Therapeutic Drug Classification

João Roque and Ricardo Margarido

University of Coimbra

Abstract. This project was made on scope of the course Introduction to Bioinformatics in the University of Coimbra. The objective of this work is to classify drugs in 18 specific therapeutic drug families, using the KEGG database. With this database it was possible to extract exact mass, mol weight and formula for 1491 different drugs. After implementing and testing various clustering and machine learning methods, it became clear that the mol weight and exact mass of the drug were not statistically significant for this problem. However, by using the formula, we were able to achieve 46% accuracy and 39% recall rates across the 18 different classes and a 87% accuracy, 89% sensibility and 80% specificity on a classifier built to differentiate only one class (Antineoplastics) from the rest of the drugs.

1 Introduction

The most used system worldwide is the ATC - Anatomic Therapeutic Chemical classification system, recommended by the World Health Organization. Therapeutic Drug classes are a way of classifying medical drugs according to their functions and chemical properties. If two drugs are in the same class, it means that they are intended to treat the same medical conditions. The therapeutic classes considered in this work were: **a)** Antineoplastics **b)** Antibacterials **c)** Antivirals **d)** Antifungals **e)** Antiparasitics **f)** Antidiabetics **g)** Hypolipidemic agents **h)** Osteoporosis drugs **i)** Cardiovascular agents **j)** Psychiatric agents **k)** Neurological agents **l)** Anti-allergic agents **m)** Anti-rheumatic and anti-gout drugs **n)** Anesthetics, analgesics and anti-inflammatory drugs **o)** Respiratory agents **p)** Gastrointestinal agents **q)** Endocrine and hormonal agents **r)** Dermatological agents.

1.1 State of the Art

When we first approached this subject, we looked at what has been done and what is a good starting point. The most accepted methods for drug-target interaction are, by far, in-vitro experiments. This biochemical procedure can give reliable results about whether a given drug interacts or not with the given target. Although these methods are reliable they are also costly and very time consuming because the reactions can take a lot of time and the molecules involved can be expensive.

So, naturally, researches turned to computers to see if this interaction could be simulated and predicted by the machines that are of disposal. This makes the overall process much cheaper and faster when compared to the in-vitro methods. Docking and machine learning are the most used approaches as of now.

Docking relies on the 3D structure of the molecules so it is able to compute the binding site of the proteins or molecules and predict if the drug does in fact interact with the target. The main issue is that the 3D structure is rarely available and when it is all the calculations are heavy and take a lot of computational power to do.

Logically another method is needed and that is where machine learning enters the scene. Offering a lighter set of calculations, it is possible to test a lot more samples in the same given time. To do this comparisons and predictions SVMs are used. When the SVMs dont perform as expected then the similarities with known compounds are calculated and predictions made on that basis.

1.2 Our Approach

To see if other methods could be used to tackle this problem we were set to try Random Forests based of the similarities between compounds.

However, when studying this classification system, we discovered that the same drug can be categorized on multiple classes. This is a very important detail, because the same drug with the same features might have different targets, which can lead to a more inaccurate classifier. Since drugs in the same class are used to treat the same system, it makes sense that the chemical reactions that occur between the drug and each system have similarities. However, this might not be true for every class, because a drug might tackle the same medical condition in different ways than other drugs in the same class. So, we know beforehand, that our classification is dependent on the way drugs in the same group react. This also leads us to believe that the most valuable feature that we were able to obtain is each drug formula, since the mol weight and exact mass are not that relevant in this type of interactions.

In the first place, we tested the mol weight and the exact mass with different clustering methods. They revealed themselves to be inefficient to distinguish all classes. We decided to leave these features, and only use the formula, since it makes more sense to tackle this problem.

After this, we used the formula for each drug as feature for several classifiers. We noticed that some drugs were being classified better than others. In this work, we try to give an answer as to why this happens, and ultimately, we prove that one single drug class can be distinguished from the rest only by it's written formula, specially if in this class, the chemical structure of the drug is one of the most important points in treating the specific class medical condition.

In reference papers, the authors tackled this type of problem by analyzing the similarity between structures in each class, in target proteins, physicochemical properties, etc. Our new approach, lies on only using the written formula from a drug to classify it.

2 Methods

2.1 Extracting KEGG drug data

To use data from KEEG, we extracted the source html and used www.html-cleaner.com to remove the tags. We then processed this code, and obtained all the drugs name tag. Using the KEEG API, we were able to extract mol weight, exact mass, formula, and target family for each drug.

