, \boldsymbol{w})}{\partial w_k} = \frac{(x_k - \mu_{\mathbf{x}}) - \frac{\mathscr{B}}{\mathscr{D}} \cdot (w_k - \mu_{\boldsymbol{w}})}{\sqrt{\mathscr{C} \cdot \mathscr{D}}} . \quad (4)$$

This formula is very important for our next tasks. As can be seen, $\mathscr{B}, \mathscr{C}$ and $\mathscr{D}$ induce interdependence of attributes, and for non-normalized data $\partial \mathsf{r}(\mathbf{x}, \boldsymbol{w})/\partial w_k \neq -\partial \mathsf{r}(\mathbf{x}, \boldsymbol{w})/\partial x_k$, as illustrated in the example below. The derivative given in Eqn. 4 can be used in Eqn. 2 in order to obtain ratings of attributes of data vectors compared by Pearson correlation.

In practice, the presented derivation for component rating in Pearson correlation can be used without restrictions. However, some remarks should be considered. For identical vector pairs the derivative is set to zero in order to avoid numerical problems. If either data vector $\mathbf{x}$ or $\boldsymbol{w}$ is constant, correlation gets zero; therefore, the derivative should be also set to zero, because of an indecisive direction. It should be noted that the computing time of the correlation variance analogon is $\mathcal{O}(n^2)$ instead of $\mathcal{O}(n)$ required for standard variance.

An example. In order to convey an impression of the properties of the variance analogon, two synthetic 100-dimensional data vectors are compared. The components of the first one, $\mathbf{x}$, are drawn from a uniform data distribution $x_i \in [0; 1]$. The second one, $\boldsymbol{w}$, adds a linear function to $\mathbf{x}$: $w_i = 0.75 \cdot i/100 + x_i$, $i = 1 \ldots 100$. A scatter plot of the components of the two data vectors is given in Fig. 1.

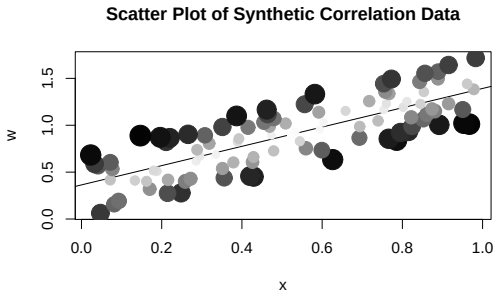

Figure 1. Scatter plot of $n = 2$ synthetically generated 100-dimensional vectors, including the regression line as reference to attribute correlation. Point shades and sizes correspond to the impact of the attributes on the regression according to Eqn. 4 plugged into Eqn. 2: the darker and bigger the point, the more important is the attribute.

While robust attributes (light small points) are found close to the regression line around (0.5,1), sensitive attributes (dark large points) are found on both ends above and below that line. Derivatives are much depending on the reference point $\mathbf{x}$ and $\boldsymbol{w}$, as Fig. 2 illustrates. Interestingly, the two vector derivatives reflect independent and dependent variable as more or less continuous lines with different levels of fluctuations. After all, meaningful information can be obtained by the new variance analogon for correlation.

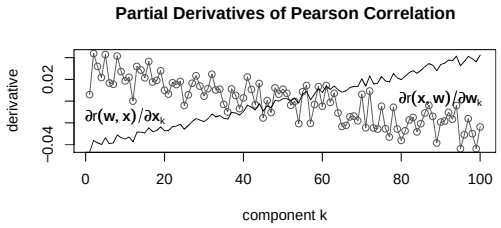

Figure 2. Correlation derivatives of the 100 data attributes contributing to the point shading in Fig. 1. While the gradient in $\mathbf{x}$ (dark upwards directed line) displays the synthetically induced linear dependence of $\boldsymbol{w}$ quite nicely, the gradient in the dependent vector $\boldsymbol{w}$ (downwards directed line with bullets) reflects more uncertainty.

Relationship between Pearson correlation and Euclidean distance. Pearson correlation provides invariance to rescalings of whole data vectors by a common multiplication factors and additive component offsets, such as the gain of measuring devices and homogeneous background signals. This feature makes Pearson correlation very suitable for dealing with data from individually calibrated devices that are often used in biomedical sciences. The invariance feature of Pearson correlation results from implicit data normalization in Eqn. 3. The denominator normalizes the centered covariance $\mathscr{B}$ of $\mathbf{x}$ and $\boldsymbol{w}$ by their individual standard deviations expressed by square roots of $\mathscr{C}$ and $\mathscr{D}$.

