Applications of SOM magnification to data mining

ERZSÉBET MERÉNYI[1], ABHA JAIN[1], WILLIAM H. FARRAND[2]
[1]Electrical and Computer Engineering
Rice University, 6100 Main Street, Houston, Texas, 77005, U.S.A.
[2] Space Science Institute
Boulder, Colorado, U.S.A.

Abstract: - Magnification in Self-Organizing Maps refers to the functional relationship between the density of the SOM weights in input space, and the density of the input space. The explicit magnification control scheme proposed by Bauer, Der and Herrmann [1] in 1996 opened the possibility to achieve specific magnifications that have attractive properties for data mining. However, the theoretical support only extends to 1- and 2-dimensional data with independent dimensions. This paper studies the scope of validity of the magnification control approach in hope to justify its application to real, high-dimensional data, which do not fall in the categories supported by the theory. We show encouraging results on synthetic as well as on real data.

Key-Words: - Self-Organizing Map, map magnification, clustering, high-dimensional data, data mining, spectral image analysis

1 Magnification in SOMs

One theoretically interesting and powerful data analysis aspect of Self-Organizing Maps (SOMs) is the so called *map magnification* which refers to the power law that relates the density of weights in the input space Q(**w**) to the *pdf* P(**v**) of the input samples:

$$Q(\mathbf{w}) = c\, P(\mathbf{v})^{\alpha}$$

where α is the *magnification exponent* and c is a constant [1]. As shown by [1] a converged SOM with $\alpha = 1$ maximizes information theoretic entropy. $\alpha = d/(d+2)$ for d-dimensional data corresponds to minimum mean squared error quantization. $\alpha < 0$ enlarges response areas in the SOM for low-frequency inputs, which is potentially useful for making discoveries as it would enhance the detectability of rare classes. Kohonen's SOM algorithm (KSOM) [2] is known to achieve $\alpha = 2/3$ (under certain conditions) [3] which is optimal in neither minimum distortion nor maximum entropy sense. A SOM variant called Conscience algorithm [4] can effect $\alpha = 1$ but not any other value. Bauer, Der and Herrmann proposed a principled SOM algorithm (which we will call BDH in this paper) for the explicit control of the magnification exponent by using adaptively adjusted local learning rates [1]. This possibility is extremely attractive because one would have a turn-of-the-knob solution for obtaining various quantization properties, appropriate for a given data mining purpose. However, the theory only guarantees success for 1-dimensional input data and for 2-dimensional data whose components are statistically independent. Most real data do not obey these conditions, yet it is data mining of real data (and especially of high-dimensional data) that would benefit the most from explicit magnification control. Validations of the BDH algorithm were given on appropriate 1- and 2-dimensional simple data in [1]. We are not aware that the validity and power of this intriguing scheme has been gauged for real data.

2 Magnification control tests on data unsupported by the theory

Our objective is to investigate whether the BDH algorithm can be applied to data outside of the category for which the theory can guarantee success. As usual, in cases where analytical proof is not available one can run carefully constructed simulations to chart the scope of validity. Our two main interests are: maximum entropy (or equiprobabilistic) quantization ($\alpha = 1$) because in this case the weights of the converged SOM achieve the most faithful representation of the input data space; and negative magnification ($\alpha < 0$) because in this case low-frequency inputs (rare clusters) could be detected much easier than from a regular KSOM or from a maximum entropy map thus facilitating specific data mining pursuits.

2.1 BDH for low-D synthetic "forbidden data"

In a recent work [5] we verified that the BDH indeed achieves the above theoretically derived magnification exponents on data supported by the theory. Then we examined the effect of BDH magnification on relatively simple synthetic data that fall into the “forbidden” category in [1]: 2-dimensional data with weakly and with strongly correlated dimensions. On these, the magnification exponent α_{achieved} obtained by the SOM showed a systematic difference from the intended magnification exponent α_{intended} within the range of α that we explored (0.6 – 1.2). These results are stronger than the theory offers: even though we did not achieve the exact value of α_{intended} the experiments suggested that the value of α_{achieved} could be predicted from the value of α_{intended} used in the BDH.

