Novelty detection using Self-Organizing Maps

Alexander Ypma and Robert P. W. Duin

Pattern Recognition Group
Faculty of Applied Physics Delft University of Technology
Lorentzweg 1, 2628 CJ, Delft The Netherlands
email: *ypma@ph.tn.tudelft.nl*

Abstract

Failure detection in process monitoring involves a classification mainly on the basis of data from normal operation. When a Self-Organizing Map is used for the description of normal system behaviour, a compatibility measure is needed for declaring a map and a dataset as matching. We propose a novel variant of one such measure and investigate usefulness of consisting and novel measures both with synthetic data and in two real world applications.

1 Introduction

A typical aspect of fault diagnosis in a process or system is the limited availability of measurement data concerning faulty situations: often it is hard to acquire a dataset representative of the whole "failure space", whereas the normal operation space can be characterized very accurately. It may be valuable to make an accurate representation of the normal (admissible or healthy) behaviour, and detect faults as significant deviations w.r.t. this admissible domain. When there are some measurements from the abnormal situation, giving a clue about the complementary space, fault diagnosis can be based on the trajectories of a system or state in the feature space [Simula and Kangas, 1995].

Since failure detection must be performed on the basis of a dataset mainly originating from a single class (which can be looked upon as unlabeled data), the use of an unsupervised learning method seems appropriate. A Self-Organizing Map (SOM) map could be selected for this purpose: this method places emphasis on domain description compared to probability density estimation methods. Moreover, the topology of the input space is being preserved (as opposed to conventional unsupervised clustering algorithms like *k-means clustering)*, giving information about the regularity of the mapping, which could be exploited in defining a more confident "compatibility measure". Since regularity is favourable when a SOM is used for visual inspection [Simula and Kangas, 1995, Kaski and Lagus, 1996], tracking a map's regularity is useful.

2 Measures for map goodness

The Self-Organizing Map involves both a vector quantization and a nonlinear projection of the input space to some (usually lower-dimensional) output space. Hence, a measure of the accuracy of the representation of a dataset by a SOM could be based on its accuracy in performing these tasks. In the most basic approach, one only takes the first task into account, by looking at the *average quantization error* for judging a map's goodness [Kohonen, 1995].

Several measures were proposed in the literature for evaluation of a map's preservation of the input space topology. In [Zrehen, 1993] proper organization is expected to be indicated by the fact that map units that are close to each other on the grid are only activated by vectors that are close in the input space. Hence, the line connecting two neighbouring prototypes should not be intersected too often by another unit's Voronoi region (the set of input vectors that are closest to the unit), since this indicates remoteness in the input space. Note that this method is independent of a certain set of input vectors, but is determined using the map only.

Alternatively, one can focus on the continuity of the nonlinear mapping from input space to map grid. In [Bauer and Pawelzik, 1992] the *topographic product* was used for this purpose. Basically, it consists of two terms that indicate the correspondence between k-nearest neighbours in the original and the projected space. However, it has been pointed out that this measure cannot distinguish between folding of the map along nonlinearities in the data manifold (which is favourable) and folding *within* a data manifold (leading to discontinuities in the mapping) [Villmann et al., 1994]. Hence, using just the topology preservation property for quantification of map goodness (like in [Martinetz and Schulten, 1994]) will not suffice for tracking a map's regularity. Also the fact that a metric in the output space is incorporated makes the measure sensitive to a certain topology.

When taking both tasks into account, we could choose a measure of a map's goodness on the basis of the cost function that is being minimized in the SOM training algorithm, which can be decomposed as [Kaski, 1997]

$$E = \sum_k \|\mathbf{x}_k - \mathbf{n}_c\|^2 + \sum_i \sum_j h_{ij} N_i \|\mathbf{n}_i - \mathbf{m}_j\|^2 \qquad (1)$$

where N_i denotes the number of data items which are closest to the reference vector $\mathbf{m}_i$ and $\mathbf{n}_i =$ the centroid of the cluster corresponding to V_i, the Voronoi region corresponding to reference vector $\mathbf{m}_i$.

However, this cost function depends on the map size and the topology of the map lattice (via the neighbourhood function h_{ij}), which is undesirable since a larger, irregular map could thus yield a lower cost then a smaller, more regular map with similar prototypes. Some other criteria are required.

