SOM Quality Measures: An Efficient Statistical Approach

Lutz Hamel

Dept. of Computer Science and Statistics
University of Rhode Island
Kingston, RI 02881
hamel@cs.uri.edu

Abstract. We are interested in practical tools for the quantitative evaluation of self-organizing maps (SOMs). Recently it has been argued that any quality measure for SOMs needs to evaluate the embedding or coverage of a map as well as its topological quality. Over the years many different quality measures for self-organizing maps have been proposed. However, many of these only measure one aspect of a SOM or are computationally very expensive or both. Here we present a novel, computationally efficient statistical approach to the evaluation of SOMs. Our approach measures both the embedding and the topological quality of a SOM.

1 Introduction

We are interested in practical tools for the quantitative evaluation of trained self-organizing maps (SOM) [10]. Here we present an efficient statistical approach to the evaluation of SOM quality. A nice overview of common SOM quality measures appears in [14]. The measures described there report on either the quality of map embedding in the data input space, sometimes called coverage, (*e.g.* quantization error [10]) or on the topological quality of the map (*e.g.* topographic error [9]). Another measure not mentioned in the above overview is the neighborhood preservation [3] which similarly to the topographic error strictly measures the topological quality of a map.

More recently it has been argued that any SOM quality measure needs to report on both the embedding of the map in the input data space as well as the topological quality of a map [2]. To this we would like to add that any practical SOM quality measure also has to be computationally efficient. Most quality measures fail these requirements: they either only measure one aspect of a SOM or they are computationally very expensive or both. Here we propose a statistical approach that measures both the embedding and the topological quality of a map and is computationally efficient even for large training data sets and/or maps. Our proposed measure computes the quality of a SOM as a pair of numbers: 1) the embedding accuracy, 2) the estimated topographic accuracy. The embedding accuracy is a quality measure we first explored in [6] as a convergence criterion and we reexamine it here in this new context. The

estimated topographic accuracy is a novel statistical approach to the topological quality of a map. Besides developing our statistical approach here we also provide a preliminary validation.

The remainder of this paper is structured as follows. Section 2 examines our notion of embedding summarizing major results. We develop the estimated topographic accuracy in Section 3. Our implementation is briefly discussed in Section 4. We provide the results of our preliminary validation in Section 5. Section 6 provides conclusions and points to further work.

2 Map Embedding Accuracy

Yin and Allinson have shown that under some mild assumptions the neurons of a large enough self-organizing map will converge on the probability distribution of the training data given infinite time [19]. This is the motivation for our map embedding accuracy:

> *A SOM is completely embedded if its neurons appear to be drawn from the same distribution as the training instances.*

This was the basic insight of our original SOM convergence criterion [6]. Here we briefly summarize and adjust our terminology with respect to embedding.

Our view of embedding naturally leads to a two-sample test [12]. Here we view the training data as one sample from some probability space $\mathbf{X}$ having the probability density function $p(x)$ and we treat the neurons of the SOM as another sample. We then test to see whether or not the two samples appear to be drawn from the same probability space. If we operate under the simplifying assumption that each of the d features of the input space $\mathbf{X} \subset \mathbb{R}^d$ are normally distributed and independent of each other, we can test each of the features separately. This assumption leads to a fast algorithm for identifying SOM embedding: We define a feature as embedded if the variance and the mean of that feature appear to be drawn from the same distribution for both the training data and the neurons. If all the features are embedded then we say that the map is completely embedded.