2.2 BDH for >2-D synthetic data

Encouraged by the apparent wider scope of BDH than guaranteed in [1] we then ran simulations to observe the effect of negative magnification on 6-dimensional data with small (5) and larger number (20) of known classes, including extremely rare clusters. Figure 1 shows the results obtained in [5] for the 20-class data consisting of 6-dimensional inputs. The class highlighted in magenta has 1 data vector in it, the aqua class contains 16 input vectors. All other classes have approximately 1024 data vectors.

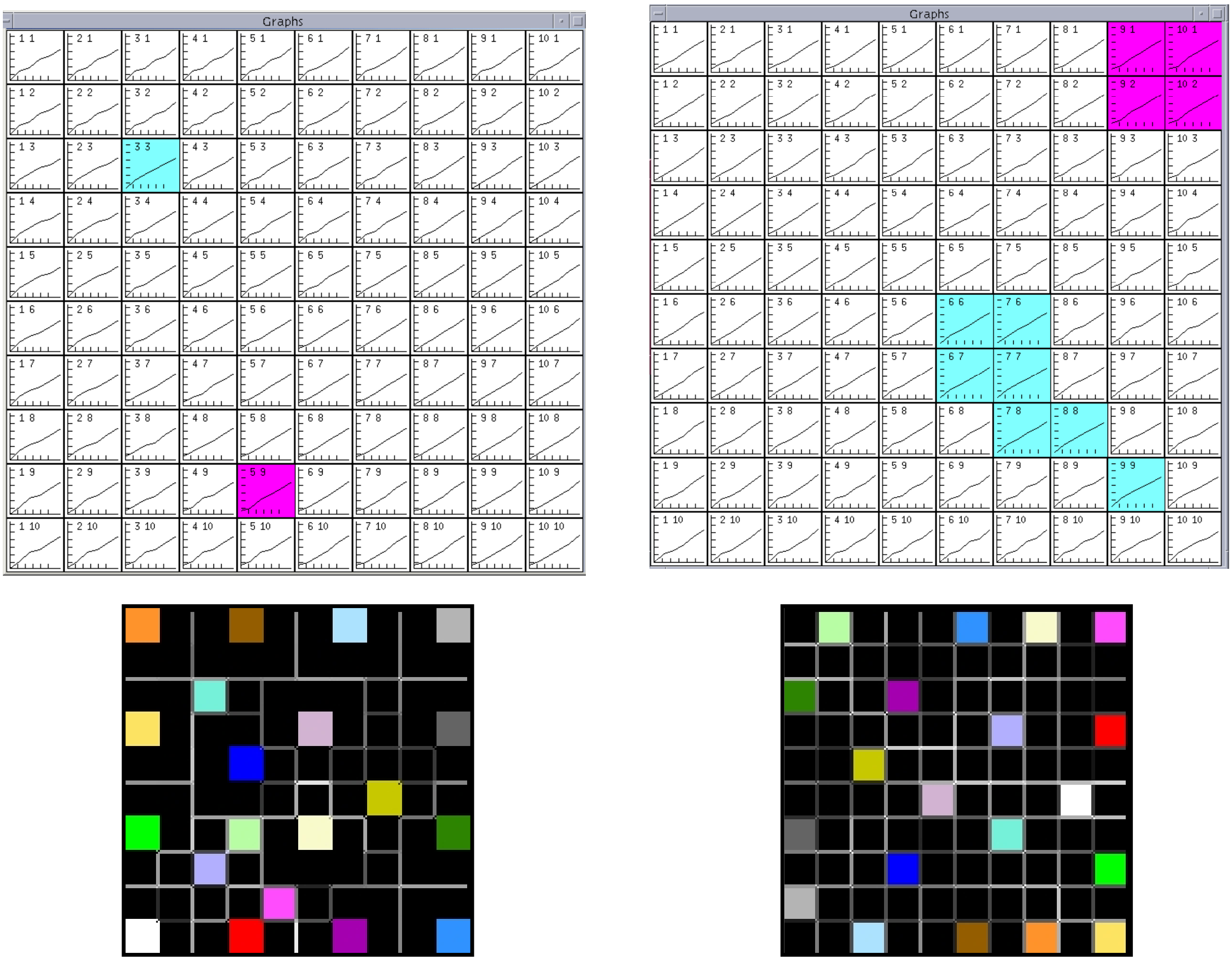

Fig. 1. *Clustering a 20-class data set (from [5]).* ***Left:*** *Using KSOM. Top: Weight vectors in the 10 x 10 SOM map. Only 1 node represents each of the rare classes, the 1-element class highlighted in magenta, and the 16-element class shown in aqua. Bottom: Clusters identified in the SOM by visualizing the distances of weights of adjacent nodes as grey scale “fences”. Lighter fence means larger difference between the weights of the corresponding nodes. The regular classes separate well. The 1-element magenta cluster mapped very close to the red class, their separation is weak. Likewise, the fences separating the aqua class from the dark blue class are less developed than fences among the non-rare classes. The separation between the purple (a 256-element moderately ‘rare’ cluster) and the light green class is well developed through the distance of their weights in the diagonal direction, in spite of the adjacency of the respective nodes.* ***Right:*** *Using BDH with α = -0.8 . Top: Weight vectors in the 10 x 10 SOM map. 4 and 7 nodes represent now the rare classes, respectively. Bottom: Clusters identified in the SOM. The magenta cluster now separates well from the red class. Original figure is in color. Paper can be downloaded from* http://www.ece.rice.edu/~erzsebet/papers/wseas-paper.pdf .

We note that evaluation of the achieved magnification exponent, similarly as was done in [6] is only possible for low-dimensional data because it involves estimation of both the *pdf* of the weights in input space and the *pdf* of the input data. Since the required number of data points for acceptable estimation grows exponentially with dimensionality, the *pdf* estimation becomes practically impossible for higher than 3-4 dimensions. Therefore, for higher dimensions we evaluate $\alpha_{achieved}$ by direct observation of the converged SOM.

Negative BDH magnification works well on fairly complex, albeit synthetic data. Figure 2 shows that maximum entropy quantization (α = 1) also works very well. We compared the performance of BDH with that of the Conscience algorithm on a known 8-class, 6-dimensional synthetic data set, where the sizes of classes were precisely known. The details are explained in the caption of Figure 2. This test was two-fold. First, it confirmed that the Conscience algorithm produced maximum entropy mapping (that the number of nodes that represent each known class is proportional to the size of the class). Second, we obtained the same result with BDH inducing α = 1, thus confirming that the BDH indeed achieved $\alpha_{achieved}$ = 1. To obtain $\alpha_{achieved}$ = 1 we used $\alpha_{intended}$ = 0.7 in the BDH based on the systematic difference we observed between $\alpha_{intended}$ and $\alpha_{achieved}$ in [5]. The comparison with the Conscience algorithm is a special case of great importance because it provides a way to evaluate the magnification for $\alpha_{achieved}$ =1.

