Visual analysis techniques for medical diagnosis support

T. RZEŹNICZAK

tomek.rzezniczak@gmail.com

Military University of Technology, Faculty of Cybernetics
Institute of Computer and Information Systems
Kaliskiego Str. 2, 00-908 Warsaw, Poland

Taking into account vast amount of data available for general practitioners, medical diagnostic procedure can be treated as a complex analytical task. A doctor has to analyze the patient's symptoms, medical test results, and medical knowledge, correlate everything and decide on the diagnosis. In order to do this more effectively, dedicated analytical tools and techniques can be used. The paper elaborates on the application of Map of Attributes (MoA) visualization technique for analysis of a patient's health and disease pattern recognition. Various modes of using MoA are proposed and discussed. Furthermore, an application of diseases ranking preparation methods in visual filtering of diseases is presented. The methods use flexible similarity indices in conjunction with a graphical presentation of the Pareto model and Multidimensional Scaling model. Their goal is to allow physicians to narrow the space of detailed analysis in an interactive visual manner.

Keywords: data analysis, data visualization, medical decision support.

1. Introduction

The Map of Attributes (MoA) visualization technique supports visual analysis and recognition of objects by producing their visual representations in a form of regular polygons [16]. The technique's primary design goal was efficient use of human visual perception capabilities [15]. Besides the fact that the polygon's visual representation already brings abstract patterns into the level of visual cognition, additionally forced shape regularities improve the speed of human perception, therefore, recognition of objects is also supported. More precisely, MoA visualization produces a square matrix of points, where each point represents some attribute. The attributes are features describing objects from the visualized set (shared attributes are presented only once on the map). Then, a single object representation is constructed by drawing a polygon whose vertices are attributes belonging to the object.

This study presents possible methods of using MoA in the medical diagnosis process. In general, medical diagnosis is a procedure during which a medical practitioner tries to identify a disease which causes the symptoms of the patient [11], [21]. Since many of the symptoms have several causes, usually they don't discriminate diseases explicitly. Therefore, the doctor has to analyze the available information, correlate them and decide on the diagnosis. This procedure can be performed in several ways, among others, pattern recognition diagnosis [21]. It is based on experience. The doctor collects all the symptoms and using his knowledge of diseases, recognizes the pattern. Unfortunately, the procedure is not always so straightforward, very often a more in-depth analysis is needed. During the analysis, possible candidate diseases are verified and eliminated until a diagnosis can be made.

The paper proposes medical data presentation modes using MoA, as well as diseases filtering techniques that support effective usage of MoA or any other detail visual analytics.

2. Prerequisites

The idea of using MoA for medical patterns is to construct visual representation of disease entities patterns (DEPs) and allow a presentation of patients' state of health (PSH) on the same map. Thanks to the perceptual characteristics of such visual representations [15], they can improve comparison of particular diseases with PSH. This can bring a positive effect and facilitate the pattern recognition diagnostic process.

To construct MoA visualization, a repository of DEPs must be provided. Such an exemplary repository was built based on Mayo Clinic diseases and conditions website

resources [7]. The repository covers selected DEPs from three medical specialties: pulmonary medicine, cardiology and gastroenterology. Prepared DEPs are simplified compared to what is proposed in [1], [2] and covers only symptoms (no risk factors). A single symptom can occur in multiple diseases. In figures, the repository contains 78 diseases which are defined by 143 symptoms.

Finally, MoA visualization of the repository was constructed using a heuristic approach described in [17]. Visualization examples presented further are based on the best map that was found, so called TopMap.

Before any PSH analysis can be started, symptoms must be collected. A list of symptoms is a result of patient interview and examination. It can be stored only in the doctor's memory, written down on paper or entered into some kind of EMR (Electronic Medical Record) system [14]. Naturally, the prerequisite to use MoA visualization technique is a digitalized version of symptoms' list. Further in the paper, it is assumed that such a list is provided and reconciled with the DEPs' repository.

3. Presentation modes

Once the symptoms describing the patient's state of health are entered, the actual visual analysis process can be started. This subsection describes basic static presentation modes (no interaction), which are called: dot-pattern mode, full figure mode, partial figures mode.

3.1. Dot-pattern mode

The dot-pattern mode is the simplest mode of PSH presentation on MoA. Relevant symptoms are marked on the map without connecting them and forming a polygon. Fig. 1 presents four variants of PSH in dot-pattern mode: including full matching with Tetralogy of Fallot disease pattern and different levels of missing symptoms. Even though, this is a very basic mode, thanks to the regularity of patterns, it already brings benefits in disease identification. Experiments on perception of dot patterns confirm that humans remember regular patters better than irregular ones [5], [9]. Obviously, the easiness of identification will depend on the complexity of the pattern and the number of reported symptoms. According to French, patterns containing 6 to 8 dots are identified with the lowest level of errors. Taking into account the medical patterns frequency vs. the number of symptoms, over 42% of diseases in the repository is represented by 6 to 8 symptoms.