The following is the formula for the $(1-\alpha)*100\%$ confidence interval for the ratio of the variances from two random samples [12],

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{f_{\frac{\alpha}{2},n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot f_{\frac{\alpha}{2},n_1-1,n_2-1}, \tag{1}$$

where s_1^2 and s_2^2 are the values of the variance from two random samples of sizes n_1 and n_2 respectively, and where $f_{\frac{\alpha}{2},n_1-1,n_2-1}$ is an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. To test for SOM embedding, we let s_1^2 be the variance of a feature in the training data and we let s_2^2 be the variance of that feature in the neurons of the map. Furthermore, n_1 is the number of training samples and n_2 is the number of neurons in the SOM. The variance of a particular feature of both training data and neurons appears to be drawn from the same probability space if 1 lies in the confidence interval denoted by

equation (1): the ratio of the underlying variance as modeled by input space and the neuron space, respectively, is approximately equal to one, $\sigma_1^2/\sigma_2^2 \approx 1$, up to the confidence interval.

In the case where $\bar{x}_1$ and $\bar{x}_2$ are the values of the means from two random samples of size n_1 and n_2, and the variances of these samples are σ_1^2 and σ_2^2 respectively, the following formula provides $(1-\alpha)*100\%$ confidence interval for the difference between the means [12],

$$\mu_1 - \mu_2 > (\bar{x}_1 - \bar{x}_2) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \tag{2}$$

$$\mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{3}$$

The mean of a particular feature for both training data and neurons appears to be drawn from the same probability space if 0 lies in the confidence interval denoted by equations (2) and (3). Here $z_{\frac{\alpha}{2}}$ is the appropriate z score for the chosen confidence interval.

We say that a feature is embedded if the above criteria for both the mean and variance of that feature are fulfilled. We can now define the *map embedding accuracy* for d features,

$$ea = \frac{1}{d}\sum_{i=1}^{d} \rho_i, \tag{4}$$

where

$$\rho_i = \begin{cases} 1 & \text{if feature } i \text{ is embedded,} \\ 0 & \text{otherwise.} \end{cases}$$

The map embedding accuracy is the fraction of the number of features which are actually embedded (i.e. those features whose mean and variance were adequately modeled by the neurons in the SOM). With a map embedding accuracy of 1 a map is fully embedded. In order to enhance the map embedding accuracy in our implementation [7], we multiply each embedding term ρ_i by the significance of the corresponding feature i which is a Bayesian estimate of that feature's relative importance [5].

The computational complexity of our map embedding accuracy is,

$$O((n+m) \times d) \tag{5}$$

with n the number of training examples, m the number of neurons, and d the number of features. For most cases we have that $d \ll n$ and $d \ll m$, therefore we can say our algorithm is quasi-linear in the sum of the number of training examples and number of neurons. This means that computing the map embedding accuracy is efficient for most cases.

In essence our map embedding accuracy measures the same thing as the quantization error: the effective representation of the training data by the neurons of a map. There is one big difference; our map embedding accuracy indicates when

a map is completely embedded, that is, it indicates when statistically there is no difference between the population of training points and the population of neurons. No such criterion exists for the quantization error. The ramification is that the map embedding accuracy can be used as a measure across different sized maps where the quantization error cannot [14]. A more in-depth statistical analysis of our map embedding accuracy can be found in [13].

3 Estimated Topographic Accuracy

Many different approaches to measuring the topological quality of a map exist, *e.g.* [11, 18]. But perhaps the simplest measure of the topological quality of a map is the *topographic error* [9] defined as:

$$te = \frac{1}{n} \sum_{i=1}^{n} err(\boldsymbol{x}_i) \tag{6}$$

with

$$err(\boldsymbol{x}_i) = \begin{cases} 1 & \text{if } bmu(\boldsymbol{x}_i) \text{ and } 2bmu(\boldsymbol{x}_i) \text{ are not neighbors,} \\ 0 & \text{otherwise.} \end{cases}$$

for training data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ where $bmu(\boldsymbol{x}_i)$ and $2bmu(\boldsymbol{x}_i)$ are the best matching unit and the second-best matching unit for training vector $\boldsymbol{x}_i$ on the map, respectively. We define the *topographic accuracy* of a map as,

