Data integration with self-organising neural network reveals chemical structure and therapeutic effects of drug ATC codes

Ken McGarry[1] and Ennock Assamoha[1]

School of Pharmacy and Pharmaceutical Sciences,
Faculty of Health Sciences and Wellbeing,
University of Sunderland, City Campus, UK.
ken.mcgarry@sunderland.ac.uk

Abstract. Anatomical Therapeutic Codes (ATC) are a drug classification system which is extensively used in the field of drug development research. There are many drugs and medical compounds that as yet do not have ATC codes, it would be useful to have codes automatically assigned to them by computational methods. Our initial work involved building feedforward multi-layer perceptron models (MLP) but the classification accuracy was poor. To gain insights into the problem we used the Kohonen self-organizing neural network to visualize the relationship between the class labels and the independent variables. The information gained from the learned internal clusters gave a deeper insight into the mapping process. The ability to accurately predict ATC codes was unbalanced due to over and under representation of some ATC classes. Further difficulties arise because many drugs have several, quite different ATC codes because they have many therapeutic uses. We used chemical fingerprint data representing a drugs chemical structure and chemical activity variables. Evaluation metrics were computed, analysing the predictive performance of various self-organizing models.

Keywords: Kohonen, prediction, ATC codes, chemical fingerprints

1 Introduction

In this paper we describe how self organizing feature maps can provide classification labels for the so called drug therapeutic and anatomical codes (ATC). Most drugs have these codes allocated/annotated manually by experts and they provide useful information pertaining to drug specifications and characteristics. Researchers developing new drugs or repositioning existing drugs for novel applications can use the guidance provided by ATC codes to assist their efforts [10,9]. The ATC classification system classifies active drug ingredients into different levels, based on the drugs chemical properties, therapeutic properties, pharmacological properties and the organ/anatomical group which they target. ATC codes are highly prevalent in drug utilisation studies but they also provide a lot of information on the drugs pharmacological, chemical and therapeutic properties. The prediction of drug ATC codes can thus further be utilised in the fields of drug discovery, adverse drug effect prediction and drug repositioning.

The Kohonen self-organizing feature map (SOM) [5] is probably the best known of the unsupervised neural network methods and has been used in many varied applications. It is particularly suited to discovering input values that are novel and for this reason has been used in industrial, medical and commercial applications. The SOM has the ability to easily visualize difficult to interpret data, this is because of its topology-preserving mapping of the input data to the output units. Allowing a reduction in the dimensionality of the input data, making it more suitable for analysis and therefore also contributing towards forming links between neural and symbolic representations [12]. Furthermore, symbolic rules can be extracted from the code-book vectors and weights providing an explanation of the cluster boundaries [8], it is this feature we use to help uncover the relationships between independent variables that describe ATC drug boundaries.

In figure 1, the overall operation of data download preprocessing and statistical model development is shown. Data was downloaded from drugbank and chembl repositories to obtain the ATC codes and chemical structures. The side-effect information from SIDER4 database was used to augment the relationships between chemical structure and drug activity. In previous work we have related side-effect information with drug action similarity for drug re-purposing opportunities [9]