In [Kraaijveld et al., 1995] the distance between two map units was defined in terms of the input space, i.e. as the minimum of summed distances between neighbouring units over all eight-connected paths on a 2D map grid from one unit to the other, where the distance between neighbours on the map is defined as the Euclidean distance of their reference vectors. In [Kaski and Lagus, 1996] this idea was used to define the *goodness* of a map w.r.t. a data set as the average quantization error over the data set plus a term that penalizes remoteness of units on the map grid that have prototypes that are close in the input space: $C = E[d(x)]$, where

$$d(x) = \|x - m_{p_1(x)}\| + \min_i \sum_{k=0}^{K_{p_2(x),i}-1} \|m_{I_i(k)} - m_{I_i(k+1)}\| \qquad (2)$$

where $I_i(k)$ denotes the index of the kth unit on a path along the map grid from the unit $I_i(0) = p_1(x)$ matching the input vector x best, to the second nearest reference vector in the input space $I_i(K_{p_2(x),i}) = p_2(x)$. Note that, using this measure, only the amount of topology preservation in the mapping from input space to map grid can be tracked. This suffices, since the SOM learning algorithm enforces the preservation of the topology in the *inverse mapping*, provided the input space is of higher dimension than the map dimension [Martinetz and Schulten, 1994] (which will almost always be the case in practice).

We used a data set generated from a two-dimensional parabolic shaped distribution to illustrate the use of the original goodness measure (2) for determination of the goodness of a map w.r.t. a certain data set. A one-dimensional SOM of 100 units was trained for 10.000 cycles, and the final neighbourhood (during the convergence phase of the training procedure) was varied from 1.0 to 5.0 and 10.0 (figure 1). Increasing the final neighbourhood in adaptation of a SOM amounts to increasing its stiffness, since its becomes less sensitive to local characteristics: the broader neighbourhood kernel makes also less local distances contribute to the error function.

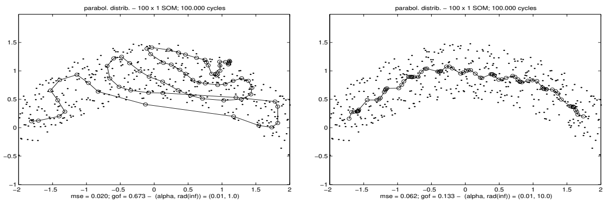

Figure 1: Map goodness with varying stiffness

Intuitively, the SOM in figure 1b gives the best representation of the input space: it does not fold itself unnecessarily into the input space, whereas the general form of the distribution is still being tracked. The SOM in figure 1a is much less stiff, and exhibits a kind of "overfitting" to the input distribution: every local characteristic is being tracked, enabling excessive tuning to local disturbances. Indeed, it can be observed that going from less to more stiffness causes a slight increase in quantization error, from 0.020 in the least stiff case, to 0.032 in the intermediate case (not shown) and 0.062 in the most stiff case (which is still reasonably low), while the goodness of the map w.r.t. the input distribution, denoting the regularity of the map, decreases with stiffness from 0.673, to 0.164 and finally 0.133.

The main difficulty with the previous goodness measure relates to the fact that it gives no indication of compatibility in its own right (i.e. without referring to the order of magnitude of the quantization error part). Moreover, only the second nearest reference vector is taken into account for the determination of a map's regularity. With sparse datasets (with only a few samples per node), this could lead to inaccurate goodness estimations, while information about third nearest neighbours etc. is neglected [Kaski, 1997]. For this reason we propose a normalized version of the measure, that is zero for perfect mappings and denotes deviations in terms of map distances. The regularity is tracked by looking at a sample's best-matching unit and its 2 to k-nearest neighbours in the input space:

$$\begin{aligned} d(x) &= \frac{\|x - m_{p_1(x)}\|}{\frac{1}{Q_2}\sum_{q \in q_2(x)} \|m_{p_1(x)} - m_q\|} \\ &+ \{\sum_{n=2}^{k} \alpha_n | \min_i \sum_{j=0}^{K_{p_k(x),i}-1} \|m_{I_i(j)} - m_{I_i(j+1)}\| \\ &- \min_l \sum_{j=0}^{K_{q_k(x),l}-1} \|m_{J_l(j)} - m_{J_l(j+1)}\| | \} \end{aligned} \qquad (3)$$

where $J_l(j)$ denotes the index of the jth unit on the shortest path l along the map grid from the unit $J_l(0) = p_1(x) = q_1(x)$ to a k-nearest map unit (in the output space) $J_l(K_{q_k(x),l}) = q \in q_k(x) = \{$ units $q|q$ is $(k-1)$-nearest neighbour in the output space of best-matching unit $p_1(x) = q_1(x)\}$. Q_k is the cardinality of $q_k(x)$.

Note that only the cost of the path *higher* than the ideal case path cost is incorporated in the penalty term. By suitable weighting of the contributions of the k nearest neighbours (terms $0 \leq \alpha_n \leq 1, \forall n = 2, \ldots, k$), emphasis can be placed on local or global map orderliness. Normalizing the *mse*-part of the formula to the map distances in the neighbourhood of the best-matching unit makes the measure more sensitive to the orderliness term.

