# АНАЛИЗ ТЕКСТОВ ПОЛИЦЕЙСКИХ ОТЧЕТОВ С ПОМОЩЬЮ ЭМЕРДЖЕНТНЫХ САМООРГАНИЗУЮЩИХСЯ КАРТ И МНОГОМЕРНОГО ШКАЛИРОВАНИЯ*

***Poelmans J.,*** *Ph.D., PostDoc researcher*
*Katholieke University Leuven*
*e-mail: Jonas.Poelmans@econ.kuleuven.be*
***Игнатов Д.И.****, к.т.н., старший преподаватель*
*Высшая школа экономики (НИУ)*
*e-mail: dignatov@hse.ru*
***Marc M. Van Hulle,*** *Ph.D., Professor*
*Katholieke University Leuven*
***Stijn Viaene,*** *Ph.D., Vlerick Mangement School*
***Guido Dedene,*** *Ph.D., Professor*
*Katholieke University Leuven*
*Universiteit van Amsterdam*
***Paul Elzinga****, Doctoral Researcher*
*Amsterdam-Amstelland Police*

## 1. ВВЕДЕНИЕ

В этой работе мы приводим основные результы исследования [8] и проводим сравнение методов топографических карт и многомерного шкалирования [10] для извлечения знаний из неструктурированных текстов. Топографические карты осуществляют нелинейное отображение многомерных данных в пространство меньшей размерности (как правило, 2-мерное), что облегчает визуализацию и изучение структуры данных. Эмерджентные самоорганизующиеся карты (ESOM) наиболее современный тип топографических карт [2]. Эмерджентные SOM в отличие от традиционных SOM имеют большее количество нейронов (по краней мере несколько тысяч). Эмерджентность – это способность системы воспроизводить явление на новом, более высоком уровне. Для достижения эмерджентности необходимо сосуществование и взаимодействие большого числа элементарных процессов, таким образом, большое число нейронов

*Jonas Poelmans поддержан как аспирант «Исследовательского фонда Фландрии» (Fonds voor wettenschappelijk onderzoek – Vlaanderen)

может представлять кластеры данных идивидуально, что упрощает их обнаружение. В традиционных SOM число нейронов слишком мало для того чтобы поддерживать эмерджентность. Многомерное шкалирование (MDS) – это метод, использующий сходство и различие среди пар объектов в исходном пространстве для представления их в пространстве меньшей размерности с целью визуализации. В нашем случае используется классический метрический алгоритм MDS [11] для визуализации в двумерном пространстве полицейских отчетов, в том смысле, что два отчета близки друг другу, если их числовое сходство высоко [10].

Мы используем MDS и ESOM для автоматического выявления случаев домашнего (бытового) насилия в неструктурированных текстах полицеских отчетов. Данные министерства юстиции Нидерландов показывают, что 45% населения сталкивались, так или иначе, с домашнм насилием. Для 27% населения такие случаи происходили еженедельно и даже ежедневно [1]. Пару лет назад полиция Амстредама-Амстелланда в Нидерландах решила сделать проблему домашнего насилия одним из приоритетных направлений [7], поэтому оперативное выявление случаев домашнего насилия и классификация соответствующих отчетов стало крайне важной задачей. К сожалению, массовые проверки полицеских баз данных, содержащих текстовые отчеты, показали, что многие отчеты неверно классифицированы.

Ранее полиция Амстердама-Амстелланда разработала систему сортировки отчетов, которая автоматически отбирает подозрительные случаи, не помеченные как домашнее насилие, для более тащательного анализа экспертами. К сожалению, количество ложных положительных случаев, обнаруживаемых системой, выше 80%, что значительно замедляет проверку вручную и классификацию. Было предпринято несколько попыток разработать систему, которая автоматически классифицирует случаи домашнего и не домашнего насилия [9]. Для этого были использованы многослойные персептроны и машины опроных векторов SVM, но, к сожалению, точность классификации оказалась ниже 80%. Более того, эти методы не дают понимания, на основе чего произведена такая классификация, оставаясь для пользователя черным ящиком.

