

A Neural Network based framework for Customer Profiling for Risk analysis

Jay B.Simha,
Abiba Systems
Bengalooru
Jay.b.simha@abibasystems.com

Raghavendra B.K.,
Department of Computer Science
Gousia College of Engineering
Ramanagaram

Abstract— Customer profiling is an important function in Risk Management. It will help the decision makers to select the most profitable group of customers for any credit activity as well as to explore the behavior. Statistical methods like k-means were frequently used for customer profiling. Such methods are not optimal in selecting the number of segments. In this work a self organizing map based clustering is used to segment the customer base and understand their behavior. The proposed approach uses SOM algorithm with multiple business relevant metrics as the cluster validation criteria to arrive at optimal number of clusters. The case study confirms the efficacy of the proposed approach.

Key words— Data Mining, SOM, Customer Profiling, Risk Analysis

I. INTRODUCTION

Many businesses today make intensive use customer information held in databases. Most of the CRM products provide operational functions. Operational purposes usually include interrogating the database to select groups of customers. It requires expertise and in-depth knowledge of the structure of the database along with anticipated customer response. Further it can pose a problem to use the vast variety of information available in a database all at the same time, as very often it is impossible to recognise the correlation between data fields and anticipated customer response. Statistical methods like k-means [4] are used in customer segmentation. However, such methods are too restrictive on the since the number of cluster are to a priory, which is not possible when data is explored. Self organizing maps (SOM) which are based on unsupervised neural network learning have been successfully used in segmentation to overcome the problems of statistical algorithms. But few of these algorithms have business metrics built into them for cluster validation. In this paper an attempt has been made to develop a framework for customer segmentation using business metrics for cluster validation.

The rest of the paper is organized as follows. Literature is reviewed in section 2. The self organizing maps are discussed briefly in section 3. In section 4, proposed frame work is discussed. In section 5, business metrics used for cluster validation are discussed. Experimental validation of the proposed framework using benchmark datasets is given in section 6.

II. LITERATURE REVIEW

Self organizing neural networks have been an integral part of the data analytics in business intelligence. The spectrum of applications ranges from finance [1], qualification analysis in business [6], as well as market segmentation in e-commerce [5] and market basket analysis [2].

As a consequence thereof this class of neural networks is the subject of continuing efforts for improvement. Corresponding research interests are devoted to both their individual design for specific areas of application and the elimination of existing methodological problems. However, the efficient determination of adequate parameter settings continues to be a crucial practical problem, which has to be solved for each data set by a more or less troublesome trial-and-error process, if the internal structure of the data is unknown. This can be minimized by incorporating cluster validation criteria.

Literature review indicates that almost no work has been done in developing the segmentation framework incorporating business metrics. In this paper an attempt has been made to develop a framework for customer segmentation using business metrics for cluster validation.

III. SELF ORGANIZING MAPS

The self-organizing map (SOM) is a subtype of artificial neural networks[3]. It is trained using unsupervised learning to produce low dimensional representation of the training samples while preserving the topological properties of the input space. The self-organizing map is a single layer feed-forward network where the output syntax are arranged in low dimensional (usually 2D or 3D) grid. Each input is connected to all output neurons. Attached to every neuron there is a weight vector with the same dimensionality as the input vectors.

The goal of the learning in the self-organizing map is to associate different parts of the SOM lattice to respond similarly to certain input patterns. The weights of the neurons are initialized either to small random values or sampled evenly from the subspace spanned by the two largest principal component eigenvectors. The training utilizes competitive learning. When a training sample is given to the network, its Euclidean distance to all weight vectors is computed. The neuron with weight vector most similar to the input is called the Best Matching Unit (BMU). The weights of the BMU and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time

and is smaller for neurons physically far away from the BMU. The update formula for a neuron with weight vector $\mathbf{W}_v(t)$ is

$$\mathbf{W}_v(t+1) = \mathbf{W}_v(t) + \Theta(v, t)\alpha(t)(\mathbf{D}(t) - \mathbf{W}_v(t))$$

where $\alpha(t)$ is a monotonically decreasing learning coefficient and $\mathbf{D}(t)$ is the input vector. The neighborhood function $\Theta(v, t)$ depends on the lattice distance between the BMU and neuron v . In the simplest form it is one for all neurons close enough to BMU and zero for others, but a Gaussian function is a common choice, too. Regardless of the functional form, the neighborhood function shrinks with time. At the beginning when the neighborhood is broad, the self-organizing takes place on the global scale. When the neighborhood has shrunk to just a couple of neurons the weights are converging to local estimates.

