Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 2179–2188

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Clustering and visualization of a high-dimensional diabetes dataset

Piotr Lasek[a]*, Zhen Mei[b]

[a]*University of Rzeszow, Poland,* [b]*Manifold Data Mining Inc., Canada*

Abstract

Data clustering algorithms have proved to be important and widely used methods of artificial intelligence and data mining for discovering unknown yet important patterns in datasets. Nevertheless, one of the additional aspects of data clustering is proper interpretation of the clustering results. In this paper we aim to investigate possibilities of using both data clustering and visualization methods to analyze a sample diabetes dataset. In the first part, we focus on how to cluster a highly-dimensional sample dataset and then, we concentrate on how to properly visually present the clustering results in the most meaningful way to uncover potentially interesting behavioral patterns or features of diabetes patients. In this work we examine two clustering algorithms (DBSCAN, k-Means) along with several different distance measures. We also present sample visualizations of clustering results generated by an application which we have developed and discuss if the proposed way of clustering results visualization can be helpful in understanding the analyzed dataset and lead a viewer to drawing valuable conclusions about it.

Keywords: data mining, data visualization, interactive data visualization, diabetes, visualization, clustering, high-dimensional clustering

1. Introduction

Chronic diseases have undoubtedly become an important aspect of today's societies [4, 7, 8, 14, 15, 16, 19] and there is still a lot to be done to improve the ways how patients are treated and even more to better understand how the

* Corresponding author. Tel.: +4-817-851-8531; fax: +4-817-851-8525.
E-mail address: lasek@ur.edu.pl

1877-0509
10.1016/j.procs.2019.09.392

diseases spread and what the most important decisive factors in those diseases are. For example, in Canada, there are eleven million people with one or more chronic conditions (heart diseases, diabetes, asthma, arthritis, cancer, depression, etc). Those patients are not only responsible for the great majority of the total health care cost but also require more physician visits, hospital admissions, total hospital days, prescriptions, and home care visits than an average patient.

Obviously, a number of methods for analyzing and discovering unknown yet interesting patterns in medical datasets is known and new ones are continuously proposed [20, 21]. Nevertheless, complexity and larger sizes of datasets require developing more interactive and easy-to-understand ways of presenting the results of algorithms to analysts. Interactive exploration and visualization become a must in the today's world and the statement that one picture is worth a thousand words is nowadays even more up to date. Interactive data visualization is an important bridge connecting data with people. It can not only dramatically (if done correctly) improve the possibilities of understanding data, it can actually turn data into knowledge. It has been even suggested recently [1] that scientists might be better off analyzing images than of actual data.

Data visualization has a long history behind [22] and can be applied to any step of data analysis. Nevertheless, for the purpose of this work we will focus on applying visualization techniques for meaningful presentation of the results of clustering high-dimensional datasets. The main goal of this work is to determine if by means of clustering and proper visualization it is possible to discover behavioral patterns or features of diabetes patients. So, we aim at finding ways of presenting the results of clustering so it can let viewers to easily understand the clustering results and draw valuable conclusions about the dataset.

Organization of the paper is as follows. In Section 2 we remind two of the most popular clustering algorithms: a density-based clustering algorithm DBSCAN [3] and a well-known k-Means [9]. In Section 3 we describe several clustering and visualization experiments which led us to implementation of an application for visualization of high-dimensional (with over 1200 attributes) dataset. We summarize the results, conclude the paper and discuss further steps in the final section.

2. Data clustering and visualization

2.1. Data clustering

Clustering data into meaningful groups has always been an important task of both artificial intelligence and data mining. Clustering is considered as an unsupervised classification of data where the results of the task depend on the algorithm used. A number of different clustering algorithms have been offered in the literature over time. Some of them are capable of discovering proper clustering of data only when the number of clusters is known in advance. Other algorithms can discover clusters of particular shapes only. There are also algorithms that are able to identify noise data. In this subsection we remind two commonly used algorithms, namely: DBSCAN and k-Means.