2.2 Mol Weight and Exact Mass - Clustering and Classification

To check if mol weight and exact mass were statistically significant features to determine each drug class, we decided to make two clusterings. First, we used a KMeans clustering with 18 centroids, to see if it was possible to divide all the drugs in the proper 18 families. In the second place, we used a hierarchical clustering method called MeanShift, that decides by himself, how many clusters the data can be divided in.

After that, we implemented a Random Forest classifier with 100 trees using both features.

2.3 Classification with Drugs Formula

To adapt the drug formula to a classifier, we first processed the formulas in the following way: we created a vector with the same length as the number of atoms in the periodic table, and then, counted how many elements there were in each formula. These elements were added to the vector in their proper position. In the end, we had a vector for each drug with their formula information, that we used as feature.

After this, we needed to take care of the fact that the same drug might have multiple targets. To solve this in a simple way, we decided to not use drugs with multiple targets, deleting them from our data. The initial number of drugs was 1491, and ended up as 1122.

Following this process, we decided to apply the newly obtained data to a Random Forest Classifier. This classifier had 1000 trees, and ran 200 times. In each run, the training and testing group were reassigned randomly. We noticed that some families were classified significantly better, which ultimately led us to do a simple A to B classification with the family of Anti-Bacterials.

2.4 A to B classification - Antibacterials and other drugs

Since Anti-Bacterials was the class with the best classification rate, we decided to change our targets. Every drug that did not belong to the Anti-Bacterials family became the class 0, and the Anti-Bacterials became the class 1. This division causes a disparity between classes, that we solved by randomly deleting drugs from class 0.

This data division allowed us to perform an A to B classification, performed with a Random Forest classifier with 250 trees.

3 Results

3.1 Mol Weight and Exact Mass - Clustering and Classification

The first method implemented was the KMeans clustering. We can see the plot of the centroids on Fig. 1.

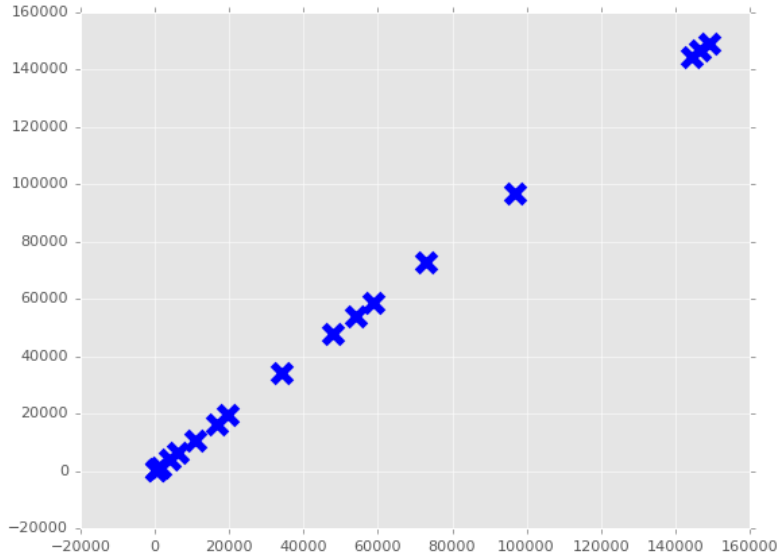

Fig. 1. Plot of the centroids of the clustering KMeans with Exact Mass and Mol Weight.

The Hierarchical clustering MeanShift with the same data only discovered 12 centroids.

The Random Forest classifier with this data only had an accuracy of 20%. We did not perform any additional metric on this classifier since the initial results were already very poor, which was according to what we expected.

3.2 Classification with Drugs Formula

Using the formulas scores as feature in a Random Forest Classifier improved the previous result by a considerable margin. The accuracy rate grew to 46% with a recall rate of 39%, across 200 runs. The accuracy by class can be seen in Fig. 2., along with the multiclass confusion matrix in Fig. 3. . In Fig. 2., some bars are non-existent, because in some runs, there were no elements of this class. This would cause a mean to be calculated with a nan (not a number) value, which would make the final result to end up as nan. This is something that can definitely be improved on future work.

