SOM Group 00

Report for assignment 1 in Self-Organizing Systems: SOM

Clemens Fickl
Technical University of Vienna
Karlsplatz 13
1040, Vienna
e1129315@student.tuwien.ac.at

Marten Sigwart
Technical University of Vienna
Karlsplatz 13
1040, Vienna
e1638152@student.tuwien.ac.at

1. INTRODUCTION

The dataset we dealt with contained hypothyroid disease records supplied by the Garavan Institute and J.Ross Quinlan, New South Wales Institute, Sydney, Australia. The dataset can be found under http://www.openml.org/d/57. Our goal was to train a Self-Organizing Map (SOM) of the data using the Java SOMToolbox developed at the University of Vienna to derive and extract relevant information. More specifically, we were especially interested in finding clusters, which might explain certain correlations between attributes as well as correlations of attributes and the class distribution.

2. CHARACTERISTICS OF DATA SET

The dataset consists of 3772 instances with each 30 attributes. Each instance represents a single patient. Each attribute represents a specific medical record of that patient. The attributes are described in Table 1.

Furthermore each instance is assigned to one of four classes. These classes are:

- primary hypothyroid
- compensated hypothyroid
- secondary hypothyroid
- negative

Basically, each class represents the final diagnosis of a patient. Class "negative" (3481 instances) means the patient has not been diagnosed with hypothyroidism. Class "compensated hypothyroid" (194 instances) represents patients whose peripheral thyroid hormone levels are within the normal range, but tyhroid stimulating hormone (TSH) is mildly elevated. According to http://patient.info/doctor/subclinical-hypothyroidism the condition is quite common occurring in 3-8% of the population, and carries a risk of progression to primary hypothyroidism. Class "primary hypothyroid" (94 instances) represents patients diagnosed with primary hypothyroidism a failure of the thyroid gland itself. Class "Secondary hypothyroid" (2 instances) is assigned to patients with diagnosis secondary hypothyroidism, a condition in which not the thyroid gland itself failed, but which is instead caused by a failure of the pituitary gland which releases thyroid stimulating hormone (TSH).

2.1 Missing Values

Overall, the dataset contains 6064 missing values. This makes up ca. 5.4% of the total number of values. More specifically we have 150 (4%) missing values for attribute 'Sex', 369 (10%) for attribute 'TSH', 769 (20%) for 'T3', 231 (6%) for 'TT4', 387 (10%) for 'T4U', 385 (10%) for 'FTI' and 3772 (100%) for 'TBG'. We note, the number of missing values is quite high. Therefore we will need a suitable missing value replacement strategy. These will be described in section 3.3.

2.2 Hypothesis about data, clusters and classes

We assume training a SOM of the described data set will reveal a couple of interesting relationships. Firstly, we are convinced that we will see a clear class distribution of our data in the SOM. Our assumption is that especially data points of class "primary hypothyroid" will lie closely together as we have enough data points of that class to reveal similarities between instances. We are not sure if we can also see a clear structure for data points of class "secondary hypothyroid" as we only have 2 of those instances. Class "compensated thyroid" instances are expected to form a kind of connection between "negative" instances and "primary hypothyroid" instances, as compensated hypothyroidism has a tendency to progress to primary hypothyroidism. We assume that the class distribution in the SOM will be mainly determined by the continuous measurement values contained in the data, in particular, we assume attribute "TSH" to have a strong impact on the final diagnosis.

Further, we assume that boolean attributes like "psych", "pregnant", "sick", etc. will not show a great influence on the class distribution/cluster structure of the SOM, as we are convinced that we do not have enough data to reveal any relations between rare events like a pregnancy and the probability of hypothyroidism diagnosis. Concerning the cluster structure, we are sure that the SOM will reveal mainly two big clusters dividing data points into male and female, and that the age of the patients will also be visible in the clusters

Table 1: Attributes

Column	Type	Description
Age	Integer	Age of the patient
Sex	Enumeration	Sex of the patient (M = male, F = female)
On thyroxine	Boolean	True, if patient is on thyroxin - a drug often used by people with hypothyroidism.
Sick	Boolean	True, if patient is currently sick.
Pregnant	Boolean	True, if patient is currently pregnant.
Thyroid surgerey	Boolean	True, if patient has had thyroid surgery
I131 treatment	Boolean	True, if patient is under I131 treatment
Query hypothyroid	Boolean	True, if patient said he/she has hypothyroid
Query hyperthyroid	Boolean	True, if patient said he/she has hyperthyroid
Lithium	Boolean	True, if patient is taking lithium
Goitre	Boolean	True, if patient has symptoms of goitre (a swelling of the neck from big thyroid gland)
Tumor	Boolean	True, if patient has a tumor
Hypopituitary	Boolean	True, if patient has hypopituitarism (decreased secretion of one or more hormone)
Psych	Boolean	True, if patient has psychological problems
TSH measured	Boolean	True, if the TSH level has been measured (will be in the next column)
TSH	Decimal number	Measured TSH (Thyroid Stimulating Hormone) level of the patient
T3 measured	Boolean	True, if the T3 level has been measured (will be in the next column)
T3	Decimal number	Measured T3 level of the patient
TT4 measured	Boolean	True, if the TT4 level has been measured (will be in the next column)
TT4	Decimal number	Measured TT4 level of the patient
T4U measured	Boolean	True, if the T4U level has been measured (will be in the next column)
T4U	Decimal number	Measured T4U level of the patient
FTI measured	Boolean	True, if the FTI level has been measured (will be in the next column)
FTI	Decimal number	Measured FTI level of the patient
TBG measured	Boolean	True, if the TBG level has been measured (will be in the next column)
TBG	Decimal number	Measured TBG level of the patient
Referral Source	Enumeration	Referral source

in some way. We think that attributes with rare events as described before will perish in these bigger clusters, and will not show a big impact. Lastly, we suspect that a smaller cluster will also emerge for data points with assigned class "primary hypothyroid".

3. PREPROCESSING

To train a SOM, the data has to be in a format that can be interpreted by the SOMToolbox. A SOM can only be trained on data of the SOMLib file format. The SOMLib file format consists of one file with file extension .vec, containing the input vectors to be used for the training of the SOM, and one file with file extensions .tv, containing a template vector providing the attribute structure of the data. For more information on the SOMLib file format check out `http://www.ifs.tuwien.ac.at/dm/somtoolbox/somlibFileFormat.html`.
Another requirement for training the SOM is that all attribute values are in numerical form. Only the class values can remain in nominal (categorical) form. As last requirement the data must not contain any missing values. Additionally, it is important to perform some kind of normalization on the data, as any unnormalized data will be very hard to compare.

The original data was in the ARFF format, contained lots of nominal attributes and contained also quite a number of missing values. Hence, to get our data ready for the SOM Toolbox, extensive preprocessing had to be done. The preprocessing was done in the following steps:

1. Transformation of nominal into numerical attributes
2. Outliar detection
3. Missing Value Replacement
4. Data Format Conversion
5. Normalization

3.1 Transformation of nominal to numerical attributes

The first step in our preprocessing was the conversion of nominal attributes to numerical values. In the most cases this affected the boolean values, where we converted 't' to 1 and 'f' to 0. This was analogically done for the sex-attribute, where the values 'M' became 1 and 'F' became 0.
In this step we also realized that the attribute 'referral source' contains 5 different discrete values. Here we thought about coding this into 5 columns, but on further investigation, we realized, that this attribute has no impact on the class and is just for traceability in the data set.

3.2 Outlier detection

While investigating on the data set we found one outlier in the age-attribute with a value of 455, which makes no sense. After a discussion about deleting or replacing the value we decided to replace it with the value 45, since we assumed that this was an input error, where someone unintentionally mistyped a second 5.

3.3 Missing Value Replacement

The next step in the preprocessing was the handling of missing values. Therefore we analyzed all attributes, how many missing values there are and discussed what to do with them. Following the attributes, where we found missing values and a description, what we did with them:

- **Age**: In the age-attribute we found exactly 1 missing value. We decided to delete the whole entry, since it was in the class negative (where most of the data is) and has therefore no big impact on the result.
- **TBG and TBG measured**: The attribute '*TBG*' has no entries (100% missing values) and the attribute '*TBG measured*' only consists of the value 'f' (false). Therefore we decided to delete these two columns completely.
- **Sex**: We had 150 missing values for the sex. They were replaced with mean value of the respectively class.
- **TSH, T3, TT4, T4U, FTI**: These 5 columns consist of measured values for different examinations. Each has between 369 and 769 missing values (which represents 10 to 20 percent of the data samples). Since we realized that the impact of these columns is quite high, we had to find a suitable replacement strategy. We decided to calculate the median values for each of our four classes for each of these 5 attributes and replace the missing values with the corresponding median.

We used MATLAB scripts to perform the more sophisticated missing value replacement. To do so we used the InputDataFileFormatConverter provided by the SOM toolbox to transform the original ARFF data file into an CSV (Comma Seperated Value) File. This could be easily imported into MATLAB using a slightly modified function for importing CSV files importThyroidData. This function would return a header vector containing our attribute names, a data array containing all data in numerical form except the class labels, and a class vector containing the class labels for all data instances. The actual missing value replacement is performed in script SOMPreprocessing.m. Afterwards the modified data is exported again to ARFF Format using the function writeThyroidARFF.m.

3.4 Format Conversion

After replacing the missing values, the ARFF file had to be transformed into the SOMLib data format. We used the InputDataFileFormatConverter provided by the SOM toolbox to do that job. Afterwards the data was ready for normalization and training of the SOM.

3.5 Normalization (Scaling)

For our normalization we used the std-score scaling. On the one hand we assumed that unit length makes no sense (we also got confirmed when trying it out, see Section 4.6). On the other hand, we saw that we have a small amount for outliers (e.g. in the age there is one value with 455 and also in the measurement values there were some quite high numbers), which would destroy the scaling for these attributes.

4. TRAINING AND ANALYSIS

4.1 Reasonably sized "regular" SOM

4.1.1 Parameters

For our "regular" SOM we used the following parameters:

- **randomSeed**: 7
- **Map-Size**: 20x20 (this is about 1/10th of our number of values and seemed to be a good point to start for us)
- **learnRate**: Default value (0.75)
- **numIterations**: 1500 (retrospectively this was a bit low, but in the first tries playing with it it seemed way enough)
- **sigma** (neighborhood radius): Default value (10 for 20x20 map)

4.1.2 Class distribution

After starting the SOMViewer the first thing that leaps out is that our map seems to consist of one very big structure in the middle stretching from top to bottom containing a lot of the input instances (see Figure 1). Also several single units containing lots of data points got mapped to deserted parts of the map, clearly separated from the big cluster (see Section 4.1.4). If we add the pie chart visualization (see Figure 2) we can observe that most instances with class labels "primary hypothyroid" and "secondary hypothyroid" are neatly grouped together at the bottom of the map with only a few exceptions. This confirms our assumptions we made in section 2.2. Data points of class "negative" are spread across the whole map with no striking structures visible. The data points of class "compensated hypothyroid" are scattered all around the map with a fair number of points building a kind of transition between the small cluster of primary and secondary hypothyroid and the big "negative"-cluster in the middle.

4.1.3 Cluster structure analysis

In Figure 3 you can see the result of running the k-means clustering algorithm for five clusters. This shows that the visually big cluster in the middle gets split into 3 clusters and most of the "primary hypothyroid" instances are together in one cluster at the bottom. When comparing this with the result we get for running k-means clustering with only four clusters (see Figure 4), we can actually observe that the cluster containing these data points is actually a sub-cluster of one of the big clusters in the center. Furthermore, data points of class "compensated hypothyroid" do not seem to lie within a specific cluster, but are scattered over the whole map.

4.1.4 Attribute influence

When comparing the class distribution to the distribution of single attributes using the component planes visualization, we can observe a couple of interesting relations. Maybe the most interesting one that leaps out almost immediately is the relation between the class distribution and the distribution of attributes containing specific medical measurement values (TSH, T3, TT4, T4U, FTI). For example, let's look at the component planes for attributes TSH and FTI (see Figure 5 and Figure 6. Blue means high values, red means low values. For TSH we can see that the data points of class