Помимо извлечения важных знаний с помощью ESOM и MDS мы проводим сравнительное исследование этих двух инструментов и показываем, что мы разработали эффективную и высокоточную модель автоматической классификации (точность до 89%). Это

основное улучшеное предыдущей модели, в которой каждый случай обрабатывался вручную.

## 2. ПОСТАНОВКА ЗАДАЧИ

База данных полции Амстердама содержит более чем 8000 полицейских отчётов за 2007 год, каждый отчет включает показания жертвы конкретного случая насилия. Сразу после того как жертва сообщила о преступлении, офцер полиции должен определить является этот случай домашним насилием или нет. К сожалению, не все случаи домашнего насилия распознаются как таковые офицерами полиции, вследствии чего многие отчеты помечены как “не домашнее насилие” (т.е. ложные отрицательные).

Анализ текстов (Text mining) является многообещающим подходом для обработки большого количества информатиции и определяется как “как обнаружение с помощью компьютера новой, ранее не известной информации, путем автоматического её извлечения из различных текстовых ресурсов” [6]. По причине отсутствия хороших тезаурусов, т.е. списков терминов используемых для индексирования полицеских отчетов, недостаточного четкого определения домашнего насилия, ошибок классификации, совершаемых офицерами полиции и недостатке инструментов, которые служат для более глубокого исследования данных, предыдущие проекты по анализу текстов для выявления случаев домашнего насилия провалились [9].

Примерно 5 лет назад система автоматической сортировки полицейских отчетов, которая отбирает подозрительные случаи для детального ручного анализа, была введена в эксплуатацию для того чтобы существенно сократить количество невыявленных случаев домашнего насилия. Однако большое количество этих выявленных случаев были неверно отобраны для последующего детального анализа и классификации. Возвращаясь к 2007 году можно отметить, что только около 20% из 1091 отобранных случаев были отнесены к категории домашнее насилие. И это при том, что на чтение и классификацию каждого случая требуется как минимум 5 минут. Необходимость более точной модели, сберегающей время, очевидна.

Согласно данным Департамента юстиции, а также Полици Нидерландов, домашнее насилие может быть описано как серъезные действия насилия совершенные кем-либо из домашнего окружения жертвы. Насилие подразумевает все формы физического насилия. Домашнее окружение включает всех партнеров, экспартнеров, членов семьи, родственников и друзей семьи жертвы. Друзья семьи – это те

люди, которые состоят в дружественных отношениях с жертвой и (регулярно) встречаются с жертвой в его/её доме [1].

Исходный набор данных состоит из выборки 4814 полицеских отчетов, описывающих случаи насилия произошедшие в 2007. Все случаи домашнего насилия, начиная с этого периода – подмножество этого набора данных. Каждый из этих отчетов содержит официальные показания жертвы, данные полиции. Из этих 4814 отчетов 1657 были помечены как домашнее насилие, остальные – нет.

Контрольная выборка состоит из отобранных 4378 отчетов за 2006. 1734 из 4738 случаев были помечены как домашнее насилие офицерами полиции. В 2006 году система сортировки отчетов отобрала 1157 из них для повторного изучения офицерами полиции. 318 отчетов были помечены как случаи домашнего насилия, в то время как 839 были помечены как «не домашнее» насилие.

Исходный тезаурус содержал 123 термина предметной области. В нашем наборе данных показано, какой термин присутствовал в конкретном полицейском отчете. Термины тезауруса – признаки в нашем наборе данных, а исходное множество терминов было получено одним из следующих двух способов: либо использованием стандартных средств анализа текстов, таких как Datadetective или после обсуждения с экспертами предметной области. Фрагмент этих данных приведен в таблице 1.

