Maps for the Visualization of high-dimensional Data Spaces

Alfred Ultsch
DataBionics Reseach Lab
Department of Computer Science
University of Marburg
D-35032 Marburg, Germany
ultsch@informatik.uni-marburg.de

Keywords: SOM, U-Matrix, Visualization, Clustering, Density Estimation, Projection, Data Mining, Knowledge Discovery

Abstract— The U-Matrix is a canonical tool for the display of distance structures in data space using emergent SOM (ESOM). The U-Matrix defined originally for planar map spaces is extended in this work to toroid neuron spaces. Embedding the neuron space in a finite but borderless space, such as a torus, avoids border effects of planar spaces. A planar display of a toroid map space disrupts, however, coherent U-Matrix structures. Tiling multiple instances of the U-Matrix solves this problem at the cost of multiple images of data points. The P-Matrix, as defined here, is a display of the density relationships in the data space using Pareto Density Estimation. While the P-Matrix is useful for clustering, it can also be used for a non-ambiguous display of a non planar neuron space. Centering the display for high density regions and removing ambiguous images of data points leads to U-Maps and P-Maps. U-Maps depict the distance structure of a data space as a borderless three dimensional landscape whose floor space is ordered according to the topology preserving features of ESOM. P-Maps display the density structures. Both maps are specially suited for data mining and knowledge discovery.

1 Introduction

In the SOM literature one can distinguish two different prototypical SOM usages. The first are SOM where the neurons are identified with clusters in the data space (k-means SOM). The second are SOM where the map space is regarded as a tool for the characterization of the otherwise inaccessible high dimensional data space. A characteristic of this SOM usage is the large number of neurons. Thousands or tens of thousand neurons are used. Such SOM allow the emergence of intrinsic structural features of the data space (ESOM)[2]. The U-Matrix is the canonical tool for the display of the distance structures of the input data on ESOM [3]. A U-Matrix is usually defined on a planar topology of the neuron space. Embedding the neuron space in a finite but borderless space such as a torus avoids the problems of borderline neurons. Such a map space has, however, the disadvantage that the display of a U-Matrix as a planar map disrupts coherent structures. In this paper we present an approach which solves this problem. The solution uses an estimation for data density, called Pareto Density Estimation which is aimed at the detection of clusters in data sets.

2 SOM Notation

It is hardly necessary to introduce Kohonen's SOM algorithm here. The interested reader is referred to [1]. To avoid misunderstandings we shortly present, however, our notation for the usage of a SOM:

data space: $D \subset R^n$: the subspace of R^n where data points of an application can be observed.

input data: $E = \{x_1, \ldots x_d\}$ with $x_i \in D$ the set of data presented to SOM leaning algorithm

data distance: distance measure defined in the data space $D \times D \rightarrow R^+$: $d_{xy} = d(x,y) \geq 0$, d_{ij} is a shorthand for $d(x_i, x_j)$

neurons:: $M=\{n_1, \ldots, n_k\}$ a set of neurons

weights: each neuron is associated with a (high-dimensional) weight vector $w_i = weight(n_i) \in D$.

weight space: $W = \{w_1, \ldots, w_n\}$

position of a neuron: each neuron n_i has a position, i.e a vector of coordinates $pos_i = pos(n_i) \in K$ in the map space K.

map space: $K \subset R^m$, $m \leq n$ a m-dimensional space with a distance measure $k: K \times K \rightarrow R^+$: $kij = k(pos(ni), pos(nj)) \geq 0$, and a neighborhood function N.

neighborhood function: a mapping $M \times M \times R^+ \rightarrow [-1\ 1]$,
$h_{ij}(r) = h(n_i, n_j, r)$ with the following properties:
$h(n_i, n_j, r) \geq h(n_s, n_j, r)\ \forall\ j \neq i$ with $k_{ij} > 0$ and $r > 0$
$h(n_i, n_j, r) = 0\ \forall\ n_j$ with $k_{ij} > r$ compact support (kernel)
r is called neighborhood radius

neighborhood: $N_i = N(n_i) = \{n_j \in M | h_{ij}(r) \neq 0\}$ the set of neurons with non vanishing neighborhood function h.
The neighborhood defines a lattice of neurons in the map space K

bestmatch: $D \rightarrow M$: $bm_i = bm(x_i)$ is the neuron $n_b \in M$ having the smallest data distance to x_i . I.e.:
$n_b = bm(x_i) \Leftrightarrow d(x_i, w_b) \leq d(x_i, w_j)\ \forall\ w_j \in W$.