$$ta = 1 - te. \tag{7}$$

Computing the topographic accuracy can be very expensive, especially for large training data sets and/or maps. If we let n be the size of the training data, m the number of neurons of the map, and d the number of features of the training data, then the complexity of computing the topographic accuracy is,

$$O(n \times m \times d). \tag{8}$$

One way to ameliorate the situation is to sample the training data and use this sample S to estimate the topographic accuracy. If we let s be the size of the sample then the *estimated topographic accuracy* is,

$$ta' = 1 - \frac{1}{s} \sum_{i=1}^{s} err(\boldsymbol{x}_i) \tag{9}$$

with $\boldsymbol{x}_i \in S$ and complexity $O(s \times m \times d)$. As we will see later in the paper we can get accurate values for ta' with very small samples. Therefore we can assume $s \ll m$. Also, in most cases we have $d \ll m$. Therefore, the complexity of ta' becomes quasi-linear in the number of neurons of the map which again represents a very efficient algorithm to compute the estimated topographic accuracy.

In addition to computing the value for the estimated topographic accuracy we use the bootstrap [4] to compute values for an appropriate confidence interval

in order to give us further insight into the estimated topographic accuracy in relation to the actual value for the topographic accuracy whose value should fall within the bootstrapped confidence interval.

It is easy to see from (9) that for topological faithful maps the estimated topographic accuracy should be close to 1. We then say that the map is *fully organized*.

4 Implementation

We maintain an R package called `popsom` [7] in the CRAN repository [15]. The functionality discussed in this paper has been implemented in that package and is available as of package version 3.0.[1] Here is a sample session using our package:

```
 1: > library(popsom)
 2: > data(iris)
 3: > df <- subset(iris,select=-Species)
 4: > labels <- subset(iris,select=Species)
 5: > m <- map.build(df, labels, xdim=15, ydim=10, train=1000)
 6: > q <- map.quality(m)
 7: > cat(sprintf("embedding: %3.2f\n",q$embedding))
 8: embedding: 0.81
 9: > acc <- q$accuracy$acc
10: > lo <- q$accuracy$lo
11: > hi <- q$accuracy$hi
12: > cat(sprintf("accuracy: %3.2f (%3.2f-%3.2f)\n",acc,lo,hi))
13: accuracy: 0.94 (0.86-1.00)
14: >
```

The first four lines deal with loading the package and the data and then preparing the data for building maps. On the fifth line we build a map with dimensions 15×10 using 1000 training iterations. On line six we compute the map quality. This computes a value with multiple components which we print out separately on the following lines. The embedding accuracy is 0.81 and the estimated topographic accuracy is 0.94. The bootstrapped 95% confidence interval for the estimated topographic accuracy is 0.86-1.00. One way to interpret this interval is that there is a 95% probability that the topographic accuracy computed on the whole training data lies within the interval 0.86-1.00.

5 Preliminary Validation

For our preliminary validation we use the same experiments as in [14]; namely we use the Iris data set [1] (4 independent variables, 150 instances, 3 classes) and the Epil data set [16] (8 independent variables, 236 instances, 2 classes). We build SOMs with the following sizes for the Iris data set:

- small Iris map: 5×3 (15 nodes)

[1] The 3.0 version should be available on CRAN by August 2015.

Table 1. Results for the Iris data set.

iter	*qerr*	*ea*	*ta*	*ta′*	(*lo-hi*)
		*** 5×3 ***			
1	43.95	0.81	0.69	0.74	(0.64-0.86)
10	16.10	0.13	0.83	0.82	(0.70-0.92)
100	5.14	0.68	0.91	0.92	(0.84-0.98)
1000	3.29	1.00	0.95	0.94	(0.88-1.00)
10000	3.36	1.00	1.00	1.00	(1.00-1.00)
		*** 11×6 ***			
1	28.36	0.96	0.09	0.06	(0.00-0.14)
10	20.01	0.28	0.47	0.44	(0.28-0.58)
100	4.10	0.00	0.95	0.88	(0.82-0.96)
1000	1.27	0.96	0.99	1.00	(1.00-1.00)
10000	1.24	1.00	0.99	1.00	(1.00-1.00)
		*** 23×11 ***			
1	36.67	0.81	0.00	0.00	(0.00-0.00)
10	18.12	0.81	0.17	0.14	(0.06-0.22)
100	3.29	0.00	0.82	0.76	(0.64-0.88)
1000	0.59	0.81	0.98	1.00	(1.00-1.00)
10000	0.46	1.00	1.00	1.00	(1.00-1.00)