In iterative algorithms it might be advisable to make parts of the implicit normalization in Pearson correlation explicit by data preprocessing: the z-score transform $\mathbf{x}^z = (\mathbf{x} - \mu_{\mathbf{x}})/\sqrt{\mathrm{var}(\mathbf{x})}$ discards the mean value of $\mathbf{x}$ and yields unit variance. For z-score transformed vectors $\mathbf{x}^z$ and $\boldsymbol{w}^z$

the correlation of x and $\boldsymbol{w}$ can be expressed as simple scalar product $\langle\cdot,\cdot\rangle$ by

$$\mathsf{r}(\mathsf{x},\boldsymbol{w}) = \langle \mathsf{x}^z, \boldsymbol{w}^z\rangle/(d-1)\,.$$

Because of invariance $\mathsf{r}(\mathsf{x},\boldsymbol{w}) = \mathsf{r}(\mathsf{x}^z,\boldsymbol{w}^z)$. When the same notation is applied to the squared Euclidean distance of z-score transformed data this yields

$$\begin{aligned}\mathrm{d}^2(\mathsf{x}^z,\boldsymbol{w}^z) &= \sum_{i=1}^{d}(x_i^z - w_i^z)^2\\ &= \langle \mathsf{x}^z,\mathsf{x}^z\rangle - 2\cdot\langle \mathsf{x}^z,\boldsymbol{w}^z\rangle + \langle \boldsymbol{w}^z,\boldsymbol{w}^z\rangle\\ &= 2\cdot(d-1)\cdot\big(1-\mathsf{r}(\mathsf{x},\boldsymbol{w})\big)\,.\end{aligned}$$

Thus, since correlation r can be expressed as distance d^2 and vice versa, are correlation-specific designs needed at all?

1. In dynamic models for which static precomputation is not possible there is no good reason to prefer explicit z-score normalization over implicit normalization in Pearson correlation.
2. It is much simpler to think of the 'derivative of correlation' rather than of the 'derivative of negative rescaled and shifted squared Euclidean distance'.
3. Most importantly, although correlations are invariant to z-score transforms, derivatives of correlations are not. Thus, data transformation would affect results of derivative-based analysis.

Therefore correlation-specific considerations should be pursued.

2.2 High-throughput multidimensional scaling (HiT-MDS) for visualization

For planar visualization of individual data points the most widely used method is principal component projection. However, PCA is restricted to linear mappings of high-dimensional data, thereby focusing on directions of maximum (Euclidean) variance. Contrary to that, in the following the goal is to obtain a display that reflects most faithfully the inter-vector similarities of the source data. A multi-dimensional scaling (MDS) framework is utilized to meet this target by optimization of free parameters $\hat{\mathbf{x}}^i \subset \hat{\mathbf{X}}$ i.e. locations of points $\hat{\mathbf{x}}^i = (\hat{x}^i_1,\ldots,\hat{x}^i_q)$ in a q-dimensional target space corresponding to $i = 1\ldots n$ input vectors $\mathsf{x}^i \in \mathbf{X}_{n\times d}$ of dimension d [1]. An advanced MDS scheme for high-throughput data (HiT-MDS) for faithful distance reconstruction, based on powers of correlation, has been proposed earlier by the authors [9]. Here, a significant improvement of the original method is obtained by a heuristic simplification described below. As a result, convergence gets much faster without noticeable loss of accuracy; moreover, only the number of training iterations and – a now quite insensitive – learning rate need to be chosen to run the new version of HiT-MDS.

