Prediction of Drug Classes based on Gene Expression Data

Master's Thesis
Li Yinghua
Faculty of Medicine and Life Sciences
University of Tampere
May 21, 2018

Master's Thesis

Place:	Faculty of Medicine and Life Sciences
	University of Tampere
Author:	Li Yinghua
Title:	Prediction of Drug Classes based on Gene Expression Data
Pages:	54
Supervisor:	Professor Frank Emmert-Streib, Dr. Juha Kesseli
Reviewers:	Professor Matti Nykter, Dr. Juha Kesseli
Date:	21.05.2018

Abstract

Nowadays, the financial investments in pharmaceutical research and development are an enormous increase. Drug safety is very important to health and drug development. Finding new uses for the approved drug has become important for the pharma industry. Drug classification accuracy helps identify useful information for studying drugs, also helps in accurate diagnosis of drugs. Gene expression data makes a possible study of biological problems and machine learning methods are playing an essential role in the analysis process. Meanwhile, many machine learning methods have been applied to classification, clustering, dynamic modeling areas of gene expression analysis.

This thesis work is using R programming language and SVM machine learning method to predict the ATC class of drugs based on the gene expression data to see how well the gene expression patterns correlate after treatment within the therapeutic / pharmacological subgroup. A dimensionality reduction method will use to reduces the dimensions of the dataset that improves the classification performance. The classifiers built using SVM machine learning technique in this thesis study had limited with detecting drug groups based on the ATC system.

Acknowledgment

This thesis work was conducted at Signal Processing department by Professor Frank Emmert-Streib at the Tampere University of Technology. I would like to thank my supervisor, Professor Frank Emmert-Streib, who gave me the opportunity to conduct this thesis research and patiently guided me through the project. Thank Tampere university lecturer Juha Kesseli and Professor Matti Nykter, who taught me during the master degree study and patiently helped me through the studies. I also want to thank Aliyu Musa for his help and suggestions.

Finally, I would like to thank my family who were always supporting and encouraging to me.

Tampere,

Li Yinghua

Table of Contents

List of Tables

List of Figures

Abbreviations

Abbreviation	Meaning
ACC	Accuracy
ATC	Anatomical Therapeutic Chemical
AUCROC	Area Under the Receiver Operating Characteristic curve
CMap	The Connectivity Map
CV	Cross Validation
C-SVC	C-Support Vector Classification
FDA	Food and Drug Administration
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
KNN	K Nearest Neighbors
LIBSVM	Library for Support Vector Machine
MAP	Maximum a Posteriori
Multiclass ROC curve	Multiclass Receiver Operating Characteristic curve
NPV	Negative Predictive Value
PC	Principal Component
PCA	Principal Component Analysis
PPV	Positive Predictive Value
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TN	True negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
v-SVC	v-Support Vector Classification
v-SVR	v-Support Vector Regression
ε-SVR	ε-Support Vector Regression
WHOCC	World Health Organization Collaborating Center for Drug Statistics Methodology

Chapter 1. Introduction

Nowadays, the financial investments in pharmaceutical research and development are an enormous increase. Finding new uses for the approved drug and reducing costs to providing new treatments for unmet medical needs has become important for the pharma industry (Napolitano et al., 2013). The prediction of drug classification provides valuable information for drug discovery and repositioning, helps new drug development and helps to understanding of drugs' chemical, pharmacological and therapeutic properties (Lin, Lin, & Weng, 2007). A large quantity of gene expression compounds which are approved by Food and Drug Administration, have been measured on several cultured human cell lines with the purpose to inquire into comparability between drugs mechanisms of action (Iorio et al., 2010; Lamb et al., 2006; Napolitano et al., 2013). Computational approaches have been designed for drug repositioning to find correlations of expression signatures between drug associated and disease associated. Recently many comprehensive methods such as molecular, biological aspects of the drug-disease interactions and account chemical have been proposed (Gottlieb, Stein, Ruppin, & Sharan, 2011; Napolitano et al., 2013). For instance, Chen et al. were able to map the chemical substructure of drugs, described by the Anatomical Therapeutic Chemical (ATC) classification system to predict chemical similarities (Chen, Zeng, Cai, Feng, & Chou, 2012). The study used different benchmarks to predict level 1 of ATC codes by collecting chemical data for 3,883 drugs with high classification performance (73% accuracy). Napolitano et al. proposed a novel computational method to predicts drug repositioning based on machine learning algorithms to predict the 2nd level (therapeutic class) of FDA-approved compounds with high accuracy levels (78%) (Napolitano et al., 2013). Therefore, drug classification accuracy helps identify useful information for studying drugs, also helps in accurate identification of drugs.

Gene expression profile has been used to draw forth show cryptic subtypes of diseases. At the same time, potentially, it forecasts biomedical issues which significantly regards to human's health (Khan et al., 2001). It measures the activity of genes and creates a global picture of cellular function. In drug discovery, gene expression profile may be used to diagnose a disease or condition. In order to study the effect of drug and small-molecule perturbations on different types of human cells, rich information is gathered from datasets generated from Library of Integrated Network-based Cellular Signatures (LINCS) (Genometry, 2017). All signatures are

represented by vectors measuring differential expression in the case of L1000 data ("LINCS Data Portal," 2017).

The Anatomical Therapeutic Chemical (ATC) classification system is an approved classification system for drugs which recommended by the World Health Organization Collaborating Center for Drug Statistics Methodology (WHOCC) (Z. Liu et al., 2015). According to the organ or system on which they act and pharmacological, therapeutic and chemical properties, drugs are classified into five different levels in the Anatomical Therapeutic Chemical (ATC) classification system (https://www.whocc.no/atc/structure_and_principles). For example, L (Antineoplastic and immunomodulating agents, the first level), L01 (Antineoplastic agents, the second level), L01X (Other antineoplastic agents, the third level), L01XE (Protein kinase inhibitors, the fourth level) and L01XE01(Imatinib, the fifth level).

In general, machine learning algorithms are a learning techniques deal with making intelligent decisions and drawing a conclusion based on the features of input data. Machine learning methods are usually used class labels to represent expression data (Alpaydin, 2010; Nilsson, 1998). The most commonly used machine learning algorithms are Decision Tree, K nearest neighbors (KNN), K-Means, Linear Regression, Logistic Regression, Naïve Bayes, Random Forest (RF), Support Vector Machine (SVM), Dimensionality Reduction Algorithms and Gradient Boost & Adaboost (Shai & Shai, 2014). The Error estimation procedure determines the validity of the resulting classifier model, thus it is essential to classification (Dougherty, Chao, Hua, Hanczar, & Braga-Neto, 2010). Moreover, the distribution of sample size, features and labels, and the classification rule is used to design the classifier have an interactive effect on the performance of an error estimator (Dougherty et al., 2010). The most common performance evaluation used in classification models are confusion matrix and receiver operating curves (ROC) (Oprea, 2014).

The aim of this thesis work is using SVM machine learning method to predict the ATC class of drugs based on the gene expression data to see how well the gene expression patterns correlate after treatment within the therapeutic / pharmacological subgroup. A dimensionality reduction method will use to reduce the dimensions of the dataset that improves the classification performance.

This thesis's chapters are organized as follow: chapter 1 gives an introduction of ATC classification system and gene expression dataset, also simply described the technologies being used in this thesis work. Chapter 2 provides the literature review of several machine learning classification methods. Chapter 3 presents the objective of this thesis research. Chapter 4 presents the material and several machine learning methods that applied to classification using gene expression data. Chapter 5 describes the prediction results of gene expression analysis. Chapter 6 discusses the experimental results based on the gene expression data. Finally, chapter 7 is the conclusion of this thesis research.

Chapter 2. Literature Review

Recently, the financial investments in pharmaceutical research are an immense increase. Drug safety is remarkably important to health and drug development. Finding new uses for approved drug and reducing costs to providing new treatments for unmet medical needs has become important for the pharma industry (Napolitano et al., 2013). Drug classification accuracy helps identify useful information for studying drugs, also helps in accurate examine of drugs.

Recently machine learning approaches contribute to drug classification problem. Mostly, chemical features were used to classify drugs (El-Hachem et al., 2016). For instance, Chen et al. were able to map the chemical substructure of drugs, described by the Anatomical Therapeutic Chemical (ATC) classification system to predict chemical similarities (Chen et al., 2012). The study used different benchmarks to predict level 1 of ATC codes by collecting chemical data for 3,883 drugs with high classification performance (73% accuracy). Napolitano et al. proposed a novel computational method to predicts drug repositioning based on machine learning algorithms to predict the 2nd level (therapeutic class) of FDA-approved compounds with high accuracy levels (78%) (Napolitano et al., 2013).

2.1 ATC-codes

The Anatomical Therapeutic Chemical (ATC) Classification system (http://www.who.int/classifications/atcddd/en/) was first published in 1976. In the ATC classification system, drugs are classified into 5 different levels. The structure of ATC system (https://www.whocc.no/atc/structure_and_principles) is illustrated using the example in Table 1. For example, the classification is organized in five levels: A (Alimentary tract and metabolism, the first level, anatomical main group), A10 (Drugs used in diabetes, the second level, therapeutic subgroup), A10B (Blood glucose lowering drugs, excl. insulins, the third level, pharmacological subgroup), A10BA (Biguanides, the fourth level, chemical subgroup) and A10BA02 (Metformin, the fifth level, chemical substance) (https://www.whocc.no/atc/structure_and_principles).

Table 1. An example of detailed information of metformin at 5 levels in ATC system (https://www.whocc.no/atc/structure_and_principles)

A	Alimentary tract and metabolism (1st level, anatomical main group)
A10	Drugs used in diabetes (2nd level, therapeutic subgroup)
A10B	Blood glucose lowering drugs, excl. insulins (3rd level, pharmacological subgroup)
A10BA	Biguanides (4th level, chemical subgroup)
A10BA02	metformin (5th level, chemical substance)

2.2 Gene expression data

Gene expression profiling has been applied to potentially predict biomedical issues which concerning human health conditions significantly (Khan et al., 2001). It measures the activity of genes and creates a global picture of cellular function. In drug discovery, gene expression profile may be used to diagnose a disease or condition. Liu et al., used the latest LINCS L1000 data for breast cancer (MCF-7) cell lines, presented a "compound signature" based approach to analyzing the pharmacological potential of compounds (C. Liu, Su, Yang, Ma, & Zhou, n.d.). L1000^{TM} expression profiling is a patented whole-genome gene-expression assay. Plentiful information is produced by datasets generated from Library of Integrated Network-based Cellular Signatures (LINCS). The information clarifies the small molecule perturbations and the response of different types of human cells to drug (Genometry, 2017). All signatures are represented by vectors measuring differential expression in the case of L1000 data ("LINCS Data Portal," 2017). It measures 978 validated landmark genes in crude cell lysate in 384 well plate format, using an algorithm trained on tens-of-thousands of historical gene-expression profiles to calculate the levels of the remaining transcripts (Genometry, 2017). Using these gene expression databases, many compounds and their properties can be acquired to enhance the predictive power of the available computational methods.

2.3 Machine learning method

In general, analyze an enormous amount of biological information by providing mathematical frameworks, machine learning as a valuable tool has shown widely used in resolving important issues in science (Genometry, 2017). A certain proportion of the whole data is used to train the predictor (classifier in classification) and then tested and validated to the newly built predictor through the rest part of the data. Support vector machines (SVM), Naïve Bayes, K Nearest Neighbors and Random Forest are popular methods of classification in machine learning (Shai & Shai, 2014).

Dimensionality reduction approach refers to the process of converting a set of data having high dimensionalities into data having smaller dimensionalities, reduce and clean the data for supervised training, and solve machine learning problems to obtain better features for a classification (Ghodsi, 2006; Ray, 2015). The most common methods to perform dimensional reduction are high correlation, backward feature elimination, low variance and principal component analysis (PCA).