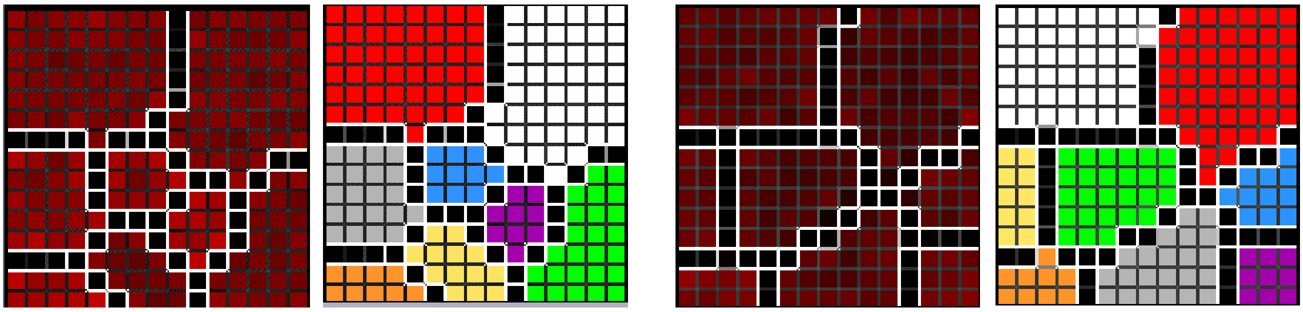

Fig. 2. *Comparison of the performance of Conscience, and BDH forcing α = 1. The synthetic data set consists of 6-dimensional patterns that fall into 8 known classes. The two panels on the left illustrate the clusters learned by the Conscience algorithm, and the two right panels illustrate the BDH clusters, respectively.* ***Left of each pair of panels****: The cluster boundaries, visualized as the distance of weights of adjacent nodes. White is high fence (large dissimilarity), black is low fence (great similarity). The SOM has a 15 x 15 rectangular grid. Each grid cell is shaded by an intensity of red proportional to the number of data points mapped to the node in that grid cell. Black grid cells between the strong fences indicate that the receptive fields of the corresponding nodes are empty.* ***Right of each pair of panels****: The known class labels, color coded and superimposed over the grid cells representing the nodes. Both representations show that the nodes (SOM weights) are divided among the classes in proportion to the sizes of the classes: A, B (red and white) contain 4096 data points each, C, O (green and grey) 2048, and D, H, I, and M have 1024 points. The corresponding number of designated weights are A:48, B:49, C:25, O:21, D:13, H:9, I:10, M:9. The deviations from the exact 4:2:1 proportions can be due to the small size of the SOM, integer arithmetic, and the formation of inter-cluster gaps. Original figure is in color. Paper posted at http://www.ece.rice.edu/~erzsebet/papers/wseas-paper.pdf .*

3 BDH on real, >2-D "forbidden data"

Naturally, the real interest is in the application of the BDH algorithm for real data, especially where it is desirable to increase the chance of discovery.

In [7] we showed a comparison between an SOM obtained by the Conscience algorithm, and an SOM obtained by BDH magnification with $\alpha_{intended}$ = -0.8, for an 8-band remote sensing spectral image of Ocean City, Maryland. This data set is outside of the category of data that the theory supports, on two counts: it is more than 2-dimensional, and because the pairwise correlations between image bands range anywhere from 0.5 to 0.95. Direct computation of the magnification achieved by the SOM was not possible (see [5]), but we could compare the areas occupied by the same clusters in the SOMs produced by Conscience and BDH. A supervised classification from an earlier study [7] that was sufficiently verified to be considered as "ground truth" (known labeling of image pixels) also served the evaluation. The Ocean City analysis showed beautifully that very small clusters gained more areal representation in the BDH SOM than in the Conscience SOM. In fact we discovered a new cluster (a roof type) that filled a very small area, left unclassified in the earlier supervised classification.