SOM learning: when a input vector x_i is learned, the weight of a neuron

n_l is modified as follows. Let $\eta \in [0, 1]$ then

$$\Delta w_l = \eta \cdot h(bm_i, n_l, r) \cdot (x_i - w_l)$$

3 U-Matrix

A U-Matrix is originally defined on planar map spaces [3]. Examples of such map spaces are rectangular or hexagonal grids. The U-Matrix is calculated in the weight space and displayed using the map space. The *vicinity* U_i of a neuron n_i is the set $U_i = \{n_j \mid k(n_j,n_i) < u,\ n_j \neq n_i\}$ for some small positive constant u. I.e. a neuron's vicinity are the closest neighbors in the map space. The *U-height* of a neuron $uh(n_i)$ is the sum of all data distances from the weight of n_i to the weight vectors of the neurons in U_i

$$uh(n_i) = \sum_{n_j \in U_i} d(n_i, n_j) .$$

A visualization of all U-heights at the neuron's coordinates in an appropriate way gives the U-Matrix[3]. Typical visualizations are colored contour plots on top of the planar SOM floor (e.g. in [9]).

3.1 Properties of the U-Matrix

The U-matrix delivers a "landscape" of the distance relationships of the input data in the data space (compare figure 5). Properties of the U-Matrix are:

- the position of the bestmatches reflect the topology of the input space, this is inherited from the underlying SOM algorithm
- weight vectors of neurons with **large** U-heights are very distant from other vectors in the data space
- weight vectors of neurons with **small** U-heights are surrounded by other vectors in the data space
- bestmatches are typically found in depressions
- outliers in the input space are found in „funnels".
- "mountain ranges" on a U-Matrix point to cluster boundaries
- „valleys" on a U-Matrix point to cluster centers

The U-Matrix realizes the emergence of structural features of the distances within the data space. Outliers, as well as possible cluster structures can be recognized for high dimensional data spaces. The proper setting and functioning of the SOM algorithm on the input data can be visually checked.

U-Matrices have been used in a number of applications to detect new and meaningful knowledge in data sets. To name a few: sea level prediction [4], DNA–array analysis [5], customer segmentation in mobile phone markets [6], stock portfolio selection [7], and many more…

3.2 Borderless U-Matrix

There is, however, one problem associated with planar map spaces: seams. The neighborhood of neurons at the edges of a planar map space contains much less neurons compared to the middle of the map space. This leads to undesired seam effects in the SOM algorithm. A possible solution for this is to connect the edges of a planar map space to form a toroid map space. For rectangular grids this may be regarded as the "Pacman universe": to the right of the rightmost neurons are the leftmost neurons; to the top of the topmost neurons are the bottom most neurons. A planar display of such a toroid U-Matrix cuts, however, through structures that cross the planar borders.

We show this using an example data set, called "Hepta". Figure 1 shows this data set. Hepta set consists of 72 points in seven clusters of ten points

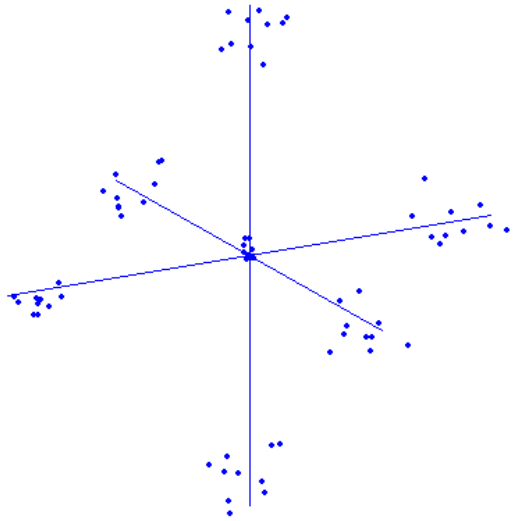

Figure 1: The Hepta data set

each, plus two additional points for the center cluster. The centroids of the clusters span the coordinate axes of the R^3. The density of the central cluster is about twice as big as the density of the other six clusters.

