Bachelor thesis
Computer Science

Radboud University

K-Means clustering of a database with PTSD and other psychiatric patients

Author:
Thomas de Bel
s3032817

First supervisor/assessor:
PhD. Twan van Laarhoven
tvanlaarhoven@cs.ru.nl

[Second supervisor:]
Associate Prof. Elena Marchiori
elenam@cs.ru.nl

February 2, 2015

0.1 Acknowledgements

I want to thank my supervisor Twan van Laarhoven for his time, assistance and guidance in the research. I would also like to thank my second supervisor associate prof. Elena Marchiori for her time and tips. Furthermore, I want to thank Laurens de Vries at ProPersona for his time and for giving me the opportunity to work on this greatly interesting subject.

Finally, I would like to thank Sanne Arts for her love and support during the whole project.

Abstract

The health care industry has become more information-rich. Large amounts of data are accumulated on a daily basis. With this development, data mining can be a useful tool for extracting useful information out of the data. In this study, a K-means clustering algorithm is used on a database of psychiatric patients. The aim of this study is to use K-means to see if a differentiation can be made between patients with post-traumatic stress disorder (PTSD) and patients without PTSD. Furthermore, we aimed to determine specific subgroups in patients that have PTSD. The database was provided by ProPersona Nijmegen. The database contained information about the medical history of the psychiatric patients. 16 dimensions were formed: age, gender and 14 disorder categories. A total of 26,769 patients were included in the study. The results of the clustering are promising for future research, but no clear differentiation between PTSD and non-PTSD was made. Some clear subgroups were formed by the clustering algorithm. This research serves as a proof-of-concept for the usefulness clustering in psychiatry. The subgroups that were found can be used for further research. Clustering on symptoms can possibly yield interesting results.

Contents

Chapter 1

Introduction

In the last decades data mining techniques have found their way into medical sciences. Data mining can be described as the area of applied mathematics that tries to extract information from large datasets, often stored in large computer databases (Tovar et al., 2012). In this research we want to look at one specific field of medicine: psychiatry. As every other branch of medicine, it's research includes patient monitoring, animal studies, and in vivo and in vitro studies, which generate large amounts of data. According to Tovar et al. (2012) this is why data mining can be proven to be a useful tool in psychiatric research. They state that even the most experienced clinical scientists fail to analyze data properly and draw safe conclusions. Due to the complications of the task, mathematical modeling and data mining should be used to assist (Tovar et al., 2012).

Since one of the major preconditions for applying data mining techniques is the existence of uniform data sets, data mining is mostly used in the biomedical field within medical research, such as in research about gene expression and regulation. It is less frequently applied in everyday medical work. Some studies, however, indicate that data mining techniques can also analyze various data collected from the patient (such as short medical history or specialist findings). In this case algorithms can be used for finding relations between different parameters, leading to more targeted therapeutic interventions (Marinic et. al., 2007). In the past few decades the health care industry has become more information-rich. Large amounts of data are produced daily from the processing of health care transactions, carried out by both patients as well as doctors. With these vast amounts of information, data mining has become more prevalent in the health care business (Wang et al., 2012).

In this research we will try to take a small step in the direction of transforming medical data, through data mining, into information that is valuable for identifying patients with post-traumatic stress disorder (PTSD). We aim

specifically using clustering to help identify (subgroups of) patients with post-traumatic stress disorder. We have a large database of psychiatric patients at our disposal that includes both patients with PTSD as well as other psychiatric patients (patients without PTSD). The contents of the database consist of the medical history of the patients. To this database we will apply clustering, which falls under unsupervised learning. This means that we aim to discover certain patterns in unlabeled data by finding associations between data points (Tovar et al., 2012). Specifically, we will use a K-means clustering algorithm to achieve this. A first goal is to identify specific clusters in which patients with PTSD differentiate from psychiatric patients without PTSD. Furthermore, we want to identify specific clusters within the group of patients with PTSD that indicate certain subgroups. With the clustering we try to achieve a better understanding of PTSD. Our purpose is to contribute to faster diagnostic procedures and more targeted therapeutic interventions. With the clustering we also try to achieve a second goal, namely to visualize the data of psychiatric patients in a way, so that it can easily be comprehended and analyzed. The main question of this research is:

- **How can K-means clustering, through analyzing and visualizing a database of psychiatric patients, contribute to identifying patients with PTSD?**

In an attempt to answer this question, the following sub-questions will be dealt with:

- How can K-means clustering be applied to analyze the database?
- How can the K-means clustering of the database be visualized in a conveniently arranged figure?
- What specific clusters can be found with K-means clustering, which differentiate patients with PTSD from patients without PTSD?
- What specific clusters can be found with K-means clustering, within the group of patients with PTSD, that could indicate certain subgroups?

Chapter 2

Preliminaries

In this chapter we will first discuss what PTSD is and how it is diagnosed (paragraph 2.1). We will also explain what K-means clustering is (paragraph 2.2).

2.1 Post-traumatic stress disorder

PTSD is a psychiatric disorder that can occur in people who have experienced or witnessed life-threatening events. This could be a natural disaster, serious accident, terrorist incident, war or a violent personal assault like rape. PTSD usually appears within three months of the trauma, but sometimes the disorder appears later. PTSD patients often relive the experience through flashbacks or nightmares, have difficulty sleeping, and feel detached or estranged. Although it was once thought to be mostly a disorder of war veterans who had been involved in combat, researchers found that PTSD also affects civilians, both male and female (American Psychiatric Association, 2011).

In some cases the symptoms of PTSD disappear with time, whereas in others they can stay for many years. Symptoms of PTSD can be grouped within three categories: intrusion, avoidance and hyperarousal. Intrusion stands for the intrusion of episodes called 'flashbacks' into the current life of a patient. Flashbacks are the unexpected re-occurrence of memories of the trauma. Avoidance symptoms affect the relationship of the patient with others. For instance, a person with PTSD often avoids close emotional ties with family, colleagues, and friends. The inability of a person with PTSD to work out grief, anger, or fear from the traumatic event can influence the person's behavior without the individual being aware of it. Hyperarousal means that a person can act as if they are constantly threatened by the trauma that caused their illness. A patient can suddenly become irritable or explosive, even when unprovoked (American Psychiatric Association, 2011).

To diagnose PTSD psychiatrists use the Diagnostic and Statistical Man-

ual of Mental Disorders (DSM). The DSM has gone through a number of revisions through the years, and in 2013 the fifth edition was published. In the fourth edition PTSD is considered a type of anxiety disorder but in the new edition PTSD is moved into a new category: 'Trauma and Stress-related Disorders' (Staggs, 2013). However in this research we use the fourth edition because the fifth edition was not yet published when the patients in the database were diagnosed. A full description of the diagnostic criteria for PTSD can be found in Appendix A.

2.2 K-means Clustering

K-means clustering is a form of unsupervised learning, where the task is to divide unlabeled data points into similar groups. The main principle is to take a certain amount of centroids. These centroids are placed in such a way that the objective function is as small as possible. This objective function in the case of K-means clustering is called the inertia or sum of squared error function (SSE) (Berkhin, 2006). Minimizing the SSE is defined as follows:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_j - \mu_i||^2)$$

The algorithm consists of the following steps to minimize the SSE:

1. Place k centroids in the space that represents the data points.
2. Assign to each data point one of the centroids that is closest to the data point.
3. Adjust the centroids according to all data-points assigned to them.
4. Repeat steps two and three, until the centroids do not move more than some threshold.

SSE is not a normalized metric. Because of this, there is no such thing as a perfect clustering. A lower SSE means a better clustering, with zero being the optimum. K-means is one of the most popular clustering tools used in scientific and industrial application. Some weaknesses of the K-means clustering are (Berkhin, 2006):

- The results are dependent on the initial positions of the centroids.
- It is not clear beforehand how many clusters should be used.
- The clustering is sensitive to outliers.

These shortcomings will be addressed in this research.

