ROBUST CENTROID-BASED CLUSTERING USING DERIVATIVES OF PEARSON CORRELATION

Marc Strickert, Nese Sreenivasulu
Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany

Thomas Villmann
Clinic for Psychotherapy, University of Leipzig, Germany

Barbara Hammer
Institute of Computer Science, University of Clausthal, Germany

Keywords: Centroid-based clustering, correlation, quantization cost optimization.

Abstract: Modern high-throughput facilities provide the basis of -omics research by delivering extensive biomedical data sets. Mass spectra, multi-channel chromatograms, or cDNA arrays are such data sources of interest for which accurate analysis is desired. Centroid-based clustering provides helpful data abstraction by representing sets of similar data vectors by characteristic prototypes, placed in high-density regions of the data space. This way, specific modes can be detected, for example, in gene expression profiles or in lists containing protein and metabolite abundances. Despite their widespread use, k-means and self-organizing maps (SOM) often only produce suboptimum results in centroid computation: the final clusters are strongly dependent on the initialization and they do not quantize data as accurately as possible, particularly, if other than the Euclidean distance is chosen for data comparison. Neural gas (NG) is a mathematically rigorous clustering method that optimizes the centroid positions by minimizing their quantization errors. Originally formulated for Euclidean distance, in this work NG is mathematically generalized to give accurate and robust results for the Pearson correlation similarity measure. The benefits of the new NG for correlation (NG-C) are demonstrated for sets of gene expression data and mass spectra.

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 data models are required in order to faithfully reduce the data complexity and to facilitate the analysis.

Centroid-based data representations provide most intuitive interpretations, because a centroid can be regarded as noise-free prototype of its surrounding cloud of data. Especially for many data vectors, centroids can be much more easily assessed than results from hierarchical clustering, for example. Some well-known clustering algorithms are freely available (de Hoon et al., 2004), implementing widely used methods of Eisen et al. (Eisen et al., 1998).

As a matter of fact, self-organizing maps (SOM) and k-means clustering are frequently used methods for extracting a pre-defined number of centroids from the data (Kohonen, 2001; MacQueen, 1967). While centroids in k-means get specialized by an iterative averaging procedure applied to data that they do currently represent, SOM is a cooperative model with updates of the best-matching centroid and also of its neighbors. Since SOM neighbors are related to lateral centroids arranged on a grid structure, the SOM update triggers a mapping of similar high-dimensional data vectors onto similar positions of that usually low-dimensional grid, often, a 2D-plane for visualization. Due to topological constraints induced by the grid, quantization accuracy, i.e. data representation, of the SOM centroids is often not optimum (Villmann and Claussen, 2006). Thus, if the dimension reduction

Strickert M., Sreenivasulu N., Villmann T. and Hammer B. (2008).
ROBUST CENTROID-BASED CLUSTERING USING DERIVATIVES OF PEARSON CORRELATION.
In *Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing*, pages 197-203
DOI: 10.5220/0001062601970203

feature of SOM is not needed, better representations are obtained without grid structure. This can be realized by a SOM-like algorithm called neural gas (NG) that will be of interest here.

Data condensation requires similarity criteria in order to gather related items. Besides Euclidean distance, Pearson correlation is one of the most often used comparison criteria in biological studies. In principle, a wide range of similarity measures, expressed as Minkowski metric or correlation, is available in self-organizing maps (SOM) and k-means.

There is a commonly overlooked problem connected to similarity rating and model update, though. SOM centroids, for example, are defined by their closeness to data points, and it is thus straight-forward to describe closeness by task-specific similarity measures. Yet, the SOM update rule 'make close centroids more similar to the data' is traditionally implemented as a claim for identity: centroids are moved on straight lines in *Euclidean* space, in portions depending on their closeness, towards presented data points. There is thus a difference between the update rule for a strict one-to-one correspondence of all centroid components with those of the represented data vectors, and the more relaxed desire of merely representing high similarity. Although, for vector pairs, identity is equivalent to maximum similarity, the situation is different for a single centroid representing many data points; then, similarity constraints do usually allow more degrees of freedom for the centroid placement than strict identity constraints. Analogous considerations apply to the k-means clustering method, in which custom measures define data assignments to centroids, but in which average data centroids are strictly computed (by averaging) in Euclidean space.