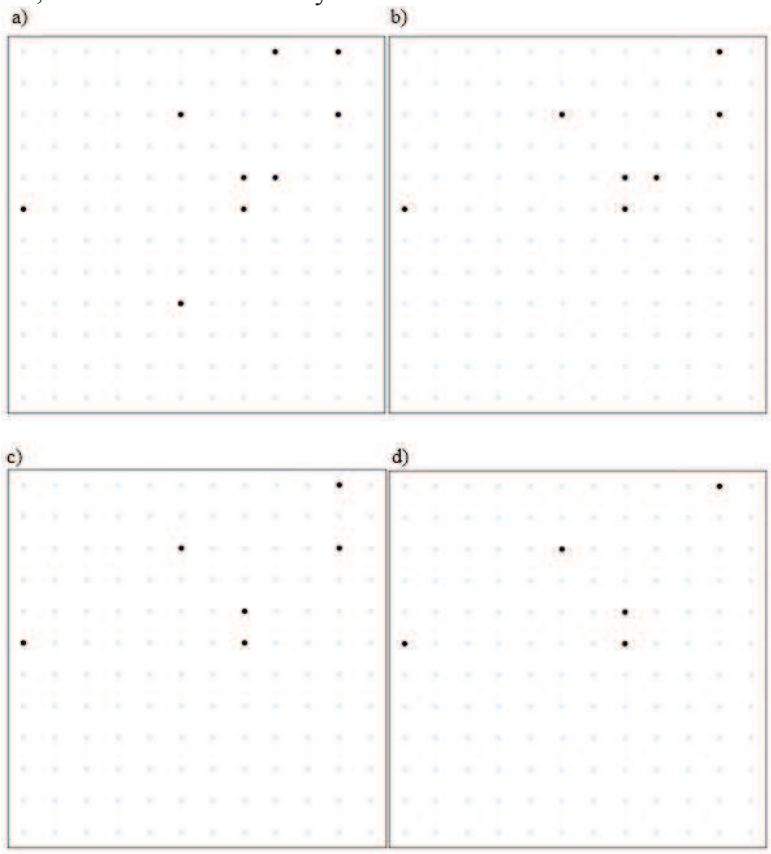

Fig. 1. Tetralogy of Fallot presented in dot-pattern mode with different levels of missing symptoms: a) all 9 symptoms of the DEP; b) 2 missing symptoms; c) 3 missing symptoms; d) 4 missing symptoms

3.2. Full figure mode

In full figure mode, instead of only marking PSH's symptoms on the map of attributes also figure contour is constructed and presented. The contour is built using the same methods as in the case of disease patterns' polygon construction using spatial contiguous condition [18]. Following this condition, the constructed polygon has the lowest circuit. Figure 2 shows an example of PSH representation using full figure mode with different levels of missing symptoms.

Additionally, since this is the same method as for DEP representation, it ensures the same contour shape in the case of exact symptoms matching between PSH and DEP. As one can expect, such an optimistic case is not something common. What should be expected in a real case is rather a partial matching to one or more DEPs, understood as a subset of symptoms that matches a subset of unmatched symptoms. The sources of unmatched symptoms can differ. First of all this can be symptoms that are not typical and are not included in the disease pattern but are still 'positive' symptoms and should not change the diagnosis. The second type are 'false' symptoms which are irrelevant for the particular disease being considered. Unmatched symptoms cause two types of distortion: in the case of 'positive' symptoms it hinders recognition of the disease pattern which is undesirable, while in the case of 'false' symptoms this is an expected behavior.

Another type of distortion originates from a small number of reported symptoms compared to the number of symptoms included in the sought disease pattern. This can result in a figure whose level of distortion can significantly limit recognition. There is no evidence to which level of distortion the disease pattern can still be recognized. On the other hand, a small number of symptoms and no clear match may indicate that additional examination is needed to identify the disease.