DBSCAN. The main feature of this algorithm is that each point of a cluster must contain at least a certain number of points (*MinPts*) within its ε-neighborhood (*Eps*). In other words, the density in the ε-neighborhood of a point belonging to a cluster must be greater or equal to a predefined threshold. The clustering process in DBSCAN is based on the following concepts of relations between points: *directly density-reachability* and *density-reachability* (please refer to [3] for the definitions and detailed descriptions of those concepts). DBSCAN discerns three types of points: *core points*, *border points* and *noise points*. A cluster in the context of the DBSCAN algorithm is a region of high density. Regions of low density constitute *noise*. A point in space is considered a member of a cluster if there is a sufficient number of points within a given distance from it. Firstly, the algorithm generates a label for the first cluster to be found. Next, the points from the dataset are processed. The initial value of the *ClusterId* attribute of the first point read is equal to UNCLASSIFIED. While the algorithm analyzes point after point, it may occur that the *ClusterId* attributes of some points may change before these points are actually analyzed. Such a case may occur when a point is *density-reachable* from a *core point* examined earlier. Such density-reachable points will be assigned to the cluster of a *core point* and will not be analyzed later. If a currently analyzed point has not been classified yet (the value of its *ClusterId* attribute is equal to UNCLASSIFIED), then the *ExpandCluster* function is called for this point. If the point

is a *core point*, then all points within its ε-neighborhood are assigned by the *ExpandCluster* function to the cluster with a label equal to the current cluster's label. Next, a new cluster label is generated. Otherwise, if the point is not a *core point*, the attribute *ClusterId* of point p is set to NOISE, which means that the point will be tentatively treated as *noise*. After analyzing all points in the dataset, each point's attribute *ClusterId* will store a respective cluster's label or its value will be equal to NOISE.

K-Means. The k-Means algorithm is the simplest and most know representative of a group of minimum-variance or partition algorithms for which the goal of the algorithm is to minimize the sum of squared error criterion function [11] and the number of clusters is known. Partitioning clustering algorithms produce single data partitions instead of creating a structure such as *dendrogram* created by hierarchical clustering algorithms. A major problem related to the fact that a user must pre-set a number of clusters to determine with partitioning algorithms is actually a proper selection an appropriate number of output clusters [5]. In k-Means algorithm, each cluster is represented by the gravity center of the cluster, so called *centroid*. Analogously, in a similar partitioning algorithm, k-Medoids [6], each cluster is represented by its center point belonging to a dataset called *medoid*. Another example of the partitioning clustering algorithms is an improved version of k-Medoids - CLARANS [2].

To process a dataset, the k-Means algorithm starts with a first group of *k* randomly selected *centroids*. Those *centroids* are used as initial center points for every cluster. Then, the algorithm, iteratively performs calculations in order to optimize the locations of centroids. The calculations are stopped when either the centroids locations have stabilized (this means that their coordinates do not change between iterations anymore; in other words, the clustering has been successful) or the pre-defined number of iterations has been achieved.

2.2. Data visualization

Interactive data visualization, as the Gartner's report confirms [13] is an important need of today's data exploration and discovery tools. The process of data visualization is often described in the literature by the Schneiderman's famous mantra: *Overview first, zoom and filter, details on demand* [26]. This means that the analyst is primarily concerned with obtaining a general visual description of the analyzed data set. In the next step, he looks for specific and interesting details. While in the case of small data sets, visualization of them is practically not a problem, then in the case of large datasets or data with a large number of dimensions you can no longer assume that you only need to display the original set of data on the screen - the number of objects can be just so large that it will exceed the number of available pixels by orders of magnitude [10]. In other words, the data has become too large to be able to see it directly. However, data visualization clearly becomes a final yet a critical step in extracting knowledge from data which is confirmed by recent calls to action [24].

Nevertheless, some of the data visualization challenges have already been and still are being addressed [22]. For example, if we are focusing on how to efficiently visualize data we can use techniques of data reduction such as summarizing (e.g. employing data clustering), condensing (using so-called binned aggregation) or bounding (visualization by the number of pixels used to display it, e.g. *nanocubes*). Similarly, we can try to reduce data to be visualized before it is even sent to a visualization client. Here, filtering, aggregation and sampling come in useful.

2.3. Data presentation challenges

In this work we are employing three of the above-mentioned methods of dealing with large amounts of data objects to visualize, namely: clustering (for summarization of data points), aggregation (for generating descriptions of data clusters) and visual condensing (for generating more meaningful visualizations).

Clustering. For clustering we employ the k-Means and DBSCAN algorithms to label points with cluster ids. We do also examine several different distance measures (Euclidean, Manhattan, Tanimoto). Discovered clusters are then aggregated and condensed to generate a meaningful visual representation of a dataset.