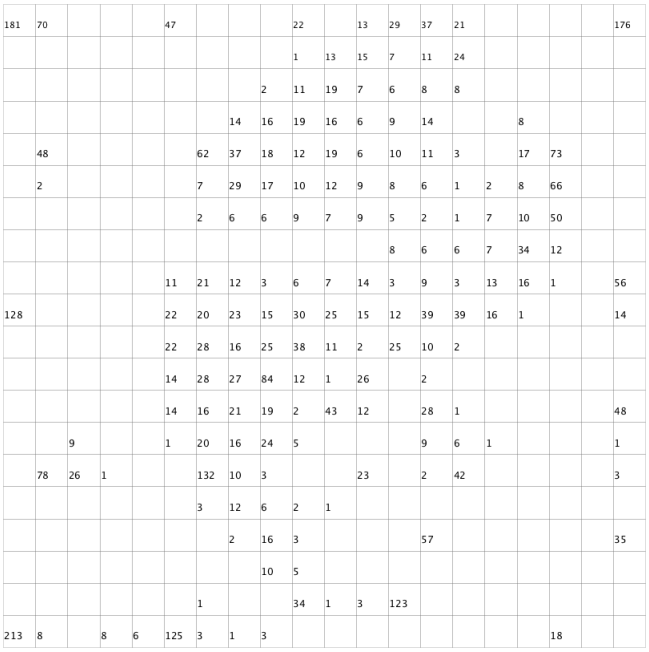

Figure 1: Standard SOM: datapoint distribution

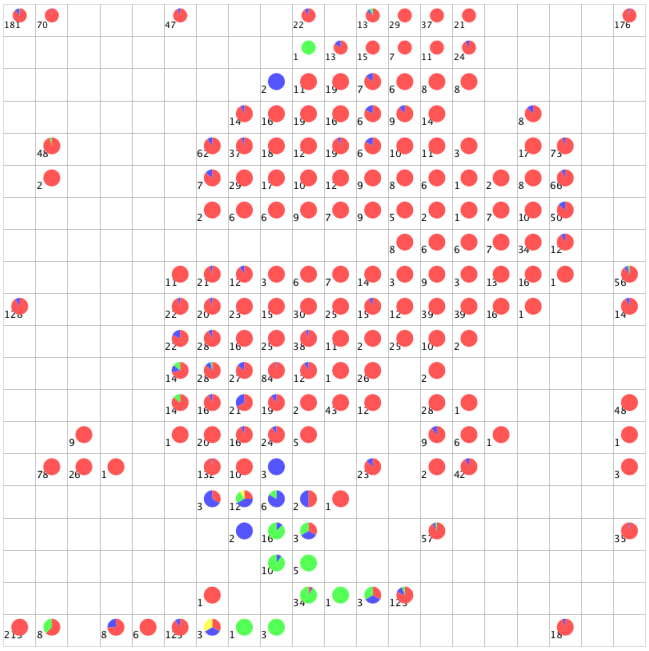

Figure 2: Standard SOM: class distribution

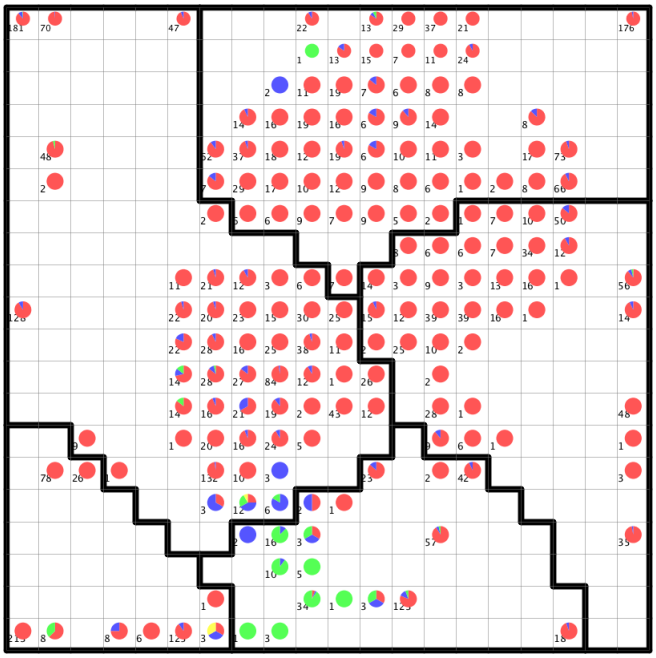

Figure 3: Standard SOM: K-means clustering with 5 clusters

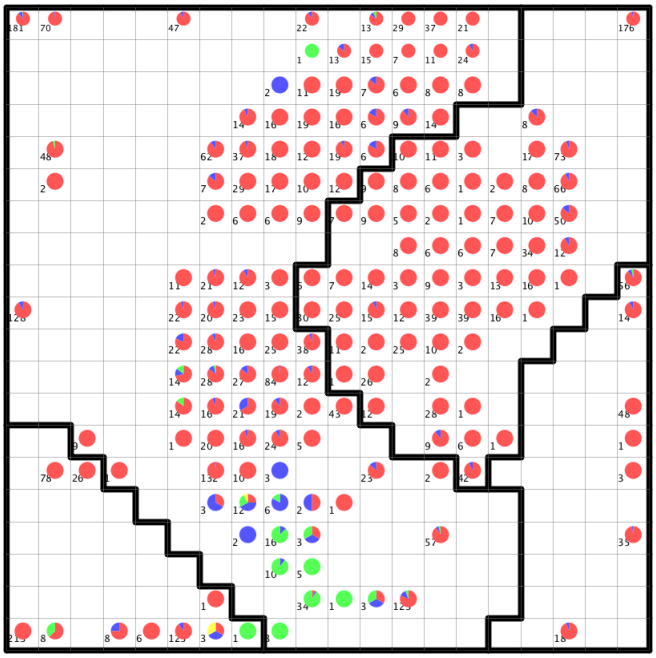

Figure 4: Standard SOM: K-means clustering with 4 clusters

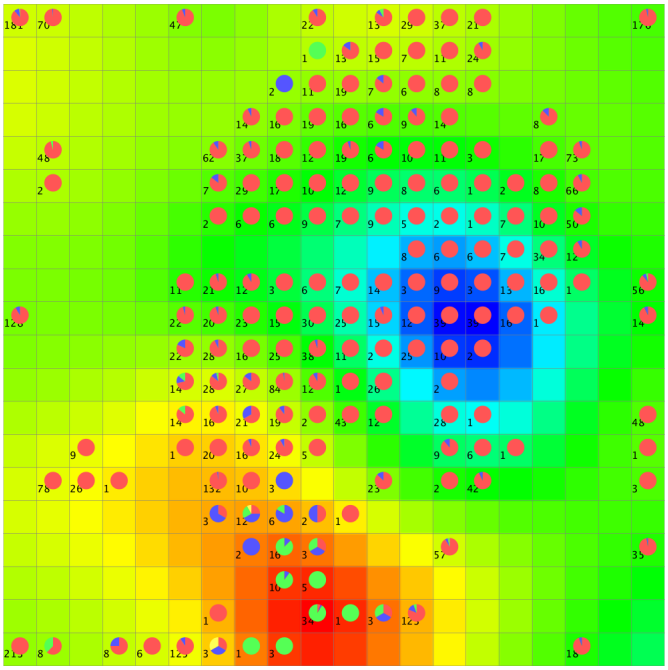

Figure 5: Standard SOM: Component Plane for "FTI"

"primary hypothyroid" lie exactly where the TSH value is high. The same goes for attribute FTI, just that here the data points lie in the area of low FTI values. Accordingly, the same observation can be made for all "measurement" attributes, where the addressed classes lie either in the area of high or low values for these attributes. The only attribute that does not seem to have any influence on the class distribution is attribute T4U. The component plane does not seem to reveal any relationship between the values of T4U and the class distribution (see fig 7. In general, this confirms our assumptions about the data we made in section 2.2.
As already mentioned before we also realized, that most of the boolean attributes seem to be the cause of the single unit clusters scattered across the map. For example, Figure 8 shows the component plane of the attribute "thyroid surgery", where basically all data points with that specific attribute being positive are mapped to the two units on the left of the map.

4.1.5 Quantization errors and topology violations

Looking at the quality measure "quantization errors" as shown in Figure 10, we can see that most units on our map have quite low quantization errors. Only some single cluster units, mostly in the edges of the map, have more quantization errors.
When we look at quality measure "Topographic Error neighborhood - 8 units", as shown in Figure 9, we see that our map has a quite low number of topology violations. This believe is further confirmed by the neighborhood graph with radius 0,8, as shown in Figure 11. The single unit clusters show almost no topology violations and the big structure in the middle is clearly divided into 2 big clusters. Notable is also that even though the data points of class "primary hypothyroid" seem to be part of the big structure, they clearly lie outside of this neighborhood.

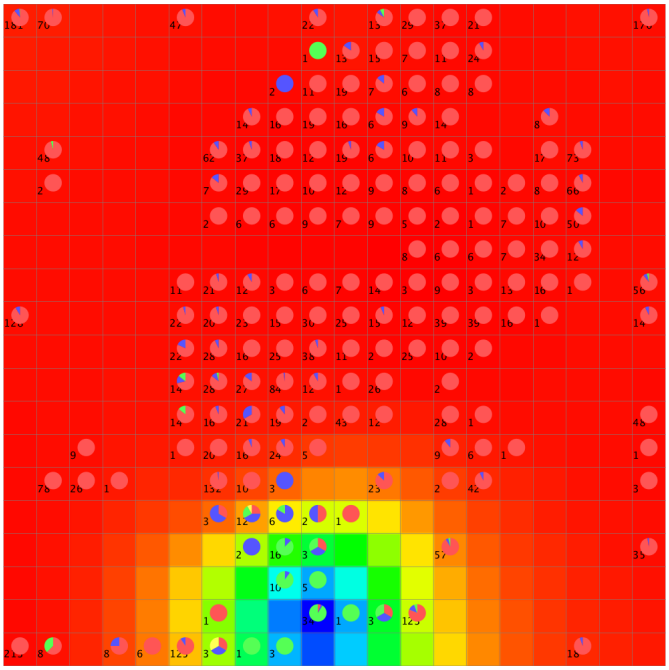

Figure 6: Standard SOM: Component Plane for "TSH"

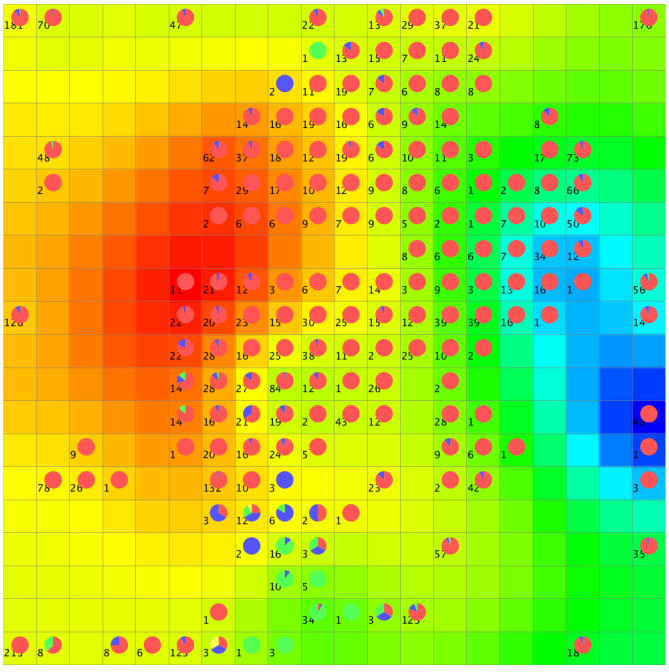

Figure 7: Standard SOM: Component Plane for "T4U"

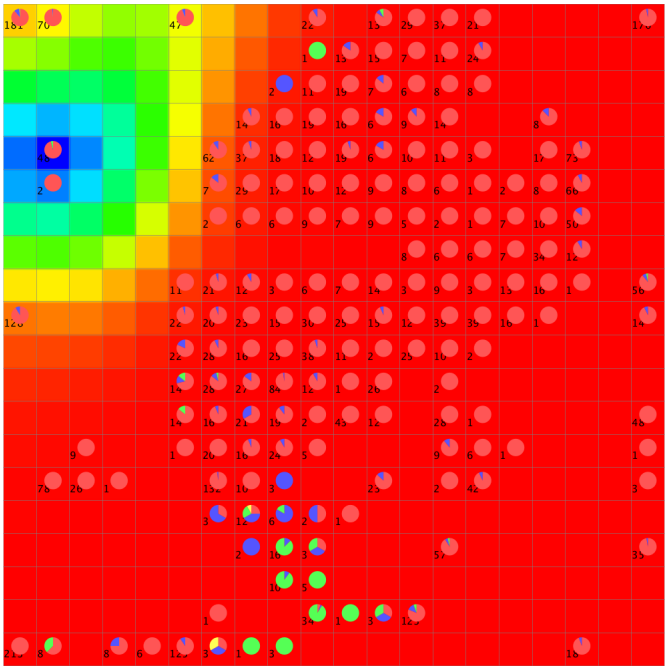

Figure 8: Standard SOM: Component Plane for "thyroid surgery"

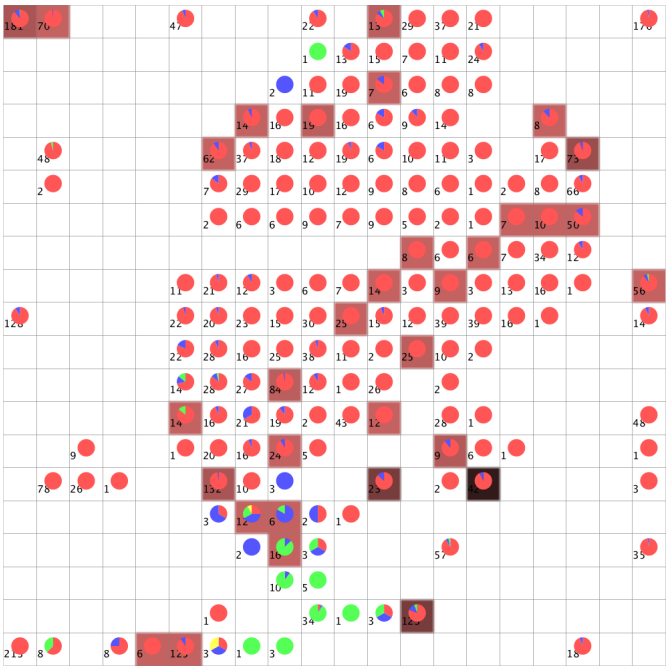

Figure 9: Standard SOM: Topographic Error neighborhood 8 units

We assume that the cluster separation is quite robust and of high quality.

4.2 Different initializations of the SOM

Next we tried out two different initializations of the SOM. One was initialized with a rather high random seed value of 100. The other one was initialized with a rather low value of 2.

High: Let's look at the SOM which was initialized by a random seed 100. Looking at the class distribution (see Figure 12) the first thing that leaps out is that the structure seemingly changed completely. On closer inspection though, we note that instances of class "primary hypothyroid are still mapped together at the bottom of the map. Also they still seem to lie at the edge of one bigger cluster with "negative" instances. The bigger cluster in the middle we saw in our standard SOM is not present anymore. There do seem to be two bigger "negative" clusters, though, which could be the bigger cluster from the standard SOM just divided into two smaller clusters. Looking at the component plane visualization of attribute "age" in Figure 13, we see that the age of the patients seem to have influenced that split of clusters. Furthermore, we can still make out multiple "single unit" clusters, which seem to have been caused by single boolean attributes (like "pregnant", "thyroid surgery", "psych",etc.). The quantization errors and topology violations show very similar results to the standard SOM, so we are not showing them here.