In our research group's U-Matrix tool, called Bionic Data Mine (BDM) we had implemented a method to depict a toroid U-Matrix: four adjoining pictures of the same U-Matrix. This is called a tiled display. In the center of this display emergent structures can be seen that would cross the edges of a simple planar display of the toroid U-Matrix.

Figure 2 shows a top view of a tiled display U-Matrix for Hepta on a toroid 64 by 64 SOM. Large U-heights are shown in white, small U-heights in black. At least in the center the structures that cross the border of a single display are visualized coherently. The limits of the planar display are added to demonstrate the disruption of coherent toroid structures.

With such a tiled display, however, U-Matrix structures are repeated and there is no obvious region to focus on. The tiled display has also the disadvantage, that each input data point is represented on multiple locations. This makes it difficult to grasp the intrinsic structure of the data set in particular for an inexperienced user.

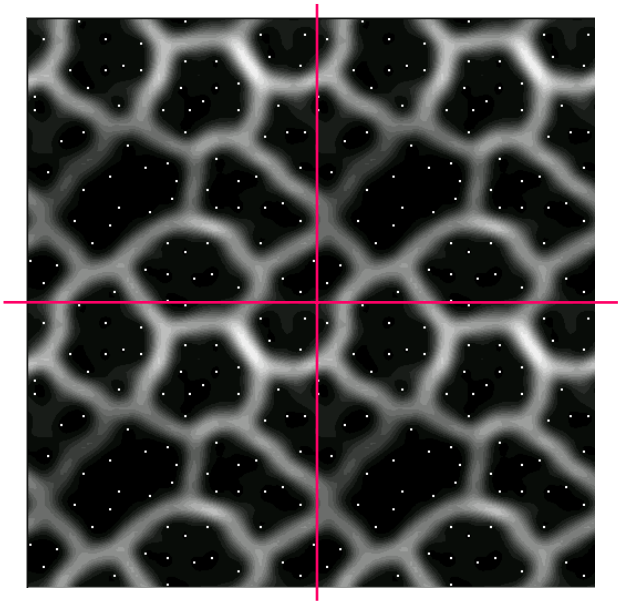

Figure 2: Tiled display of a U-matrix of Hepta on a toroid ESOM

4 Pareto Density Estimation

In the next chapter, we present a new visualization for structural features of data spaces using ESOM. This tool is called P-Matrix. For this, the definition of a density estimation method especially designed for the usage with ESOM.

Density estimation is the construction of an estimate of the probability density that generated the data. Within the context of SOM a distance measure in data space is given.. Data density may thus be estimated by dividing the number of points within a hypersphere by the volume of the hypersphere. Volume calculation is, however, problematic for high dimensional data due to the so called "curse of dimensionality"[16]. If the radius of the hypersphere is kept constant, the number of points included is proportional to the density. This type of density estimation is a special case of kernel density estimation using a fixed kernel bandwith[10]. Such uniform kernel estimates can approximate the true probability up to any desired degree of accuracy, if the true probability is known[11]. The choice of this radius is, however, critical. A too small radius overfits, a too large radius oversmoothes the density estimation.

We propose to use a radius that fulfills an optimality criterion with respect to information. Let S be a subset of a set of n points with $|S| = s$ the number of elements in S, then $p = s/n$ is the relative size of the set. If there is an equal probability that any point x is observed then $p = p(x \in S)$. Information theory defines the (partial) information I(S) of a set using p. Scaled to the range [0,1], the information of a set is calculated as $I(S) = - e\, p \ln(p)$ [13]. See figure 3 for a graph of I(S).

To find an optimal set size, define the unrealized potential URP(S) of a set as the Euclidian distance from the ideal point, i.e. an empty set producing 100% of information. This definition of URP(S) gives: $URP(S) = \sqrt{p^2 + (1 + e \ln(p)^2)}$ [8]. URP(S) can be seen in figure 3 as the length of the line starting at point (0,1) and ending at I(S).