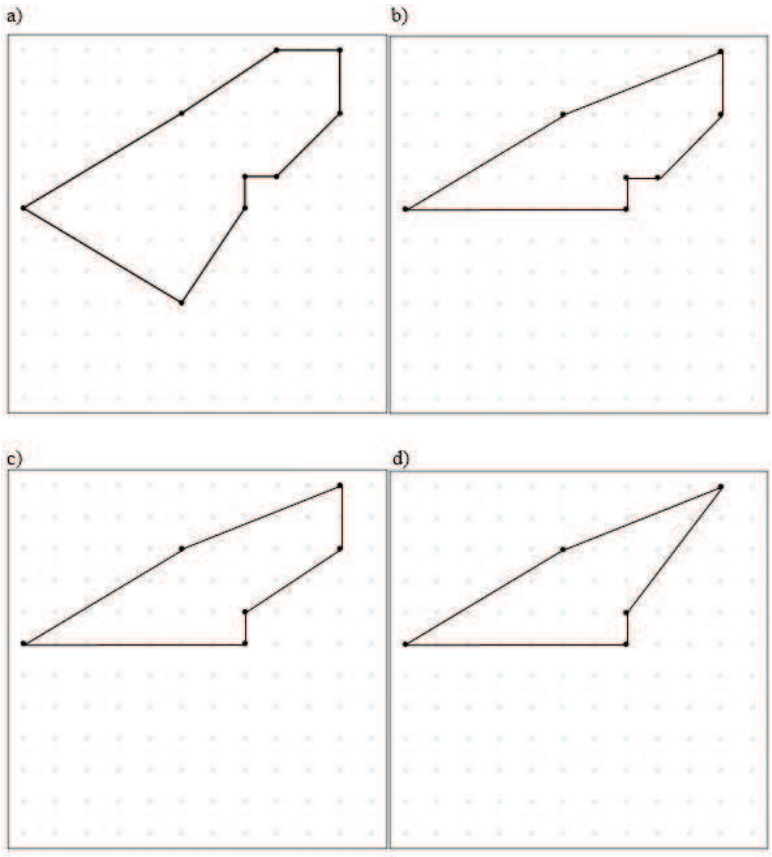

Fig. 2. Tetralogy of Fallot presented in full figure mode with different levels of missing symptoms:
a) all 9 symptoms of disease pattern; b) 2 missing symptoms;
c) 3 missing symptoms; d) 4 missing symptoms

3.3. Partial figures mode

Last of the static visualization modes is partial figures mode. In this case, a set of visualizations is generated, where each represents a partial figure of a certain DEP. A partial figure is a not fully connected polygon, only a set of polygon sides that reflects common parts of the DEP and PSH. Compared to the full figure mode, the main difference is related to the way in which possible distortions are handled (see previous section). In full figure mode, some unknown/ artificial patterns could be constructed, whereas, in the partial figures mode several maps are plotted, one for each of the diseases that fulfill certain selection criteria. The selection criteria should be related to some sort of similarity measure between PSH and DEPs. The most obvious would be presentation of partial figures for all the diseases that share at least two of patient's symptoms that are two vertices of the same side of a disease pattern (a more sophisticated model of the criteria will be elaborated in section 4). An example of a partial figure mode is presented in Figure 3, where a set of hypothetical symptoms fulfills the basic selection criteria for several DEPs. Additionally, Fig. 4 shows the same examples in an extended partial figure mode. The extended mode also displays symptoms and polygons' sides which are missed to matching the pattern fully.

It was proven that even in the case of partial or occluded view, an observer is still able to recognize an object surprisingly easily [8]. This phenomenon is called perceptual closure. Human perception is capable of 'filling-in' the missing information. It is strongly related to perceptual learning and occurs when the observer was exposed to the object in the past. According to Doniger et al. even for the proportion of deleted segments of a line drawing reaching 88% the observer is able to recognize correctly the drawing with 41% of mean accuracy, when the drawing was repeated (seen earlier by the observer) [8]. While at the level of 50% and 30% of deleted segments the cumulative proportion of pictures recognized rises to 95% and close to 100% respectively.