Aggregation. We aggregate points assigned to clusters by calculating attribute-value descriptors representing how often specific values of attributes occur in a cluster. This is later drawn in a visual form of bars where values are represented by lines of length corresponding to how frequently they occur in a cluster.

id	geogprv	geodpmf	geodbcha	adm_prx	adm_n09
1	BRITISH COLUMBIA	NORTH SHORE/C. G	VANCOUVER	NO	N/A
2	QUEBEC	RÉG. MAURICIE	N/A	NO	IN PERSON
3	BRITISH COLUMBIA	NORTH SHORE/C. G	VANCOUVER	NO	N/A
4	MANITOBA	WINNIPEG RHA	N/A	NO	IN PERSON
5	NEW BRUNSWICK	ZONES 6 AND 7	N/A	NO	IN PERSON
6	ONTARIO	CITY OF OTTAWA H	N/A	YES	N/A
7	YUKON/NWT/NUNA.	YUKON/NORTHWEST	N/A	NO	ON TELEPHONE
8	ALBERTA	CALGARY ZONE	N/A	NO	IN PERSON
9	PEI	PEI	N/A	NO	N/A
10	ONTARIO	SIMCOE MUSK. H	N/A	NO	IN PERSON
11	ALBERTA	SOUTH ZONE	N/A	NO	N/A
12	QUEBEC	RÉG. ABITIBI	N/A	NO	N/A
13	ONTARIO	SIMCOE MUSK. H	N/A	NO	N/A
14	SASKATCHEWAN	SASKATOON	N/A	NO	N/A
15	ONTARIO	BRANT COUNT	N/A	YES	N/A
16	ONTARIO	SIMCOE MUSK. H	N/A	NO	N/A
17	SASKATCHEWAN	SUNRISE/KELSEY	N/A	NO	N/A
18	ONTARIO	RENFREW COUNTY H	N/A	NO	N/A
19	QUEBEC	RÉG. MAURICIE	N/A	NO	ON TELEPHONE
20	BRITISH COLUMBIA	NORTHERN INTER.	NORTHEN	NO	IN PERSON

Fig. 1. A sample extract from the original dataset used for the experiments.

Condensing. We plot the results of clustering and aggregation using a special form of a bar-chart where bars represent frequency of values of attributes. This approach allows us to efficiently represent a large number of attributes showing at the same time aggregated distribution of values of a given attribute within a cluster. Additionally, we can plot several clusters in the same figure so that one can easily visually compare the clusters. Such a way of visualization makes possible to quickly determine substantial or outlying attributes defining a cluster. Based on this an analyst can draw conclusions and gain knowledge about the dataset's objects.

3. Experiments

In this section we start with a description of the dataset we have used for our experiments. Then, we describe three groups of experiments that we have conducted. First, we tried to find a way of showing the results of clustering of a dataset with over 1200 dimensions. Then, by means of clustering, we tried to confirm a hypothesis concerned with a manual way of segmenting the dataset. Finally, we present a method employing DBSCAN (with the Tanimoto distance measure) and our visualization system for presenting and analyzing high-dimensional results of clustering.

3.1. The dataset

The original dataset we use contains 65,000 of records from a 2010 Canadian Community Health survey[†]. The dataset includes information on various aspects of people's life such as lifestyle factors and socio-demographics, attitudes, stress level, satisfaction, exercise, diet and smoking (Figure 1). The attributes were of different types: e.g. numerical, nominal, intervals. In some cases, we focused on analyzing a subset of the original dataset containing about 17 000 object representing different patients with diabetes and heart disease. In this case each patient was described using about 30 – 200 attributes, both numerical, nominal or ordinal. The main goal of the analysis was to try to find groups of diabetes patients and determine what the main factors and attributes characterizing them were.

3.2. Mapping all attributes to numerical values

In our initial experiment we tried to overcome the issue concerned with different types of attributes. We prepared a procedure for mapping all types of attributes to numerical values from the interval [0, 10]. For example, Boolean values like 'Yes' and 'No' were mapped to integer values such as 10 and 0 respectively. For nominal values, such as

[†] http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=81424

for example ‘In person’, ‘On telephone’, ‘N/A’ we to assigned integer values which would be as adequate as possible. In this case, the mapped values were 10, 5 and 0. Further, for intervals we assigned values also based on our preliminary analysis, e.g. to an age interval ‘0 – 10 years old’ we assigned 0, to ‘11 – 20 years old’ we assigned 1, to ‘21 – 30 years old’ we assigned 2, etc.