Low: Now let's have a look at the initialization with a random seed of 2. Similarly, as in the initialization with random seed 100, the overall structure of the SOM changed completely (see Figure 14). Again, the bigger cluster from the Standard SOM seems to have split again, this time into 3 distinct clusters. Same as before it is notable that most instances of class "primary hypothyroid" lie again closely

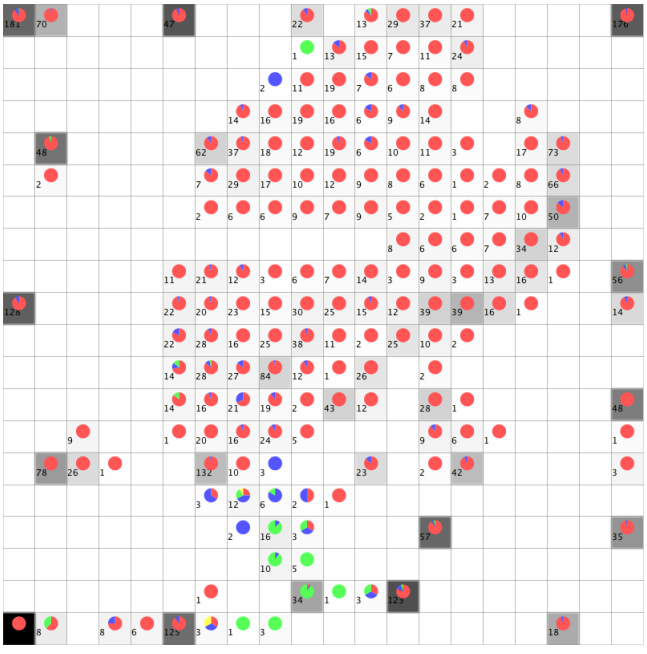

Figure 10: Standard SOM: Quantization Error

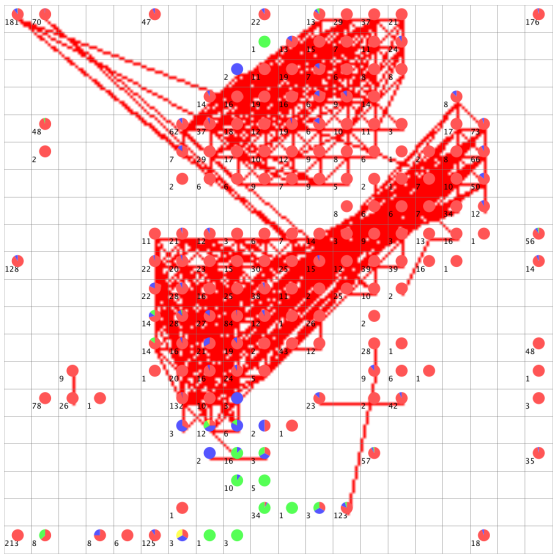

Figure 11: Standard SOM: Neigborhood Graph with radius 0.8

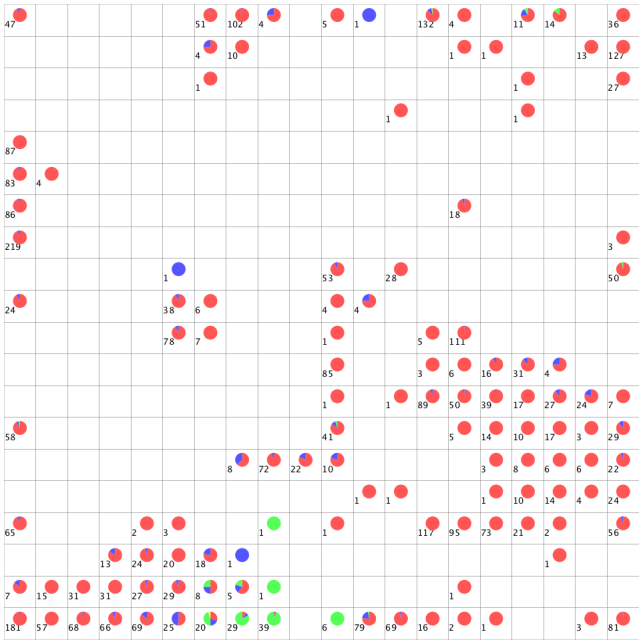

Figure 12: SOM with random seed 100: class distribution

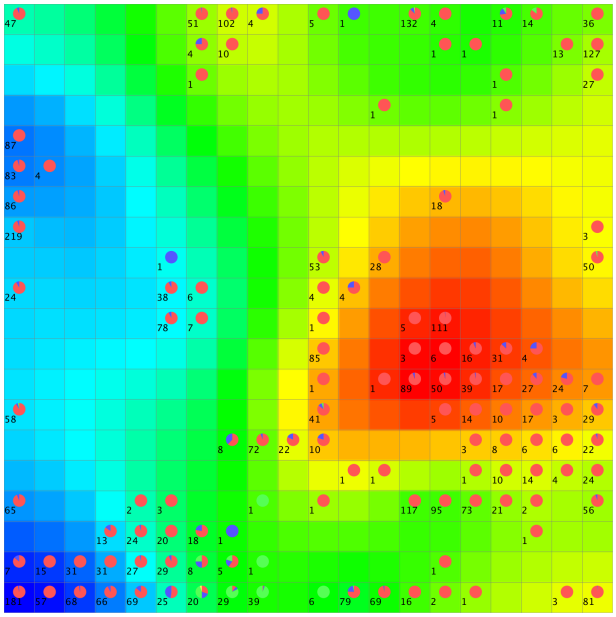

Figure 13: SOM with random seed 100: Component Plane "age"

together at the edge of one of the bigger clusters. This time the split of the big cluster doesn't seem to be the impact of attribute "age". A look at the corresponding compoenent plane (Figure 15 reveals that. On further inspection with different attributes we couldn't really make out any single attributes which might have been the cause of this split, so our assumption is that it must have either been caused by an interplay of multiple attributes or there does not really exist a clear cluster separation. A look at the neighborhood graph with radius 0.8 (Figure 17) revealed that, in fact, there does not seem to exist a clear division as lots of instances seem to have neighbors in the other clusters. Just as before, single attributes do seem to be the cause of "single unit" clusters, e.g. attribute "sick" caused the small cluster in the bottom-left corner of the map (see Figure 16).
The quantization errors and topology violations show very similar results to the standard SOM, so we are not showing them here.

Conclusion: As conclusion, we can say that the random seed has a huge impact on the overall structure of the SOM. SOMs of the same data that are initialized with different seeds will show no obvious similarities. This makes sense as the random seed just causes a different initialization of the weight vector, which causes data points to be mapped to be different parts of the map. However, as further observations revealed, the class distribution remains the same. In our examples this was shown by the data points of class "primary hypothyroid" which remained closely grouped together no matter the initialization. We conclude by saying the random seed has an influence on the initialization of the SOM but will not have a big impact on the training process of the SOM, the trained SOM will produce the same kind of relations independently of the inital random seed.

4.3 Different Map sizes

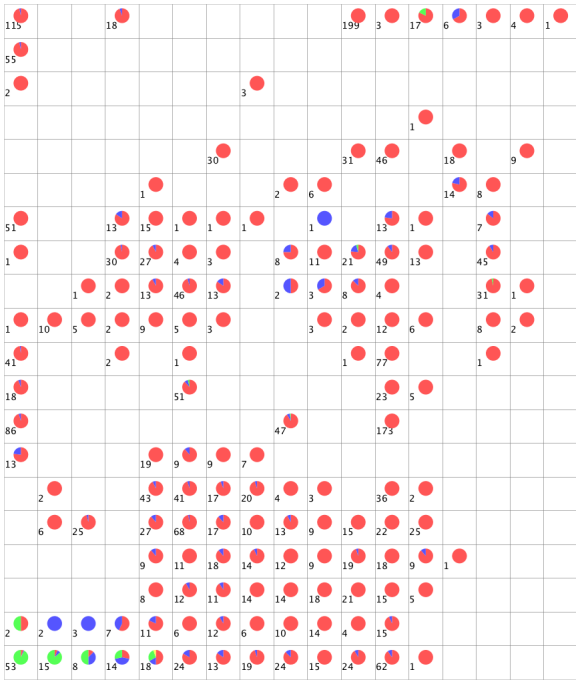

Figure 14: SOM with random seed 2: class distribution

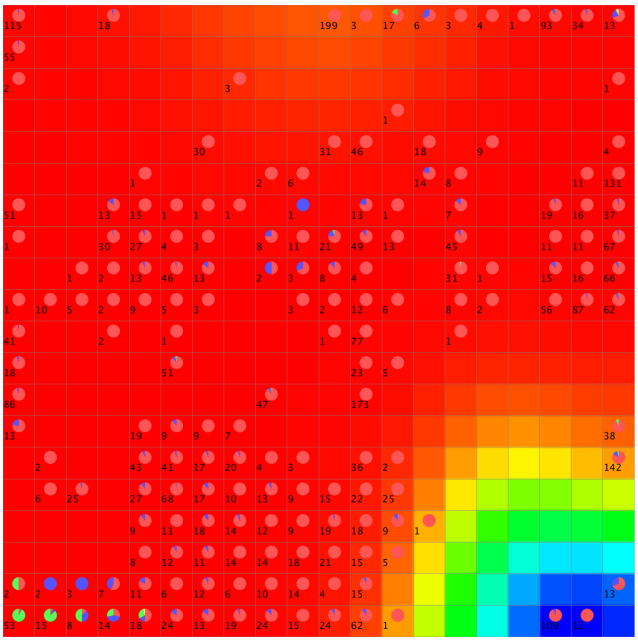

Figure 16: SOM with random seed 2: Component Plane "sick"

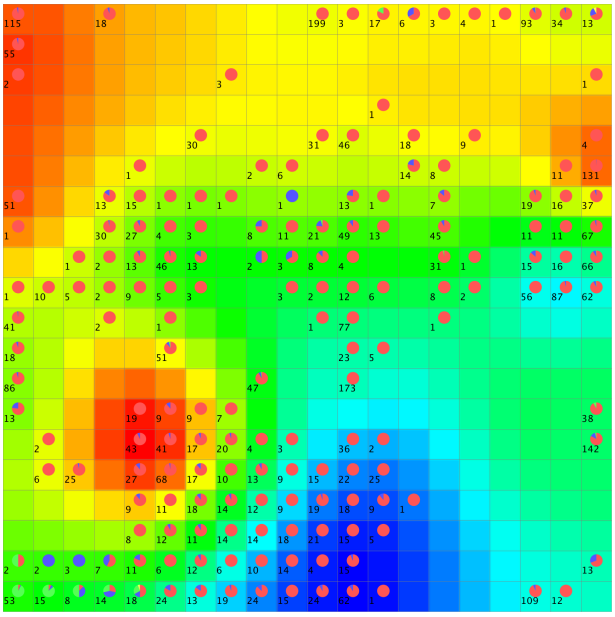

Figure 15: SOM with random seed 2: Component Plane "age"

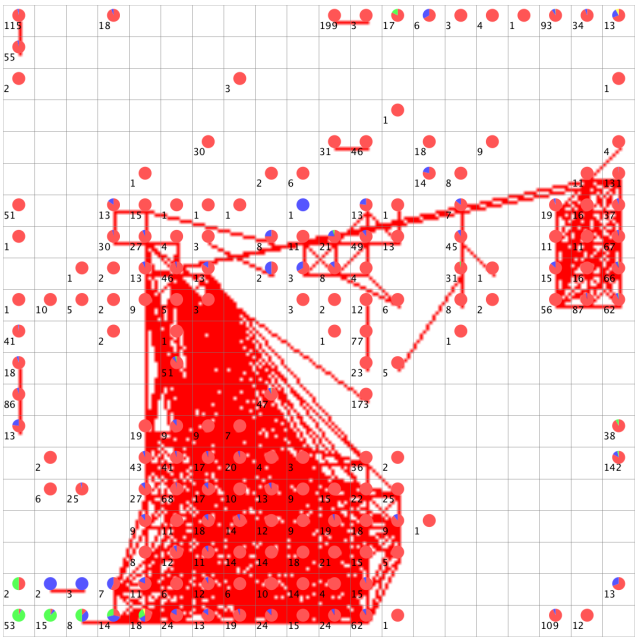

Figure 17: SOM with random seed 2: Neighborhood graph with radius 0.8

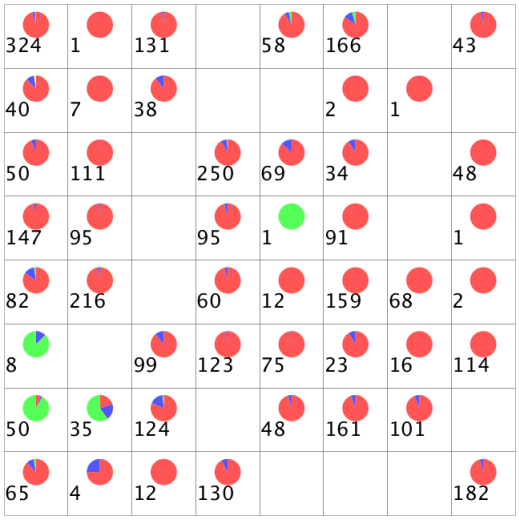

Figure 18: SOM with 8x8 map: class distribution

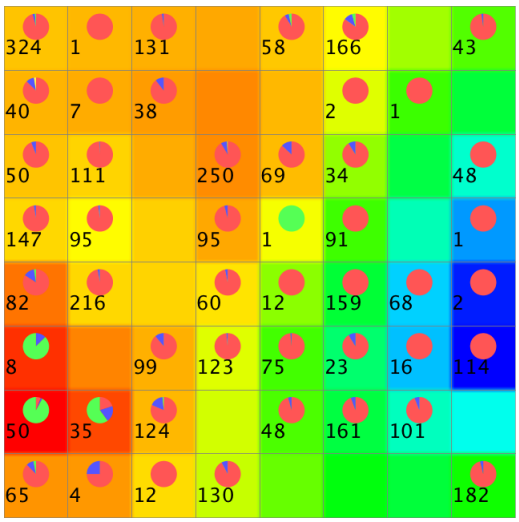

Figure 19: SOM with 8x8 map: Component Plane "TT4"

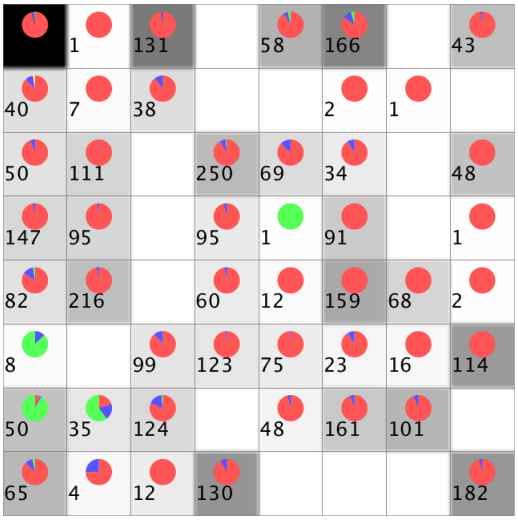

Figure 20: SOM with 8x8 map: Quantization error

In this section we analyzed different map sizes of the SOM. We wanted to examine in what way the cluster structure of the data is defined by the size of the map. For that we analyzed two different map sizes: a small map and a big map, with 8x8 units and 60x60 units, respectively. Furthermore, each map had to be trained with rather large neighborhood radius and high learning rate. As high learning rate we defined 0.9, as default value is 0.75. We determined the "high" neighborhood radius by looking at the default value of the respective map size.

Small:

- Mapsize: 8x8 –> 64 units, which is about 1% of 3700 values.
- neighborhood radius (sigma): 6 (default is 4 for 8x8)

The class distribution (Figure 18) reveals that classes "primary hypothyroid" still lies neatly packed together. The map seems to show some kind of cluster structure, but because the map is so small, and therefore almost every unit contains several data points, it is difficult to make out any distinct clusters. Looking at the component planes visualization the map seems to show similar relations as the standard SOM though in a much smaller scale. This can be seen for example for attribute "TT4" in Figure 19.

The quantization error (Figure 20 of the map seems to be a bit higher in comparison to the standard SOM, with a huge error in the unit in the top-left corner. In general, it seems to be fairly well distributed. As our map is too small to show any topology, it is unnecessary to have a look at topology violations.

Large:

- Mapsize: 60x60 –> 3600 units, which is nearly as high as our number of values (3700)
- neighborhood radius (sigma): 40 (default is 30 for 60x60)

Looking at the pie charts visualization in Figure 21) we can make out some similarities (after rotating and flipping of the map) in cluster structure and class distribution between the standard SOM and the new SOM with the large map size. But clusters in the big map seem to be spread out more, and 3 clusters have emerged from the big cluster in the standard SOM. In the figure we can see that clusters generally are well divided you can see clear borders between them. Further you can see that class "primary hypothyroid" still is grouped together on the edge of more bigger "negative" cluster.

