Using Self-Organizing Maps for object classification in Epo image analysis

Dorothea Heiss-Czedik, Ivan Bajla

Dept. of High Performance Image Processing
ARC Seibersdorf research GmbH, A-2444 Seibersdorf, Austria
e-mail: dorothea.heiss@arcs.ac.at

Abstract. *Erythropoietin (Epo) is a hormone which can be misused for doping. The detection of its recombinant form (rEpo) involves analysis of Epo chemiluminescence images containing bands. Within a research project, granted by the World Anti-Doping Agency, we are developing the GASepo software to serve for Epo testing. For detection of the bands we have developed a segmentation procedure. Whereas all true bands are properly segmented, a relatively high number of artifacts is generated. The goal is therefore to separate the artifacts from the bands. In the paper an alternative classification method, based on self-organizing map, is proposed to solve the task of separation. The method performs well, when compared with other classification methods. In addition, it provides valuable insight into the properties of the data, their dependencies and their relevance for the classification task.*

Keywords: Epo doping control, image segmentation, self-organizing map, classification

1. Introduction

The peptide hormone recombinant Erythropoietin (rEpo) can be misused as a doping agent, because it increases the oxygen-carrying capacity of the blood. Gas-chromatography and mass-spectrometry are not capable to clearly prove rEpo application. To accomplish this task, a methodology of isoelectric focusing (IEF) in electrophoretic gels is used [1]. Briefly, the IEF involves separation of sample proteins in a polyacrylamide gel, their transfer to a thin support membrane and detection by chemiluminescence image, containing a typical pattern of lanes (vertical stripes). As can be seen in Fig.1, the lanes comprise bands (deposits of individual protein glycoforms), which have been separated by pH gradient in the gel. When a sample containing known rEpo is subjected to IEF, some bands at different pH positions appear in the upper part of the gel (the first standard lane in Fig.1). When urine with natural Epo is subjected to the same process, the spots are concentrated in the middle part of the lane (lanes 5 to 8). In a positive doping case (3), these bands substantially overlap the region of those bands which belong to rEpo. The doping positivity criterion is based on measuring the individual characteristics of overlapping bands in reference to the cut-off-line and the bands in the standard lane.

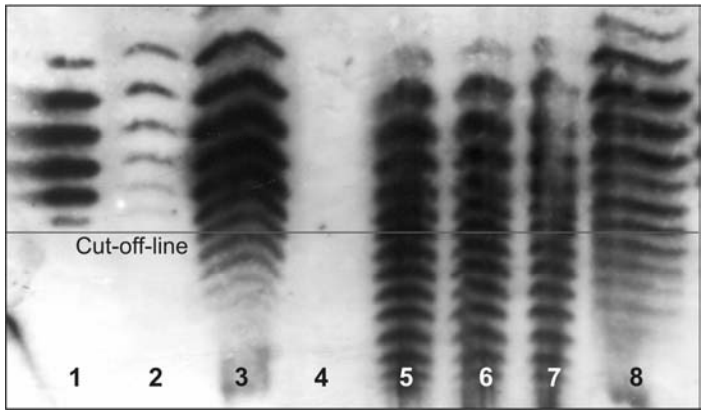

Fig.1 Typical Epo image with standard lane (1) and sample lanes.

To enable this measurement, a segmentation of bands in Epo images is required. We have developed a segmentation method [2] which consists of a sequence of operations specifically tailored to the following crucial problems of band degradation: (i) blurred bands, (ii) bands which are merged into one blob object, and (iii) bands which are represented by separate individual objects. The segmentation method properly segments all true bands. However, it generates a relatively high number of false positives (artifacts). The goal is therefore to separate the artifacts from the bands using a classification procedure and to gain more insight into the data, the object properties and their dependencies. Self-organizing maps (SOMs) [3] seem to be a suitable classification method to achieve this goal.

2. Self-organizing maps

Biological Background

The original motivation behind the development of self-organizing neural nets has been the modeling of basic information processes in the cortex as they are known from neuro-physiological experiments. The synaptic connections among the neurons are built depending on the type and frequency of sensorial stimuli. During this process individual neurons (or neuron groups) become sensitive to specific patterns occurring in sensorial signals. In particular, neighboring neurons always "learn" similar signal patterns. The abstract relation between the input (sensorial) signal and the synaptic adaptation of neurons was mathematically described by T. Kohonen in 1982 [3]. His "learning rule" (now called "Kohonen Algorithm") is surprisingly simple.