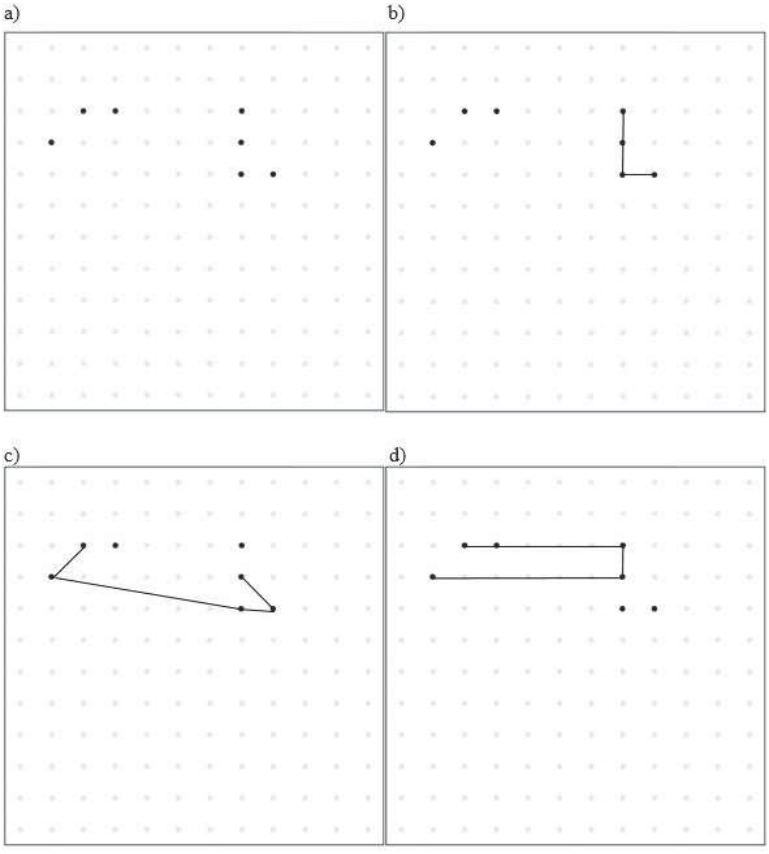

Fig. 3. A set of hypothetic symptoms presented in partial figure mode:
a) original symptoms' set; b) part of matched Asbestosis pattern;
c) part of matched Esophageal Cancer pattern;
d) part of matched Gastroesophageal reflux pattern

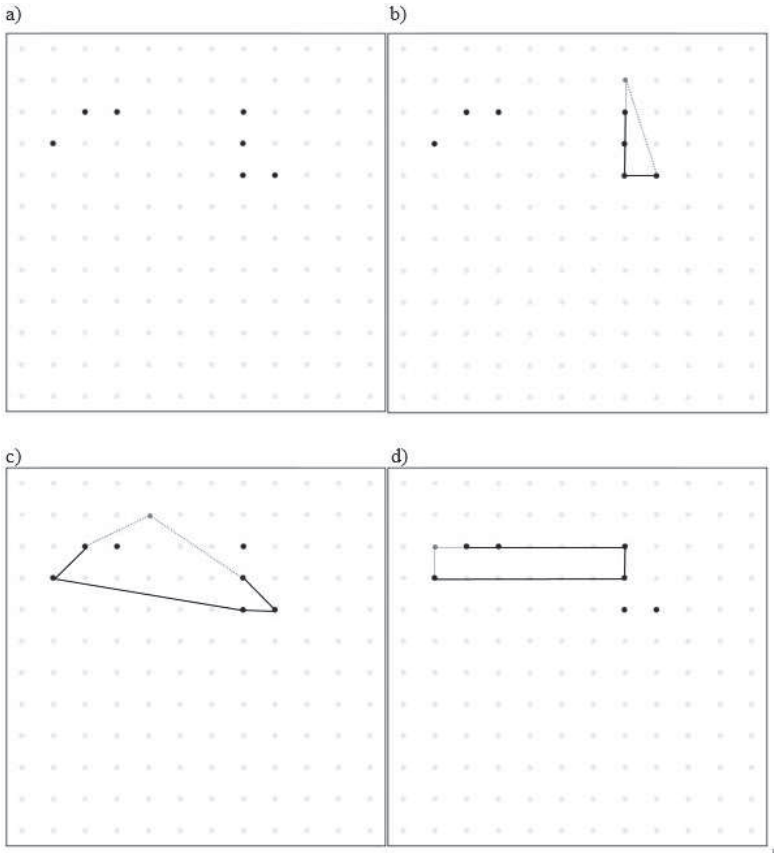

Fig. 4. A set of hypothetic symptoms presented in extended partial figure mode:
a) original symptoms' set; b) part of matched Asbestosis pattern;
c) part of matched Esophageal Cancer pattern;
d) part of matched Gastroesophageal reflux pattern

4. Patterns filtration

Recalling the discussion on the partial figures mode, one of the important aspects was the selection criteria for patterns that are visualized. Criteria that filtrate diseases and limit the space of further detailed analysis are essential. DEPs for analysis can be selected explicitly by a doctor, but the preferred method would be to narrow the set of disease patterns by applying some similarity metric and creating a ranking of DEPs. This section presents a proposition of the ranking construction methods.