Next we looked at the component planes of single attributes to see if any of the attributes are the cause of some of the clusters we see in the map. For the smaller clusters all around the map, the same observation could be made as before. Each smaller cluster seemed to be caused by instances which have shown a specific condition like "pregnant", "psych", etc. Also we could see that the bigger cluster on the right-hand side of the map seemed to contain almost all of the data points which have attribute "sex" set as male (see Figure 22). Figure 23 shows the component plane of attribute "thyroid surgery" and what is interesting is that a lot of the instances who have this attribute set, lie in the area where of class "primary hypothyroid". This could mean that thyroid surgery and hypothyroidism may be related. In general, as the map is so big, the impact of single attributes on clusters of the SOM can be seen quite clearly.

Lastly, we observed the quantization error and topology violations. The quantization error (Figure 24) seems to be even lower than in the standard SOM with almost no error in the center of the map, and higher values in the corner and some clusters on the edges. Looking at topology violations with k-nn (Figure 25) we see that our map, in general does not consist of many topology errors, but: one cluster on the left hand side, seems to produce many topology violations as we can see many red lines spanning across the map to unit on the top-right corner.

Conclusion: In general we can say that the SOM reveals similar structures of the data for different map sizes, although clusters may have shifted/moved around the map. Also clusters in the big map a far more distinguishable and seem more spread out than in a small map.

4.4 Different initial neighborhood radius settings

In this sections we tried out different neighborhood radius settings. This means we played around with the parameter determining on how many neighbors an activated weight vector has influence. We trained one "regular" (standard size) with a much too large neighborhood setting, and one very large map (50x50) with very small radius setting.

Regular map, large neighborhood: We trained a 20x20 map with neighborhood radius sigma = 220. Looking at the class distribution of the SOM (Figure 26) we notice that all data points seem to have been pushed to the edges and corners of the SOM. This is an expected behavior as the large neighborhood will cause weight vectors of units to be influenced by almost all units around it. As the units in the middle have more neighbors than the units on the edges and in the corners, the weight vectors of these units in the mid-

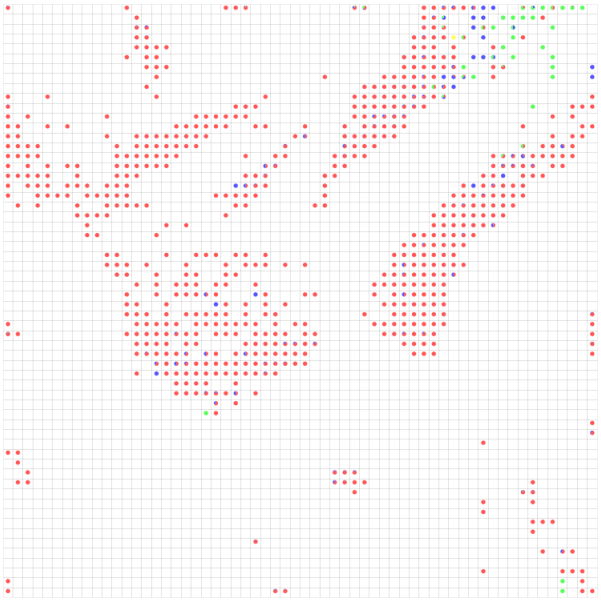

Figure 21: SOM with 60x60 map: class distribution

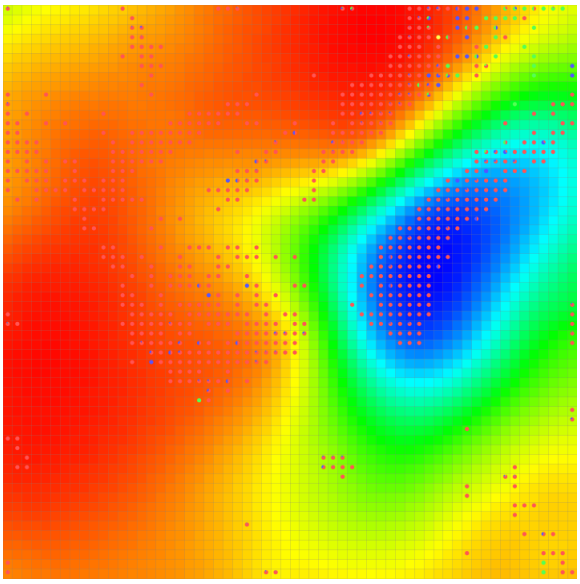

Figure 22: SOM with 60x60 map: Component Plane "sex"

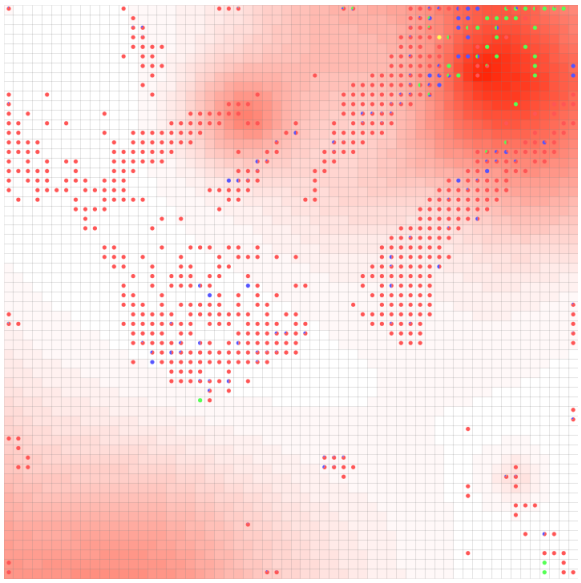

Figure 23: SOM with 60x60 map: Component Plane "thyroid surgery"

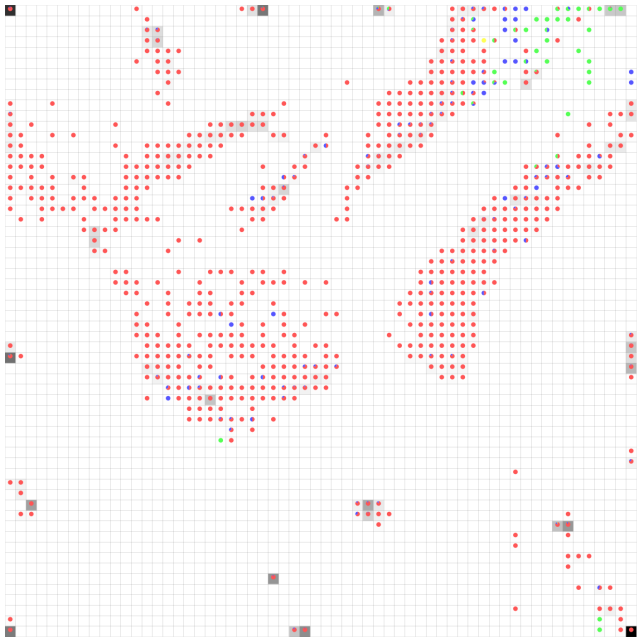

Figure 24: SOM with 60x60 map: Quantization error

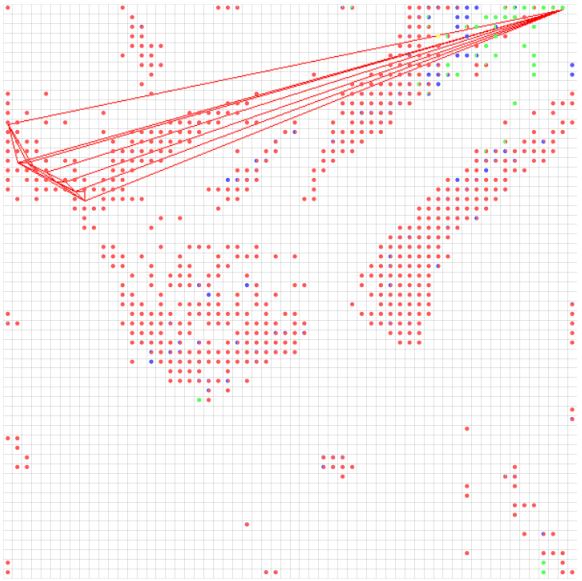

Figure 25: SOM with 60x60 map: Topology violations k-nn

dle will behave like a "FÃd'hnlein im Winde", changing direction whenever one of the neighboring weight vectors gets activated. Units with not as many neighboring units (edges and corners) do not get influenced that strongly which is why they will pull more strongly on the data points. This is known as boundary effect. This can be observed even more when looking at the component planes of data. For example the component plane for the attribute "TSH" reveals that data points have been pulled to the corners. Interesting to note hereby, is that high and low values seem to end up in opposite corners. A look at the neighborhood graph with radius 0.5 also shows many topology violations, especially from one edge to the other (see Figure 28).

Large map, small neighborhood: We trained a standard 50x50 map with neighborhood radius sigma = 2. Looking at the pie-chart visualization (Figure 29) the map seems to contain clear cluster structures. A look at the component planes of the attributes quickly reveals that these structures do not really show the underlying structure of the data. Especially the component plane for attribute "sex" in Figure 30 shows this quite well: Instead of two bigger clusters dividing the gender of the patients, we see multiple smaller clusters which divide by gender. This conforms to the too small neighborhood parameter we have chosen, as activated weight vectors will have nearly no influence on the neighboring units. Therefore if there are similar weight vectors in the map they will not converge with time, but create isolated units with the same structures on different parts of the map. Clusters will not converge, but we will have several clusters representing the same relationship throughout the map. This is also confirmed by the neighborhood graph with radius 0.5 in Figure 31 as we see lots of topology violations between the smaller clusters.

Conclusion: To conclude we can say that the neighborhood radius has a great influence on the outcome of the map. Hereby, it is important to choose the parameter not too small, so that clusters showing the same kind of relationship converge, as well as not too high, as this will inevitably

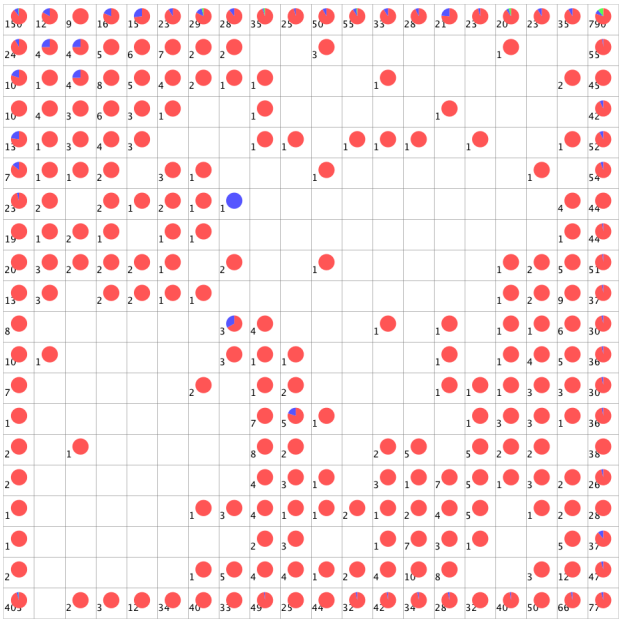

Figure 26: SOM with large neighborhood radius: class distribution

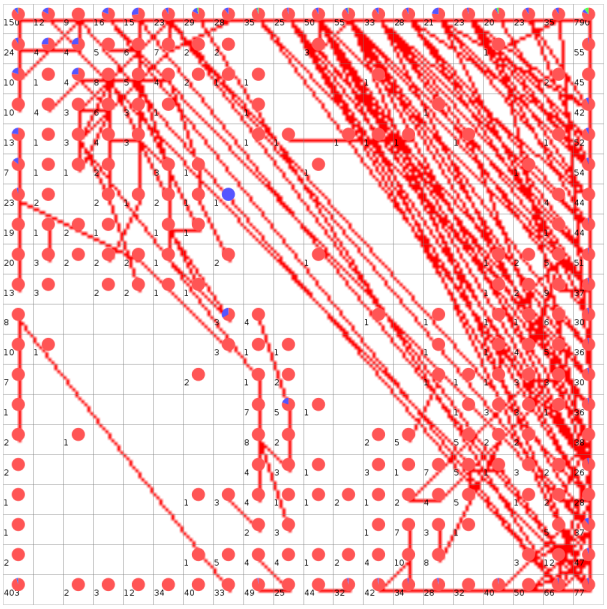

Figure 28: SOM with large neighborhood radius: Neighborhood graph with radius 0.5

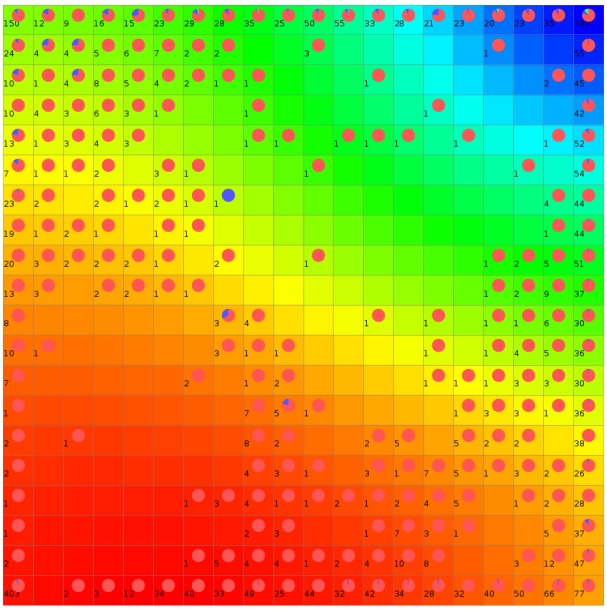

Figure 27: SOM with large neighborhood radius: Component Plane "TSH"