Attribute group	Attribute	Description
ADL	ADL_01	Needs help - preparing meals
ADL	ADL_02	Needs help - getting to appointments
ADL	ADL_03	Needs help - doing housework
ADL	ADL_04	Needs help - personal care
ADL	ADL_05	Needs help - moving about inside house
ADL	ADL_06	Needs help - looking after finances
ADL	ADLF6R	Help needed for tasks - (F)
ADM	ADM_N09	Interview by telephone/in person
ADM	ADM_N10	Respondent alone during interview
ADM	ADM_N11	Answers aff./presence of another person
ADM	ADM_PRX	Health Component completed by proxy
ADM	ADM_RNO	Sequential record number

Fig. 2. Two sample groups of attributes. First (ADL) is a group of attributes representing different types of help needed (preparing means, getting to appointments, doing housework, personal care, etc.), the latter is a group corresponding to a way how the survey was filled in (by a respondent himself or with a help of other people or systems).

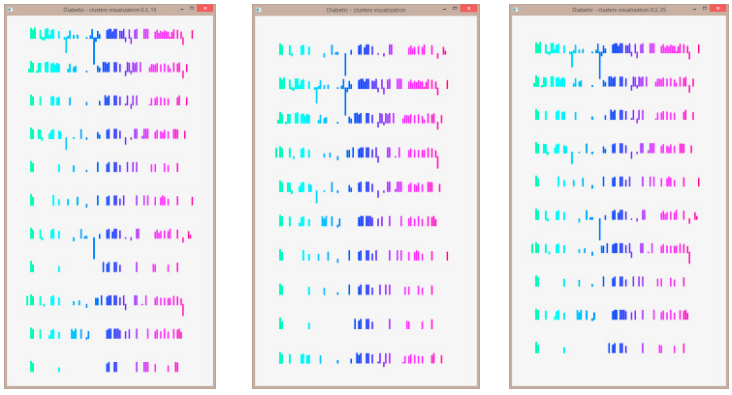

Fig. 3. Sample visualization of clustering results. Bars represent averaged values of condensed attributes, lines represent different clusters discovered.

When the attributes are of different types and when they are numerous, it is difficult to properly choose a distance measure for the clustering algorithm. However, mapping all types of attributes into numerical ones, allowed us to use the Tanimoto measure – a measure which finds numerous applications e.g. in bio- and chemical-informatics, but also in web and text mining. The Tanimoto similarity measure proved to be very efficient for calculating distances between high-dimensional representations of objects [25]. The measure for two vectors u and v can be calculated using the following formula:

$$T(u,v) = \frac{u \cdot v}{u \cdot u + v \cdot v - u \cdot v}.$$

For the preliminary experiments we used the DBSCAN algorithm. We run it several times so that we could adjust values of *MinPts* and *Eps* parameters. The number of clusters changed from 8 to 13. Then, we wanted to visualize the results of clustering. However, visual analysis of more than 1200 attributes simply plotted in a form of graphical bars would not result in providing an analyst with a visualization easy to understand and interpret. Thus, we rather focused on how to simplify (condense) the visual representation of the clustering results and it turned out, that in the analyzed dataset attributes can actually be grouped into segments of similar attributes (We show two sample groups of similar attributes In Figure 2). This observation allowed us to condense data to be visualized even more and allowed us to reduce the number of graphical objects representing a single class at least by one order of magnitude. We have managed to group attributes belonging to similar categories and calculate average values of attributes within attributes groups. This led to the reduction of dimensions to be shown to c.a. 100 dimensions.

The visualized results of three sample clustering experiments can be found in Figure 3. The DBSCAN parameters we used differed from 0.3 to 0.5 (*Eps*) and from 15 to 25 (*MinPts*). As one can see, there are groups of attributes (represented by single colored bars) in each of the clusters which clearly stand out (compared to the same attribute

groups from different clusters). Moreover, one can easily point which features make the cluster distinct from the other clusters by visually comparing graphical patterns of groups of similar attributes (bars). For the number of discovered clusters which is relatively small (not much greater than 10) this is an easy task and can lead the viewer to understand which features or attributes constitute particular clusters.

Table 1. The mapping of two chosen attributes (self-perceived health and self-perceived health – compared to 1 year ago) to a new attribute – a segment of a patient (Anxious, Depressed, Gloomy, Happy, Optimistic, Rejuvenated, Satisfied).

GEN_01 - Self-perceived health	count
VERY GOOD	46780
GOOD	37581
EXCELLENT	23126
FAIR	12844
POOR	4359
DON'T KNOW	217
REFUSAL	22

GEN_02 - Self-perceived hlth - compared 1 yr ago	count
ABOUT THE SAME	84759
SOMEWHAT BETTER	15193
SOMEWHAT WORSE	12772
MUCH BETTER	9639
MUCH WORSE	2317
DON'T KNOW	233
REFUSAL	16

mapped into 8 segments

Anxious
Depressed
Gloomy
Happy
Optimistic
Rejuvenated
Satisfied

Table 2. The comparison of clusters generated by DBSCAN and k-Means to test if the manual segmentation (by means of the additional attribute - the patient's segment defined in Table 1) reconciles with the results of clustering. In the tables below, each cluster has been split into rows corresponding to numbers of objects in each of the manually created segments.