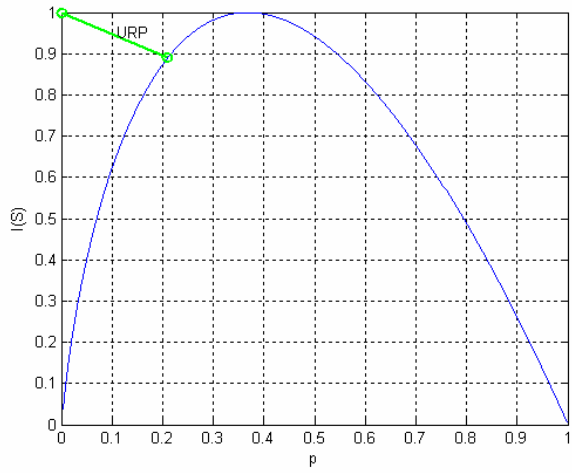

Figure 3: Information I(S) and unrealized potential URP(S)

Minimizing the unrealized potential results in an optimal set size of p_u = 20.13%. This set size produces 88% of the maximum information. The optimality of this set at about (20%, 80%) might be the reason behind the so called Pareto 80/20 law, which is empirically found in many domains [14]. Therefore density estimation using volumes with an expected average of p_u points are called Pareto (probability) Density Estimation (PDE). The radius of hyperspheres with this property (Pareto spheres) is called the Pareto radius r_p.

Subsets (volumes) which contain in the average p_u data points are optimal in the sense that they yield with minimal set size as much information as possible.

5. Density estimation for cluster in data

This chapter presents some empirical properties for the usage of PDE to measure data density in data sets which contain clusters.

We found in 1000 repeated experiments with random N(0,1) distributed data points that the 18 percentile of all distances is the Pareto radius. This radius may therefore be used for data sets with an presumably Gaussian inner cluster structure.

The quality of PDE was tested on a two cluster experiment: in1000 experiments 500 data points were drawn form a N(0,1) distribution and 500 from a N(20,1) distribution. This gives data sets which contain two natural clusters of known data density. Data density in these sets was estimated by hypersphere density estimations using all percentiles of the data distances as radius. Figure 4 shows the mean +/- standard deviation of the mean sum of squared differences (MSSE) between the true density and the estimated density. It turned out, that PDE is the optimal data density estimation.

Since data clusters may overlap, hypersphere density estimations for overlapping clusters were performed. Overlap is defined as the integral of the probability density function common to both clusters. A t-test with alpha level of 5% rejects the hypothesis that PDE is the

best density estimation up to an overlap ≥ 18%. It can be concluded that up to about 20% of common points, PDE is optimal to identify clusters.

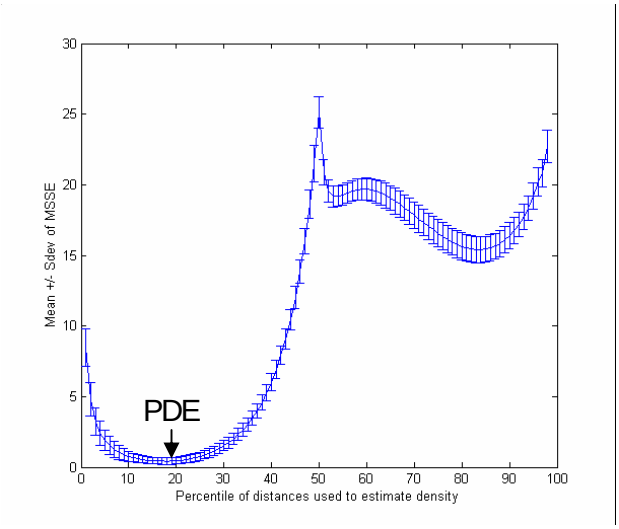

Figure 4: Quality of density estimation using hyperspheres

The distribution of the distances is, however, strongly influenced by the inner cluster variances. In 1000 repeated experiments with different inner cluster variances we found that within a range of 0.1 to 20 PDE stays very close to the best hypersphere density estimation. For high data dimensions we found that PDE, while still being close to the best density estimation, systematically overestimates low densities and overestimates high densities. This is, however, a well known property of all density estimations with fixed kernel width [10].All these experiments show that PDE is a very good density estimation especially suited for the detection of clusters. This is in particular true within the setting of SOM usage.