The discrepancy between similarity computation and subsequent update of data models can be circumvented by coupling the update procedure with analytic properties of the selected similarity measure. In cost function frameworks the model parameters can be adapted by optimization of similarity relationships. Here, gradient-based optimization of centroid locations is discussed for Pearson correlation similarity. Correlation is often used in biomedical analysis tasks. It has got favorable pattern matching characteristics, and it allows to calculate formal derivatives and can be directly used in gradient methods such as the Heskes variant of SOM (Heskes, 1999), neural gas (Martinetz and Schulten, 1991), and generalized learning vector quantization (Sato and Yamada, 1995). The subsequent derivative is integrated into the highly accurate neural gas clustering method, for which superior performance is demonstrated for gene expression data and mass spectrum data.

2 METHODS

Faithful data representation requires robust centroid locations within the data. Self-organizing maps (SOM) realize a cooperative centroid placement strategy by iterative presentation of data points that trigger further improvements of previously placed centroids. A general formulation of this simple procedure is given in Algorithm 1.

Algorithm 1 SOM / NG centroid update

```
repeat
  chose randomly a data vector $\mathbf{x}$
  $k \leftarrow \arg\min_i \{ \mathrm{d}(\mathbf{w}^i, \mathbf{x}) \}$
      $\{ \mathbf{w}^k$ is closest centroid to data vector $\mathbf{x} \}$
  for all $m$ centroids $j$ do
    $\mathbf{w}^j \leftarrow \mathbf{w}^j + \gamma \cdot \mathrm{h}_\sigma \left( D(\mathbf{w}^k, \mathbf{w}^j) \right) \cdot U(\mathbf{x}, \mathbf{w}^j)$
      $\{ \gamma, \mathrm{h}, \sigma, D, U$: see text $\}$
  end for
until no more major changes
```

SOM Mode of Algorithm 1. Since SOM centroids cooperate laterally on a grid structure, updates imply spatial specialization with similar grid neighbors. Grid dependencies between centroids k and j are expressed by the neighborhood index $D(\mathbf{w}^k, \mathbf{w}^j)$. For example, rectangular 2D grids possess four direct neighbors $\mathcal{N}_k$ of non-boundary centroids with $D(\mathbf{w}^k, \mathcal{N}_k) = 1$. The σ-range of neighborhood cooperation is expressed by the decreasing function h_σ, with maximum value at $\mathrm{h}_\sigma(0) = 1$. Often a Gaussian bell $\mathrm{h}_\sigma(D) = e^{-D^2/\sigma^2}$ is put upon the grid, contracted during update by shrinking $\sigma \to 0$. In addition to neighborhood characterization, the update strategy of centroid $\mathbf{w}^j$ facing data vector $\mathbf{x}$ is described by $U(\mathbf{x}, \mathbf{w}^j)$. As said above, centroids are most often moved on straight Euclidean lines towards the data vector, i.e. by the term $U(\mathbf{x}, \mathbf{w}^j) = (\mathbf{x} - \mathbf{w}^j)$, in small steps depending on the update rate $\gamma < 1$.

NG Mode of Algorithm 1. The neural gas algorithm works exactly the same as described in the previous SOM mode, except for one crucial exception: the centroid neighborhood is no longer defined on a pre-defined grid structure. Instead, the neighborhood changes dynamically in course of data presentation. The centroid closest to the currently presented data vector $\mathbf{x}$ is assigned a rank of zero, the runner-up gets a rank of one, and so forth. In general, the neighborhood is defined by the ranks relative to only the data vector: $D(\mathbf{w}^k, \mathbf{w}^j) = D(\mathbf{w}^j) = \mathrm{rnk}(\mathbf{x}, \mathbf{w}^j)$ with

$$\mathrm{rnk}(\mathbf{x}, \mathbf{w}^j) = \left| \{ \mathrm{d}(\mathbf{x}, \mathbf{w}^i) < \mathrm{d}(\mathbf{x}, \mathbf{w}^j), i = 1 \ldots m \} \right| .$$

In contrast to SOM, the best-matching centroid $\mathbf{w}^k$ does not induce a specialized structure on the grid neighbors, and the rank-based neighborhood is always data optimum. Centroid update profits from ranks, because they are useful for breaking ties, i.e. for differentiation of very similar data. Ranks are exponentially wrapped by $\mathrm{h}_\sigma(D) = e^{-D/\sigma}$, again $\sigma \to 0$ during update iterations. As for SOM, $U(\mathbf{x},\mathbf{w}^j) = (\mathbf{x} - \mathbf{w}^j)$ and $\gamma < 1$.