Theory of Kohonen Nets

Kohonen nets are artificial neural networks which adapt themselves in response to input signals and on the basis of the Kohonen algorithm. They consist of nodes, which are spread uniformly on a grid and are "functionally" connected to their neighboring nodes. In most cases, two dimensional grids are used, however, Kohonen nets with one-dimensional or multi-dimensional grids are possible as well. A Kohonen net can be viewed as an elastic membrane having an inner tension and moving freely in data space. It becomes "attracted" to the nearest data vectors. As a result, the net (which is initially placed on a plane in the corresponding data space) bends during the training process and finally finds an average position. During the training process, all data vectors are repeatedly presented to the net. The aim is to approximate the grid of the nodes to the data distribution as closely as possible and to correctly model the data distribution.

At the beginning of the training process the tension of the net is high. This means that the net covers a small portion of the data space. At the final stages of the training process the tension is lower and, thus, the net stretches towards the data clusters. The final locations of the nodes correspond to the data distribution in the data space. In other words, many nodes can be found in regions with high data density.

Mathematical description

This self-organization process can be described in mathematical form. The input consists of a sample of n-dimensional data vectors, each of the form

$$x(t) = [x_1(t), x_2(t), ..., x_n(t)] ,$$

where t is regarded as the index of the data vectors in the sample and also the index of the iterations ($t = 1, 2, ..., T$).

The goal of the training process is to determine the n-dimensional weight vectors of the nodes (neurons)

$$m_i(T) = [m_{i1}(T), m_{i2}(T), ..., m_{in}(T)] ,$$

where *i* denotes the index of the node in the self-organizing map (*i = 1, 2, ..., I*) and T is the final iteration of the training process. *I*, the number of nodes, is determined empirically.

For each intermediate iteration *t*, the training process performs the following steps:

1. The best matching node $m_c(t)$ most closely resembling the current data vector $x(t)$ is selected, for which the following is true:

$$\|x(t) - m_c(t)\| = \min_i \{\|x(t) - m_i(t)\|\} .$$

2. The nodes m_i are updated, using the formula:

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}(t)[x(t) - m_i(t)] ,$$

where the adjustment is monotonically decreasing with the number of iterations. This is controlled by the learning rate factor $\alpha(t)$ ($0 < \alpha(t) < 1$), which is usually defined as a linearly decreasing function over the iterations. The neighbors of the best matching node are also adjusted, but the adjustment is decreasing as the distance from the best matching node in the grid increases. This adjustment is determined by the neighborhood function $h_{ci}(t)$.

The neighborhood function often has the Gaussian form

$$h_{ij}(t) = \exp\left[-\frac{\|r_i - r_j\|^2}{2\sigma(t)^2} \right] ,$$

where r_i denotes the place of this node in the map and $\sigma(t)$ is some monotonically decreasing function over the iterations. Sometimes a simpler form of the neighborhood function is used, e.g. the bubble function which just denotes a fixed set of nodes around the best matching node (in the map). The Gaussian form ensures the global best ordering of the map [3].

3. An alternative classification using self-organizing maps

As mentioned above, the segmented objects of the Epo images need to be classified, to separate the artifacts from the bands. The self-organizing maps (SOMs) are trained to represent the distribution of the input data which are constituted by vectors of property values of the segmented objects. The new representation implicitly or explicitly shows relationships in the data, thus attributing to a greater knowledge of the underlying domain. SOMs are an example of unsupervised learning neural networks: The network is trained without taking a target value into account. Nevertheless, SOMs can also be used for classification. In contrast to [5], we do not use the class membership as a training variable. We try to classify the objects as bands or artifacts on the basis of some properties.

Quantitative properties of segmented objects

We have chosen the following eight geometrical and shape properties of the segmented objects, which might be useful for the classification: five common properties for binary images and three additional ones characterizing specific nature of the bands in the Epo images. The definition of all measures requires certain terms to be defined first.

BoundingBox	is the smallest rectangle containing the segmented object (shortly: object).
x_width, *y_width*	are the sizes of the *BoundingBox*.
LaneArea	is the number of all pixels in the given lane image.
Area	is the number of pixels belonging to the given object.
Perimeter	is the number of boundary pixels of the given object.

Based on these terms the following properties of the objects are defined:

Relative Area	***RelArea*** = *Area*/*LaneArea*.
BandBox Ratio	***BBRatio*** = *y_width*/*x_width*.
Eccentr*icity*	is the ratio of the distance between the foci and the major axis length of the ellipse with the same second moments as the object.
Orient*ation*	is the angle (in degrees) between the *x*-axis and the major axis of the ellipse that has the same second moments as the object.
Solidity	is the ratio of the pixels in the *BoundingBox* being also in the object.
Cent*roid* ***Ecc****entricity*	is the absolute distance of the vertical axis of the given lane from the vertical axis of the given object.
ObjectBoundary Complexity	***BndComx*** = 2 * (*x_width* + *y_width*)/*Perimeter*.
BandBox Fullness	***BBFulln*** = *Area*/(*x_width* * *y_width*).