2.4 Classification methods

In machine learning, classification is one of the techniques used to predict which categorical class a data sample belongs to. It is a mapping from samples to predicted classes (David & Balakrishnan, 2010). This technique is based on supervised learning, which means using the unique features of the data to construct a classification model to identify the correct class by analyzing one or several sets of labeled training data. Then use new data that's like the testing data in this classification model to predict which class they belong to. Some popular approaches for the classification techniques include Support Vector Machines, Random Forest, K-Nearest Neighbors and Naïve Bayes

2.4.1 Support Vector Machine techniques

Support vector machine (SVM) is a popular type of machine learning. Support vector machine (SVM) algorithm classifies each data based on its value obtained by the classifier function. SVM

builds a classification model that maximizes the margin between the data points that belong to each class. In 1963, Vapnik and Lerner presented the main concept of SVM is to constructs a hyperplane as the separator of the two classes (Cortes & Vapnik, 1995). Cortes & Vapnik (1995) developed SVMs for binary classification. The basic SVM supports only binary classification, when have more than two classes, by running the algorithm several times and testing all the one-against-one class-combinations, then using the support vector machine classifier classifying the data which had the highest performance. With a linearly separable data, a hyperplane can divide a data set into two classes correctly (Karatzoglou, Meyer, & Hornik, 2006). The distance between SVM classifier and data points called margin that illustrates the SVM solution. Support vector machine is defined by kernel function use an implicit mapping Φ of input data into a high-dimensional feature space (Karatzoglou et al., 2006). RBF kernel is the first choice in general, this nonlinear kernel maps samples into a higher dimensional space (Hsu, Chang, & Lin, 2016).

2.4.2 Naïve Bayes

A Naïve Bayes classifier it is a classification technique for binary and multi-class classification problem, with an assumption of independence among predictors based on Bayes' Theorem. The calculation of the probabilities for each hypothesis are simplified to make their calculation tractable (Brownlee, 2016). It is a simple model and easy to builds for very large data sets, with no complicated parameter estimation. It is a family of an algorithm that all Naïve Bayes classifiers assume that the presence of a feature in a class is unrelated to the presence of any other feature (Ray, 2015). This classifier can be trained in a supervised learning setting efficiently and is known to perform highly sophisticated classification methods.

2.4.3 Random Forest technique

Random Forest (RF) is one of the commonly used predictive modeling and machine learning technique. RF can be used for classification and regression applications, such as performance score (Merrett, 2016). It reduces chances of over-fitting and gives higher model performance or accuracy. A random forest can be mathematically written as $\{h\ (x, \Theta_k), k = 1, \ldots\}$ where k means kth tree, Θ_k is a random vector, resulting in a classifier $h\ (x, \Theta_k)$ where x is an input

vector (Breiman, 1999; Merrett, 2016). It mainly performs well on classification model and listed as a top algorithm in Kaggle competitions. The random forest starts with a decision tree. On the top is an input data and it traverses down the tree, so the data gets divided into smaller sets (Benyamin, 2012).

2.4.4 K-Nearest Neighbors

K nearest neighbors (KNN) is a non-parametric algorithm that can be used for classification or regression. KNN is widely used in data science analytics as a supervised machine learning algorithm. KNN algorithm stores all available instances and classifies new cases by majority vote of its k neighbors (Ray, 2015; Vladimirov, 2013). The algorithm finds the k closest observations for each data point, then classifies the data point to the majority. The training dataset is used to classify each member of a target dataset in KNN classification.

2.5 Evaluation methods

2.5.1 K-fold Cross Validation techniques

K-fold Cross Validation is the most commonly used method. Firstly, the initial dataset was partitioned randomly into a number (k) of subsets. Each subset has an approximate number of subsets. Afterward, the remaining subsets are combined to perform the role of the training partition while each subset is used as the test partition (Dernoncourt, 2015; Hastie & Tibshirani, 2009; Jensen, 2016). k models are all trained on a different subset of the initial dataset during the whole process, and each of the subsets has been used as the test partition. The models' accuracy can be computed as the average accuracy across the k model. K models are fit and k validate statistics are obtained and giving the best validation statistics is chosen as the final model.

2.5.2 ROC

The receiver operating characteristic (ROC) curves are widely used in evaluating the discrimination ability of statistical methods for predictive purposes (Hanley & McNeil, 1982). The receiver operating characteristic curve (ROC curve) can be plotting with the fraction of false positives (FP) out of false positives rate (FPR) on the x-axis and the fraction of true positive (TP) out of true positive rate (TPR) on y-axis obtained at different threshold values. The receiver operating characteristic analysis enables us in choosing optimal models or parameters for classification. One point corresponds to an individual result of the classifier in the ROC space. The accuracy is measured by the area under the ROC curve. The resulting curve is called ROC curve, and the metric which needs to consider is the AUC of this curve called AUROC. Receiver-operating characteristic (ROC) plots to capture a unifying or central position in the process of test evaluation (Zweig & Campbell, 1993).

2.6 Dimensionality Reduction method

Dimensionality reduction method refers to the process of converting a set of data having high dimensions into smaller dimensions which simplify, reduce and clean the data for supervised training and solve machine learning problems to obtain better features for a classification (Ghodsi, 2006; Ray, 2015). It is helpful in noise removal, takes care of multicollinearity that improves the performance of models, removes redundant features and reducing data dimensions (Ray, 2015). The most common methods to perform dimension reduction are backward feature elimination, and principal component analysis (PCA).

2.6.1 Principal Component Analysis

Important variables were extracted from a large set of available variables in the data set for principal components analysis. The low dimensional set is extracted from a high dimensional data with the motive of capturing as much information as possible. At the same time, the visualization also becomes much more meaningful (Analytics vidhya content team, 2016).

The main use of PCA is to reduce the size of the feature space while retaining as much of the information as possible by retaining the explained variance ratio of the PCs. If the full variance

of data set defines as $\sigma = \sum_j \lambda_j$ then the explained variance ratio of component j is defined as $r_j = \frac{\lambda_j}{\sigma}$ (Otterbach, 2016).

Chapter 3. Objectives

Expression data have been used to classify biological samples in many novel ways such as by pharmacological mechanism (Gunther, Stone, Gerwien, Bento, & Heyes, 2003). Meanwhile, ATC classification prediction is helpful for new drug development. This thesis evaluates the quality of the prediction of data. Several machine learning methods will use to predict the ATC class of drugs based on the gene expression data to see how well the gene expression patterns correlate after treatment within the 2nd level (2nd level: therapeutic / pharmacological subgroup). A dimensionality reduction method will use to reduces the dimensions of the dataset that improves the classification performance.

The main objectives of thesis research are:

- Use gene expression data predicts the ATC classes of the drugs
- Apply suitable machine learning method to predict ATC classes based on gene expression data from the perturbation effects of the drugs.
- Apply the same machine learning method after dimensionality reduction method to improves the accuracy.

Chapter 4. Materials and Methods

4.1 Drug Target Data

A standard is provided by the Anatomical Therapeutic Chemical (ATC) classification system applied for classifying medical substances for drug utilization research. Moreover, the WHO Collaborating Centre (WHOCC) for drug statistics methodology has maintained the standard since 1976 (https://www.whocc.no/atc_ddd_index). It is a widely-approved classification system for drugs. The system divides drugs into 5 different groups according to their act and pharmacological, therapeutic and chemical properties. Recently, researches are usually focused on the first and second level of ATC groups, indicate the anatomical and the therapeutic main groups. This thesis mainly focuses on the 2nd level of ATC groups which indicate the therapeutic/pharmacological subgroup. The drug target data is obtained from DrugBank database (Wishart et al., 2006). DrugBank is a database containing bioinformatics and cheminformatics information. It combines detailed information about drugs and their targets. Table 2 presents some example target data for selected drugs.

Table 2. Some example target data for selected drugs

drugbank_id	name	type	groups	atc_codes	categories
DB00001	Lepirudin	biotech	approved	B01AE02	Antithrombins|Fibrinolytic Agents
DB00002	Cetuximab	biotech	approved	L01XC06	Antineoplastic Agents
DB00003	Dornase alfa	biotech	approved	R05CB13	Enzymes
DB00004	Denileukin diftitox	biotech	approved|investigational	L01XX29	Antineoplastic Agents
DB00005	Etanercept	biotech	approved|investigational	L04AB01	Immunosuppressive Agents

The “atc_codes” column is considered as a target (label) and it is being used to train the classifier and to evaluate the performance of the tested methods using cross-validation.

4.2 Drug Profile Dataset

978 validated landmark genes were measured by L1000 expression profiling in crude cell lysate in 384-well plate format. In order to enable scaling and normalization, 80 invariant genes are explicitly measured additionally. The crude lysates of human cells in 384-well plates is the input, and the tab-delimited text files of log-transformed expression values for 22,000 genes × 380 samples is the output (Genometry, 2017). The gene expression profiles of LINCS L1000 dataset are available to download via clue.io of the NCBI Gene expression Omnibus (accession number for phase I is GSE92742, accession number for phase II is GSE70138) (http://www.lincscloud.org). Raw fluorescence intensities were converted by the data into differential gene expression signatures processed through a computational system. The data are available at each stage of the preprocessing. Level 1 (LXB) refers to the raw data. For each well of a 384-well plate, one LXB file is generated. Level 2 (GEX) refers to the gene expression values per 1000 genes after the processing of deconvolution. Level 3 (Q2NORM) refers to the gene expression profiles which cover both directly measured landmark transcripts as well as inferred genes. They are normalized by quantile normalization using invariant set scaling. Level 4 (Z-SCORES) refers to the signatures with differentially expressed genes. The signatures are computed by robust z-scores with the purpose to control for each profile relative. Level 5 (SIG) consists of the replicates (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742).

The L1000 profiled gene expression was measured at different time points in cells treated with different dosages (Z. Wang, Clark, & Ma'ayan, 2016). The quantile-normalized gene expression profiles (level 3) were used for this thesis analysis. This thesis uses 2nd level of ATC codes as targets, combined with L1000 gene expression profiles. The gene expression data for the thesis research contains 1209 rows (drug samples) and 978 columns (features). Each gene is considered as one feature as well as one dimension and the ATC class (the 2nd level) is called target from the perspective of statistical learning. Considering a large number of features in the dataset, some of those features are useless information which would give rise to overfitting and consequential poor performance of the classifier. Table 3 presents a part of the dataset containing selected drug samples and genes. Row names are the drug samples, columns names are the features (genes). The last column contains the different classes at 2nd level of ATC codes. For example, X16 referring to alanyl-tRNA synthetase (Homo sapiens (human)); X10058 referring to ATP binding cassette subfamily B member 6 (langereis blood group) (Homo sapiens (human)); X51385 referring to zinc finger protein 589 (Homo sapiens (human)) and X9183

referring to zw10 kinetochore protein (Homo sapiens (human)). DB00014 is an accession number of Goserelin. Goserelin is a synthetic hormone. This drug may stimulate the growth of cancer cells. When stop using this medication, hormone levels return to normal level. DB00091 is an accession number of Cyclosporine. It used for the treatment of transplant rejection, rheumatoid arthritis, severe psoriasis, such as kidney, liver, and heart. DB00121 is an accession number of Biotin. Biotin is using for treating dietary shortage or imbalance, also for nutritional supplementation. DB00130 is an accession number of L-Glutamine. It is the principal carrier of nitrogen in the body and an important energy source for many cells. It used for reducing the acute complications of sickle cell disease in adult and the age of 5 and older pediatric patients. DB00131 is an accession number of Adenosine monophosphate. The adenine nucleotide that contains on phosphate group esterified to the sugar moiety locating in the 2,3,5-position. (https://www.drugbank.ca/drugs)

Table 3. Part of the dataset containing selected drug samples and genes

	X16	**X10058**	**X51385**	**X9183**	…	**Level_2**
DB00014	-1.1242093	0.5199319	-1.3167174	0.7545028	…	L02
DB00091	169.6886998	3.6738522	11.8406073	2.7333477	…	C10
DB00121	0.2777322	-2.8542680	0.7421095	-1.5298809	…	C10
DB00130	1.4300881	-1.1594369	1.5580980	-2.8821927	…	N06
DB00131	2.5383233	-1.1943148	1.6071767	1.6233860	…	C09

Each gene is considered as one feature as well as one dimension, and the ATC class (the 2nd level) is called target (Figure 8). Regarding the big number of a feature in the dataset, some of those features introduce useless information which would give rise to overfitting and poor performance of the classifier.