Chapter 3

Research

In this chapter we will first address the acquisition and the preprocessing of the data (paragraph 3.1). After this we will look at the general statistics of the database to get accustomed to the database and to see how this database relates to literature (paragraph 3.2). We will then cluster the data and show the results graphically (some of the code used for the clustering and visualization can be found in Appendix B) (paragraph 3.3). We will discuss the results of the first clustering, that includes patients with PTSD as well as psychiatric patients without PTSD (paragraph 3.4). With this we attempt to differentiate between patients with PTSD and psychiatric patients without PTSD. In the second clustering we only include the patients with PTSD and we will also discuss these results (paragraph 3.5). With this cluster analysis we attempt to find certain subgroups within the group of patients with PTSD.

3.1 Data acquisition and preprocessing

3.1.1 Data acquisition

The data analyzed in this research originates from the patient database of the institute ProPersona Nijmegen. It is obtained through a tenant of ProPersona Nijmegen. Before the data was supplied it was made anonymous for privacy purposes. The database includes all patients of ProPersona that are diagnosed with PTSD, a total amount of 7,543. However, it does not include all psychiatric patients that are diagnosed with other psychiatric disorders, because not all cases were available. In the database 19,226 of psychiatric patients without PTSD are included. Because this is a substantial part of the psychiatric patients without PTSD we assume that this selection is a good representation of the psychiatric patients without PTSD. ProPersona included the gender, age and medical diagnosis (presented as DSM-codes) of the patients in the database.

3.1.2 Preprocessing

Preprocessing was done on the database to allow for direct input into the classification algorithm. We included 16 dimensions in this research. The database of patients with PTSD was converted to a two-dimensional array of 19,43 rows by 16 columns, where each tuple corresponded with one patient. The same was done for all psychiatric patients without PTSD. The two datasets were merged together to form one complete dataset. The 16 dimensions in this research are:

- Gender - from three patients the gender was unknown, these are removed from the database. Gender is a nominal value and therefore male gender was converted to '1' and female gender to '0'.
- Age - age is a continuous variable. Because a Euclidean Distance metric was used, age was normalized to fall between '0' and '1' as well, to give it an equal weight compared to the other dimensions.

Psychiatric disorders
In consultation with ProPersona the amount of dimensions is reduced by dividing a total of 194 psychiatric disorders into 14 categories. The division of disorders into categories can be viewed in Appendix C. The categories that were chosen are similar to the categories specified in DSM-IV. A few of the psychiatric disorders are not represented in the database, this is because of the rarity of these disorders. The medical diagnosis represents a nominal value and therefore a '1' was scored if the psychiatric disorder was present in the lifetime prevalence of a patients and a '0' was scored if it was not.

- Psychotic disorder - mental disorders that cause abnormal thinking and perception. For example: schizophrenia.
- Mood disorder - group of diagnoses where a disturbance in the person's mood is the main underlying feature. For example: depression
- Bipolar disorder - disorder characterized by periods of elevated mood and periods of depression.
- Anxiety disorder (PTSD excluded) - mental disorders characterized by feelings of anxiety and fear. For example: panic disorder. Note that PTSD falls originally under this category but is not included in a category because this is the disorder we want to differentiate on.
- Personality disorder – a class of disorders characterized by patterns of behavior that deviate from the accepted cultural standards. For example: antisocial personality disorder.
- Borderline personality disorder – a disorder characterized by impulsive behavior, and an unstable affect and self-image.

- Drug/medication bound disorder – disorders characterized by substance abuse, for example drugs or alcohol.
- Dissociative disorder – a group of conditions that involve disruption in identity and memory. For example: dissociative fugue.
- Pervasive developmental disorder – a class of disorders characterized by delay in development. For example: autism.
- Attention deficit- and behavior disorder – a group of conditions that are characterized by the inability to concentrate. For example: ADHD.
- Sleep disorder – a group of conditions that revolve around the sleeping pattern. For example: narcolepsy.
- Somatoform disorder – a class of disorders with physical symptoms that can not be explained by physical illness. For example: conversion disorder.
- Factitious disorder – disorders where patients hurt themselves or others to generate attention. For example: Münchhausen by proxy.
- Eating disorder - group of conditions that revolve around the eating pattern. For example: anorexia (APA, 2000).

3.2 General Statistics of the database

In this section we will relate the characteristics of this database to recent literature about PTSD. We will discuss where the population of this database differs from the general population of patients with PTSD.

3.2.1 Gender and age

	PTSD(%)	Non-PTSD(%)
Database Population	28.17	72.81

Table 3.1: The percentage of patients with and without PTSD in the complete database.

	Total population(%)	PTSD(%)	Non-PTSD(%)
Male	43.27	35.3	46.4
Female	56.73	64.7	53.6

Table 3.2: Percentages of the male and female population.

The database consists of 7,543 patients with PTSD and 19,226 psychiatric patients without PTSD (Table 3.1). If we categorize the database by gender a high amount of female patients in the group of PTSD catches the eye: 64.7% is female, as opposed to 35.3% being male (Table 3.2). This seems to correspond with other research, because one of the most consistently reported risk factors for PTSD is being female (Shansky, 2015). Haskell et al. (2010) for example state that women are twice as likely as men to develop PTSD after a trauma. The reasons for this discrepancy however are still poorly understood (Shansky, 2015).

Speculations have been made that the increased risk of PTSD among females is due to the higher likelihood of females to experience specific trauma types that appear to be particularly traumatic or PTSD inducing. However it has been reported that the increased prevalence of PTSD in women remains even when trauma type is corrected for (Ditlevsen and Elklit, 2010). For example, the study of Tolin and Foa (2006) shows that the twofold risk of PTSD among women can not be attributed to a higher risk of sexual traumas. Another possible explanation is that gender has been found to be an important biological determinant of vulnerability to psychosocial stress (Ditlevsen and Elklit, 2010). Furthermore, some arguments have been made that the increased PTSD prevalence among women is due to a report bias because men tend to under-report and women tend to over-report symptoms of PTSD (Saxe and Wolfe, 1999). This could be influenced by the social expectancy related to the male and female gender role. Where women are expected to be vulnerable, men are expected to be tough and more resilient to trauma (Tolin and Foa, 2006).

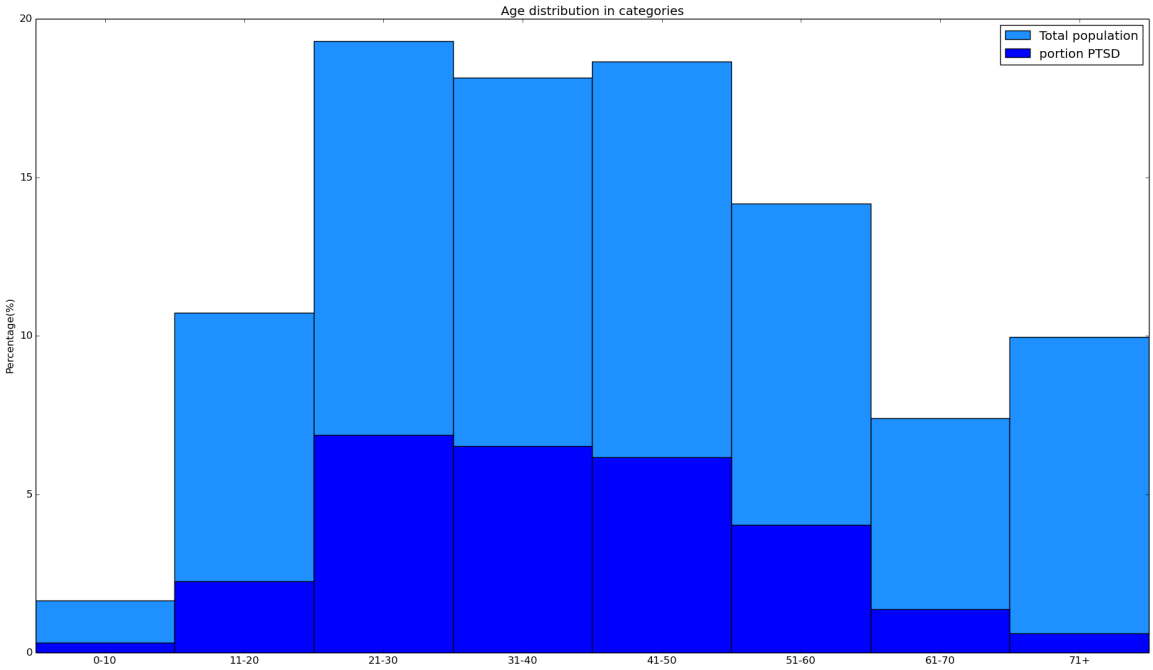

Figure 3.1: Age-distribution of the total database population by category. The portion of patients with PTSD is depicted in a darker color.