Its is known that the NG algorithm asymptotically realizes a stochastic gradient descent on the cost function (Martinetz et al., 1993):

$$\mathrm{E}(\mathbf{W},\sigma) = \frac{1}{C(\sigma)} \cdot \sum_{j=1}^{m}\sum_{i=1}^{n} \mathrm{h}_\sigma(\mathrm{rnk}\left(\mathbf{x}^i,\mathbf{w}^j\right)) \cdot \mathrm{d}(\mathbf{x}^i,\mathbf{w}^j)\,. \tag{1}$$

The scaling factor $C(\sigma) = \sum_{i=0}^{m-1} \mathrm{h}_\sigma(i)$ is used for normalization. In the limit $\sigma \to 0$, the NG mode of Algorithm 1 leads to a centroid placement that minimizes the total quantization error, defined by $\mathrm{d}(\mathbf{x}^j,\mathbf{w}^i)$, between m centroids and n data vectors. This property does not hold for the SOM version. Even worse, in general the mathematical optimization target of SOM is undefined (Cottrell et al., 1994), unless the costly modification proposed by Heskes is implemented (Heskes, 1999).

The benefits of neural gas are: mathematical understanding of centroid specialization, high reproducibility of results, neighborhood cooperation for robustness against initialization, and easy implementation. Very importantly, the generic formulation of the neural gas algorithm allows to create modifications with respect to the choice of the data similarity measure. A minor drawback of NG is the sorting operation, i.e. a computing complexity of $\mathcal{O}(n \log n)$, required for rank calculation. Therefore, a fast batch version of neural gas with quadratic convergence based on Newton's method has been proposed recently (Cottrell et al., 2006), complementing the iterative online approach discussed here. The authors do also present a method for clustering data only defined by a similarity matrix. For its simplicity, we stick to Algorithm 1 in the following, and we introduce a derivation making full use of the analytic properties of Pearson correlation for an improved centroid update rule.

Neural Gas Clustering with Pearson Correlation. Pearson correlation is our focus of choice, because it provides a certain degree of invariance to additive or multiplicative effects induced by measuring devices or biochemical probe concentrations. Thus, pattern-based analysis is enhanced by choosing Pearson similarity for data vectors and centroids, mathematically described with abbreviation $\mathrm{r}(\mathbf{x},\mathbf{w}) = \frac{\mathscr{B}}{\sqrt{\mathscr{C}\cdot\mathscr{D}}}$ by

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

In principle, the covariance of $\mathbf{x}$ and $\mathbf{w}$ gets standardized by the product of the individual variances of $\mathbf{x}$ and $\mathbf{w}$. However, due to dynamic centroid update, there is no much use in making the implicit standardization explicit by data preprocessing, such as z-score transformation. Furthermore, in cases when correlation is just a building block, like in the dissimilarity measure $(1-\mathrm{r})^p$ (Zhou et al., 2002), it is much more natural to think in terms of a self-contained equation (Eqn. 2) than in terms of statically preprocessed data.

Correlation described by Eqn. 2 can be plugged into the cost function Eqn.1 being optimized by gradients along partial derivatives of E with respect to coordinates of all centroids $\mathbf{w}$. In general, these derivatives indicate contributions of the k-th centroid component of $\mathbf{w}$ to the distance or similarity measure.

For the squared Euclidean distance $\mathrm{d}^2(\mathbf{x},\mathbf{w}) = \sum_{i=1}^{d}(x_i - w_i)^2$ this corresponds to the previously mentioned term $U(\mathbf{x},\mathbf{w}) = (\mathbf{x} - \mathbf{w})$:

$$\frac{\partial \mathrm{d}^2(\mathbf{x},\mathbf{w})}{\partial w_k} = -2 \cdot (x_k - w_k) \propto U(x_k, w_k)\,.$$

For Pearson correlation the derivative is

$$\frac{\partial \mathrm{r}(\mathbf{x},\mathbf{w})}{\partial w_k} = \frac{(x_k - \mu_\mathbf{x}) - \frac{\mathscr{B}}{\mathscr{D}} \cdot (w_k - \mu_\mathbf{w})}{\sqrt{\mathscr{C}\cdot\mathscr{D}}}\,. \tag{3}$$