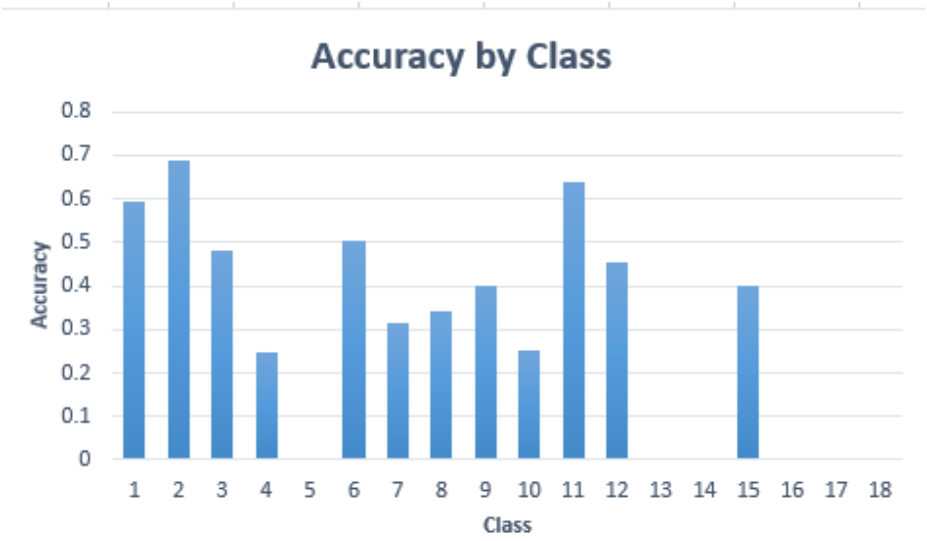

Fig. 2. Accuracy by class using the Random Forest classifier with Formulas score as feature across 200 runs.

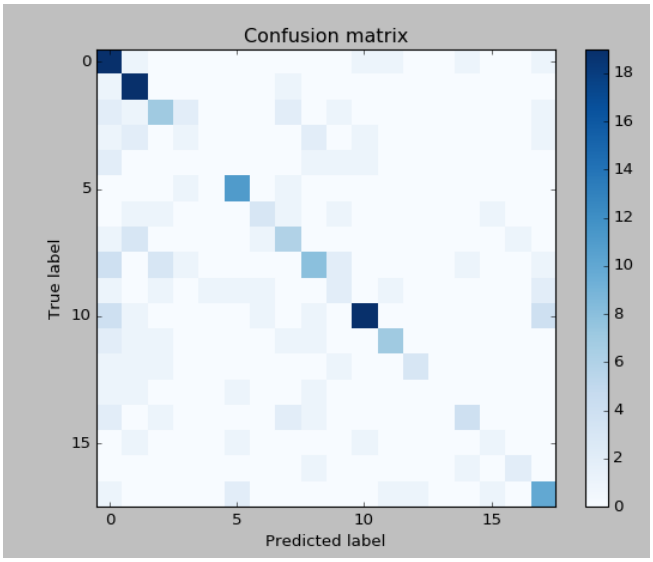

Fig. 3. Confusion matrix from the Random Forest classifier with Formulas score as feature across 200 runs.

3.3 A to B classification - Antibacterials and other drugs

The classification using only antibacterials and other drugs turned out to have very good results. One of the best runs showed an accuracy of 87%, a sensibility

of 89% and a specificity of 80%. In the Fig. 4., we can see the confusion matrix for this specific run, and in the Fig. 5., it's ROC curve.

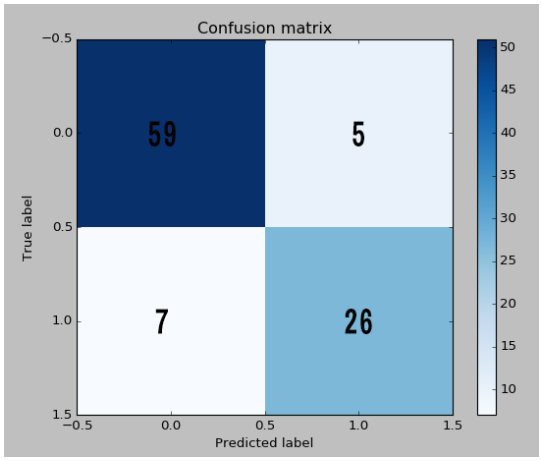

Fig. 4. Confusion matrix from classification with Antibacterials and other drugs.