4.1. Contrast model

Similarity between objects can be defined as the mental representation proximity of these objects. Tversky proposed a flexible model of similarity, defined by objects' characteristics, where similarity between object a and b is expressed by function of common and distinctive features [19], [20]. Denoting by A set of features of object a and the B set of features of object b, the similarity is a function of three arguments, measuring the level that two sets of features fit together:

- $A \cap B$ – common features of a and b;
- $A - B$ – features of a not occurring in b;
- $B - A$ – features of b not occurring in a.

Tversky defined an interval similarity scale $s(a,b)$ (contrast model) expressed as a linear combination of measures of common and distinctive features:

$$s(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A) \quad (1)$$

where: $\theta, \alpha, \beta \geq 0$, and f is a function (usually a weighted sum) representing contribution of different features of objects in their similarity.

This model does not define a unique index of similarity, but their family, as defined by parameters: θ, α, β, thereby allowing the introduction of various relations of similarity between the same objects, such as:

- if $\theta = 1, \alpha = \beta = 0$, then $s(a,b) = f(A \cap B)$ and it is symmetric with the interest focused on similarities;
- if $\theta = 0, \propto = 1, \ \beta = 0.5$, then $-s(a,b) = f(A - B) + 0.5 * f(B - A)$ and it is symmetric with the interest focused on differences.

The contrast model can be applied as a similarity measure for building diseases patterns ranking in terms of their similarity with PSH [3]. It fits very well into our representation of patterns, where each object is described by a set of attributes. Its elasticity allows to define different preferences of a physician. Assuming that f is a cardinality of sets, a represents patient's condition and b is a disease pattern, the general preferences in diagnosing procedure are [3]:

- Pref1: $f(A \cap B) \rightarrow max$;
- Pref2: $f(B - A) \rightarrow min$;
- Pref3: $f(A - B) \rightarrow min$, although this is not obvious considering that the patient may suffer from several diseases.

These preferences can be modified while the ranking is built by adjustment of θ, α, β parameters. If the doctor wants to verify different hypotheses, like focusing more on common symptoms or conversely emphasize the differences during analysis of dissimilarity, it can be achieved by modifying the parameters. For instance rising the: β and minimizing α, means that only DEP is important in dissimilarity assessment (neglecting any symptoms that can be present because of coexisting diseases).

The contrast model was a basis for a normalized similarity index called ratio model bounded between 0 and 1:

$$ratio(a,b) = \frac{\theta f(A \cap B)}{\theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)} \qquad (2)$$

Actually, setting $\theta = 1, \alpha = \beta = -1$ produces Jaccard index, if $\theta = 1, \alpha = \beta = -0.5$ it is a Dice index [6].

4.2. Pareto model

The contrast model can also be seen as a multi-criteria similarity model and considered using Pareto similarity relation, like in [4]. The criteria are equal to general preferences in diagnosis procedure discussed in the previous section. To produce a ranking of diseases one has to deal with trade-off between different criteria of the model. It can be facilitated using Pareto-optimality known from multi objective optimization. Let $\{f_1(x), f_2(x), \dots, f_k(x)\}$ represent a set of criteria functions, where $x \in X$ is a vector of decision variables from a possible decisions set X. If the optimization problem is to minimize the objective function

$$f(x) = (f_1(x), f_2(x), \dots, f_k(x)) \qquad (3)$$

then the Pareto-optimal solution is defined as $x^* \in X$, if there is no $x \in X$ such that $f_i(x) \leq f_i(x^*)$for all $i = 1,2, \dots, k$ and if $f_j(x) < f_j(x^*)$ for at least one index j (in the case of maximization of some of the criteria it is equal to $-f_i$) [13]. All the Pareto-optimal solutions form a Pareto-optimal front, which can be visualized very easily for two criteria and it is a little more difficult for three criteria.