4.3 Classification methods

In machine learning, classification is a technique used to predict which categorical class of a data example belongs to. It is a mapping from samples to predicted classes (David & Balakrishnan, 2010). This means by analyzing one or several labeled training dataset and using the unique features of the data to construct a classification model to identify the correct class.

Then using testing data identical to the training data in this classification model to predict which class they belong to. Some popular methods for the classification techniques include Random Forest, K-Nearest Neighbors, Naïve Bayes and Support Vector Machines.

4.3.1 Support Vector Machine(s) (SVMs)

In 1990s, the computer science community has developed the support vector machine(s) (SVMs) approach for classification (Witten et al., 2013). In an n-dimensional space, a hyperplane is a flat affine subspace of dimension n-1. In n > 3 dimensions, the notion of (n -1) dimensional flat subspace applies, but it can be hard to visualize a hyperplane (Witten et al., 2013). It is particularly applicable for large dimensionalities ($\geq 10^6$) of document classification and sentiment analysis (Halls-Moore, 2014).

libSVM

Chang & Lin developed the LIBSVM package as a library for the support vector machines approach. LIBSVM can prohibitive for datasets with hundreds of thousands of examples and achieve faster training times (Ben-Hur & Weston, 2010). LIBSVM involves two steps: training a data to obtain a model in the first, then predicts information of testing data with the trained model. LIBSVM solving SVM multi-class classification, parameter selection, theoretical convergence, probability estimates and optimization problems (Chang & Lin, 2011). LIBSVM supports C-support vector classification (C-SVC), distribution estimation (one-class SVM), ε-support vector regression (ε-SVR), ν-support vector classification (ν-SVC) and ν-support vector regression (ν-SVR) (Chang & Lin, 2011).

The basic ideas to utilize LIBSVM are: randomly split dataset into training data and testing data format to be classified according to the LIBSVM requirement; choose the appropriate kernel function k(x_i, x_j) and determine the optimal parameter C and gamma through cross-validation; train the training data set using the computed C and gamma to get the SVM model; perform future predictions by utilize the computed model (Song, 2013). Using linear kernel (for the most common classes) and radial kernel (for 32 classes), parameter C and gamma to get SVM model.

The label is the target classes which corresponds to the 2nd level of ATC class (the 2nd level: therapeutic / pharmacological subgroup).

Multiclass SVM

One major method for the multi-class problem is the one-against-all. This approach is an alternative procedure for applying SVMs in the case of k > 2 classes (Witten et al., 2013). The idea for multi-class problems is building k two-class rules where the mth function $w_m^T \emptyset\ (x) + b$ separates training vectors of the class m from the other vectors (Bottou et al., 1994; Hsu & Lin, 2015). The mth SVM is trained with all the instances in the mth class which have positive labels while all other with negative labels (Fazli, Afrouzian, & Seyedarabi, 2009).

The one-against-one is another major method for the multi-class problem. The idea of this method is to builds k (k – 1) / 2 classifiers, so each one is trained on data from two classes (Hsu & Lin, 2015). For instance, one such SVM might compare the k'th class, codes as +1, to the k'th class, codes as -1. Assigning the test observation to the class performed the final classification. It is most frequent to assign in these pairwise classifications (Witten et al., 2013).

LIBSVM implements the "one-against-one" method to gain the accuracy of cross-validation (CV) for multi-class classification. The parameter selection tool proposes the same (C, γ) for all k (k-1) / 2 decision functions (Chang & Lin, 2011; Knerr, Personnaz, & Dreyfus, 1990). Before minimizing the negative log likelihood, they conduct 5-fold cross-validation to obtain decision values. It first estimates pairwise class probabilities

$$r_{ij} \approx P(y = i \mid y = i \text{ or } j, \boldsymbol{x}) \quad (1)$$

using an improved implementation of Platt (2000) (Lin et al., 2007).

4.4 Evaluation methods

4.4.1 K-fold Cross-Validation

The data is randomly split into k segments of equal size, one is used as test data for validating the model and remaining k-1 segments are used for predicting training model. Each of the k subsamples is being used as test data and the rest used as training dataset by repeated k times. Thus, it takes the average of k results from each fold of k-fold cross-validation and these k results to obtain a single measure of the performance of the trained model for the given data set. All the samples are used for testing (validation) and training where each observation is used during the whole process.

In this thesis, 5-fold cross-validation is used for predicting. For each k=1,2,3…K, fit the model with parameter λ to the other K-1 parts, giving $\hat{\beta}^{-k}(\lambda)$ and compute its error in predicting the K part (Hastie & Tibshirani, 2009):

$$E_k(\lambda) = \sum_{i \in kth\ part}(y_i - X_i \hat{\beta}^{-k}(\lambda))^2 \quad (2)$$

The cross-validation error is given as (Hastie & Tibshirani, 2009)

$$CV\ (\lambda) = \frac{1}{K}\sum_{k=1}^{K} E_k\ (\lambda) \quad (3)$$

After obtaining the new interaction, calculate the performance of the trained model.

4.4.2 Classification based on the machine learning

In the data, each column represents the instances in a predicted class, and each row represents the instances in an actual class. The performance of a classification model on test dataset is often used confusion matrix to describe (Markham, 2014). The confusion matrix shows for each pair of classes. Table 4 shows an example of a confusion matrix for a two-class classifier.

Table 4. An example of Confusion Matrix for two-class classifier

	Predicted Condition Positive	Predicted Condition Negative
Prediction Positive	True Positive (TP)*	False Negative (FN)*
Prediction Negative	False Positive (FP)*	True Negative (TN)*

*True Positive (TP): the number of positive samples which correctly predicted
*False Negative (FN): the number of practical positive samples marked as negative

*False Positive (FP): the number of practical negative samples marked as positive
*True Negative (TN): the number of negative samples which correctly predicted

Accuracy is calculated as the whole true predictions divided by the overall number of the dataset (Saito & Rehmsmeier, 2015). Accuracy (ACC):

$$\text{ACC} = (TP + TN)/(P + N) \quad (4)$$

The proportion of instances classified positive in relation to all instances tested positive is the calculation of true positive rate (Oprea, 2014; Saito & Rehmsmeier, 2015). Sensitivity (or True Positive Rate (TPR)):

$$\text{TPR} = TP/P = TP/(TP + FN) \quad (5)$$

The number of true negative predictions divided by the overall number of negatives is the calculation of the true negative rate (Saito & Rehmsmeier, 2015). Specificity (or True Negative Rate (TNR)):

$$\text{TNR} = TN/N = TN/(FP + TN) \quad (6)$$

The number of false negative predictions divided by the overall number of positives is the calculation of the false negative rate. Miss Rate (or False Negative Rate (FNR)):

$$\text{FNR} = FN/P = FN/(FN + TP) = 1 - \text{TPR} \quad (7)$$

The number of false negative predictions divided by the overall number of negatives is the calculation of false positive rate (Saito & Rehmsmeier, 2015). Fall-out (or False Positive Rate (FPR)):

$$\text{FPR} = FP/N = FP/(FP + TN) = 1 - \text{TNR} \quad (8)$$

The number of correctly predicted positive observation divided by the overall number of positive predictions is the calculation of positive predictive value (Saito & Rehmsmeier, 2015). Precision (or Positive Predictive value (PPV)):

PPV = $TP/(TP + FP)$ (9)

The number of correctly predicted negative observation dived by the number of negative predictions is the calculation of negative predictive value. (Myatt & Johnson, 2009). Negative Predictive value (NPV):

NPV = $TN/(TN + FN)$ (10)

F-measure is the harmonic mean of the two parameters which combines precision and sensitivity (Oprea, 2014). F-measure can be calculated using the formula:

F-Measure = $(2 \times TPR \times Precision) / (TPR \times Precision)$ (11)

The error rate is the number of every incorrect prediction divided by the overall number of the dataset (Saito & Rehmsmeier, 2015). Error rate (ERR):

ERR = $(FP + FN)/(P + N)$ (12)

When the dataset is unbalanced it tends to be the misleading result, thus, the accuracy is not an affirmatively reliable metric for the actual performance of the predictor. A biased predictor would put all samples into the major class if one class size is significantly larger amount than the others, so the minor class will have a performance of 0%.

4.4.3 ROC Analysis

The binary classifier performance can be measured by using Receiver Operating Characteristic (ROC). The ROC curve is plotted with the fraction of positive samples correctly classified (TPR / Sensitivity) on the x-axis and fraction of negative samples incorrectly classified (FPR / Specificity) on the y-axis (Grigorev, 2015). *Figure 1* is an example of the ROC curve. The dotted line indicates the average performance of samples and the curve closer to left top corner indicates a good performance of the given classifier (Greg, 2015).

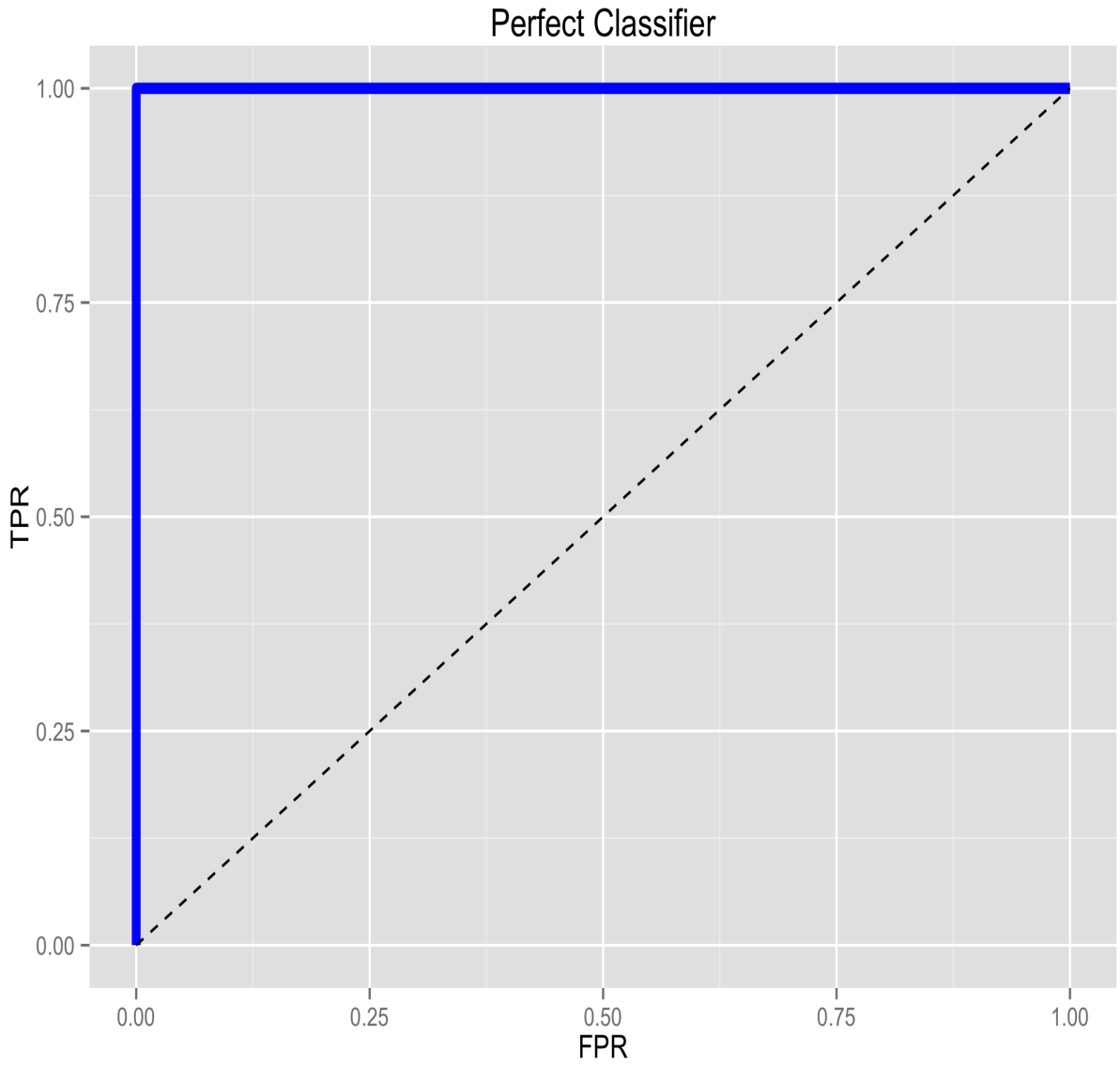

Figure 1. An example of the perfect classifier of Receiver Operating Characteristic Curve (ROC) (Greg, 2015).

