Analyzing domestic violence with topographic maps: a comparative study

Jonas Poelmans[1], Paul Elzinga[2], Stijn Viaene[1,3], Guido Dedene[1,4], Marc M. Van Hulle[5]

[1]K.U.Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium
[2]Police Amsterdam-Amstelland, James Wattstraat 84,
1000 CG Amsterdam, The Netherlands
[3]Vlerick Leuven Gent Management School, Vlamingenstraat 83,
3000 Leuven, Belgium
[4]Universiteit van Amsterdam Business School, Roetersstraat 11
1018 WB Amsterdam, The Netherlands
[5]K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie
Campus Gasthuisberg 3000 Leuven, Belgium
{Jonas.Poelmans, Stijn.Viaene, Guido.Dedene}@econ.kuleuven.be
Paul.Elzinga@amsterdam.politie.nl
marc@neuro.kuleuven.be

Abstract. Topographic maps are an appealing exploratory instrument for discovering new knowledge from databases. During the recent years, several variations on the Self Organizing Maps (SOM) were introduced in the literature. In this paper, the toroidal Emergent SOM tool and the spherical SOM are used to analyze a text corpus consisting of police reports of all violent incidents that occurred during the first quarter of 2006 in the police region Amsterdam-Amstelland (The Netherlands). It is demonstrated that spherical topographic maps provide a powerful instrument for analyzing this dataset. In addition, the performance of the toroidal Emergent SOM is compared to that of the spherical SOM, and it turned out to be superior to that of an ordinary classifier, applied directly to the data.

Keywords: Topographic maps, domestic violence, knowledge discovery in databases, Emergent SOM, BLOSSOM

1 Introduction

According to the department of Justice of the Netherlands, domestic violence can be characterized as serious acts of violence committed by someone of the domestic sphere of the victim. Violence includes all forms of physical assault. The domestic sphere includes all partners, ex-partners, family members, relatives and family friends of the victim. Family friends are those persons who have a friendly relationship with the victim and who regularly meet the victim in his/her home [1].

Research has proven that domestic violence is a largely underestimated problem in our modern society [2,3,4,5]. Pursuing an effective policy against offenders is one of the top priorities of the police organization of the region Amsterdam-Amstelland in the Netherlands. Of course, in order to pursue an effective policy against offenders, being able to swiftly recognize cases of domestic violence and label reports accordingly is of the utmost importance. Still this has proven to be problematic. In the past, intensive audits of the police databases related to filed reports have established that many reports tended to be wrongly classified as domestic or as non-domestic violence cases. One of the conclusions was that there was a need for an in-depth investigation of this problem area.

In the current paper, we develop an application in the problem area of topographic maps [7], which are particularly suited for high-dimensional data visualization. Two recent tools will be considered, the Emergent SOM and the spherical SOM, and their performances compared. The remainder of this paper is composed as follows. In section 2, we discuss the essentials of topographic map theory and in particular the Emergent SOM and Spherical SOM. In section 3, we elaborate on the dataset. In section 4, the results of the comparative analysis of the toroidal ESOM (using the Databionics tool) and the Spherical SOM (using the BLOSSOM tool) are presented. Section 5 concludes the paper.

2 Topographic Map essentials

From a practitioner's point of view, topographic maps are an especially appealing technique for knowledge discovery in databases [15]. It performs a non-linear mapping of a high-dimensional space to a low-dimensional one, usually a two-dimensional one. It offers the user a useful tool for exploring the dataset [12]. It can be used to detect clusters and it maintains the neighborhood relationships that are present in the input space. It also provides the user with an idea of the complexity of the dataset, the distribution of the dataset (e.g. spherical) and the amount of overlap between the different classes. The lower-dimensional data representation is also an advantage when constructing classifiers.