3 Mechanical fault detection

Early detection of faults in rotating mechanical machinery can both increase safety of the system and decrease maintenance costs, since maintenance can be scheduled when appropriate and convenient. Failures like misalignment and imbalance will be visible in the spectrum of machine casing vibrations as amplified harmonics of machine running speed (in revolutions per minute).

From a submersible pump, rotating at 1500 RPM, measurements were acquired at three normal modes (with different loads) and two failure modes (imbalance, different running speeds), and the spectrum was determined over segments of 500 samples. Normalization w.r.t. mean and standard-deviation and determination of the principal components resulted in 5 classes of 256-dimensional feature vectors. Two 5x5 SOMs with a rectangular 4-connected neighbourhood topology were trained for 10.000 cycles on data from one of the normal classes and one of the fault classes, denoted by *net1* and *net2* respectively (using a box-shaped neighbourhood kernel). The initial update neighbourhood was 10.0 and the final update neighbourhood 1.0.

The distance of each dataset to the map was computed using the average quantization error *mse* and goodness measures (2) and (3). In the latter case, the penalty term is computed for 2- and 3-nearest neighbours, with weighting coefficients $\alpha_2 = \alpha_3 = 1.0$. For a final neighbourhood of 5.0, the results are given in the table below.

net1:	*mse*	*gof* (2)	*gof* (3)
train normal	0.010248	0.077419	0.173853
test normal 1	0.014796	0.083955	0.235532
test normal 2	0.008649	0.052118	0.186613
train imbalance 1	0.072720	0.116546	1.580411
test imbalance	0.017222	0.017222	0.559953
net2:			
train normal	0.034181	0.061685	0.852837
test normal 1	0.033888	0.058335	0.851931
test normal 2	0.023353	0.023353	0.766289
train imbalance 1	0.048503	0.147683	0.539826
test imbalance	0.026366	0.072205	0.594898

From these figures it is clear that only using the normalized goodness measure normal and abnormal behaviour can be distinguished. For a final neighbourhood of 1.0 (yielding a less stiff map) discrimination results were similar for all three distance measures. Normalization of the *mse*-part of measure (3) could hence also account for the improved discriminatory effect.

4 Pipeline leak detection

Detection of leaks in pipelines using hydrophone measurements is an example of a fault identification problem, where a representative set of measurements characterizing all possible leak scenarios can be easily acquired, while the possible set of background noise is much more diffuse and hard to measure.

A training set of 100 18-dimensional feature vectors was formed out of a representative subset of the available data, where the measured time series were preprocessed by computing the logarithm of the power spectrum and normalizing w.r.t. mean and standard-deviation. Leaks of varying remoteness were simulated by superposing appropriately attenuated data, acquired near a leak, with background noise. A test set from different measurements with similar charateristics as the train set was used to track the map's interpolation capabilities, a validation set with different characteristics was used to look into its extrapolation power. To verify the map's capability to distinguish between leak and noise, a representative equal sized set of background noise vectors was constructed, originating from (proximal and remote) ship noise and normal background noise. Out of each set of leak vectors, two more sets were constructed by attenuation with 10 and 20 dB before superposition with a representative set of noise vectors.

A Self-Organizing map of the same topology as in the previous section was trained on the set of train vectors for 10.000 cycles. The (*mse*) was determined for the leak train-, interpolation- and extrapolatio sets, and for the set of noise vectors. Next, the goodness of the fit (*gof*) between the trained map and the data set under investigation was determined for all sets. This process was repeated for the attenuated versions of the leak sets. Since training of a SOM is a stochastic approximation method, every experiment was replicated for 20 times. Scatter-plots of the *mse* and the *gof* resulting for the (progressively attenuated) leak and noise vectors are shown in figure 2. Here, '+'-signs denote the *mse* and 'o'-signs the *gof* figures. Figure 2a shows the outcomes for train leak and noise sets, figure 2b for the interpolation sets and 2c for the extrapolation sets.

It can be observed that the leaksets used to train the SOM can well be discriminated from the noise set: the distance to the straight line $error(noise) = error(leak)$ is fairly large, even for a large amount of attenuation. The same holds for the interpolation data (figure 2b). Extrapolation to data with very different characteristics leads to worse discrimination: attenuation with 20 dB results in an error which is larger for leak than for noise data.

