ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Self-Organization Algorithm to Process Data With Missing Values and Estimation

Mukta Agarwal

Lecturer, Dept. of Computer Science, Vidya Bhawan Rural Institute, Udaipur (Raj.), India

ABSTRACT: We show how you can use the self-organization algorithm Kohonen to process data with missing values and estimate them. After methodological reminder, we illustrate our subject from three applications to actual data

KEYWORDS: Imputation Data analysis, Kohonenmaps, Missing Data

I.INTRODUTION

The processing of data with missing observations is a concrete problem and always embarrassing when it comes to actual data. Indeed in applications, it is very often in the presence of observations for which is not available to all Descriptive values of variables, and this happens for many reasons: errors seizure not indicated topics in surveys; outliers are preferred delete data easily collected, not available official statistics, etc. Most statistical software (such as eg SAS) suppress purely and simply incomplete observations, but if it has no practical consequences when has extensive data, this can remove any interest in the study if the number of remaining data is too low. To prevent and remove data can be replaced by a missing value average of the corresponding variable, but this average may be a very bad approximation in the case where the variable has a high dispersion.. We focus particularly here in the Self organization algorithm

II.RELATED WORK

In the following, we presenting an example of real data. This is a classic example of data analysis, taken from Bouroche and Saporta, "Data analysis" (1980). This is the structure State spending, measured over 24 years between 1872 and 1971 by a vector of dimension 11. In this example, we have artificially suppressed values in the original data, worth about 11-8 of 11 values, randomly selected for assess the accuracy of the estimates obtained by replacing these values with the values associated vectors corresponding codes.

III.ADAPTATION OF A Self organization Algorithm

We assume that the observations are real-valued vectors of dimension p
When exhibit incomplete data vector x, one first determines the set Mx numbers of the missing components. Mx is a subset of {1, 2, ..., p}. Si (C1,C2... Cn) Is the set of code vectors at this time, the code vector is calculated winner $C_{i(0)(x)}$associated with x and its class, placing

$$i_0(C,x) = Arg\min_{i} \|x - C_i\|$$

where the distanceis calculated on these componentsin the vector x.

They can be used $\|x - C_i\|^2 = \sum_{k \in M_x} (x_k - C_{i,k})^2$ vectors with missing data in two ways.

If you want to use when building code vectors, eachstep, once determined the number of the winning unit, the update of code vectors (theWinner and neighbors) concerns only the components present in the vector.
If there is sufficient data to dispense incomplete vectors to build the map, you can also simply classify, after construction of the map,incomplete vectors affecting the class whose code vector is the closest,the sense of distance restricted to these components.

This gives excellent results in the course that the variable is not or almost completely absent, and also insofar as the variables are correlated,which is the case in most real data sets.

IV.ESTIMATING MISSING VALUES

Whatever the method used to use the data with missing values, the most interesting properties of the algorithm, and it is possible to estimate a posterior I missing values If *Mx*is the set of numbers of the missing components of the observation *x*, and *x* is classified in class *i* for each index *k M* It is estimated *x k* b.

$$\hat{x}_k = C_{i,k} \cdot$$

As the end of the Self organization learning algorithm is to "zero neighbor," we know the code vectors are asymptotically close to the average of their class. This estimation method is thus to estimate the missing values of a variable by its average in the class.

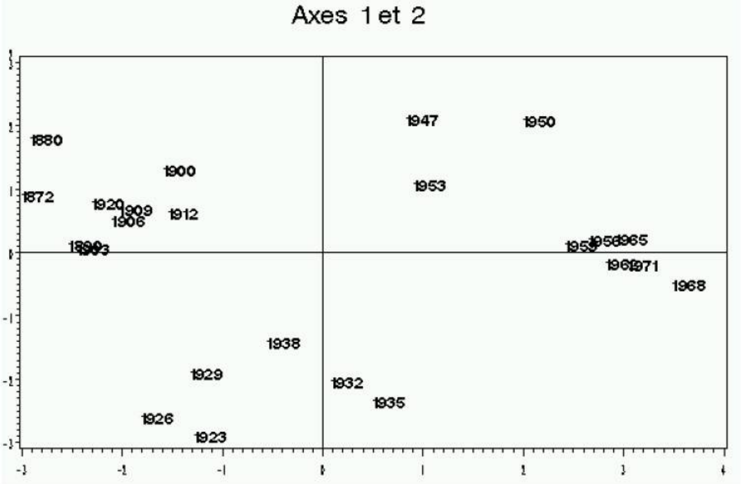
Clearly, this estimate is more accurate than classes formed by the algorithm are homogeneous and well separated from each other. Numerous simulations showed as in the case of synthetic data than real data, in the presence of correlated variables, the accuracy of these estimates is remarkable.