DBSCAN
34 attributes
Tanimoto measure
4 clusters

cid	segment	count
1	Satisfied	127
1	Depressed	106
1	Gloomy	74
1	Happy	43
1	Optimistic	10
1	Anxious	6
1	Rejuvenated	2
2	Satisfied	2892
2	Optimistic	886
2	Gloomy	868
2	Happy	788
2	Anxious	631
2	Rejuvenated	616
2	Depressed	335
3	Satisfied	1619
3	Gloomy	550
3	Optimistic	538
3	Happy	519
3	Anxious	322
3	Rejuvenated	237
3	Depressed	113
4	Satisfied	404
4	Optimistic	300
4	Happy	157
4	Gloomy	98
4	Anxious	73
4	Rejuvenated	30
4	Depressed	14
noise	Satisfied	1529
noise	Depressed	928
noise	Gloomy	768
noise	Happy	494
noise	Optimistic	317
noise	Rejuvenated	306
noise	Anxious	136

DBSCAN
7 attributes (with highest weights)
Tanimoto measure
2 clusters

cid	segment	count
1	Satisfied	5167
1	Gloomy	1827
1	Happy	1409
1	Optimistic	1356
1	Depressed	1230
1	Rejuvenated	1097
1	Anxious	932
2	Satisfied	1399
2	Optimistic	693
2	Happy	587
2	Gloomy	529
2	Depressed	260
2	Anxious	235
2	Rejuvenated	88
noise	Depressed	6
noise	Rejuvenated	6
noise	Happy	5
noise	Satisfied	5
noise	Optimistic	2
noise	Gloomy	2
noise	Anxious	1

K-Means
7 attributes (with highest weights)
Manhattan dist. measure
k = 7

cid	segment	count
0	Satisfied	2166
0	Gloomy	642
0	Happy	526
0	Optimistic	417
0	Rejuvenated	413
0	Depressed	402
0	Anxious	393
1	Depressed	520
1	Satisfied	512
1	Gloomy	408
1	Rejuvenated	162
1	Happy	134
1	Anxious	54
1	Optimistic	26
2	Optimistic	826
2	Satisfied	643
2	Anxious	414
2	Rejuvenated	326
2	Happy	260
2	Gloomy	109
2	Depressed	9
3	Satisfied	474
3	Gloomy	273
3	Happy	157
3	Depressed	152
3	Rejuvenated	83
3	Optimistic	66
3	Anxious	55
4	Satisfied	1209
4	Gloomy	294
4	Happy	283
4	Depressed	121
4	Rejuvenated	64
4	Optimistic	19
4	Anxious	12
5	Satisfied	223
5	Gloomy	148
5	Depressed	114
5	Happy	65
5	Rejuvenated	62
5	Anxious	18
5	Optimistic	8
6	Satisfied	1344

K-Means
7 attributes (highest weights)
Euclidean dist. measure
k=7

cid	segment	count
0	Satisfied	458
0	Gloomy	233
0	Happy	140
0	Rejuvenated	77
0	Depressed	74
0	Anxious	67
0	Optimistic	50
1	Satisfied	1990
1	Gloomy	556
1	Happy	458
1	Depressed	377
1	Optimistic	375
1	Rejuvenated	361
1	Anxious	342
2	Depressed	496
2	Satisfied	459
2	Gloomy	366
2	Rejuvenated	182
2	Happy	123
2	Anxious	42
2	Optimistic	17
3	Satisfied	1402
3	Optimistic	694
3	Happy	588
3	Gloomy	531
3	Depressed	261
3	Anxious	235
3	Rejuvenated	91
4	Satisfied	1557
4	Happy	405
4	Gloomy	339
4	Depressed	114
4	Rejuvenated	66
4	Optimistic	21
4	Anxious	11
5	Optimistic	832
5	Anxious	411
5	Rejuvenated	321
5	Satisfied	232
5	Happy	127
5	Gloomy	46
5	Depressed	9
6	Satisfied	473

6	Optimistic	689
6	Happy	576
6	Gloomy	484
6	Anxious	222
6	Depressed	178
6	Rejuvenated	81

6	Gloomy	287
6	Depressed	165
6	Happy	160
6	Rejuvenated	93
6	Optimistic	62
6	Anxious	60

3.3. Finding correlation between segments and clusters

The goal of this experiment was to check whether the manual segmentation based on a new attribute (the patient's segment) (Table 1) reconciles with the assignment of patients to clusters generated by the clustering algorithms. The intuition behind introducing this attribute was that the patients are able to describe by themselves the state of their health quite well (especially when comparing to how they felt a year ago).