Таблица 1. Фрагмент набора данных, используемых в работе.

| | 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 | |

Мы имеем бинарный вектора документов, в которых 0 и 1 показывают, встречается ли данный признак в отчете или нет. Использование частоты термов не было бы реально полезным, учитывая, что они короткие примерно равной длины.

## 3. ЭМЕРДЖЕНТНЫЕ SOM (ESOM)

Сохранение топологии традиционных SOM проекций не дает преимуществ в случае, когда карта мала: производительность небольших SOM почти идентична кластеризации *k*-средних, при *k* равному числу узлов карты [2]. ESOM особенно полезны для

визуализации данных большой размерности и дают интуитивное представление об их структуре [3].

ESOM состоит из множества нейронов $I$, имеющих гексагональную топологию (решетку). Нейрон $n_j \in I$ является набором $(w_j, p_j)$ карты, состоящим из вектора весов $w_j=(w_{j1},\ldots,w_{jm})$ с $w_j \in \mathbb{R}^m$, где m размерность пространства вектора весов, а $p_j \in P$ его дискретное положение ($P$ – пространство карты). Пространство данных $D$ – метрическое подпространство $\mathbb{R}^m$. Обучающее множество $E = \{x_1,\ldots,x_k\}$, такое, что $x_1,\ldots,x_k \in \mathbb{R}^m$ состоит из входных экземпляров, предъявляемых ESOM на этпае обучения. Мы используем алгоритм онлайн обучения, который находит лучшее совпадение для входного вектора и соотвествующего вектора весов и также для его соседних нейронов на карте, которые обновляются сразу же.

Для входного вектора $x_i$, попадающего на вход обучающегося алгоритма, вес $w_j$ нейрона $n_j$ изменяется следующим образом:

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

где $\eta \in [0,1]$, $r$ – радиус окрестности и $h$ – не стремящяяся к нулю функция окрестности. Лучший по совпадению нейрон $x_i \in D$

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

это такой нейрон, который имеет наименьшее Евклидово расстояние до $x_i$:

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

Здесь $d(x_i, w_j)$ обозначает Евклидово расстояние от входного вектора $x_i$ до вектора весов $w_j$. Окрестность нейрона

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

состоит из множества нейронов окружающих нейрон $n_f$ и определяется как множество окрестности $h$. Окрестность определяет подмножество на пространстве карты нейронов $K$, а $r$ называется диапазоном соседства.

ESOM используют отношение соседства во входном пространстве, что дает пользователю представление о структуре данных, характере их распределения (например, сферическое) и степени перекрытия различных классов. Существует программное средство Databionics ESOM Tool, разработанное для построения таких карт [4].

## 4. ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

На первом шаге ESOM карты с тороидальной и плоской топологией нейронов обучались на входных данных. В качестве программного обеспечения мы использовали Databionics [4]. На карте, показанной на рисунке 1, наилучшие совпадения узлов (ближайше соседи) помечены как два различных класса для наших данных (красный для случаев домашнего насилия, зеленый для «не домашнего» насилия). Красные квадраты на всех рисунках представляют нейроны, которые в основном содержат отчеты по случаям домашнего насилия, а зеленые – «не домашнего». U-матрица используется как фон на карте ESOM. U-матрица показывает локальные расстояния структуры для каждого нейрона как значение высоты, создающее трехмерный ландшафт многомерного пространства данных. Высота рассчитывается как сумма расстояний до всех непосредственных соседей, нормированная по наибольшей высоте. Это значение будет большим для областей, где нет или очень мало точек (белый цвет) и мало в областях высоких плотностей (синий и зеленый цвет).

Анализ ESOM на рисунке 1 показал, что существует кластер случаев домашнего насилия, расположенный в центре карты, и кластер домашнего насилия, который простирается вверх влево на карте. Этот последний кластер проходит вдоль ребра карты (карта на самом деле тороидальная) и имеет выброс справа на карте.

