SIMPLE COMPETITIVE LEARNING, KOHONEN S.O.M. AND SCHMITTER-STRAUB'S METHOD: A PROPOSAL IN FINDING "GOOD" TARIFF CLASSES[1]

Renato Pelessoni

Liviana Picech

Dipartimento di Matematica Applicata alle
Scienze Economiche Statistiche ed Attuariali "Bruno de Finetti"

UNIVERSITÀ DEGLI STUDI DI TRIESTE

1. Introduction

In rate making process the statistical information on claim experience are combined with observable variables describing the risks, in order to build a tariff. The observable variables considered in the tariff structure are called tariff variables and the premium for a new risk is simply estimated from the observed values of these variables. In this process, many statistical methods and mathematical algorithms are applied to select the variables and to build the tariff.

Many tariff methods require the values of the tariff variables to be collected in classes. Even if it is not necessary, commercial reasons often suggest making use of a low number of tariff classes. For instance, if "age of the insured" is a motor vehicle insurance tariff variable, it may be better to classify the risks into age classes instead of considering the single ages. Obviously, a question arises: which values of the tariff variable should be grouped together and which not, and also how many classes should be formed. This paper concentrates on these particular aspects of the rate making process.

We suppose to have selected a set of tariff variables. For each of these

[1] This research work was partially supported by C.N.R. (Research Project n. 97.01047.CT10115.21688: "Modelli e metodi per valutazioni in assicurazioni non-vita").

variables, according to notation in [7], we call basic classes the elements of the finest partition of its values (e.g. the age of the insured in whole years) and we call tariff classes the clusters grouping the basic classes for rate making purposes. The determination of tariff classes can be viewed as a clustering problem.

Cluster analysis techniques (for an extensive review see [1], [8], [10]) have a wide range of applications but, as pointed out in [7], they are rarely applied in insurance field. However, suitable adaptations allow their application to collect the basic classes and form homogenous groups, according to one or more characteristic variables describing the claim experience. In the following, for the sake of simplicity and unless otherwise stated, we will consider only one characteristic variable. Turning back to the variable "age of the insured", the claim experience can be described, for instance, by the claim frequency. Ages with "quite similar" claim frequency will be allocated in the same cluster. In [7] Van Eeghen *et al.* discussed the methods proposed by H. Dickmann and by K. Loimaranta *et al.*, which are adaptations of a hierarchical clustering method and of a non-hierarchical method of mixtures respectively.

Partitioning methods of cluster analysis have a wide range of applications, but they also require some adaptations to give suitable solutions to this type of problems. In [9] the Authors proposed to apply, in a neural network framework, some of these methods and introduced the essential adaptations in the involved algorithms.

Generally, different cluster analysis techniques produce different subdivisions of the basic classes in clusters, among which a choice must be done. An actuarial approach to solve this problem was suggested by H. Schmitter and E. Straub [20].

In this paper we examine the results obtained applying the neural clustering techniques developed in [9]. Comparison among different classification results are then performed by means of Schmitter and Straub's method.

An outline of the paper is the following.

In Section 2 some applications of clustering techniques in determining tariff classes are delineated.

In Section 3 the Schmitter and Straub's method is discussed.

In Sections 4 and 5 we briefly recall two neural network algorithms frequently used in clustering problems: simple competitive learning and Kohonen's self-organising map.

Section 6 is devoted to an application of the algorithms quoted in Sections 4 and 5 to collect in clusters the basic classes described by the fiscal power in a motor insurance portfolio. For this purpose the technique developed in [9] is applied.

In Section 7 some final remarks and suggestions for further investigations are resumed.

2. Clustering methods proposed for the determination of tariff classes.

The basic classes can be seen as objects that have to be joined together according to the values of the characteristic variable; from this point of view the problem of determining the tariff classes may be treated simply as a clustering problem. However the observed values of the characteristic variable in each basic class are not immediately comparable by means of a similarity or dissimilarity measure, as usually done in traditional clustering procedures. In fact, these values arise from observations on the insured risks and then they are affected by the exposures in each basic class. For instance, if the characteristic variable is the claim frequency and the basic classes are the values of fiscal power, we will make use of observations from a portfolio of motor insurance risks. For each value of fiscal power, namely for each basic class, we observe the number of claims reported by the risks having that value of fiscal power. Then, we consider their total exposure measured by the policy-years and we can determine their claim frequency, say the observed value of the characteristic variable. It is clear that the basic classes cannot be compared simply by looking at the values of the characteristic variable but also the exposures must be taken into account.

In actuarial literature, two methods have been proposed to determine the tariff classes, respectively by H. Dickmann and by K. Loimaranta, J. Jacobsson & H. Lonka (see [7] for a review). Both methods are adaptations of traditional clustering algorithms. The first one is a hierarchical agglomerative clustering method, whereas the second is a non-hierarchical method of mixtures.

At the beginning of Dickmann's algorithm each basic class can be viewed as a group containing one object so that the within-clusters variance of the characteristic variable can be assumed to be zero. The method consists of a hierarchical agglomerative procedure in which the merging of two groups, at each stage, is done to minimise the increase of the total within-cluster variance. The procedure is repeated until all basic classes are located in one cluster.

As pointed out by Dickmann himself, this method can be seen as an adaptation of the clustering method proposed by Ward [21] to the actuarial problem of determining the tariff classes, taking account of the different exposures of the basic classes simply by means of the definition of within-cluster variance.

For a short description of the algorithm, let us consider a single stage with the basic classes joined together in K clusters. Let

m_k be the number of basic classes located in cluster k;

x_{ik} be the observation of the characteristic variable with respect to the i-th basic class located in cluster k;

t_{ik} be the value which reflects the exposure of the i-th basic class located in cluster k (e.g. the number of observed policy-years);

$n_k = \sum_{i=1}^{m_k} t_{ik}$ be the total exposure in cluster *k*.

