Description of Input Patterns by Linear Mixtures of SOM Models

Teuvo Kohonen
Helsinki University of Technology, Adaptive Informatics Research Centre
P.O. Box 5400, FI-02015 TKK, Finland
email: teuvo.kohonen@tkk.fi

Keywords: least-squares fitting, linear mixture, self-organizing map

***Abstract*—** This paper introduces a novel way of analyzing input patterns presented to the Self-Organizing Map (SOM). Instead of identifying only the "winner," i.e., the model that matches best with the input, we determine the linear mixture of the models (reference vectors) of the SOM that approximates to the input vector best. It will be shown that if only nonnegative weights are allowed in this linear mixture, the expansion of the input pattern in terms of the models is very meaningful, contains only few terms, and provides a better insight into the input state than what the mere "winner" can give. If then the models fall into classes that are known a priori, the sums of the weights over each class can be interpreted as expressing the affiliation of the input with the due classes.

1 Introduction

When an unknown input pattern is presented to the *Self-Organizing Map*, called briefly the *SOM* (Kohonen, 2001), the algorithm returns and displays the location of the best-matching model, the "winner."

However, it has long been felt that a single response, the "winner," to an input pattern is not sufficient for the description of input information that is derived from two or more independent sources, for instance when the input describes features of several distinct objects that occur simultaneously. Consider also that if the SOM is used to track sequences of the states of a system or a machine, the input vectors are usually constructed as averages of measurements made over time windows of considerable length, say, minutes or even hours. If the system is just then undergoing a state transition, the input vector, averaged over the time window, is expected to be a linear mixture of the samples, eventually representing different states that have occurred within that window.

Another reason for mixtures of input states to occur is that the system from which the input observations are derived is defined by several more or less independent state variables. Their deviations from regular values are caused by faults or noise. The simultaneous occurrence of two or more independent faults in different parts of the system might then be reflected as a mixed input state of the SOM, and thus as a mixture of input signal components, each of which corresponds to an individual fault.

A very intriguing case occurs in document analysis that is based on the usage of typical words in the text. An interdisciplinary or a multidisciplinary document is expected to contain words from the different vocabularies of its subtopics. One may then be interested in the relative contributions of the mixed topics in the document.

It will be shown in this article that a more thorough description of the input is obtainable if, instead of determining only the "winner" among the models, one is able to fit a *linear mixture of the models* (reference vectors) to the input. It should be noted that I do not mean "K winners," which are rank-ordered according to their matching, nor a set of "parallel winners," each of which is defined over a local area of the SOM. Instead, the input pattern is approximated by the linear mixture of *any* subset of models (i.e., reference vectors) that approximates to the particular input best.

In this paper we shall show, however, that if only *nonnegative weights* are allowed in the fitting of the models, there will be left only a relatively very small number of models in the mixture, i.e., the fitting is very selective.

Since we are going to deal with linear-fitting problems in the sequel, it may seem proper that the SOM should be of the *dot-product* type, in which the matching of the input vector with the model vectors is measured in terms of their dot products. Then the input vector and all of the weight vectors shall be *normalized*, say, to unit length. Nonetheless there are no restrictions in principle to the application of the same ideas to SOMs that have been constructed on the basis of the Euclidean or any other metric. Even in these cases, however, before using this method, it will be mandatory to normalize the SOM model vectors.

Let us regard the input as a Euclidean vector $\mathbf{x}$, and let its dimensionality be n. In matrix calculations $\mathbf{x}$ shall be regarded as an n-dimensional *column vector*. Let us then

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

denote the SOM by a p times n matrix **M**, where p is the number of the *models*. If the *model vectors* (regarded as *column vectors*) of dimensionality n are denoted as $\mathbf{m}_i$ = 1, 2, … , p, they constitute the *rows* of **M** and must then be denoted by $\mathbf{m}_i'$, where the prime (´) denotes the *transpose* of a vector or a matrix. All of the $\mathbf{m}_i$ shall have an identical Euclidean norm. In the *dot-product SOM*, the "activation" of the models, or the *degree of matching* of **x** with the $\mathbf{m}_i$, is thought to be represented by the p-dimensional *activation vector* (column vector) **y**,

$$\mathbf{y} = \mathbf{M}\,\mathbf{x}\,. \qquad (1)$$

The largest component of **y** identifies the best-matching model, the "winner."