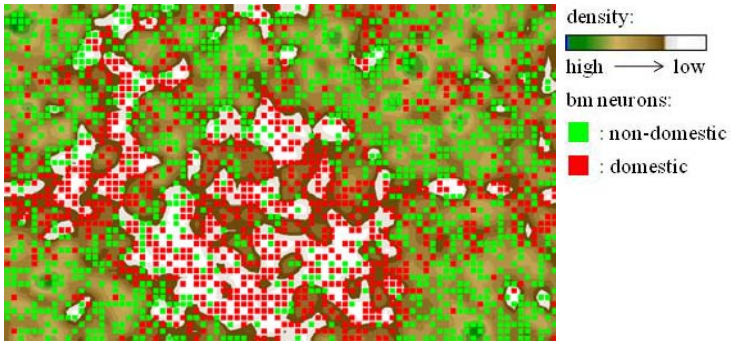

**Рис. 1. ESOM карта для исходных данных**

Карта, изображенная на рисунке 1 посторена на основе 123 признаков. Рисунок 1 ясно показывает, что профиль плотности ESOM карты не совпадает с пометками случаев домашнего насилия. Более того, нет четкой границы на карте, которая отделяет классы домашнего от «не домашнего» насилия. Следовательно, метод «водораздела» не приводит к корректному обнаружению границ классов.

Оба метода MDS и ESOM могут быть использованы для обнаруженя близко связанных точек данных, но каждый из них имеет свою специфику. В противоположность к ESOM, которые принимают на вход векторы документов, для применения MDS мы должны построить матрицу различий. В нашем случае это (симметричная) 4814x4814 матрица, содержащая Евклидовы расстояния между каждой парой нормализованных векторов. Алгоритм MDS [13] начинает с вычисления матрицы расстяний и использует минимизацию функции для того чтобы найти лучшую конфигурацию в пространстве низкой размерности, т.е. отображение исходного пространства в двумерное пространство, таким образом минимизируя общую ошибку. Ошибка определяется как сумма квадратов разностей между расстояниями в исходном пространстве (прдставлены в Евклидовой матрице) и соответствующими расстояниями в пространстве низшей размерности. Мы использовали cmdscale алгоритм из пакета R для вычисления MDS карты [11].

Выход ESOM алгоритма отличается от выхода метрического MDS. Метрический MDS алгоритм уделяет внимание большим различиям, в то время как ESOM концентрирует внимание на больших сходствах. ESOM пытается воспроизвести строение данных на двумерной решетке, а не само расстояние. Сходные документы представлены соседними нейронами в ESOM, в том время как расстояния на MDS карте могут быть интерпретированы как оценки истинного расстояния между ними [12].

Мы покажем, как ESOM и MDS могут быть использованы в приложениях, работающих с реальными данными для информационного поиска в большом количестве полицейских отчетов содержащих неструктурированный текст. Подробное изучение вручную полицейских отчетов, соответствующих нейронам выбросам на рисунке 2 привело к некоторым интересным открытиям. Мы обнаружили, что только небольшая часть этих полицейских отчетов была некорректно классифицирована как домашнее или не домашнее насилие. К нашему удивлению многие из этих отчетов содержали

большое количество признаков и понятий, которые отсутствовали в исходном представлении экспертами предметной области. Примеры таких новых признаков гомосексуальные отношения, до- и внебрачные отношения, перцевый спрей, сексуальное насилие и т.п. После нескольких последовательных итераций уточнения тезауруса, обучения новой карты и анализа результирующей ESOM, наш тезаурус содержал более 800 терминов предметной области, их комбинации и кластеры.

Перед тем как подать данные на вход классификатору мы применили эвристическую процедуру отбора признаков, известную как минимально-избыточная-максимальная-релевантность (mRMR) [5]. Качество классификации показано на рисунке 2. Результат работы mRMR алгоритма – ранжированный список лучших признаков. Ось *x* показывает, какое количество этих лучших признаков было использовано для обучения классификатора. Ось *y* показывает качество классификации для этих различных подмножеств признаков. Мы обнаружили, что точность классификаторов SVM, нейронных сетей, Naïve Bayes и kNN значительно улучшилась после добавления вновь выявленных признаков в тезаурус. Например, для метода SVM лучшая точность классификации на исходном множестве составляла 83%, в том время как для набора данных с уточненным тезаурусом она составила 89%.