Since the cost function should be minimized, correlation r is turned by negative sign into a dissimilarity measure. Therefore, the term $U(x_k,w_k) = -\partial\mathrm{r}(\mathbf{x},\mathbf{w})/\partial w_k$ is inserted into Algorithm 1 which constitutes the new version of neural gas for correlation-based centroid placement, NG-C for short. It can be shown that this correlation-based update rule yields a valid gradient descent also at the boundaries of the receptive fields. A proof, originally for the Euclidean case, is provided by (Martinetz et al., 1993), where a vanishing contribution of the ranks was presented. Since the proof does not rely on specific properties of the Euclidean metric, a direct transfer to Pearson correlation is possible. Therefore, Eqn. 1 is still a cost function that gets optimized by the neural gas algorithm.

Usually, good convergence is reached after 50–1000 repeated data cycles, depending of the size n of the data set and the number m of centroids. Thereby, the neighborhood range σ is exponentially decreased from a starting size of $\sigma = m$ to a small value of $\sigma = 0.001$. This involves all prototypes strongly in the beginning, contracting centroids towards the data

’center’, and it leads to a fine-tuning of data-specific centroids in the final phase.

3 RESULTS

The following three applications show the superiority of NG-C clustering over traditional methods with Pearson correlation. As demonstrated, cost function optimization by NG-C provides better data representations and higher reproducibility of results.

3.1 Single Cluster Representation of Gene Expression Data

A first proof of concept is given for the simple, but illustrative task of finding only a single centroid position. This points out structural differences between Euclidean- and correlation-based centroid update. We use an exemplary 14-dimensional gene expression data set, where macroarrays were used to cover 14 temporal developmental stages in the endosperm tissue of developing barley grains, sampled from day 0 after flowering in steps of two days to day 26. After quality-based filtering, 4824 highly reliable genes were obtained. Conforming to standards, expression values were quantile normalized and $\log_2$-transformed. However, for maintaining overall expression levels, z-score was not applied to the 14-dimensional expression series. For illustration, the set was further reduced to 344 genes of prominent temporal up-regulation with more than 10 transitions $x^j_t < x^j_{t+1}$.

Neural gas has been run with Euclidean update $U(\mathbf{x},\mathbf{w}) = (\mathbf{x}-\mathbf{w})$ and with updates based on the derivative of correlation according to Eqn. 3. Both approaches have been re-run 50 times with random centroid initialization. Each run has been carried out with 100 update iterations using $\gamma = 0.001$ for the approach Euclidean and $\gamma = 0.01$ for the correlation-based one. Neighborhood size σ does not have any influence and even d is not important for data assignments, because there is only one centroid to be assigned to. Thus, only the effect of the derivative of d on the centroid specialization is studied here.

The results are displayed in the plot panel of Fig. 1. The plots show the 14-dimensional expression series together with their centroids, projected by PCA and embedded by multi-dimensional scaling (MDS) in two dimensions. PCA represents the Euclidean view on the data, MDS the correlation-based view. To summarize the displayed results, Euclidean update is very stringent in both data views, the top left panel indicating that all 50 centroids are almost perfectly located in the center of gravity at point (0,0), which is the k-means solution for $k = 1$. Complementary to that, correlation-based update exhibits many degrees of freedom in Euclidean view, but shows very high specificity in the correlation view – which is exactly what is has been designed for.

In addition to visual validation, which might suffer from shortcomings of the built-in dimension reduction, quantization errors have been calculated. For the average data vector, analog to the deterministic k-means result with $k = 1$, an average correlation of $\mathsf{r} = 0.96226$ to the data vectors is found. The Euclidean NG-update yields a result with an average correlation of the generated centroids of $\mathsf{r} = 0.96222 \pm 5.583 \cdot 10^{-5}$, which is virtually the result of the avarage vector, affected by minor update-specific fluctuations. Correlation-based centroid update yields the best results with an average correlation of $\mathsf{r} = 0.96403 \pm 8.173 \cdot 10^{-5}$. In combination with the bottom left panel in Fig. 1 it can be concluded that there are non-unique solutions that can be reached only, if Euclidean constraints are relaxed to updates operating in correlation space. Despite of the small differences for the presented data set, the results are quite fundamental, because they show that better solutions exist beyond averages. On a good mathematical basis, similarity-specific updates induce less constraints on the cost function and yield better data representations.

