CHEMICAL ENGINEERING TRANSACTIONS

VOL. 43, 2015

A publication of
AIDIC
The Italian Association
of Chemical Engineering
Online at www.aidic.it/cet

Chief Editors: Sauro Pierucci, Jiří J. Klemeš
Copyright © 2015, AIDIC Servizi S.r.l.,
ISBN 978-88-95608-34-1; ISSN 2283-9216

DOI: 10.3303/CET1543211

# Workplace Accidents Analysis with a Coupled Clustering Methods: S.O.M. and K-means Algorithms

Lorenzo Comberti*, Gabriele Baldissone, Micaela Demichela

SAfeR, Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, Corso Duca degli Abruzzi, 24 – 10129 Torino, Italia
lorenzo.comberti@polito.it

Occupational accident databases are widely used by Workers' Compensation Authorities and private Safety and prevention Management with different purposes. A systematic accidents reporting leads to a large and complex data base where each element is characterized by many parameters and dealing with this amount of information becomes hard. Data mining techniques represent an efficient toll for locate useful information from large databases and in the last 20 years several techniques have had a wide applications in many classification and analysis problems. Among these methods, a coupled clustering, constituted by a projection of data from high-dimensional space to a low dimensional space and a numerical clustering, was presented in 2009 and performed promising results. Carrying on with this approach, this paper introduce a new release of the method that allows to exceed some lacks regarding the numerical clustering stability and the result visualisation. The method has been applied successfully to a data base of occupational accident with the purpose of grouping the element according with their similarity and making a clear visualisation of this classification. The capability of this method of grouping data and visualize them represent a powerful toll for analyst that have to deal with large occupational data base and it can represent a flexible support in the preventive measurement designing.

## 1. Introduction

Occupational accidents still represent a national problem despite the legislative and technological actions made in the last 20 years, the Italian Workers'Authority (INAIL) still reports in 2013 over 660 fatalities occurred in working activities.

The accident investigation is a crucial field for a safety and prevention management system because understanding how an accident occurred can help how to avoiding that type of accident in the future.

Analysing event by event and consequently designing and activating single countermeasure represent a preventive strategy commonly adopted but with the limit of not allowing a systematic approach.

The systematic reporting and collecting of occupational accidents in an organised data base represent a measurement of the safety level of a system and not only a collection of single events.

A system of preventive countermeasures should be better designed after a systematic analysis of all the occupational accident reported in the domain of interest.

This approach collides against the complexity of the system that must be analysed.

An occupational accident data base is generally characterised by tree factors of complexity: the high data number and the elevated number of descriptive parameters for each accident generate an high dimension domain in addition to this, the categorical nature of the descriptive parameters makes more difficult to deal efficiently with.

Manage this large amount of information with traditional statistics methods like multivariate analysis or linear regression require plenty of a-priori assumptions over the distribution of the variables involved and became heavy to apply.

A consolidated alternative for extracting information from a complex data base is presented by data mining approach that in the last 20 years have had a wide applications in many classification and analysis problems

Please cite this article as: Comberti L., Baldissone G., Demichela M., 2015, Workplace accidents analysis with a coupled clustering methods: som and k-means algorithms, Chemical Engineering Transactions, 43, 1261-1266 DOI: 10.3303/CET1543211

regarding databases of occupational accidents, blazes, road and flight accidents as reported in Edelstein (1999) and Zaiane (1999) and more recently in Larose (2005).

Many different algorithms were proposed to approach this problem and belongs them a promising toll was presented by Piccinini and Palamara (2011).

The method was originally designed in two level of clustering: the first level was represented by a S.O.M. (Self Organising Map), an unsupervised learning algorithm for generating topology preserving transformation from a high-dimensional data vector space to a low-dimensional map space.

The second level was represented by a classical numerical clustering algorithm (k-means) and it was used to make a quantitative partition of the domain depending on the level of similarity of the data.

The authors applied this method to a large data base of occupational accident occurred in Italian wood processing industry.