To increase accuracy, it offers in his thesis to produce several versions of the Self organization algorithm, and take the average of the estimates obtained in each card.

V. STRUCTURE OF THE EXPENSES OF THE STATE, FROM 1872 TO 1971

For 24 years, separated into 3 categories (14-18 before the war, between the wars, after WWII) was measured 11 variables representing the state expenditure in different sectors: Public authorities, Agriculture, Commerce and Industry, Transport, Housing and Spatial Planning, Education and Culture, Social Welfare, Veterans fighters, Defense, Debt, Miscellaneous? So this is a small example, with 24 observations of 11 dimensions, without values missing.

A simple analysis of main components provides excellent representation even in two dimensions (64% of variance explained). See Figure 6, the representation of variables and FIG, the representation of years.



Representation of years on the first principal plane. We distinguish
Perfectly all three groups.

DOI: 10.15680/IJIRCCE.2016. 0406194

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

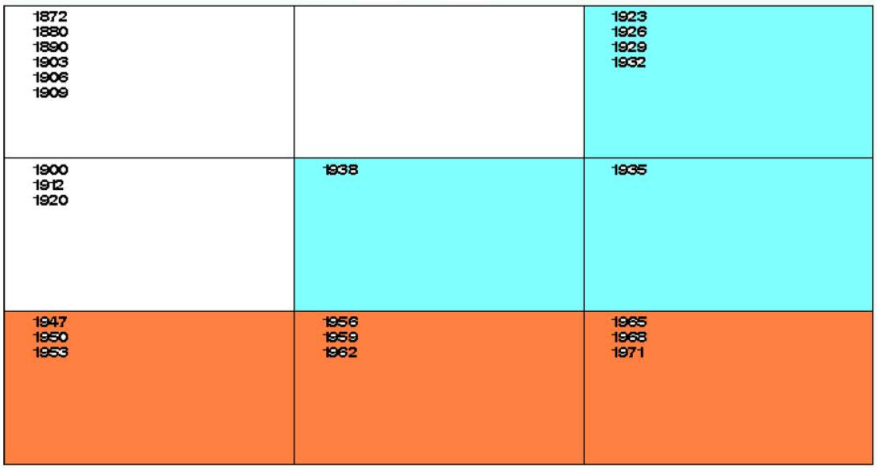
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

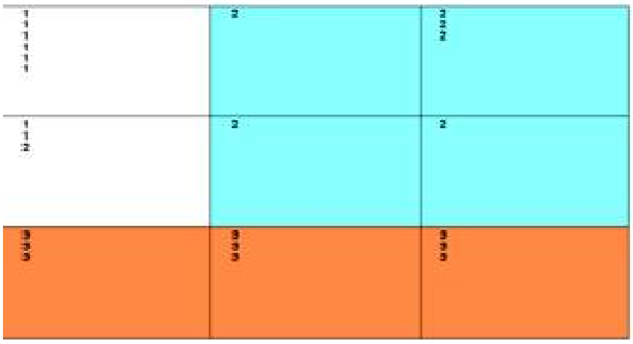
Note that years fall into three groups, corresponding to the three clearly defined periods (before World War I, between the wars, after WWII). Only the year 1920, the first year it appears post Expenditure devoted to veterans is placed with the first group, whereas it belongs to the second.

The years are shown in a Self-organization map of size 3 3 and observes the same groupings. Years were grouped into three classes with a hierarchical classification code vectors



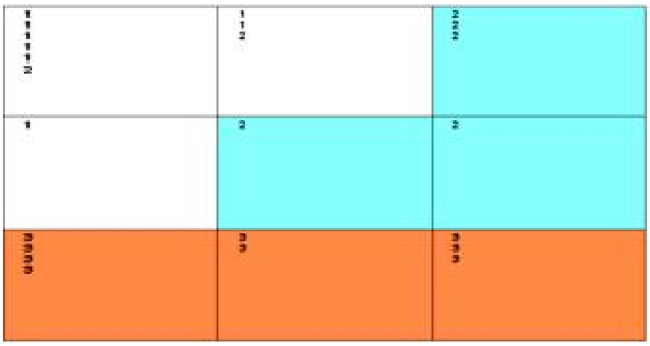
1872 1880 1890 1903 1906 1909		1923 1926 1929 1932
1900 1912 1920	1938	1935
1947 1950 1953	1956 1959 1962	1965 1968 1971

In the following figure shows the Self organization maps obtained after artificially deleted a number of values. Instead of writing the precise date, we Note 1, 2 or 3, depending on the period. Is removed from January to August values per year (about 11 in total) randomly. The maps obtained with the grouping into three classes by the method of hierarchical classification are presented in Figure.4