This process is repeated for each input vector, over and over, for a (usually large) number of cycles. The network winds up associating output nodes with groups or patterns in the input data set. If these patterns can be named, the names can be attached to the associated nodes in the trained net, which is very helpful in profiling.

IV. PROPOSED FRAMEWORK

The proposed framework consists of five phases as shown in fig 1. In data preparation stage the required data from different sources and preprocessed for transformation into a standard form. In the next stage, the features which are important for analysis are selected.

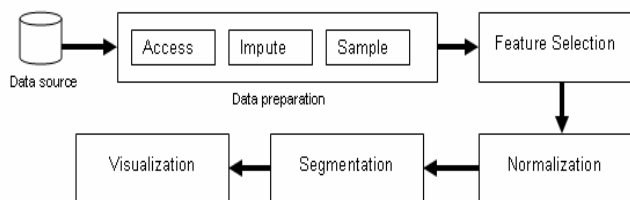


Fig 1. Segmentation framework

In the third stage the selected data are normalized, since SOM can work only with normalized data for better results. In the next stage the normalized data are clustered using SOM. In the visualization stage the derived segments are explored with multiple criteria as discussed in the subsequent section. .

V. BUSINESS METRICS FOR CLUSTER VALIDATION

Most of the cluster validation depends on homogeneity of the cluster elements computed using some form of distance. However, one problem with such methods is that they may not be readily acceptable by business users, as they have different set of validation criteria. In this paper we discuss two widely used business metrics namely: response rate and profit maximization. Response rate is useful when the any cost of misclassification is acceptable. However profit maximization tries to identify the segments which are profitable by a pre-

defined function. In this research both the approaches are used to analyze the data.

VI. EXPERIMENTAL VALIDATION

Credit scoring, is an attempt to classify applicants for credit into either “good” or “bad” risk classes. Majority of the credit scoring models use supervised classification methods. Although these methods have made considerable progress in risk prediction, they do not utilize all the information available in the data set. Therefore, it is worthwhile to investigate the applicability of unsupervised learning methods in risk classification. In this research we have used the proposed framework based on SOM for risk classification. Two publicly available datasets – Australian credit and German credit are used in the analysis. The details of the datasets used are shown in table 1.

Table 1. Dataset descriptions

	<i>Australian</i>	<i>German</i>
Number of instances	690	1000
Number of variables	14	20
Positive percentage	28	30
Number of missing values	67	0
Number of nominal variables	8	13
Number of ordinal variables	0	0
Number of interval variables	6	7

The proposed framework was applied to these two datasets. The cluster validation based on distance, response rate and profit (based on the minimum risk) were tested on the benchmark datasets. The results of the analysis are given in table 2. It can be observed that business metric based validation of clusters has different values than classical cluster validity based on purity of clusters.

Table 2. Results of exploratory analysis

	<i>Purity</i>	<i>Response rate</i>	<i>Profit</i>
Australian			
Number of clusters	5	6	6
Purity	95%	90%	87%
Response rate	12%	35%	27%
Profit	32%	56%	73%
German			
Number of clusters	5	6	6
Purity	95%	86%	88%
Response rate	24%	46%	39%
Profit	29%	60%	78%

VII. CONCLUSIONS AND SCOPE FOR FURTHER WORK

Segment based risk analysis is another way of aggregate and exploratory analysis applied in data mining before using supervised learning algorithms for classifications. In this research work an attempt has been made to develop a framework based on self organizing maps (SOM) and business metrics for cluster validation. Preliminary results on benchmark datasets have shown promising results. Work is under progress to extend the framework to include the supervised learning component on the segmented dataset.

REFERENCES

- [1] DeBoeck, G. and Kohonen, T. (1998). *Visual Explorations in Finance*. Springer.
- [2] Decker, R. and Monien, K. (2003), "Market basket analysis with neural gas networks and self organizing maps", *Journal of Targeting, Measurement and Analysis for Marketing* **11**, 373-386.
- [3] http://en.wikipedia.org/wiki/Self-organizing_map
- [4] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*", Berkeley, University of California Press, 1:281-297
- [5] Vellido, A., Lisboa, P. J. G., and Meehan, K. (1999). "Segmentation of the on-line shopping market using neural networks", *Expert Systems with Applications* **17**, 303-314.
- [6] Wagner, R. (2004). "Mining promising qualification patterns", In *Innovations in Classification, Data Science, and Information Systems* (Edited by D. Baier and K.-D. Wernecke). Springer, 249-256.