3.1 Rare clusters in spectral images of Mars

One of the most important things in the exploration of other planets is the discovery of new, surprising or rare materials. Spectral imagery is a frequently used type of data for mapping planetary surface composition, similarly to earth remote sensing. Encouraged by the success with terrestrial imagery we analyzed a spectral image of the Martian surface taken by the Imager for Mars Pathfinder (IMP) in 1997. This image is one of the so-called octants of the SuperPan (360 degree panorama) image obtained by the left eye of the IMP. (The IMP had a left and a right eye, with different but somewhat overlapping wavelength sets.) The left eye images consist of 8 bands taken at wavelengths from 0.44 to 1.001 microns. The spatial size is nearly 1,000 x 1,000 pixels, with a large area occupied by the lander's ramp. The Martian surface shows in about 600,000 pixels. We clustered this image in a previous work [8] with a Conscience SOM, and found known, very rare occurrences of a "black rock" type that is of mafic composition (fairly pristine, olivine and/or pyroxene rich) and of great interest from a geologic point of view. We also split the rare black rock type into two subtypes the spectral signatures of which are distinct: one subtype has pyroxene absorption at approximately 0.93 microns (consistent with orthopyroxene), the other has an absorption at a longer wavelength, around 1 micron (consistent with clinopyroxene or olivine). One subset of the clustered image is in Figure 3, with the inset enlarging an occurrence of both rare classes. The full image can be seen in [8], posted at www.ece.rice.edu/~erzsebet/publications.html .