Define the within-cluster variance for cluster *k* as:

$$ {}_k\sigma_W^2 = \sum_{i=1}^{m_k} \left(x_{ik} - \bar{x}_k \right)^2 \frac{t_{ik}}{n_k} $$

where

$$ \bar{x}_k = \sum_{i=1}^{m_k} x_{ik} \frac{t_{ik}}{n_k} . $$

The total within-cluster variance with *K* clusters is defined as:

$$ (2.1) \qquad {}^{(K)}\sigma_W^2 = \sum_{k=1}^{K} {}_k\sigma_W^2 \frac{n_k}{n} $$

where $n = \sum_{k=1}^{K} n_k$.

We pass from *K* to *K-1* clusters by merging two of the existing clusters so that the increase of the within-clusters variance

$$ {}^{(K-1)}\sigma_W^2 - {}^{(K)}\sigma_W^2 $$

is minimum.

In the clustering method proposed by Loimaranta-Jacobsson-Lonka it is assumed that *N* basic classes belong to *K* clusters and that the characteristic variables are independent random variables with probability distribution a mixture of *K* distribution, one for each cluster: from this fact the identification of the method as non-hierarchical method of mixtures.

More precisely, let $X_1,...,X_N$ be the random characteristic variables of the *N* basic classes and $x_1,...,x_N$ their observations. $X_1,...,X_N$ are assumed to be independent and X_i (*i=1,...,N*) distributed as:

$$ P_{mix}\left(x; t_i, \vartheta^{(1)},..., \vartheta^{(K)} \right) = \sum_{k=1}^{K} p_k \; P\left(x; t_i, \vartheta^{(k)} \right) $$

where

t_i is a value which reflects the exposure of the basic class *i;*

$\vartheta^{(1)},..., \vartheta^{(K)}$ are parameters to be estimated;

p_k is the *k*-th weight in the mixture and can be seen as "a priori" probability that a basic class is located in cluster *k*; $\sum_{k=1}^{K} p_k = 1$;

$P\left(x; t_i, \vartheta^{(k)} \right)$ is the probability distribution of the characteristic variable conditionally to the belonging of the basic class *i* to the cluster *k* and dependent on the exposure t_i.

The posterior probability $P(k|x_i)$ for the *i*-th basic class to belong to the *k*-th cluster can be derived:

$$P(k|x_i) = \frac{p_k \; P(x_i; t_i, \vartheta^{(k)})}{\sum_{h=1}^{K} p_h \; P(x_i; t_i, \vartheta^{(h)})} \qquad k=1,...,K.$$

After having assigned the "a priori" probabilities p_k $(k = 1,\ldots, K)$ and having estimated the parameters $\vartheta^{(1)},\ldots,\vartheta^{(K)}$ by the maximum likelihood method, the posterior probabilities $\hat{P}(k|x_i)$ ($k=1,...,K$; $i=1,...,N$) can be estimated. As long as the probability distribution $\hat{P}(k|x_i)$ $k=1,...,K$ is "sufficiently" concentrated on the value $\bar{k}$ then the *i*-th basic class will be clearly allocated in cluster $\bar{k}$.

Loimaranta, Jacobsson and Lonka estimated these probabilities assuming for X_i ($i=1,...,N$) a mixture of Poisson distribution and they suggested the possibility of extending the method to multivariate characteristic variables with different distributions. Some particular cases (for instance normal multivariate distribution) have been treated in cluster analysis literature (e.g. see [8], p.34).

Besides hierarchical clustering methods and cluster analysis methods based on mixtures of probability distributions, another important class of cluster analysis techniques is known as partitioning methods (among which the well-known k-means algorithms). In these methods the number of clusters *K* is fixed in advance or, in some variant, determined through the procedure. Moreover, unlike the hierarchical techniques, they allow the relocation of the objects. In this way, bad initial partitions can be improved. Most of these techniques consist of two distinct procedures:

- the determination of an initial allocation of the objects into the clusters;
- the relocation of some or all of the objects to the clusters.

An essential feature of these methods is the calculation of the centroids of the clusters. Many clustering algorithms have been proposed; among them those proposed by E.W. Forgy, by J.B. MacQueen and a variant of the latter method (see [1]) are reported in [9].

In his discussion on the main characteristics of different clustering algorithms, B. Everitt [8] pointed out that:

"Hierarchical clustering techniques have a general disadvantage since they contain no provision for reallocation of entities who may have been poorly classified at an early stage in the analysis. In other words there is no possibility of correcting a poor initial partition."

For this reason hierarchical techniques are best suited for data in which a hierarchical structure can be assumed to exist, as for instance in biological data, so that no reallocation is needed. On the other hand partitioning

techniques seems to be particular valuable, even though some difficulties may arise: the possibility of determining suboptimal solutions and heavy computation in case of large date sets.
Moreover, Loimaranta *et al*. state that, in their opinion, the hierarchical clustering techniques only seldom are appropriate in actuarial applications and, as far as the determination of tariff classes is concerned, a method that searches for the optimal partition is preferable. However, in this case another difficult arises: how to take account of the exposures of the basic classes, since simple modifications of Forgy's or MacQueen's algorithms seem not to be easily available.
In [9], the Authors discussed some techniques, in NN framework, by which some partitioning algorithms can be implemented in a more flexible environment, allowing also the exposures to be taken into account.