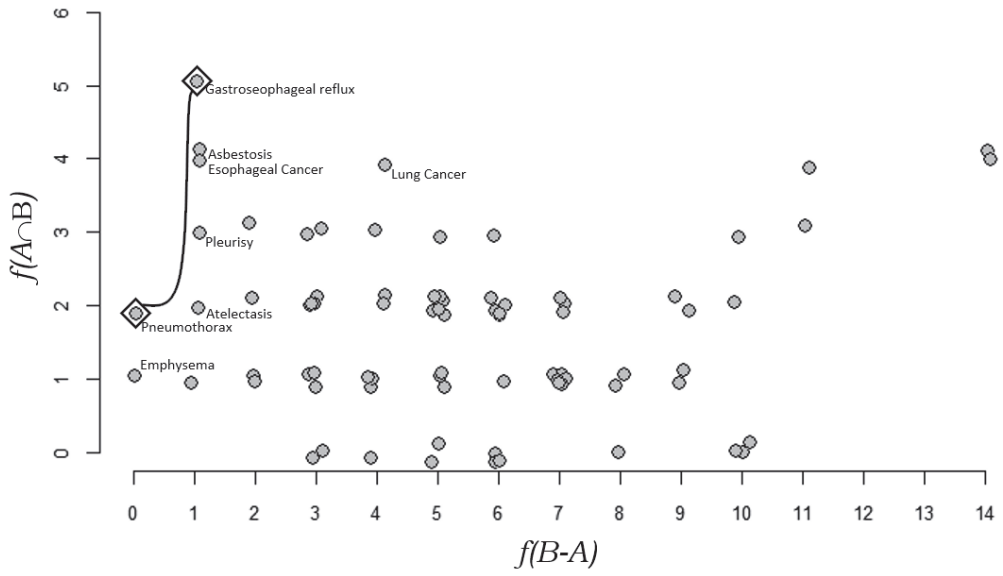

Fig. 5. Visualization of Pareto-optimal front (the marked points) based on the example repository of DEPs and hypothetical PSH (the same PSH as used in Fig. 3). The preferences are: $f(A \cap B) \rightarrow max$, $f(B - A) \rightarrow min$. For clarity of presentation the diseases positions are jittered and only some of the labels are showed

The Pareto-optimal front can be used for disease patterns ranking and visualization of the detection space [1], [3], [4]. The detection space in our study spans on two or three dimensions corresponding to the preferences: Pref1, Pref2, Pref3. An example visualization of the Pareto-optimal front for a virtual PSH and the example repository of DEPs is presented in Figure 5. The graphical representation (with marked front) can be used for interactive filtering of diseases for further investigation using MoA or another visualization technique.

4.3. Multidimensional Scaling model

Another method of building interactive visual filtering is to use geometric model of similarity relation. This type of a model represents an object as a point in space (usually Euclidean space), and distance between points corresponds to similarity of objects. An example of such a model is MDS (Multidimensional Scaling) [10]. Input data for the model are similarities/differences between all objects under consideration. The result is a geometric model representing objects as points in an n-dimensional space.

Formally, MDS can be described as follows: let k be the number of all objects under consideration and n is the number of attributes of each object. Matrix X with a dimension of $k \times n$ will contain spatial coordinates of the objects, where row i indicates the coordinates of object i. While, the difference between objects i and j will be described by δ_{ij}. The distance in the Euclidean space between objects i and j is defined as the shortest line connecting i with j and takes on the form:

$$d_{ij}(X) = \left(\sum_{s=1}^{n}(x_{is} - x_{js})^2\right)^{1/2} \quad (4)$$

The purpose of MDS is to find such a matrix X that $d_{ij}(X)$ corresponds to δ_{ij}. This assumption can be presented in various forms, including in the least squares MDS model proposed by Kruskal [12]:

$$\sigma^2(X) = \sum_{i=2}^{k}\sum_{j=1}^{i-1} w_{ij}(\delta_{ij} - d_{ij}(X))^2 \quad (5)$$

where w_{ij} is a non-negative weight. For example, many MDS implementations take $w_{ij} = 0$ for the missing differences.