The age distribution of the database is shown in figure 3.1. Because not all psychiatric patients (that are not diagnosed with PTSD) are included in the database, the real portion of patients with PTSD (compared to all psychiatric patients) is smaller in the general population of psychiatric patients. In this database PTSD is less common in the youngest and oldest psychiatric patients (younger than 20 years and older than 61 years).

Fewer age studies than gender studies are represented in the PTSD literature and the conclusions of these researches do not always appear to be the same (Dillevsen and Elklit, 2010). Also, in a substantial part of other research age extremities (childhood or late life) don't seem to be included. For example Norris et al. (2002) include participants between 15 and 45 years and examined the effects of age on PTSD in a cultural context and compared the effects of age after similar disasters in three different parts of the world. The findings showed no consistent effect of age on PTSD. It was concluded that PTSD depended upon other factors more than it depended on age. Our results do seem to coincide with the research of Creamer & Parslow (2008) that included participants beyond the age of 54. They state that both male and female participants above the age of 65 reported negli-

gible rates of PTSD. The findings also suggested that the highest rates of PTSD prevalence among both men and women are found between the age of 18 and 24 years old. This corresponds with the age distribution in our dataset.

3.2.2 Comorbidity

Amount of diagnoses	Percentage(%)
0	27.18
1	33.13
2	22.19
3	11.04
4	4.91
5	1.31
>6	0.24

Table 3.3: Amount of diagnoses apart from PTSD, within the group of patients with PTSD.

In table 3.3 the amount of diagnoses in the PTSD group are shown. More than 7 out of 10 (72.82%) of the PTSD-diagnosed patients are also diagnosed with at least one other psychiatric disorder, which are included in the categories. Almost 2 out of 10 (17.50%) are diagnosed with three other psychiatric diseases or more. This does not necessarily mean that a patient suffers from more than one psychiatric disease at the time (note that the database includes the diagnoses from a patients' lifetime), it however does indicate that co-morbidity is likely. Other research also state that PTSD frequently appears accompanied by other psychiatric disorders (Dadic-Hero e.a., 2009). For example, Brady e.a. (2000) state that the vast majority of individuals with PTSD meet criteria for at least one other psychiatric disorder. They also state that a substantial percentage of the PTSD-diagnosed patients have three or more psychiatric diagnoses apart from their PTSD. They call co-morbidity in PTSD rather the rule than the exception (Brady e.a., 2000).

In other research a number of different hypothetical constructs have been posited to explain the high co-morbidity of PTSD. For example, O'Donnell e.a. (2004) asked the question whether PTSD and depression are separate disorders in the aftermath of trauma or part of a single general traumatic stress conduct. Based on their findings they can not answer this question but their findings suggest that when PTSD and a depression occur together, they reflect a shared vulnerability with similar predictive variables. This seems to correlate with the statement of Brady e.a. (2000) that depressive disorder

can be a common and independent consequence of exposure to trauma and having a previous depressive disorder is a risk factor for the development of PTSD once exposure to a trauma occurs.

	Total(%)	PTSD(%)	Non-PTSD(%)
psychotic disorder	7.71	6.26	8.29
mood disorder	35.93	46.32	31.85
bipolar disorder	2.75	1.51	3.24
anxiety disorder (PTSD excluded)	19.31	22.22	18.16
personality disorder	21.2	23.33	20.37
borderline personality disorder	7.06	12.94	4.75
drug/medication bound disorder	9.63	11.12	9.05
Dissociative disorder	0.81	2	0.34
pervasive developmental disorder	5.03	1.62	6.37
attention deficit- and behavior disorder	8.17	4.76	9.5
sleep disorder	1.09	1.05	1.11
somatoform disorder	4.29	5.17	3.94
factitious disorder	0.05	0.04	0.05
eating disorder	2.68	3.77	2.26

Table 3.4: *The percentages of each category for the total population of the database, the PTSD patients and the non-PTSD patients.*

The most common co-morbid (occurring at the same time) diagnoses of PTSD are depressive disorders, substance use disorders and other anxiety disorders (Brady e.a., 2000). The data from our database seems to support this statement partially. Depression falls under the category 'mood disorder' and almost half of the PTSD-diagnosed patients are also diagnosed with a mood disorder during their lifetime (46.32%). This is also significantly more if we compare it to the psychiatric patients that are not diagnosed with PTSD: 31.85% of the patients are diagnosed with a mood disorder within this group. The percentages of patients with a drug/medication bound disorder (substance use disorder) and anxiety disorder are (slightly) higher within the group of PTSD-diagnosed patients when compared to the non-PTSD diagnosed psychiatric patients. Our database suggests a high co-morbidity with other psychiatric disorders. For example the borderline personality disorder and the personality disorder seem to be more prevalent in PTSD-diagnosed patients than substance abuse disorder (drug/medication bound disorder).

Brady e.a. (2000) state that a substance use disorder may often develop as an attempt to self-medicate the painful symptoms of PTSD and that withdrawal states exaggerate these symptoms. In our database the percentage of drug/medication bound disorders is not higher for PTSD patients.

3.3 Clustering and visualization

3.3.1 Clustering

Firstly we determined the optimal amount of clusters. A heuristic for determining the optimal amount of clusters is the 'Elbow Criterion'. The elbow criterion says that you should not add any more clusters if there is no gain of information, a so-called elbow can be seen in the data plot(Madhulatha, 2012). In our case the sum of squared errors(SSE) will be used to determine this gain. Other more sophisticated cluster validation algorithms are available (Sugar et al.), but these are mainly tested on small dimension datasets. For the purpose of this study, the 'elbow' heuristic is sufficient. To find the optimal amount of clusters, we plotted SSE against the amount of clusters. This resulted in the following graph: As can be seen in the graph, the SSE

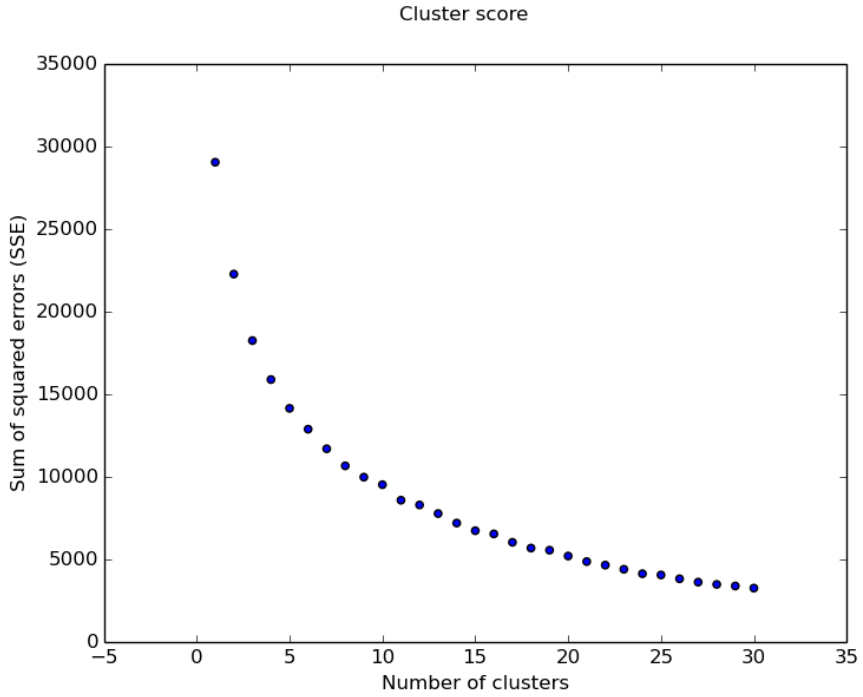

Figure 3.2: The sum of squared errors (SSE) plotted against the amount of clusters.

starts dropping linearly at around 10-15 clusters, the elbow point. Because of this, we chose to use an amount of fifteen clusters. The resulting SSE of the clustering was: 12082. This seems high, but because of the high dimensionality, this was expected. Similar to the total dataset, the amount of clusters in patients with PTSD was found to be optimal around 15 clusters.