3. How to choose the number of tariff classes: Schmitter and Straub's method.

A problem common to all clustering techniques is to decide the number of clusters in which the data are to be grouped and the determination of the tariff classes is not exception.
Dickmann (1978) suggests a criterion to choose the number of tariff classes, based on the "loss of information" caused by joining the basic classes in clusters. The loss of information is defined as a function of the number of clusters K (following the notation in Section 2):

$$g(K) = \frac{{}^{(K)}\sigma_W^2}{\sigma^2}$$

where

$$\sigma^2 = \sum_{k=1}^{K}\sum_{i=1}^{m_k}\left(x_{ik} - \bar{x}\right)^2 \frac{g_{ik}}{n} \qquad \text{and} \qquad \bar{x} = \sum_{k=1}^{K}\sum_{i=1}^{m_k} x_{ik}\frac{g_{ik}}{n}.$$

Since $g(K)$ is a decreasing function, we could choose K as the smallest number of clusters for which, for instance, $g(K) \le 0.05$.
Also Loimaranta-Jacobsson-Lonka (1980) dealt with the determination of the number of clusters and, under the assumptions of the Poisson mixture model, they derived asymptotic results to test the hypothesis on the probability distribution of the characteristic variable mixture of K distributions ([16]).
Some years before the papers published by H. Dickman and by K. Loimaranta, J. Jacobsson and H. Lonka, S. Schmitter and E. Straub (1975) introduced a method to find the "best" subdivision of an insurance portfolio in tariff classes. They assumed the existence of a "natural subdivision" and derived two statistics to single out this subdivision, or possibly the "closest" one from a set of "admissible subdivisions".

For "admissible subdivisions" they mean a subset of all the subdivisions of the portfolio, which can be actually considered for practical and commercial reasons.
According to [7], in Section 6 we apply the Schmitter and Straub's (S-S) model, originally designed to subdivide a portfolio in tariff classes, as a criterion to choose among different allocations of the basic classes in clusters. For this purpose, we will discuss the hypotheses and the model revisited in our perspective.
It is assumed that a "natural subdivision" of N basic classes in K clusters exists. These clusters are characterised by the risk parameters $\left(\Theta^{(1)},\ldots,\Theta^{(K)}\right)$, assumed to be a vector of random variables. Observations of the characteristic variables of the basic classes over I years are available.
With reference to the natural subdivision let:

t_{ik} be the exposure of the k-th cluster in the i-th year;

x_{ik} be the mean of the observations in the year i of the characteristic variable of the basic classes located in the cluster k, weighted with their exposures.

For the sake of simplicity in the exposition, but without losing in generality, we will consider as characteristic variable the claim frequency. In this case, t_{ik} is the number of policy-years and x_{ik} is the observed claim frequency in the k-th cluster and i-th year.
Let us define

X_{ik} the random characteristic variable claim frequency of the k-th cluster in year i

and assume the following hypotheses:

1. for $k = 1,\ldots,K$, conditionally to $\Theta^{(k)} = \vartheta^{(k)}$ the random variables $X_{1k},\ldots,X_{Ik}$ are independent and a couple of functions μ and σ exists such that:
$$E\left(X_{ik}|\Theta^{(k)} = \vartheta^{(k)}\right) = \mu\left(\vartheta^{(k)}\right) \qquad i=1,\ldots,I$$
$$Var\left(X_{ik}|\Theta^{(k)} = \vartheta^{(k)}\right) = \frac{\sigma^2\left(\vartheta^{(k)}\right)}{t_{ik}} \qquad i=1,\ldots,I$$

2. the random vectors $\left(\Theta^{(k)}, X_{1k},\ldots, X_{Ik}\right)$, $k=1,\ldots, K$, are independent; the random variables $\Theta^{(k)}$, $k = 1,\ldots, K$, are i.i.d. and we call:
$$Var\left[\mu\left(\Theta^{(k)}\right)\right] = w \quad \text{for } k=1,\ldots,K.$$

Let:

$t_{\cdot k} = \sum_{i=1}^{I} t_{ik}$ be the total exposure of the k-th cluster over the I years,

$t = \sum_{k=1}^{K} t_{\cdot k}$ be the total exposure of the K clusters over the I years,

and define:

$X_{\cdot k} = \dfrac{\sum_{i=1}^{I} X_{ik} t_{ik}}{t_{\cdot k}}$ the random variable claim frequency in the k-th cluster over the I years and

$X = \dfrac{\sum_{k=1}^{K} X_{\cdot k} t_{\cdot k}}{t}$ the random variable claim frequency over the I years.

Let

$$W = \frac{1}{K-1} \sum_{k=1}^{K} \frac{t_{\cdot k}}{t} \left(X_{\cdot k} - X \right)^2$$

$$V = \frac{1}{K} \sum_{k=1}^{K} \frac{t_{\cdot k}}{t} \frac{1}{I-1} \sum_{i=1}^{I} \frac{t_{ik}}{t_{\cdot k}} \left(X_{ik} - X_{\cdot k} \right)^2 \tag{3.1}$$

and

$$T = (K-1)(W-V) .$$

Schmitter and Straub show that T is (apart from a multiplicative factor) an unbiased estimator of the variance w. It can be interpreted as a measure of the heterogeneity among the classes of the natural subdivision.

Now we consider L admissible subdivisions and let $K_g \ (g = 1,\ldots, L)$ be the numbers of clusters of the g-th subdivision.

Of course, we can introduce for every admissible subdivision the same quantities we defined for the natural one. We will mark with $^{(g)}$ (e.g. $t_{ik}^{(g)}$, $X_{ik}^{(g)}$, etc.) the quantities corresponding to the g-th admissible subdivision. In particular we define:

$$T^{(g)} = \left(K_g - 1 \right) \left(W^{(g)} - V^{(g)} \right)$$

where:

$$W^{(g)} = \frac{1}{K_g - 1} \sum_{k=1}^{K_g} \frac{t_{\cdot k}^{(g)}}{t} \left(X_{\cdot k}^{(g)} - X \right)^2$$

$$V^{(g)} = \frac{1}{K_g} \sum_{k=1}^{K_g} \frac{t_{\cdot k}^{(g)}}{t} \frac{1}{I-1} \sum_{i=1}^{I} \frac{t_{ik}^{(g)}}{t_{\cdot k}^{(g)}} \left(X_{ik}^{(g)} - X_{\cdot k}^{(g)} \right)^2 . \tag{3.2}$$

The observed value of $W^{(g)}$ is

$$w^{(g)} = \frac{1}{K_g - 1} \sum_{k=1}^{K_g} \frac{t_{\cdot k}^{(g)}}{t} \left(x_{\cdot k}^{(g)} - x \right)^2$$

where $x_{\cdot k}^{(g)}$ and x are the observed values of the random variables $X_{\cdot k}^{(g)}$ and X respectively. $w^{(g)}$ measures the variability between the clusters of the *g*-th subdivision.