2.1 Emergent SOM

An Emergent Self Organizing Map (ESOM) is a very recent type of topographic map [8]. It is argued to be especially useful for visualizing sparse, high-dimensional datasets, yielding an intuitive overview of its structure [10]. An Emergent SOM differs from a traditional SOM in that a very large number of neurons (at least a few thousand) are used [9]. Alfred Ultsch argues that the topology preservation of the traditional SOM projection is of little use when using small maps: the performance of a small SOM is almost identical to that of k-means clustering, with k equal to the number of nodes in the map [8]. An additional advantage of an ESOM is that it can be trained directly on the available dataset without first having to go through a feature selection procedure [11]. ESOM maps can be created and used for data analysis by

means of the publicly available *Databionics ESOM Tool*. This tool allows the user to construct both flat and unbounded (i.e., toroidal) ESOM maps.

2.2 Spherical SOM

In a spherical SOM, the neurons are arranged on a sphere. Recently, several spherical self-organizing topographic maps have been introduced in the literature [6]. These maps are spherical or toroidal and, thus, not bounded as in the case of e.g. the traditional SOM and its many versions, and thus should not suffer from the border effect. The border effect is a phenomenon which occurs in flat maps because the number of neighborhood neurons of a neuron at the border of the map is smaller than the number of neighborhood neurons of a neuron at the center of the map [14]. This might cause distortions of the map, e.g. leading to a too small area for cluster detection near the edges of the map. The spherical SOM tool used here is BLOSSOM [13].

3 Dataset

The dataset consists of 4146 police reports describing all violent incidents from the first quarter of 2006. All domestic violence cases from that period are a subset of this dataset. Unfortunately, many of these 4146 police reports did not contain the reporting of a crime by a victim, which is necessary for establishing domestic violence. This happens for example when a police officer was sent to an incident and later on wrote a report in which he/she mentioned his/her findings, while the victim did not make an official statement to the police. Therefore, we only retained the 2288 documents in which the victim reported a crime to a police officer. From these 2288 documents, we removed the follow-up reports referring to previous cases. This filtering process resulted in a set of 1794 reports. From these reports, the person who reported the crime, the suspect, the persons involved in the crime, the witnesses, the project code and the statement made by the victim to the police were extracted. Of these 1794 reports, 462 were cases of domestic violence; the others not. These data were used to generate the 1794 html-documents that were used during the research.

We also have at our disposal a thesaurus – a collection of terms – that was obtained by performing word frequency analyses on these police reports. The relevant terms that occurred most often were retrieved and added to the initially empty thesaurus. This resulted in a set of 123 terms. In the categorical dataset, it is indicated for each police report which ones of these terms appear in the report. In the continuous dataset, the relevance of each term is indicated for each police report by means of a continuous value between 0 and 1. This value was calculated on the basis of the number of times the term appeared in the report.

For each police report, some additional information is available. This information includes whether or not the suspect of the criminal offence is known, the gender of the victim, the age of the victim, whether the perpetrator and victim lived at the same address, etc.

4 Experiment

In a first step, a toroidal ESOM map was trained on the basis of these 2 datasets, in order to discover the distribution of the dataset. In the map displayed in Fig. 1, the best matching (nearest-neighbor) nodes are labeled in the two classes for the given test data set (red for domestic violence, green for non-domestic violence). By analyzing the ESOM map based on the categorical dataset, we found that there is one large domestic violence cluster running vertically through the center of the map, and one less clearly demarcated domestic violence cluster running to the left. The latter continues over the edge of the map and has an outlier on the right of the map. Therefore, it seems natural to use a spherical or toroidal SOM for visualizing this dataset.