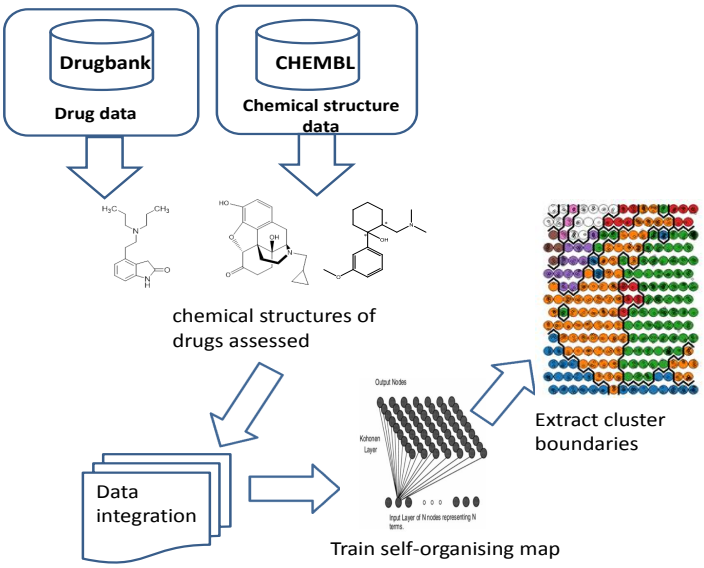

Fig. 1: System overview: data download, preprocessing and model building

1.1 Related work

Dunkel et al (2008) devised the *superpred* webserver, which constructs a structural fingerprint from a user defined molecule. The fingerprint was then compared to approximately 6000 drugs which had been enriched by approximate 7000 links to molecular drug targets; procured though text mining. However, Chen et al extended this method to the lower levels of ATC classification. In a later study, Chen et al introduced ontology information as well as chemical interaction and chemical structure information for the prediction of drug ATC codes [2,1]. Gurulingappa et al combined the techniques of information extraction and machine learning in order to assign ATC codes to drugs [4]. The method had good predictive accuracy but was only tested on drugs which had an ATC classification code for the cardiovascular anatomical group.

Wang et al, utilized a Support Vector Machine (SVM) learning in a method named Netpred ATC [14]. The Netpred method utilised chemical structure information and drug target protein information The results from Wang et al's method was deemed to outperform the superpred method of Dunkel [3]. The method was able to successfully predict the ATC codes of unclassified and classified drugs. A web service called SPACE (Similarity-based Predictor of ATC CodE) was also developed to predict a range of ATC codes for a given drug compounds and their probability scores [17]. Other methods incorporate data from various sources such as gene ontology, chemical and compound structures [7].

The R code used to perform the analysis and the datasets we have used are freely available on GitHub from: **https://github.com/kenmcgarry/ATC**

The remainder of this paper is structured as follows; section two describes our methods, indicating the types of data used and how we download and preprocessed it, along with the details of the self-organising feature map used to model this data, section three presents the results, section four provides the discussion and finally section five summarizes the conclusions and future work.

2 Methods

The ATC system is regulated by the World Health Organisation (WHO) and is the most renowned classification system in existence, used in a wide array of drug utilization studies. The ATC system consists of 5 grouping levels: with the detail of classification increasing as you progress from the top level (level 1) to through to the bottom level. The top level of the system classifies drugs based on the anatomical groups it targets. This level consists of 14 anatomical group which a drug may possibly target. The second level depicts the pharmacological/ therapeutic sub group of a drug. The third level and the fourth level consist of the chemical/pharmacological/therapeutic subgroups. Moreover, the fifth level pertains to the chemical substance.

For example, the Drug Amlodipine has an ATC code of C08CA01 see table 1. It is not necessary for a drug to only have one ATC code. A drug substance can have

Code	Description
C	Cardiovascular System (1st level, anatomical main group)
C08	Calcium Channel Blockers (2nd level, therapeutic subgroup)
C08C	Selective Calcium Channel Blockers with Mainly Vascular Effects (3rd level, pharmacological subgroup)
C08CA	Dihydropyridine Derivatives (4th level, chemical subgroup)
C08CA01	Amlodipine (5th level, chemical substance)

Table 1: Example of ATC for the drug Amlodipine

more than one ATC code assigned to it depending on whether it is available in different dosage formulations or strengths with therapeutic indication that are clearly different from one another. This can be seen in the drug acetylsalicylic acid. Acetylsalicylic acid has three different ATC codes: B01AC06 for when it is used as a platelet aggregation inhibitor, A01AD05 for when it is used for local oral treatment and N02BA01 for when it is used as an analgesic and antipyretic. The level 1 code for all categories is shown in table 2.

Code	Description	Code	Description
A	Alimentary tract and metabolism	L	Antineoplastic and immunomodulating
B	Blood and blood forming organs	M	Musculo-skeletal system
C	Cardiovascular system	N	Nervous system
D	Dermatologicals	P	Antiparasitic products, insecticides and repellents
G	Genito-urinary system and sex hormones	R	Respiratory system
H	Systemic hormonal preparations	S	Sensory organs
J	Anti-infectives for systemic use	V	Various

Table 2: Level 1 ATC codes

The basic Kohonen SOM has a simple 2-layer architecture. Since its initial introduction by Kohonen several improvements and variations have been made to the training algorithm. The SOM consists of two layers of neurons, the input and output layers. The input layer presents the input data patterns to the output layer and is fully interconnected. The output layer is usually organised as a 2-dimensional array of units which have lateral connections to several neighbouring neurons. The architecture is shown in figure 2.