The observed value of $V^{(g)}$ is

$$v^{(g)} = \frac{1}{K_g} \sum_{k=1}^{K_g} \frac{t_{\cdot k}^{(g)}}{t} \frac{1}{I-1} \sum_{i=1}^{I} \frac{t_{ik}^{(g)}}{t_{\cdot k}^{(g)}} \left(x_{ik}^{(g)} - x_{\cdot k}^{(g)} \right)^2$$

where $x_{ik}^{(g)}$ is the observed value of the random characteristic variable $X_{ik}^{(g)}$. $v^{(g)}$ is a weighted mean of the empirical variances within the clusters.

Under the assumptions 1. and 2. above, Schmitter and Straub show that:

A. $E\left[T^{(g)}\right] \leq E[T]$ for all *g*=1,...,*L* and $E\left[T^{(g)}\right] = E[T]$ if the *g*-th subdivision is the natural one or a subdivision of it;

B. $E\left[W^{(g)}\right] < E[W]$ if the *g*-th subdivision is a proper subdivision of the natural one.

These results suggest a practical decision rule:

> choose the subdivision *g* that shows the highest value of $W^{(g)}$ among those with the highest $T^{(g)}$ values.

However, the subdivision with the highest value of $T^{(g)}$ will be discarded if another subdivision with a slightly lower value of $T^{(g)}$ and with a higher value of $W^{(g)}$ can be formed joining some clusters of the former subdivision.

From a methodological point of view, it is also important to note that the results A. and B. are formulated in terms of expectations, whereas in the decision rule the observed values of the statistics $T^{(g)}$ and $W^{(g)}$ are considered. Some troubles arise in the application of the decision rule when we have observations over one year only, since (3.1) and (3.2) are not defined when *I*=1. In this case, following [7] we can set $V = 0$ and $V^{(g)} = 0$, $g = 1, \ldots, L$. So we have:

$$T^{(g)} = \left(K_g - 1 \right) W^{(g)}$$

and

$$W^{(g)} = \frac{1}{K_g - 1} \sum_{k=1}^{K_g} \frac{t_{\cdot k}^{(g)}}{t} \left(X_{\cdot k}^{(g)} - X \right)^2 .$$

In [7] it is pointed out that, although the method is clear and valuable, since a good subdivision of the basic classes in clusters should reflect the heterogeneity of the portfolio, the decision rule cannot ensure to find the natural subdivision. In fact it could not belong to the family of admissible subdivisions and moreover, since in practical situations the boundary among the clusters may be rather vague, it could not be identified by the decision

rule.

4. Simple competitive learning.

The structure of a neural network is usually described as a connected graph, whose vertices, called units or neurons, are disposed into layers. We will consider only a two-layer network with *m* units on the first layer (input units), *K* units on the second layer (output units) and connections linking each input unit with each output unit. A real number, called weight, is associated at each connection. Each output unit will be represented by the weight vector $m_j = (\mu_{j1},\ldots,\mu_{jm})$, where μ_{ji} is the weight corresponding to the connection between the input unit *i* and the output unit *j*,. We denote by d a distance (not necessarily Euclidean) in $\Re^m$.
In simple competitive learning (SCL) a network is used to classify a set of data in clusters. We will suppose to have a set S (input space) of N real vectors of $\Re^m$ denoted by x_i $(i = 1,\ldots,N)$, which have to be classified in clusters. If we present a vector of data $x \in S$ to the network, it can be compared with all the weight vectors. We call winner unit, $c = c(x)$, the unit satisfying the condition

(4.1) $\quad d(x,m_c) \le d(x,m_j)\ \forall j = 1,\ldots,K$.

The SCL algorithm ([11], [12]) carries out a vector classifier according to the criterion (4.1). In order to minimise the number of misclassifications, the algorithm updates the weights of the network by means of a learning rule (step 4 in the following description of the algorithm).
More precisely, denoting by $m_j(t)$ and $x(t)$ the weight vectors at time *t* and the input vector presented at the same time respectively, the simple competitive learning algorithm consists of the following steps.

Simple competitive learning (SCL) algorithm

1. put $t = 0$ and initialise the vectors $m_j(0)$ $(j=1,\ldots,K)$;
2. choose an input vector $x(t) \in S$;
3. find the index *c* such that $d(x(t),m_c(t)) = \min_j \{d(x(t),m_j(t))\}$;
4. update the weight vector m_c according to the rule
   $m_c(t+1) = m_c(t) + \alpha(t)(x(t) - m_c(t))$, with $\alpha(t) \in]0,1[$;
5. stop if the stopping rule is satisfied;
   otherwise replace t with $t+1$, go back to step 2 and repeat for the next input vector.

At the end of the learning process the network is able to classify the input

vectors: input vectors that make the same output unit winner belong to the same cluster and the corresponding weight vector can be chosen as "representative" of the cluster itself.

Several methods can be used to initialise the vectors $m_j(0)$ in step 1. The simplest one is the so-called random guess method: the initial values are chosen randomly in the "right" domain, according to the values of the input vectors. Another method is to initialise the weights by the average of the minimum and the maximum values of the elements of the vectors which have to be classified. In the experiments presented in Section 6 both methods have been used. Nevertheless, other more sophisticated methods are available.

The term $\alpha(t)$, called learning rate, is a non-increasing function of the variable t. A good choice of the learning rate can speed up and improve significantly the convergence of the algorithm. For our experiments we chose individual learning rates for each weight vector in the form of

$$\alpha_j(t+1) = \begin{cases} \dfrac{\alpha_j(t)}{1+\alpha_j(t)} & \text{if j = c} \\ \alpha_j(t) & \text{if j} \neq \text{c} \end{cases} \qquad j = 1,\ldots,K \tag{4.2}$$

so that, in every training cycle, only the learning rate corresponding to the winner unit c is updated. A discussion about (4.2) and the choice of an "optimal" learning rate can be found in [14].