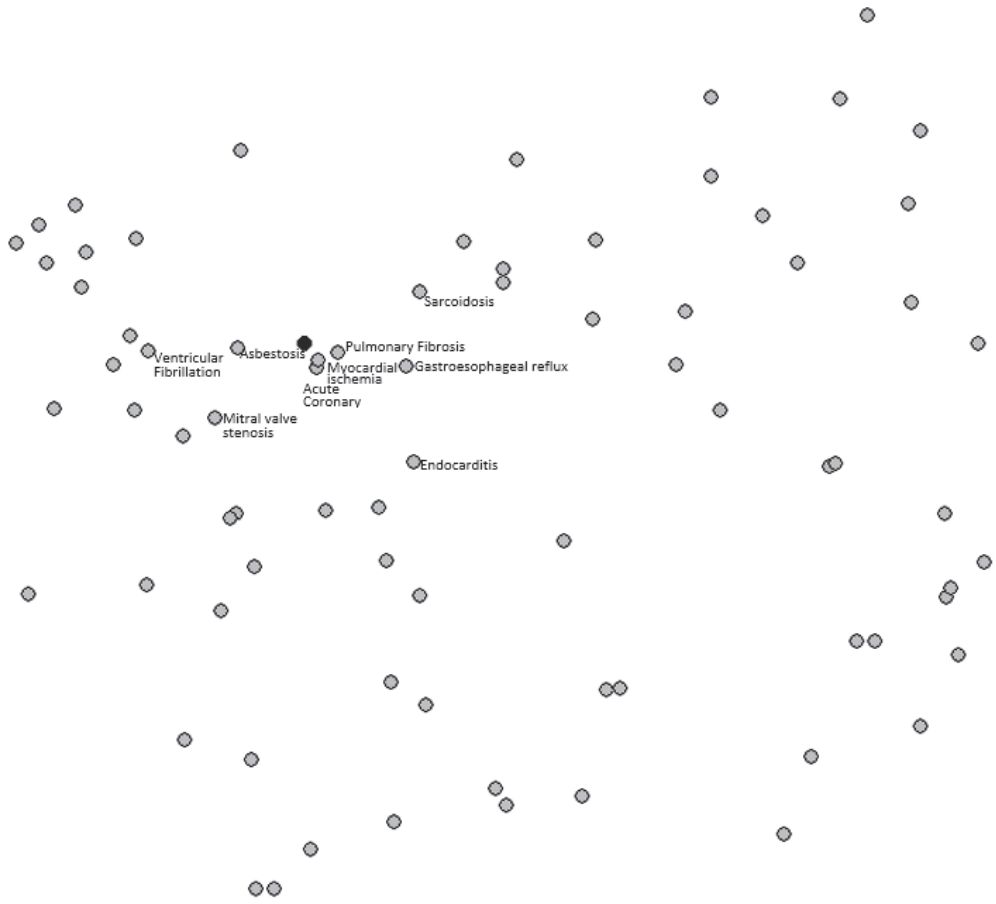

Fig. 6. Visualization of MDS based on the example repository of DEPs and hypothetical PSH (black point – the PSH as used in Figures 6.3). The distance is calculated using Jaccard index. For clarity of the presentation only some of the labels are showed

Application of MDS to medical patterns domain assumes calculation of distances δ_{ij}, for all objects from the repository including the PSH that is analyzed. The calculation can be done using one of described earlier similarity indices derived from the contrast model. Next, if the number of dimension n is set to 2, than the matrix X can be represented on a scatter plot. Such a plot can be used as a visual filter, for example filtering through selecting the nearest neighbors (the most similar diseases patterns according to the selected index) of the dot symbolizing the patient's condition. Figure 6 presents an example of MDS based on Jaccard index distance applied to the example repository of DEPs and hypothetical description of PSH.

4.4. Simultaneous analysis of multiple diseases patterns

Whatever filtration method is chosen, a doctor should be able to simultaneously analyze all the chosen DEPs and compare them with the PSH. Presentation of different diseases is basically embedded in the partial figures mode, while the two other modes: dot-pattern and full figure, can be easily extended to include such a feature. The possible techniques are:

- **single map** – DEPs are plotted on the same map with PSH;
- **multiple maps** – each DEP is plotted on a separate map next to the map with PSH.

Furthermore, DEPs' displaying techniques can be matched with the mode selected for PSH presentation, but there is no impediment to mix them and draw a full figure together with a dot-pattern.

5. Summary

The idea of medical visual analysis presented in this paper is to give physicians an efficient mechanism that facilitates interactive study of patient's data, especially their comparison to DEPs. Therefore, in the first part possible basic methods of using Map of Attributes visualization technique were presented, this includes: dot-pattern mode, full figure mode, partial figures mode. They were supported by reference to scientific research on human perception; however, this can be treated as a first step, in order to decide which mode is preferred in which circumstances, conducting an empirical study would be necessary.

The second part of the paper discussed visual filtration methods that can narrow the space of diseases for detailed analysis. The presented methods are based on flexible similarity indices that can be adjusted by doctor's preferences. The indices are then used to prepare diseases visual filtration using either Pareto model or MDS model. The methods are agnostic to the actual type of further analysis, therefore, their application is not limited to use with the MoA visualization technique.

6. Bibliography

[1] Ameljańczyk A., "Matematyczny model wspomagania diagnozowania medycznego na podstawie symptomów chorobowych i czynników ryzyka", *Bio-Algorithms and Med-Systems*, Supplement No. 1, Vol. 6, No. 12, 33 (2010).

[2] Ameljańczyk A., "Wielokryterialne mechanizmy wspomagania podejmowania decyzji medycznych w modelu repozytorium w oparciu o wzorce", *Biuletyn Instytutu Systemów Informatycznych*, Nr 5, 1–6 (2010).

[3] Ameljańczyk A., "Multicriteria similarity models for medical diagnostic support algorithms", *Bio-Algorithms and Med--Systems* , 9(1), 1–7 (2013).