In brief, let matrix $\mathbf{D} = (\mathrm{d}_{ij})_{i,j=1\ldots n}$ contain pattern distances, and $\hat{\mathbf{D}} = (\hat{\mathrm{d}}_{ij})_{i,j=1\ldots n}$ those of the reconstructions. Then HiT-MDS maximizes the correlation $\mathsf{r}(\mathbf{D},\hat{\mathbf{D}})$ between entries of the source distance matrix $\mathbf{D}$ and the reconstructed distances $\hat{\mathbf{D}}$ by minimizing the embedding cost function:

$$\mathrm{s} = -\mathsf{r}\circ\hat{\mathbf{D}}\circ\hat{\mathbf{X}} \to \min \qquad (5)$$

$$\Rightarrow \frac{\partial \mathrm{s}}{\partial \hat{x}^i_k} = -\sum_{j=1\ldots n}^{j\neq i}\frac{\partial \mathsf{r}}{\partial \hat{\mathrm{d}}_{ij}}\cdot\frac{\partial \hat{\mathrm{d}}_{ij}}{\partial \hat{x}^i_k} \to 0 \qquad (6)$$

Thus, locations of n points $\hat{\mathbf{x}}^i$, $i = 1\ldots n$, in target space are obtained by gradient descent on the stress function s using the depicted chain rule. For that purpose, randomly drawn points $\hat{\mathbf{x}}^i$ are updated into the direction of the sign $\mathrm{sgn}(x) = x/|x|$ of the steepest gradient of s, scaled by the learning rate γ: $\Delta\hat{x}^i_k = -\gamma\cdot\mathrm{sgn}(\frac{\partial \mathrm{s}}{\partial \hat{x}^i_k})$. Since the dynamic range of mere correlation is poorly suited for stress function optimization, in the previous formulation of HiT-MDS, some efforts were put into the design of appropriate power transforms for stress function enhancement, however, the new sign-based formulation is much more effective. Thereby, the speed-ups for incremental updates of the similarity matrix and the tracking of changes in the correlation $\mathsf{r}(\mathbf{D},\hat{\mathbf{D}})$, detailed in [9], are still valid.

For Euclidean output spaces with metric

$$\hat{\mathrm{d}}_{ij} = \sqrt{\sum\nolimits_{l=1}^{d}(\hat{x}^i_l - \hat{x}^j_l)^2}$$

the required derivatives in Eqn. 6 are

$$\frac{\partial \mathsf{r}}{\partial \hat{\mathrm{d}}_{ij}} = \frac{(\mathrm{d}_{ij}-\mu_{\mathbf{D}}) - (\hat{\mathrm{d}}_{ij}-\mu_{\hat{\mathbf{D}}})\cdot\mathscr{B}/\mathscr{D}}{\sqrt{\mathscr{C}\cdot\mathscr{D}}}$$

$$\frac{\partial \hat{\mathrm{d}}_{ij}}{\partial \hat{x}^i_k} = (\hat{x}^i_k - \hat{x}^j_k)\,/\,\hat{\mathrm{d}}_{ij}$$

While, for planar and intuitive plotting purposes, target distances $\hat{d}_{ij}$ are usually Euclidean, input distances can be general Minkowski metrics or mere dissimilarities, like (powers of) flipped Pearson correlation $d_{ij} = (1 - r(\mathbf{x}^i, \mathbf{x}^j))$. Note that these data vector correlations are completely independent of the target of the correlation-based cost function optimization in HiT-MDS.

3 Analysis of medical data

The presented methods are applied to mass spectroscopy data from a clinical cancer study. Frozen tissue material was sliced by a microtom from which, subsequently, 1050 mass spectra where taken using a linear MALDI-TOF MS, Autoflex, in a range of 2000-10000Da (devices and spectra preparation software by Bruker Daltonik GmbH, Bremen). The data preparation protocol of the measured spectra followed the default workflow in ClinProTools 2.1 software (Bruker) for baseline correction, alignment and peak picking. For robustness, peaks with signal to noise ratio $S/N > 5$ were used for further analysis, and only maxima of the extracted peaks were considered. Finally, a data set of 1050 high-quality spectra with 32 values per mass spectrum was obtained.

Due to the overall complexity of tissues – being a mixture of different cell types – the data are quite inhomogeneous and multi-modality can be expected. In cancer identification, however, a proper characterization of high-density regions in the multi-dimensional spectra is required, because such modes might indicate healthy or malignant tissues. In the following, it is demonstrated how the presented methods can be used for data outlier identification (variance analogon) and for faithful visualization of spectrum relationships (HiT-MDS). Additionally, it is shown how the variance analogon helps to extract only those peaks which are meaningful for proper reconstructions of the spectra relationships.