In this case we have performed the experiments using only 34 chosen attributes (selected by statistical importance of those attributes). The algorithms we used were DBSCAN with the Tanimoto measure and k-Means with Euclidean distance measures. Regardless of initial intuition, the results we obtained did not confirm that such a correlation exists. In each of the discovered groups we could see representatives of all of the segments which means that the segments were not in line with any of the clustering results we have obtained. We came to the same conclusions clustering the data set using only 7 selected attributes and the same algorithms with similar settings. The results has been presented in Table 2.

3.4. High-dimensional visualization

In this series of experiments, we have decided not to preprocess the dataset but to use all dimensions available. Additionally, we did not try to map the values of nominal and interval attributes into numerical ones. Instead we have employed a special distance measure which can be used to calculate distances between objects described with attributes of different types. However, we have limited our experiments to the patients with diabetes and heart diseases as those two often occur together. Below we describe the details of the experiments starting from the distance measure we used, parameters of the clustering algorithm. Then we describe the system we have implemented and discuss its possibilities for data visualization of high dimensional datasets.

Distance measure. The distance measure ($dist\_hd(u, v)$) we used to handle multiple types of attributes can be express with the following formula $dist\_hd(u, v) = \sum_{i=1}^{m} dist(u_i, v_i)$, where u, v are multidimensional vectors. The $dist(u_i, v_i)$ function is defined as follows:

$$dist(u_i, v_i) = \begin{cases} 0 & \text{if } u_i \text{ and } v_i \text{ are nominal and } u_i = v_i \\ 1 & \text{if } u_i \text{ and } v_i \text{ are nominal and } u_i \neq v_i, \\ |norm(u_i) - norm(v_i)| & \text{if } u_i \text{ and } v_i \text{ are continuous} \end{cases}$$

where m is the number of attributes.

Clustering with DBSCAN. We used DBSCAN to cluster dataset. This time we have performed a number of experiments with different values of *MinPts* and *Eps* parameters of DBSCAN (using similar values to those used in the experiments described earlier). However, in this case, we used more attributes, starting from 37 selected ones (based on their statistical significance) up to all 1207 attributes. Also, we did not try to reduce the number of attributes to make clustering results visualizations easier to understand. Instead, we have implemented a visualization client capable of presenting all attributes at the same time.

The visualization client. The application we have developed is capable of presenting all the attributes using for clustering in one single visualization. It may seem, and we have actually pointed this out earlier, that a visualization of a large number of attributes can be difficult to analyze but we have tried to visualize them so that the viewer would actually able to visually analyze them easily (Figure 4). First, every cluster is represented by a set of bars (attributes). Each bar represents a single attribute which is then divided into sub-bars, where each sub-bar corresponds to a single value of an attribute (the length of the sub-bar is calculated based on how often the particular value occurs within objects assigned to the cluster). Additionally, when a user moves the cursor over a particular bar (attribute) or a sub-

bar (value), the application display the name of the attribute and its value over which the cursor is located). The user of the application can easily configure a subset of attributes to display, run the clustering with different parameters, save current experiment into a database as well as open previously saved clustering results as a visualization. The application is available for cloning from a GitHub repository: https://github.com/piotrlasek/cv-hd.