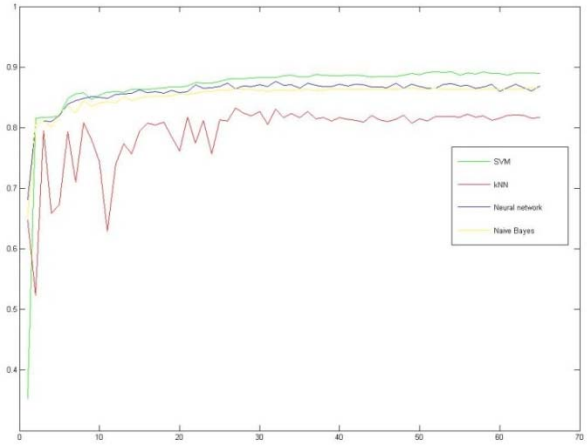

**Рис. 2. Качество классификации**

Мы также «обучали» новую тороидальную ESOM карту и карту MDS на наборе данных с уточненным тезаурусом. Результирующая ESOM карта показана на рисунке 3, а MDS карта на рисунке 4.

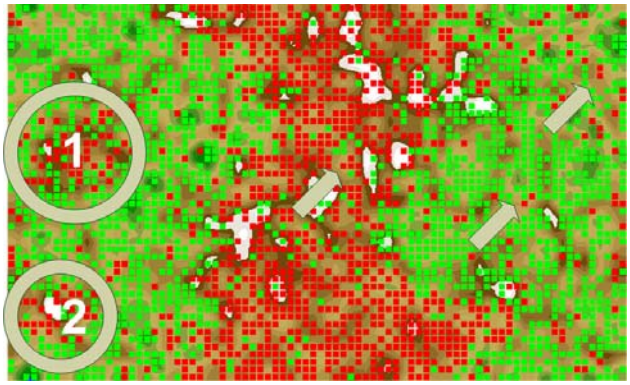


**Рис.3. Тороидальная ESOM карта для исходных данных**

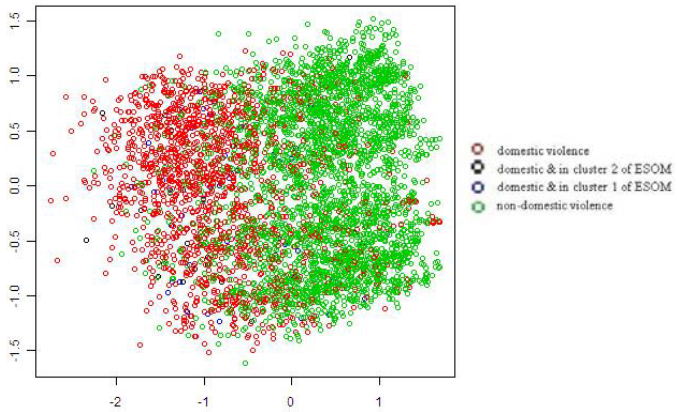


**Рис. 4. MDS карта для исходного набора данных**

Сравнение исходной ESOM карты с картой на рисунке 4 показало, что количество перекрытий между двумя классами было значительно уменьшено после того, как был введен улучшенный тезаурус. Карта на рисунке 4 показывает три различных кластера, которые в основном и содержат случаи домашнего насилия. После изучения случаев

содержащихся в левом верхнем кластере (окружность с 1 внутри), мы обнаружили, что этот кластер в основном содержит случаи краж со взломом, которые по неизвестным причинам были неверно помечены полицеийскими. В ходе анализа кластера расположенного слева внизу карты (окружность с 2) и нескольких выбросов (стрелки), мы обнаружили большое количество ситуаций, которые были оценены противоречиво полицейскими. Их мнения о том, как эти случаи должны быть оценены, отличались. Такие кластеры не были обнаружены в случае MDS карт. В итоге мы обнаружили, что выбросы в основном содержали случаи, которые были неверено помечены либо как домашнее, либо как «не домашнее» насилие. Мы пришли к выводу, что ESOM более пригодны в нашем случае для выявления знаний.