The using of a such coupled method of clustering in risk assessment field has had recently further development in the works of Alikhani (2013) where an hybrid clustering-classification method has been adapted by using k-means and SOM as clustering methods to improve accuracy of classification.

The method proposed by Palamara (2011) was investigated and a sensitivity analysis was performed in order to evaluating his limit and performances.

The results of this studies reveals some heavy numerical stability problems related to the application of the second level of the methods that nullify the reproducibility of the clustering. In addition to this some lacks on the visualisation outputs were noted.

A new release of the method was projected and called SKM (SOM K-means method) and successfully tested.

In section 2 a short description of SKM structure is provided while Section 3 present the results of the application over a case study and then section 4 provided some discussion.

Conclusions and prospects for future work end the paper.

## 2. Methodology

SKM was developed, as proposed by Vesanto (2000), in two level of data elaboration with the aim of combining the SOM analysis in the Kohonen's approach (1990-1997) with a classification algorithm. The SKM was implemented in Matlab 7.0 coding with the support of SOM (2007) training tool.

SKM is structured as it shown in Figure 1.

The method requires a pre-processing phase where each element of the incident matrix must be coded in a numerical vector. The coding is based on the replacement of each descriptive parameters with a vector containing a sequence of zeros and a one according with a fixed tables of conversion. Using of this type of coding ensures that during the analysis the various categories are equidistant and the coding system does not affect the follow steps.

The complete code is given by the union of the vectors that describe the attributes used for the analysis, consequently the result is a vector with many 1 as the attributes and many 0 as the number of categories less the number of attributes.

At the end of the pre-processing phase the Input Matrix(IM) that originally contained a group of described occupational accidents is coded in the Input matrix that contains an equivalent number of numerical vectors.

The first level of clustering represent the SOM elaboration.

In this phase a projection process is made up in order to reflect the data similarity.

By this way accident with equal characteristic are projected in the same map point and accidents strongly different are projected in distant points.

At the end of this process the map obtained is characterized by dark areas, that represent areas with a large number of data projected, and grey hill that represent areas empty of data (Figure 2).