In the algorithm described above, the training is continuous, since the weights are updated after the presentation of each pattern. On the contrary, in the batch version of the algorithm, known as Linde-Buzo-Gray (LBG) algorithm of vector quantisation ([15]), the weights are updated after all patterns have been presented. As observed in [12] and [15], there is a strong relation between SCL, LBG and k-means algorithms.

Vector quantisation algorithms have been originally designed as encoding/decoding processes in data compression. An unified framework of most of those algorithms can be found in [3].

As pointed out in [17] and [18], if we denote by p_j $(j = 1,\ldots,N)$ a probability distribution over the set S, the LBG algorithm converges to a local minimum of the quantity (average distortion)

$$D = \sum_{i=1}^{N} (x_i - m_{c(x_i)})^2 \, p_i \; . \tag{4.3}$$

Nevertheless, continuous training is frequently used, because the random presentation order of the input vectors can help to avoid poor local minima (see [11] at page 168).

There are different types of stopping rules that we can use in step 5. A natural stopping rule ([3]) is

$$\Delta_n = \frac{D_{n-1} - D_n}{D_n} < \varepsilon$$

where D_n denotes the average distortion after the n-th training cycle and ε is a fixed threshold. In our experiments we followed the suggestion of several authors and we stopped the algorithm after a quite large number of iterations.

5. Self-Organising Maps.

In [13] T. Kohonen introduced an unsupervised technique to construct topology-preserving mappings from the input space into a low dimensional lattice (usually a one- or two-dimensional array of units). This algorithm is called self-organising map (SOM) and it is implemented by a network whose architecture is similar to that of the SCL networks. The most important difference between SCL and SOM is the following: while simple competitive learning modifies only the weight vector of the winner unit, self-organising map updates the weight vectors of the units placed in a suitable neighbourhood of the winner unit too.

Let

I be the set of output units

d' be a distance defined on $I \times I$

λ_t be a family of positive non-increasing real functions defined on $\Re^+$, where t is a real non-negative number and $\lambda_t(0) = 1 \ \forall t$

The SOM algorithm coincides with the SCL algorithm described in Section 4, except for step 4, which is replaced by:

4'. update the weight vectors according to the rule

$$m_j(t+1) = m_j(t) + \alpha(t)\lambda_t(d'(i,j))(x(t) - m_j(t)) \quad \forall j \in I \ .$$

Here too the learning rate α $(0 < \alpha(t) < 1)$ is a non-increasing function of t and $\alpha(0)$ is close to 1 (typically 0.8). Since λ_t is a non-increasing function, weights of units close to the winner unit and of the winner unit itself are changed significantly. On the contrary, weights of units placed further away from the winner unit are not updated appreciably. After the convergence of the algorithm, input vectors that are close in the input space are assigned to clusters corresponding to output units which are close in the lattice. A definition of this property of topology preservation can be found in [6].

The choice of the functions λ_t is crucial for the topology preservation. In our experiments we used the well-known gaussian function

$$\lambda_t(r) = \exp\left(-\frac{r^2}{2\sigma^2(t)}\right)$$

where σ is a decreasing function and $\sigma(0)$ is large enough. According to Ritter and Schulten (see [11] at page 114) we made the following choice for α and σ:

$$(5.1) \qquad \alpha(t) = \alpha_0 \left(\frac{\alpha_{t_{\max}}}{\alpha_0} \right)^{\frac{t}{t_{\max}}}$$

$$(5.2) \qquad \sigma(t) = \sigma_0 \left(\frac{\sigma_{t_{\max}}}{\sigma_0} \right)^{\frac{t}{t_{\max}}}$$

where $t_{\max}$ is the maximum value for t (fixed in advance) and α_0, $\alpha_{t_{\max}}$, σ_0, $\sigma_{t_{\max}}$ are the fixed initial and final values of α and σ respectively.

It must be noted that, despite the extensive use of SOM, the mathematical theory of Kohonen's algorithm is so far unsatisfactory. A fundamental book on the theory of SOM is [14], while a review on main results can be found in [4]. A wide investigation of the connections between neural networks and pattern recognition can be found in [2] and [19].

6. Experiments

In this section we present an application of the algorithms described in Section 4 and 5 to the data in Table 6.a, where we have reported the claim frequencies in a motor vehicle insurance portfolio. The basic classes are the fiscal powers of the vehicles and we want them to be allocated in clusters according to their claim frequencies. In [9] it is pointed out how an analogy between the expression of the average distortion (4.3) and of the total within cluster variance (2.1) in Dickmann's method suggests to consider the relative exposures as probability distribution on the basic classes. Assuming this hypothesis, we applied SCL and SOM to our data and tried to solve the problem of choosing the right number of clusters by means of S-S method. We also applied the same procedure to the same data yet grouped in a fine partition, reported in Table 6.b.

We used Matlab version 4.2c.1 and the Neural Network Toolbox version 2.0b [5] to perform the experiments described in this section. For this purpose we had to modify the programs provided in the toolbox. Among the main modifications we carried out, it has to be pointed out the implementations of the recursive formula (4.2) in the SCL program and of the gaussian function in the SOM program. In the latter program, α and σ are given by (5.1) and (5.2) and different initial choices of the parameters are allowed.

6. 1. Clustering by SCL

In Table 6.1.1 are reported the best results obtained by means of the SCL

algorithm in several trials carried out with different numbers of output units and various initial learning rates.
Note that the subdivisions reported in Table 6.1.1 (and in the following analogous tables) refer to the order in the data: e.g. (3 1 3 5 4 7) characterises the subdivision where the first cluster contains the first three elements in Table 6.a (fiscal powers : 22, 31 and 23), the second cluster contains the fourth element (fiscal power 39) and so on (see also Table 6.1.2 for some other examples).

Table 6.a: Policy-years (exposure) and relative and absolute claim frequencies in automobile insurance for different fiscal powers.