The objective is to build a self-organising feature map in a two stage process of training the network and then passing a vector of test data through the network and observing the active neurons. A well trained network will have different neurons respond to specific input patterns. During training the network requires exposure to patterns which will modify the inter-neuron connections during the learning phase. Competitive learning is ensures that the best matching unit (neuron) will be activated (competes) in preference to the other units in the network. In each training step, the algorithm will calculate the changes to the synapses for every new input, D_j for each neuron:

The competitive learning process is presented in equation 1 and the best matching neuron is derived from equation 2.

$$D_j = \sum_{i=0}^{p} ||I_i - W_{ij}|| \tag{1}$$

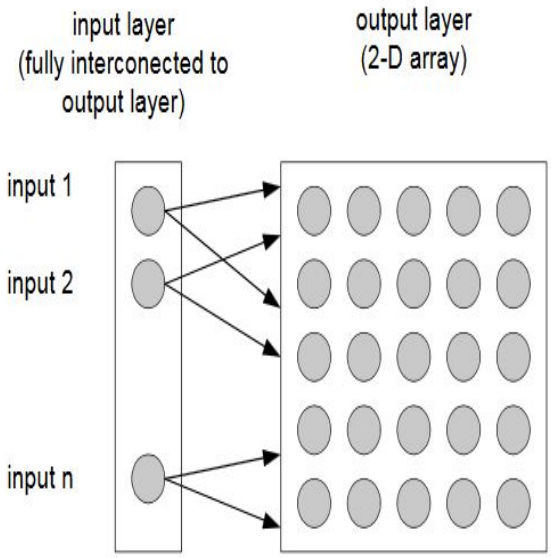

Fig. 2: Architecture of self-organising map

Where: D_j is the distances for each neuron, I_i is the current input vector and W_{ij} is the weight vector.
We then select the Best Match Unit (BMU) and update the weight vectors of the map according to Equation 2.

$$W_i(t+1) = W_i + h_{ci}(t) * (I(t) - W_i(t)) \tag{2}$$

where t denotes the time and h_{ci} is the neighborhood kernel around the BMU. W_i is the weights attached to that unit.

However, key to understanding the kohonen network is the so called unified or U-matrix decomposition method. This enables the cluster boundaries to be made visible to the eye [13]. The matrix output uses relative distance between reference vectors to find cluster boundaries. Given an M x N lattice, the Euclidean distances associated with the reference vectors of the adjacent cells, such as M_{i-1}, M_{i+1} are summed, $M_{i,j}$ is designated as M adjacent (i, j), and d represents the Euclidean distance, then, according to U is now plotted using equation 3.

$$U_{(ij)} = \sum d(M_{adjacent}(i, j), M_{i,j})) \tag{3}$$

When the matrix is plotted, the cluster boundaries are generally dark colours while the clusters form lighter coloured spaces in between.

CHemBL is an important database containing chemical compounds usually represented as string data called the SMILES (Simplified Molecular Input Line Entry Sys-

tem) format, various algorithms have been developed that can generate numbers describing the compound [16]. The majority of our chemical data was downloaded and stored in SDF format, these are plain text files with a specific internal format. The format is extensible, a single file can contain a single chemical structure or millions of structures. The SDF text files have a fairly straightforward structure although can be inefficient in memory storage since a lot of whitespace and unnecessary characters are used to represent the chemical structures, see figure 3. This redundancy dates in part from earlier legacy file systems, programming languages and parsing techniques.

```
1
2    175579
3  -OEChem-02060822402D
4
5  19 20  0     1  0  0  0  0  0999 V2000
6    0.7103   -1.0586    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
7   -2.1414   -0.2414    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
8   -0.0034    1.8241    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
9    0.7138   -0.2345    0.0000 C   0  0  3  0  0  0  0  0  0  0  0  0
10   0.0000   -0.6517    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
```

Fig. 3: SDF chemical structure file, showing first few lines for the Tramadol drug. The first line is the drug id, second line refers to the software package that created it. The next few lines describe the properties which state the number of atoms and bonds, each atom and bond on a separate line.

In algorithm 1 the training data generation, processing and neural network processing is clarified, as a series of steps.