3.2 Clustering of Gene Expression Data

Mining for principal shapes in large lists of gene expression patterns is a central tool for the identification of co-expressed genes. Neural gas with correlation is used to meet this purpose for the data set described in the last paragraph containing 4824 gene expression levels at 14 time points. For comparison, Eisen’s implementation of k-means and Gasch’s and Eisen’s fuzzy k-means are taken as reference models (de Hoon et al., 2004). Both make use of Pearson correlation for creating sets of similar patterns for centroid calculation, but they compute centroid positions in Euclidean space. Calculations were done with 100 cycles for neural gas, i.e. 482,400 centroid updates, and 100 cycles for the k-means models.

A number of 23 centroids was used in all models, because fuzzy k-means is, due to its built-in PCA, limited to 3×#experiments+2=$3\times 14 + 2 = 44$ prototypes of which only 23 were identified as unique by fuzzy k-means (Gasch and Eisen, 2002). Contrary to the k-means methods, unused prototypes do not occur in NG-C, because of its neighborhood cooperation. The exponential NG-C neighborhood influence is realized as exponential decay from $\sigma = 23$ to $\sigma = 0.001$,

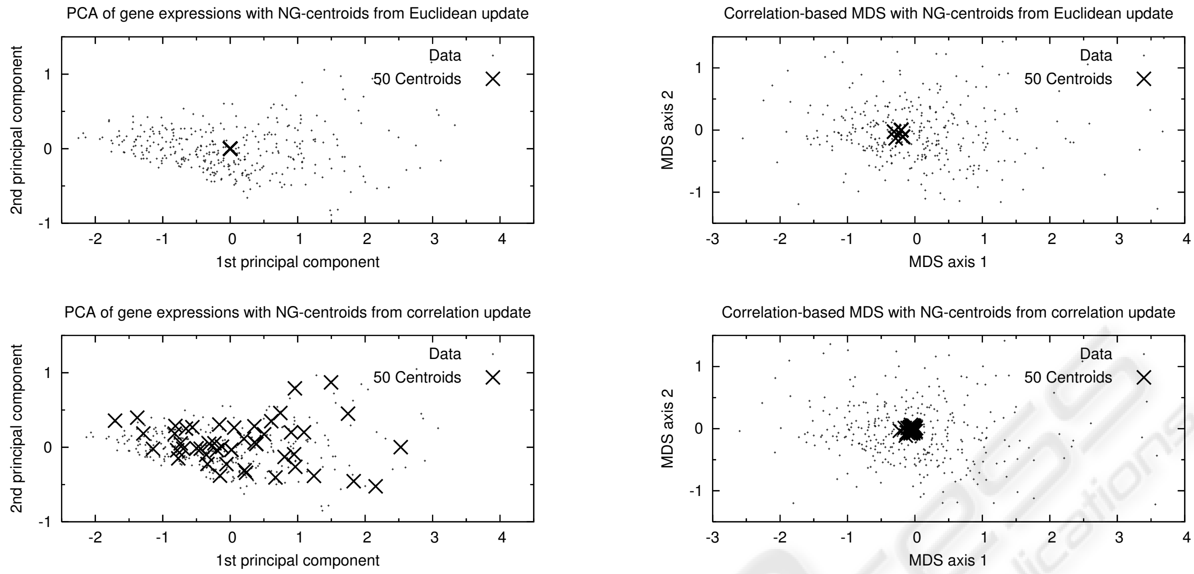

Figure 1: Centroid specialization for single cluster representation. Top row: Euclidean update rule, bottom row: update based on derivative of Pearson correlation. Left column: principal component plots, right column: multidimensional scaling of $(1-r)$ data relationships. In 50 individual NG runs, Euclidean updates (top row) show high specificity for both the Euclidean data view (shown as PCA) and the correlation-based view (shown as MDS). Correlation-based updates (bottom row) exhibit large diversity in Euclidean view (PCA) and high specificity in correlation view (MDS).

the update rate is set to $\gamma = 0.001$. Two quality criteria are considered for model comparison: reproducibility of the obtained centroids for different runs of the algorithms and quantization accuracy.