FISCAL POWER	No. OF CLAIMS	EXPOSURE	CLAIM FREQUENCY
22	38	221.17	0.171810
31	3	17.76	0.168957
23	270	1608.08	0.167903
39	1	6.72	0.148898
26	22	172.89	0.127247
19	497	4056.32	0.122525
21	11	93.31	0.117890
15	1636	14425.51	0.113410
18	1221	10917.02	0.111844
20	1004	9184.99	0.109309
17	1474	13543.50	0.108834
29	6	56.33	0.106523
14	1023	9985.27	0.102451
16	630	6319.19	0.099696
13	1856	18773.85	0.098861
28	2	20.60	0.097102
12	2122	23710.77	0.089495
32	2	22.37	0.089397
10	780	8798.13	0.088655
11	46	604.43	0.076105
9	34	462.59	0.073499
37	2	35.61	0.056172
8	6	108.32	0.055394
30	3	58.08	0.051655
27	1	20.04	0.049913
25	1	29.77	0.033590
24	0	8.39	0
33	0	1.24	0
35	0	0.00	0
36	0	8.25	0
38	0	3.11	0
40	0	0.60	0
41	0	8.13	0

(Data provided by an Insurance Company)

From subdivisions in Table 6.1.1 three hierarchies show up. All the subdivisions of the first and the second hierarchy present similar values of T, whereas the third shows worse results.
According to the S-S criterion, subdivision (4 8 4 17), which is present both in the first and in the second hierarchy, seems to be the best one (see the corresponding value of W). More details on this subdivision are reported in Table 6.1.2. We note that the weights of the neurons and the centroids of the clusters are approximately equal. We can deduce the convergence of the algorithm to a possibly local minimum of the average distortion; therefore the number of iteration used can be considered sufficient.

Table 6.1.1: Best subdivisions in clusters obtained by SCL from data in Table 6.a.

No. of clusters	Clusters	W x 10^{-5}	T x 10^{-4}	
6	3 1 3 5 4 17	3.2618	1.6309	R
5	3 1 8 4 17	3.9683	1.5873	R
4	4 8 4 17	5.2904	1.5871	R
6	4 3 5 3 1 17	3.2614	1.6307	R
5	4 8 3 1 17	3.9678	1.5871	R
4	4 8 4 17	5.2904	1.5871	R
5	11 5 3 2 12	2.9598	1.1839	
4	11 5 3 14	3.8964	1.1689	
3	11 5 17	5.6281	1.1256	

(R=weight initialisation by random guess method)

Table 6.1.2: Details on a subdivision obtained by SCL from data in Table 6.a.

4 clusters - W=5.2904 x 10^{-5} T=1.5871 x 10^{-4}				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	4	22 23 31 39	0.1682	0.1683
2	8	15 17 18 19 20 21 26 29	0.1119	0.1119
3	4	13 14 16 28	0.1001	0.1000
4	17	8 9 10 11 12 24 25 27 30 32 33 35 36 37 38 40 41	0.0886	0.0885

It is worthwhile to note that the best results have been obtained by means of the random guess initialisation of the weights. Obviously a number of trials large enough has been necessary.
However, we believe that these subdivisions could be unsatisfactory for actual rate making purposes, since the basic classes are not contiguously grouped. Moreover, we observe that basic classes characterised by low exposure are anyhow classified according to their claim frequencies. For instance, the fiscal power basic classes 31 and 39 are classified in the first

cluster (see Table 6.1.2) because of their high claim frequency, even if their exposures are very low. Whereas other basic classes "quite" near to 31 and 39 (e.g. 33, 35, 36, 38, 40, 41), with low exposure too, show low frequency instead and are therefore classified in the forth cluster. A way to try to avoid this kind of problems is to group the data in a fine partition before clustering (Table 6.b).

Table 6.b: Policy-years (exposure) and relative and absolute claim frequencies in automobile insurance for different fiscal powers (a fine partition of data in Table 8.a).

FISCAL POWER	No. OF CLAIMS	EXPOSURE	CLAIM FREQUENCY
22	38	221.17	0.171810
23	270	1608.08	0.167903
19	497	4056.32	0.122525
21	11	93.31	0.117890
15	1636	14425.51	0.113410
18	1221	10917.02	0.111844
20	1004	9184.99	0.109309
24-26	23	211.05	0.108976
17	1474	13543.50	0.108834
14	1023	9985.27	0.102451
16	630	6319.19	0.099696
13	1856	18773.85	0.098861
27-29	9	96.96	0.092824
12	2122	23710.77	0.089495
10	780	8798.13	0.088655
11	46	604.43	0.076105
9	34	462.59	0.073499
30-	11	161.86	0.067958
8	6	108.32	0.055394

(Data provided by an Insurance Company)

The results are reported in Table 6.1.3, where two hierarchical structure can be identified. Both present subdivision (2 7 3 7), that could be considered interesting according to the S-S method. Nevertheless, in the first hierarchy also subdivision (2 2 5 3 7) is remarkable. In the third part of Table 6.1.3 also other subdivisions not identified in a hierarchical structure are reported.
In Table 6.1.4 subdivisions (2 7 3 7) and (2 2 5 3 7) are displayed in detail. Since the latter is a subdivision of the former, (2 7 3 7) should be preferred, because its W value is higher, provided that the difference observed in the T values can be considered negligible. However, it does not seem to be so. In fact, if we look also at the centroids of (2 2 5 3 7), we note that the difference between the second and the third cluster is appreciable and therefore we conclude that this cannot be considered a subdivision of the natural one. According to S-S method, we prefer subdivision (2 2 5 3 7) to (2 7 3 7).

Table 6.1.3: Best subdivisions in clusters obtained by SCL from data in Table 6.b.