Хотя мы находим графический вид ESOM интуитивно понятным в использовании, количественный анализ выявил, что разделение данных MDS алгоритмом значительно лучше. Мы разработали kNN классификатор, основанный на ESOM и MDS картах для предсказания класса новых случаев, которые выявляет система предварительной сортировки. Точность классификации MDS для контрольной выборки (80,5%) была на 3% выше, чем для ESOM (77,5%). Эта вероятно вызвано тем, что ESOM алгоритм нашел два дополнительных кластера, помимо крупного в середние. Разделение этих данных в различные кластеры уменьшает точность классификации, но является полезным в предварительном анализе.

## 5. ВЫВОДЫ

В этой работе мы провели сравнение методов ESOM и MDS для анализа большого количества неструктурированных текстов. Мы показали достоинства этих средств для выявления новых признаков и для обнаружения противоречивых и неправильных случаев классификации. В отличие от ранее разработанных методов анализа данных для домашнего насилия, часто работающих как черный ящик без вмешательства пользователя, наша методология вовлекает эксперта предметной области в процесс поиска и позволяет понимать данные глубже. Ключевой момент состоит в инкрементальном характере пополнении знании эксперта о предметной области с помощью интуитивно понятного интерфейса навигации по данным. В то время как ранее эксперт предметной области был прегружен данными, сейчас наш метод позволяет интегрировать его знания в процесс исследования данных. Офицеры полиции, которые тестировали оба подхода, удовлетворены интерфейсом ESOM средств

и считают его более удобным для анализа большого количества полицейских отчетов, чем MDS. Более того ESOM распознал два важных дополнительных кластера данных, которые не были найдены MDS. Количественное сравнение показало превосходство MDS. Точность kNN классифиактора, построенного на основе MDS карты, на 3% выше, чем в случае ESOM. Опираясь на наши исследования, мы смогли улучшить точность классификации SVM до 89%. Тема дальнейших исследований – примение ESOM к другим типам криминальных случаев и построение системы для их классификации.

## Литература


1. Keus, R., Kruijff, M.S. (2000) Huiselijk geweld, draaiboek voor de aanpak. Directie Preventie, Jeugd en Sanctiebeleid van de Nederlandse justitie.
2. 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
3. Ultsch, A. (2004) Density Estimation and Visualization for Data containing Clusters of unknown Structure. In proc. GfKI 2004 Dortmund, pp 232-239
4. http://databionic-esom.sourceforge.net/
5. 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.
6. Fan, W., Wallace, L., Rich, S., Thang, T. (2006) Tapping the power of text mining. Communications of the ACM, Vol. 49, no. 9
7. http://www.politie-amsterdam-amstelland.nl/get.cfm?id=86
8. Poelmans, J., Van Hulle, M., Viaene, St., Elzinga, P., Dedene, G. (2011) Text mining with emergent self organizing maps and multi-dimensional scaling: A comparative study on domestic violence. Journal of Applied Soft Computing, 11(4), 3870–3876
9. Proc. Second IEEE Computational Systems Bioinformatics Conf., pp. 523-528, Aug. 2003.
10. Elzinga, P. (2006) Textmining by fingerprints. Onderzoeksrapport huiselijk geweld zaken. IGP project Activiteit 0504
11. Borg, I., Groenen, J.F. (2005) Modern multidimensional scaling: theory and applications. Springer series in statistics.
12. Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325–328.
13. 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.
14. 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.