Comment. Graphically, the SOM is most often defined as a two-dimensional array of nodes where with each node i, a model $\mathbf{m}_i$ is associated. During learning, the models in this array interact in such a way that the highest-activated cell imposes corrections on its neighboring models in the array in the same direction. One must clearly realize the distinction between the *rows* of *matrix* **M** and the *2-D geometry* of the *SOM array*, however.

2 Failure of the Unconstrained Linear Fitting

Let us tentatively try to fit the *best linear mixture* of any given *reference vectors* $\mathbf{m}_i$ to a given vector **x**. In other words, we want to determine the optimal scalar coefficients k_i in the following error expression **e**, whereupon the Euclidean norm of **e** shall be minimized:

$$\mathbf{e} = k_1\,\mathbf{m}_1 + k_2\,\mathbf{m}_2 + \ldots + k_p\,\mathbf{m}_p - \mathbf{x}\,. \qquad (2)$$

Let **k** be the *column vector* formed of the k_i,

$$\mathbf{k} = [k_1, k_2, \ldots, k_p]'. \qquad (3)$$

The linear mixture of the $\mathbf{m}_i$ can be written, using matrix expressions, as

$$k_1\,\mathbf{m}_1 + k_2\,\mathbf{m}_2 + \ldots + k_p\,\mathbf{m}_p = \mathbf{M}'\,\mathbf{k}\,, \qquad (4)$$

where **M´** is the transpose of the matrix **M** that has the reference vectors as its rows; the latter are identified as the $\mathbf{m}_i'$. Now **M´ k** is the *estimate* of **x**. If the fitting error is written as

$$\mathbf{e} = \mathbf{M}'\,\mathbf{k} - \mathbf{x}\,, \qquad (5)$$

the square of the Euclidean norm of **e** is

$$\mathbf{e}'\,\mathbf{e} = \mathbf{k}'\,\mathbf{M}\,\mathbf{M}'\,\mathbf{k} - \mathbf{k}'\,\mathbf{M}\,\mathbf{x} - \mathbf{x}'\,\mathbf{M}'\,\mathbf{k} + \mathbf{x}'\,\mathbf{x}\,. \qquad (6)$$

It is generally known that **e´e** is minimized when its *gradient* with respect to **k** is equal to zero. This gradient (cf., e.g., Kohonen, 2001, Ch.1) is

$$\mathrm{grad}_{\mathbf{k}}\,(\mathbf{e}'\mathbf{e}) = 2\,\mathbf{M}\mathbf{M}'\,\mathbf{k} - 2\,\mathbf{M}\,\mathbf{x}\,. \qquad (7)$$

From the expression (7) we may try to solve for **k**:

$$\mathbf{k} = (\mathbf{M}\mathbf{M}')^{-1}\,\mathbf{M}\,\mathbf{x}\,. \qquad (8)$$

Unfortunately, the expression (8) can only be computed if $(\mathbf{M}\mathbf{M}')^{-1}$ exists, i.e., if the determinant of **MM´** is nonzero. A necessary condition for it is that all of the $\mathbf{m}_i$, i = 1, 2, …, p are *linearly independent*. For the SOM matrices this is usually not the case.

Even though we would have a SOM in which the dimensionality of the vectors and the number of the models were identical, and even though $(\mathbf{M}\,\mathbf{M}')^{-1}$ would exist, we may discern a weird result. If the input **x** is an arbitrary vector, some of the k_i in its linear mixture may attain very large values (say, some thousands when the vectors are normalized). *The fitting may be perfect, but it makes no sense.* This kind of a problem is called "overfitting."

3 Fitting with Nonnegative Weighting Coefficients

Much attention has recently been paid to least-squares problems where the fitting coefficients are constrained to *nonnegative* values. Such constraints are natural, when the negatives of the samples have no meaning, for instance, when the input consists of statistical indicators that can have only nonnegative values, or is a weighted word histogram of a document. In these cases at least, the constraints contain additional information that is expected to make the fits more meaningful. In any case we are able to circumvent the above "overfitting" problem, if we forbid the (eventually large) negative weights.

The mathematical problem is formulated as follows: for general dimensionalities of **M** and **k**,

$$\text{minimize}\;\; \mathrm{norm}(\mathbf{M}'\,\mathbf{k} - \mathbf{x}) \qquad (9)$$

subject to the condition that all of the elements of **k** are nonnegative. In this work the norm is Euclidean.