No. of clusters	Clusters	W x 10^{-5}	T x 10^{-4}	
6	2 2 2 3 3 7	3.2358	1.6179	
5	2 2 5 3 7	4.0112	1.6045	
4	2 7 3 7	5.2129	1.5639	
3	2 10 7	6.6232	1.3246	
6	2 1 3 3 3 7	3.2358	1.6179	
5	2 4 3 3 7	3.9770	1.5908	R
4	2 7 3 7	5.2129	1.5639	
3	9 3 7	5.5312	1.1062	
5	2 3 4 3 7	3.9829	1.5932	
5	2 7 3 3 4	3.9826	1.5931	R
5	2 7 1 2 7	3.9262	1.5705	R
4	2 4 5 8	4.8243	1.4473	
4	2 4 4 9	4.8172	1.4452	
4	2 3 5 9	4.7897	1.4369	

(R=weights initialisation by random guess method)

Table 6.1.4: Details on some subdivisions obtained by SCL from data in Table 6.b

4 clusters - W=5.2129 x 10^{-5} T=1.5639 x 10^{-4}				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	2	22 23	0.16774	0.16838
2	7	15 17 18 19 20 21 24-26	0.11189	0.11188
3	3	13 14 16	0.10005	0.10003
4	7	8 9 10 11 12 27-29 30-	0.08862	0.08862
5 clusters - W=4.0112 x 10^{-5} T=1.6044 x 10^{-4}				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	2	22 23	0.16746	0.16838
2	2	19 21	0.12224	0.12242
3	5	15 17 18 20 24-26	0.11099	0.11097
4	3	13 14 16	0.10008	0.10003
5	7	8 9 10 11 12 27-29 30-	0.08857	0.08862
5 clusters - W=3.9829 x 10^{-5} T=1.5932 x 10^{-4}				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	2	22 23	0.16746	0.16838
2	3	15 19 21	0.11549	0.11542
3	4	17 18 20 24-26	0.10993	0.10993
4	3	13 14 16	0.1001	0.10003
5	7	8 9 10 11 12 27-29 30-	0.08865	0.08862

We note that subdivision (2 3 4 3 7), the most interesting among those in the third group in the Table 6.1.3, is dominated by (2 2 5 3 7) according to the values of T and W. In fact, the difference between the centroids of the second and of the third cluster is lower than the same difference in (2 2 5 3 7) (see Table 6.1.4).
It is remarkable that the initial partitioning of the data does not influence significantly the result, since the obtained T values in this case are not so far from those obtained from original data.

6. 2. Clustering by SOM

In Table 6.2.1 are reported the best results obtained by means of Kohonen's SOM using the data in Table 6.a.
A Kohonen network with a one-dimensional array of output units and the Euclidean distance has been considered.
We observe that the choice of the parameters σ_0 and $\sigma_{t_{\max}}$ affects significantly the results obtained in the trials. More precisely, by putting σ_0 equal to half the number of neurons (a choice recommended by several authors, e.g. [11]), the results are substantially worse than by using a lower initial value of σ.
Two hierarchies are manifest. We note that the T values presented by the subdivisions of the first hierarchy (Table 6.2.1) are better than the T values obtained by SCL (Table 6.1.1). Subdivision (4 8 4 17) was found by both methods. Following S-S method we choose (4 8 4 17) instead of (3 1 8 4 17) in the SCL case.

Table 6.2.1: Best subdivisions in clusters obtained by SOM from data in Table 6.a.

No. of clusters	Clusters	W x 10^{-5}	T x 10^{-4}	σ_0	$\sigma_{t_{\max}}$
8	4 3 2 3 1 3 3 14	2.4201	1.6941	0.75	0.25
7	4 3 2 3 4 3 14	2.8124	1.6874	0.75	0.25
6	4 3 5 4 3 14	3.3480	1.6740	0.75	0.25
5	4 3 5 4 17	4.0767	1.6307	0.75	0.25
4	4 8 4 17	5.2904	1.5871	0.75	0.25
3	4 8 21	6.9993	1.3999	0.75	0.25
6	7 2 3 2 2 17	2.8578	1.4289	3	0.5
5	7 4 2 3 17	3.5440	1.4176	2.5	0.5
4	9 3 4 17	4.0005	1.2001	2	0.5
3	12 4 17	5.6278	1.1256	1.5	0.5

Here also (4 3 5 4 17) and (4 3 5 4 3 14) seem to be very interesting. In Table 6.2.2 details on these subdivisions are reported and compared with the

subdivision in seven clusters (4 3 2 3 4 3 14). The T values relative to the subdivisions in six and seven clusters are sensibly higher than that of the subdivision in five clusters and, by comparing the W values between the subdivision in seven and in six clusters, we note that the first is remarkably lower. In fact, in this subdivision the centroids of the third and the fourth cluster are quite the same. So the subdivision in six clusters seems to be more advisable.

Table 6.2.2: Details on some subdivisions obtained by SOM from data in Table 6.a

5 clusters - $W=4.0767 \times 10^{-5}$ $T=1.6307 \times 10^{-4}$				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	4	22 23 31 39	0.1674	0.1683
2	3	19 21 26	0.1224	0.1226
3	5	15 17 18 20 29	0.1112	0.1110
4	4	13 14 16 28	0.1003	0.1000
5	17	8 9 10 11 12 24 25 27 30 32 33 35 36 37 38 40 41	0.0881	0.0885
6 clusters - $W=3.348 \times 10^{-5}$ $T=1.6740 \times 10^{-4}$				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	4	22 23 31 39	0.1678	0.1683
2	3	19 21 26	0.1223	0.1226
3	5	15 17 18 20 29	0.1110	0.1110
4	4	13 14 16 28	0.1003	0.1000
5	3	10 12 32	0.0893	0.0893
6	14	8 9 11 24 25 27 30 33 35 36 37 38 40 41	0.0695	0.0690
7 clusters - $W=2.8124 \times 10^{-5}$ $T=1.6874 \times 10^{-4}$				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	4	22 23 31 39	0.1653	0.1683
2	3	19 21 26	0.1223	0.1226
3	2	15 18	0.1127	0.1127
4	3	17 20 29	0.1090	0.1090
5	4	13 14 16 28	0.1001	0.1000
6	3	10 12 32	0.0893	0.0893
7	14	8 9 11 24 25 27 30 33 35 36 37 38 40 41	0.0758	0.0690

The same algorithm has been applied also to the grouped data in Table 6.b. The main results are reported in Table 6.2.3. Also in this case the best results are obtained using initial values of σ not too high. The subdivisions in four and in five clusters are the same obtained by SCL (Table 6.1.3). However, because of the sensibly higher T value with respect to the subdivision in six

clusters (2 2 5 3 3 4), this seems to be preferable. Details of subdivisions (2 2 5 3 3 4) and (2 2 2 3 3 3 4) are reported in Table 6.2.4.