Reproducibility of Clusters. One major aspect of clustering is the consistency of the results. This has been tested by running NG-C and k-means 10 times from random starting configurations of the 23 centroids. For fuzzy k-means the standard initialization is fixed, which makes repeats unnecessary. Visual comparison is thus restricted to NG-C and k-means. An informative comparison between both methods is displayed in Figure 2, created using the free TreeView software. Both horizontal intensity bars contain the 23 centroids of 10 runs, i.e. 230 columns. Shades of gray denote specific gene expression intensities. Patterns of temporal up- and down-regulations present in the underlying data set are nicely captured by centroids of both models. The tic marks attached to the bottom of the NG-C bar point out 23 prominent bands that reflect a high reproducibility of the centroids contained therein, independent of their random initialization. For k-means, displayed in the row above, the result is very different: an unspecific continuous range of final states is obtained, which supports the experience of many users of k-means who complain about the poor reproducibility of results.

Quantization Accuracy. Table 1 provides a summary of the quantization accuracy of the found clusters. For each run, the average correlation of expression patterns with their corresponding centroids are measured, and the respective standard deviations are also calculated. These two values are averaged over all centroids. Finally, mean values for the 10 experiments are determined and listed in Table 1. As a major outcome, NG-C shows a superior data representation over k-means and fuzzy k-means. The fuzzy k-means is a little better than simple k-means, but its major disadvantage is the limitation to 44 centroids of which 21 are even unused. The good results of NG-C, however, are not too much surprising, because neural gas has been mathematically designed to optimize the goal of

Table 1: Average correlations between data samples and their centroids for 10 independent runs of NG-C and k-means. The deterministic result of the fuzzy k-means is 0.9335 ± 0.07216. In terms of quantization accuracy and data assignment variability NG-C performs best. Both k-means and its fuzzy k-means yield slightly worse quantizations, but fuzzy k-means covers data more homogeneously.

NG-C	
mean	std.-dev.
0.9516±0.0001	0.0573±0.0004

k-means	
mean	std.-dev.
0.9329±0.0017	0.0881±0.0038

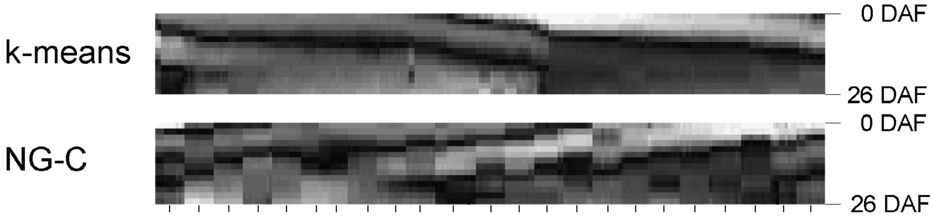

Figure 2: Cluster reproducibility for correlation-based neural gas (NG-C) and k-means. Both models, corresponding to the upper and lower bar, were run 10 times with random initialization. For the considered number of 23 centroids this yields a total of 230 centroids (gray columns) for comparison. While the final states of k-means cover a continuum of centroid locations, exhibiting only one major break, the final states of NG-C are highly conserved and displayed as 23 characteristic bands containing approx. 10 specific centroids.

maximum quantization accuracy (minimum quantization error), while the k-means methods are acting on assumptions about Euclidean data centers iteratively assessed by averaging.

3.3 Clustering of Mass Spectra

The last application concerns clustering of mass spectrum data from a clinical cancer study with 1050 mass spectra taken from sliced frozen tissue probes, using a linear MALDI-TOF MS, Autoflex, in a range of 2000-10000Da (by courtesy of Bruker Daltonik GmbH, Bremen). The data preparation protocol of the measured spectra followed the default workflow for baseline correction, alignment and peak picking. Robust peaks with signal to noise ratio $S/N > 5$ were used for further analysis, and only maxima of the extracted peaks were considered. This led to a high quality data set of 1050 samples, each described by 32 peaks. Clustering assists in tasks of data inspection and hypothesis generation.

Neural gas is applied in two manners to address the task of deriving tissue-specific spectrum centroids from the 32-dimensional data: one with Pearson correlation for centroids assignment, but with Euclidean update, the other fully correlation-driven for both pattern matching and update.

A small number of 11 clusters has been chosen in order to force sparse representations and to make the constraints of stringent Euclidean updates apparent. Both approaches have been trained in 10 independent runs using 1000 data cycles, i.e. 1000 x 1050 iterations, starting with randomly initialized centroids. Euclidean update was performed with an update rate of $\gamma = 0.01$. A value of $\gamma = 10^4$ was used for the correlation-based update. This large value compensates for the very small variability of the derivatives of correlations, which are caused by very similar mass peak profiles.

