Knowledge Discovery in Microbiology Data: Analysis of Antibiotic Resistance in Nosocomial Infections

Alexey Tsymbal[1], Mykola Pechenizkiy[2], Seppo Puuronen[2],
Michael Shifrin[3], Irina Alexandrova[3]

[1] Dept. of CS, Trinity College Dublin, Ireland, tsymbalo@cs.tcd.ie
[2] Dept. of CS and ISs, Univ. of Jyväskylä, Finland, {mpechen, sepi}@cs.jyu.fi
[3] N.N.Burdenko Institute of Neurosurgery, Russian Academy of Medical Sciences, Moscow, Russia {Shifrin, IAlexandrova}@nsi.ru

Abstract. The goal of this paper is to address the currently serious problem of antibiotic resistance applying knowledge discovery techniques to real hospital data. In this paper we introduce our approach to that problem and the first results of our project aimed to perform exploratory analysis of microbiology data. While discussing preliminary findings we consider possible directions of further research.

1 Introduction

It is known that 3 to 40 percent of patients admitted to hospital acquire an infection during their stay, and that the risk for hospital-acquired infection, or *nosocomial infection,* has risen steadily in recent decades. The frequency depends mostly on the type of conducted operation being greater for "dirty" operations (10-40%), and smaller for "pure" operations (3-7%). For example, such serious infectious complication as postoperative meningitis is often the result of nosocomial infection.

Antibiotics are the drugs that are commonly used to fight against infections caused by bacteria. However, according to the Center for Disease Control and Prevention (CDC) statistics, more than 70 percent of the bacteria that cause hospital-acquired infections are resistant to at least one of the antibiotics most commonly used to treat infections.

Analysis of the microbiological data included in antibiograms collected in different institutions over different periods of time is considered as one of the most important activities to restrain the spreading of antibiotic resistance and to avoid the negative consequences of this phenomenon.

Knowledge discovery in databases (KDD) is a combination of data warehousing, decision support, and data mining that indicates an innovative approach to information and knowledge management. KDD is an emerging area that considers the process of finding previously unknown and potentially interesting patterns and relations in large databases. In this paper we apply KDD techniques to the selected part of real clinical database trying to evaluate possibilities to reveal some interesting patterns of antibiotic resistance.

The paper is organized as follows: in Section 2 we consider the phenomenon of nosocomial infection, in Section 3 the problem of antibiotic resistance is discussed, in

Section 4 data collection and organization is described, in Section 5 we present the preliminary results of data analysis, and in Section 6 we conclude with a brief summary and further research directions.

2 Nosocomial Infections

Infections acquired during a hospital stay are called nosocomial infections. Formally, they are defined as infections arising after 48 hours of hospital admission. For earlier periods it is assumed that the infection arose prior to admission, though this is not always true. [4].

Nosocomial infections are the inevitable consequence of long treatment, especially in Intensive Care Units (ICUs). The first step of this process is the colonization of skin and mucous tunic by hospital microorganism cultures. The peculiarity of these cultures is the acquisition of unpredictable antibiotic resistance according to the policy of the use of antimicrobial medications in the present department or institution.

Factors, contributing to nosocomial infections, include the defects of mucous tunics and skin, long lasting artificial ventilation of lungs, long catheterization of vessels and urinary tracts, implantation of foreign bodies and prosthetic devices, insufficient feeding, decrease in the resistance of organism etc.

Multiple investigations, conducted in different institutions, have shown the possibility of reduction of the number of nosocomial infections by about one third at maximum, even when optimal organization of the treatment process is used. The use of antibiotics with the objective of prophylaxis of nosocomial infections has proven to be ineffective, as pathogens become resistant to the used antibiotics. Normally, antibiotics are used strictly according to indications only.

To treat nosocomial infections, at first a microbiological investigation is normally conducted. In this investigation pathogens are isolated and for each isolated bacterium, an antibiogram is built (represents bacterium's resistance to a series of antibiotics). The user of the test system can define the set of antibiotics used to test bacterial resistance. The result of the test is presented as an antibiogram that is a vector of couples (antibiotic/resistance). The information included in the antibiogram is used to prescribe an antibiotic with a desired level of resistance for the isolated pathogen.

The antibiogram is not uniquely identified given the bacterium species, but it can sometimes vary for bacteria of the same species. This is due to the fact that bacteria of the same species may have evolved differently and have developed different resistances to antibiotics. However, still quite often groups of antibiotics have similar sensitivity when tested on a given bacterium species, despite its strains [5].

3 The Problem of Antibiotic Resistance

Antibiotics, also known as antimicrobial drugs, are drugs that are used to fight against infections caused by bacteria. After their discovery in 1940's they transformed medical care and dramatically reduced illness and death from infectious diseases. However, over the decades the bacteria that antibiotics control have developed resistance

to these drugs. Today, virtually all important bacterial infections throughout the world are becoming resistant. Infectious microorganisms are developing resistance faster than scientists can create new drugs. This problem is known as *antibiotic resistance*, also known as antimicrobial resistance or drug resistance [6].

Antibiotic resistance is an especially difficult problem for nosocomial infections in hospitals because they harbor critically ill patients who are more vulnerable to infections than the general population and therefore require more antibiotics. Heavy use of antibiotics in these patients hastens the mutations in bacteria that bring about drug resistance [6]. Persons infected with drug-resistant organisms are more likely to have longer hospital stays and require treatment with second or third choice drugs that may be less effective, more toxic, and more expensive [6]. In short, antimicrobial resistance is driving up health care costs, increasing the severity of disease, and increasing the death rates of some infections.