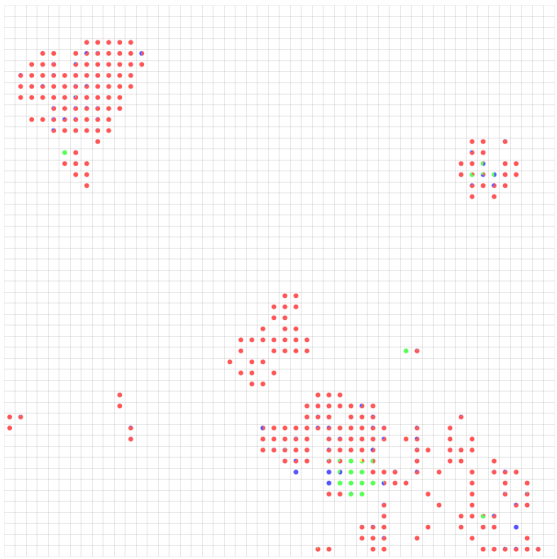

Figure 29: SOM with small neighborhood radius: class distribution

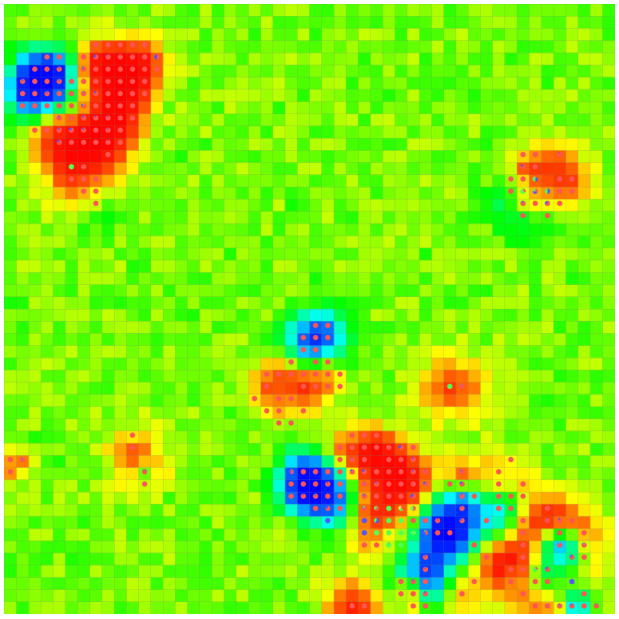

Figure 30: SOM with small neighborhood radius: Component Plane "sex"

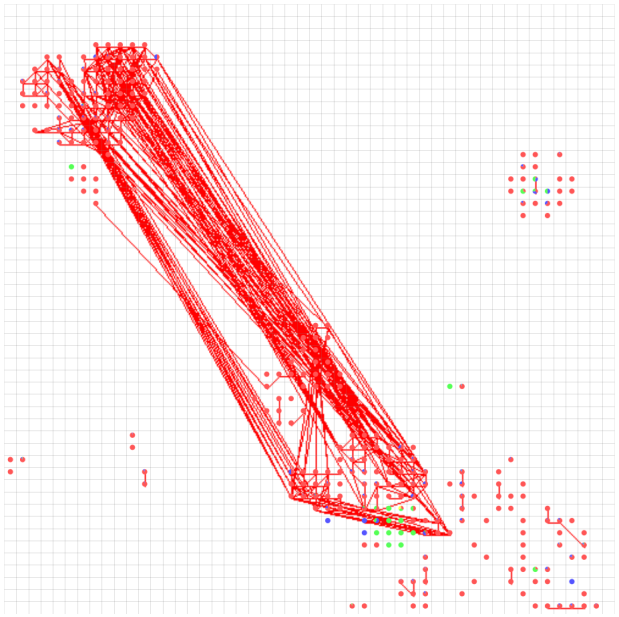

Figure 31: SOM with small neighborhood radius: Neighborhood graph with radius 0.5

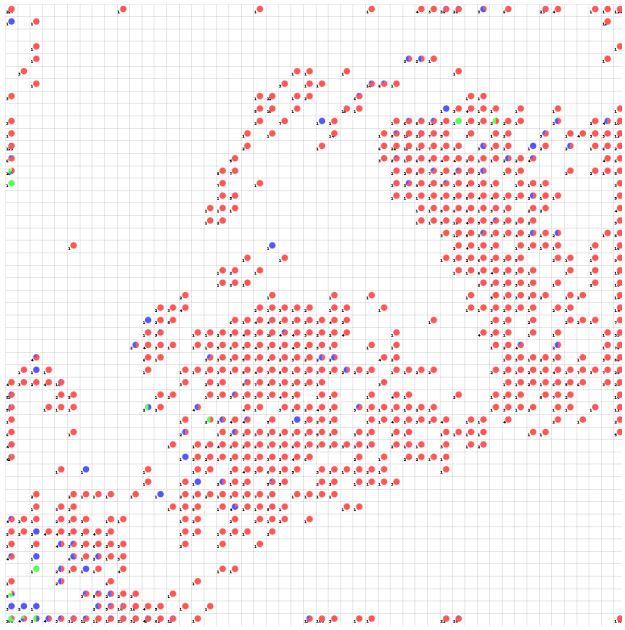

Figure 32: SOM with low learning rate: Class distribution

lead to boundary effect, eliminating cluster structures and important insight of the data.

4.5 Different initial learning rates

In this section we analyzed different initial learning rates for our SOM training. We did this with a very low learning rate of 0.1 on a very large map with a size of 50x50 and a very high learning rate of 0.99 on a "regular" sized map, as specified in the assignment.

Low learning rate (0.1) on big map (50x50: When working with a low learning rate we get most of our primary thyroid classed data points into one single unit on the bottom left, as we can see in Figure 32. Also the two large clusters are not built around these unit but especially one of them is nearly on the other end of the map. When we look at the component planes, we also see that the values are more distributed. For example, when we look at the component plane for "TSH" in Figure 34 we can see, that most of the positive values are still on the bottom left, but wide areas of the map are green, yellow and orange, which indicates, that the values are not that concentrated on one spot, but are more distributed than for our "regular" SOM (Figure 6) and our optimal SOM (Figure 63).
When looking at the quantization errors in Figure 33, we can see that there is a huge amount of quantization erros, especially from the bigger cluster on the top right (which as already mentioned is normally quite close to most of the primary thyroid classed data points.

High learning rate (0.99): When looking at our SOM trained with a high learning rate in Figure 35 we can see, that the two big clusters are not separated that clearly and also one of them is quite far away from our data points with class primary thyroid. This gets even more clear, when we look at the component plane for the attribute "Sex" in Fig-

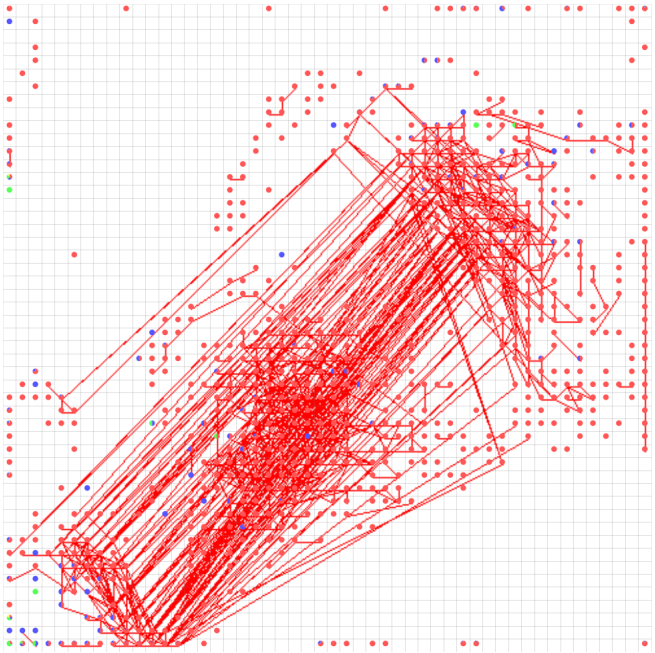

Figure 33: SOM with low learning rate: Neighborhood radius 0.5

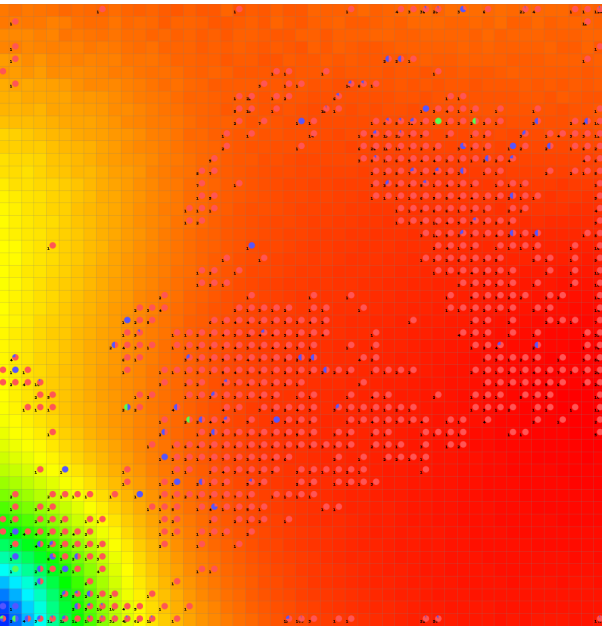

Figure 34: SOM with low learning rate: Component plane for "TSH"

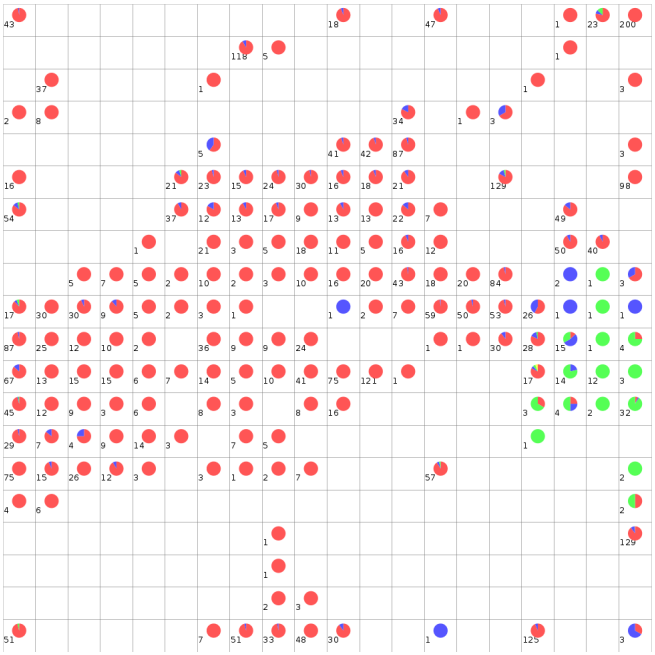

Figure 35: SOM with high learning rate: Class distribution

ure 37, where we can see that the cluster with most of our male patients is on the left side. This cluster is in most of our other maps quite close to the primary thyroid classed data points, as we can see on the optimal SOM in Figure 62.
Additionally the number of topology violations (see Figure 36) is very high compared to other SOMs.

Conclusion: In general we can say that the learning rate should be handled carefully, otherwise our good cluster structures get lost and we get a lot of topology violations.

4.6 Different scaling

In this section we trained our SOM with obviously wrongly scaled data. In our case we used the unit length normalization. This would reduce the influence of the boolean attributes to a great deal, as unit length normalization is done by taking attribute value divided by the total sum on all attributes for each instance.

Looking at the pie-chart visualization in Figure 38 we see immediately that the structure of the SOM has not only changed completely but also there does not seem to exist any structure at all. There are data points in almost every unit of the map, with almost equal density everywhere. only on the edges and in the corners there are some units containing more data points. The only thing that remained the same is the class distribution of class "primary hypothyroid". We assume this is because the class is mostly determined by the continuous measurement values like TSH, FTI, etc. which will not be influenced as much by the unit length normalization as the boolean values. Looking at some of the component planes of some of the boolean attributes, our assumption about the influence of boolean attributes on cluster structure is confirmed. For example, a look at

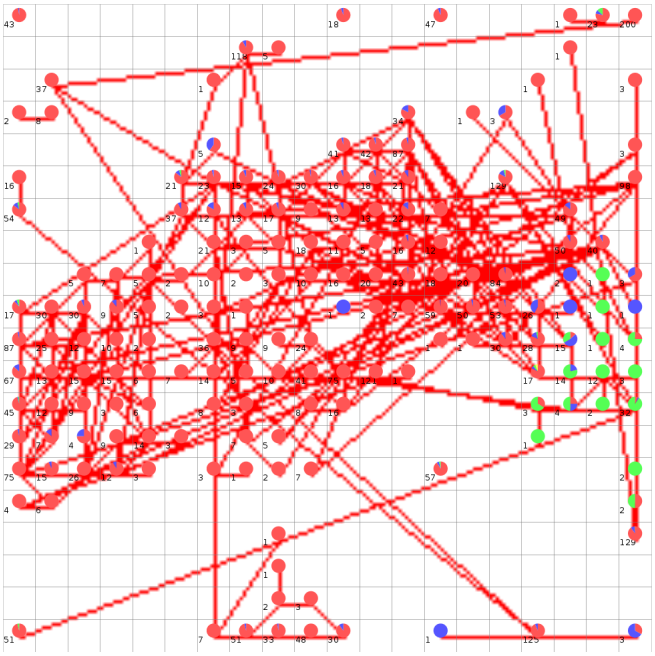

Figure 36: SOM with high learning rate: knn neighborhood

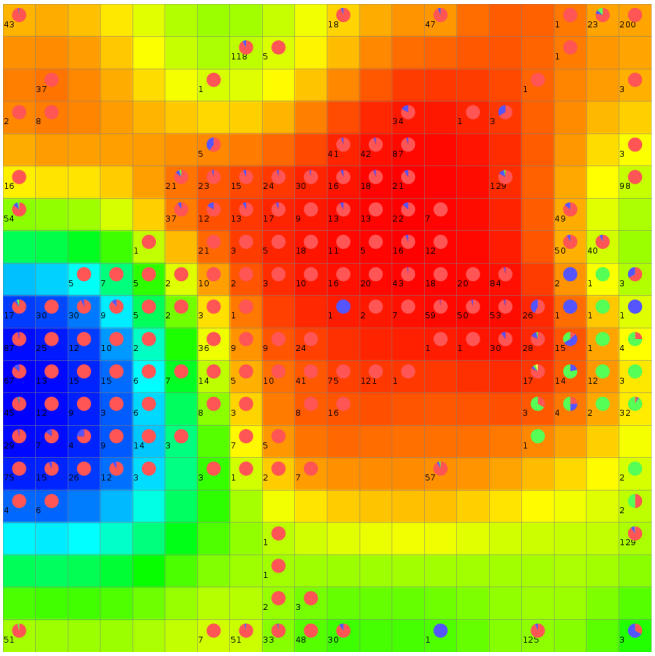

Figure 37: SOM with low learning rate: Component plane for "Sex"

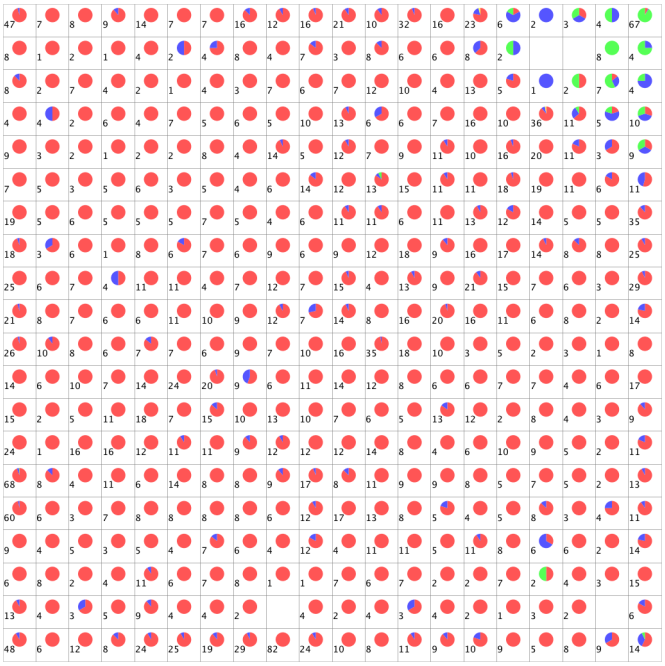

Figure 38: SOM with wrong scaling: class distribution

the component plane of attribute "psych" (Figure 39) shows that instances which had the attribute got mapped to almost anywhere around the map, where as in our standard SOM, those instances got mapped mostly to a unique single cluster containing all data points of that type.