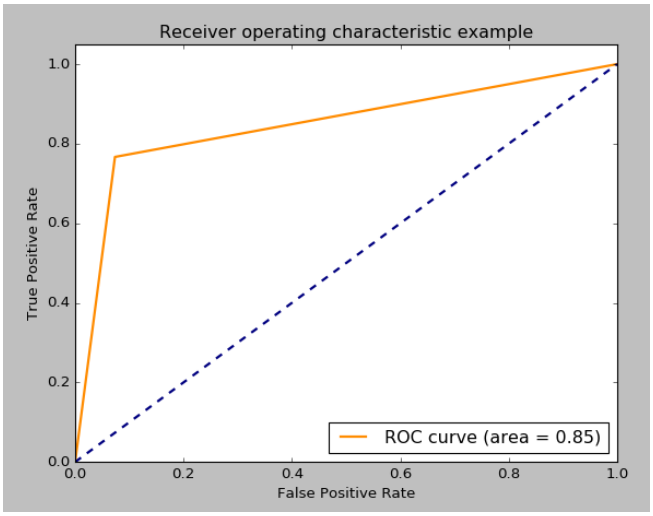

Fig. 5. ROC curve from classification with Antibacterials and other drugs. The AUC is 0.85 .

4 Discussion

When we started this work, we intended to do a drug-target study of the drugs available in the KEGG database. However, after spending some time investigating and learning about the website, we noticed that these drugs were divided in 18 families. This matter caught our attention, and after downloading each drug feature, we decided to investigate it further. We learned that the Therapeutic Drug division had to do with the medical condition each drug treats and it's chemical characteristics, which means drugs in the same class might bind to the same places to treat some human systems. It is also generally accepted that compounds with similar physicochemical properties exhibit similar biological activity. By acknowledging that, we realized that drugs in the same family could be distinguishable by it's formula. From this line of thought, we already knew that the mol weight and exact mass for each drug were not very good to classify each class. We still performed some clusterings and classifiers with both these features, but after accuracies hovering around 20%, we decided that investigating the formulas was definitely the way to go.

The authors from the referenced articles analyzed this matter through chemical similarities or chemical interactions. In our work, we wanted to know if it was possible to properly classify drugs according to the ATC classes only using the formula, which means, by chemical similarities. By starting the problem with a general multiclass classifier, we had some interesting discoveries. By only using the formula, we achieved 46% accuracy and 39% recall. However, when analyzing class by class accuracy, we noticed that some classes had much higher accuracies than others. This means that not all classes can be divided by it's chemical similarities. The 3 classes that were classified better were the Antibacterials, the Antineoplastics and Neurological Agents. Since we mostly studied the Antineoplastics, this is the class we'll mostly discuss.

Antineoplastics or anticancer drugs affect the process of cell division. They mostly damage the DNA and initiate the cell apoptosis. Since these drugs mostly bind to the DNA, their base structures are pretty similar, which makes it easier to classify them all into one class by only using the formula. As we can see in Fig. 4., the classification of this drug went very well and matched what was expected. It also proves that some drug groups can indeed be distinguished only by their formula. This type of work can help doctors and researchers, for example, to find the system a newly discovered drug can treat, that they are not yet able to test on humans.

5 Conclusion

In this work, we were able to distinguish with an 87% accuracy an Anti-neoplastic drug from other drugs, only by using their written chemical formula. This led us to conclude that some types of drug classes from the ATC have drugs with very similar chemical compositions, that are ultimately very important in treating the designated medical condition.

In the future, we hope to improve aspects such as: try other classifiers that might adjust better to our data, find another effective way to deal with the drugs that have multiple targets and use other features to improve classifier accuracy on all classes.

References

1. Mathias Dunkel, Stefan Gnther, Jessica Ahmed, Burghardt Wittig and Robert Preissner: SuperPred: drug classification and target prediction. Nucl. Acids Res. (2008) 36 (suppl 2): W55-W59.
2. Lei Chen, Wei-Ming Zeng, Yu-Dong Cai , Kai-Yan Feng, Kuo-Chen Chou: Predicting Anatomical Therapeutic Chemical (ATC) Classification of Drugs by Integrating Chemical-Chemical Interactions and Similarities. April 13, 2012
3. Zhongyang Liu, Feifei Guo, Jiangyong Gu, Yong Wang, Yang Li, Dan Wang, Liang Lu, Dong Li, and Fuchu He: Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources. Bioinformatics (2015)