The clustering was done with the skicit-learn library in Python. In

appendix B the full code can be viewed. Because the starting points of the clustering can determine the quality of the clustering, the clustering algorithm was initialized more than once. The best clustering was chosen out of 10,000 initializations. The threshold value for the centroids at which to stop iterating was set at 0.0005. The first clustering (on the complete dataset) focuses on identifying clusters in which a significantly high or low percentage of PTSD is shown. The second clustering (on the dataset with patients with PTSD) focuses on finding significant subgroups within the group of PTSD patients.

3.3.2 Visualization

Bar graphs were chosen to visualize the data. With bar graphs we were able to compress all information from the clustering into one graph. Another graph-sort that we considered were pie-charts. Because nominal values were used and because of the high dimensionality, a plot of all possible two-dimensional combinations was not viable. We tried this, but the graphs were unreadable.

3.4 Clustering of the complete dataset

3.4.1 Clustering of the complete dataset

The results of the first clustering are graphically shown on the next page (Figure 3.3). The percentage of patients with PTSD in the whole population is 28,18%. Some clusters differed significantly from this number. The distribution of PTSD over the clusters was as follows:

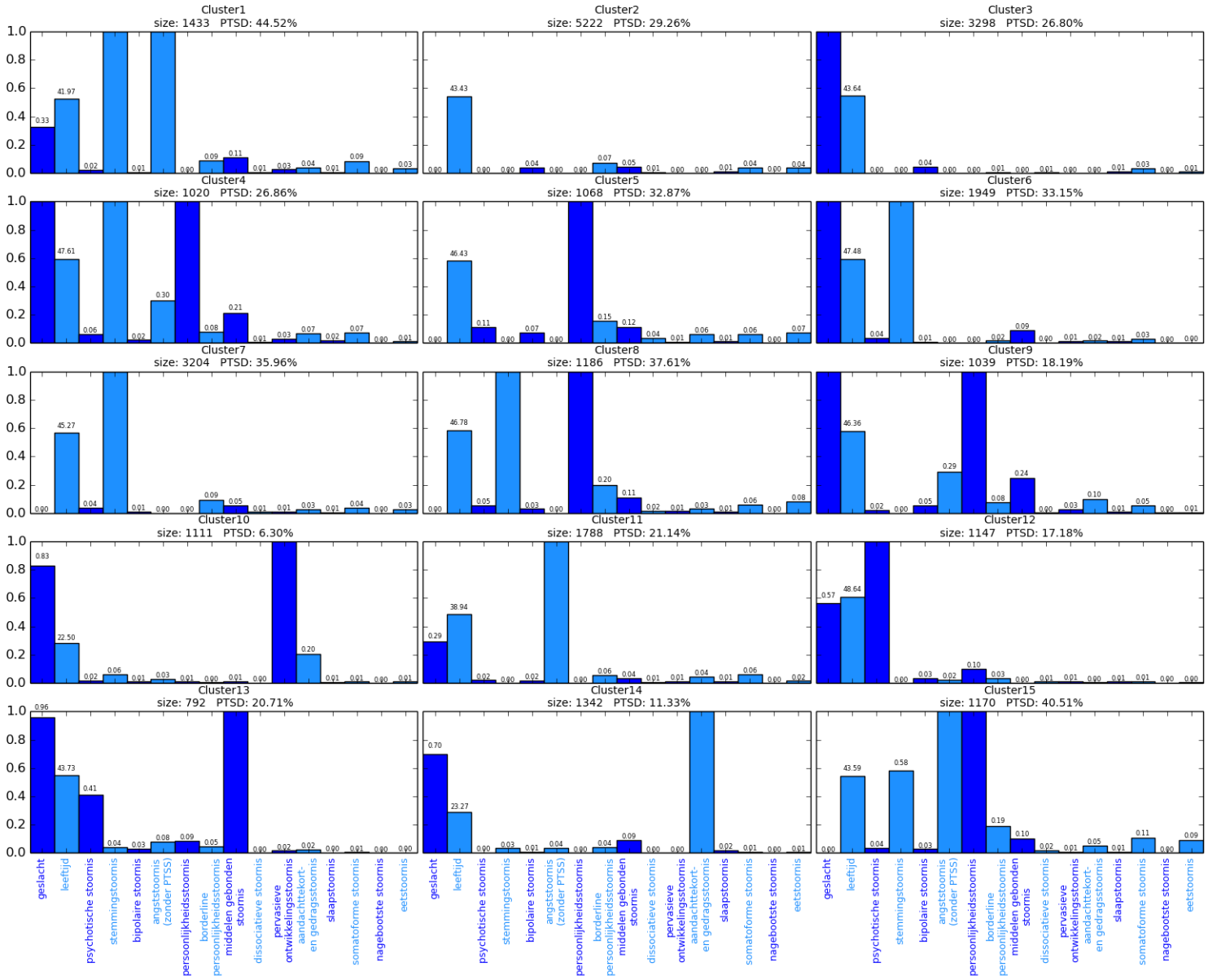

Figure 3.3: The results of the clustering are shown in the above figure. Each of the smaller graphs represents one cluster. The size as well as the percentage of patients with PTSD in each cluster is shown above the graph of each cluster.

95%-Confidence intervals (CI) will be given with relevant percentages. The confidence intervals are based on the standard error of the mean. The mean that is taken here is the percentage of patients with PTSD in the

Cluster Nr.	Cluster size (N)	PTSD(%)
1	638	44.52
2	1528	29.26
3	884	26.8
4	274	26.86
5	351	32.87
6	646	33.15
7	1152	35.96
8	446	37.61
9	189	18.19
10	70	6.3
11	378	21.14
12	197	17.18
13	164	20.71
14	152	11.33
15	474	40.51

Table 3.5: *The size of each cluster with the percentage of patients with PTSD in that cluster.*

whole population of the database. Most interesting were clusters 1 and 15. In cluster 1, 44.52% (CI[44.49%, 44.54%]) of the patients have PTSD and in cluster 15 there are 40.51% (CI[40.48%, 40.54%]). The characteristics of cluster 1 reveal that 100% of this cluster has had a mood disorder and an anxiety disorder (other than PTSD) in their lifetime prevalence. The majority of this cluster is female (67%). Cluster 15 is characterized by all cases having a personality disorder and an anxiety disorder. Another cluster that yielded interesting results, was cluster 10. In this cluster only 6.3%(CI[6.27%, 6.33%]) had PTSD. Characteristics of this group are that they all have a pervasive developmental disorder. The group mainly consists of male patients (83%). The clusters are roughly of equal size. Outliers are cluster 2 (N=5222) and cluster 3 (N=3298). These clusters score low on each category.

The standard deviation of the ages is presented in the following table: From this table we can derive that there are no significant differences in age per cluster.

Cluster number	Standard deviation
1	14.59
2	16.96
3	13.75
4	12.9
5	15.42
6	12.78
7	12.53
8	15.17
9	11.48
10	11.1
11	14.58
12	13.02
13	13.53
14	11.13
15	11.36

Table 3.6: *The standard deviation of the age for each cluster of the complete dataset.*

3.5 Clustering of PTSD patients

Clustering on the PTSD cases was carried out similar to the clustering of the whole dataset. The results are shown in figure 3.4. A few clusters show significant results. Cluster 1 shows a group of 987 female patients that all have a mood disorder. This is the same as cluster 7 (N=373), except for that patients in cluster 7 also have an anxiety disorder. Cluster 3 shows that 426 patients have a mood disorder as well as a personality disorder. This cluster consists mainly of female patients (73%). In cluster 4, all patients are of male gender and have a drug/medication bound disorder. Cluster 6 shows a group of 289 male patients with a mood disorder and an anxiety disorder. Borderline personality disorder is seen in all patients (female) in cluster 14.

