Text Mining with Emergent Self Organizing Maps and Multi-Dimensional Scaling: A comparative study on domestic violence

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

[1]K.U.Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium
[2]Vlerick Leuven Gent Management School, Vlamingenstraat 83,
3000 Leuven, Belgium
Amsterdam-Amstelland Police, James Wattstraat 84,
1000 CG Amsterdam, The Netherlands
[4]Universiteit van Amsterdam Business School, Roetersstraat 11
1018 WB Amsterdam, The Netherlands
[5]K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie
Campus Gasthuisberg O&N2, Bus 1021, Herestraat 49
3000 Leuven, Belgium
{Jonas.Poelmans, Stijn.Viaene, Guido.Dedene}@econ.kuleuven.be
Paul.Elzinga@amsterdam.politie.nl
marc@neuro.kuleuven.be

Abstract

In this paper we compare the usability of ESOM and MDS as text exploration instruments in police investigations. We combine them with traditional classification instruments such as the SVM, Naïve Bayes, etc. We perform a case of real-life data mining using a dataset consisting of police reports describing a wide range of violent incidents that occurred during the year 2007 in the Amsterdam-Amstelland police region (the Netherlands). We compare the possibilities offered by the ESOM and MDS for iteratively enriching our feature set, discovering confusing situations, faulty case labelings and significantly improving the classification accuracy. The results of our research are currently operational in the Amsterdam-Amstelland police region for upgrading the employed domestic violence definition, for improving the training of police officers and for developing a highly accurate and comprehensible case triage model.

Keywords: Emergent Self Organizing Map (ESOM), multi-dimensional scaling (MDS), domestic violence, exploratory data analysis, knowledge discovery in databases, text mining

1. Introduction

In this paper, we compare a methodology based on topographic maps [4, 10] with multi-dimensional scaling [24], for discovering new knowledge from unstructured texts. Topographic maps perform a non-linear mapping of a high-dimensional data space to a low-dimensional one (usually 2-dimensional), which facilitates the visualization and exploration of the data structure [9]. An Emergent Self Organizing Map

(ESOM) is a more recent type of topographic map [5]. An Emergent SOM differs from a traditional SOM in that a very large number of neurons (at least a few thousands) is used. According to [15] "Emergence is the ability of a system to produce a phenomenon on a new, higher level." In order to achieve emergence, the existence and cooperation of a large number of elementary processes is necessary, so that large numbers of neurons can represent data clusters individually which facilitates their detection. In the traditional SOM, the number of nodes is too small to support emergence.

Multi-dimensional scaling (MDS) is a method that uses the similarity or dissimilarity among pairs of objects in the original space to represent the objects in a lower dimensional space for visualization purposes. In our case, we have used the classical metric MDS algorithm [25] to visualize, in a two-dimensional space, the distribution of police reports in the sense that two reports will be close to each other when their correlation is high [24].

We use MDS and ESOM for automating the detection of domestic violence from the unstructured texts comprising the police reports. An inquiry of the Ministry of Justice of the Netherlands showed that 45% of the population had experienced non-incidental domestic violence at some point. For 27% of the population, the incidents even occurred on a weekly or daily basis [2]. A couple of years ago, the Amsterdam-Amstelland police in the Netherlands decided to make domestic violence to one of its top priorities [16]. By consequence, the swift recognition of cases of domestic violence, and assigning these reports the appropriate label, are of the utmost importance. Unfortunately, intensive audits of police databases related to filed reports have shown that many reports tended to be wrongly classified.

In the past, the Amsterdam-Amstelland police adopted a case triage system that automatically filters out suspicious cases, which did not receive the domestic violence label, for in-depth manual inspection. Unfortunately the false positive rate of this system is over 80% and since this manual inspection and labeling is a time consuming task, several attempts have been made to develop a system that automatically assigns a domestic or a non-domestic violence label to these incoming cases [20,21]. A multi-layer perceptron and an SVM were used but, unfortunately, the classification accuracy was below

80%. Moreover, these techniques did not provide any insight into the performed classification, since they are black-box approaches.

In this paper, we show how we extracted new and important knowledge regarding domestic violence from the unstructured text present in police reports by using the ESOM and MDS. In addition, we perform a comparative study of both instruments and show how we developed an efficient and highly accurate automated classification model. A label can be automatically assigned to incoming cases with an accuracy of 89%. This is a major improvement over the previous situation where each retrieved case had to be dealt with manually.

The remainder of this paper is as follows. In section 2, the problem setting, with our motivation for using the ESOM and MDS in this study, and a description of the dataset, will be given. In section 3, the ESOM will be briefly discussed, after which, in section 4, the setup we used for ESOM and MDS will be detailed. In section 5, the results of our comparative study are discussed. Finally, section 6 concludes the paper.

2. Problem setting

2.1. Motivation

The Amsterdam police databases contain more than 8000 police reports from 2007 that contain a statement made by the victim of a violent incident. Immediately after a victim reported a crime, the police officer has to judge whether or not it is a case of domestic violence. Unfortunately, not all domestic violence cases are recognized as such by police officers and, by consequence, many police reports are wrongly assigned the "non-domestic violence" label (i.e. false negatives).

Text mining seemed an interesting approach for processing this large amount of information. Text mining has been defined as "the discovery by computer of new, previously unknown, information by automatically extracting information from different written resources" [13]. Unfortunately, due to the lack of a good thesaurus, which is lists the terms used for indexing the police reports, the vague definition of

domestic violence, the classification errors made by police officers, and the lack of a tool that facilitates an in-depth exploration of the data, previous text mining projects for the domestic violence case had failed [20,21].