6 P-Matrix

The P-Matrix on a SOM is defined analogously to a U-Matrix using a measure of data density, such as PDE(w_i), as P-height at the coordinates of neuron n_i. The P-heights are displayed at the neuron's coordinates. This means, at the position of each neuron a density estimation for the data space is displayed. The P-Matrix on a ESOM shows a landscape of density relationships ordered by the topology preserving properties of the ESOM.

6.1 Properties of a P-Matrix

The P-Matrix displays the number of input data points in a hypersphere with the Pareto radius around each weight vector of a neuron. Properties of the P-Matrix are:

- the position of the bestmatches reflect the topology of the input space, this is inherited from the underlying SOM algorithm
- neurons with **large** P-heights are situated in dense regions of the data space
- neurons with **small** P-height are "lonesome" in the data space
- outliers in the input space are found in „funnels".
- "ditches" on a P-Matrix point to cluster boundaries
- „plateaus" on a P-Matrix point to cluster centers

One can see, that many, but not all, properties of the P-matrix are the inverse of the U-matrix. In contrast to the U-matrix , which is based on the distance structure of the data space, the P-Matrix is based on the data's density structure. This gives a new and complementary insight into a high dimensional data space.

7 U-Maps

The highest regions on a P-matrix provide a natural starting point for a visual investigation into a high dimensional data space. Such regions correspond to the most dense regions in the data space. In order to remove duplicates from a tiled display of a borderless U-Matrix , the region with the largest P-heights is used as the center region. The adjoining region of the U-Matrix with the second largest integrated P-height is added next. Other regions are added successively according to this criterion. The algorithm terminates if there is precisely one image of each bestmatch. This leads to a U-Matrix landscape with curved boundaries. The resulting landscapes resemble islands or continents (see figure 6). The center of this landscape represents the most dense data regions.

The obvious resemblance with geographical landscapes led us to call this display a U-Map. A P-Map is constructed from a U-Map using the corresponding P-Matrix. The resemblance of U-Maps to geographical maps or landscapes may be enhanced by computer graphical means such as texturing, coloring and lightening. The following shows the U-Map for the Hepta data. One can see that the regions are grouped around the central cluster containing the 12 points.

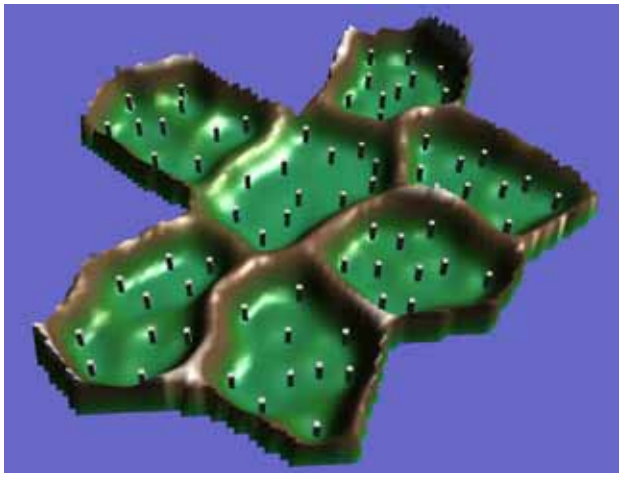

Figure 5: U-Map of the Hepta data

Figure 6 shows a U-Map of a real world data set. The data used consists of a sample of about 300.000 customer data records. The data was provided by Swisscom AG, Bern, Switzerland (see [6] for details). Twenty variables concerning the usage of the Swisscom Mobile telecommunication network were used. A toroid U-Matrix with neurons of a on a rectangular grid ESOM of size 128 by 128 was constructed. The U-Map yielded valuable insights into the customer structure of the mobile phone markets. It has been used for the prediction of churning and for market segmentation [6].