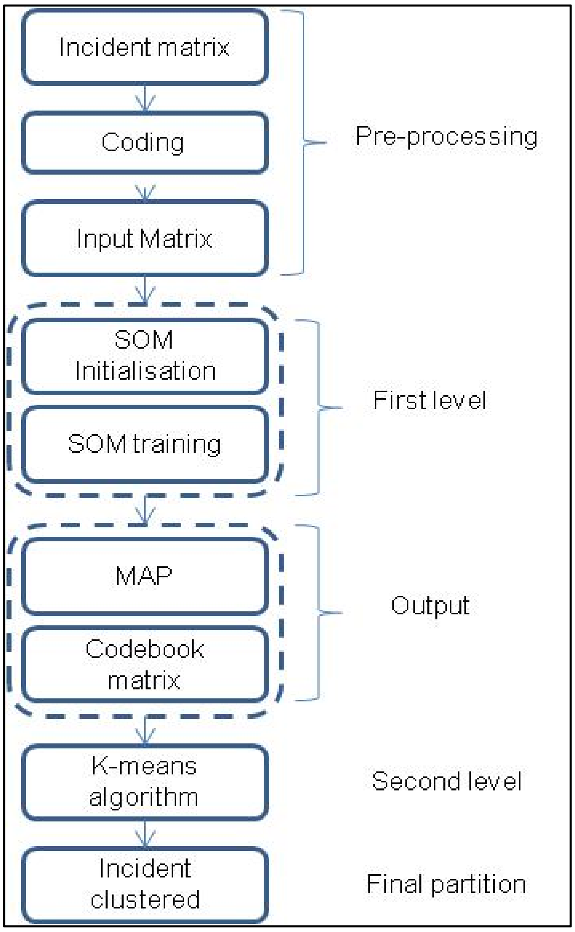


*Figure 1: SKM structure*

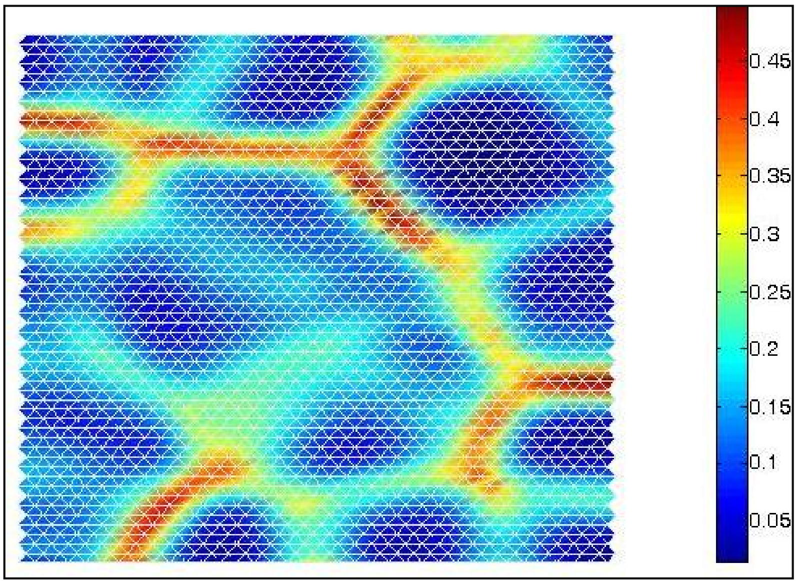


*Figure 2: SOM map elaborated at the end of the first level of clustering*

The number of the dark areas can reveal the number of groups of similar accident and counting them it is possible to estimate the number of cluster for following phase.

The second level is applied on a Clustering Matrix (CM) which was elaborated with a specific routine.

CM contains a number of element equal to the number of IM and each element is defined by a numerical vector, called "prototype", generated by the SOM process.

Each prototype vector, according with Vesanto (2000), contains some numerical values called "weights" that are substantially proportional to the numbers and types of data that are projected in the correspondent unit, as a consequence accident projected in the same unit are characterized by the same prototype vector.

CM matrix and the cluster number estimated by the SOM map evaluation are the input for the K-means clustering phase that provides a data partition that is resumed by a chart where each occupational accident is attributed to a specific cluster.

The second level of the method is also focused on a partition visualization that is shown in figure 3.

The graph illustrate, in the SOM map domain, the correspondent distribution of activated units.

Only activated unit are represented in the graph, units belongings to the same can be identified by a common color and the same cluster number.

In addition to this each unit is marked by the number of element projected.