About 5 years ago, a case triage system that automatically filters out suspicious cases for in-depth manual inspection and classification was introduced to substantially reduce the number of domestic violence cases that were not recognized as such. However, a large number of these retrieved cases were wrongly selected for further in-depth analysis and classification. Going back to 2007, only about 20% of the 1091 retrieved cases were reclassified as domestic violence. Given that it takes at least 5 minutes to read and classify a case, it is clear that a more accurate case triage model will result in major time savings.

2.2. Dataset

According to the department of Justice and the Netherlands police, domestic violence can be characterized as serious acts of violence committed by someone from 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].

The dataset consists of a selection of 4814 police reports describing violent incidents that occurred in 2007. All domestic violence cases from that period are a subset of this dataset. Each of these reports contains an official statement made by the victim to the police. Of these 4814 reports, 1657 were labeled as domestic violence; the others were not. An example of such a report, which was originally written in Dutch, is displayed in Figure 1.

The validation set consists of a selection of 4738 cases describing violent incidents from the year 2006 where the victim made a statement to the police. 1734 of these 4738 cases were labeled as domestic violence by police officers. In 2006, the in-place case triage system retrieved 1157 police reports, containing a victim statement, which had to be manually re-inspected by police-officers. 318 reports were relabeled as domestic violence, while 839 were labelled as non-domestic violence.

Title of incident	Violent incident xxx
Reporting date	24-10-2008
Project code	Domestic violence against ex-partner
Crime location	Amsterdam Wibautstraat yyy
Suspect (male) Suspect (18-45yrs)	Zzz
Address	Amsterdam Waterlooplein yyy
Involved (male) Involved (>45yrs)	Neighbours
Address	Amsterdam Wibautstraat www
Victim (female) Victim (18-45yrs)	Uuu
Address	Amsterdam Waterlooplein vvv

victim statement

Yesterday morning I was in my kitchen preparing the breakfast. My daughter was playing in the garden. Suddenly I heard my daughter screaming for help. I ran outside where I saw my ex-husband. He threatened me with a knife while he wanted to kidnap my daughter. I ran to him to save my daughter. At that moment he cut me with the knife in my hand. He stabbed me three times once in my arm en two times in face. The neighbours heard me yelling and came to help me while one of the neighbours called the police, etc.

Figure 1. Example police report.

The initial thesaurus contained 123 domain-specific terms. In our dataset, it is indicated for each police report which terms are present. The terms in the thesaurus are the features in our dataset and the initial set of terms was obtained in one of the following 2 ways: either by using standard text mining tools

such as Datadetective or after discussions with domain experts. An excerpt of this dataset is displayed in table 1.

Table 1. Excerpt of the dataset used during the research.

	kicking	Dad hits me	Stabbing	cursing	scratching	maltreating
Report 1	X	X				X
Report 2			X	X	X	
Report 3	X	X	X	X	X	
Report 4						X
Report 5				X	X	

We have binary document vectors in which for each report is indicated with a 0 or 1 if the term from the thesaurus appears in the report or not. Using term frequencies was not really useful since all documents are rather short and of equal length.

3. Emergent SOM

According to Ultsch and co-workers, the topology preservation of the traditional SOM projection is of little use when the map is small: the performance of a small SOM is argued to be almost identical to that of a *k*-means clustering, with *k* equal to the number of nodes in the map [5]. The ESOM is especially useful for visualizing sparse, high-dimensional datasets, yielding an intuitive overview of its structure [7]. An ESOM map is composed of a set of neurons I, arranged in a hexagonal topology map or lattice. A neuron $n_j \in I$ is a tuple (w_j, p_j) in the map, consisting of a weight vector w_j=(wj1,…,wjm} with wj $\in \mathbb{R}^m$, where *m* is the dimensionality of the weight vector space and a discrete position $p_j \in P$, where P is the map space. The data space D is a metric subspace of $\mathbb{R}^m$. The training set $E = \{x_1, ..., x_k\}$ with $x_1, ..., x_k \in \mathbb{R}^m$ consists of input samples presented during the ESOM training. The training algorithm used is the online training algorithm in which the best match for an input vector is searched for, and the

corresponding weight vectors, and also those of its neighboring neurons of the map, are updated immediately.

When an input vector x_i is supplied to the training algorithm, the weight w_j of a neuron n_j is modified as follows,:

$$\Delta w_j = \eta h(bm_i, n_j, r)(x_i - w_j)$$

with $\eta \in [0,1]$, r the neighborhood radius and h a non-vanishing neighborhood function. The best-matching neuron of an input vector $x_i \in D$

$$D \to I : bm_i = bm(x_i)$$

is the neuron $n_b \in I$ having the smallest Euclidean distance to x_i :

$$n_b = bm(x_i) \Leftrightarrow d(x_i, w_b) \le d(x_i, w_b) \forall w_b \in W .$$

Where $d(x_i, w_j)$ stands for the Euclidean distance of input vector x_i to weight vector w_j. The neighborhood of a neuron

$$N_f = N(n_f) = \{n_j \in M \mid h_{fj}(r) \neq 0\}$$

is the set of neurons surrounding neuron n_f and determined by the neighborhood set h. The neighborhood defines a subset in the map space of the neurons K, while r is called the neighborhood range.

ESOM maps maintain the neighborhood relationships that are present in the input space and provide the user with an idea of the structure of the dataset, its distribution (e.g. spherical) and the degree of overlap between the different classes. ESOM maps can be created and used for data analysis by means of the publicly available Databionics ESOM Tool [14].

4. Experiment setup

An initial analysis of the data revealed that the two cases (domestic and non-domestic violence) do not appear in separate clusters; hence, the use of a density estimation method followed by a labeling of the high density peaks as separate classes is not a viable approach (see Figure 2 where the high density

regions (darker pixels) do not correspond to separate labels). Topographic maps such as the ESOM rather approximate the data manifold onto which both cases can be projected.