- medium Iris map: 11×6 (66 nodes)
- large Iris map: 23×11 (253 nodes)

and SOMs of the following sizes for the Epil dataset:

- small Epil map: 5×4 (20 nodes)
- medium Epil map: 10×8 (80 nodes)
- large Epil map: 22×15 (330 nodes)

Map quality does depend largely on two factors: the map size and the number of training iterations applied to a map. Therefore, the big difference between our study and the original study is that we not only track map sizes but also the number of training iterations applied to each map. This allows us to observe the respective quality measures with regards to map sizes and training iterations. Table 1 shows our results for the Iris data set. Here we have the following abbreviations:

- *iter*: training iterations
- *qerr*: the quantization error defined as

$$qerr = \frac{1}{n} \sum_{i=1}^{n} ||bmu(\boldsymbol{x}_i) - \boldsymbol{x}_i||^2, \tag{10}$$

  where $||bmu(\boldsymbol{x}_i) - \boldsymbol{x}_i||$ represents the Euclidean distance between point $\boldsymbol{x}_i$ and its best matching unit $bmu(\boldsymbol{x}_i)$ on the map

Table 2. Results for the Epil data set.

iter	*qerr*	*ea*	*ta*	*ta′*	(*lo-hi*)
		*** 5×4 ***			
1	21.06	0.91	0.37	0.34	(0.24-0.48)
10	12.08	0.30	0.54	0.50	(0.36-0.66)
100	5.50	0.23	0.92	0.90	(0.82-0.98)
1000	2.53	0.98	1.00	1.00	(1.00-1.00)
10000	2.01	0.91	1.00	1.00	(1.00-1.00)
100000	2.17	0.91	1.00	1.00	(1.00-1.00)
		*** 10×8 ***			
1	20.67	0.00	0.23	0.10	(0.02-0.18)
10	18.49	0.00	0.06	0.04	(0.00-0.10)
100	4.27	0.30	0.90	0.88	(0.78-0.96)
1000	1.02	0.91	0.98	1.00	(1.00-1.00)
10000	0.82	0.91	0.98	0.98	(0.92-1.00)
100000	0.93	0.91	0.97	0.98	(0.94-1.00)
		*** 22×15 ***			
1	17.76	0.00	0.00	0.00	(0.00-0.00)
10	16.99	0.30	0.06	0.02	(0.00-0.06)
100	8.52	0.30	0.62	0.62	(0.48-0.74)
1000	0.45	0.53	0.93	0.98	(0.94-1.00)
10000	0.27	0.68	1.00	1.00	(1.00-1.00)
100000	0.33	0.99	0.98	1.00	(1.00-1.00)

- *ea*: embedding accuracy as defined by (4)
- *ta*: topographic accuracy as defined by (7)
- *ta′*: estimated topographic accuracy as defined by (9)
- (*lo-hi*): bootstrap estimate of the 95% confidence interval of *ta′*

We can observe that the quantization error decreases for the most part for all map sizes as the number of training iterations applied to the maps increases. One of the big issues with the quantization error as a quality measure is to determine when it is sufficiently small for the map to be considered to be a good map. That is, with the quantization error there is no indication when a map is completely embedded. Reducing the quantization error to zero is usually not the solution as then the map will likely overfit the data as is usual with statistical models whose training error was reduced to zero. Notice that the quantization error is non-zero for fully embedded and fully organized maps.