Algorithm 1 Data generation, processing and Kohonen training

procedure TRAINKOHONEN($CheMBL$ chemical structures, ATC codes from $DrugBank$)
  do initialize
  $cStruc \leftarrow$ only get drugs with chemical structures[$CheMBL$]
  $aCodes \leftarrow$ only get drugs with ATC codes[$DrugBank$]
  $numStruc \leftarrow$ length[$cStruc$] ▷ How many useful chemicals do we have?
  $Kohonen \leftarrow$ [N x M] ▷ Setup small 5 x 5 matrix of nodes
  end initialize

  for $i \leq numStruc$ **do** ▷ process every drug with a chemical structure
    $FP_i \leftarrow Fingerprint(cStruc)$ ▷ Convert from SDF to binary fingerprints
    $FP_i \leftarrow Fingerprint(aCodes)$ ▷ Attach ATC code as class label to binary fingerprints
    $FP_{train} \leftarrow randomsample(FP_i, 80)$ ▷ split data 80/20% train/test
    $FP_{test} \leftarrow randomsample(FP_i, 20)$ ▷ split data 80/20% train/test
  end for
  repeat
    Train Kohonen $[FP_{train}]$
    Test Kohonen $[FP_{test}]$
    Modify Kohonen architecture, $[NxM]$ ▷ Manual intervention, increase until 15x15
  until $Kohonenerror \leq 0.01$ ▷ stop training when error reaches cutoff point
  $GCs \leftarrow CalcNetworkStatistics(Cw)$ ▷ call
  inspect Mapping ▷ visually inspect patterns to neurons
  inspect Umatrix ▷ visually inspect umatrix
end procedure

The chemical database was then imported into the system using SDF format data. The database consisted of 7,759 drugs with 12 variables relating to chemical structure. Chemical fingerprints were created from these structures - each fingerprint is a binary matrix with 1 = structure present or 0 = structure absent. A random sample of 80% was taken from the dataset of 1,334 drugs and 20% for test, with chemical fingerprints assigned to each drug (representing the drug chemical structure). The training dataset consisted of 1,067 drugs with chemical fingerprints. This was presented to a Kohonen network with and architecture of 15 x 15 nodes. We use the Kohonen in a semi-supervised way, i.e. we have class labels (ATC codes).

We implemented the system using the R language with the RStudio programming environment, on an Intel Xenon 64-bit CPU, using dual processors (3.2GHz) with six cores, and 128 GB of RAM. R is primarily a statistical data analysis package but is gaining popularity for various scientific programming applications and is very extendable using packages written by other researchers [11]. We used the following R packages: Kohonen [15]. Since it is an interpreted language, R is generally quite slow compared with a compiled language. However, we used the MicroSoft R Open system (https://mran.microsoft.com/) because it is optimized to take advantage of processor cores and the majority of its mathematics/matrix operations are rewritten in C++ to speed up operation. It is fully compatible with the oringinal CRAN version of R.

3 Results

The DrugBank database was integrated into our system because it contains the majority of drugs that are currently prescribed, or have been withdrawn or are at the clinical trial stage. This resource is widely used by those developing drugs, chemists, pharmacologists and others involved in pharmaceutics research [6]. Every drug is listed with its main targets, known off-targets along with chemical structure and other important characteristics.

Figure 4 contains the plot showing the training progress, during training, the codebook vectors are becoming more and more similar to the closest objects in the dataset.

When the test data is passed to the self-organising map we obtain the following confusion matrix based on the success or otherwise. It is evident that the ATC classes A, J, M and N consistently recorded high class accuracy, recall, precision and f1 scores for the self-organising map. The confusion matrix relating the accuracy of the various ATC codes is displayed in figure 5.

The correct classifications for the test data are on the upper left to bottom right diagonal, any misclassification's are located off-diagonal and reveal which class they were misclassified as. For example the ATC code 'A' has 12 correctlty identified test cases but one sample is misclassified as 'C', five misclassified as 'D' etc.

In figure 6 we reproduce the effects of the influence of the independent variables on the neurons. Here, for simplicity a 5 x 5 grid is extracted from the 15 x 15 grid of neurons. Each of the 12 independent variables will be colour coded and similar to a pie-chart the magnitude and orientation of the slice indicates its influence on the neuron.