4.1. ESOM and MDS parameter settings

In a first step, an ESOM map with a toroidal topology of the neurons as well as a flat topology were trained using this dataset, in order to capture the distribution of the dataset. To simulate the ESOM, we used the Databionics software and the effect of the parameter settings on the simulation is described in detail in the Databionics ESOM user manual [14]. We did not attempt to optimize them. 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. In the map displayed in Figure 2, 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). The red squares in all figures represent neurons that mainly contain domestic violence reports, whereas the green squares represent neurons that mainly contain non-domestic violence reports. The U-Matrix [12] is used as background visualization in the ESOM. The U-matrix displays the local distance structure at each neuron as a height value creating a 3D landscape of the high-dimensional data space. The height is calculated as the sum of the distances to all immediate neighbors normalized by the largest occurring height. This value will be large in areas where no or few data points reside (white colour) and small in areas of high densities (blue and green colour).

An analysis of the ESOM map in Figure 2 reveals that there is a domestic violence cluster located at the center of the map, and a domestic violence cluster running upward and to the left of the map. The latter continues over the edge of the map (note that the map is actually toroidal) and has an outlier on the right of the map. When we compared the flat ESOM maps to the toroidal ESOM maps, we found that the toroidal maps yield a slightly better visualization of the dataset (for example, the reader can compare

Figure 6 with Figure 8). The border effect was present in the flat map resulting in undesired distortions of the map. Most of the observed clusters were located at the left border of the map, and the two groups of domestic violence cases that can be observed in Figure 6 was less clearly distinguishable from the non-domestic violence cases and these cases were spread over the right side of the map. Therefore, it seemed more adequate to use a toroidal ESOM for visualizing this dataset.

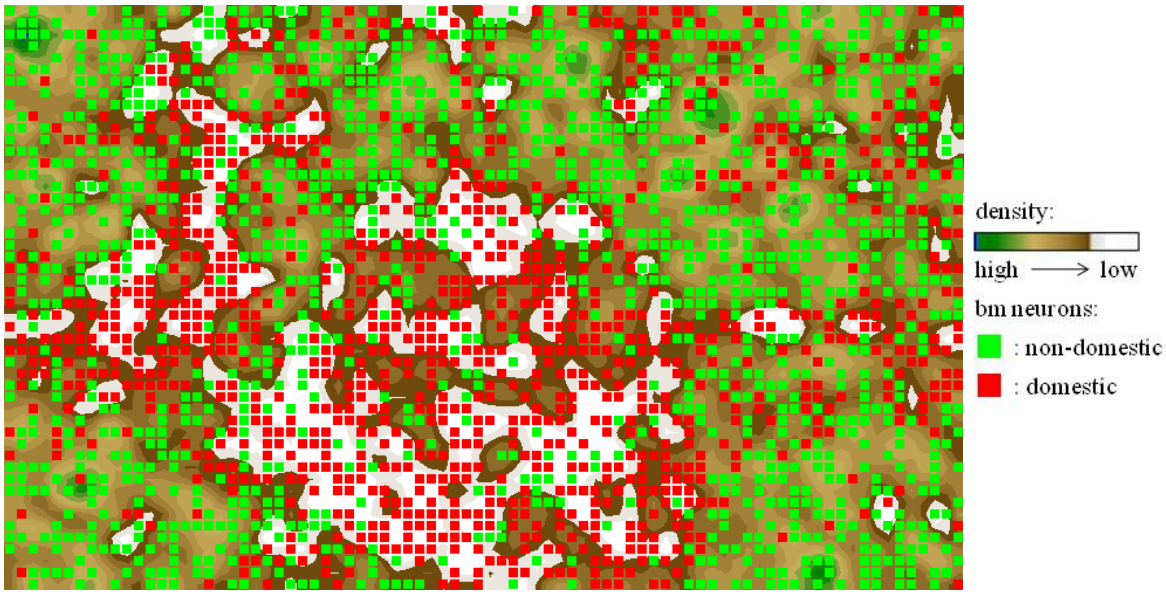

Figure 2. Toroidal ESOM map trained on the dataset

The map displayed in Figure 2 was trained directly on the entire dataset with 123 features. Figure 2 clearly shows that the density profile of the ESOM 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 class boundaries.

Both the MDS and ESOM can be used for detecting closely related data points, but each one has its own focus. Contrary to the ESOM, which starts directly from the document vectors, we first have to construct a dissimilarity matrix prior to the MDS calculation. In our case, it is a (symmetric) 4814 x 4814 matrix containing the Euclidean distances between each pair of normalized document vectors. The MDS algorithm [27], starts from this calculated distance matrix and uses a function minimisation algorithm to

find the best configuration in a lower dimension, i.e. a mapping of the original space on a two-dimensional space, thereby minimising the overall error. The error is defined as the sum of the squared differences between the distances in the original space (as present in the Euclidean distance matrix) and the corresponding ones in the lower dimensional space. We used the cmdscale algorithm from the R package for calculating the MDS map [25].

The output of an ESOM calculation is different from that of a metric MDS. The metric MDS algorithm concentrates on the largest dissimilarities whereas ESOM concentrates on the largest similarities. ESOM tries to reproduce the topology of the data in a 2D grid, instead of reproducing distances. Similar documents are represented by neighboring neurons in an ESOM, while a distance in an MDS map can be interpreted as an estimate of the true distance between both [26]. The MDS map trained on the same initial dataset is displayed in Figure 3. The red dots indicate police reports labelled as domestic violence, whereas green dots indicate police reports labelled as non-domestic violence.