Both the embedding accuracy (*ea*) and topographic accuracy (*ta*) increase with the number of training iterations applied to a map until both reach 1 indicating that the map is fully embedded and completely organized, respectively. There is phenomenon where the random initialization of an untrained map can look like a fully embedded map except that it is completely unorganized according to the topographic accuracy.

We can observe that the estimated topographic accuracy (ta') is a good estimate for the topographic accuracy (ta) as it usually falls within a couple of 1/100's of the actual value.

Finally, the topographic accuracy value ta falls within the bootstrap estimate of the 95% interval except for the cases where the map is completely unorganized or the map is fully organized. In these boundary cases the 95% confidence interval does not fully predict the value of ta. In all the computations we use a sample size of 50 to both compute the value of ta' and to compute the bootstrap estimate of the confidence interval. We take a look at the effects of the sample size on the value of ta' and the bootstrap estimate in the next section.

Table 2 shows the results of our experiments for the Epil data set. We can make observations very similar to the observations we made on the Iris data set: The quantization error decreases with training, both ea and ta increase with training until they both reach 1, ta' is a fairly accurate estimate of ta, and the bootstrap estimate of the range of the actual value ta is correct except for the boundary cases. However, the Epil data set seems to be inherently more complex than the Iris data set because even with 100,000 iterations the embedding accuracy never quite reaches 1 even for the small map.

It is interesting to see that in most cases the topographic accuracy converges on 1 much faster than the embedding accuracy, that is, in those cases ta indicates that a map is fully organized without being fully embedded. Also, as we observed earlier, an untrained map can appear to be fully embedded without being fully organized. Therefore, both quality measures are necessary to fully evaluate the goodness of a map and of course we prefer maps where both indices are close to 1. In our implementation we could have created some sort of linear combination of both indices in order to come up with a single quality index. However, we prefer the additional information separate embedding and topographic accuracies purvey.

5.1 Sample Size and Estimated Topographic Accuracy

In order to see the effect the sample size has on the estimated topographic accuracy and the corresponding bootstrap estimate of the confidence interval we trained the respective medium sized maps for both the Iris and the Epil data set using 1000 iterations. We then computed the topographic accuracy ta (7), the estimated topographic accuracy ta' (9), and the bootstrap estimate of the 95% confidence interval using sample sizes k that roughly corresponded to 10%, 30%, 60%, and 100% of the training data. Table 3 shows the results. What is surprising that even with very small samples we obtain accurate estimates of the topographic accuracy. On the other hand, the bootstrap estimate of the confidence interval improves with larger sample sizes.

With a sample size that corresponds to 100% of the data the interpretation of the confidence interval slightly shifts. Here we see that the precise value of the topographic accuracy and in turn the value of the topographic error is data depend. The confidence interval at 100% of the training data tells us that if we were to select another set of data points from the same distribution as the

Table 3. Effects of the sample size on the estimated topographic accuracy.

k	ta	ta'	(lo-hi)
		*** Iris ***	
15	0.95	1.00	(1.00-1.00)
50	0.95	0.96	(0.90-1.00)
100	0.95	0.94	(0.89-0.98)
150	0.95	0.95	(0.91-0.98)
		*** Epil ***	
25	0.97	1.00	(1.00-1.00)
100	0.97	0.96	(0.92-0.99)
200	0.97	0.97	(0.94-0.99)
236	0.97	0.97	(0.94-0.99)

training data in order to compute the topographic accuracy we would expect a value within the given interval.

6 Conclusions and Further Work

We are interested in practical tools for the quantitative evaluation of self-organizing maps. Here we presented a novel statistical approach to the evaluation of SOMs which directly measures the embedding accuracy or coverage of a map and its topographic accuracy. Both quality indices can be computed in quasi-linear time for most cases making them computationally very efficient. We have provided an implementation of our quality measure in form of an R package.