Gradient-descent optimization. There exist several ways for the solution of (9). The simplest and most straightforward is the *gradient-descent optimization*. An iterative algorithm that takes into account the nonnegativity constraint can be specified in the following simple way. Denoting the component i of the gradient in eq.(7) by G_i we write

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

$$\text{for all } i, \quad k_i(t+1) = \max(0, k_i(t) - \alpha G_i), \qquad (10)$$

where t is the integer-valued index of the iteration step, and α is a small scalar factor that defines the size of the gradient step. Typically, a few tens of thousands of iteration steps (10) are necessary to reach a reasonably stable solution with a numerical accuracy of, say, three significant digits.

The Matlab function *lsqnonneg*. The present fitting problem belongs to the broader category of *quadratic programming* or *quadratic optimization*, for which numerous methods have been developed over the years. A recent one-pass solution of (9) is based on the *Kuhn-Tucker theorem* (cf. Lawson & Hanson, 1974), but it is too complicated to be reviewed here. Let it be mentioned that it has been implemented as the function named the *lsqnonneg* in the Matlab:

$$\mathrm{K} = \mathrm{lsqnonneg}\,(\mathrm{M}', \mathrm{X}, \mathrm{K}(\mathbf{1}))\,, \qquad (11)$$

where K stands for $\mathbf{k}$, $\mathbf{M}$ has been denoted by M and $\mathbf{x}$ by X, respectively, and K$(\mathbf{1})$ is the initial value of $\mathbf{k}$ used by the algorithm.

Limitations of the methods. For both of the above methods, the *rank* of the matrix $\mathbf{M}$ can be arbitrary, and solutions exist even though $(\mathbf{MM}')^{-1}$ does not. Nonetheless one can easily see that there are theoretical cases where the optimal solution is not unique, for instance, when some of the $\mathbf{m}_i$ *in the final optimal mixture* are *linearly dependent*. This condition can be checked by first forming the matrix $\mathbf{M}_1$ of those models, the corresponding coefficients k_i of which are positive, and then computing the determinant det($\mathbf{M}_1\,\mathbf{M}_1'$). If it is zero, the components in the linear mixture are linearly dependent. Such a case will also be manifested by a slow-down of the computing time in both of the above methods.

4 Applications

4.1 Cellular-Phone Data

The first practical example describes the performance of a cell in a cellular-telephone network. The input vector to the SOM was defined by 22 variables that describe the *key performance indices (KPI*) such as signal qualities in inward and outward transmission, frequencies of breaks in operation relating to different kinds of faults, and loadings of the cell. We had data from 110 cells available, and each one of the records was an average of the respective measurement or evaluation over an hour. This example is from cell No. 50 during 879 hours of uninterrupted operation. The particular SOM constructed for this study consisted of 80 models with the dimensionality of 22.

A comparison of the different responses to the input vector has been presented in Fig. 1 for five successive sampling intervals (Nos. 10 through 14). The algorithm thereby used was the *lsqnonneg.* On the first row we see the degree of matching (dot product) of the various models with the input. The second row shows the location of the ”winner” on the SOM. The third row illustrates the weighting coefficients k_i, displayed on the SOM groundwork. The shade of the dots corresponds to the value of k_i. One can discern that the resulting mixtures that describe the input states are not very complex: on the average, they contain only about five per cent of all models in this application.

Table 1 shows the quantization and fitting errors for five samples that were not involved in the computation of the SOM. Let us recall that the expression of the former error is norm($\mathbf{m}_c - \mathbf{x}$), where $\mathbf{m}_c$ is the “winner,” and that of the latter is norm($\mathbf{M}'\mathbf{k} - \mathbf{x}$), respectively.

Table 1: The quantization error and the fitting error for five successive sampling intervals of the mobile-phone data.

Interval No.	Quantization error	Fitting error
10	.5659	.3072
11	.2996	.2272
12	.1777	.1270
13	.1550	.0907
14	.1694	.1435

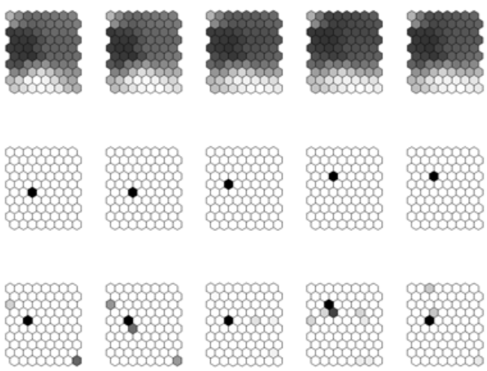


Fig. 1. Comparison of different displays. Horizontal direction: five successive sampling intervals. First row: The activation vectors $\mathbf{y} = \mathbf{Mx}$ averaged over one-hour sampling periods. Second row: The locations of the “winners” on the SOM. Third row: the fitting coefficients k_i, shown on the SOM groundwork.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

4.2 Document Analysis

In the second and third application, the input item was a weighted histogram of the vocabulary of a given document. The objective was to perform a *text analysis*, in order to discover to what extent the vocabularies of the given document and those of the other documents overlap. It must be emphasized that only the *most distinctive words* were taken into account in the overlapping vocabularies.