Although tissues are labeled by two cancer classes and one normal class, this additional information is only used for visual annotation in this study. There are suitable supervised data models to obtain prototype-based classification [8, 10], but since, for the present data set, there is still some uncertainty in the manual class assignment, the introduced unsupervised methods are intended to assist in tasks of data inspection and hypothesis generation.

3.1 Attribute characterization

An interesting starting point is the variance analysis of the data set with respect to the importance of peaks. Standard variance yields a peak rating displayed in Fig. 3, top panel, while the new variance analogon based on Eqn. 2 yields a different rating, as shown in the bottom panel. The connected different rankings become much obvious, for example, for attribute number 11. In contrast to the very large numbers connected to variance, the variance analogon rating produces rather small values. This indicates that attribute rating for squared Euclidean distances and Pearson correlation are methods for quite different quantities. In a feature selection scenario, described in the HiT-MDS visualization section below, it will be shown that the feature rating carries very meaningful information for correlation-based data characterization.

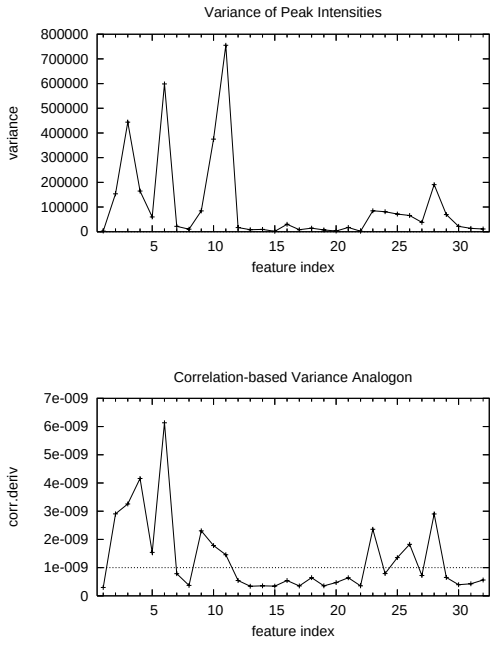

Figure 3. Standard variances of peak intensities (top) and variance analogon for correlation-based comparison (bottom) of spectra. Different ratings between both methods are observed: for example, different peaks 6 and 11 are ranked as most influential. The horizontal line at 10^{-9} in the bottom panel separates 12 top-rated peaks from 20 low-rated peaks used in a subsequent feature selection scenario.

3.2 Visualization of the data correlation space

Data screening, outlier and cluster validation are much enhanced by appropriate data visualizations. When data relationships should be communicated, the need for proper data displays is obvious.

Attribute characterization. Much structure of data attributes, i.e. the considered peaks, is revealed by combining the scatter plot display of their correlatedness with their variance analogon ratings. The HiT-MDS embedding in Fig. 4 arranges peaks by their (1-r)-pseudo-distance, thus, highly correlated masses are closely grouped, while discorrelated masses are distantly arranged. Such a plot gives an intuitive glimpse on attribute redundancy and importance, which strongly helps in attribute characterization. Using a C-implementation, it takes less than a second to compute the 2500 data cycles, using a learning rate of 1 until convergence to the displayed point configuration.

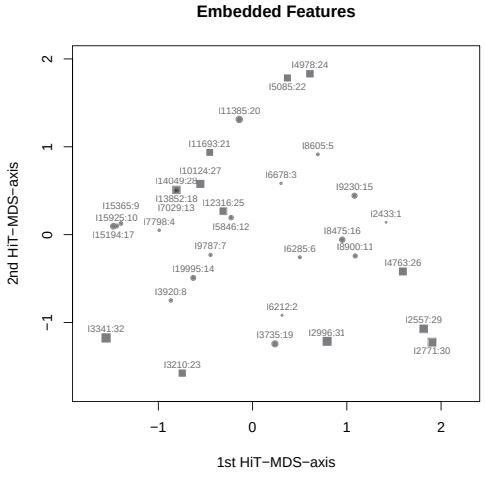

Figure 4. Visualization of the correlatedness of 32 mass spectrum peaks. The embedding shows a good spreading of the peaks, i.e. non-redundant information. Only peaks at disjoint masses of 7029Da, 13852Da, and 14049Da do highly correlate, as well as at similar masses of 15194Da and 15925Da. Colon-separated numbers 1–32 denote the importance of the peak in terms of variance analogon; these numbers are proportional to the point size. Circles refer to the 20 least-rated peaks, while squares correspond