Table 6.2.3: Best subdivisions in clusters obtained by SOM from data in Table 6.b.

No. of clusters	Clusters	W x 10^{-5}	T x 10^{-4}	σ_0	$\sigma_{t_{max}}$
7	2 2 2 3 3 3 4	2.7452	1.6471	0.75	0.25
6	2 2 5 3 3 4	3.2674	1.6337	0.75	0.25
5	2 2 5 3 7	4.0112	1.6045	0.75	0.25
4	2 7 3 7	5.2129	1.5639	0.75	0.25
3	2 7 10	6.9078	1.3816	0.75	0.25
6	4 2 3 2 1 7	2.8099	1.4049	3	0.5
5	4 5 1 2 7	3.4841	1.3936	2.5	0.5
4	5 4 3 7	4.0481	1.2144	1.5	0.5

Table 6.2.4: Details on some subdivisions obtained by SOM from data in Table 6.b.

6 clusters - W=3.2674 x 10^{-5} T=1.6337 x 10^{-4}				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	2	22 23	0.1672	0.1684
2	2	19 21	0.1222	0.1224
3	5	15 17 18 20 24-26	0.1109	0.1110
4	3	13 14 16	0.1000	0.1000
5	3	10 12 27-29	0.0893	0.0893
6	4	8 9 11 30-	0.0746	0.0725
7 clusters - W=2.7452 x 10^{-5} T=1.6471 x 10^{-4}				
Cluster	No. of elements	Fiscal Powers	Weights	Centroids
1	2	22 23	0.1664	0.1684
2	2	19 21	0.1221	0.1224
3	2	15 18	0.1128	0.1127
4	3	17 20 24-26	0.1090	0.1090
5	3	13 14 16	0.1002	0.1000
6	3	10 12 27-29	0.0893	0.0893
7	4	8 9 11 30-	0.0757	0.0726

7. Ending remarks

In this paper we were concerned with the problem of determining the tariff classes by means of particular cluster analysis techniques. Two clustering algorithms, simple competitive learning and Kohonen's self organising-map,

have been applied in a neural network framework to the same set of data. In order to make a choice among the different subdivisions obtained we have applied a criterion proposed by H. Schmitter and E. Straub.
The SCL algorithm belongs to the partitioning methods family of clustering algorithms, which generally provide good results. In our case, Kohonen's SOM has produced even better results for a suitable choice of the parameters. This fact suggests a deeper investigation on the effect of the parameters. Moreover, also the meaning of the weights produced by Kohonen's SOM is worthwhile of further studies.

Bibliography

[1] Anderberg M. R., *Cluster Analysis for applications*, Academic Press, New York, N.Y., 1973.

[2] Bishop C.M., Neural networks for pattern recognition, Clarendon Press, 1995.

[3] Black J.V., "A Unified Framework for Vector Quantisers", *Defence Research Agency Malvern Memorandum 4670*, 1992.

[4] Cottrell M., J. C. Fort, G. Pagès, "Two or three things that we know about the Kohonen algorithm", *Proceedings of ESANN'94*, Bruxelles, 1994.

[5] Demuth H., M. Beale, *Neural Network Toolbox For Use with MATLAB - User's Guide*, The Mathworks Inc., 1994.

[6] Der R., M. Herrmann, T.M. Martinetz, T. Villmann, "Topology preservation in self-organizing feature maps: exact definition and measurement", *IEEE Transactions on Neural Networks*, 8 No.2, 256-266, 1997.

[7] van Eeghen J., E. K. Greup, J. A. Nijssen, "Rate making", *Surveys of Actuarial Studies*, 2, Nationale-Nederlanden N.V., 1983.

[8] Everitt B., *Cluster analysis*, Heinemann Educational Books, London, 1974.

[9] Giulini S., R. Pelessoni, L. Picech, "Determination of tariff classes: cluster analysis methods and unsupervised neural networks", *Proceedings of the XXVIII International Astin Colloquium* (to appear), 1997.

[10] Hartigan J.A., *Clustering Algorithms*, John Wiley & Sons, New York, N.Y., 1975.

[11] Hassoun M.H., Fundamentals of artificial neural networks, The MIT Press, 1995.

[12] Hertz J., A. Krogh, R. G. Palmer, *Introduction to the theory of neural computation*, Addison-Wesley Publishing Company, 1991.

[13] Kohonen T., *Self-Organization and Associative Memory*, Springer, 1984.

[14] Kohonen T., *Self-Organizing Maps*, Springer, 1995.
[15] Linde Y., A. Buzo, R. Gray, "An Algorithm for Vector Quantizer Design", *IEEE Transactions on Communications*, COM-28, 84-99, 1980.
[16] Loimaranta K., J. Jacobsson, H. Lonka, "On the use of mixture models in clustering multivariate frequency data", *Transactions of the 21st International Congress of Actuaries*, 2, 147-161, 1980.
[17] Luttrell S.P., "Hierarchical Self-Organising Networks", *Proceedings of the 1st International Conference on Artificial Neural Networks*, London, 1989.
[18] Luttrell S.P., "Vector Quantisation of K-Distributed Data", *RSRE Research Note SP4/110*, 1990.
[19] Ripley B.D., *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.
[20] Schmitter H., E. Straub, "How to find the right subdivision into tariff classes", *Astin Bulletin*, 8 No.2, 257-263, 1975.
[21] Ward J.H., "Hierarchical grouping to optimize an objective function", *J. Am. Statist. Ass.*, 58, 236-244, 1963.