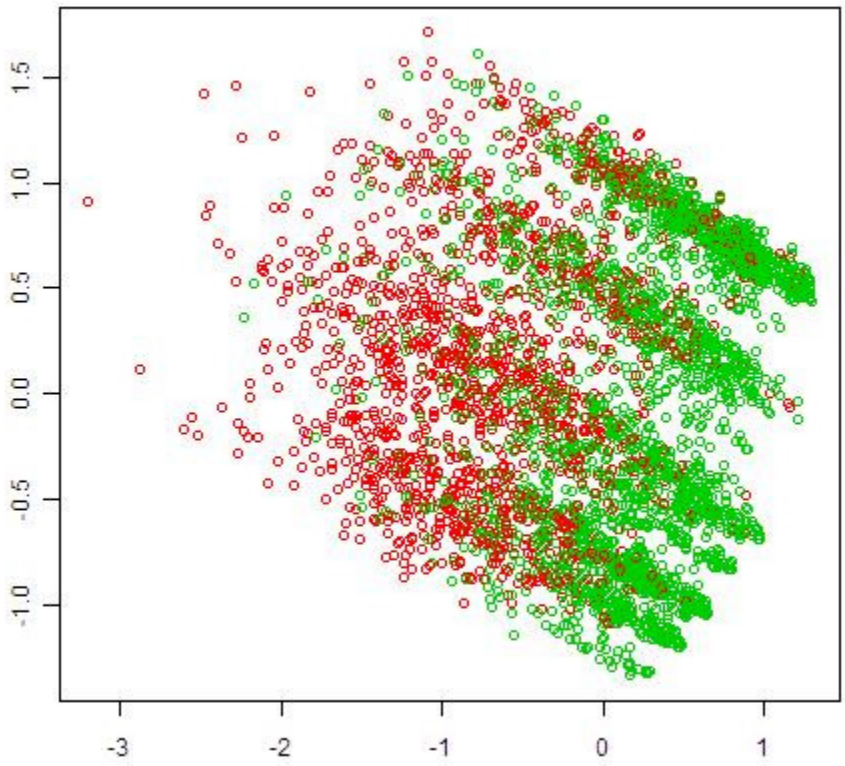

Figure 3. MDS map trained on the dataset

4.2. Data preprocessing and feature selection

We have applied feature selection to reduce the input space dimensionality, for the different classifiers discussed in section 4.3. We chose to select the 65 most relevant features. Feature selection comprises the identification of the most discriminative features of the observed data. Given the input data D consisting of N samples and M features $F = \{f_i,\ i = 1 \ldots M\}$, and the target classification variable c, the feature selection problem is to find from the M-dimensional observation space, R^M, a subspace of m features, R^m, that optimally characterizes c. A heuristic feature selection procedure, known as minimal-redundancy-maximal-relevance (mRMR), as described in [11], was considered. In terms of mutual information I, the purpose of feature selection is to find a subset S with m features $\{f_i\}$, which jointly have the largest dependency on the target class c. This is called the Max-Dependency scheme:

$$\text{Max } D(S,c),\ D = I(f_1,\ldots,f_m;c) \quad (1)$$

As the Max-Dependency criterion is hard to implement, an alternative is to select features based on maximal relevance criterion (Max-Relevance). Max-Relevance is to search features satisfying (2), which approximates $D(S,c)$ in (1) with the mean value of all mutual information values between individual feature f_i and class c:

$$\max D(S,c), D = \frac{1}{|S|}\sum_{fi \in S} I(f_i;c) \quad (2)$$

Features selected according to Max-Relevance could have redundancy, i.e. , the dependency among these features could be large. When two features highly depend on each other, the respective class-discriminative power would not change much if one of them was removed. Therefore, the following minimal redundancy (Min-Redundancy) condition can be added to select mutually exclusive features [19].

$$\min R(S), R = \frac{1}{|S|^2}\sum_{fi,fj \in S} I(f_i, f_j) \quad (3)$$

The criterion combining the above two constraints is called "minimal-redundancy-maximal-relevance) (mRMR). The operator $\Phi(D,R)$ is defined to combine D and R and the following is the simplest form to optimize D and R simultaneously:

$$\max \Phi(D,R), \Phi = D - R \quad (4)$$

The outcome of this filter approach is a ranked list of features. To decide on where to cut off this list we use the classifiers discussed in the next section.

4.3. Initial classification performance

To obtain the optimal feature set, an SVM, a Neural Network, a kNN (k-nearest-neighbor with k=3) and a Naïve Bayes classifier were used to measure the classification performance for an increasing number of features.

Naïve Bayes is based on the Bayes rule and assumes that feature variables are independent of each other given the target class. Given a sample $s=\{f_1,\ldots, f_m\}$ for m features, the posterior probability that s belongs to class c_i is

$$p(c_i \mid s) \propto \prod_{j=1}^{m} p(f_j \mid c_i)$$

where $p(f_j \mid c_i)$ is the conditional probability table learned from examples in the training process. Despite the conditional independence assumption, Naïve Bayes has been shown to have good classification performance for many real data sets [3]. We have used the WEKA package [22]. We used 10-fold cross-validation.

The Support Vector Machine (SVM) [6] is a more modern classifier that uses kernels to construct linear classification boundaries in high dimensional spaces. We make use of the LibSVM package [17]. A Radial Basis Function (RBF) was chosen as kernel, after sensitivity analysis 0.01 was found to be the optimal kernal parameter, 1 the optimal c value and 10-fold cross-validation was used.