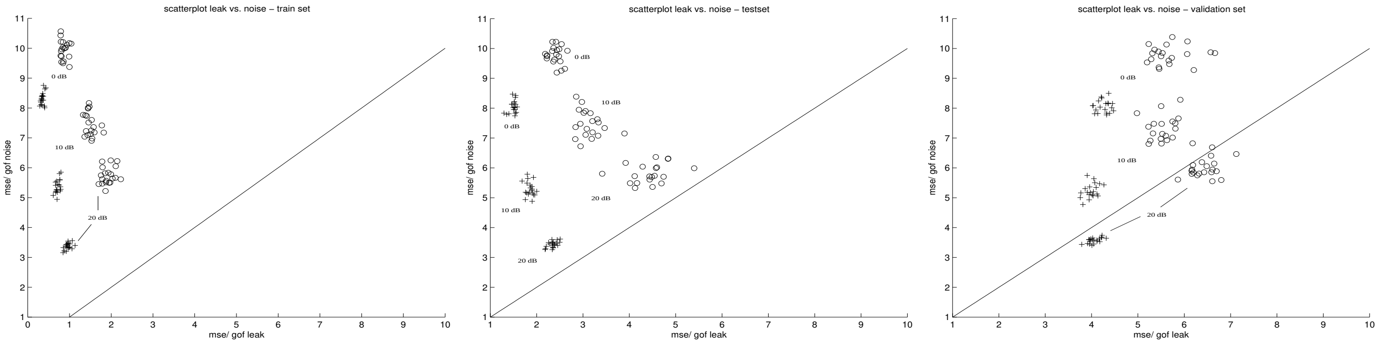

Figure 2: Scatterplots of error on noise vs. leak set

Both error figures (*mse* and *gof*) have similar discrimination capacity: the clusters of goodness-figures are slightly more diffuse and translated from the origin according to the straight line, indicating a nearly constant offset w.r.t. the quantization error for both leak and noise data. This can be understood by noting that the goodness measure of equation (2) computed over a batch of samples consists of the average quantization error over the batch plus an "orderliness" term that is a mainly a feature of the map. Differences in input data are mainly reflected in the *mse*-part, orderliness penalties will be comparable for incompatible data. For data used to train the map, the situation is different: the map has become ordered w.r.t. this dataset, so the orderliness penalty will be somewhat smaller. This can be seen in figure 2a as the almost vertical offset of the *gof*-clusters from the *mse*-clusters. With the training set, map nodes will occur as starting point for the computation of the orderliness term according to the input distribution, opposed to incompatible datasets that will select border nodes more often than inner nodes. The border effect that occurs in SOM training (*the map is less well ordered near the borders of the map* [Kohonen, 1995]) will lead to higher orderliness penalties in the latter case. Moreover, the probability that an input vector will have one- and two-nearest neighbours in the input space that are not neighbours on the map is expected to become larger as the distance of the vector to the map increases. Hence, map regularity w.r.t. a certain dataset could be employed for determination of a map's compatibility to the dataset.

5 Conclusions

Incorporation of a penalty term for map complexity in evaluation of SOM performance can be valuable for detection of datasets incompatible to the SOM, because of sensitivity for remoteness between dataset and map and the border effect in SOM training. Application of the novel measure to mechanical fault detection indicated that also a normalized quantization error term in the formula may be valuable in unsupervised novelty detection.

References

[Bauer and Pawelzik, 1992] Bauer, H. U. and Pawelzik, K. R. (1992). Quantifying the neighbourhood preservation of self-organizing feature maps. *IEEE Transactions on neural networks*, Volume 3, No. 4:570–579.

[Kaski, 1997] Kaski, S. (1997). *Data exploration using Self-Organizing Maps.* PhD thesis, Helsinki University of Technology.

[Kaski and Lagus, 1996] Kaski, S. and Lagus, K. (1996). Comparing self-organizing maps. In Von der Malsburg, C., Von Seelen, W., Vorbrüggen, J. C., and Sendhoff, B., editors, *ICANN '96 Proceedings*, volume 1112 of *Lecture notes in computer science*, pages 809–814. Springer-Verlag.

[Kohonen, 1995] Kohonen, T. (1995). *Self-Organizing Maps.* Springer Series in Information Sciences. Springer.

[Kraaijveld et al., 1995] Kraaijveld, M. A., Mao, J., and K., J. A. (1995). A nonlinear projection method based on Kohonen's topology preserving maps. *IEEE Transactions on neural networks*, Volume 6, No. 3:548–559.

[Martinetz and Schulten, 1994] Martinetz, T. and Schulten, K. (1994). Topology representing networks. *Neural networks*, Volume 7, No. 3:507–522.

[Simula and Kangas, 1995] Simula, O. and Kangas, J. (1995). *Process monitoring and visualization using Self-Organizing Maps*, chapter 14, pages 377–390. Elsevier Science B.V.

[Villmann et al., 1994] Villmann, T., Der, R., and Martinetz, T. (1994). A new quantitative measure of topology preservation in Kohonen's feature maps. In *ICNN'94 Proceedings*, pages 645–648. IEEE Service Center, Piscataway, NJ.

[Zrehen, 1993] Zrehen, S. (1993). Analyzing Kohonen maps with geometry. In Gielen, S. and Kappen, B., editors, *ICANN '93 Proceedings*, pages 609–612. Springer-Verlag.