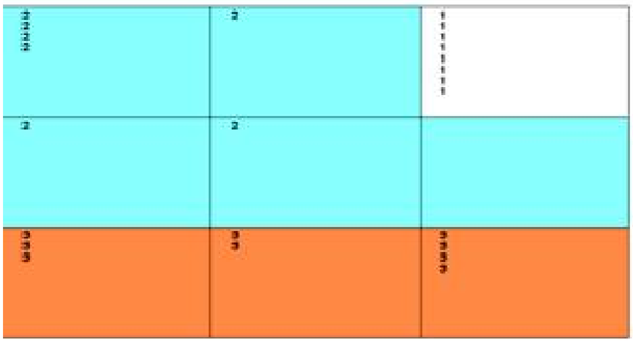
1 1 1 1 1 1	2	2 2 2
1 1 2	2	2
3 3 3	3 3 3	3 3 3

1 suppressed value

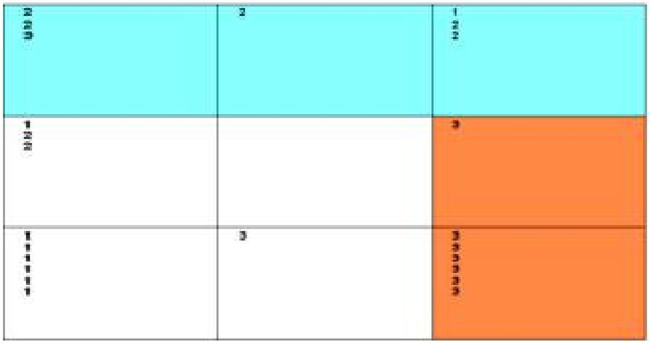


1 1 1 1 1 2	1 1 2	2 2 2
1	2	2
3 3 3 3	3 3	3 3 3

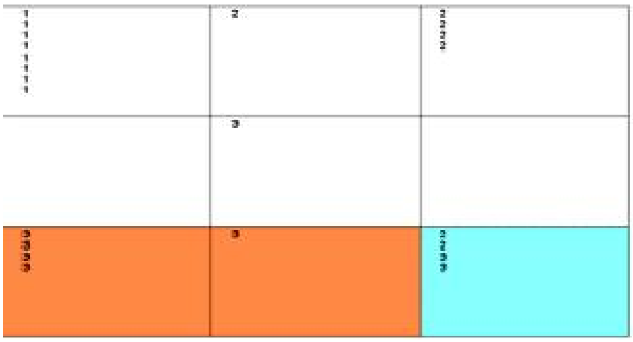
2 suppressed values



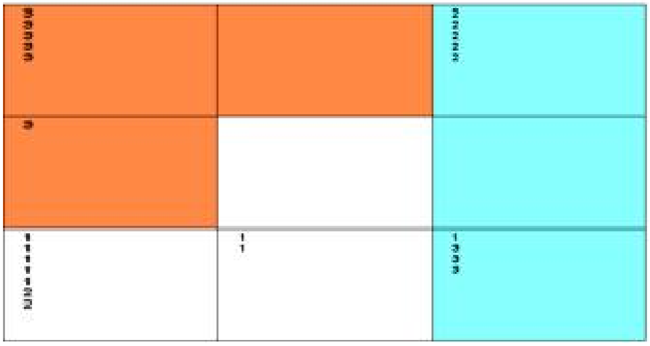
3 suppressed values



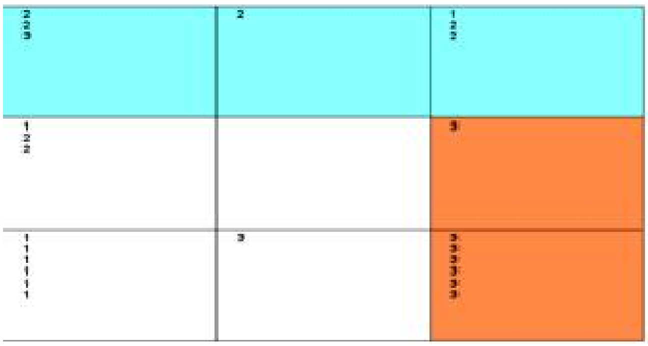
4 suppressed values



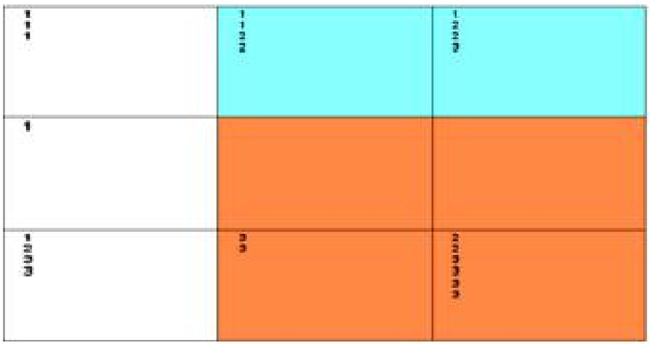
5 suppressed values



6 suppressed values



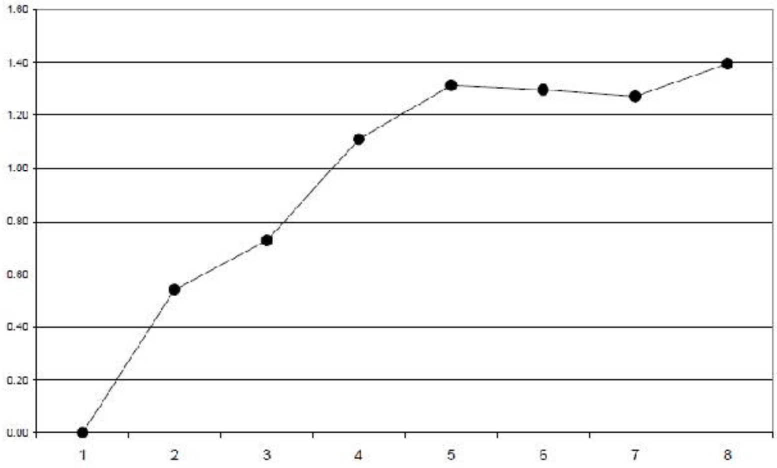
7 suppressed values



8 suppressed values

We see that the super-classes remain consistent as long as does not remove more 3 values per year, or 27% of values. Then classes mingle years. The years are marked 1, 2, 3, following the period.
Then estimated in each case the values that had been deleted. The following table shows in each case the mean square error. Figure 10 shows the evolution of this error based on the number of deleted values

Val Number	1	2	3	4	5	6	7	8
	0.39	0.54	0.73	1.11	1.31	1.30	1.27	1.39



Mean squared error estimation based on the numberdeleted values in each year

it is found that the error remains very low when no longer removes 3 values per year. Then the error ceiling, this is due to the fact that the number of data (24) and vectors codes (9) are both low. The components of vectors are available codes limited number, and the estimates are in this case in all these components.

VI.CONCLUSION

So we showed these three examples how it is possible and desirable to use Self organization maps when the available data have missing data. Good certain estimates and classes obtained will be more relevant than variables Descriptive data are well correlated.
Example shows how this method allows estimating missing data accuracy. The data thus completed can then be subjected to any conventional treatment.

REFERENCES