For two classes AUC is computed with ROCR package (Sing, Sander, Beerenwinkel, & Lengauer, 2015). Use prediction function to transform the predicted input data into a standardized format; then use performance function to get tpr, fpr and auc. Finally, use plot function to get ROC plot of two classes.

Use SVM classification method to predict data, then attribute probabilities from svm prediction result. In the prediction function, predictions and labels need to be the same length. To get the ROC plot, the computed prediction is then put into performance function with tpr and fpr measures. Finally, plot method to plot the objects of class performance (Sing et al., 2015). The

AUROC measurement equals the value of the Wilcoxon-Mann-Whitney test statistic. This measure cannot be combined with other measures into a parametric curve since the output of auc is cutoff-independent (Sing et al., 2015).

For multi-class classifier, assume that the multiple classes were labeled as 0,1, 2, 3, …, c -1 (c > 2), estimates the probability of every test point belongs to each class $\hat{p}\ (i|x)$ for i = 0, 1, 2, 3, …, c – 1. Using either $\hat{p}\ (i|x)$ or $\hat{p}\ (j|x)$ to compute the measure $\hat{A}$ for any pair of classes i and j, so that a randomly selected member of class j will have a higher probability of belonging to class i than a randomly selected member of class i is the probability of $\hat{A}\ (j|i)$ (Y. H. Wang & Cheng, 2010). For more than two classes $\hat{A}\ (i|j)\ \neq\ \hat{A}\ (j|i)$ (Hand & Till, 2001; Y. H. Wang & Cheng, 2010).

The average of all pairs (M) is calculated by the whole performance of the classification rule in separating the c classes (Hand & Till, 2001; Y. H. Wang & Cheng, 2010):

$$M = \frac{2}{c\ (c-1)} \sum_{i<j} \hat{A}\ (i,j) \tag{13}$$

In the Mann-Whitney-Wilcoxon two-sample test, $\hat{A}$ is an equivalent used to the test statistic. The measure $\hat{A}$ is an overall measure of how well separated are the estimated distributions of $\hat{p}\ (x)$ for different classes (Hand & Till, 2001).

An average AUC is computed for multiclass AUC as defined (equation 14) (Hand & Till, 2001; Robin et al., 2011):

$$\text{auc} = \frac{2}{\text{c}\ (\text{c}-1)} \sum \text{aucs} \tag{14}$$

with aucs all the pairwise roc curves.

The Area Under the Receiver Operator Curve (AUCROC) is a metric of model performance commonly used in machine learning and binary classification or prediction problems. It generates a threshold independent measure of the model able to differentiate between two outcomes. This means that for any model which makes continuous predictions of binary

outcomes, an arbitrary threshold above will be a prediction of 1, below will be 0. By integrating cross all possible thresholds, AUC gets around of this threshold. It is calculated by plotting the ROC curve and calculating the area under the curve.

4.5 Dimensionality reduction methods

Dimensionality reduction method is the process of converting a dataset having high dimensions into lesser dimensions which simplify, reduce and clean the data for supervised training (Ghodsi, 2006; Ray, 2015). It is used to obtain better features for classification by solving machine learning problems. It is helpful in noise removal, takes care of multicollinearity that improves the performance of models, removes redundant features and reduces data dimensions (Ray, 2015). The most common methods to perform dimension reduction are backward feature elimination and principal component analysis (PCA).

4.5.1 Principal Components Analysis (PCA)

Principal components analysis (PCA) method is one of the most popular techniques for dimensionality reduction. Principal components analysis extracts important variables from a large set of available variables in a dataset. These new variables are known as PCs. To catch as much information as possible is the motive of extracting low dimensional dataset from a high dimensional dataset (Analytics vidhya content team, 2016). For example, the attribute space is reduced with losing only the smallest possible amount of information in the original data (Janecek & Gansterer, 2008). PCA is performed on a symmetric correlation or covariance matrix and usually deals with 3 or higher dimensional data (Analytics vidhya content team, 2016). The main use of PCA is to reduce the size of the feature space while retaining as much of the information as possible by retaining the explained variance ratio of the PCs. If the full variance of data set defines as $\sigma = \sum_j \lambda_j$ then the explained variance ratio of component j is defined as $r_j = \frac{\lambda_j}{\sigma}$ (Otterbach, 2016).

For example, we have a set of predictors as $X_1, X_2 \ldots, X_p$. The data is first centered on the means of each variable. The first principal component can be written as:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \quad (15)$$

The accounts for the greatest possible variance in the dataset is the first principal component. Used constraint that their sum of squares is 1 to calculated weights (Holland, 2017).

Uncorrelated with the first principal component which calculated with the conditions is the second principal component. It accounts for the second highest variance, it can be written as:

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \quad (16)$$

Alternatively, the total variation on all the principal components will be equal to the total variation among all the variables. The correlations of the original variables with the principal components are often useful. The correlation of variable X_i and principal component Y_i is

$$r_{ij} = \sqrt{a_{ij}^2 \, Var(Y_j) / \, s_{ii}} \quad (17)$$

When the principal components have been calculated, use the scree plot of the eigenvalues to decide which principal components should be retained. The eigenvalues' scree plot indicates whether there is an obvious cut-off between large or small eigenvalues (Cornish, 2007).

Chapter 5. Results

The whole data set is a 1137 × 979-dimensional matrix, used Five-fold cross-validation to construct a multi-label classifier. The classification is organized in five levels. Figure 8 shows how those levels are organized. The first level describes an anatomical main group; the second level refers to the therapeutic main group; the third and fourth levels refer to therapeutic/chemical/pharmacological subgroups. The individual chemical substances are identified in the fifth level. This thesis uses the 2nd level of ATC codes (at 2nd level: therapeutic / pharmacological subgroup) as targets for the classifier. I removed all the empty classes obtaining a 779 × 979-dimensional matrix representing 73 levels, because of several classes appeared highly under-represented or empty at that level in our dataset. *Figure 9* shows the distribution of the numbers of ATC-classes and the drug samples may belong to.

5.1 Two high-iterated classes classification

In the very beginning, I used the most common two classes L01 (antineoplastic agents) and J01 (Antibacterials for systemic use) with 106 samples for the first prediction to see how well those two classes classified based on the gene profile data. There are 57 samples in L01 class and 49 samples in J01 class. The obtained dataset for these two classes classification is 106 × 979-dimensional matrix. The data split into 70% training data and 30% testing data. First use 70% training data and svm model (without bootstrap or feature selection), with parameters 5 cross validation, C support vector, 0.5 cost and linear kernel to train data. Then use the result of svm model and 30% testing data for prediction. Additionally, perform dimensionality reduction method to improves the performance.

The accuracy calculated obtained by using the SVM method is 87.1%. Figure 2 shows ROC curve by use SVM method, the area under the curve is 0.90.

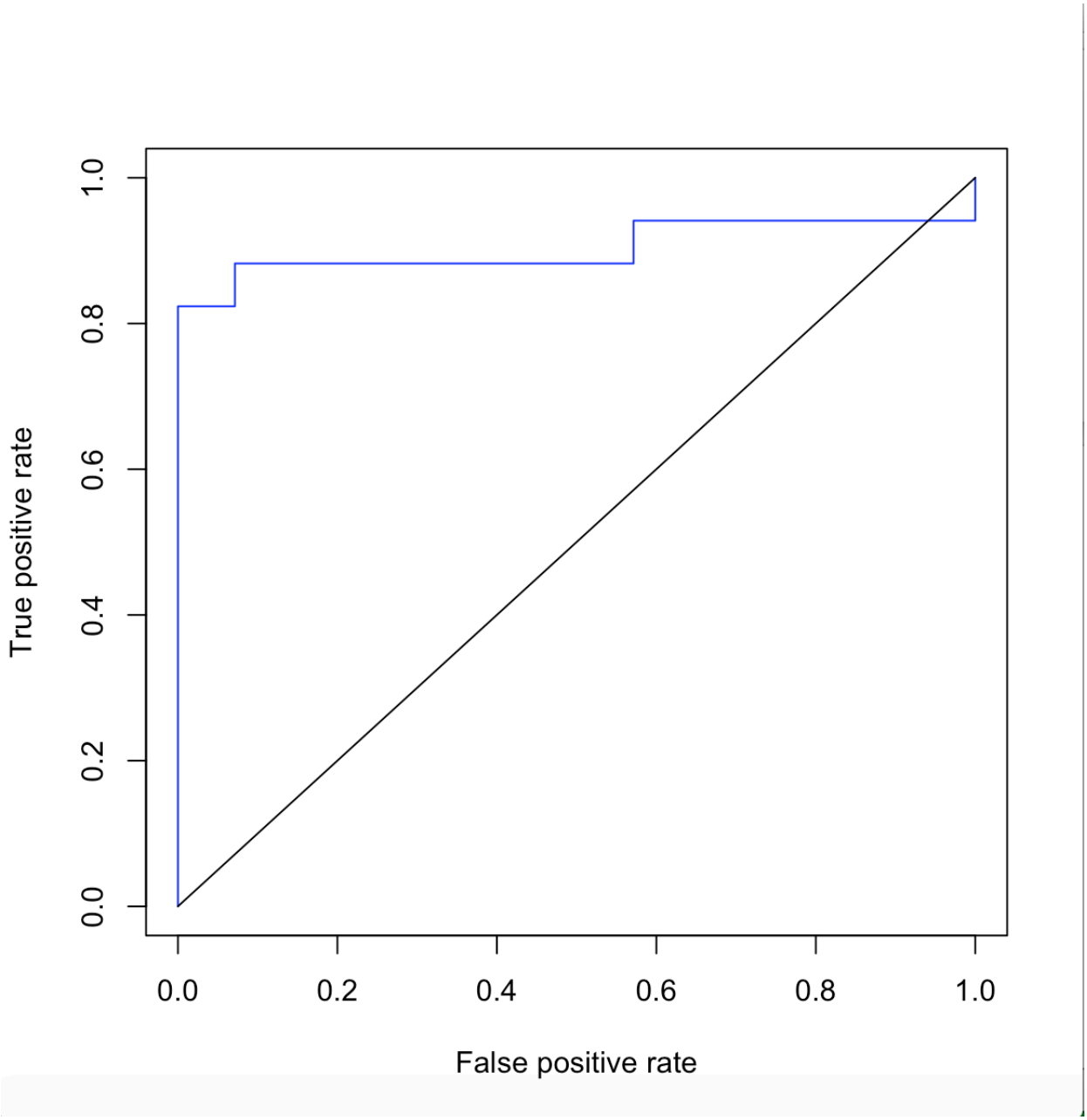

Figure 2. ROC curve by SVM method. The graph indicates ROC curve by SVM method, the area under the curve is 0.90.

The k-fold cross validation calculated the mean of 5-fold accuracies is 76%. Next, I used dimensional reduction method (PCA) to reduces the data dimensions and to improves the classification performance. Table 5 shows the proportion of variances explained by each component. For example, the first principal component explains 73.17% variance, the second principal component explains 18.90% variance, the third principal component explains 4.00% variance and so on.

Table 5. The proportion of variances result

0.7316880550	0.1889474112	0.0399863290	0.0104919835
0.0072207719	0.0044320653	0.0033602759	0.0028150862
0.0025134913	0.0014972181	0.0011944287	0.0007778557
0.0005495822	0.0004687764	0.0003834742	0.0003480590
0.0003029497	0.0002953833	0.0002721552	0.0002556968

Figure 3 shows the first component has the highest variance, something value around 70% while the 3rd component is around 0%. So, it indicates that we should pick up the first three components.

