Submission for the CoiL Challenge 2000

Juha Vesanto and Janne Sinkkonen
Neural Networks Research Centre
P.O.Box 5400, 02015 HUT, Finland
Juha.Vesanto@hut.fi

24th May 2000

Task 1: prediction

First, the data was preprocessed. All variables, except customer maintype and subtype, were scaled to range [0,1] and then multiplied with a weighting factor which was based on estimated importance of the variable. Categories of the subtype were manually scored on five dimensions, and the scores were factored with PCA. Three components associated with the largest eigenvalues were used as input values for the model, instead of the original customer subtype values. Customer maintype was omitted.

The prediction model was an RBF network, the first layer consisting of non-normalized, gaussian, spherically symmetric kernels of common width. The second layer was a regularized log-linear model. 200 kernels were used, and the centers of the kernels were computed by the k-means algorithm. The kernel width and the regularization parameter of the log-linear model were selected by maximizing log-likelihood of an independent test set, which was of equal size with the learning set.

Of the submitted set of 800 indices from the test set, 110 were caravan policy owners.

Task 2: description

A typical caravan policy owner can be characterized as having

1. private third party insurances, and especially car (and fire) policies,
2. and high or medium level income (and associated demographic features).

But there are exceptions to all these properties. By dividing the data set to a set of clusters, and describing these clusters in a hierarchical manner one can see that there are actually several different types of customers that have a caravan policy. Clusters are also well-suited for explaining the prediction results, because the model was basically based on kernels, i.e. clusters.

The clustering was done using a Self-Organizing Map of the same data set as was used for training the prediction model. Figure 1 shows how the map has been divided to clusters, and the dendrogram in Figure 2 shows the hierarchy

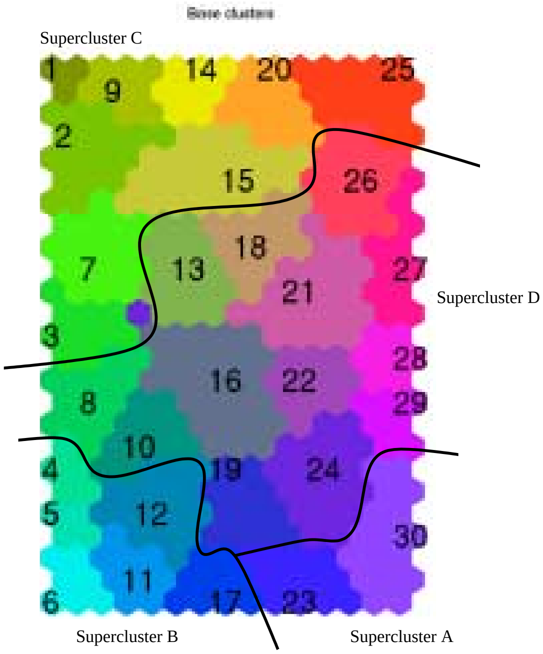

Figure 1: Superclusters depicted on a Self-Organizing Map.

of these clusters. First, the map (and thus the original data) is divided to four superclusters, and each of these is further divided to several lower-level clusters.

The response of the map to the caravan policy owners can be seen in Figure 3. The clusters that get the highest response from the caravan-policy owners are the most interesting ones. Thus, the most interesting clusters are:

- supercluster A consisting of subclusters 23 and 30
- supercluster B, and there subclusters 4, 5, 6, and 17
- subclusters 7 and 25 within supercluster C
- subcluster 24 within supercluster D

In the prediction task the 800 most probable caravan policy owners were primarily picked from superclusters A and B. However, the other clusters are also interesting as they indicate further potential target groups. Below, the four superclusters and the interesting low-level clusters are characterized in terms of what are their properties and how they differ from each other.

Supercluster A: insurance policy owners

The most characteristic property of this supercluster is that the contribution and number of insurance policies is high, especially contributions of private third party insurance policies, car policies and fire policies. Other characteristic properties are a car, and a house of their own. The subclusters 23 and 30 differ in that the latter exhibits lower income levels.

People in supercluster A clearly have a desire to insure their property.

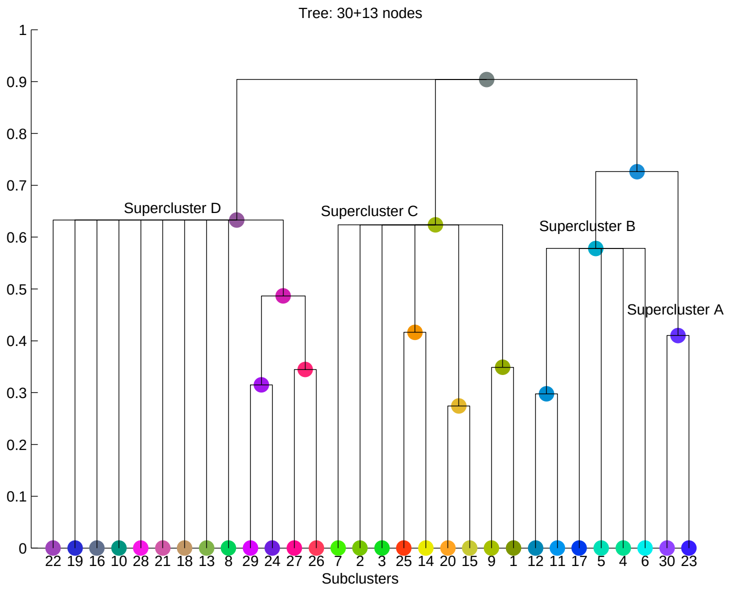

Figure 2: Subclusters belonging to the superclusters.

Supercluster B: high purchasing power

Typical of this supercluster are high, or at least medium, education, status, social class and income levels. Besides these, this supercluster differs from A in that the contribution of policies is typically not as high. Also people in supercluster B have a car, and are typically house owners.

The subcluster 6 exhibits highest levels of social class, income, education, etc. Subcluster 4, 5 and 17 exhibit middle levels of income, education, etc. They differ in that subcluster 17 are house owners, while people in subcluster 4 and 5 may live on rent.

People in supercluster B are not necessarily enthusiastic about insuring their property, but they do have quite enough wealth to own a caravan, even if using it were not their prime hobby.

Supercluster C: rent

Supercluster C has a number of different subclusters. The unifying thing for most of them, however, is that the people live in rented houses. The economical and status levels are low or medium, and many don't own a car.

The subclusters having caravan policy owners are clusters 7 and 25, for former having higher education levels etc. than the latter. They differ from the rest of the subclusters of C in that they have car insurance policies and often private third party insurance policies.

Supercluster D: others

Supercluster D is a large and diverse group which contains the rest of the subclusters. The common factor is that they don't have very many or very valuable

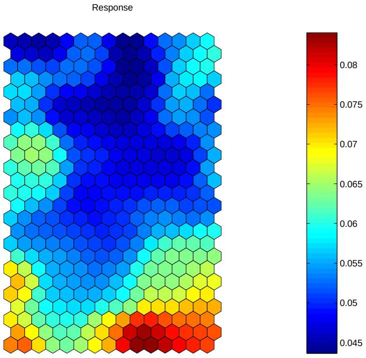

Figure 3: Response of the Self-Organizing Map to the caravan policy owners: the greater the response, the higher the probability that a caravan policy owner can be found.

insurance policies. The only exception is subcluster 24 which exhibits high levels of car policies.