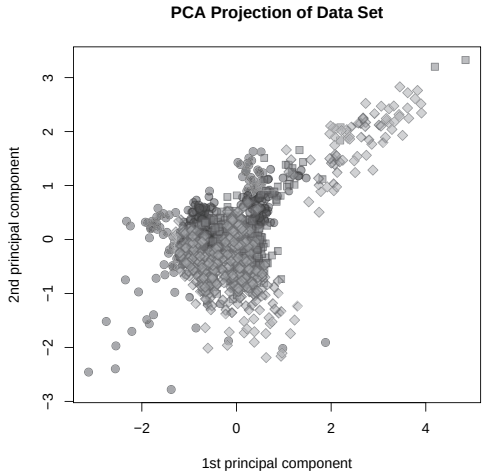

Figure 5. Principal component projections of the mass spectrum data set. No prominent structure is revealed: the three tissue classes represented by diamonds, bullets, and squares are projected into heavily overlapping scatter regions.

Data browser. A common starting point for assessing the arrangement of prominent data patterns is the projection of multidimensional data onto the two major principal components. Figure 5 shows such a plot for the mass spectrum data. Apart from a diamond-rich region on the top right, characteristic of one of the cancer tissues, the plot does not exhibit much valuable information.

Much more data structure is obtained by the nonlinear HiT-MDS embedding procedure. A number of 500 HiT-MDS cycles has been trained, taking less than a minute on a 3GHz P4 computer; a learning rate of $\gamma = 100$ has been used. The resulting correlation-based visualization shown in Fig. 6 contains many more details than the previous PCA projection. The pairwise relationships of the displayed 2D-points correspond at a high level of $r = 0.934$ with the original data relationships of the 32-dimensional spectra. The cancer class, related to diamond shapes, is clearly multimodal which hints at the need for a more refined pathological characterization. However, the visual separation of the manually labeled classes indicates that the corresponding spectra do contain meaningful differences which could be exploited by a classification system. Additionally,

tissue, can be easily identified. Finally, the size of the symbols, again representing outstanding samples for the correlation-based characterization of the complete data set, can be used to specify the support of the data set.

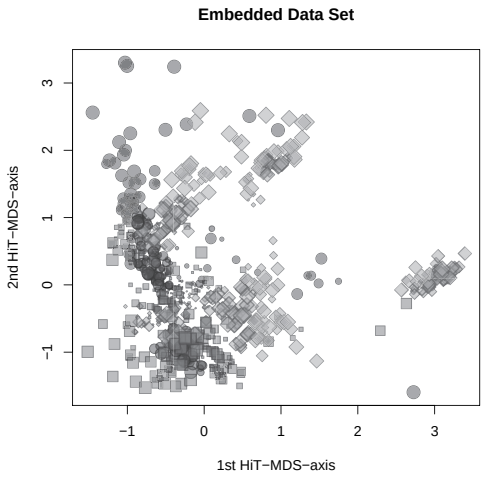

Figure 6. Visualization of the mass spectrum data set using HiT-MDS with (1-r) data similarity measure.

Feature selection. In multi-dimensional data sets, one of the interesting questions concerns the selection of important data attributes. Such a focus helps to reduce vast amounts of data coming from spectroscopic devices and it also helps in noise cancellation.

In the previous section about variance analogon, a subset of 12 top-rated features out of 32 has been identified (cf. Fig. 3 and Fig. 4), but a confirmation of this subset is still missing. Hence, we give results for exactly the same embedding experiment as in the last paragraph, but only for the 12-dimensional feature subset. The obtained visualization in Fig. 7 shows quite a good structural agreement with the display in Fig. 6 for the complete set of attributes. Without affecting too much the visual class separability, the scattering is partially twisted and a little sharpened in contrast to Fig. 6. Despite the heavy reduction from 32 to 12 dimensions, the pairwise relationships of the displayed 2D-points correspond at level of $r = 0.939$ with the original data relationships. This is even higher than the level for the unreduced feature set. This surprising result is easily explained, because it is more difficult to embed high-dimensional data containing spurious fluctuations than to embed clearly expressed data.