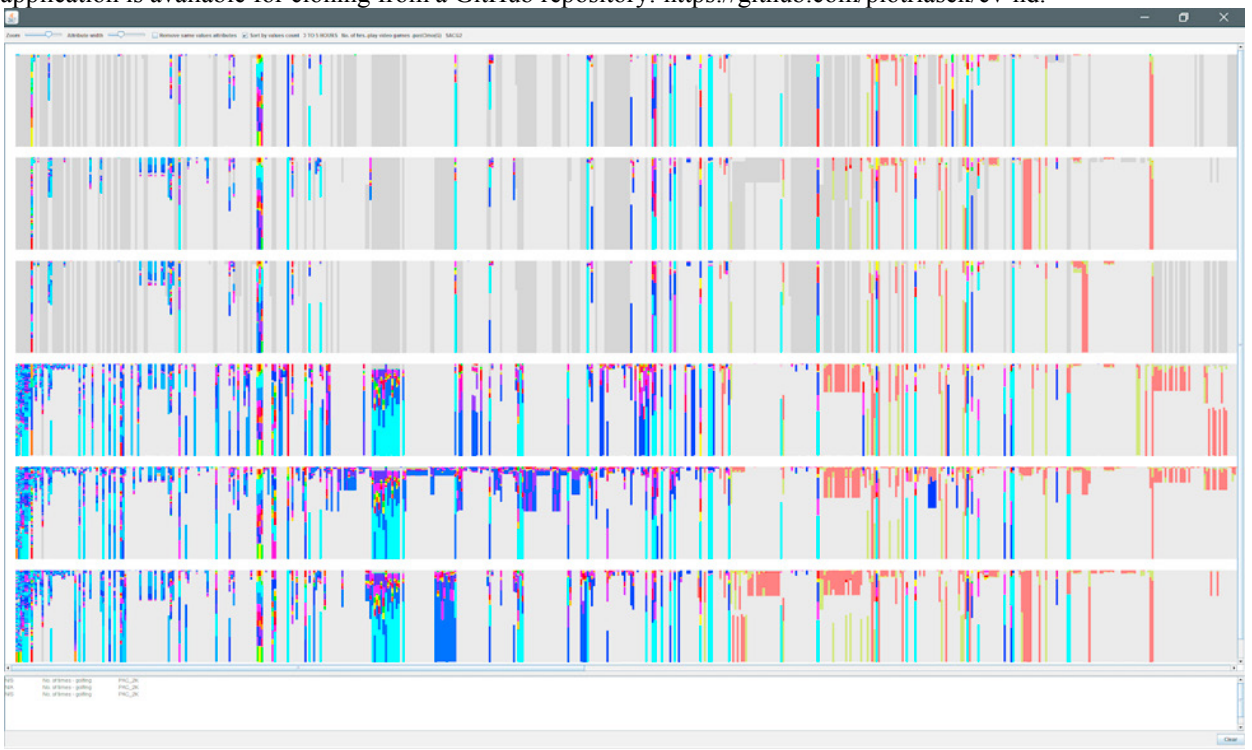

Fig. 4. A sample visualization of clustering of a high-dimensional dataset.

Table 3. Visual interpretation of clustering results (experiment 3.4.1).

Cluster	Distress scale - past month	Daily consumption - fruit	Age	Province of residence of respondent	Knowledge languages	Perceived Health	Main source of personal income	Marital status
1				NS, BC, SA, NB, MA				
2	H	H		ON				
3				QU	FRENCH			Widow
4			> 80	ON		POOR	Seniors benefits	
5				MA				

Table 4. Visual interpretation of clustering results (experiment 3.4.2).

Cluster	BMI	Daily energy expenditure	Height (metres)	Satisfaction with life in general	Province of residence of respondent	Total income	Perceived Health
1	34,5		1.664 TO 1.688 M		QU, BC, NB, SA	20 000	Good, Very good
2	34,5		1.664 TO 1.688 M	H	ON	20 000	Good, Very good
3					ON	20 000	Good, Poor
4	27	H	1.816 TO 1.841 M		AL	80 000	Good, Very good
5	28		1.816 TO 1.841 M		MA	80 000	Good, Very good

Table 5. Visual interpretation of clustering results (experiment 3.4.3).

Cluster	Age	Income	Province	Living arrangements	Knowledge languages	Perceived Health	Household size	Marital status	First language	Type of drinker	Sex
1			BC, NB, SA, NS			Good, Poor	1,2	Married, Widow	Eng	No	M/F

2			ON			Good, Very good	1,2	Married, Widow	Eng	No	M/F
3			QU		French only	Good, Very good	1,2	Married, Widow	Fr	No	M/F
4	> 80	DECILE 5	ON	With spouse		Good, Poor	1	Married	Eng	Regular	M

Interpretation of results. In Tables 3-4 we have collected several findings based on visual analysis of sample clustering results of three chose experiments which we have conducted. The results of the analysis seem to provide valuable knowledge about the analyzed dataset. The algorithm discovered that, depending on the values of parameters we used, there exist several different groups of diabetes patients which can be characterized with several different attributes. Some of the interesting examples are:

- There exists a group of older people in Ontario whose main source of personal income is senior benefit, whose perceived health is poor (Table 3).
- High daily energy expenditure and good or very good perceived health are characteristic features of diabetes patients from Alberta with income higher than 80000 dollars (Table 4).
- Patients who live in the Quebec province, speak French only, are married or widowed, do not drink, perceive their health as good or very good (Table 5).