An expert manual segmentation of Epo images from various doping control laboratories was carried out. To generate a training set for our classification method, we have calculated the values of eight measures (properties) for each segmented object.

SOM ordering and visualization

The 506 objects of the training set have been ordered by SOMs using the software package Viscovery SOMine 4.0 [6] because of its visualization capabilities. The output of the algorithm consists of the map, which is a flattened representation of the grid in the data space.

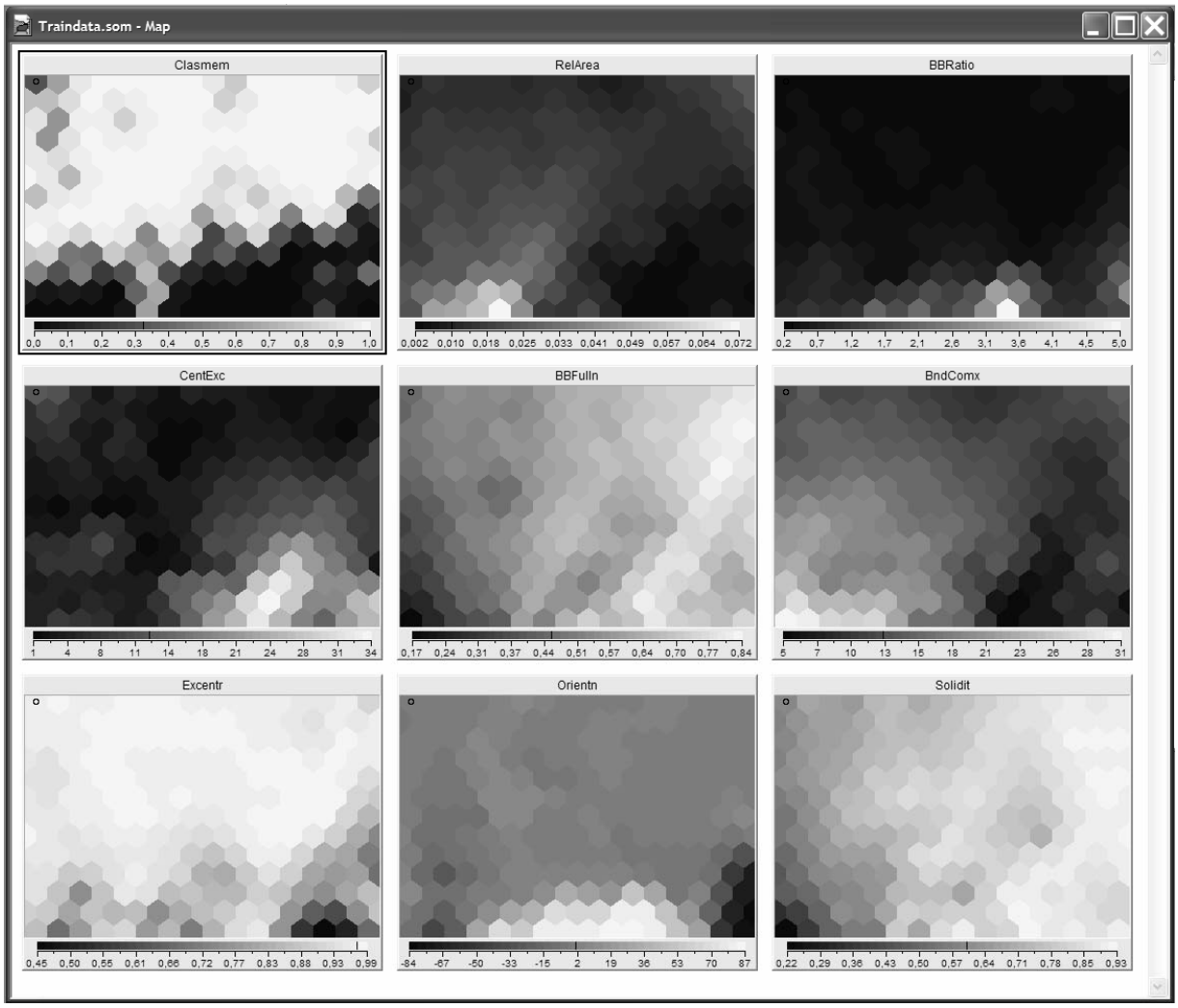

Fig.2 SOM trained with the training set of segmented Epo objects.

To visualize this map, several pictures are shown, one for each property, and one extra picture for the class membership (where 1 means band and 0 means artifact). Each node appears at

the same location in all property pictures. A property gray-level picture represents the (local average) property value at each node. The gray scale below the picture shows the relationship between intensities and property values. For example, in the property picture of ***Excentr*** (Fig. 2) we see that nodes with high eccentricity (0.99) are white and those with low eccentricity (0.45) are black.