4.2.1 The Reuters Corpus

The first more extensive experiment of this type was based on the text corpus collected by Reuters. No original documents were made available, but Lewis et al. (2004) have preprocessed the textual data by removing the *stop words*, and reducing the words into their *stems*. J. Salojärvi from our laboratory selected a 4000-document subset from this preprocessed corpus, restricting only to such documents that were classified into one of the following categories:

1. Corporate-Industrial.
2. Economics and Economic Indicators.
3. Government and Social.
4. Securities and Commodities Trading and Markets.

Salojärvi then picked up those 1960 words that appeared at least 200 times in the selected data and weighted the word i (”term”) of document j by the factor

$$w_{ij} = (1 + \log(TF_{ij})) \log(N / DF_i), \tag{12}$$

where TF_{ij} is the ”term frequency” (frequency of word i in document j), DF_i (”document frequency”) tells in how many documents word i appears, and N is the total number of documents (Manning & Schütze,1999).

Labeling of the SOM models. The number of input items used for training was only about twice the number of models, so it was not reasonable to label the models according to the majority of hits on them. Instead, each model was labeled according to its K nearest neighbors in the input space of the training data, where K was initially taken as 10 and, for each model separately, increased gradually only in the case of *ties* in the determination of the majority of labels.

Fig. 2 shows the labeled class regions on the SOM, using four different shades.

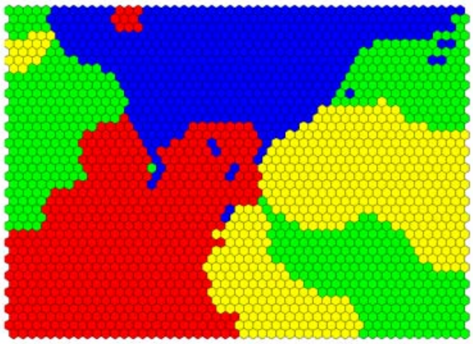


Fig. 2. Labeling of the SOM nodes according to the four document classes. The shades correspond to the classes.

Fitting results. The best-fitting linear mixtures of models for three randomly chosen documents are shown in Figs. 3, 4 and 5 on the SOM groundwork, indicating the value of the due k_i by the shade of the corresponding SOM location, respectively.

The quantization error and the fitting error for the three documents are given in Table 2. Notice that these documents were excluded from the corpus when training the SOM.

Table 2: The quantization error and the fitting error for three documents.

Document No.	Quantization error	Fitting error
101	.8541	.8036
201	.8084	.7767
901	.8686	.8082

Class mixtures. If it is wanted to evaluate, e.g., the degree of *multidisciplinarity* of a document, one approach is to sum up the k_i of each class separately. Figs. 6, 7 and 8 illustrate, in relation to this approach, the affiliations of the individual documents with the classes 1 through 4, respectively. The first and second documents belonged to category No. 1, and the third document was from category No. 4.

Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

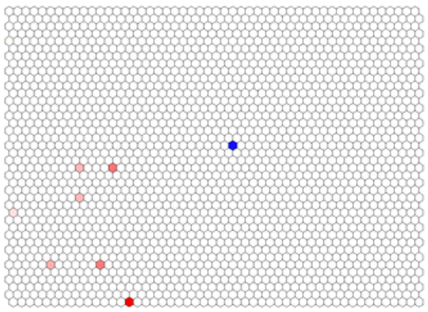


Fig. 3. The linear mixture (with nonnegative coefficients shown as shaded dots) of models that fits best with document No. 101 in the corpus.

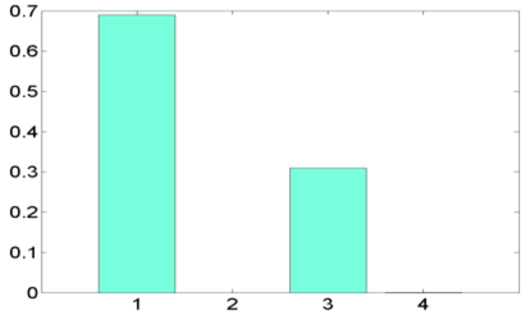


Fig. 6. Relative contributions of the vocabularies of the four categories to the vocabulary of document No. 101. This document belonged to category No. 1.