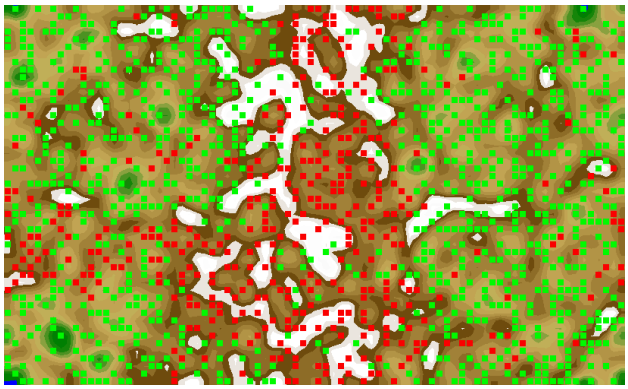

Fig. 1. Toroid ESOM map trained on the categorical dataset with all features

For both tools, it was possible to train a map directly on the entire dataset with more than 123 features. However, in order to prevent distortions on the map caused by irrelevant and redundant features, it was chosen to apply feature selection. A heuristic feature selection procedure called minimal-redundancy-maximal-relevance (mRMR), as described in [16], was considered. The aim was to select the 50 most relevant features. To obtain the optimal feature set, an SVM, a Neural Network, a kNN (with k=3) and a Naïve Bayes classifier were used to measure the classification performance for an increasing number of features. The classification performance was plotted as a function of the number of features and it was decided to retain the best 18 features.

For the ESOM, a SOM with a lattice containing 50 rows and 82 columns of neurons was used (50x82=4100 neurons in total). The weights were initialized randomly by sampling a Gaussian with the same mean and standard deviation as the corresponding features. A Gaussian bell-shaped kernel with initial radius of 24 was used as a neighborhood function. Further, an initial learning rate of 0.5 and a linear cooling strategy for the learning rate were used. The number of training epochs was set to 20. Both a map with a toroidal topology of the neurons as well as a flat topology were used. For BLOSSOM, a network consisting of 642 neurons was used. The weights were initialized randomly. A Gaussian kernel with initial radius π was used

as a neighborhood function. Further, an initial learning rate of 0.9 and a linear cooling strategy for the learning rate were used. The number of training epochs was set to 50.

The BLOSSOM map trained on the categorical dataset is displayed in Fig. 2. The BLOSSOM map trained on the continuous dataset is displayed in Fig. 3. The toroidal ESOM map trained on the categorical dataset is displayed in Fig. 4. The flat ESOM map trained on the categorical dataset is displayed in Fig. 5.

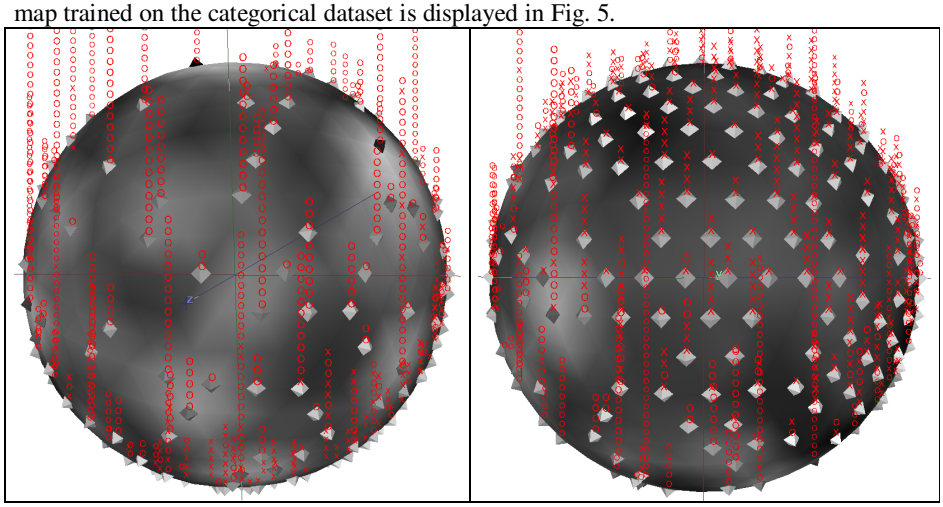

Fig. 2. Two views of the BLOSSOM map trained on the categorical dataset with 18 features. The grayscales on the surface indicate local densities (white= high density). The small tetrahedrons indicate the nearest-neighbor neurons for the two types of labels; "x" indicates a domestic violence case, "o" a non-domestic violence case.