4. Results

Interpretation of the SOM

To find out the dependencies or correlations between properties, it is sufficient to inspect the corresponding property pictures: highly correlated properties will have very similar property pictures. We are especially interested in the property ***Clasmem*** (class membership): the upper part of the map consists mostly of bands, whereas the lower part contains mostly artifacts. This means, that the bands and artifacts were separated quite well, although the class membership information was not used during the training of the SOM. This proves that at least some of the eight properties used provide effective information for the classification task at hand. To find out, which properties are useful, we inspect the property pictures. We see that ***BBRatio***, ***Excentr*** and ***Orientn*** have all strong dependencies with the class membership: ***Excentr*** is correlated with ***Clasmem***, ***BBRatio*** is anti-correlated with ***Clasmem***, and very high values of ***Orientn*** as well as very low values of ***Orientn*** correspond with ***Clasmem*** = 0 (artifacts).

Testing the SOM classification and comparison of the results

Apart from providing understanding of the data, SOMs can be used for classification of the objects in bands and artifacts: The trained SOM is the model and each object of the testing set is matched into the SOM, i.e. the object is associated to its best matching node (the nearest node which is most similar to the object). Then the ***Clasmem*** value of this node is taken, which is a weighted average of the ***Clasmem*** values of the objects of this node in the training set and also in the neighboring nodes. To determine the class membership of the object in the testing set, we simply round the ***Clasmem*** value of its best matching node. A value ≥ 0.5 indicates bands, a value <0.5 indicates artifacts.

The SOM-based classifier has been compared to some other classifiers:

- nonparametric statistical classifier based on weighted ranks (WR)
- Fisher linear classifier (FLC)
- k-nearest neighbor decision rule (kNN)
- Logistic regression (LogReg)
- neural network approach using two-layer perceptron (NNC)
- fuzzy decision tree (FDT).

To ensure a unified basis for the comparison, all classifiers used the identical training set in the learning process. Then they were applied to classification of the testing set (338 objects) which was identical for all methods. The conditional probabilities of misclassification have been used as characteristics of the classification quality. The measures (i) $Pr(A|B)$ - misclassification of bands B as artifacts A, given as the ratio of the number of misclassified bands and number of all bands in the testing set, (ii) $Pr(B|A)$ - misclassification of artifacts A as bands B, given as the ratio of misclassified artifacts and number of all artifacts in the testing set, and finally, (iii) $Pr(Err)$ - total misclassification error, that is defined as the ratio of the number of classified objects and number of all objects in the testing set. The following

table of the misclassification errors calculated for the different classifiers shows that the SOM-based classifier performs better than average.

Method	**SOM**	**WR**	**FLC**	**5NN**	**LogReg**	**NNC**	**FDT**
Pr(A|B)	0.0494	0.0330	0.0950	0.0210	0.0329	0.0530	0.0247
Pr(B|A)	0.2421	0.1370	0.1580	0.3260	0.3368	0.2650	0.3684
Pr(Err)	0.1036	0.0620	0.1120	0.1070	0.1183	0.1124	0.1213

5. Conclusions

Although the SOM-based classifier did not use the class membership as a property for the model, the results are comparable or even better than the results of such well-known methods as logistic regression, Fisher linear classifier or logistic regression. A reason is probably the non-linear behavior of this classification problem. Methods, which rely on a simple correlation dependency of a property with the target value (like: the higher the property value, the higher the target value) cannot model this sort of non-proportional dependency, where, for example, all extreme values (high or low) cause a certain target value (in this case, an artifact). SOMs divide the data space in many different homogeneous sub-spaces (i.e. nodes). Each node can be seen as a model for the corresponding data sub-space, and all these models together can predict a target value more accurately than one model alone.

References

[1] Lasne, F. and de Ceaurriz, J.: Recombinant erythropoietin in urine. Nature (2000), 405, 635.

[2] Ramoser, H., Biber, J., Bajla, I., Holländer, I.: Segmentation of Electrophoretic Images in Doping Control. In: Proceedings of the Intern. Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, METMBS'04, 2004, p.467-470, Las Vegas, USA, 15-16 June, CSREA Press.

[3] Kohonen, T.: Self-Organizing Maps. 2nd ed. Springer, Heidelberg, Germany, 1997.

[4] Deboeck, G.: Visual explorations in finance with Self-Organizing Maps. Springer-Verlag, London, Great Britain, 1998.

[5] Tan, R., v.d.Berg, J., v.d.Bergh, W.: Credit rating classification using Self-Organizing Maps. In: Neural Networks in Business: Techniques and Applications. Smith, K., Gupta J., Editors, Idea Group Publishing, 2002.

[6] Eudaptics: Viscovery SOMine 4.0 User's Manual. www.eudaptics.com, 2002.