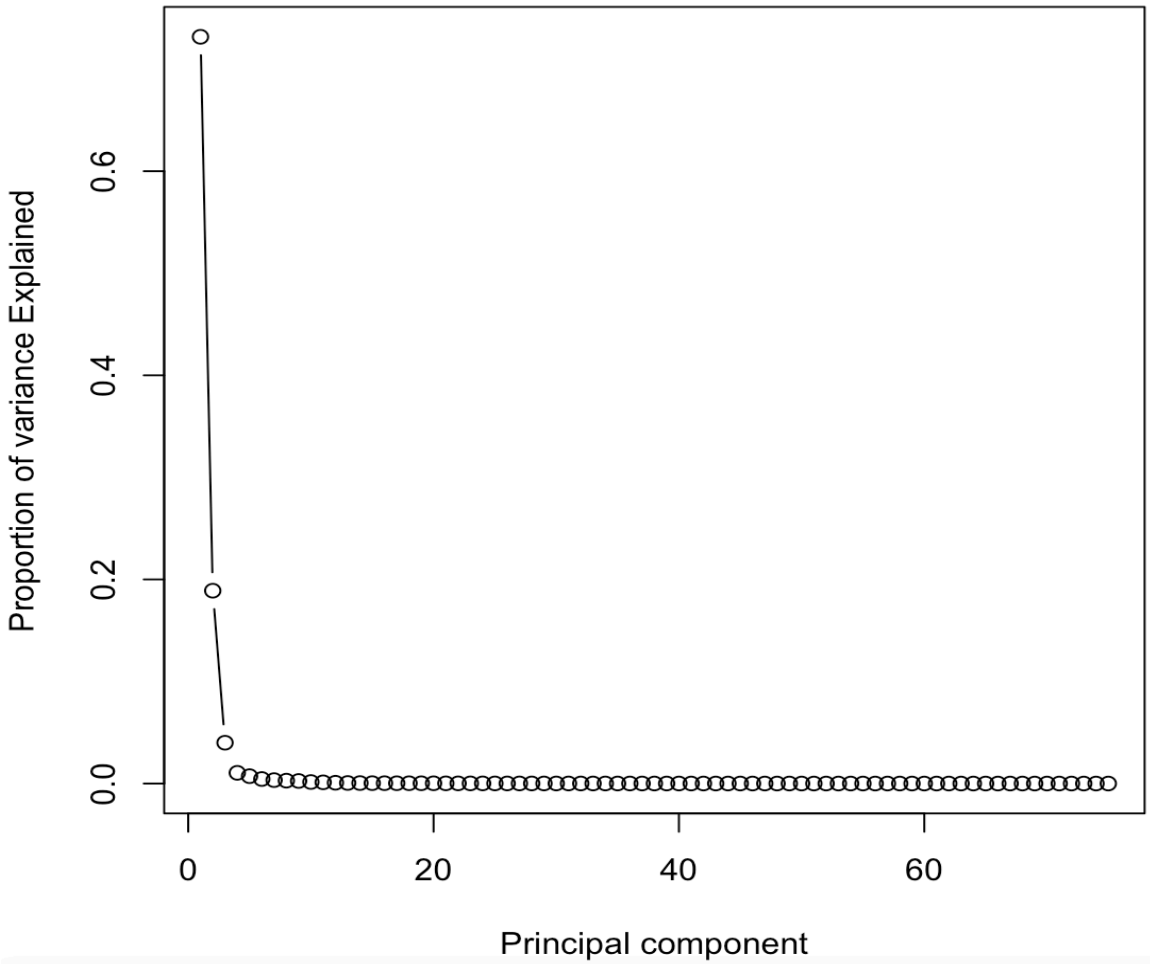

Figure 3. The graph shows the percentage of variance of each feature. the first component has the highest variance, something value around 70% while the 3rd component is around 0%.

Use the eigenvalues to construct a useful graph to decide which principal components need to choose. Figure 4 shows how many principal components should choose.

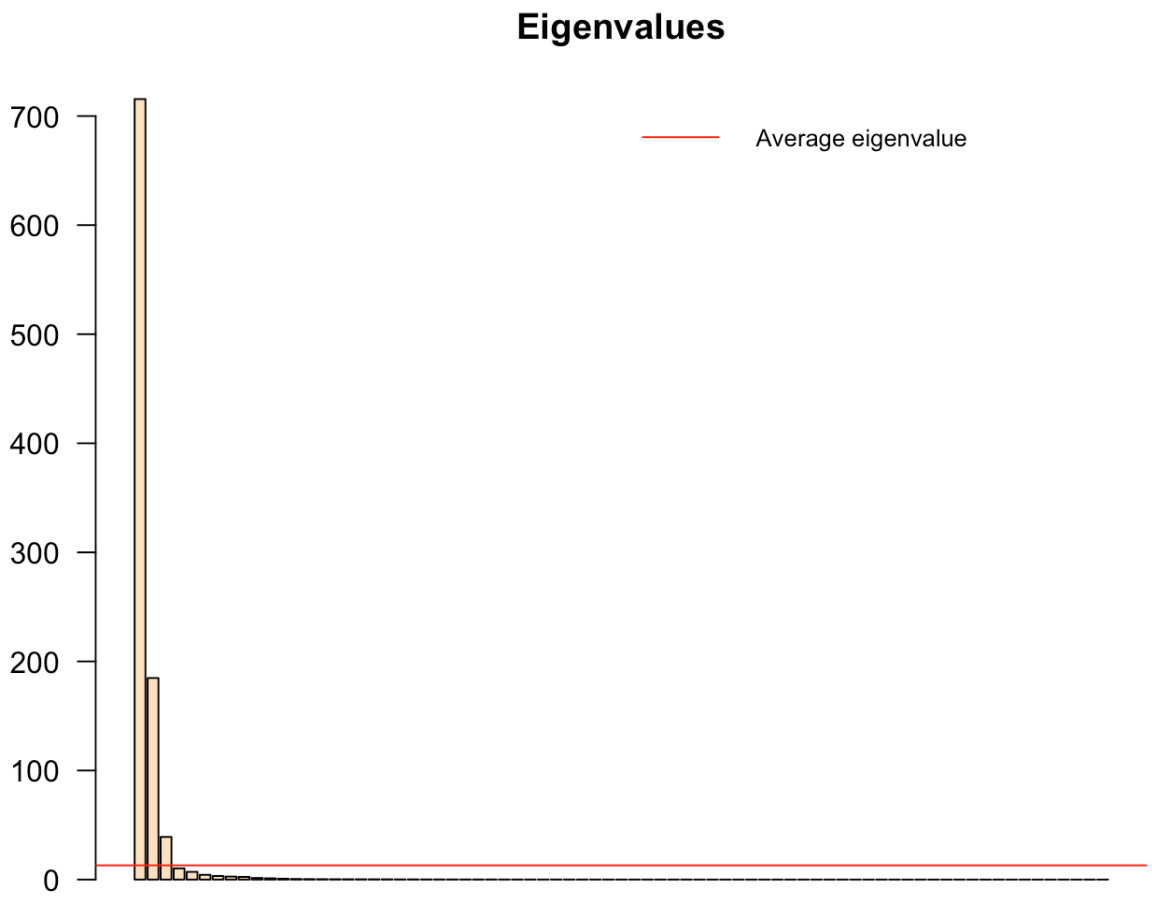

Figure 4. Barplot of eigenvalues and percent of the proportion of variations of two classes. From this plot, there are only 3 components above the average line (red line).

Figure 4 shows that only three principal components should be chosen, because their bars are above the average eigenvalue line (red line). Figure 5 presents the ROC curve of SVM method after PCA. The accuracy calculated by SVM method is 64.52%, the area under the curve is 0.77.

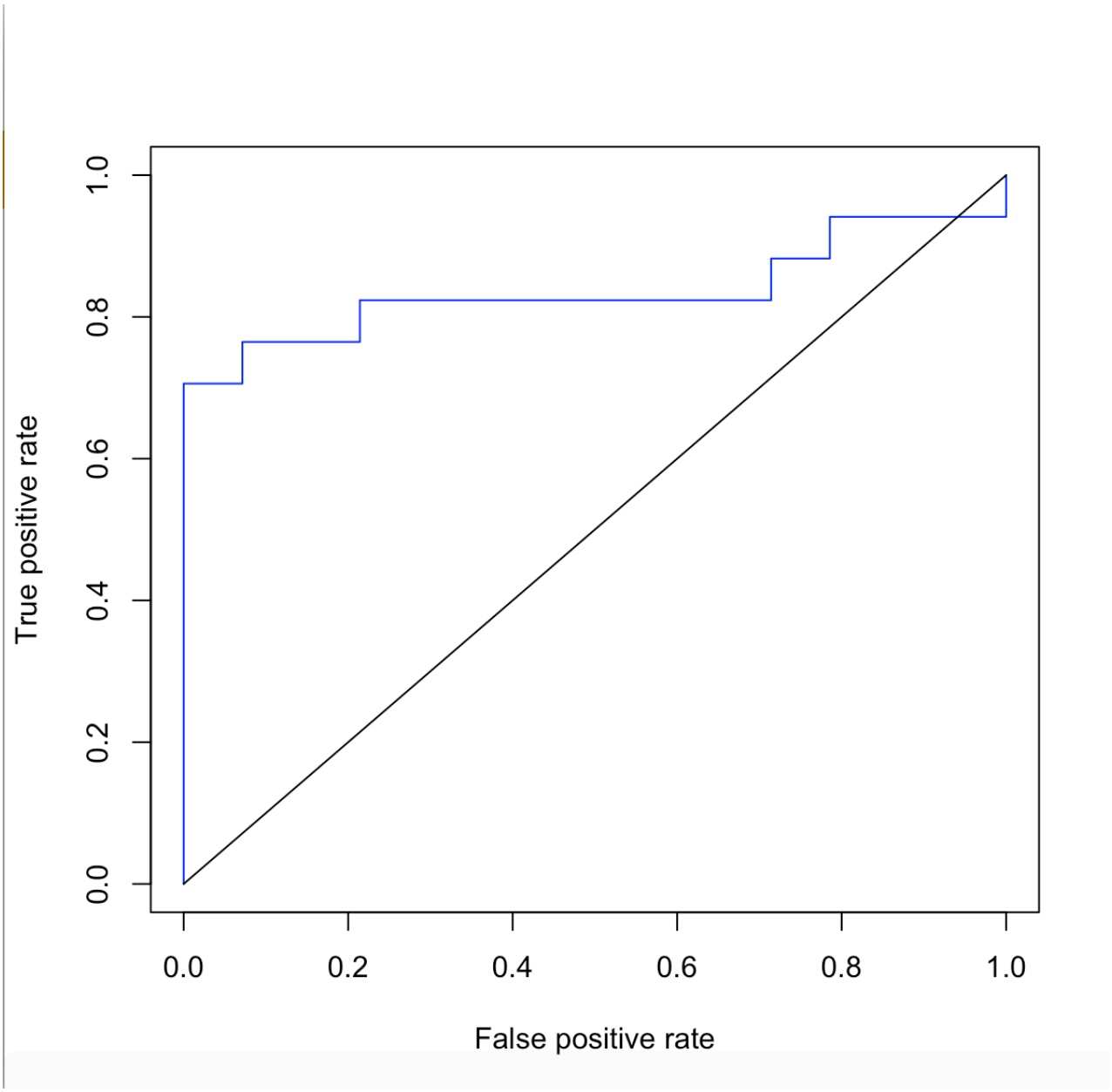

Figure 5. ROC curve of SVM method after PCA. The x-axis represents FPR (specificity) and y-axis represents TPR (sensitivity) from the prediction object. The graph presents the area under the curve of SVM method is 0.77.