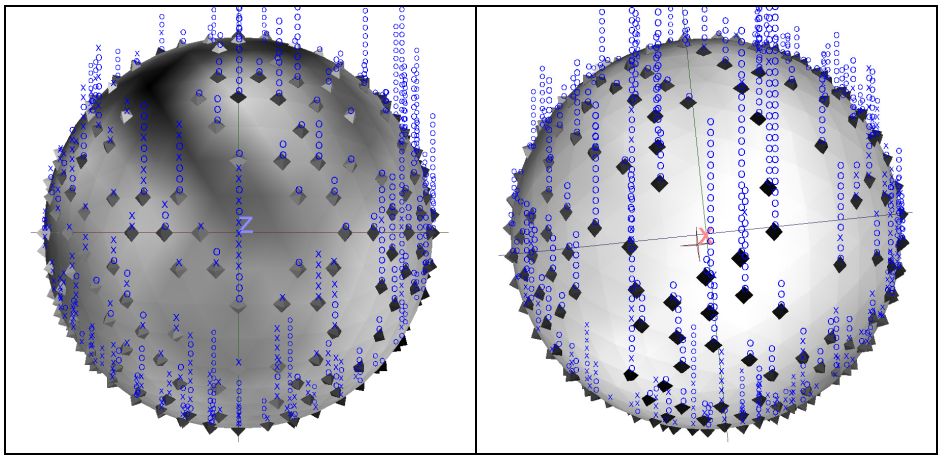

Fig. 3. Two views of the BLOSSOM map trained on the continuous dataset with 18 features

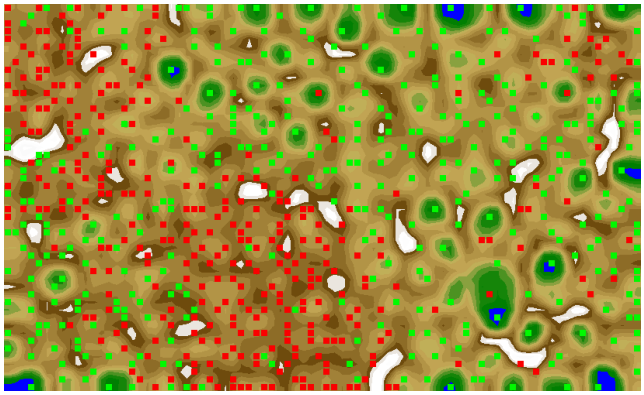

Fig. 4. Flat ESOM map trained on the categorical dataset with 18 features

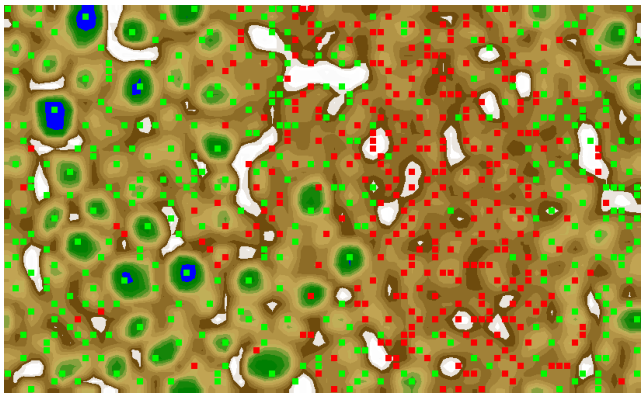

Fig. 5. Toroid ESOM map trained on the categorical dataset with 18 features

Finally, a kNN classifier was built for the ESOM and BLOSSOM maps. For BLOSSOM, k was set to 1. In order to obtain the misclassification error of the BLOSSOM map, the Euclidean distance of each input vector to each weight vector was measured. For each weight vector (corresponding to a node of the map) it was calculated how many of the domestic and non-domestic violence cases had this weight vector as a best match. If the node dominantly contained domestic violence cases, it was labeled as a domestic violence node and the non-domestic violence cases that best matched to this node were considered to be wrong classifications. Because the ESOM map contained about 2 times as many best-matched neurons as the BLOSSOM map (680 vs. 316), k was set to 2 for the ESOM map. A best-matched neuron is a neuron for which there exists at least one input vector for which the Euclidean distance to the weight vector of this node is minimal.