Nearest neighbor methods estimate the probability $p(t|x)$ that an input vector $x \in R^n$ belongs to class $t \in \{0,1\}$ by the proportion of training data instances in the neighborhood of x that belong to that class. The metric used for evaluating the distance between $a, b \in R^n$ is the Euclidean distance:

$$dist(a,b) = \| a-b \|_2 = \sqrt{(a-b)^T (a-b)}$$

This version of the *k*-nearest neighbor was was chosen because it is especially appropriate for handling discrete data [18]. The problem with discrete data is that several training data instances may be at the same distance from a test data instance x as the kth nearest neighbor, giving rise to a non-unique set of *k*-nearest neighbors. The *k*-nearest neighbor classification rule then works as follows. Let the number of training data instances at the distance of the kth nearest neighbor be n_k, with n_{k1} data instances of class $t = 1$ and n_{k0} data instances of class $t = 0$. Let the total number of training data instances within, but excluding this distance be N_k, with N_{k1} data instances of class $t = 1$ and N_{k0} data instances of class $t = 0$ if

$$N_{k1} + \frac{k - N_k}{n_k} \times n_{k1} \geq N_{k0} + \frac{k - N_k}{n_k} \times n_{k0}$$

where $N_k < k \leq N_k + n_k$. Now all training data instances at the distance of the kth nearest neighbor are used for classification, albeit on a proportional basis. The parameter k was set to 2 and 10-fold cross-validation was used.

We also used a feed-forward multiplayer perceptron (MLP) with one hidden layer consisting of 5 neurons and an output layer consisting of one neuron [23]. The weight decay parameter was set to 0.2 and the number of training cycles to 10. Again we used 10-fold cross-validation.

The classification performance is plotted as a function of the number of features in Figure 4. The result of the mrmr algorithm is a ranked list of the best features. The x-axis indicates how many of these best features were used to train the classifiers. The y-axis shows the classification performance for these different feature subsets. We opted to retain the best 44 features which is a compromise for the 4 classifiers. 44 features was one of the points in the curve where the sum of classification performances for

the different classifiers was highest. We also tested other maxima such as 15 and 30 but this resulted in a less qualitative graphical image. A toroidal ESOM map was trained on this dataset with a reduced number of features and was compared to that of Figure 2. It shows that the density problem (one class label for each density peak) was not solved by lowering the number of features (result not shown).

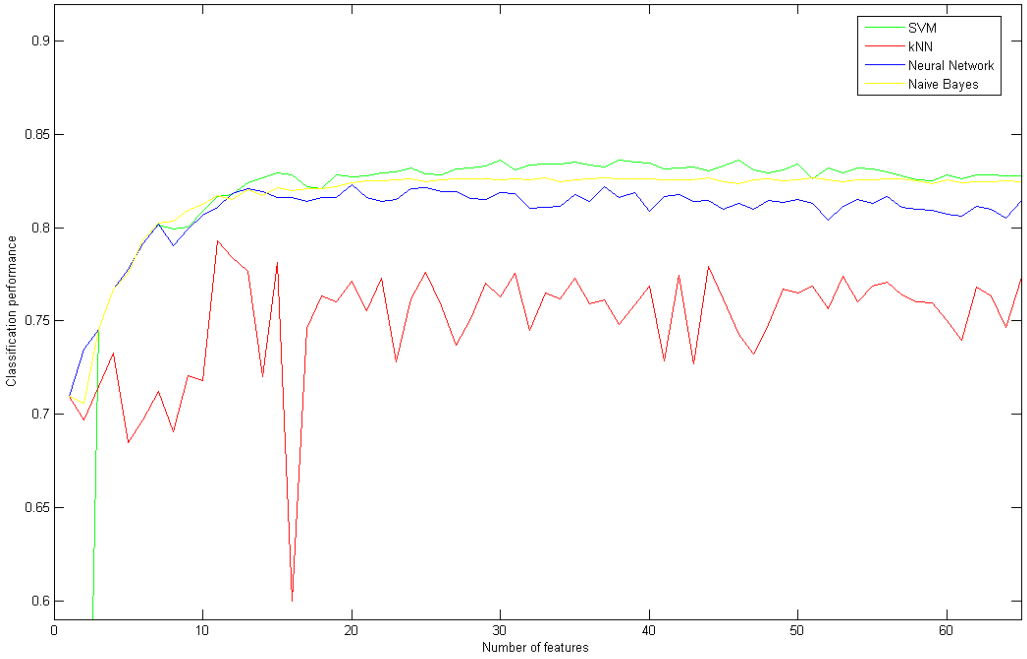

Figure 4. Classification performance for different subsets of the ranked list of features produced by the mrmr algorithm

5. Results

5.1 ESOM analysis and knowledge discovery method

In this section, we describe our method for knowledge discovery from unstructured text using the ESOM and MDS tools. We use the maps to efficiently browse through the underlying data and select interesting reports for in-depth manual inspection. We focus on how the domain expert can use this visual instrument for iteratively enriching his knowledge on the problem domain. We particularly make use of the color coding functionality offered by the ESOM and MDS tools for analyzing maps. The distribution of the best matching neurons gives the exploratory data analyst clear indications of where to look. The intuitive interface of the ESOM and MDS tool allows the domain expert to manually select data clusters and interesting neurons. Interesting neuroms are typically defined as those neurons for which the label assigned after training does not match the label of a high percentage of neighboring neurons. The cases that have these neurons as a best match can be shown. We then select representative reports for in-depth manual inspection. In these texts, the terms contained in the thesaurus are highlighted. We search these reports for interesting new domain-specific terms and concepts for enriching our thesaurus and our prior knowledge about domestic violence. We also discover situations that are confusing for police officers i.e. situations to which different police officers assign different labels and we detect situations that are typically assigned a faulty case label by police officers. We presented these doubtful cases to the board members, responsible for the domestic violence policy, to enrich and refine the employed domestic violence definition for better training the police officers. We, finally, used the ESOM and MDS map to develop a comprehensible case classification model that is currently operational in the Amsterdam-Amstelland police department to automatically assign a domestic or non domestic violence label to incoming cases.