5.2 32 classes classification

In this dataset, a number of classes appeared highly under-represented at this level. For this reason, I removed all others ATC classes with less than 10 samples per class, obtaining a data of 619 × 979 dimensions with 32 classes. *Figure 9* shows sample number per each class and detail information of ATC-codes.

Table 6. Sample numbers per each class and detail information

Samples	**ATC-codes**	**Detail information**
57	L01	Antineoplastic agents
49	J01	Antibacterials for systemic use
43	N05	Psycholeptics
42	N06	Psychoanaleptics
30	C01	Cardiac therapy
22	M01	Anti-inflammatory and antirheumatic products
22	R06	Antihistamines for systemic use
20	J05	Antivirals for systemic use
18	G03	Sex hormones and modulators of the genital system
17	A03	Drugs for functional gastrointestinal disorders
17	A10	Drugs used in diabetes
17	C03	Diuretics
17	C07	Beta blocking agents
16	C10	Lipid modifying agents
16	N03	Antiepileptics
15	N07	Other nervous system drugs
14	C09	Agents acting on the renin-angiotensin system
13	A07	Antidiarrheals, intestinal antiinflammatory / antiinfective agents
13	C08	Calcium channel blockers
13	D01	Antifungals for dermatological use
13	N04	Anti-parkinson drugs
13	P01	Antiprotozoals
13	R03	Drugs for obstructive airway diseases
12	A02	Drugs for acid related disorders
11	C02	Antihypertensives
11	G04	Urologicals
11	L02	Endocrine therapy
11	N02	Analgesics
11	R01	Nasal preparations
10	B01	Antithrombotic agents
10	M03	Muscle relaxants

The data then split into 70% training set and 30% testing set. The same method as the previous section is then used for prediction.

The accuracy calculated by SVM method with the radial kernel is 16.37%. The area under the curve of SVM method is 0.66. The k-fold cross validation calculated the mean of 5-fold accuracies is 46.48%.

Table 7 shows some information of proportion of variances. For example, the first principal component explains 66.01% variance, the second principal component explains 22.04% variance, the third principal component explains 3.93% variance and so on.

Table 7. Some information of Proportion of variances results

0.6601848424	0.2204012434	0.0393501227	0.0139816991
0.0094117530	0.0084948216	0.0066827641	0.0043391264
0.0040842748	0.0027268552	0.0021862980	0.0020204416
0.0017985616	0.0016041122	0.0015450434	0.0012603670
0.0010444086	0.0009174959	0.0008801428	0.0007918687

Figure 6 shows the first component has the highest variance, something value around 66% while the 3rd component is around 0%. So, it indicates that we should pick up the first three components.

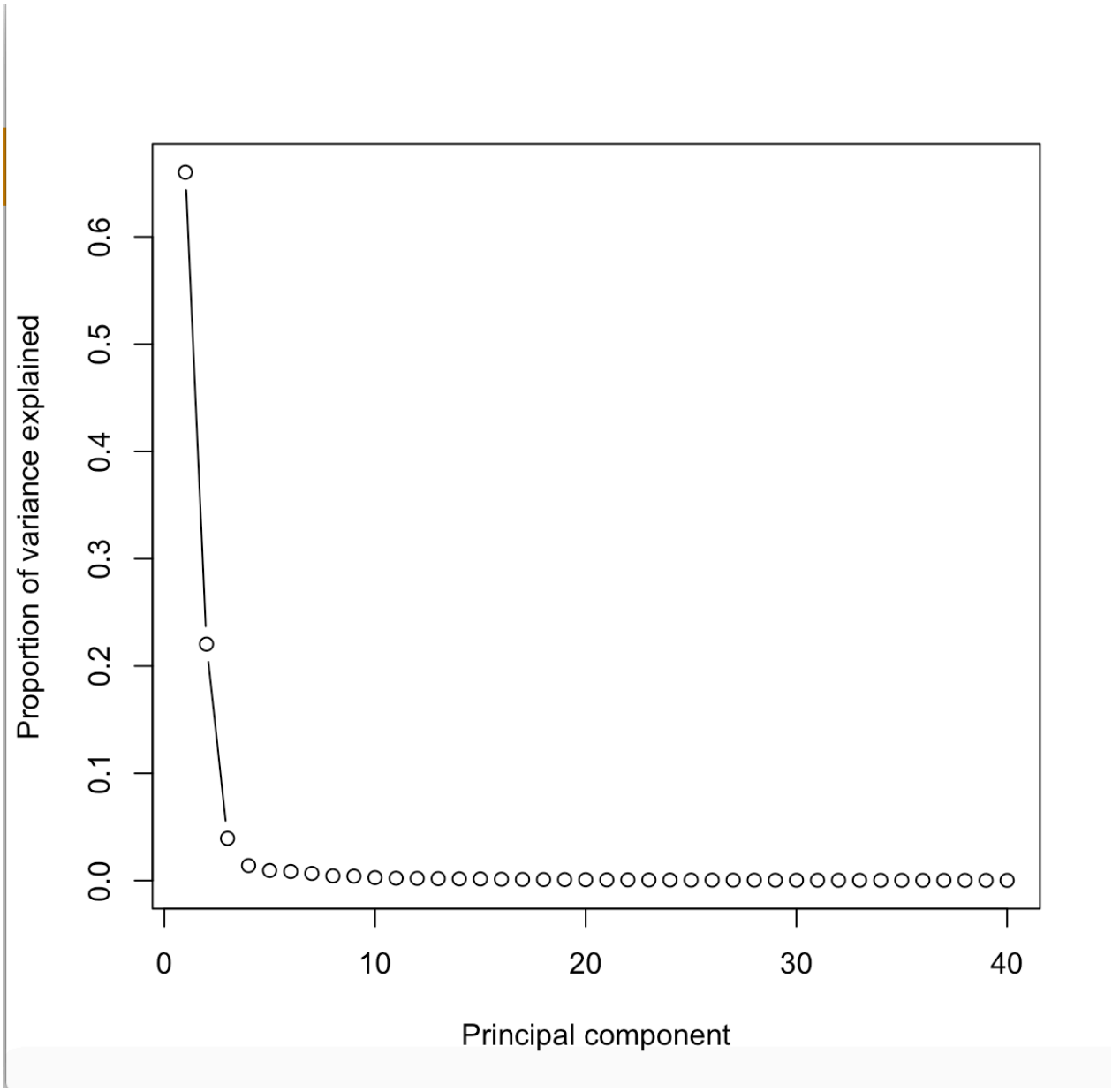

Figure 6. The graph shows the percentage of variance of each feature. the first component has the highest variance, something value around 66% while the 3rd component is around 0%.

Figure 7 shows that only the first three principal components should be chosen, because those bars are above the average eigenvalue line.

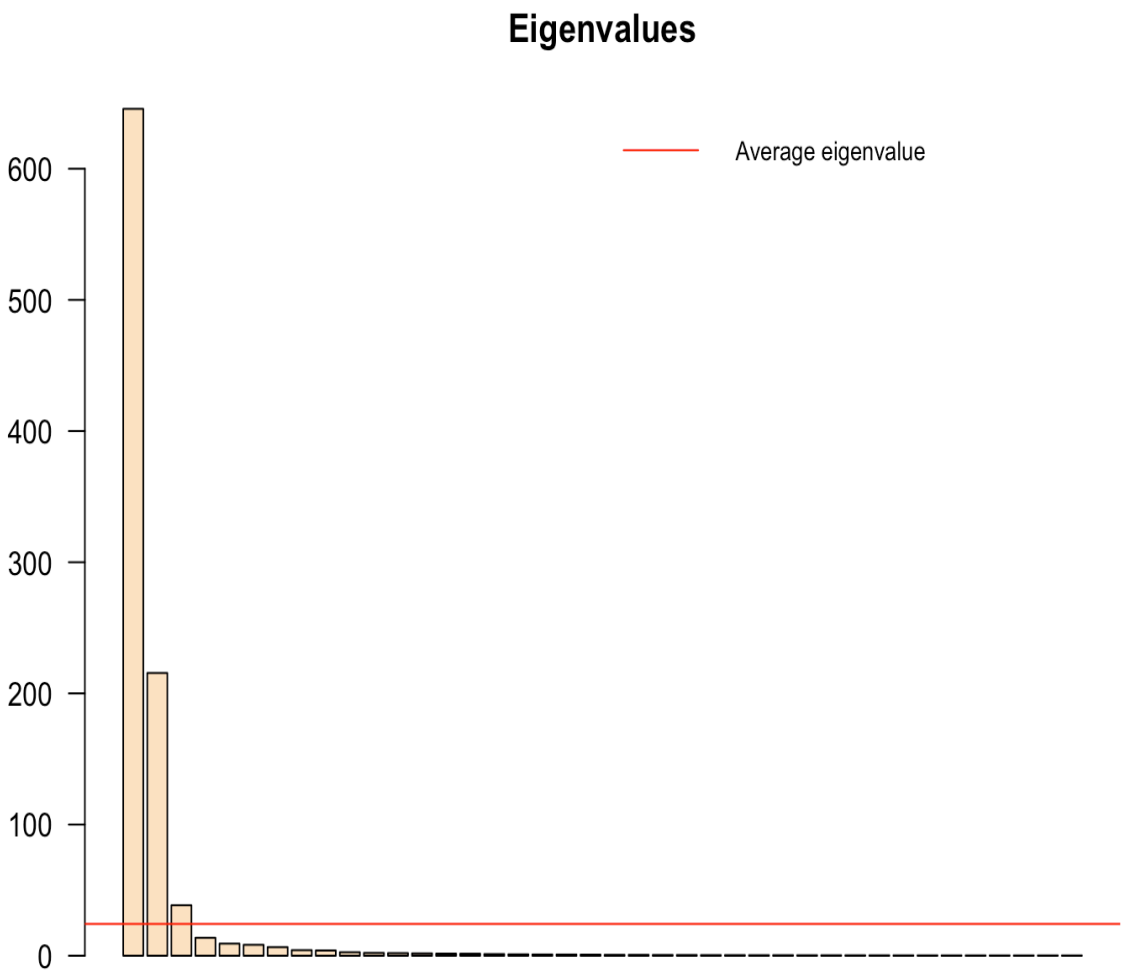

Figure 7. Barplot of eigenvalues and percent of proportion of variation for the 32 classes. The barplot shows that only the first three principal components should be chosen, because they are above the average line (red line).

The accuracy calculated after dimensional reduction by SVM method is 16.37. The area under the curve calculated by SVM method is 0.65.

Chapter 6. Discussion

Drug targets can be restricting the efficiency of clinical treatment and tolerance development (Iskar et al., 2010). Drug classification accuracy helps identify useful information for studying drugs, also helps in accurate diagnosis of drugs. Extracted the therapeutic class of each drug from the ATC classification (level 2: pharmacological/therapeutic subgroups) (https://www.whocc.no/atc/structure_and_principles/). This thesis research presented computational approaches that used newly LINCS L1000 dataset and ATC second level. The data obtaining an expression matrix of dimension 1137 × 979. I removed all the drugs falling into ATC classes with less than 10 drug samples per subgroup as our final prediction, because of The number of different ATC codes (Table 1) at level 2 is large compared with the number of drugs and highly under-represented classes or empty classes appeared at this level. By using the ATC 2nd level (pharmacological/therapeutic subgroups) and the classes with at least 10 drugs, with SVM classification method, I have applied random re-sampling and re-training iterations both by using dimensional reduction method to improves classification performance and to reduce over-fitting. Table 8 presents the general interpretation of AUC (Morgan, 2016). For example, when AUC is ≤ 0.5, the result is not considerable; when AUC result is between 0.6 to 0.7, the accuracy is poor; when AUC is between 0.7 to 0.8, the accuracy is fair; when AUC is between 0.8 to 0.9, the prediction result is good; when AUC is greater than 0.9 to 1.0, then the accuracy result is perfect.