5 Analysis and Results

The ESOM tool was first applied to the dataset containing all features for quickly obtaining an overview of the structure of the dataset. We observed that some of the clusters continued over the edges of the map, thereby making BLOSSOM an interesting candidate tool. A problem with the ESOM map is that the density profile of the map does not match the uniform distribution of the labeled data vectors. Moreover, there is no ridge in the map that separates the domestic- from the non-domestic violence cases. Therefore, the 'watershed' technique will not lead to a correct identification of the classes. This problem was not solved by lowering the number of features. Nevertheless, much more density variations can be observed in Fig. 5. BLOSSOM was problematic in that not all labels are visible on the map. Many of the labels at the upper side of the map were impossible to see because of the small window size used by the tool.

By examining the BLOSSOM map shown in Fig. 2, one can conclude that there are no clearly demarcated clusters of domestic violence cases available in the categorical dataset. This was also the case for the spherical map trained on the dataset containing all 123 features. However, several clearly demarcated clusters of non-domestic violence cases can be observed in the map of Fig. 2. These clusters correspond to the white (= high density) areas of the map.

We found that these clusters correspond to types of incidents that can be clearly distinguished from other types of incidents. In burglary cases e.g., the suspect is typically not known, neither a description of the suspect is provided, and one or more locations inside the house are mentioned. These typical characteristics result in a grouping of such cases by the BLOSSOM tool. From the map displayed in Fig. 3, one can conclude that this is also the case for the continuous dataset. However, it is conspicuous that the latter contains much less density variations.

Another interesting result is that the map provides a good division between domestic and non-domestic violence cases. Many of the best matched nodes dominantly contain either domestic or non-domestic violence cases. This indicates that there is only a small amount of overlap between them. The observed overlap probably indicates that the feature set is not sufficiently refined to discriminate between the two classes. However, it should be considered that some cases might have been wrongly classified by police officers. The latter might be due to the vagueness of the domestic violence definition.

When the flat ESOM map is compared to the toroidal ESOM map, one concludes that the toroidal map provides a better visualization of the dataset. The border effect is clearly present in the flat map resulting in undesired distortions of the map. Most of the observed clusters are located at the border of the map, which makes them smaller in area, and the large group of domestic violence cases is less clearly demarcated from the non-domestic violence cases.

Finally, the results of the nearest neighbor classifiers based on the ESOM and BLOSSOM maps are displayed in table 1 and 2. These values are averages of the accuracy obtained during 40 runs of each method.

Table 1. Classification performance on the categorical dataset.

	Overall accuracy	False Positive Rate	False Negative Rate
BLOSSOM 1NN	90.4%	2.9%	29%
Toroid ESOM 2NN	90.8%	8%	12.6%
Flat ESOM 2NN	91.4%	6.8%	13.9%
Toroid ESOM 1NN	95%	4%	7.6%
Flat ESOM 1NN	95.1%	4%	7.4%

An interesting result is the striking difference in performance of the traditional kNN classifier (65%) and the kNN classifier based on the spherical BLOSSOM map (90,4%). This is due to the topographic map being a model of the data distribution: it forms an approximation of the data manifold, offering interpolating facilities, and it spends more neural hardware at clusters in the data, leading to a modeling of the local density.

Table 2. Classification performance on the continuous dataset.

	Overall accuracy	False Positive Rate	False Negative Rate
BLOSSOM 1NN	87%	3,7%	40%
Toroid ESOM 2NN	88,7%	8,6%	20,6%
Flat ESOM 2NN	88,9%	8,8%	18%
Toroid ESOM 1NN	94,4%	0,3%	21%
Flat ESOM 1NN	94,7%	0,3%	20%