5.2 Exploring the concept of domestic violence

In this section we show how ESOM and MDS can be used in a real-life knowledge discovery setting to explore a large amount of police reports containing unstructured text. First, in-depth manual inspection of the police reports corresponding to some of the outlier neurons in the map in Figure 2 led to some interesting discoveries. We found that only a small part of these police reports had been incorrectly

labelled by police officers as domestic or as non-domestic violence. Surprisingly, many of these reports turned out to contain a large number of important features and concepts that were lacking in the expert's initial understanding of the problem domain. Examples of such newly discovered features are homosexual relationships, extramarital affairs, pepper spray, sexual abuse, etc. After multiple successive iterations of refining the thesaurus, training a new map, and analyzing the resulting ESOM, our thesaurus contained more than 800 domain-specific terms, term combinations and term clusters.

Before we gave the data as input to the classifiers of section 4.3, we again applied the mrmr as feature selection procedure. We found that the classification accuracy of the SVM, Neural network, Naïve Bayes and kNN classifiers improved significantly after adding the newly discovered features to the thesaurus. For example, for the SVM, the best classification accuracy on the initial dataset was around 83%, while the best classification accuracy on the dataset with the refined thesaurus was around 89%.

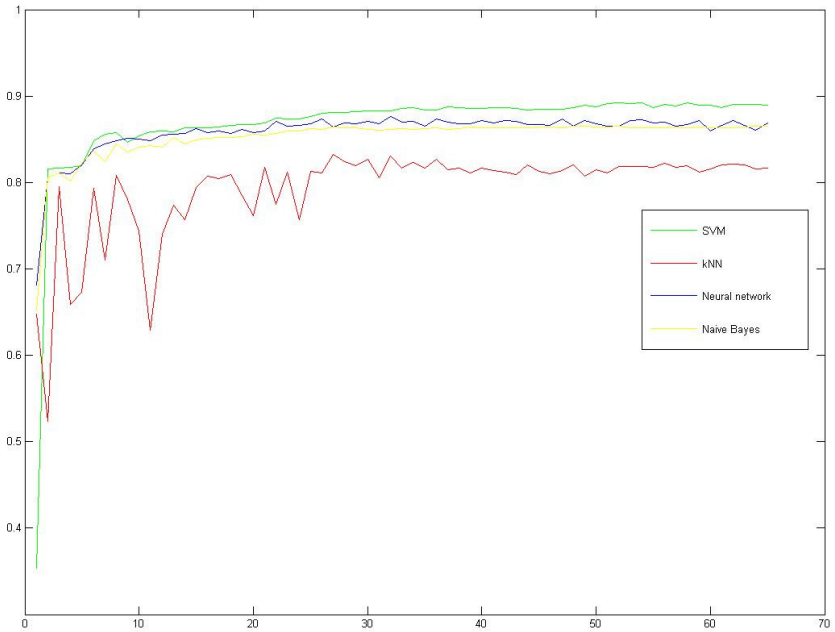

Figure 5. Classification performance

We also trained a new toroidal ESOM map and MDS map on the dataset based on the refined thesaurus. The resulting map is displayed in Figure 6. The resulting MDS map is displayed in Figure 7.

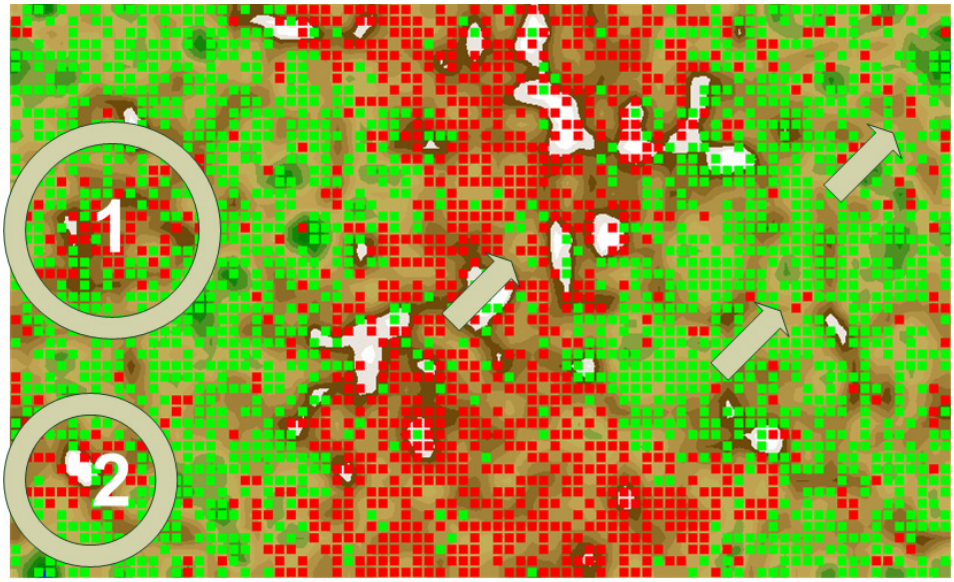

Figure 6. Toroid ESOM map trained on the dataset. See text for an explanation of the arrows and circles.