Table 8. General interpretation of AUC

AUC	Interpretation
0.9 to 1.0	Excellent accuracy
0.8 to 0.9	Good accuracy
0.7 to 0.8	Fair accuracy
0.6 to 0.7	Poor accuracy
0.5	No discriminating ability

This thesis performed filtering and normalization steps. A total of 979 features were left for further processing. After pre-processing, using the average of all experiments in the corresponding calculated mean centered of each probe in the gene expression profiles (Iskar et al., 2010). The area under the Receiver operator characteristic (ROC) curve (AUC) for drugs (average over all), and the 2nd level of the available ATC code is low (Iskar et al., 2010).

The first prediction used the two most high-represent classes with 106 samples for testing how well those classifiers classified. The ROC plot (Figure 2) presented by SVM method before use dimensional reduction. The area under the curve by SVM method is 0.77. The 5-fold cross validation result shows the mean of accuracy is 0.76. According to the table of general interpretation of AUC (Table 8), the results are fair.

Next, used the classes with at least 10 drug samples class obtaining data matrix of dimensions 619×980, using the same computational approach (SVM method) as the final prediction. The result demonstrates a poor test. The area under the curve by SVM method is 0.65.

The proper projections of the data by Principal component analysis (PCA) method is used to improves the performance of the machine learning modeling. There are two options to choose important and representative PCA axes, which are Kaiser-Guttman criterion and Broken stick model (Borcard, Gillet, & Legendre, 2011; MacArthur, 1960). They are both based on the eigenvalues which tell how much a principal component is able to explain from initial dataset.

Figure 5 demonstrates a theoretically fair test for two classes. The area under the curve by SVM method is 0.77. For 32 classes, the area under the curve by used SVM method is 0.65. The PCA method has not improved the performance of the machine learning modeling, it may lose spatial information. So, the classification accuracy decreases.

Chapter 7. Conclusion

This thesis presented a SVM approach to predict ATC classes of drugs based on their characteristic gene expression signature, alternatively evaluate the performance of machine learning result to see how much the ATC similarity of drugs is related to the gene expression patterns. Subtract different classifiers with different matrix dimensions, then applied SVM and Naïve Bayes methods, used multiclass ROC curve and k-fold cross validation to evaluated the classification accuracies. In addition, the proper projections of the data by Principal component analysis (PCA) method is used to improves the performance of the machine learning modeling.

To improve the prediction performance, PCA method has been used to removes multi-collinearity and improve the performance of the machine learning modeling. Kaiser-Guttman criterion and Broken stick model (Borcard et al., 2011; MacArthur, 1960) are performed based on the eigenvalues, which tell how much a principal component is able to explain from initial dataset, and how many principal components need to choose. To evaluate the prediction performance, the area under the ROC curve has been measured, which measures the sensitivity (TPR) and the specificity (FPR). Sensitivity (TPR) defined as TPTP + FN, corresponds to the correctly predicted positive data points while testing the model and specificity (FPR) defined as FPFP + TN, corresponds to the mistakenly predict negative data points during the test (Owen, 2013). From the result of final prediction after PCA, the area under the curve of SVM method is 0.65. The result shows the effective of dimensional reduction. The classification accuracy decreased and it may lose spatial information.

Recent research usually focuses on the anatomical main group and the therapeutic main group (first level) of ATC groups. For example, Chen L. et. al. were able to map the chemical substructure of drugs, described by the Anatomical Therapeutic Chemical (ATC) classification system to predict chemical similarities (Chen et al., 2012). The study used different benchmarks to predict ATC codes (level 1) by collecting chemical data for 3,883 drugs with high classification performance (73% accuracy). For this thesis, 65% accuracy is predicted by SVM method. Additionally, the evaluation is not significantly affected by the number of folds in cross-validation (Z. Wang et al., 2016). From the final prediction result, I believe that target based predictive models are strongly dependent on the completeness, data size and drug target datasets qualities (Z. Wang et al., 2016). According to the General interpretational of AUC table (Table 8), the accuracy result from this thesis computation approaches is low. The possible reasons

caused low accuracy can be: adverse drug reactions can be a result of a systemic malfunction, which cannot manifest at the gene expression level. Before inducing its global effect, the drug may be processed into various forms (Z. Wang et al., 2016). Because here used different gene expression dataset than Chen and Napolitano (Chen et al., 2012); thus, we cannot compare the performance directly from the result to recent research.

The main conclusion is that classifiers built used support vector machine approach in this thesis study had limited with detecting drug groups, shows the relations between gene expression profile and the 2nd level of ATC system (Anatomical Therapeutic Chemical Classification) (Table 1). (https://www.whocc.no/atc/structure_and_principles/). Although this thesis mainly focused on the prediction of drugs, the similar methods can be used for drug repurposing to predict novel indications. With some simple modifications of prediction approach, new drugs and small molecules for the treatment of many diseases can be identify (Z. Wang et al., 2016).

References

Alpaydin, E. (2010). *Introduction to Machine Learning (Second edition)* (2nd ed.). Massachusetts Institute of Technology. Retrieved from http://cs.du.edu/~mitchell/mario_books/Introduction_to_Machine_Learning_-_2e_-_Ethem_Alpaydin.pdf

Analytics vidhya content team. (2016, March 21). Practical Guide to Principal Component Analysis (PCA) in R & Python. Retrieved May 9, 2017, from https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

Ben-Hur, A., & Weston, J. (2010). A User's Guide to Support Vector Machines. In *Data Mining Techniques for the Life Sciences* (pp. 223–239). Humana Press. https://doi.org/10.1007/978-1-60327-241-4_13

Benyamin, D. (2012, November 10). A Gentle Introduction to Random Forests, Ensembles, and Performance Metrics in a Commercial System. Retrieved May 9, 2017, from http://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics

Borcard, D., Gillet, F., & Legendre, P. (2011). *Numerical Ecology with R*. New York: Springer. Retrieved from http://www.ievbras.ru/ecostat/Kiril/R/Biblio/R_eng/Numerical%20Ecology%20with%20R%20(use%20R).pdf

Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., & Jackel, L. D. (1994). Comparison of Classifier Methods: A Case Study in Handwritten Digit Recognition. Retrieved from http://yann.lecun.org/exdb/publis/pdf/bottou-94.pdf

Breiman, L. (1999). RANDOM FORESTS--RANDOM FEATURES. Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/567.ps.Z

Brownlee, J. (2016, April 11). Naive Bayes for Machine Learning. Retrieved May 9, 2017, from http://machinelearningmastery.com/naive-bayes-for-machine-learning/

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, *2*(3), 27:1–27:27. https://doi.org/10.1145/1961189.1961199

Chen, L., Zeng, W.-M., Cai, Y.-D., Feng, K.-Y., & Chou, K.-C. (2012). Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PloS One*, *7*(4), e35254. https://doi.org/10.1371/journal.pone.0035254

Cornish, R. (2007). Statistics: factor analysis. *Mathematics Learning Support Centre*. Retrieved from http://www.lboro.ac.uk/media/wwwlboroacuk/content/mlsc/downloads/3.3_Factoranalysis.pdf

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*, 273–297.

David, J. M., & Balakrishnan, K. (2010). Significance of Classification Techniques in Prediction of Learning Disabilities, *1*(4), 111–120.

Dernoncourt, F. (2015, January 9). classification - What does AUC stand for and what is it? - Cross Validated. Retrieved May 9, 2017, from https://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it

Dougherty, E. R., Chao, S., Hua, J., Hanczar, B., & Braga-Neto, U. M. (2010). Performance of Error Estimators for Classification. *Current Bioinformatics*, *5*, 53–67.

El-Hachem, N., Gendoo, D. M. A., Ghoraie, L. S., Safikhani, Z., Smirnov, P., Isserlin, R., … Haibe-Kains, B. (2016). Integrative pharmacogenomics to infer large-scale drug taxonomy. *BioRxiv*, 046219. https://doi.org/10.1101/046219

Fazli, S., Afrouzian, R., & Seyedarabi, H. (2009). A Combined KPCA and SVM Method for Basic Emotional Expressions Recognition. In *2009 Second International Conference on Machine Vision* (pp. 84–88). https://doi.org/10.1109/ICMV.2009.67

Genometry, I. (2017). L1000TM Expression profiling services. Retrieved from http://genometry.com

Ghodsi, A. (2006). Dimensionality Reduction A Short Turorial.

Gottlieb, A., Stein, G. Y., Ruppin, E., & Sharan, R. (2011). PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, *7*(1). https://doi.org/10.1038/msb.2011.26

Greg. (2015, June 15). ŷhat | ROC Curves in Python and R. Retrieved May 10, 2017, from http://blog.yhat.com/posts/roc-curves.html

Grigorev, A. (2015, January 18). ROC Anslysis. Retrieved from http://mlwiki.org/index.php/ROC_Analysis

Gunther, E. C., Stone, D. J., Gerwien, R. W., Bento, P., & Heyes, M. P. (2003). Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9608–9613. https://doi.org/10.1073/pnas.1632587100

Halls-Moore, M. (2014, September 12). Support Vector Machines: A Guide for Beginners. Retrieved from https://www.quantstart.com/articles/Support-Vector-Machines-A-Guide-for-Beginners

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multipe Class Classification Problems. *Machine Learning*, *45*, 171–186. https://doi.org/10.1023/A:1010920819831

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Hastie, T., & Tibshirani, R. (2009, February 25). K-Fold Cross-Validation. Retrieved from http://statweb.stanford.edu/~tibs/sta306bfiles/cvwrong.pdf

Holland, S. (2017). Principal Components Analysis [edu]. Retrieved from http://strata.uga.edu/8370/lecturenotes/principalComponents.html

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2016). A Practical Guide to Support Vector Classification. Retrieved from http://www.csie.ntu.edu.tw/~cjlin

Hsu, C.-W., & Lin, C.-J. (2015). A Comparison of Methods for Multi-class Support Vector Machines. Retrieved from https://www.csie.ntu.edu.tw/~cjlin/papers/multisvm.pdf

Iorio, F., Bosotti, R., Scacheri, E., Belcastro, V., Mithbaokar, P., Ferriero, R., … Bernardo, D. di. (2010). Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, *107*(33), 14621–14626. https://doi.org/10.1073/pnas.1000138107

Iskar, M., Campillos, M., Kuhn, M., Jensen, L. J., Noort, V. van, & Bork, P. (2010). Drug-Induced Regulation of Target Expression. *PLOS Computational Biology*, *6*(9), e1000925. https://doi.org/10.1371/journal.pcbi.1000925

Janecek, A. G. K., & Gansterer, W. N. (2008). A comparison of Classification Accuracy Achieved with Wrappers, Filters and PCA. Retrieved from http://www.ecmlpkdd2008.org/sites/ecmlpkdd2008.org/files/pdf/workshops/fsdm/7.pdf

Jensen, K. (2016, March 2). k-fold Cross-validation in IBM SPSS Modeler. Retrieved April 4, 2018, from https://developer.ibm.com/predictiveanalytics/2016/03/02/k-fold-cross-validation-ibm-spss-modeler/

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support Vector Machines in R, *15*(9). Retrieved from Support Vector Machines in R

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., … Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*(6), 673–679. https://doi.org/10.1038/89044

Knerr, S., Personnaz, L., & Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing* (pp. 41–50). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-76153-9_5

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., … Golub, T. R. (2006). The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, *313*(5795), 1929–1935. https://doi.org/10.1126/science.1132939

Lin, H.-T., Lin, C.-J., & Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines, *68*(3), 267.

LINCS Data Portal. (2017, March 15). Retrieved from http://support.lincscloud.org/hc/en-us/articles/203133687-Protocol-Overview-L1000-

Liu, C., Su, J., Yang, F., Ma, J., & Zhou, X. (n.d.). Compound signature detection on LINCS L1000 big data, (Mol Biosyst. 2015 Mar), 714–722. https://doi.org/10.1039/c4mb00677a

Liu, Z., Guo, F., Gu, J., Wang, Y., Li, Y., Wang, D., … He, F. (2015). Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources. *Bioinformatics*, *31*(11), 1788–1795. https://doi.org/10.1093/bioinformatics/btv055

MacArthur, R. H. (1960). On the relative abundance of species, 25–36.