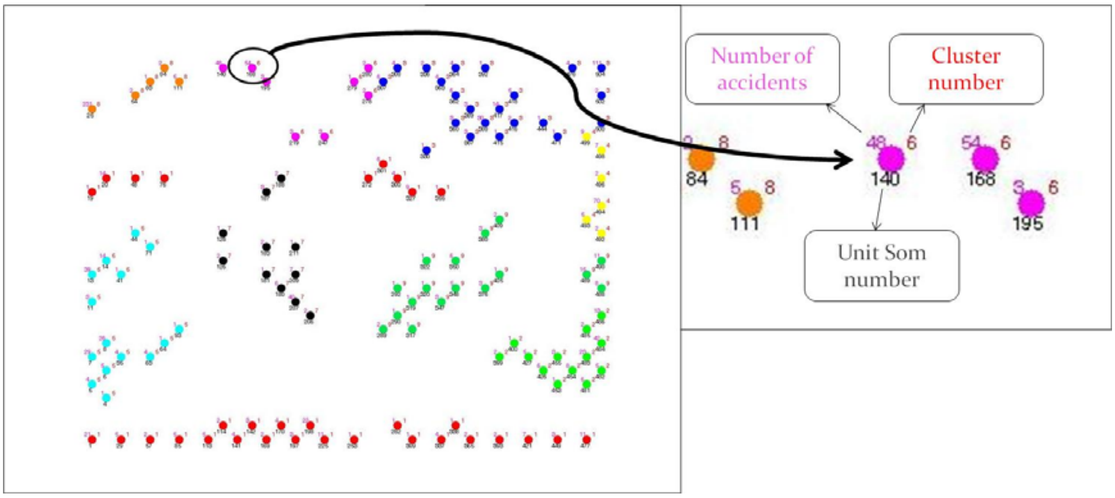


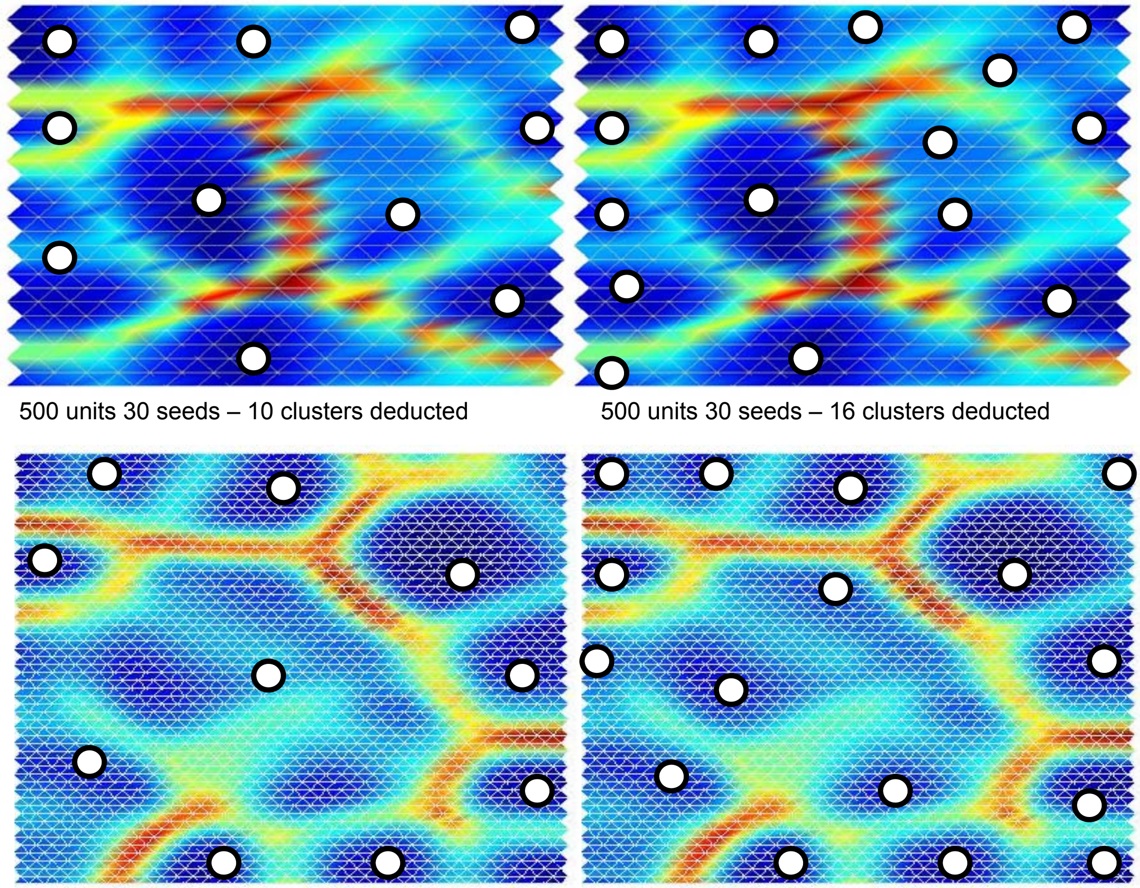
*Figure 3: Clustering partition visualization.*

## 3. Experiments

In order to evaluate the validity of SKM, a set of test was performed with two main several purposes:
verification of the clustering reproducibility and numerical stability of k-means and evaluation of clustering goodness.
The tests were applied versus the same case study presented by Palamara (2011).
A sample of 1247 occupational accident in Italian wood industry was selected from INAIL DB and the attention of the research was focused on the accident dynamic.
As a consequence from the original data set selected from INAIL DB a subset was realized defining each element as a sequence of six categorical parameters that define the sequence of events that lead to the accident as shown in the following sentence:
'During (activity) with (material 1), because of (deviation) with (material 2), the worker injured because he (contact) with (material 3)' .
Each parameter can assume 8 possible values, for example "Activity" can take the label of: "working with machinery", "working with tools', 'driving', 'handling woodworks', 'opening closing packages', 'spilling/filling', 'manual transport', 'moving'.