As with the whole dataset, there is no cluster that is significantly determined by age. This can be seen in table 3.5.

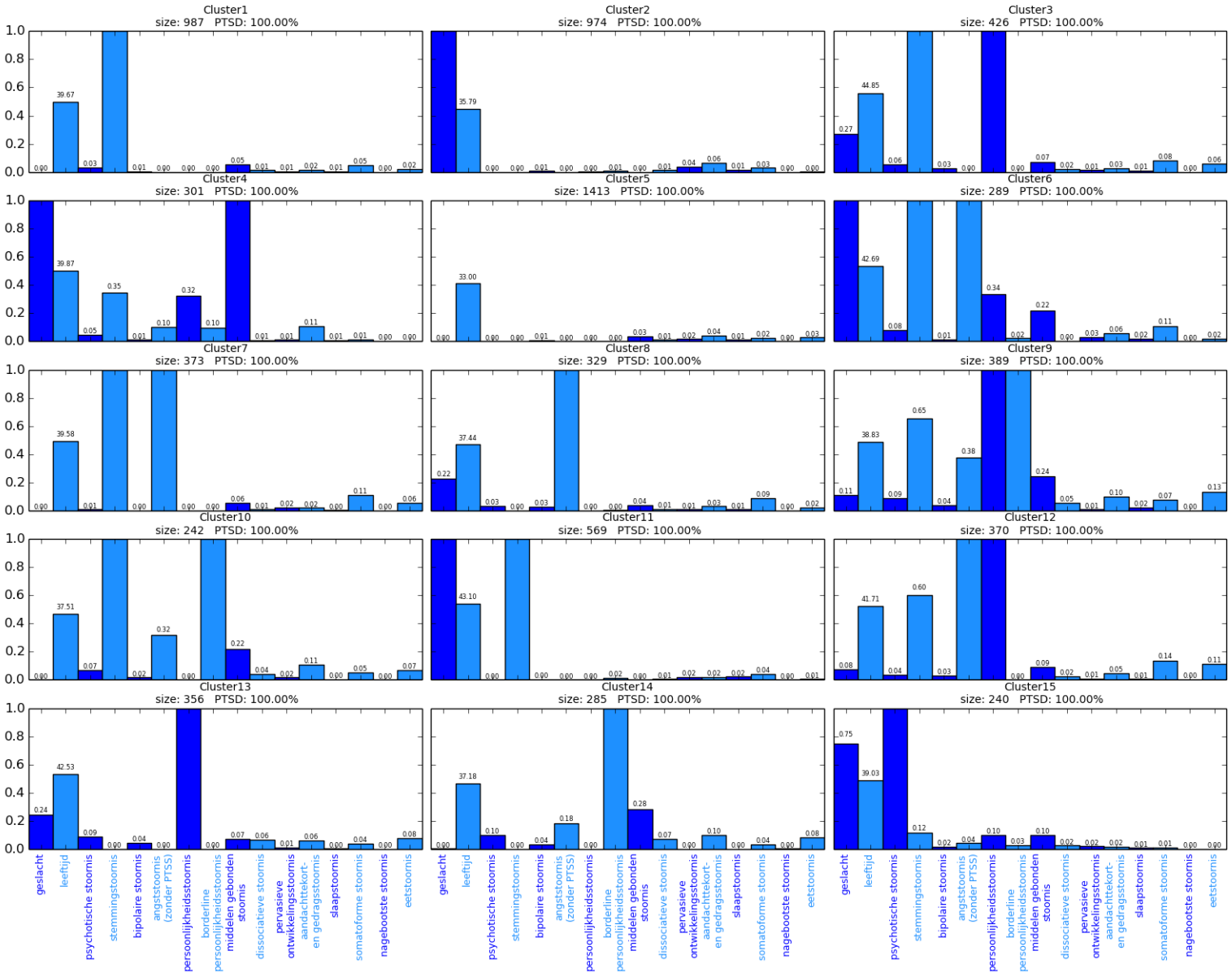

Figure 3.4: *The results of the clustering are shown in the above figure. Each of the smaller graphs represents one cluster. The size of the clusters is shown above each graph.*

Cluster number	Standard deviation
1	15.52
2	22.43
3	22.4
4	14.63
5	16.68
6	17.22
7	19.45
8	16.25
9	15.25
10	12.01
11	18.08
12	17.28
13	15.54
14	12.01
15	15.06

Table 3.7: *The standard deviation of the age for each cluster of the dataset with only PTSD patients.*

Chapter 4

Related Work

In this chapter we will give an overview of studies related to this research. Data mining in psychiatry is still a small field of research, but some interesting studies have been published.

4.1 Data Mining in Psychiatry

In a study by Hyejoo Lee et al. a k-means clustering was done to generate a hypothesis about the differences between Paternal Age Related Schizophrenia (PARS), a subgroup of schizophrenia, and other cases of schizophrenia. The study uses a very similar method to this study. A k-means clustering algorithm was run on different amounts of clusters to identify subgroups in which PARS was significantly high. A few clusters yielded significant results, meaning they found clusters with a high percentage of cases with PARS. One analysis revealed a cluster containing 83% PARS cases. Interesting about this research is that they compared a subtype of one disorder with the disorder as a whole. The study shows that this can yield interesting results. One of the big differences is the number of patients included in the research. In the study by Hyejoo Lee et al. the number of patients included in the study was 170, a smaller number than in this study. This shows that clustering is also viable with a small number of patients.

In a study by Igor Marinić et al. about patients with PTSD, a Random Forrest classifier was used to diagnose patients. The Random Forrest classifier is an ensemble learning method that combines multiple decision trees to predict the classes. Patients were diagnosed before the study. The classifier achieved an accuracy between 70% and 80% depending on the attributes that the classifier was used on. These results were considered moderate. In one of the analyses the importance of data from psychiatric scales were identified to be more relevant attributes than the data from structured interviews. It was shown that data about the patient's medical history, social and economical status were relevant, but data about previous and current

symptoms were of greater importance in the model design. The conclusion of the study was that data mining can be useful in clinical practice, more research was advised on larger groups of patients and using several different data mining techniques. In this study PTSD was compared to other disorders, a relevant similarity to this study, although different attributes were used.

Chapter 5

Conclusions

In this chapter we will first concisely answer our subquestions. Based on this we will finely answer the main question of this research.

- How can K-means clustering be applied to analyze the database?
  After the data is preprocessed sixteen dimensions were formed. This included the categorizing and normalizing the dimensions. With the Elbow technique the optimal amount of clusters is found. An implementation of the algorithm in Python was used. Clustering was done both on the complete database as well as on the set of patients with PTSD. Both clusterings were successful.

- How can K-means clustering be used to summarize the database of psychiatric patients in a well-ordered figure?
  We chose to visualize the clustering through bar charts. This method gave a clear view of the clustering.

- What specific clusters can be found with K-means clustering, which differentiate patients with PTSD from patients without PTSD?
  We judged that the optimal amount of clusters was fifteen. There were no clusters that had a 100% purity in having only PTSD patients or no PTSD patients. Before clustering, the database consisted out of 28.17% psychiatric patients with PTSD. All of the fifteen clusters contained a portion of PTSD patients that differed significantly from this number, however only three of them showed interesting results. Cluster 1 and 15 showed an amount of 44% and 40% respectively of patients with PTSD. Cluster 10 showed a low amount of 6% of patients with PTSD. We can conclude from this that in some of the groups there is a heightened or lowered chance of PTSD. However, these results are not significant enough to make conclusions for diagnosis. This means that based on clustering on the chosen sixteen dimensions we can not identify PTSD patients with enough certainty.

- What specific clusters can be found with K-means clustering, within the group of patients with PTSD, that could indicate certain subgroups?
  Several subgroups were found with the clustering. In clusters 1, 4, 6, 7 and 14 clear subgroups were formed by the algorithm.