8. Conclusions

This paper defines some new tools for the discovery of structures in high dimensional data sets based on emergent SOM (ESOM) technology. The well known U-Matrix defined originally for planar map spaces is extended to toroid neuron spaces. Planar maps exhibit border effects. Borderless neuron spaces, such as toroid spaces, avoid this unwanted effect. In this work we present a new method for the display of borderless toroid neuron spaces. For this we introduce a data density estimation method, called Pareto Density Estimation (PDE), which is based on the distribution of the data distances given in the context of SOM usage. While PDE optimizes an information theoretic criterion, it turned out that PDE is also a optimal for density estimation in data sets containing cluster. Experimental results confirmed this for a wide variation in cluster number, overlap and dimensionality.

With PDE a new visualization tool called P-Matrix could be defined displaying the density structure of the input data set. The P-Matrix can be used as a complementary tool for the detection and definition of clusters in the data. Furthermore it can be used to define centers and borderlines on a toroid U-Matrix display. This usage leads to U-Maps and P-Maps which possess all the unique properties of a U-Matrix but avoid ambiguous images. U-Maps depict the distance structure of a data space as a three dimensional landscape whose floor space is ordered according to the topology preserving features of a SOM. P-Maps show the data's density structures. Dense data spaces are displayed in the very center of these maps. The combination of U-Maps and P-Maps facilitate the detection of clusters. These properties render these maps unique tools for data mining in high dimensional spaces.

A nontrivial example from the domain of customer relationship management in telecom markets was presented. This demonstrated the effectiveness and beauty of the approach. The maps defined here are computer generated landscapes of important properties of the data space. Modern three dimensional picture techniques on the maps allow an excellent and appealing view into high dimensional spaces.

Acknowledgements

My special thanks belongs to my co-workers and students at the DataBionic reseach lab who provided programs, pictures and suggestions for this work. Special thanks to Ulrich Penndorf und Fabian Moerchen for comments on the text.

References

[1] T. Kohonen, "Self-Organized formation of topologically correct feature maps", Biological Cybernetics, Vol.43, pp.59-69, 1982.

[2] A.Ultsch, "Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series" , in: Oja, E., Kaski, S. (Eds.): Kohonen Maps, pp. 33 - 46.

[3] A.Ultsch, "Self-Organizing Neural Networks for Visualization and Classification", Proc. Conf. Soc. for Information and Classification, Dortmund, April 1992.

[4] A.Ultsch, F. Röske, "Self-Organizing Feature Maps Prediciting Sea Levels", in Information Sciences 144/1-4, Elsevier, pp 91 - 125, Amsterdam, 2002

[5] A.Ultsch, M..Eilers, "DNA Microarrays of tumors diagnosed with databionic methods" (in German) in Kooperationspartner in Forschung und Innovation, pp 19 - 20, Wiesbaden, 2002

[6] A.Ultsch, "Emergent Self-Organizing Feature Maps used for Prediction and Prevention in Mobile Phone Markets", in Journal of Targeting 10/4, Steward, pp 401 - 425, London, 2002

[7] G. J. Deboeck, A. Ultsch, "Picking Stocks with Emergent Self-Organizing Value Maps", in: Novak,M.(Ed): Neural Networks World,Vol 10,Nr. 1-2, pp 203 - 216 ,.

[8] A.Ultsch, "The reasons behind Pareto's 80/20 law and limits for an ABC analysis (in German), Technical Report, Nr 30, Department of Computer Science, Univerisity of Marburg, 2001

[9] J. Vesanto et al.,"Self-organizing map in matlab: the SOM toolbox", Proceedings of the Matlab DSP Conference, pp 35–40, Espoo, Finland, November, 1999

[10] D.W. Scott, "Mulivariate Density Estimation", Wiley-Interscience, 1992.

[11] L. Devroye, G. Lugosi, "Non-asymptotic universal smoothing factors kernel complexity and Yatracos classes", Annals of Statistics, vol. 25, pp. 2626–2637, 1997.

[12] L. Devroye, G. Lugosi, "Variable kernel estimates: on the impossibility of tuning the parameters", in: E. Giné and D. Mason (editors), High- Dimensional Probability, Springer-Verlag, New York, 2000.