Looking at the quantization error (Figure 40) we further see almost no error for most of the units except a huge error for a single unit in the top-right corner which contains almost only data points of class "primary hypothyroid". Our explanation for the error in that unit and the rest of the map showing almost no kind of error is as follows: As almost all vectors look the same because of the unit-length normalization of boolean attributes, only the ones containing different measurement values because of hypothyroidism will sort of "stand out" and will get mapped to a specific part of the map, causing the high quantization error.

Lastly we can have a look at the k-Nearest Neighbor Neighborhood graph of the SOM (Figure 41). The vast number of red lines crossing all around the map, from left to right, and from bottom to top, confirm us that, in fact, a structure/topology is non-existent.

Conclusion: We can conclude the normalization method chosen before training the map has a massive impact on the outcome of the SOM. Data analyst need to be careful about the method they choose, as the method of scaling may determine if one can see the underlying structures of the data or not.

4.7 Different max iterations

In this section we trained the SOM with an increasing number of iterations. The iterations used were 2, 5, 10, 50, 100, 1000, 5000, 10000. We were especially interested to find out at what number of iterations cluster structures seem to emerge, and when they stabilize. This should help us find an optimal number of iterations for the last analysis in Section

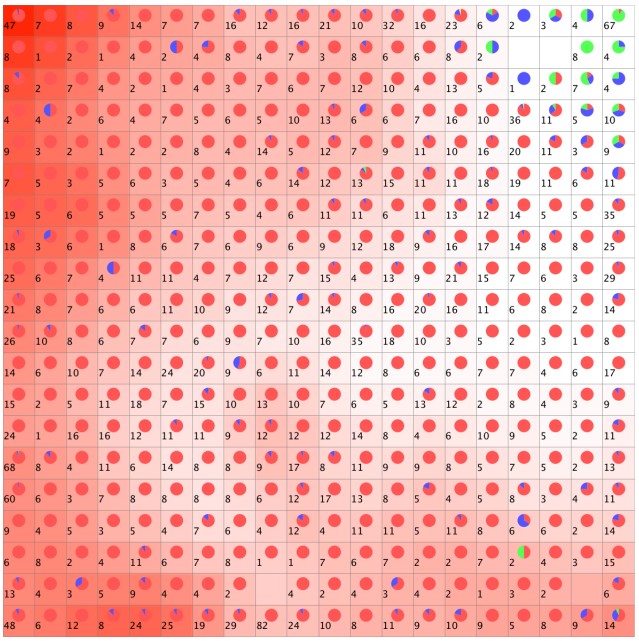

Figure 39: SOM with wrong scaling: Component Plane "psych"

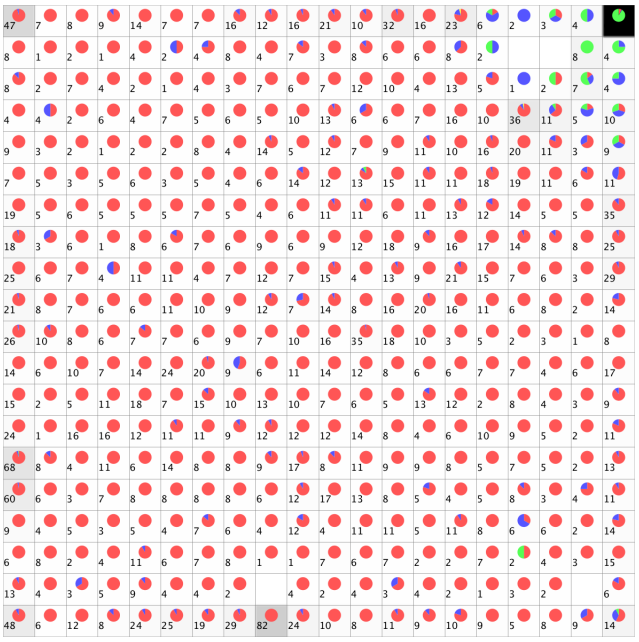

Figure 40: SOM with wrong scaling: Quantization Error

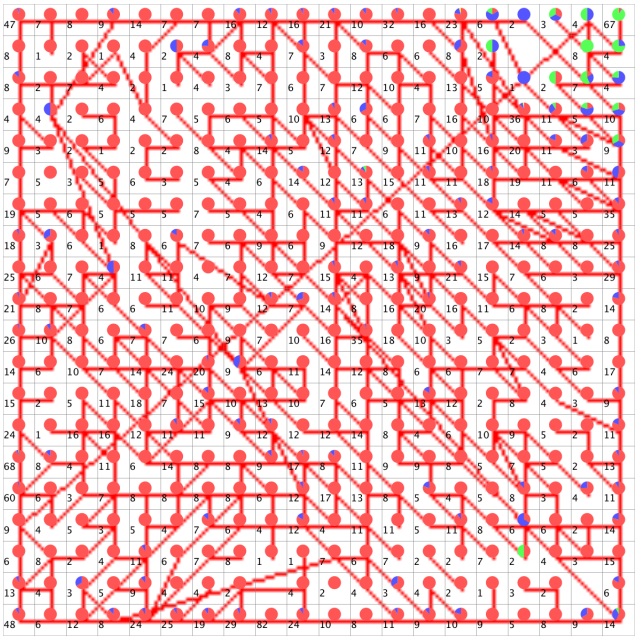

Figure 41: SOM with wrong scaling: Neighborhood graph with k-nn

4.8.

2 iterations: After just 2 iterations no real structure has emerged yet (see Figure 42). A huge amount of data points are concentrated in the units in the bottom-left and bottom-right corner of the map. A few data points (especially those of class "primary hypothyroid") fly around and about the map with no kind of pattern visible. As expected a look at the neighborhood graph with k-nn (Figure 43) reveals many topology violations.

5 iterations: After 5 iterations the SOM looks almost identical to the SOM of 2 iterations (see 44). Only a few data points have moved from the bottom-left to the bottom-right corner, as the data point count reveals.

10 iterations: 10 iterations still draw a similar image (see Figure 45). There are still most data points concentrated in corners at the bottom. Some more data points have moved from bottom-left to bottom-right. One change we do notice though is that the cluster in the bottom right together with most of the data points of class "primary thyroid" have begun to spread out a little.

50 iterations: Looking at the pie-chart visualization after 50 iterations (Figure 44) we can see that the data points from the bottom-left corner have dissolved even more. The data points of the bottom-right corner have fully emerged into a fully grown cluster carrying the data points of class "primary hypothyroid" on its top. Furthermore the data points at the upper edge of the map have moved closer together and begin to form another cluster. A look at the neighborhood graph (Figure 47) quickly reveals that the map still consists of many topology violations and probably has not stabilized yet.

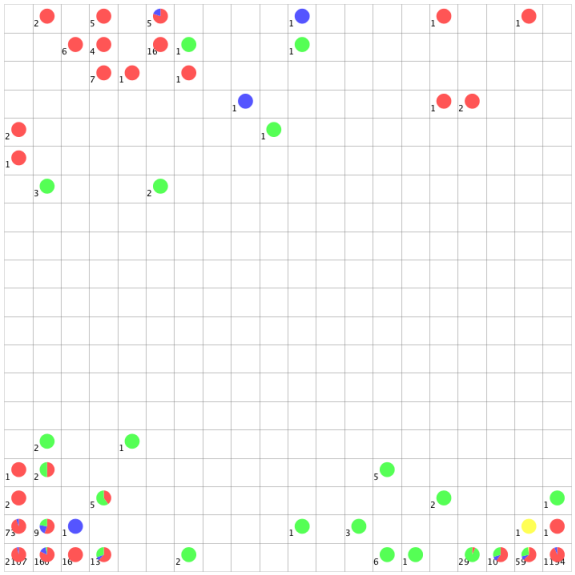

Figure 42: SOM after 2 iterations: class distribution

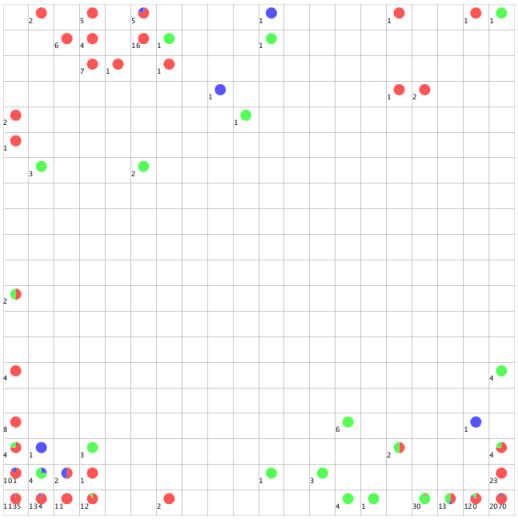

Figure 44: SOM after 5 iterations: class distribution

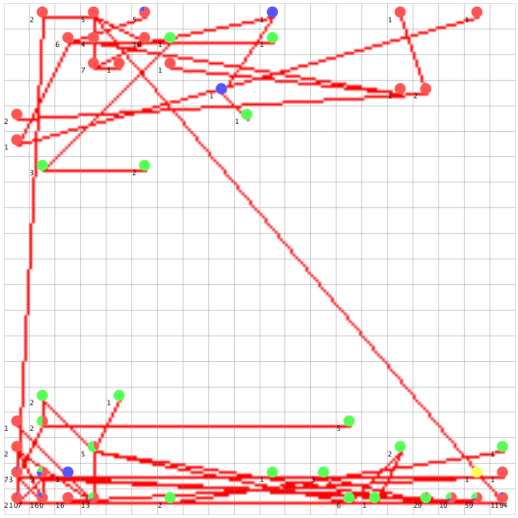

Figure 43: SOM after 2 iterations: Neighborhood Graph with k-nn

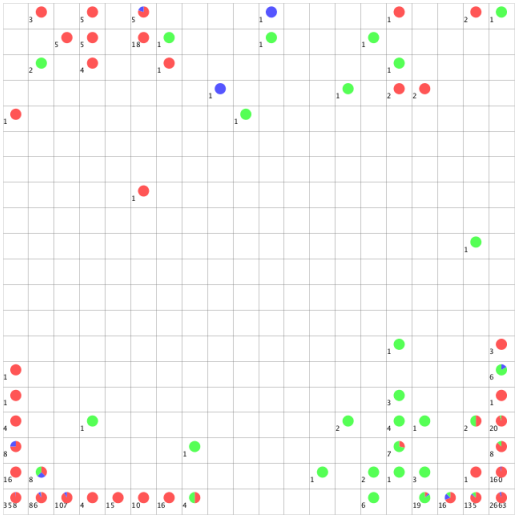

Figure 45: SOM after 10 iterations: class distribution

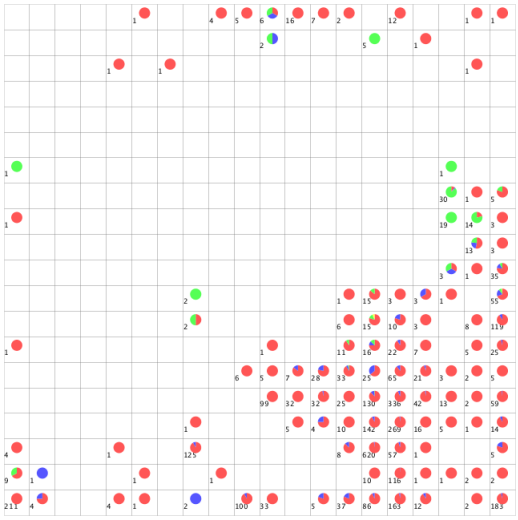

Figure 46: SOM after 50 iterations: class distribution

Figure 47: SOM after 50 iterations: Neighborhood Graph with k-nn

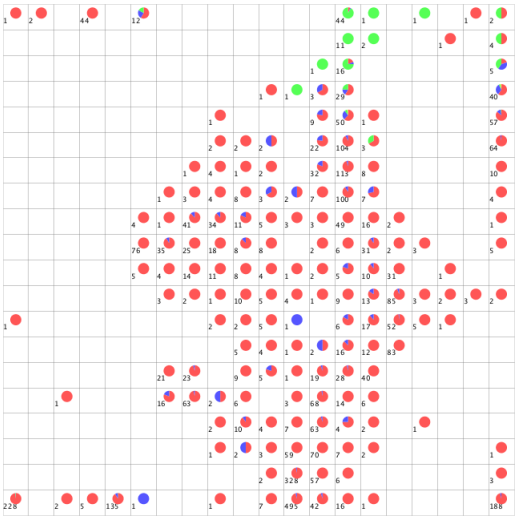

Figure 48: SOM after 100 iterations: class distribution

100 iterations: After 100 iterations the big cluster we saw starting to emerge after 50 iterations has expanded even further, now spanning pretty much from bottom to the top of the map. The data points of class "primary hypothyroid are still being carried on its top. The structure already looks kind of similar to the the structure of our standard SOM we saw in Figure 2 just flipped horizontally, also containing smaller clusters in the corner. We cannot make out as many "single unit" clusters as in the standard SOM though. Looking at the neighborhood graph of k-nn (Figure 49) we can still see many topology violations all around the map. The neighborhood graph of radius 0.8 in Figure 50 seems to show that the bigger cluster in the middle is actually made up of 2 clusters.

1000 iterations: After 1000 iterations the cluster structure of the SOM has changed completely. This is probably due to the fact that the map was not stable yet. In Figure 51 we can now see multiple clusters of medium size spread around the map. The data points of class "primary thyroid" still seem to lie at the edge of a cluster even though this cluster does not seem to resemble the cluster from 100 iterations at all. Looking at the neighborhood graph with radius 0.6 (Figure 52) reveals to us that the bigger cluster from before has actually fallen into 3-4 smaller clusters, two of them still quite dense. The topology violations in the same graph tell us also that the SOM is not stable yet.

5000 iterations: After 5000 iterations the structure of the SOM seems to be kind of similar to the structure we saw after 1000 iterations (with rotating and flipping). If we ignore the fact that cluster structures have shifted, rotated and moved around a bit, in Figure 46 we can see that the map consists of two bigger clusters, and multiple smaller ones spread across the map, just as in the map after 1000 iterations. The moving around of clusters might be an indicator that the map has not stabilized yet. A look at the neighborhood graph with radius 0.8 in Figure 54, although now showing two clear clusters, still confirm this belief as we can still observe multiple topology violations.