Markham, K. (2014, March 26). Simple guide to confusion matrix terminology. Retrieved May 9, 2017, from http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

Merrett, R. (2016). Random Forest — the go-to machine learning algorithm. Retrieved May 9, 2017, from https://www.techworld.com.au/article/594916/random-forest-go-to-machine-learning-algorithm/

Morgan, M. A. (2016, June 2). Receiver operating characteristic curve | Radiology Reference Article | Radiopaedia.org. Retrieved May 9, 2017, from https://radiopaedia.org/articles/receiver-operating-characteristic-curve

Myatt, G. J., & Johnson, W. P. (2009). *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*. John Wiley & Sons.

Napolitano, F., Zhao, Y., Moreira, V. M., Tagliaferri, R., Kere, J., D'Amato, M., & Greco, D. (2013). Drug repositioning: a machine-learning approach through data integration. *Journal of Cheminformatics*, *5*(1), 30. https://doi.org/10.1186/1758-2946-5-30

Nilsson, N. J. (1998). *Introduction to Machine Learning*. Robotics Laboratory, Standfor University. Retrieved from http://robotics.stanford.edu/people/nilsson/MLBOOK.pdf

Oprea, cristina. (2014). Performance evaluation of the data mining classification methods, *1*, 249–253.

Otterbach, J. (2016, March 24). Principal Component Analysis (PCA) for Feature Selection and some of its Pitfalls. Retrieved from http://jotterbach.github.io/about/

Owen, S. (2013, September 3). Machine Learning: What is an intuitive explanation of AUC? Retrieved from https://www.quora.com/Machine-Learning-What-is-an-intuitive-explanation-of-AUC

Ray, S. (2015, July 28). Beginners Guide To Learn Dimension Reduction Techniques. Retrieved May 9, 2017, from https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shai, S.-S., & Shai, B.-D. (2014). *UNDERSTANDING MACHINE LEARNING-From theory to Algorithms*. Cambridge University Press. Retrieved from http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2015). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940–3941.

Song, Q. (2013). Support Vector Machine and its realization through LIBSVM: A Superficial Review. Retrieved from http://www.songqiuestc.com/uploads/svm_personal_survey.pdf

Vladimirov, M. (2013, November 24). Using the k-Nearest Neighbors Algorithm in R « Web Age Dev Zone. Retrieved May 9, 2017, from http://blog.webagesolutions.com/archives/1164

Wang, Y. H., & Cheng, X. (2010). Heuristics for Multiple Class Classification Problems via ROC Hypersurface. In *2010 Third International Conference on Information and Computing* (Vol. 3, pp. 135–138). https://doi.org/10.1109/ICIC.2010.218

Wang, Z., Clark, N. R., & Ma'ayan, A. (2016). Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics (Oxford, England)*, *32*(15), 2338–2345. https://doi.org/10.1093/bioinformatics/btw168

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., … Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, *34*(Database issue), D668-672. https://doi.org/10.1093/nar/gkj067

Witten, D., James, G., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning-with Applications in R* (Vol. 103). New York: Springer New York. Retrieved from http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577.

Appendix

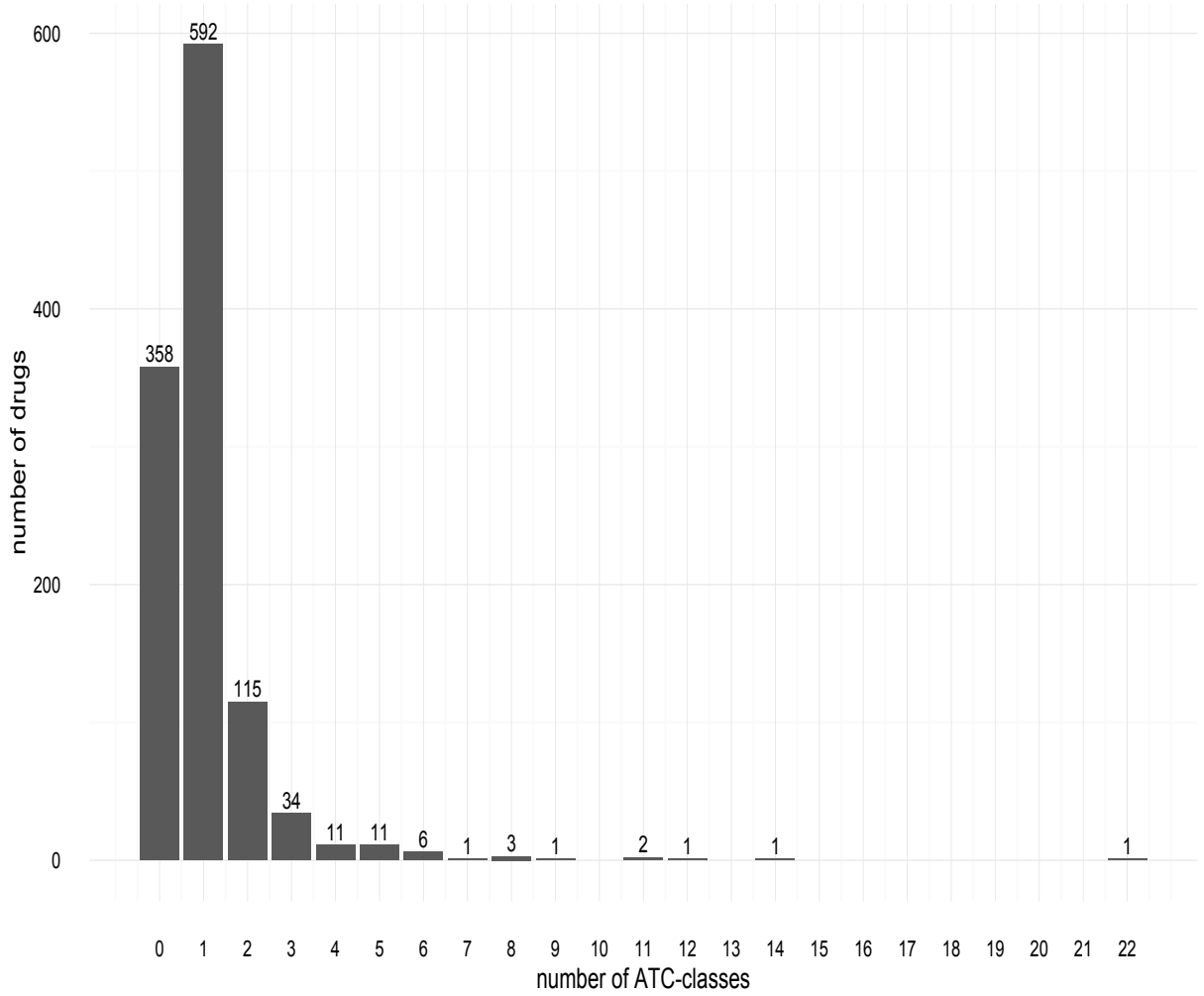

Figure 8. Distribution about the numbers of ATC-classes and drug samples may belong to. There are 358 samples belong to empty class, 592 belong to one class, 115 belong to two classes, 34 belong to three classes, 11 belong to four classes and five classes, 6 belong to six classes, 1 belong to seven classes, 3 belong to eight classes, 1 belong to nine classes, 1 belong to eleven classes, 1 belong to twelve classes, 1 belong to fourteen classes and 1 belong to twenty-two classes.

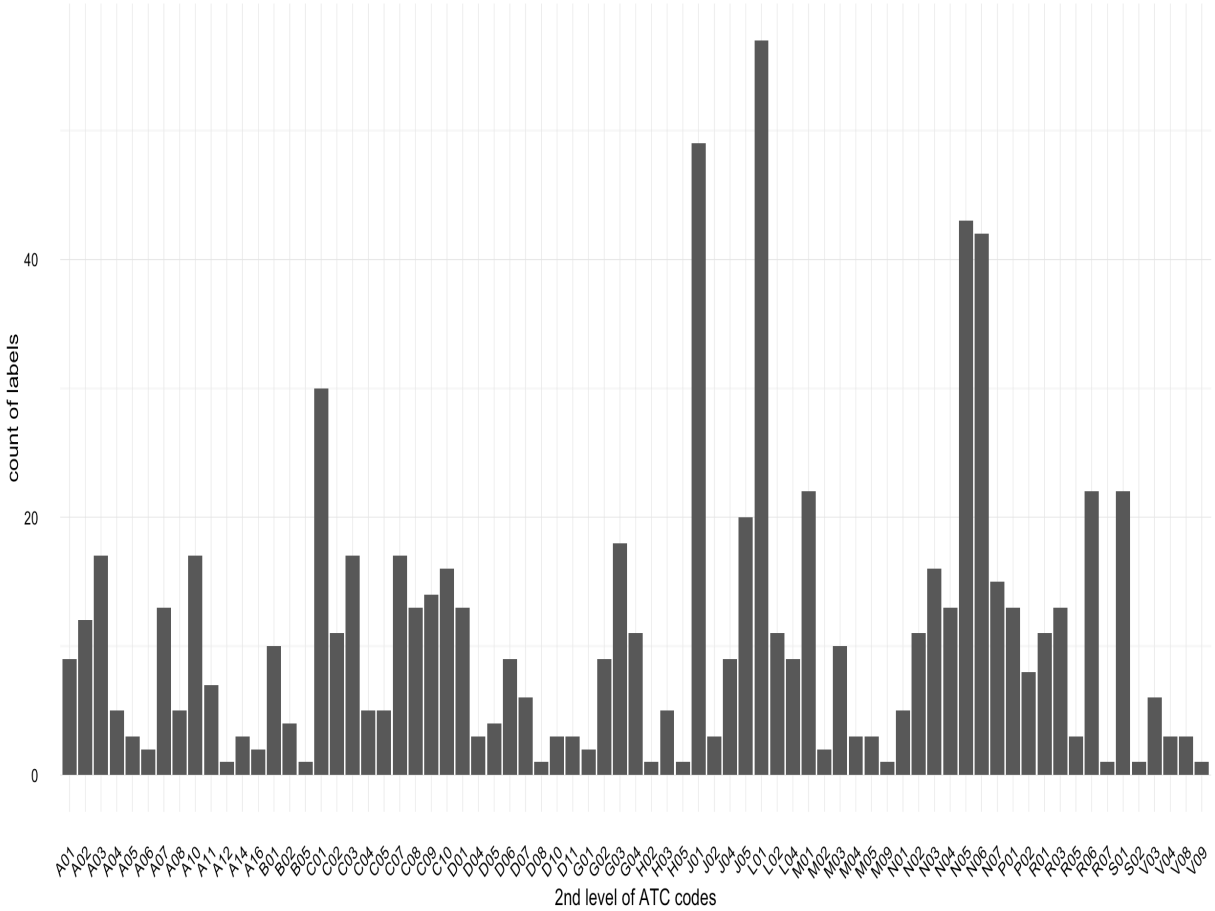

Figure 9. Distribution of 2nd level ATC-classes and drug samples may belong to. The x-axis presents all classes at 2nd level of ATC codes. The y-axis presents the count per class. There are 57 samples belong to L01 class, 49 samples belong to J01 class, 43 samples belong to N05 class, 42 samples belong to N06 class, 30 samples belong to C01 class, 22 samples belong to M01, R06 and S01 classes, 20 samples belong to J05 class, 18 samples belong to C10 and N03 classes, 15 samples belong to N07 class, 14 samples belong to C09 class, 13 samples belong to A07, C08, D01, N04, P01 and R03 classes, 12 samples belong to A02 class, 11 samples belong to C02, G04, L02, N02 and R01 classes, 10 samples belong to B01 and M03 classes, 9 samples belong to A01, D06, G02, J04 and L04 classes, 8 samples belong to P02 class, 7 samples belong to A11 class, 6 samples belong to D07 and V03 classes, 5 samples belong to A04, A08, C04, C05, H03 and N01 classes, 4 samples belong to B02 and D05 classes, 3 samples belong to A05, A14, D04, D10, D11, J02, M04, M05, R05, V04 and V08 classes, 2 samples belong to A06, A16, G01 and M02 classes and 1 sample belong to A12, B05, D08, H02, M09, R07, S02 and V09 classes.