### 3.1 Clustering reproducibility

The SOM allows to define the clusters number by the visual evaluation of areas with activated units (dark areas).This operation remains strictly subjective because it is strictly related to the skill and the experience of the analyst. The cluster number deduction is still a difficult operation because, as it is shown in figure 4, a map interpretation can suggests different solutions.



500 units 30 seeds – 10 clusters deducted

500 units 30 seeds – 16 clusters deducted

2000 units 30 seeds – 10 clusters deducted

2000 units 30 seeds – 16 clusters deducted

*Figure 4: Cluster number interpretation*

From the previous parametric study the maps generated with 500 and 2000 units and 30 or 2000 seeds were selected in order to be analysed with the k-means clustering.
For each map the CM matrix was calculated and the number of possible clusters fixed, according with the visual interpretation, in 10 and 16.The clustering for each configuration was repeated independently 7 times with the purpose of testing the numerical stability of the methods and different result were obtained.

To compare the clustering results and evaluate easily the numerical stability two new index were defined: the "sequence stability" (Ss) and the "sequence membership" (Sm). The Ss index is calculated for each element and represent the sequence of cluster attribution of that element related to the multiple repetition, the Ss index represent the number of element that have the same Sm index.
Table 1 reports for 5 element the progressive attribution in seven clustering repetition.

*Table 1: Sequence membership*

| | Sequence membership | | | | | | |
|---|---|---|---|---|---|---|---|
| | Clustering repetition | | | | | | |
| **Record** | **1°** | **2°** | **3°** | **4°** | **5°** | **6°** | **7°** |
| 5 | A | A | A | A | A | A | A |
| 2 | A | A | A | A | B | A | C |
| 3 | A | A | A | A | B | A | A |
| 4 | A | A | A | A | B | A | A |
| 1 | A | A | A | A | A | A | A |

The Sm index for the record n. 5 is: AAAAAAA while the Sm for record n. 2 is AAAABAC. A level of Ss equal to 100% is represented by all the elements that have an Sm without changes in attribution.
In other words all the element that are denoted by a stables sequence of clustering. A level of Ss equal to 85% correspond to the number of elements that haves an Sm with at least one variation in cluster attribution. Table 2 shows the Ss index calculated in 4 different initial condition after 7 clustering repetition.

*Table 2: Sequence Stability index*

| | 30 Seeds | | 2000 Seeds | |
|---|---|---|---|---|
| Ss index | 10 cluster | 16 cl. | 10 cl. | 16 cl. |
| 100% | 68 | 85 | 80 | 79 |
| more 85% | 81 | 92 | 95 | 91 |
| more 71% | 88 | 93 | 98 | 96 |

### 3.2 Clustering identification

*Table 3: Clusters quality for 2000 seeds case*

| Cluster | Seq. | Description |
|---|---|---|
| Cl1 | 132 | Movements with woodworks handling and falls to various material |
| Cl2 | 19 | Working with hand tools and damage for impact with projected parts |
| Cl3 | 39 | Transport load loss of control and impact with something in motion |
| Cl3-1 | 29 | Movements of opening/closing doors and crushing impact |
| CL4 | 69 | Working with tolls, incorrect movements , and damage for contact with cutting parts |
| Cl4-1 | 50 | Working with tolls, incorrect movements and damage for contact with cutting machinery |
| CL5 | 103 | Objects handling and damage for loosing control and incorrect movements |
| CL6 | 76 | Manual transport and strains for incorrect actions |
| CL7 | 32 | Dynamic or static action wirh torsional incorrect movement and damage |
| CL8 | 233 | Working with tools, loosing control and damage for contact with cutting elements |
| CL9 | 153 | Load and transport and damage for crushing or impact |
| CL10 | 118 | Working with machinery, loss of control or incorrect movements and heavy cutting damage |
| CL10-1 | 75 | Working with hand tools, loss of control or incorrect movements and heavy cutting damage |