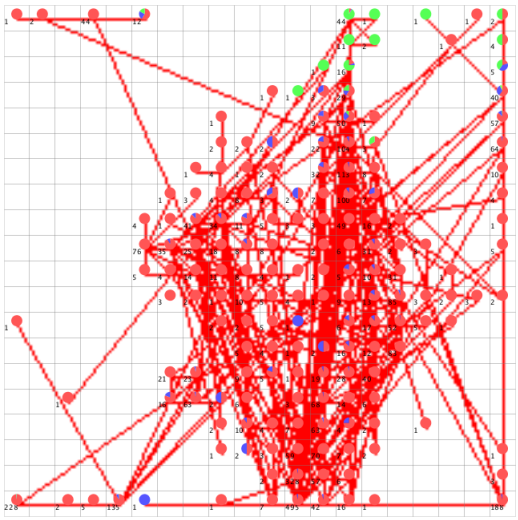

Figure 49: SOM after 100 iterations: Neighborhood Graph with k-nn

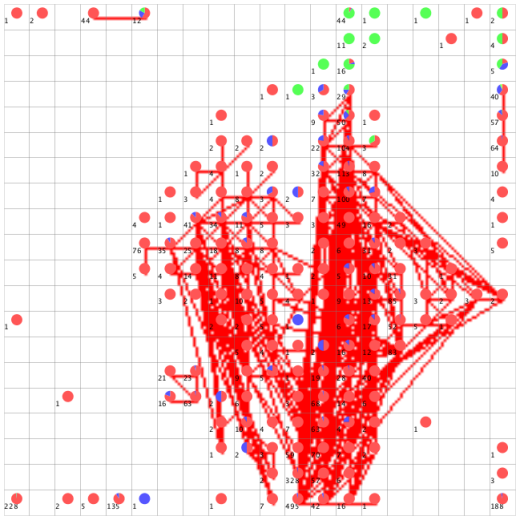

Figure 50: SOM after 100 iterations: Neighborhood Graph with radius 0.8

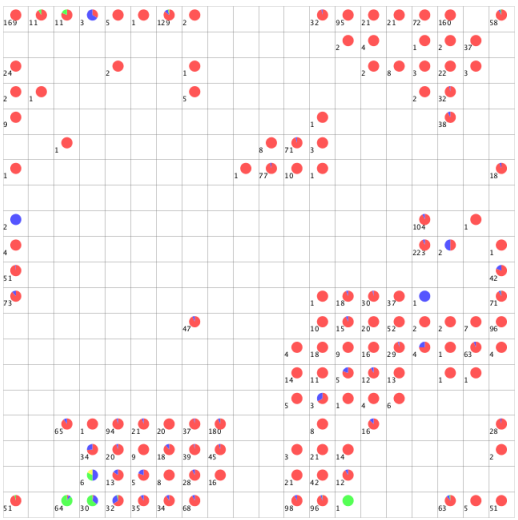

Figure 51: SOM after 1000 iterations: class distribution

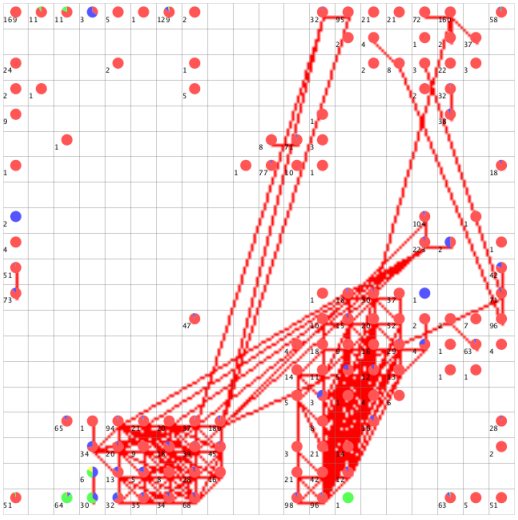

Figure 52: SOM after 1000 iterations: Neighborhood Graph with radius 0.6

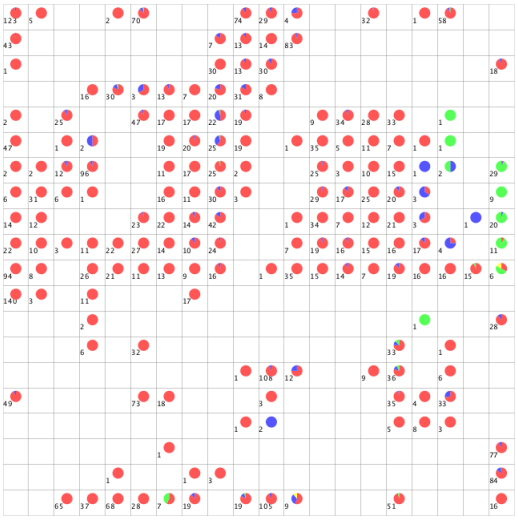

Figure 53: SOM after 5000 iterations: class distribution

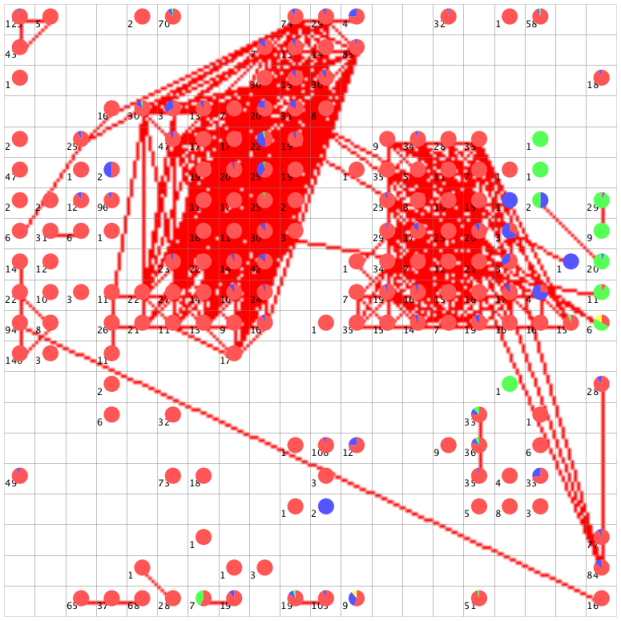

Figure 54: SOM after 5000 iterations: Neighborhood Graph with radius 0.8

10000 iterations: 10000 iterations show again similar cluster structures, if you ignore the fact that clusters have moved around, spread out, or rotated. Figure 55 shows the pie-chart visualization. We can clearly see two bigger clusters, with smaller ones lying all around the map. Component plane of attribute "sex" in Figure 56 might suggest that the two clusters actually divide the data points by sex. A look at the neighborhood graph with radius 0.8 (Figure 57 draws a similar picture as the neighborhood graph for 5000 iterations, only that the two big clusters might seem to be even more distinguishable. As you can also see in the figure a fair amount of topology violations still exist, which might suggest that the map even after 10000 iterations is not fully stabilized yet.

Conclusion: In all SOMs we saw a lot of structure changes, only after 1000 iterations structures seem to emerge that persist throughout the SOMs of 5000 and 10000 iterations. After 5000 iterations we clearly saw two big clusters emerge, which became even denser and more clearly divided after 10000 iterations. Smaller clusters as caused by single boolean attributes seem to persist as well. The neighborhood graph of 10000 iterations still revealed a number of topology errors, which is why we might choose an even higher number of iterations for the training of our optimal SOM.

4.8 Optimal SOM

Finally this section will analyze an "optimal" SOM. This means we trained a SOM trained with parameters which we think are best suited for the data set. The parameters were chosen based on our experiences we made in the previous sections where we tested different configurations of said parameters. Our optimal SOM was trained with following parameters:

- **Random Seed 7:** As we saw in Section 4.2 the random seed is mainly responsible for the initialization of the SOM. Which is why different random seeds might not create exactly the same SOM, but structures and

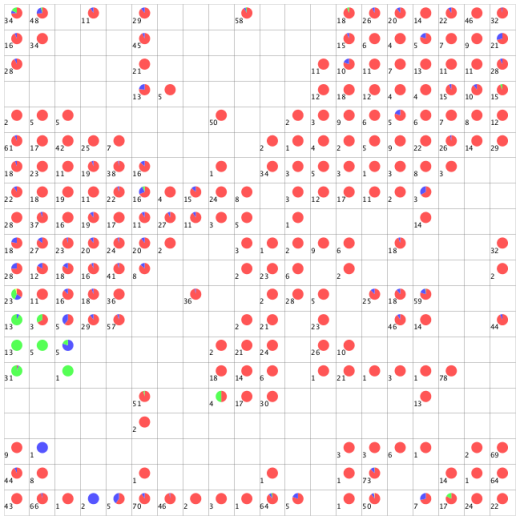

Figure 55: SOM after 10000 iterations: class distribution

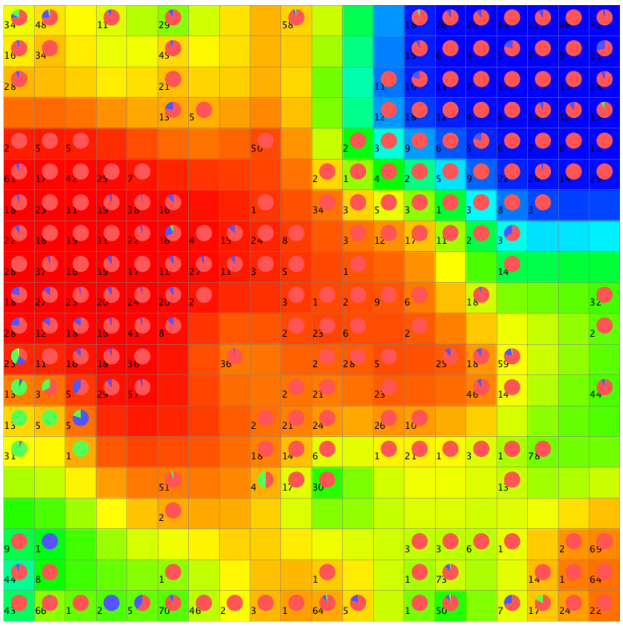

Figure 56: SOM after 10000 iterations: Component Plane "sex"

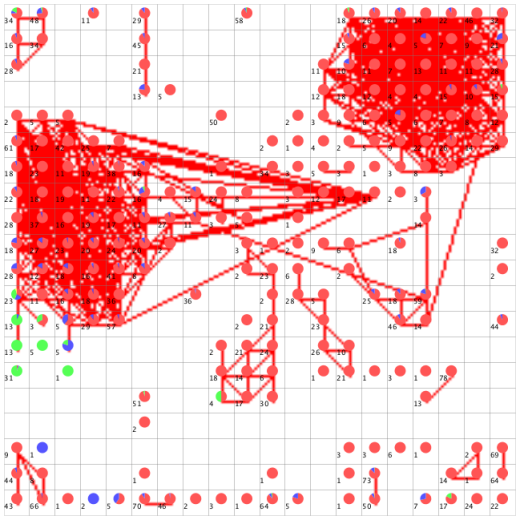

Figure 57: SOM after 10000 iterations: Neighborhood Graph with radius 0.8

relations should emerge in a trained SOM no matter the initialization. As we tried out some random seeds, we agreed that the structure looked most cleanly on random seed 7, so we decided to use this seed also for our optimal SOM.

- **Map Size 40x40:** We chose this as our map size, as we found our standard map size of 20x20 a bit small for showing clear cluster separation. In section 4.3 we tested a smaller and a bigger map size. The smaller one (8x8) turned out to be way too small to show clear distinct cluster relations, the bigger one (60x60) generally showed very good distinction between clusters, but also contained quite a few big "holes", meaning areas of the map containing no data points at all. This is why we decided 40x40 shall be the map size for the optimal SOM.
- **Neighborhood radius 15:** In sections 4.4 we showed that choosing neighborhood radius as two low results in no cluster convergence, as weight vectors will not be influenced as much. On the other hand choosing a neighborhood radius which is too large, results in boundary effects as edges and especially corners influence the other units in great deal but are themselves not that influenced by the other ones. Therefore we decided to start from the default value (20 for 40x40 maps). Testing around this default value, revealed that the more beautiful map resulted from a neighborhood radius of 15, which is why we decided to leave it at that.
- **Learning Rate 0.75:** In section 4.5 we tried SOM with different learning rates. Neither the small learning rate (0.01) nor the big one (0.99) resulted in better cluster structure and class distribution. Also while playing around in the middle, we couldn't get any better results than with the default value of 0.75.
- **Number of iterations: 20000** In section 4.7 we analyzed different number of iterations for the SOM training process. Generally, from 5000 to 10000 iterations stable clusters seem to emerge. However, even at 10000 iterations the Map revealed some topology violations. This is why for our optimal SOM we raised the number of iterations even more to 20000. This also corresponds the recommended value of choosing the number of iterations as 5 times the number of data vectors, which we oversaw at the starting of our lab execution.

After training the "optimal" SOM with above parameters, we can see in the pie-chart visualization (Figure 58) many clear cluster structures which hold a lot of interesting information about the underlying data. First thing we notice are the two very big clusters almost right in the center of the map. An expection with the smooth histogram visualization (Figure 66) tells us that these two clusters are quite dense. This means most of our data points are settled there. The component plane visualization reveals us at once the cause of the separation of the two big clusters. Component plane for attribute "sex" (Figure 62) shows us that the two clusters seem to divide the two genders quite distinctively. Also it is interesting that the instances of class "primary hypothyroid" seem to lie at the end of each cluster, building sort of like a bridge between them. This could be an indicator that both

genders tend to hypothyroidism equally, in that it is not determined by sex, whether or not a person is more likely to get hypothyroidism. As not all our data points lie within these two clusters, though it is also clear that these clusters won't contain all instances of the respective gender. Rather, we assume that the instances that do lie within them are the points which do not show any signs of rare conditions like pregnancy, sickness, etc. The data points which do have a rather rare condition can be found in the many smaller clusters all around the map, e.g. data points with attribute "psych" seem to be all included in the cluster at top-right corner of the map (see Figure 61).

Another interesting observation can be made by looking at the component plane of attribute "age". In Figure 59 we see that the age attribute seems to stretch through the "male" and "female" clusters in a similar fashion, red values being lower ages, and blue values being higher ages. As the connection of "primary hypothyroid" instances lie at the end with higher age values, we suspect old people being at risk for hypothyroidism.