1) Blayo, F., Demartines, P. (1991): Data analysis: How to compare neural networks Kohonen --other to techniques? In Proceedings of IWANN'91 Ed. A.Prieto, Lecture Notes in Computer Science, Springer-Verlag, 469-476.
2) Bouroche and G. Saporta (1980): Analysis of the data, What do I know? PUF, Paris.
3) Barron A. Universal approximation bounds for superpositions of a sigmoidal function I.E.E.E. Transactions on Information Theory, 39, pp. 390-945. (1993).
4) Benaïm M., Fort J.C, Pages. G., Convergence of the one-dimensional Kohonen algorithm, advanted appl. Prob. 30, pp. 850-869
5) Bishop C.M., Svensén M., Williams C.K.I. GTM: a principled Alternative to the self-organizing map In advances in Neural Information Proceedings Systems, Editor : MC Mozer and M.I Jordan and T. Pitche, pp. 354-360 (1997).
6) Bishop C. M., Tipping M.E., Jordan M. I. Hierarchical latent variable model for a data visualization IEEE Transactions on Pattern Analysis and Machine Intelligence 20, pp. 281-293
7) Bolzern, P. and Fronza, G. Role of Weather Inputs in Short-Term Forecasting of Electric Load, Electric Power and Energy Systems, 8. 1, pp. 42-46. (1986).
8) Bouton C., Pagès G. Self-organization of the one-dimensional Kohonen algorithm with non uniformly distributed stimuli, Stochastic Process. Appl. 47, pp. 249-274 (1993).
Bouton C., Pagès
9) Cottrell M., E. de Bodt, Verleysen M. (2002) A statistical tool for Assessment to the reliability of Self- Organizing Maps, Neural Networks, Vol 15, No. 8-9, p967-978.
10) Mr Cottrell, S. ibbou, Letrémy P. Rousset P. (2003): self-organizing maps for analysis Exploratory data and visualization, to be published in the Journal of the French Society Statistics.
11) Ibbou S., (1998): Classification, analysis of neuronal connections and methods, thesis, University Paris 1.
12) Gaubert, P., ibbou, S., Tutin, C. (1996): Real Estate Segmented Markets and Price Mechanisms: the Case of Paris, International Journal of Urban and Regional Research, 20, No. 2, 270-298.