- **How can K-means clustering, through analyzing and visualizing a database of psychiatric patients, contribute to identifying patients with PTSD?**

K-means clustering has proven to be a useful tool to visualize and analyze the database in this research. However, it is not clear yet how this can exactly contribute to the diagnostic procedure of PTSD. We can conclude that the dimensions we used were not sufficient to differentiate PTSD patients from non-PTSD patients, but was more successful in differentiating between certain subgroups within PTSD. These subgroups could benefit from targeted interventions. More research on these subgroups is advised.

Chapter 6

Discussion

We have shown that some significant subgroups regarding lifetime prevalence appear when clustering is done on the group of PTSD patients. We mainly found that mood disorders and personality disorders form large subgroups in the database of PTSD cases. Further research can be done into looking more closely at these subgroups and trying to compare them to subgroups that appear in the literature.

The clustering yielded no significant results for age. A reason for this may have been that it was the only continuous variable in the database. The age attribute was normalized to have an equal weight, but because of this most values were to close to each other to differentiate with the clustering algorithm. Another option would have been to divide the age variable into different categories, although this would have given the age variable a slightly bigger weigh in the clustering.

It must be taken into account that all of the data is based upon the psychiatrists perspective. All of the diagnoses were carried out by the psychiatrists of Pro Persona Nijmegen. Even though DSM is used, there may be variation in judgment between different psychiatrists. The clustering was done on these diagnoses in the form of lifetime prevalence. So if a patient once has had a disorder in the 'mood disorder' category, he will score a 'one' for this attribute. In this research it can not be seen whether two disorders have occurred at the same moment in time. This is a shortcoming of this research. A solution for further research is that the disorders a patient has, are taken at a point in time.

Some weaknesses of K-means clustering have been named in literature(Berhkin et al.). K-means is sensitive to outliers. We think that this was not an issue in our research, because all data was normalized and mostly categorical data was used. Another shortcoming of the K-means algorithm is that it does not always work well with categorical data. The clustering task could have been addressed by other clustering algorithms, such as K-medoid or hierarchical clustering methods. In further research it is interesting to see if

these algorithms yield better results. In biomolecular research it was noted that, in classification algorithms, it is more important how the attributes are encoded than the specific method used(Tong W et al.). The same may hold for clustering algorithms.

In conclusion, by using a k-means clustering technique, we were able to describe features that may have clinical significance for PTSD. These results support the idea that data mining techniques can be helpful in everyday clinical diagnosing in the future. This research served as a proof-of-concept that clustering can yield interesting results. Research is needed to look into other data mining techniques and the use of different attributes. We think that looking at specific symptoms can yield better results than looking at specific disorders.

Chapter 7

Bibliography

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Berkhin, P. (2006). A Survey of Clustering Data Mining Techniques. In Kogan, J. Nicholas, C. Tebouile, M. (ed.), *Grouping Multidimensional Data.* (pp. 25-71). Berlin Heidelberg: Springer.

Bloom, F. Nelson, C. and Lazerson, A. (2001). *Brain, Mind and Behavior*. New York: Worth Publishers.

Brady, K. Killeen, T. Brewerton, T. and Lucerini, S. (2000). Comorbidity of psychiatric disorders and posttraumatic stress disorder. *Journal of Clinical Psychiatry* 61(7):22-32.

Creamer, M. Parslow, R. O'Donnell, M. Elliott, P. et al. (2008). A predictive screening index for posttraumatic stress disorder and depression following traumatic injury. *Journal of Consulting and Clinical Psychology* 76(6):923-32.

Dadić-Hero, E. Torić, I. Ruzić, K. Medved, P. Graovac, M. (2009). Comorbidity - a troublesome factor in PTSD treatment. *Psychiatria Danubina* 21(3):420-4.

Ditlevsen, D. and Elklit, A. (2010).The combined effect of gender and age on post traumatic stress disorder: do men and women show differences in the lifespan distribution of the disorder? *Annals of General Psychiatry*, 9(32):1-12.

Haskell, S. Gordon, K. Mattocks, K. Duggal, M. Erdos, J. Justice, A. and Brandt, C. (2010). Gender differences in rates of depression, PTSD,

Pain, Obesity, and Military Sexual Trauma Among Connecticut War Veterans of Iraq and Afghanistan. *Journal of Women's Health*, 19(2):267-271.

Lee, H. Malaspina, D. Ahn, H. Perrin, M. Mark, G. Kleinhaus, O. Harlap, S. Goetz, R. and Antonius, D. (2011). Paternal age related schizophrenia (PARS): latent subgroups detected by k-means clustering analysis. *Schizophrenia Research.* 128(1-3): 143–149.

Madhulatha, T. (2012). An overview on Clustering Methods. *IOSR Journal of Engineering.* 2(4):719-725

Marinic, I. Supek, F. Kovacic, Z. Rukavina, L. Jendricko, T. Kozaric-Kovavic, D. (2007). Posttraumatic stress disorder: diagnostic data analysis by data mining methodology. Croatian Medical Journal, 48(2),185-197.

Norris, F. Friedman, M. Watson, P. Byrne, C. Diaz, E. Kaniasty, K. (2002). *60,000 disaster victims speak: Part I. An empirical review of the empirical literature.* Psychiatry 65, 207–239.

O'Donnell, M. Creamer, M. Pattison, P. (2004). Posttraumatic stress disorder and depression following trauma: understanding comorbidity. *American Journal of Psychiatry.* 161(8):1390-6.

Saxe, G. and Wolfe, J. (1999). Gender and posttraumatic stress disorder. In P.A. Saigh & J. D. Bremner (Eds.), *Posttraumatic stress disorder: A comprehensive text.* (pp. 160-182). Boston

Shansky, R. (2015). Sex differences in PTSD resilience and suspectibility: Challenges for animal models of fear learning. *Neurobiology of Stress*,1(1), 60-65.

Staggs, S. (2013). Symptoms & Diagnosis of PTSD. Psych Central. Retrieved on January 25, 2015, from http://psychcentral.com/lib/symptoms-and-diagnosis-of-ptsd/000158

Tolin, D. and Foa, E. (2006). Sex differences in trauma and posttraumatic stress disorder: a quantitative review of 25 years of research. *Psychological Bulletin*, 132(6):959-992.

Tong, W. Hong, H. Fang, H. Xie, Q. Perkins, R. (2003). Decision forest: combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences.* 43:525–31.

Tovar, D. Cornejo, E. Xanthopoulos, P. Guarracino, M. and Pardalos,

P. (2012). Data mining in Psychiatric Research. In F. Kobeissy (ed.). *Psychiatric Disorders: Methods and Protocols, Methods in Molecular Biology.* Volume 829 (pp. 593-602). New York: Humana Press.

What is Posttraumatic Stress Disorder? (2011). Brochure American Psychiatry Association. Arlington: American Psychiatry Association.

Wang, J. Zhou, B. and Yan, R. (2012). Benefits and Barriers in Mining the Healthcare Industry Data. *International Journal of Strategic Decision Sciences*, 3(4):51-67

Chapter 8

Appendix

8.1 A

DSM-IV-TR Criteria for Posttraumatic Stress Disorder:
A. The person has been exposed to a traumatic event in which both of the following were present:
(1) The person experienced, witnessed, or was confronted with an event or events that involved actual or threatened death or serious injury, or a threat to the physical integrity of self or others.
(2) The person's response involved intense fear, helplessness, or horror. Note: In children, this may be expressed instead by disorganized or agitated behavior.

B. The traumatic event is persistently reexperienced in one (or more) of the following ways:
(3) Recurrent and intrusive distressing recollections of the event, including images, thoughts, or perceptions. Note: In young children, repetitive play may occur in which themes or aspects of the trauma are expressed.
(4) Recurrent distressing dreams of the event. Note: In children, there may be frightening dreams without recognizable content.
(5) Acting or feeling as if the traumatic event were recurring (includes a sense of reliving the experience; illusions, hallucinations, and dissociative flashback episodes, including those that occur on awakening or when intoxicated). Note: In young children, trauma-specific reenactment may occur.
(6) Intense psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event.
(7) Physiological reactivity on exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event.