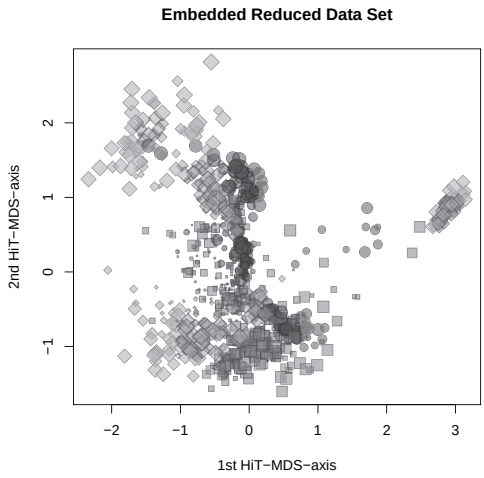

Figure 7. Visualization of the mass spectrum data set after attribute reduction to the 12 top-rated peaks. This visualization can be directly compared with Fig. 6.

4 Conclusions

A uniform approach has been presented for providing an harmonic interplay of characterization and visualization in screening of biomedical data. Based on the mathematical derivative of the Pearson correlation coefficient, a new approach to attribute rating, analog to the concept of statistical variance, has been presented. With this measure, qualitatively new aspects of the data become visible, namely, the influence of attributes on the correlation relationships between data vectors. The derivative of correlation can be plugged as gradient into data models utilizing cost functions, such as SOM and neural gas, or, as shown, into the embedding stress function of the multidimensional scaling. In the latter case, fast and accurate visualization of the correlation structure in data has been obtained by a modification of high-throughput multidimensional scaling (HiT-MDS). The new version of HiT-MDS compensates the usually extremely small contributions of correlation derivatives to the embedding stress function by sign-based updates of the embedded data points, which, after all, provides excellent convergence properties.

Although Pearson correlation is one of the gold standards in biomedical data analysis, the above concept can be easily generalized by replacing the derivative of Pearson correlation by that of any other suitable similarity measure. Obviously, the methods are not restricted to mass spectrum data, but there is also much potential for application to massive gene expression data sets, proteomic data, but even for data from technical domains. As indicated in this work for the case of Pearson correlation, the implementation of alternative data-specific similarity measures points into an interesting direction of future research.

C, MATLAB (GNU Octave), and R source code of high-throughput multidimensional scaling is online available at `http://hitmds.webhop.net/`.

Acknowledgements

This work is supported by BMBF grant FKZ 0313115 (GABI-SEED-II) and the Ministry for Cultural matters in Saxony-Anhalt, Germany, grant XP 3624HP/0606T.

References

[1] A. Buja, D. Swayne, M. Littman, N. Dean, and H. Hofmann. Interactive Data Visualization with Multidimensional Scaling. Report, University of Pennsylvania, URL: http://www-stat.wharton.upenn.edu/~buja/, 2004.

[2] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

[3] T. Heskes. Energy functions for self-organizing maps. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 303–316. Elsevier, Amsterdam, 1999.

[4] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2nd edition, 2002.

[5] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, 3rd edition, 2001.

[6] T. Martinetz and K. Schulten. A "neural-gas" network learns topologies. *Artificial Neural Networks*, I:397–402, 1991.

[7] A.S. Sato and K. Yamada. Generalized Learning Vector Quantization. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7 (NIPS)*, volume 7, pages 423–429. MIT Press, 1995.

[8] M. Strickert, U. Seiffert, N. Sreenivasulu, W. Weschke, T. Villmann, and B. Hammer. Generalized relevance LVQ (GRLVQ) with correlation measures for gene expression data. *Neurocomputing*, 69:651–659, 2006.

[9] M. Strickert, S. Teichmann, N. Sreenivasulu, and U. Seiffert. High-Throughput Multi-Dimensional Scaling (HiT-MDS) for cDNA-array expression data. In W. Duch et al., editor, *Artificial Neural Networks: Biological Inspirations, Part I, LNCS 3696*, pages 625–634. Springer, 2005.

[10] Thomas Villmann, Frank-Michael Schleif, and Barbara Hammer. Fuzzy labeled soft nearest neighbor classification with relevance learning. In *ICMLA '05: Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA'05)*, pages 11–15, Washington, DC, USA, 2005. IEEE Computer Society.