Our preliminary validation seems to show that in essence our embedding accuracy measures the same thing as the quantization error: the effective representation of the training data by the neurons of a map. However, the embedding accuracy has the advantage that it indicates when a map is fully embedded, *i.e.*, statistically there will be no improvement to the map with further training. Our preliminary validation also seems to show that our estimated topographic accuracy is very accurate with respect to the topographic accuracy computed on the whole training data set even when using very small samples.

In terms of a more rigorous validation we would like to test our quality measures against standard test suites such as FCPS [17] and on large real-world data sets. Finally, in order to dispense with our normality and independence assumptions of our data we consider switching to a multi-variate, non-parametric Kolmogorov-Smirnov goodness of fit test [8]. Experiments with the univariate Kolmogorov-Smirnov test seem promising.

Acknowledgements

The author would like to thank Gavino Puggioni for suggesting the non-parametric goodness of fit tests.

References

1. UCI machine learning repository: Iris data set (Feb 2012), http://archive.ics.uci.edu/ml/datasets/Iris
2. Beaton, D., Valova, I., MacLean, D.: Cqoco: A measure for comparative quality of coverage and organization for self-organizing maps. Neurocomputing 73(10), 2147–2159 (2010)
3. De Bodt, E., Cottrell, M., Verleysen, M.: Statistical tools to assess the reliability of self-organizing maps. Neural Networks 15(8-9), 967978 (2002)
4. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. CRC press (1994)
5. Hamel, L., Brown, C.W.: Bayesian probability approach to feature significance for infrared spectra of bacteria. Applied Spectroscopy 66(1), 48–59 (2012)
6. Hamel, L., Ott, B.: A population based convergence criterion for self-organizing maps. In: Proceedings of the 2012 International Conference on Data Mining. Las Vegas, Nevada (Jul 2012)
7. Hamel, L., Ott, B., Breard, G.: popsom: Self-Organizing Maps With Population Based Convergence Criterion (2015), http://CRAN.R-project.org/package=popsom, r package version 3.0
8. Justel, A., Peña, D., Zamar, R.: A multivariate Kolmogorov-Smirnov test of goodness of fit. Statistics & Probability Letters 35(3), 251–259 (1997)
9. Kiviluoto, K.: Topology preservation in self-organizing maps. In: IEEE International Conference on Neural Networks. pp. 294–299. IEEE (1996)
10. Kohonen, T.: Self-organizing maps. Springer series in information sciences, Springer (2001)
11. Merényi, E., Tasdemir, K., Zhang, L.: Learning highly structured manifolds: harnessing the power of SOMs. In: Similarity-based clustering, pp. 138–168. Springer (2009)
12. Miller, I., Miller, M.: John E. Freund's Mathematical Statistics with Applications (7th Edition). Prentice Hall, 7 edn. (2003)
13. Ott, B.H.: A Convergence Criterion for Self-Organizing Maps. Master's thesis, University of Rhode Island (2012)
14. Pölzlbauer, G.: Survey and comparison of quality measures for self-organizing maps. In: Proceedings of the Fifth Workshop on Data Analysis (WDA-04). pp. 67–82. Elfa Academic Press (2004)
15. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013), http://www.R-project.org/, ISBN 3-900051-07-0
16. Thall, P.F., Vail, S.C.: Some covariance models for longitudinal count data with overdispersion. Biometrics pp. 657–671 (1990)
17. Ultsch, A.: Clustering with SOM: U* C. In: In Proc. Workshop on Self-Organizing Maps. pp. 75–82. Paris, France (2005)
18. Villmann, T., Der, R., Herrmann, M., Martinetz, T.M.: Topology preservation in self-organizing feature maps: exact definition and measurement. Neural Networks, IEEE Transactions on 8(2), 256–266 (1997)
19. Yin, H., Allinson, N.M.: On the distribution and convergence of feature space in self-organizing maps. Neural computation 7(6), 1178–1187 (1995)