C. Persistent avoidance of stimuli associated with the trauma and numbing of general responsiveness (not present before the trauma), as indicated

by three (or more) of the following:
(8) Efforts to avoid thoughts, feelings, or conversations associated with the trauma
(9) Efforts to avoid activities, places, or people that arouse recollections of the trauma
(10) Inability to recall an important aspect of the trauma
(11) Markedly diminished interest or participation in significant activities
(12) Feeling of detachment or estrangement from others
(13) Restricted range of affect (e.g., unable to have loving feelings)
(14) Sense of a foreshortened future (e.g., does not expect to have a career, marriage, children, or a normal lifespan)

D. Persistent symptoms of increased arousal (not present before the trauma), as indicated by two (or more) of the following:
(1) Difficulty falling or staying asleep
(2) Irritability or outbursts of anger
(3) Difficulty concentrating
(4) Hypervigilance
(5) Exaggerated startle response

E. Duration of the disturbance (symptoms in Criteria B, C, and D) is more than 1 month.

F. The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.

8.2 B

A list of all disorders per category. The DSM-code is listed for each disorder.

1 SCHIZOFRENIE E.A. PSYCHOTISCHE STOORNISSEN
295.30 Eenmalige episode gedeeltelijk in remissie
295.30 Episodisch met restsymptomen tussen de episoden
295.30 Korter dan 1 jaar na begin eerste actieve fase
295.30 Ononderbroken
295.30 Paranoïde type
295.40 Schizofreniform Met gunstige prognostische kenmerken
295.40 Schizofreniform Zonder gunstige prognostische kenmerken
295.70 Schizo-affectieve stoornis Bipolaire type
295.70 Schizo-affectieve stoornis Depressieve type
295.70 Schizo-affectieve stoornis
295.90 Ongedifferentieerde type
297.1 Waanstoornis
297.1 Achtervolgingstype
297.1 Gemengd type
297.1 Niet gespecificeerd type
297.1 Somatisch type
297.3 Geïnduceerde psychotische stoornis
298.8 Kortdurende psychotische stoornis
298.8 Kortdurende psychotische stoornis
298.8 Zonder duidelijke stressveroorzakene factor(en)
298.9 Psychotische stoornis NAO
2 STEMMINGSSTOORNISSEN
311 Depressieve stoornis NAO
311 Depressieve stoornis NAO
296.20 Depressie in engere zin, eenmalige episode, niet-gespecifice
296.20 Niet-gespecificeerd
296.21 Depressie in engere zin, eenmalige episode, licht
296.22 Depressie in engere zin, eenmalige episode, matig
296.23 Depressie in engere zin, eenmalige episode, ernstig zonder p
296.24 Depressie in engere zin, eenmalige episode, ernstig met psyc
296.25 Gedeeltelijk in remissie
296.26 Depressie in engere zin, eenmalige episode, volledig in remi
296.26 Volledig in remissie
296.30 Depressie in engere zin, recidiverend, niet-gespecificeerd
296.31 Depressie in engere zin, recidiverend, licht
296.32 Depressie in engere zin, recidiverend, matig
296.33 Depressie in engere zin, recidiverend, ernstig zonder psycho
296.34 Depressie in engere zin, recidiverend, ernstig met psychotis
296.35 Gedeeltelijk in remissie
296.36 Volledig in remissie

300.4 Dysthyme stoornis
300.4 Dysthyme stoornis
300.4 Laat begin
300.4 Met atypische kenmerken
300.4 Vroeg begin
293.83 Met depressieve kenmerken
293.83 Stemmingsstoornis door .... [Vermeld de algemene lichamelijk
296.90 Stemmingsstoornis NAO
292.84 Stemmingsstoornis door cocaïne
292.84 Stemmingsstoornis door een ander (of onbekend) middel
3 BIPOLAIRE STOORNISSEN
296.01 Licht
296.03 Bipolaire I stoornis, eenmalige manische episode, ernstig zo
296.05 Gedeeltelijk in remissie
296.40 Laatste episode hypomaan
296.40 Niet-gespecificeerd
296.43 Ernstig zonder psychotische kenmerken
296.44 Bipolaire I stoornis, meest recente episode manisch, ernstig
296.46 Volledig in remissie
296.52 Matig
296.53 Bipolaire I stoornis, meest recente episode depressief, erns
296.63 Ernstig zonder psychotische kenmerken
296.7 Laatste episode niet-gespecificeerd
296.80 Bipolaire stoornis NAO
296.89 Bipolaire II stoornis
296.89 Depressief
296.89 Hypomaan
301.13 Cyclothyme stoornis
4 ANGSTSTOORNISSEN
300.21 Paniekstoornis met agorafobie
300.22 Agorafobie zonder anamnese met paniekstoornis
300.23 Gegeneraliseerd
300.23 Sociale fobie
300.29 Bloed-injectie-verwonding type
300.29 Diertype
300.29 Natuurtype
300.29 Overig type
300.29 Situationeel type
300.29 Specifieke fobie
300.29 Specifieke fobie
300.3 OCS Met gering inzicht
300.3 Obsessieve-compulsieve stoornis
300.00 Angststoornis NAO
293.84 Met paniekaanvallen

300.01 Paniekstoornis zonder agorafobie
300.02 Gegeneraliseerde angststoornis
300.1 Paniekstoornis zonder agorafobie
308.3 Acute stress-stoornis
5 PERSOONLIJKHEIDSSTOORNISSEN
301.0 Paranoïde persoonlijkheidsstoornis
301.20 Schizoïde persoonlijkheidsstoornis
301.22 Schizotypische persoonlijkheidsstoornis
301.4 Obsessieve-compulsieve persoonlijkheidsstoornis
301.50 Theatrale persoonlijkheidsstoornis
301.6 Afhankelijke persoonlijkheidsstoornis
301.7 Antisociale persoonlijkheidsstoornis
301.81 Narcistische persoonlijkheidsstoornis
301.82 Ontwijkende persoonlijkheidsstoornis
301.84 Passief-agressieve persoonlijkheid
301.9 Persoonlijkheidsstoornis nao
301.9 Persoonlijkheidsstoornis NAO
302.22 Schizotypische persoonlijkheidsstoornis
6 BORDERLINE PERSOONLIJKHEIDSSTOORNIS
301.83 Borderline persoonlijkheidsstoornis
7 MIDDELEN GEBONDEN STOORNISSEN
303.00 Alcoholintoxicatie
303.90 Alcoholafhankelijkheid
304.00 Afhankelijkheid van opiaten
304.10 Afhankelijkheid van sedativum, hypnoticum of anxiolyticum
304.20 Cocaïne-afhankelijkheid
304.30 Cannabisafhankelijkheid
304.40 Amfetamineafhankelijkheid
304.80 Afhankelijkheid van verschillende middelen
304.90 Afhankelijkheid van een ander (of onbekend) middel
305.00 Misbruik van alcohol
305.00 Misbruik van alcohol
305.10 Nicotine-afhankelijkheid
305.20 Misbruik van cannabis
305.30 Misbruik van hallucinogeen
305.40 Misbruik van sedativum, hypnoticum of anxiolyticum
305.60 Misbruik van cocaine
305.70 Misbruik van amfetamine
305.90 Misbruik van een ander (of onbekend) middel
291.89 Stemmingsstoornis door alcohol
292.11 Met wanen (amfetamine/cannabis
292.12 Met hallucinaties
292.12 Met hallucinaties
292.89 Opioïde-intoxicatie