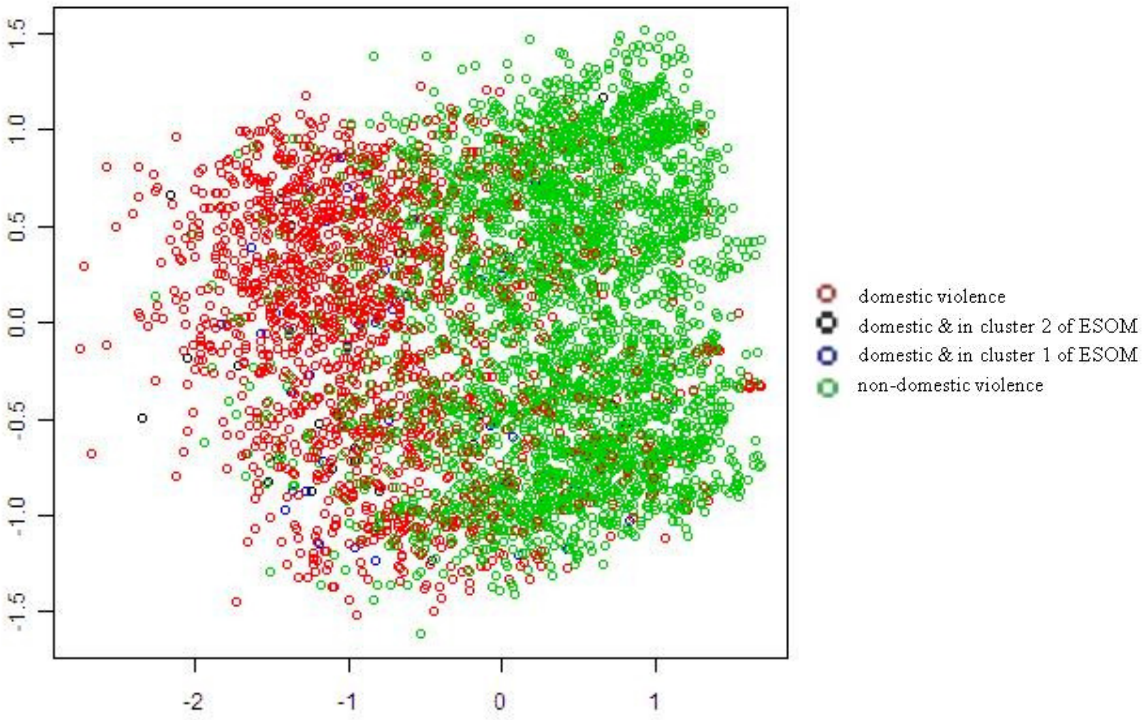

Figure 7. MDS map trained on the dataset.

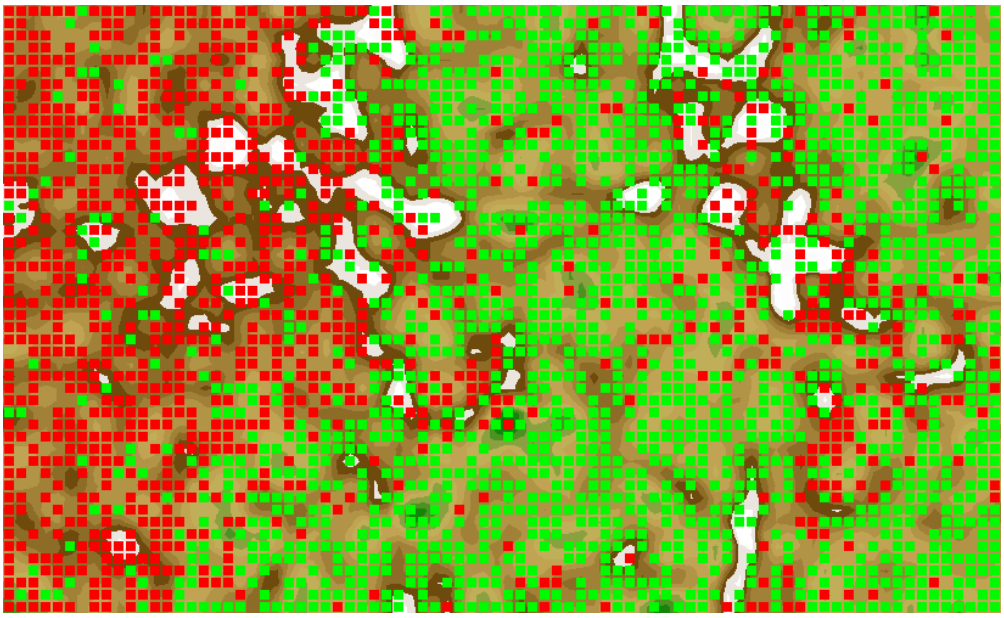

Figure 8. Flat ESOM map trained on the dataset.

Comparing the ESOM map of Figure 2 to that of Figure 5 reveals that the amount of overlap between the two classes has decreased significantly after the refined thesaurus was introduced. The map in Figure 5 shows 3 different clusters that mainly contain cases labelled as domestic violence. When we inspected the cases contained in the top left cluster (circle marked as 1), we found that this cluster mainly contained burglary cases that for some unknown reason were wrongly labeled by police officers. During analysis of the cluster located at the left and bottom of the map (circle marked as 2) and some of the outliers (arrows), we discovered a large number of situations that were found to be confusing for police officers. Their opinions differed on how these cases should be labelled. No such clusters were found in the MDS map. Finally, we found that the outliers mostly contained cases that were wrongly labeled as either domestic or non-domestic violence. We conclude that ESOM is better suited in our case for knowledge discovery purposes.