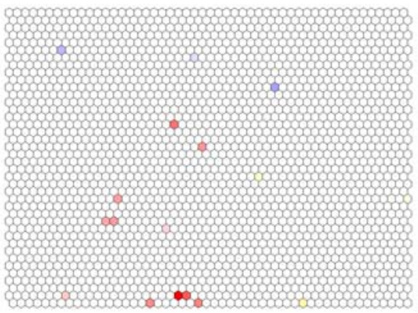


Fig.4. The linear mixture (with nonnegative coefficients shown as shaded dots) of models that fits best with document No. 201 in the corpus.

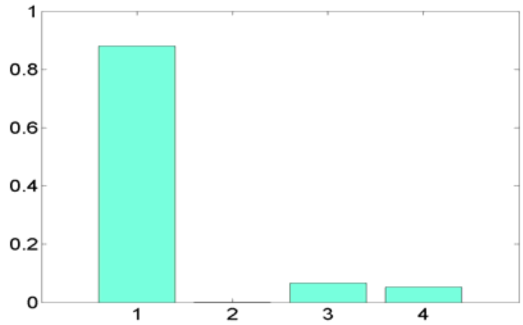


Fig. 7. Relative contributions of the vocabularies of the four categories to the vocabulary of document No. 201. This document belonged to category No. 1.

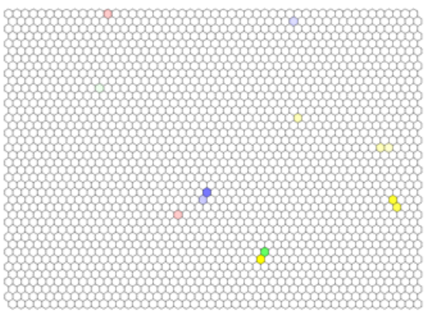


Fig. 5. The linear mixture (with nonnegative coefficients shown as shaded dots) of models that fits best with document No. 901 in the corpus.
.

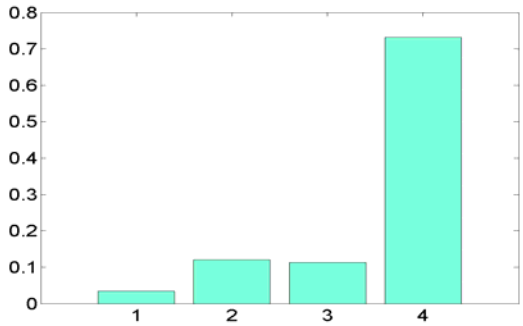


Fig. 8. Relative contributions of the vocabularies of the four categories to the vocabulary of document No. 901. This document belonged to category No. 4.

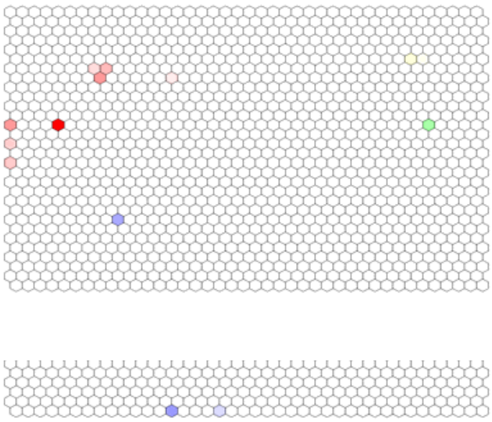
Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

4.2.2. Applications to the Academy of Finland

The Academy of Finland made 3224 grant applications from the years 2005 and 2006 available to us, in order to find out, e.g., to what extent the area of research is reflected in the textual content of the research plan, and to analyze the degree of multidisciplinarity of the applications. There were 44 ”official” research areas into which the applications were classified.

The text mining procedure was almost similar to that used with the Reuters corpus. The selected vocabulary consisted of 1200 words, and the size of the SOM was 1200, too.

Figs. 9 and 10 exemplify the linear mixtures of models that describe the applications Nos. 109706 and 109707, respectively.



109706	.8820	.7722
109707	.6332	.5482

Examples of Classification into Research Areas. Figs. 11 and 12 give the degrees to which a given application is confused with the different research areas, as analyzed by text mining. The columns of a histogram indicate to what extent the weighted vocabulary of an application resembles to those of the 44 research areas.