292.9 Aan cafeïne gebonden stoornissen NAO
292.9 Aan cannabis gebonden stoornissen NAO
292.9 Aan opioïde gebonden stoornissen NAO
8 DISSOCIATIEVE STOORNISSEN
300.12 Dissociatieve amnesie
300.13 Dissociatieve fugue
300.14 Dissociatieve identiteitsstoornis
300.15 Dissociatieve stoornis NAO
300.6 Depersonalisatiestoornis
9 PERVASIEVE ONTWIKKELINGSSTOORNISSEN
299.00 Autistische stoornis
299.80 Pervasieve ontwikkelingsstoornis NAO
299.80 Stoornis van Asperger
299.80 Stoornis van Rett
10 AANDACHTTEKORT- EN GEDRAGSSTOORNISSEN
314.00 Gedeeltelijk in remissie
314.00 Overwegend onoplettendheid type
314.01 Aandachtstekortstoornis met hyperactiviteit, gecombineerde t
314.01 Gecombineerde type
314.01 Gedeeltelijk in remissie
314.01 Overwegend hyperactief-impuls type
314.9 Aandachtstekortstoornis met hyperactiviteit NAO
313.81 Oppositioneel-opstandige gedragsstoornis
11 SLAAPSTOORNISSEN
307.42 Insomnia in samenhang met (vermeld de As I of As II stoornis
307.42 Primaire dyssomnia
307.44 Primaire hypersomnia
307.45 Niet gespecificeerd type
307.45 Slaapstoornis gebonden aan de circadiane ritmiek
307.45 Uitgestelde slaapfase type
307.46 Pavor nocturnus
307.46 Slaapwandelen
307.47 Dyssomnia NAO
307.47 Nachtmerries
307.47 Parasomnia NAO
12 SOMATOFORME STOORNISSEN
300.11 Conversiestoornis
300.11 Met gemengd beeld
300.11 Met motorisch symptoom of uitvalverschijnselen
300.11 Met sensorisch symptoom of uitvalverschijnselen
300.11 Met toevallen of convulsies
300.7 Hypochondrie
300.7 Met gering inzicht
300.7 Stoornis in de lichaamsbeleving

300.81 Ongedifferentieerde somatoforme stoornis
300.81 Somatisatiestoornis
300.82 Ongedifferentieerde somatoforme stoornis
300.82 Somatoforme stoornis NAO
307.80 Pijnstoornis Chronisch
307.80 Pijnstoornis Gebonden aan psychische factoren
307.89 Pijnstoornis Chronisch
307.89 Pijnstoornis Gebonden aan zowel psychische factoren als een somatische aa
13 NAGEBOOTSTE STOORNISSEN
300.19 Met hoofdzakelijk lichamelijke verschijnselen en klachten
300.16 Nagebootste stoornis NAO
14 EETSTOORNISSEN
307.1 Anorexia nervosa
307.1 Anorexia nervosa
307.1 Beperkende type
307.1 Purgerende type
307.1 Vreetbuien
307.50 Eetstoornis NAO
307.51 Bulimia nervosa
307.51 Niet-purgerende type

8.3 C

The code that was used to do the computations for the k-means clustering. First the code for the clustering itself:

```
#!/usr/bin/env python
import scipy.io as scio
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import numpy as np

#loading patientdata
sets = ['niet-ptss', 'ptss']

hoofdgroepen = scio.loadmat("variant1.mat")
print scio.whosmat("variant1.mat")
data = hoofdgroepen['merged']



#Start Values
nHoofdgroepen = 16
nClusters = 15

if nHoofdgroepen == 16:
    data= data[:,0:nHoofdgroepen]
elif nHoofdgroepen == 8:
    data = data[:,[0,1,2,3,5,6,7,8]]

#Kmean clustering
k_means = KMeans(n_clusters=nClusters, init='k-means++',
    max_iter= 10000, n_init=1000, tol = 0.00005)
k_means.fit(data)

print k_means.inertia_

results = {}
results['centers'] = k_means.cluster_centers_
results['labels'] = k_means.labels_
results['inertia'] = k_means.inertia_


scio.savemat("results.mat",results)
print scio.whosmat("results.mat")
```

The code for building the graph:

```
#!/usr/bin/env python
import scipy.io as scio
import matplotlib.pyplot as plt
import numpy as np

np.set_printoptions(threshold=np.nan)

#loading patientdata
resultaten = scio.loadmat("results.mat")
hoofdgroepen = scio.loadmat("variant1.mat")
trueClass = hoofdgroepen['trueClass']
clusters = resultaten['centers']
labels = resultaten['labels']

nHoofd = len(clusters[0])

#creating barplot from cluster centers.
#colors = ['CornSilk', 'Tan']
colors = ['LightBlue','DodgerBlue']

ind = np.arange(0,nHoofd)
widthbar = 1

def percPtss(cluster):
    nonptss = 0.0
    ptss = 0.0
    sum = 0
    for i in range(0,len(labels[0])):
        if labels[0,i] == cluster:
            sum += 1
            if trueClass[0,i] == 1:
                ptss += 1
            elif trueClass[0,i] == 0:
                nonptss += 1
    return sum,ptss / (nonptss+ptss)



def autolabel(rects,i):
    # attach some text
    a = 0
    for rect in rects:
        height = rect.get_height()
        if a == 1:
            ax[i/4,i % 4].text(rect.get_x()+rect.get_width()/2.,
                1.05*height, '%.2f'%(height*80),
        ha='center', va='bottom', fontsize='6')
        else:
            ax[i/4,i % 4].text(rect.get_x()+rect.get_width()/2.,
                1.05*height, '%.2f'%height,
        ha='center', va='bottom', fontsize='6')
        a += 1
```

```
fig , ax = plt.subplots(4,4)

for i in range(0,len(clusters)):
#for i in range(0,3):

    sum, perc = percPtss(i)
    rects = ax[i/4,i % 4].bar(ind, clusters[i], widthbar, color
        = colors)
    autolabel(rects,i)

    ax[i/4,i % 4].set_title('Cluster' + str(i+1) + '\n size: ' +
         str(sum) + '   PTSD: ' + str('%.2f'%(perc*100))+'%',
        fontsize = 10)
    ax[i/4,i % 4].set_xticks(ind+widthbar/2.0)

    ax[i/4,i % 4].set_ylim(0,1.2)
    ax[i/4,i % 4].set_xlim(0,nHoofd)


Labels = ['geslacht'
,'leeftijd'
,'psychotische stoornis'
,'stemmingstoornis'
,'bipolaire stoornis'
,'angststoornis(zonder PTSS)'
,'persoonlijkheidsstoornis'
,'borderline persoonlijkheidsstoornis'
,'middelen gebonden stoornis'
,'dissociatieve stoornis'
,'pervasieve ontwikkelingsstoornis'
,'aandachttekort- \n en gedragsstoornis'
,'slaapstoornis'
,'somatoforme stoornis'
,'nagebootste stoornis'
,'eetstoornis']


for i in range(0,3):
    plt.setp([a.get_xticklabels() for a in ax[i, :]], visible=
        False)
    plt.setp([a.get_yticklabels() for a in ax[:, i+1]], visible=
        False)

xTickMarks = [str(Labels[i-1]) for i in range(1,nHoofd+1)]
for i in range(0,4):
    xtickNames = ax[3,i].set_xticklabels(xTickMarks)

    plt.setp(xtickNames, rotation=90, fontsize=10)

x= ax[2,3].set_xticklabels(xTickMarks)
plt.setp(x,rotation =90, fontsize =10)
```

```
plt.setp(ax[2,3].get_xticklabels(), visible = True)
ax[3,3].set_visible(False)

plt.show()
```

The code for determining the optimal amount of clusters:

```
#!/usr/bin/env python
import scipy.io as scio
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import numpy as np

#Start Values
nHoofdgroepen = 16

hoofdgroepen = scio.loadmat("variant1.mat")
print scio.whosmat("variant1.mat")
data = hoofdgroepen['merged']

#data= data[:,0:nHoofdgroepen]
data[:,1] = data[:,1]/80.0
data = data[:,[0,1,2,3,5,6,7,8]]



#determine optimal amount of clusters by inertia

a = np.arange(1,31)
inertiaClusters = np.empty([30,2])

inertiaClusters[:,0] = a

print inertiaClusters

#Kmean clustering
#k_means = KMeans(n_clusters=nClusters, init='k-means++',
    max_iter= 1000, n_init=100,tol=0.0005)
#k_means.fit(data)

for n in range(1, 31):
    k_means = KMeans(n_clusters=n, init='k-means++', max_iter=
        1000, n_init=10, tol=0.00005)
    k_means.fit(data)
    inertiaClusters[n-1,1] = k_means.inertia_

plt.figure().suptitle('Cluster score')
plt.ylabel('Sum of squared errors (SSE)')
plt.xlabel('Number of clusters')
plt.scatter(inertiaClusters[:,0], inertiaClusters[:,1])
plt.show()
```