The clustering goodness is evaluated in relation with the evidence that the clusters include accidents with similar descriptive parameters or not. According with the definition of Ss index, it is assumed that are considered stable the cluster that have an Ss of 85% that means in the case reported in Table 2 the clusters with just one changes in attribution. Considering the initial condition of 2000 seeds and 10 clusters the correspondent Ss index of 85% cover the 95% of data examined and Table 3 reports the clusters contents.

The partition (2000 seeds) illustrated in Table 3 was characterized by an high level of clusters homogeneities, generally each clusters was described by a common and synthetics accidental dynamics.

A better partition related to the increasing value of Ss indices was generally observed on the other set tested and to the lower values of Ss index corresponds generally to a more heterogeneous clustering contents.

## 4. Conclusions

A new release of a coupled clustering methodology to deal with occupational accident databases was designed and successfully tested. The work done was also addressed in the research of a quantified description of the clustering results with the purpose of making the analysis less qualitative.

In order to quantify and to manage the numerical variability of repeated clustering two indices were defined: the “sequence stability” (Ss) and the “sequence membership” (Sm).

The using of this parameters can give a quantitative description of clustering and leaded to a new definition of cluster as a “group of element with an assigned sequence stability”.

Ss can be proposed actually as a good test in a relative comparison between more CM and clustering analysis but it is far away to be used as absolute scale of evaluation. These descriptive indices were applied in order to quickly quantify the quality partition and better address the analyst, the results obtained had shown a good performance. The proposed methodology is able to identify families of accident with similar attributes from large and complex databases. This results help analysts in finding the most critical and recurring states in occupational accidents field, this information contribute to designing and better addressing the corrective actions.Future works should be addressed on the application of this methodology to larger and more complex data set.

### References

Alikhani, M., Nedaie, A., & Ahmadvand, A. (2013). Presentation of clustering-classification heuristic method for improvement accuracy in classification of severity of road accidents in Iran. Safety Science, 60, 142-150.

Asgary, A., Sadeghi Naini, A., & Levy, J. (2012). Modeling the risk of structural fire incidents using a self-organizing map. Fire SafetyJournal, 49, 1-9.

Edelstein, H., 1999. Introduction to Data Mining and Knoledge Discovery. Two Crow Corporation, Potomac, Maryland

Kohonen, T. (2001). Self-Organizing Maps. NewYork: 3rdEdition, Springer-Verlag, Inc.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11(3) , 574–585.

Larose, D.T.,2005,Discovering Knowledge in data-An introduction to data mining. JohnWiley & Sons,Inc., New York

López Iturriaga, F. J., & Pastor Sanz, I. (2013). Self-organizing maps as a tool to compare financial macroeconomic imbalances:The European, Spanish and German case. The Spanish Review of Financial Economics, 11, 69-84.

Özkan, G., & Inalb, M. (2014). Comparison of neural network application for fuzzy and ANFIS approaches for multi-criteria decision making problems. Applied Soft Computing, 24, 232–238.

Palamara, F., Piglione, F., & Piccinini, N. (2011). Self-Organizing Map and clustering algorithms for the analysis of occupational accident databases . Safety Science, 49, 1215-1230.

Sèverin, E. (2010). Self organizing maps incorporate finance: Quantitative and qualitative analysis of debt and leasing. Neurocomputing, 73, 2061-2067.

Smith, K. A., & Ng, A. (2003). Web page clustering using a self organizing map. Decision Support Systems, 35 , 245-456.

Vesanto, J. Alhoniemi, E., 2000. Clustering of self-organizing map. IEEE Transactions on Neural Networks 11 (2), 586-600

Zaiane, B., 2003. Binary vector dissimilarities for handwriting identification. In: Proceedings SPIE, Document Recognition and Retrievial X, SantaClara, CA, pp 155-166.