The Jchem variables identified in figure 6 contained information on: acceptor count, average polarizability, donor count, ALOGPS LogP, Jchem LogP, ALOGPS LogS,

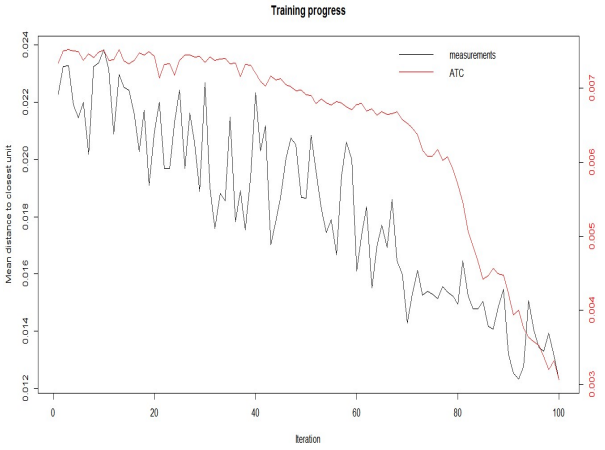

Fig. 4: Training progress of kohonen self-organizing map

number of rings, physiological charge, strongest basic pka, polar surface area, refractivity and rotatable bond count.

Plotting the training data to the 15x15 grid of neurons reveals why a classification algorithm would have difficulty. In figure 8 we can see that only a few neurons have unique ATC codes mapped to them. Most units have 2-3 sometimes more ATC codes, ideally for 100% accuracy would would need to have each ATC code mapped to its own unique neuron. There are also several neurons that could not get codes mapped to them. These seem to form a border or boundary isolating the upper right part of figure 8. The locations of the circles indicate the neurons to which the samples have been mapped. The relationship between figure 8 and figure 9 involves the mapping of particular input vectors to specific neurons thus enabling the cluster boundaries to be made visible.

Despite their excellent capability of visualization, SOMs cannot provide a full explanation of their structure and composition without further detailed analysis. One method towards filling this gap is the unified distance matrix or U-matrix technique of Ultsch [12]. The U-matrix technique calculates the weighted sum of all Euclidean distances between the weight vectors for all output neurons. The resulting values can be used to interpret the clusters created by the SOM. The rather confusing picture presented by figure 8 indicates that several neurons have many different ATC codes assigned to them (in fact the preference is for one ATC code for each neuron). The unmatrix shown in figure 9 indicates that several large boundaries exist.

Although, primarily used for situations where there are no class labels, and hence the natural structure of the data is important - the self-organizing feature map can also be used where class labels exist and semi-supervised operation is useful. This will provide more information to explain the key features of a dataset and a limited explanation of why the Kohonen organized the data they way it did [8].

The U-matrix in figure 9 is colour coded for each main cluster and has the umatrix boundary drawn to emphasize these. Any neuron near a class boundary can be expected

actual \ predicted	A	B	C	D	G	H	J	L	M	N	P	R	S	V
A	12	0	1	5	2	1	1	4	0	4	0	1	1	1
B	3	1	3	0	1	0	1	1	0	0	0	0	0	0
C	5	0	4	0	3	1	4	5	0	8	1	1	2	3
D	2	0	0	3	1	0	2	2	0	1	1	1	0	2
G	0	1	2	1	1	0	1	2	1	4	0	0	0	0
H	0	0	0	1	1	0	1	0	0	0	0	0	1	0
J	5	2	2	1	1	0	13	3	0	3	0	0	1	0
L	0	1	2	0	3	4	1	3	2	0	0	2	0	2
M	1	0	1	1	1	1	0	0	7	5	0	1	0	0
N	4	0	4	1	1	2	3	1	2	16	2	2	2	1
P	2	1	0	1	0	0	0	1	0	1	1	1	0	0
R	0	1	3	1	0	2	3	1	3	1	1	3	0	1
S	0	1	1	0	1	1	1	1	0	1	0	0	1	0
V	0	1	0	0	1	0	3	0	0	0	1	1	0	2