4. Conclusions and further works

In this paper we have aimed at investigating possibilities of using data clustering and visualization methods to analyze a sample high-dimensional diabetes dataset. We have performed a number of clustering experiments using two commonly known clustering algorithms (DBSCAN, k-Means) with different distance measures (Euclidean, Manhattan, Nominal-Numeric). We have also experimented and proposed a method for visual analysis of clustering results using our developed visualization application which main feature is that it is capable of visualizing results of clustering of high-dimensional datasets so that the graphical representation is not disturbed with additional information (attributes names or values) and the viewer can focus on analyzing the general view of clustering results but can get detailed information on interesting attributes according to the known Schneiderman's mantra: *Overview first, zoom and filter, details on demand*.

Our further works will aim at further development of the visualization application and focusing on implementing and experimenting with new methods of interactive data exploration and visualization.

References

[1] J. Ahrens, S. Jourdain, P. O'Leary, J. Patchett, D. H. Rogers, and M. Petersen, "An image-based approach to extreme scale in situ visualization and analysis," in Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal., 2014, pp. 424–434.

[2] Ng, Raymond T., and Jiawei Han. "CLARANS: A method for clustering objects for spatial data mining." IEEE Transactions on Knowledge & Data Engineering 5 (2002): 1003-1016.

[3] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. No. 34. 1996.

[4] Centers for Disease Control and Prevention (2010). Chronic disease and health promotion. National Center for Center for Chronic Disease Prevention and Health Promotion.

[5] Dubes, Richard C. "How many clusters are best?-an experiment." Pattern Recognition 20.6 (1987): 645-663.

[6] Anderberg, M. (1973). Cluster Analysis for Applications. Academic Press, Inc., New York, NY.

[7] Crossing the Quality Chasm: A new health system for the 21st century. Institute of Medicine. Accessed at www.IOM.edu

[8] Educate before you medicate: Enhancing prescription medication adherence: A national action plan. National Council on Patient Information and Education, August 2007

[9] K-means Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.

[10] Shneiderman, Ben. "Science 2.0." *Science* 319.5868 (2008): 1349-1350.

[11] McQueen, R. G., S. P. Marsh, and J. N. Fritz. "Hugoniot equation of state of twelve rocks." Journal of Geophysical Research 72.20 (1967): 4999-5036.

[12] Choi, Seung-Seok, Sung-Hyuk Cha, and Charles C. Tappert. "A survey of binary similarity and distance measures." Journal of Systemics, Cybernetics and Informatics 8.1 (2010): 43-48.

[13] R. L. Sallam, B. Hostmann, K. Schlegel, J. Tapadinhas, J. Parenteau, and T. W. Oestreich. Magic quadrant for business intelligence and analytics platforms. Gartner report, 2015.

[14] Partnering with patients to drive shared decisions, better value and care improvement. Workshop Proceedings. Institute of Medicine, August 15, 2013

[15] Preventing and managing chronic disease: Ontario's Framework. 2007 Ministry of Health and Long-term Care. Accessed at www.health.gov.on.ca/en/pro/programs/cdpm/pdf/framework_full.pdf

[16] Sabate, E. (2003). Adherence to Long Term Therapies: Evidence for Action. World Health Organization (WHO).

[19] Think outside the pillbox: Six priorities for action to support improved medication adherence. New England Health Institute (NEHI), July 19, 2013

[20] Lin, Rongheng, et al. "Chronic diseases and health monitoring big data: A survey." IEEE reviews in biomedical engineering 11 (2018): 275-288.

[21] Archenaa, J., and EA Mary Anita. "A survey of big data analytics in healthcare and government." Procedia Computer Science 50 (2015): 408-413.

[22] Godfrey, Parke, Jarek Gryz, and Piotr Lasek. "Interactive visualization of large data sets." IEEE Transactions on Knowledge and Data Engineering 28.8 (2016): 2142-2157.

[23] Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

[24] Wu, Eugene, Leilani Battle, and Samuel R. Madden. "The case for data visualization management systems: vision paper." Proceedings of the VLDB Endowment 7.10 (2014): 903-906.

[25] Kryszkiewicz, Marzena. "Using non-zero dimensions and lengths of vectors for the tanimoto similarity search among real valued vectors." Asian Conference on Intelligent Information and Database Systems. Springer, Cham, 2014.

[26] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In Visual Languages, 1996. Proceedings., IEEE Symposium on, pages 336–343. IEEE, 1996.