Both methods yield accurate data abstractions, as shown in Fig. 3. The MDS visualization faithfully displays the correlation relationships of the 32-dimensional centroids and the data. Since similar scatter points correspond to highly correlated data vectors, excellent reproducibility of the final configurations and a good data coverage can be observed. With respect to quantization, centroids from Euclidean update correlate on average at a level of $r = 92.8106 \pm 0.0043$ with the represented data. Update by Pearson correlation yields an improvement to $r = 93.4854 \pm 0.0790$ for the same number of prototypes. The small standard deviation for Euclidean update again points out (indirectly) the very strong attraction to the final centroid configuration, which is, however, not optimum in terms of quantization accuracy (data representation), for which the correlation update is clearly a better choice.

4 CONCLUSIONS

Based on the mathematical derivative of the Pearson correlation coefficient, we developed a new approach to maximize correlation in prototype-based data models. Particularly, the derivative can be directly plugged into the update step of a generalized version of the neural gas clustering method. Well-reproducible high-quality clusters were obtained by the new NG-C method. For the data clustered here, k-means and fuzzy k-means, although offering correlation similarity, are clearly outperformed by NG-C. In general, correlation-based centroid matching combined with Euclidean update, as usually realized in k-means and SOM implementations, leads to suboptimal data representations.

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 other suitable similarity measures. This opens directions to process

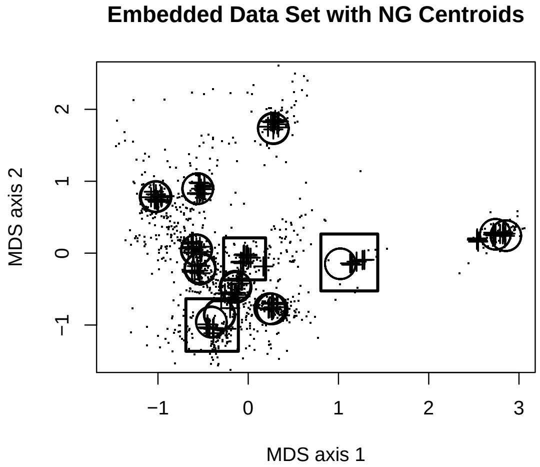

Figure 3: Visualization of data (small dots) and neural gas (NG) centroids (circles and crosses). Centroids correspond to 10 independent runs of NG, each run comprising 11 centroids, for two different update rules. Circles correspond to NG centroids obtained by Euclidean update; they do exhibit an extremely high reproducibility. Crosses correspond to centroids with correlation-based update; their final states are less stringently fixed, but their quantization quality is better (see text). In many cases, both update rules yield similar final configurations, but the boxes highlight regions with sytematic differences.

data from wide scientific fields where domain knowledge needs to be carefully considered.

ACKNOWLEDGEMENTS

Thanks to Frank-Michael Schleif (University of Leipzig) for the mass spectrum data and for fruitful discussions. This work is supported by the Ministry of Culture of Saxony-Anhalt, grant XP3624HP/0606T.

REFERENCES

Cottrell, M., Fort, J., and Pagès, G. (1994). Two or three things that we know about the Kohonen algorithm. In Verleysen, M., editor, *European Symposium on Artificial Neural Networks (ESANN)*, pages 235–244. D-facto Publications.

Cottrell, M., Hammer, B., Hasenfuss, A., and Villmann, T. (2006). Batch and median neural gas. *Neural Networks*, 19(6–7):762–771.

de Hoon, M., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. *Bioinformatics*, 20(9):1453–1454.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868.

Gasch, A. and Eisen, M. (2002). Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11):1–22.

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

Kohonen, T. (2001). *Self-Organizing Maps*. Springer-Verlag, Berlin, 3rd edition.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. University of California Press.

Martinetz, T., Berkovich, S., and Schulten, K. (1993). “Neural-gas” network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569.

Martinetz, T. and Schulten, K. (1991). A “neural-gas” network learns topologies. *Artificial Neural Networks*, pages 397–402.

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

Villmann, T. and Claussen, J. (2006). Magnification control in self-organizing maps and neural gas. *Neural Computation*, 18(2):446–469.

Zhou, X., Kao, M.-C., and Wong, W. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. *PNAS*, 99(20):12783–12788.