[13] Shannon, C.E., A Mathematical Theory of Communication, The Bell System Technical Journal, Vol 27, pp 379-423, 1948

[14] Juran, J.M., "Pareto, Lorenz, Carnot, Bernoulli, Juran and Others, Industrial Quality Control, October 1950, p. 25

[15] Bellman, R., Adaptive Control Processes: A Guided Tour, Princeton,University Press, 1961

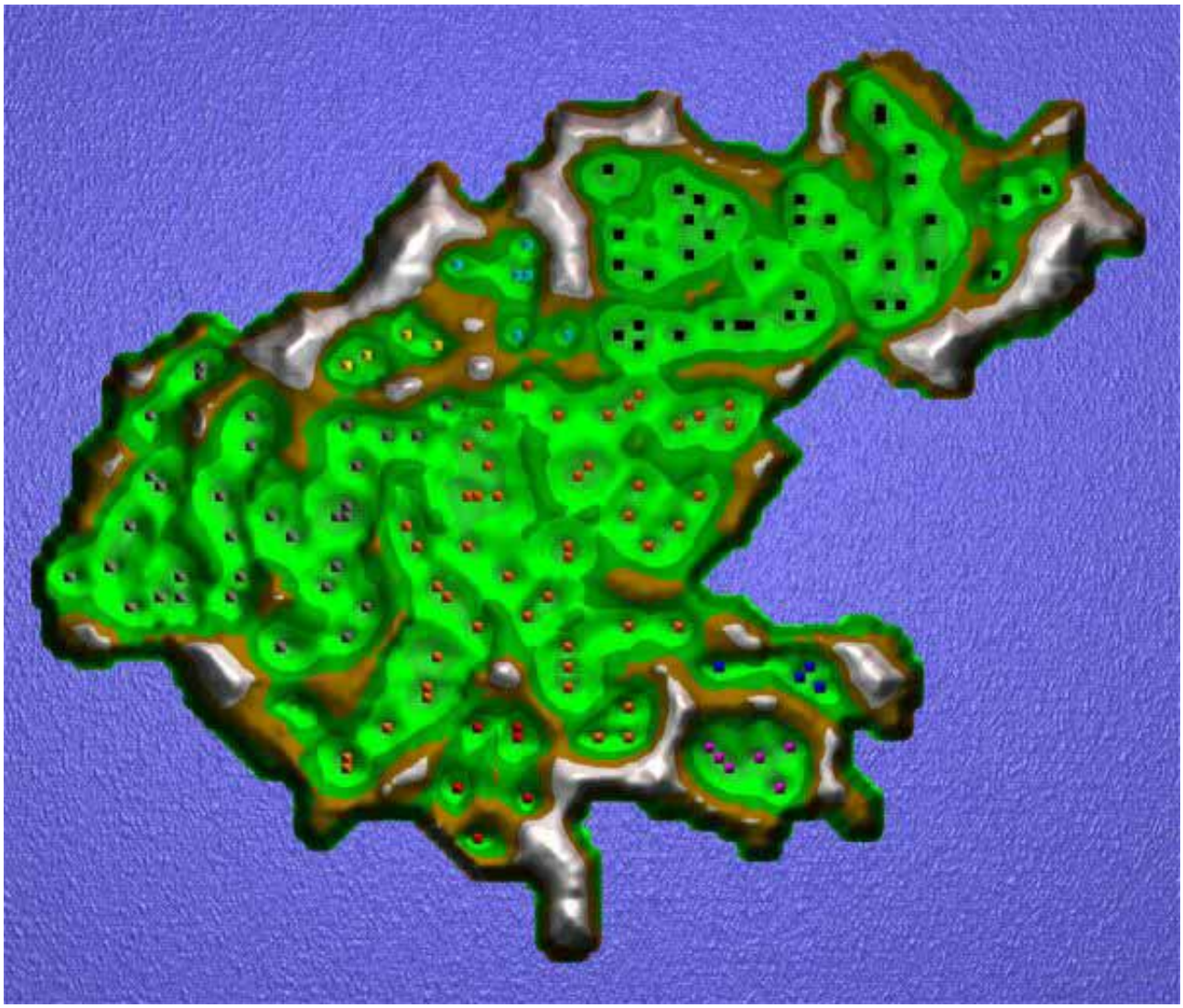

Figure 6: U-Map of telecommunication customer data