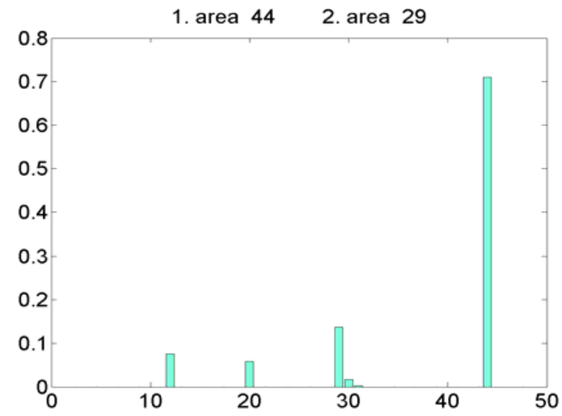
The applicants were also asked to give the primary and the secondary research area of their proposal, which have been exemplified on the top of Figs. 11 and 12. It turned out that their choices coincided with the results given by the text mining to an accuracy of about 70 per cent. There were several reasons for the remaining discrepancies:

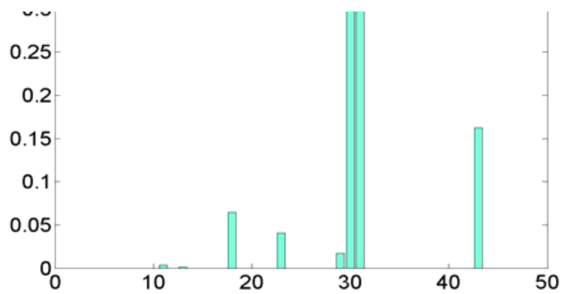
(i) Some of the 44 research areas were very similar, e.g., there were three areas connected with biochemistry, and five areas connected with the environmental sciences, and the terminologies in each subgroup were rather similar.

(ii) The choice for the primary and secondary research areas made by an applicant was completely subjective.

(iii) Each area was compared with 43 other areas, so even though there is only a small mutual overlap in the weighted vocabularies of two areas, there are many possibilities for the confusion of their vocabularies with the 43 other areas. All of these paired differences, however, sum up to the total error. A more realistic measure might be the average confusion with a different area.

Correlation coefficient. Another evaluation of the accuracy of this method is obtained if one compares the distribution of the primary research areas of the 3224 applications with the sum of all histograms. The correlation coefficient of these two “distributions” was .98.





Proceedings of the 6th International Workshop on Self-Organizing Maps (WSOM 2007)
Published by the Neuroinformatics Group, Bielefeld University, Germany, ISBN 978-3-00-022473-7
All contributions to WSOM 2007 are available online at: http://biecoll.ub.uni-bielefeld.de

5 Conclusion

The purpose of this paper has been to extend the use of the SOM by showing that instead of a single ”winner,” one can define several “outputs” that together describe the input pattern more accurately. These “outputs” are defined to be the components in the linear mixture of SOM models that approximate to the input best in the sense of least squares. Only nonnegative weights in the fitting are allowed.

In the light of the above experiments it looks evident that the approximation of the input by the optimized nonnegative-coefficient linear mixture of the SOM models contains more information than the mere location of the best-matching model can give. It is striking how few nonzero components are contained in the optimized mixture.

If the models fall into classes that are known *a priori*, the sums of weights of the models over each class can be interpreted as expressing the degree to what the input is affiliated with the various classes.

An early version of this work has been published in (Kohonen, 2007).

Acknowledgements

The author is obliged to Kimmo Raivio and Jarkko Salojärvi for making the cellular-phone and Reuters data available for these experiments, and Lasse Lindqvist for helping in the processing of them. The Academy of Finland made their data available for the experiment reported in Sec. 4.2.2 , and Timo Honkela as well as Mikaela Klami preprocessed these data. This work was carried out under the auspices of the Academy of Finland and Helsinki University of Technology.

References

[1] Kohonen, T., Self-Organizing Maps, 3rd Edition, Springer-Verlag, 2001.

[2] Lawson, C.L. and Hanson, R.J., Solving Least-Squares Problems, Prentice-Hall, 1974.

[3] Lewis, D.D., Yang, Y., Rose, T.G., and Li, F.: RCV1: A New Benchmark Collection for Text Categorization Research. J. Machine Learning Research 5, 361-397, 2004.

[4] Manning, C.D. and Schütze, H., Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, Mass., 1999.

[5] Kohonen, T., Description of Input Patterns by Linear Mixtures of SOM Models, Report E8, Helsinki University of Technology, Publications in Computer and Information Science, Espoo, Finland, 2007.