When looking at the neighborhoods graphs for knn (Figure 64) and radius 0.8 (Figure 65) we can see a lot of small lines within the two clusters and especially with the radius only very few really long lines, which indicate topology errors. Overall we are quite satisfied with the amount of errors and assume that most of them are not from a badly built map but rather from the fact, that we have a lot amount of attributes with quite dominating attribute values (we have mostly boolean attributes, where most of the values are false and only a few percent are true). Since most entries of positive values in one of these attributes are grouped together to a small cluster on the edges/corners, one entry with positive values in several attributes can either be grouped into one of these clusters or lies in the middle of them. Either way we assume that these data points lead to topology errors as seen in the neighborhood graphs.

As we saw in the previous sections the quantization errors are quite low overall, but there are some high values in the corners or edges. This also holds for the optimal SOM (but since it is really similar over all our SOMs, we didn't show another figure of it again). We assume that this also comes from the clustering of single attributes being positive (like seen in Figure 61). Although they share this one attribute in common, they might be quite different in all the other attributes and are therefore "quite far away" from each other.

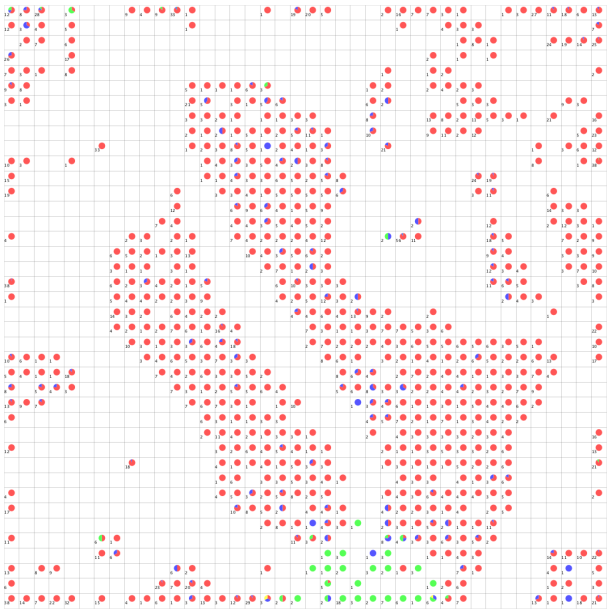

Figure 58: Optimal SOM: class distribution

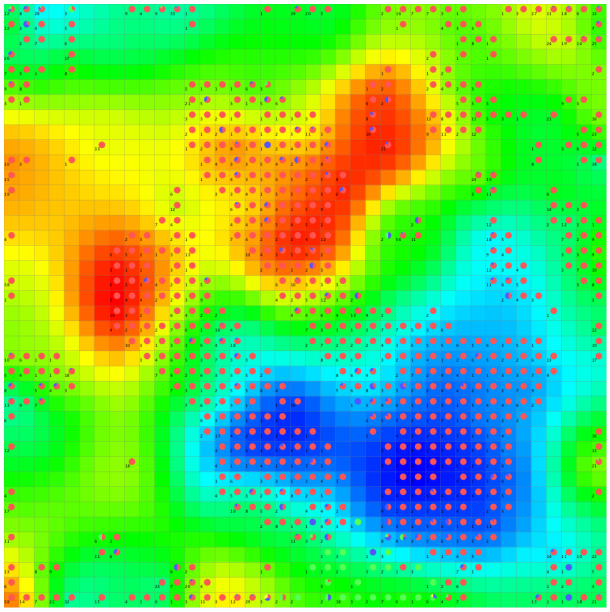

Figure 59: Optimal SOM: Component Plane "age"

5. SUMMARY

5.1 Parameter influence

While playing with the SOM in the previous sections we learned a lot about the parameters and their influence. Following our investigations after long discussions about it:

- **randomSeed** (initialization): The Random seed changed the structure of the map completely, however in the most cases the formed clusters are quite similar. For our optimal SOM training this was the last parameter to play with, since the other attributes had more comprehensible influence on the SOM and we just tried some random seeds at the end to see which one we like

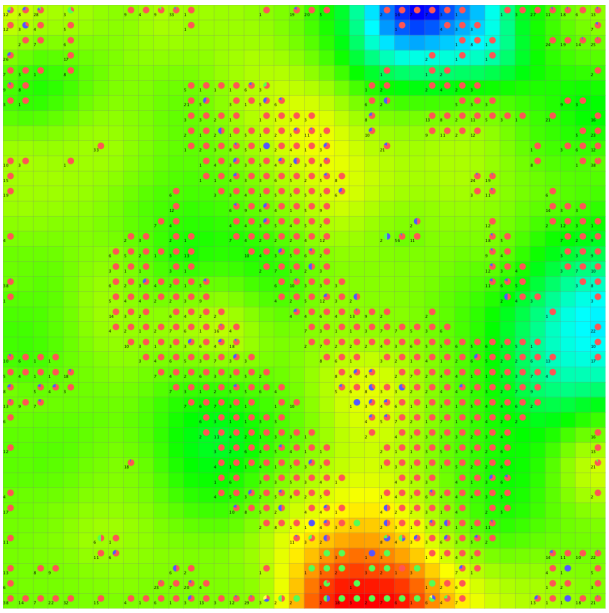

Figure 60: Optimal SOM: Component Plane "FTI"

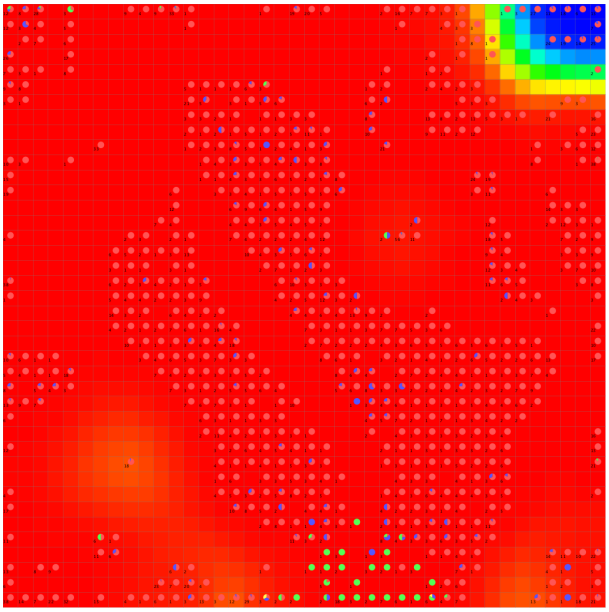

Figure 61: Optimal SOM: Component Plane "psych"

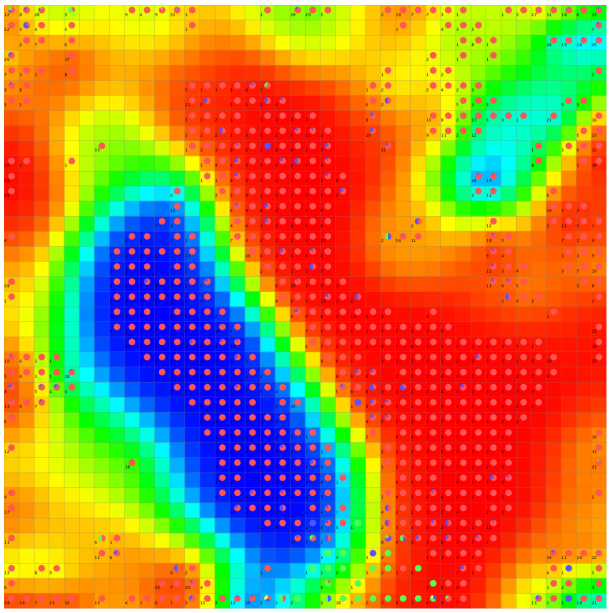

Figure 62: Optimal SOM: Component Plane "sex"

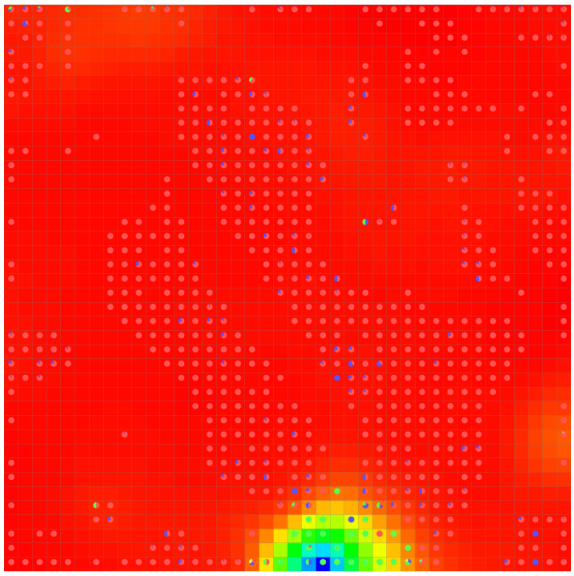

Figure 63: Optimal SOM: Component Plane "TSH"

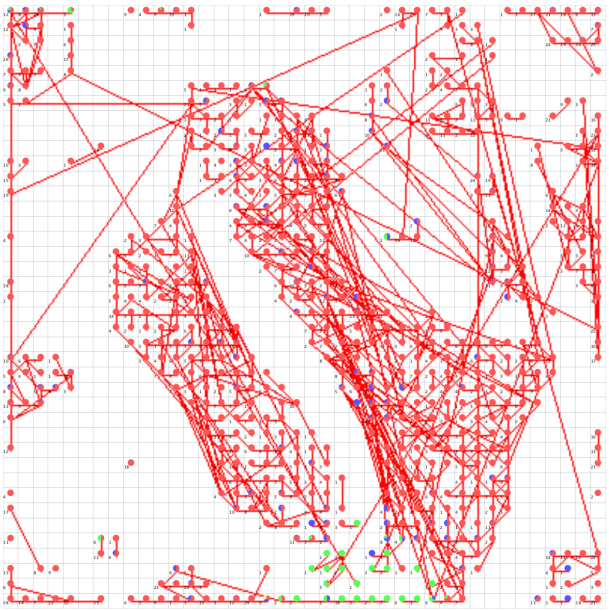

Figure 64: Optimal SOM: Neighborhood graph k-nn

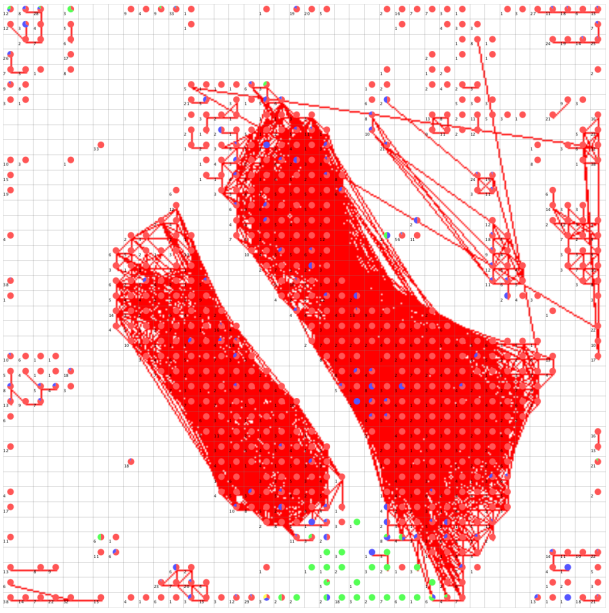

Figure 65: Optimal SOM: Neighborhood graph with radius 0.8

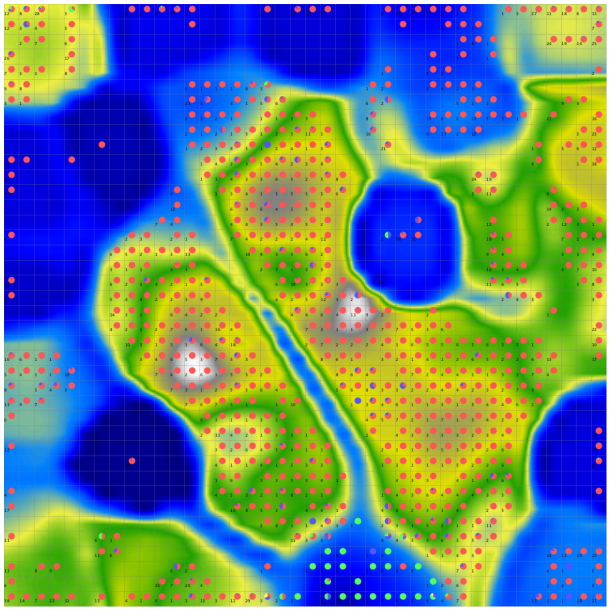

Figure 66: Optimal SOM: Smooth Histogram

the most.

- **Map-Size**: The map size is quite important for the visualization. When the map is too small, it is quite hard to distinguish clusters. However, we also realized, that the structure doesn't really change that much, it just gets clearer on bigger maps. Too big is also not good of course, since you cannot see anything on this scale and have to zoom a lot, which is quite painful. Overall the map size is only sensitive on small values for the cluster convergence, but is especially important for the visualization.

- **learnRate**: The learning rate is a quite sensitive parameter, which influences the cluster size and structure. If the value is too low, there is no cluster convergence, with too high values there are no clear clusters too see.

- **numIterations**: The number of iterations has quite huge impact if chosen too small. The clusters might not have converged already and the cluster structure changes rapidly in between iterations. Once the clusters have converged and are stable, bigger numbers of iterations have no more real impact, but just take longer to train. Therefore we recommend to take a quite high number of iterations and decrease it if the training takes too long. The number of iterations is very sensitive on the lower values.

- **sigma** (neighborhood radius): The neighborhood radius has a quite high impact on the result and is also quite sensitive. If it is too low the neighbors are not drawn with the best matching unit and therefore no clusters convergence. With too high values the boundary effect gets very strong and the correlation of single attributes blur out, which also means that we don't get the small clusters for single attributes being positive.

- **scaling**: The scaling is one of the most important parts of training a SOM. If the wrong scaling is chosen we get completely useless SOMs. This "property" is very sensitive and has to be chosen correct based on the knowledge about the data. Otherwise this can also lead to completely misinterpretation of the data.

5.2 Visualizations

While playing with the visualizations of the SOM we encountered the following views as useful:

- **Neighborhood graph**: The neighborhood graph gives a good overview about topology violations. When there many long lines for a small radius or with the knn of 1, this indicates a high number of topology violations. If only lines within clusters are visible we assume that we are on the right way with our topology.
- **Component planes**: The component planes helped us a lot finding correlations in our data. For example we could find out which parameters have a huge impact on the classes and which not. Additionally it helped us to understand the clusters, where they come from and what they tell us.
- **Class distribution**: The standard view with the small pie charts in every unit is of course one of the most important views to start with. It showed us the clustering of the primary hypothyroid patients and if that wasn't give we assumed that we have built a quite bad map.

Not very useful for us was the view **cluster connections**. It only showed us that the two big clusters are quite good connected and also between single values within the two clusters there seems to be not that big difference (mostly only the sex). However, this was information, we already gained from other views or our own interpretations. The metromaps where a visualization which didn't help us at all. We assume that this is due to the uniqueness of our data set.