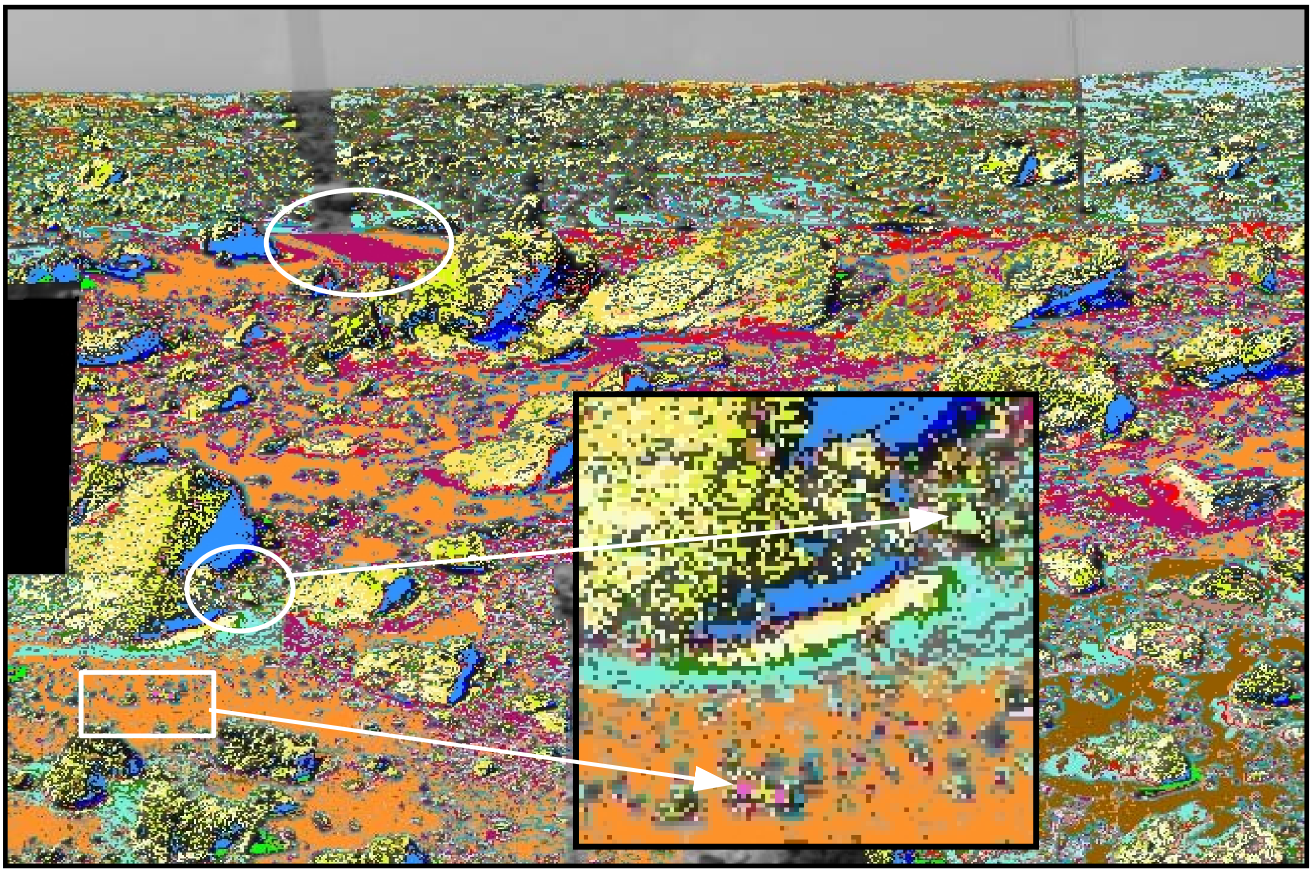

Fig. 3. *Clusters obtained with a Conscience SOM, from IMP SuperPan octant S0184 of the Martian surface at the Mars Pathfinder landing site. The full image and discussion of clusters is given in [8]. Here we point out two extremely rare clusters that represent geologically relevant materials: the pink class (label O in Figure 6) and the pale aqua class (label R in Figure 6), occupying tiny areas within the white rectangle and the white small oval, respectively, and which are enlarged and pointed at by arrows in the inset. These occurrences contain less than 50 pixels each.*

Figure 4 shows that the same clusters were found by BDH SOM clustering. (Since figures 1 through 5 are in color, this paper as well as others that we referenced for related color imagery are posted for easy access at http://www.ece.rice.edu/~erzsebet/publications.html .)

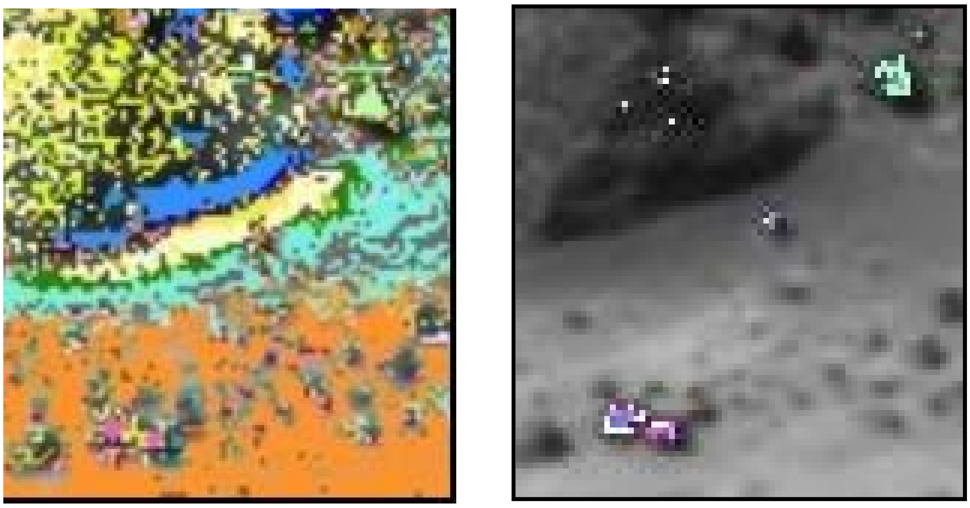

Fig. 4. ***Left:*** *The inset from Figure 3.* ***Right:*** *The same area showing the clusters found by BDH SOM clustering. For easier comparison only the rare clusters were highlighted in the BDH cluster map. The pink and aqua clusters match very well those in the Conscience SOM cluster map at left. In addition, there are further variations within each one. The pink class has distinct parts highlighted in white and blue and the aqua class also has a blue subarea. Note that blue and white here and in Figure 5 are not the same clusters as in Figure 3, these colors were recycled to make visible distinctions within the tiny pink and aqua clusters. As seen from Figure 6 the spectral signatures of these four clusters group into two types: O (pink) and O1 (white) are close to the original O signature from the Conscience mapping (Figure 6, left), while R is close to the original R signature. R1 has an absorption band around 1 micron, but deeper than in R.*