Fig. 5: Confusion matrix for the 14 ATC codes

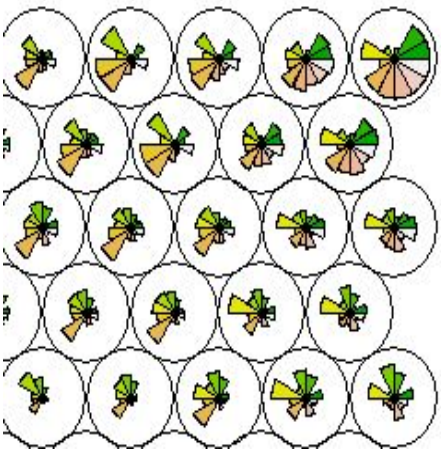

Fig. 6: Relative weights and effects of independent variables

Fig. 7: Colour scheme for identifying independent variables shown in figure 6

to have higher average distances to their neighbours, indicating structural consistencies in the data. Interpreting the u-matrix diagram with figure 8 which highlights the mapping of the ATC labels has produced an area of 'no mans land' denoted by the empty neurons not allocated to a code appears to run counter to the u-matrix in the sense that it does not seem to be part of a boundary. It is probably a sign that the 15 x 15 map is too large for the available data but changing the map to smaller grids did not improve the training or classification issues. The majority of the boundaries can be explained by the simple fact that although the drugs have different ATC classifications, their chemical structure is very similar in many cases. In fact only slight changes to chemical structure are required for very different pharmacological properties to be exhibited by a drug. Therefore, some neurons think they doing a good job of identifying drugs based on the input vectors, however since we have access to the ATC labels we know this is not quite the case.

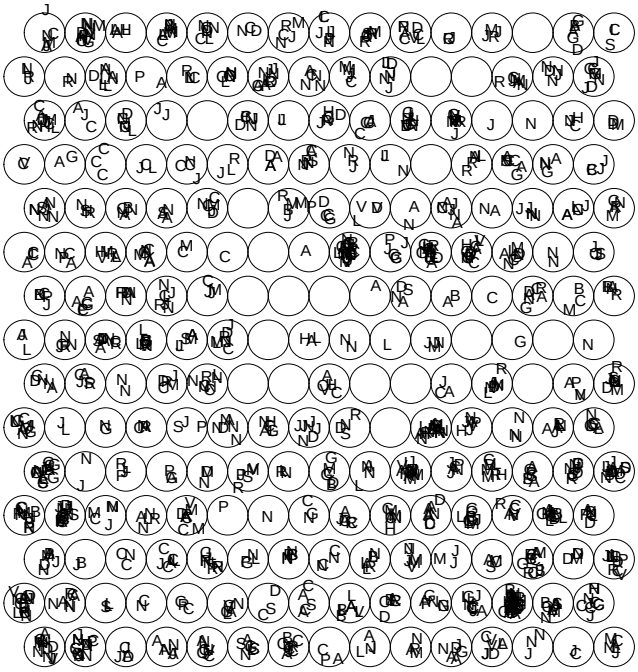

Fig. 8: Mapping ATC codes to 15x15 self-organising map

4 Discussion

The training of the Kohonen map clearly showed a decrease in the mean distance to the closest unit, as should be the case in a successfully trained Kohonen neural network. Even so, an improvement in the training process can be made as even after 100 iterations, it could be seen that the mean distance to the closest unit was still decreasing. Thereby indicating that the training sequence may not have been fully complete and that further iterations were needed. The optimum number of iterations needed can be known when the mean distance to the closest unit becomes relatively stable (flat) and begins to merely fine tune the value. The mean distances of objects that were mapped to a unit to the codebook vectors (of the units these objects were mapped to) was generally very low across all units; with only one unit/neuron displaying a mean distance above 1.0. In most cases, this meant that the objects were well represented by the codebook vectors and thus highlights the success of the Kohonen neural network method. Future work must consider adding further information to the chemical data to resolve the mapping issues.