Although we found the graphical display of the ESOM tool to be more intuitive to use, quantitative analysis revealed that the separation of the data by the MDS algorithm is slightly better. We developed a kNN classifier based on the ESOM and MDS maps for predicting the label of the new incoming cases such as the ones retrieved by the in-place case triage system. The classification performance of the MDS

for unseen cases (80,5%) was 3% higher than that of the ESOM (77,5%). This is probably caused by the ESOM algorithm which recognized two extra clusters besides the large one in the middle. Separating these data in separate clusters reduced classification performance but was found to be useful for exploratory analysis.

6. Conclusions

In this paper, we have performed a comparative study of ESOM and MDS for analyzing large amounts of unstructured text. We showcased the exploratory capabilities offered by the tools for discovering new features, and for detecting confusing situations and faulty case labelings. Whereas previously introduced data mining techniques for domestic violence often run as a black-box without user intervention, our methodology engages the domain expert in the discovery process and immerses him/her in the data. The focus lies on incrementally enriching his/her knowledge about the problem area by offering the expert an intuitive interface for browsing through the data. Whereas in the past, this domain expert used to be numbed by the overload of data, our method allows for the integration of his expert knowledge in the discovery process. The police officers who tested both techniques are more satisfied of the user interface of the ESOM tool to be more appealing for exploring this vast amount of police reports than the MDS. Moreover the ESOM was able to recognize two extra data clusters that were of significant importance but not found by MDS. Quantitative comparison turned out to be in favor of the MDS. A kNN classifier built on the MDS map had a 3% higher accuracy than that on the ESOM map. Based on the findings of our research, we managed to improve a classification performance of the SVM to 89% accuracy. A topic for future research is to apply the ESOM to other types of criminal incidents and to build a triage system for automatically distinguishing between them.

7. Acknowledgements

The authors would like to thank the Amsterdam-Amstelland region police and in particular Deputy Chief Reinder Doeleman and Chief Hans Schönfeld for supporting this 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" or Fonds voor wetenschappelijk onderzoek – Vlaanderen. We would like to express special thanks to Gert Van Dijck for sharing his insights in feature selection and Patrick De Mazière for his help with the R package. We would like to thank Karel Dejaeger for his help with the dissimilarity matrix calculation.

8. References

[1] Keus, R., Kruijff, M.S. (2000) Huiselijk geweld, draaiboek voor de aanpak. Directie Preventie, Jeugd en Sanctiebeleid van de Nederlandse justitie.
[2] Dijk, V.T. (1997) Huiselijk geweld, aard, omvang en hulpverlening (Ministerie van Justitie, Dienst Preventie, Jeugdbescherming en Reclassering, oktober 1997)
[3] Cover, T., Thomas, J. (1991) Elements of information theory. New York: Wiley..
[4] Kohonen, T. (1982), "Self-Organized formation of topologically correct feature maps", Biological Cybernetics, Vol. 43, pp 59-69.
[5] 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
[6] Vapnik, V. (1995) The Nature of Statistical Learning Theory. New York: Springer.
[7] Ultsch, A. (2004) Density Estimation and Visualization for Data containing Clusters of unknown Structure. In proc. GfKI 2004 Dortmund, pp 232-239
[8] Ultsch, A. (2003) Maps for visualization of high-dimensional Data Spaces. In proc. WSOM'03, Kyushu, Japan, pp. 225-230
[9] Ultsch, A., Siemon, H.P. (1990) Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Proc. Intl. Neural Networks Conf., pp305-308
[10] Van Hulle, M. (2000) Faithful Representations and Topographic Maps from distortion based to information based Self-Organization. Wiley: New York
[11] 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.
[12] Ultsch, A., Hermann, L. (2006) Automatic clustering with U*C, Technical Report, Dept. of Mathematics and Computer Science, Philippe-University of Marburg.
[13] Fan, W., Wallace, L., Rich, S., Thang, T. (2006) Tapping the power of text mining. Communications of the ACM, Vol. 49, no. 9
[14] http://databionic-esom.sourceforge.net/
[15] Ultsch, A. (1999) Data mining and knowledge discovery with Emergent SOFMS for multivariate Time Series. In Kohonen Maps, pp. 33-46
[16] http://www.politie-amsterdam-amstelland.nl/get.cfm?id=86
[17] Hsu, C.W., Lin, C.J. (2002) A comparison of methods for Multi-Class Support Vector Machines. IEEE Trans. Neural Networks, vol. 13 pp. 415-425.
[18] Webb, A.R. (1999) Statistical pattern recognition. Arnold.
[19] Ding, C., Peng, H.C. (2003) Minimum Redundancy feature Selection from MicroArray Gene Expression Data. Proc. Second IEEE Computational Systems Bioinformatics Conf., pp. 523-528, Aug. 2003.
[20] Raaijmakers, S.A., Kraaij, W., Dietz, J.B. (2007) Automatische detectie van huiselijk geweld in processen-verbaal. TNO-rapport 34293.
[21] Elzinga, P. (2006) Textmining by fingerprints. Onderzoeksrapport huiselijk geweld zaken. IGP project Activiteit 0504
[22] http://www.cs.waikato.ac.nz/ml/weka/
[23] http://www.informedia.cs.cmu.edu/yanrong/MATLABArsenal/MATLABArsenalDoc/MATLABArsenal/

NeuralNet.html
[24] Borg, I., Groenen, J.F. (2005) Modern multidimensional scaling: theory and applications. Springer series in statistics.
[25] Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325–328.
[26] Wehrens, R., and Buydens, L. M. (2007). Self- and super-organizing maps in R: The kohonen package. Journal of Statistical Software, 21 (5), 1–19.
[27] Kruskal, J. B., and Wish, M. (1978). Multidimensional Scaling, Sage University Paper series on Quantitative Application in the Social Sciences. Beverly Hills and London: Sage Publications.