From table 1 and table 2, one may conclude that the overall accuracy of the 1NN classifier based on the BLOSSOM map and the overall accuracy of the 2NN classifier based on the ESOM map are almost equal. However, a clear difference can be observed in the false positive rates and the false negative rates. The false positive rate (i.e. the number of non-domestic violence cases that were incorrectly classified as domestic violence, divided by the number of non-domestic violence cases contained in the dataset) for the BLOSSOM map more than twice as good as the false positive rate for the ESOM map. The opposite is true for the false negative rate (i.e. the number of domestic violence cases that were not classified as such by the NN classifier, divided by the number of domestic violence cases contained in the dataset). Surprisingly, there is almost no difference in classification performance for the flat and the toroidal ESOM map. Although the former map contains many undesired distortions, this does not result in a lower classification accuracy. Another interesting result is that, although the overall classification accuracy on the continuous dataset is only slightly worse, there is a very large difference between the false negative rates for both datasets. Since false negatives are critical, the ESOM map is better suited for our case than BLOSSOM. Finally, it should be noted that more complex classifiers such as the SVM did not perform better than the ESOM or BLOSSOM, and that the currently used system for our case is a multi-layer perceptron, which does not provide any insight into the problem (since it is a black-box), and its performance is around 80% only.

6 Conclusions and future work

In this paper, the usefulness of two recent SOM tools for studying an interesting police dataset was showcased. By applying the ESOM tool, it was possible to discover that the distribution of the dataset is spherical. By consequence, the spherical SOM tool BLOSSOM seemed natural to apply. By using this spherical SOM technique, interesting results for exploratory purposes of the data were discovered. Finally, a comparison between the ESOM and the BLOSSOM maps was performed by means of a nearest-neighbor classifier. However, it should be noted that a full-fledged benchmarking of the ESOM and BLOSSOM tools, using the full dimensionality of 123 features, is beyond the scope of this paper and is a topic for future research.

The authors are grateful to the Amsterdam-Amstelland Police, for providing us with the data. Jonas Poelmans is aspirant of the Research Foundation – Flanders.

References

[1] Keus, R., Kruijff, M.S. (2000) Huiselijk geweld, draaiboek voor de aanpak. Directie Preventie, Jeugd en Sanctiebeleid van de Nederlandse justitie.
[2] Watts, C., Timmerman, C.: Violence against women: global scope and magnitude. The Lancet 359 (9313): pp. 1232-1237. RMID 1155557.
[3] Waits, K. (1984-1985). The criminal Justice System's response to Battering: Understanding the problem, forging the solutions. Washington Law Review 60: pp. 267-330
[4] Catriona Minleer-Black (1999) Domestic violence: Findings from a new British Crime Survey self-completion questionnaire. London: Home Office Research Study.
[5] Vincent, J.P., Jouriles, E.N. (2000) Domestic violence. Guidelines for research-informed practice. Jessica Kingsley Publishers London and Philadelphia.
[6] Ritter, H. (1999) Non-Euclidean Self-Organizing Maps, pp. 97–109. Elsevier, Amsterdam.
[7] Kohonen, T. (1982), "Self-Organized formation of topologically correct feature maps", Biological Cybernetics, Vol. 43, pp. 59-69.
[8] Ultsch, A., Moerchen, F. (2005) ESOM-Maps: Tools for clustering, visualization, and classification with Emergent SOM. Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, No. 46.
[9] Ultsch, A., Hermann, L. (2005) Architecture of emergent self-organizing maps to reduce projection errors. In Proc. ESANN 2005, pp. 1-6.
[10] Ultsch, A. (2004) Density Estimation and Visualization for Data containing Clusters of unknown Structure. In proc. GfKI 2004 Dortmund, pp. 232-239.
[11] Ultsch, A. (2003) Maps for visualization of high-dimensional Data Spaces. In proc. WSOM'03, Kyushu, Japan, pp. 225-230.
[12] Ultsch, A., Siemon, H.P. (1990) Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Proc. Intl. Neural Networks Conf., pp. 305-308.
[13] Tokutaka, H., BLOSSOM Software Tool, http: //www.somj.com
[14] Nakatsuka, D., Oyabu, M. (2003) Application of Spherical SOM in Clustering. Proc. Workshop on Self-Organizing Maps (WSOM '03), pp. 203-207
[15] Van Hulle, M. (2000) Faithful Representations and Topographic Maps from distortion based to information based Self-Organization. Wiley: New York.
[16] Peng, H., Long, F., Ding, C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on pattern analysis and machine intelligence, Vol. 27, no. 8.