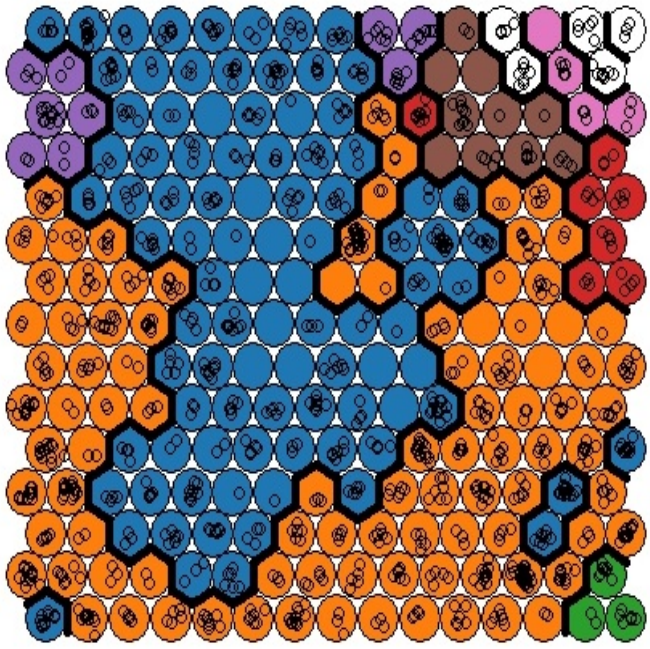

Fig. 9: Umatrix revealing cluster boundaries

5 Conclusions

The novel contribution of this work relates to the explanatory ability of the Kohonen network to reveal the internal structure of the clusters and input to output class label mapping. Future work will address integrating other sources of drug/chemical information to improve accuracy. The issue of drugs with multiple ATC codes in different therapeutic areas also needs to be resolved as it is a source of bias and class fuzziness.

References

1. Chen, F., Jiang, Z.: Prediction of drug's anatomical therapeutic chemical (ATC) code by integrating drug-domain data. Journal of Biomedical Informatics 58, 80–88 (2015)
2. Cheng, X., Zhao, S., Xiao, X., Chou, K.: iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic codes. Bioinformatics 33(3), 341–346 (2016)
3. Dunkel, M., Gunther, S., Ahmed, J., Wittig, B.: Superpred: drug classification and target prediction. Nucleic Acids Research 36, W55–W59 (2008)
4. Gurulingappa, H., Kolarik, C., Hofmann-Apitius, M., Fluck, J.: Concept-based semi-automatic classification of drugs. Journal of Chemical Information Modeling 49(8), 1986–1992 (2009)
5. Kohonen, T., Oja, E., Simula, O., Visa, A., Kangas, J.: Engineering applications of the self-organizing map. Proceedings of the IEEE 84(10), 1358–1383 (1996)

6. Law, V., Knox, C., et al, Y.D.: Drugbank 4.0: shedding new light on drug metabolism. Nucleic Acids Research 42, D1091–D1097 (2014)
7. Liu, Z., Guo, F., Gu, J., Wang, Y., Li, Y., Wang, D., Li, D., He, F.: Similarity-based prediction for anatomical therapeutic chemical classification of drugs by integrating multiple data sources. Bioinformatics 31(11), 1788–1795 (2015)
8. Malone, J., McGarry, K., Bowerman, C., Wermter, S.: Rule extraction from kohonen neural networks. Neural Computing Applications Journal 15(1), 9–17 (2006)
9. McGarry, K., Daniel, U.: Data mining open source databases for drug repositioning using graph based techniques. Drug Discovery World 16(1), 64–71 (2015)
10. McGarry, K., Slater, N., Amaning, A.: Identifying candidate drugs for repositioning by graph based modeling techniques based on drug side-effects. In: The 15th UK Workshop on Computational Intelligence, UKCI-2015. University of Exeter, UK (7th-9th September 2015)
11. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2015), https://www.R-project.org/
12. Ultsch, A., Korus, D.: Automatic acquisition of symbolic knowledge from subsymbolic neural nets. In: Proceedings of the 3rd European Conference on Intelligent Techniques and Soft Computing. pp. 326–331 (1995)
13. Ultsch, A., Mantyk, R., Halmans, G.: Connectionist knowledge acquisition tool: CONKAT. In: Hand, J. (ed.) Artificial Intelligence Frontiers in Statistics: AI and statistics III, pp. 256–263. Chapman and Hall (1993)
14. Wang, Y., Chen, S., Deng, N., Wang, Y.: Network predicting drug's anatomical therapeutic chemical code. Bioinformatics 29(10), 1317–1324 (2013)
15. Wehrens, R., Buydens, L.: Self and super-organising maps in r: the Kohonen package. Journal of Statistical Software. 21(5) (2007), http://www.jstatsoft.org/v21/i05
16. Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of Chemical Information and Modeling 28(1), 316 (1988)
17. Wu, L., Liu, N., Wang, Y., Fan, X.: Relating anatomical therapeutic indications by the ensemble similarity of drug sets. Journal of Chemical Information and Modeling 53, 2154–2160 (2013)