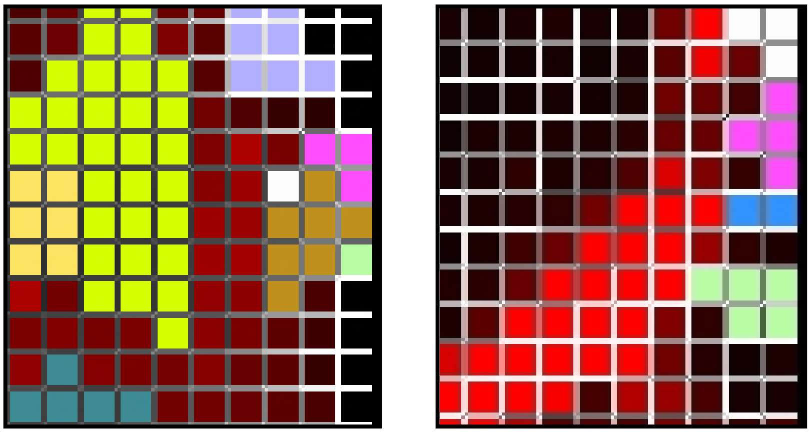

Fig. 5. ***Left:*** *Detail from the 40 x 40 Conscience SOM used in [8], showing the learned representation of the pink (label O on Figure 6, left) and aqua (label R on Figure 6, left) clusters.* ***Right:*** *Detail from the 40 x 40 BDH SOM showing the learned representation of the same two rare clusters. Notice that the separation between pink and aqua clusters is much more definite than in the Conscience SOM. The areal representation of the aqua class is much larger in the BDH SOM. The original O spectral type is now represented by the pink and white areas together, doubling the representation area. The R1 type (blue) is in between, closer to the pink cluster because of the similarity of the overall spectral shape, but separated by strong fences because the center of the absorption band is different from that of the the pink/white clusters.*

The corresponding mean spectral signatures in Figure 6 confirm these clusters. In the Conscience map the rare clusters did not receive enough spatial representation to make further distinctions within them. Significant mixing remained unresolved. With BDH magnification control, both the separation and distinction increased.

4 Conclusion

We showed that the explicit magnification control scheme of Bauer, Der and Herrmann [1] produced predictable behavior on some data for which the theory does not guarantee applicability. While this is very encouraging for data mining pursuits, one must proceed with caution: there are several aspects of the

BDH that we need to understand better in relation to "forbidden" data, for confident applications. We plan to perform further systematic investigations in order to determine the validity of the BDH for high-dimensional complex data and evaluate its benefits.