[4] Ameljańczyk A., "Teoretyczne aspekty badania podobieństwa obiektów w problematyce rozpoznawania wzorców", w: *Problemy modelowania i projektowania opartych na wiedzy systemów informatycznych na potrzeby bezpieczeństwa narodowego*, pod redakcją naukową Tadeusza Nowickiego, Zbigniewa Tarapaty, Wojskowa Akademia Techniczna, Warszawa, 9–22 (2014).

[5] Attneave F., "Symmetry, information, and memory for patterns", *The American Journal of Psychology*, 209–222 (1955).

[6] Daylight, Chemical Information System Inc., http://www.daylight.com/dayhtml/doc/theory/theory.finger.html, accessed November 2015.

[7] Disease and Conditions, Mayo Clinic, http://www.mayoclinic.org/diseases-conditions, accessed November 2015.

[8] Doniger G.M., Foxe J.J., Schroeder C.E., Murray M.M., Higgins B.A., Javitt D.C., "Visual perceptual learning in human object recognition areas: a repetition priming study using high-density electrical mapping", *NeuroImage*, 13(2), 305–313 (2001).

[9] French R. S., "Identification of dot patterns from memory as a function of complexity",

Journal of Experimental Psychology, 47(1), 22-26 (1954).
[10] Groenen F., Patrick J., Velden M., *Multidimensional scaling*, John Wiley & Sons Ltd., 2005.
[11] Groopman J.E., Prichard M. J., *How doctors think*, Vol. 82, Boston: Houghton Mifflin, 2007.
[12] Kruskal J.B., "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", *Psychometrika*, 29(1), 1–27 (1964).
[13] Madetoja E., Ruotsalainen H., Monkkonen V.M., Hamalainen J., Deb K., "Visualizing multi-dimensional Pareto-optimal fronts with a 3D virtual reality system", in: *Proceedings of the International Multiconference on Computer Science and Information Technology*, IMCSIT 2008, Wisła, Poland, 20–22 October 2008, 907–913, IEEE 2008.
[14] Rudowski R., *Informatyka medyczna*, Wydawnicto Naukowe PWN, 2003.
[15] Rzeźniczak T., "Implementation aspects of data visualization based on map of attributes", *Journal of Theoretical and Applied Computer Science*, Vol. 6(4), 24–36 (2012).
[16] Rzeźniczak T., "Applying the Concept of Figural Goodness for Automatic Design of Dataset Visualization", *Biuletyn Instytutu Systemów Informatycznych*, Nr 11, 23–31 (2013).
[17] Rzeźniczak T., "Evolutionary algorithms for Map of Attributes optimization", *Computer Science and Mathematical Modelling*, No. 1, X–X (2015).
[18] van Tuijl H.F., "Perceptual interpretation of complex line patterns", *Journal of Experimental Psychology*: *Human Perception and Performance*, 6(2), 197-221 (1980).
[19] Tversky A., "Features of Similarity", *Psychological Review*, American Psychological Association, Vol. 84, 4, 327–352 (1977).
[20] Tversky A., Gati I., "Studies of similarity", *Cognition and categorization* 1, 79–98 (1978).
[21] Wulff H.R., *Racjonalna diagnoza i leczenie*, Aktis, 2006.

Techniki analizy wizualnej dla wsparcia diagnostyki medycznej

T. RZEŹNICZAK

Biorąc pod uwagę ilość danych dostępnych dla lekarzy, diagnostyka medyczna może być traktowana jako złożone zadanie analityczne. Lekarz musi przeanalizować symptomy pacjenta, wyniki jego badań oraz wiedzę medyczną, a następnie skorelować wszystko i zdecydować o diagnozie. W celu przeprowadzenia tego efektywnie, można zastosować dedykowane narzędzia i techniki analityczne. Artykuł omawia zastosowanie techniki wizualizacji Mapa Atrybutów – MoA (ang. *Maps of Attributes*) do analizy stanu zdrowia pacjenta oraz rozpoznawania wzorców jednostek chorobowych. W artykule przedstawione i przedyskutowane zostały różne możliwe tryby użycia MoA. Ponadto, zaprezentowane jest zastosowanie metod budowy rankingu jednostek chorobowych do wizualnego filtrowania chorób. Metody te wykorzystują elastyczne indeksy podobieństwa w połączeniu z graficzną prezentacją modelu Pareto oraz modelu MDS (ang. *Multidimensional Scaling*). Ich celem jest umożliwienie lekarzowi zawężania przestrzeni szczegółowej analizy w sposób wizualny i interaktywny.

Słowa kluczowe: analiza danych, wizualizacja danych, wspomaganie decyzji medycznych.