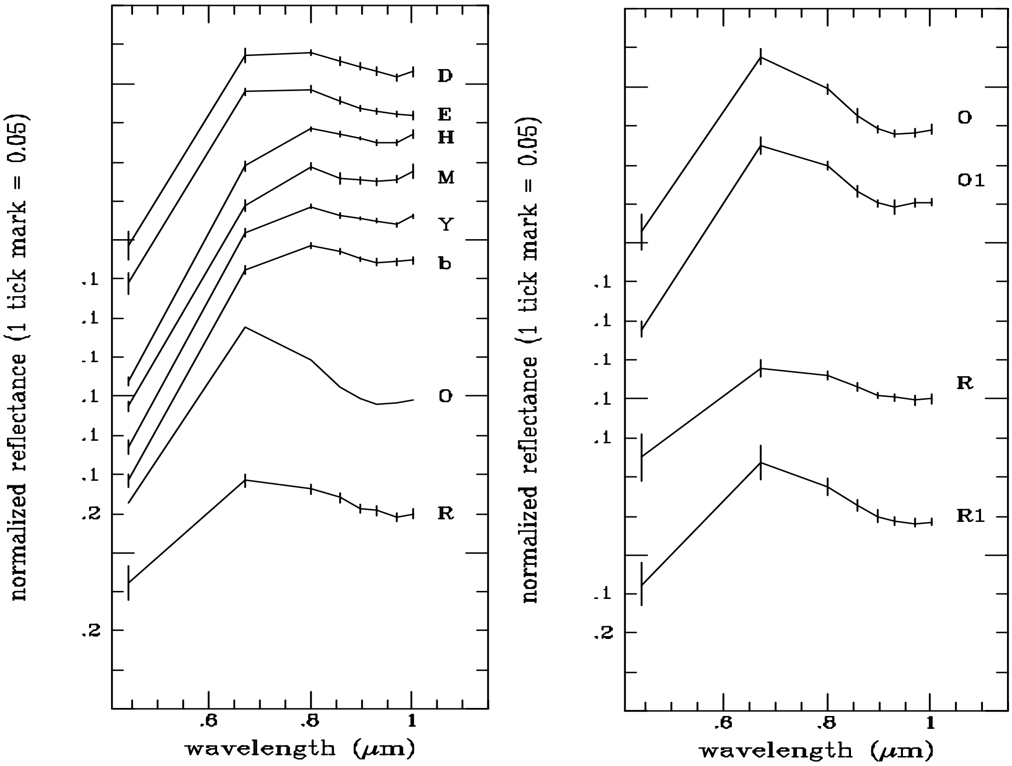

Fig. 6. *The mean spectral signatures of the clusters obtained with Conscience SOM (left) and with BDH SOM (right). (The left plot is from [8] and contains additional cluster signatures.)*

5 Acknowledgements

E. Merényi and A. Jain were partially supported by the Applied Information Systems Research Program of NASA, Office of Space Science, NAG9-10432. W. Farrand and E. Merényi was funded by the Mars Data Analysis Program of NASA, OSS, NAG5-13294, for work on the IMP SuperPan data. We thank Dr. Scott Murchie of the Johns Hopkins Applied Physics Lab for the SuperPan imagery.

References:

[1] Bauer, H.-U., Der, R.., and Herrmann, M., Controlling the magnification Factor in Self-Organizing Feature Maps, *Neural Computation*, Vol.8, No.4, 1996, pp. 757-771.

[2] Kohonen, T., The Self-Organizing Map. *Proc. IEEE*, vol. 78, 1996, pp 1464-1480.

[3] Ritter. H. and Schulten, K., On the Stationary State of Kohonen's Self-Organizing Sensory Mapping. *Biol. Cyb*., Vol. 54., 1986, pp 99-106.

[4] DeSieno, D. Adding a conscience to competitive learning. *Proc. IEEE Int'l Conf. On Neural Networks*, Vol I, 1988, pp 117-124.

[5] Jain, A., Merényi, E, Forbidden Magnification? I. *Proc. European Symposium on Artificial Neural Networks (ESANN'04) Bruges, Belgium*, 2004. in print. Also posted at URL www.ece.rice.edu/~erzsebet/papers/ESANN04-1.pdf

[6] Ritter, H. Asymptotic Level Density for a Class of Vector Quantization Processes. *IEEE Trans. Neural Networks*, Vol. 2, No. 1, 1991, pp 173 – 175.

[7] Merényi, E, Jain, A., Forbidden Magnification? II. *Proc. European Symposium on Artificial Neural Networks (ESANN'04) Bruges, Belgium*, 2004. in print. Paper also posted at URL www.ece.rice.edu/~erzsebet/papers/ESANN04-2.pdf

[8] Merényi, E, Farrand,. W.H., Tracadas, P.J. (2004) Mapping Surface Materials on Mars From Mars Pathfinder Spectral Images With HYPEREYE. *Proc. Int'l Conf. on Information Technology (ITCC 2004)*, April 5-7, 2004, Las Vegas, NV, USA. vol II, pp 607 – 614. URL: www.ece.rice.edu/~erzsebet/papers/paper-ICCT04-withhdr.pdf