4 Data Collection and Organization

The data of our analysis were collected in the N.N. Burdenko Institute of Neurosurgery using the bacterial analyzer "Vitek-60" (developed by "*bioMérieux"*) over the years 1997-2003 and information systems "Microbiologist" (developed by the Medical Informatics Lab of the institute) and "Microbe" (developed by Russian company "MedProject-3").

Each instance of the data used in analysis represents one sensitivity test and contains the following features: *pathogen* that is isolated during the bacterial identification analysis, *antibiotic* that is used in the sensitivity test and the *result of the sensitivity test* itself (sensitive S, resistant R or intermediate I), obtained from "Vitek" according to the guidelines of National Committee for Clinical Laboratory Standards (NCCLS) [3]. The information about sensitivity analysis is connected with *patient,* his or her demographical data (*sex*, *age*) and hospitalization in the Institute (*main department*, *days spent in ICU*, *days spent in the hospital before test*, etc.).

Each instance of microbiological test in the database corresponds to a single specimen that may be blood, cerebrospinal fluid (liquor), urine, etc. In this pilot study we focus on the analysis of meningitis cases only, and the specimen is liquor.

For the purposes of this exploratory analysis we picked up 1423 instances of sensitivity tests including the meningitis cases of the year 2002.

We formed two new features: one for antibiotics and one for pathogens. For antibiotics the 35 different classes of them were grouped into one feature with 4 major categories and for pathogens the 16 different classes were grouped into one feature with 7 major groups.

Each instance had 34 features that included information corresponding to a single sensitivity test augmented with data concerning the used antibiotic, the isolated pathogen, microbiology test result and clinical features of the patient and his/her demographics. These features are summarized in Table 1.

Table 1. Dataset's characteristics

Name	Type
Patient and hospitalization related	
patient	
Sex	{Male, Female}
Age	Integer
recurring stay	{True,False}
days of stay in NSI	Integer
days of stay in ICU	Integer
days of stay in NSI before specimen was received	Integer
bacterium is isolated when patient is in ICU	{True,False}
main department	{1,…,10}
department of stay (departments + ICU)	{1,…,11}
Pathogen and pathogen groups	
pathogen name	{Pat_name1, …, Pat_name17}
gram(+)	{True,False}
staphylococcus	{True,False}
enterococcus	{True,False}
enterobacteria	{True,False}
nonfermenters	{True,False}
Antibiotic and antibiotic groups	
antibiotic name	{Ant_name1, …, Ant_name39}
group1	{True,False}
…	…
group15	{True,False}
sensitivity	{Sensitive, Intermediate, Resistant}

5 Data Analysis

In our experimental studies we used various data-mining techniques available in the machine learning library with Java implementation "WEKA 3.4.2" (available at http://www.cs.waikato.ac.nz/~ml/weka/), which is currently perhaps the most popular library of machine learning algorithms [8].

As one stage of our preliminary analysis we formulated a classification problem aimed to predict the sensitivity of a pathogen to an antibiotic based on data about the antibiotic, the isolated pathogen, and the demographic and clinical features of the patient.

On the whole set of features nonparametic approaches like *3-Nearest Neighbor* (*3NN*) classifier resulted in better accuracy (76.1%) in comparison with parametric approaches like Naïve Bayes. Nearest neighbour classification was the best in comparison with the other classification approaches too (Bayesian Network, BN, a deci-

sion tree C4.5, and the rule-based JRip). In general, recursive partitioning approaches like the C4.5 decision tree and the rule-based JRip performed better than the Bayesian approaches (NB and BN), but they were worse than lazy learning approaches like 3NN with distance weighting.

In our data, classes with instances related to sensitive and resistant cases of pathogens are balanced (47% and 48% correspondingly) and easier to predict. On the contrary, there were too few instances of sensitivity tests where the pathogens' sensitivity was intermediate (5%), and it was difficult for classifiers to make good predictions for this group of instances.

However, Naïve Bayes was able to achieve the accuracy of 70.6% when feature selection was undertaken and the classification model was built on the selected subset of four features, which shows that the features in the original feature set are highly correlated, violating the basic assumption of NB.

Feature ranking according to the *Relief* measure [8] shows that most of information is concentrated in the features related to antibiotics, less information in the features that describe pathogen and even less information is in the features that describe demographics of the patients and the hospitalization context. However, Relief always selects *antibiotic_name, years_old, days_in_icu, pathogen_name, days_before_test, dept_of_stay,* and *main_dept* among the top seven features, which have feature merit values significantly greater than those of the other features.

Association and classification rules mining is a common approach in microbiological data that helps to discover new knowledge about the phenomena or to find support for already known relations between concepts and their features [5]. The rules generator JRip [2] produces a very compact set of classification rules (22 rules) that allow to predict sensitivity in 73.9% of instances.

Beside many interesting rules, several expected relationships between pathogens and antibiotics were found during the expert evaluation of discovered rules. Some of these relationships are stated also by NCCLS. We give below a few examples of discovered rules. It is important to note here, that there were some rules discovered that found associations between sex and antibiotic resistance, age and antibiotic resistance, location of a patient in the hospital and antibiotic resistance. So, some rules were valid only for children departments, some – only for men or women. Some rules were valid for patients who were staying in hospital for a long period and who, therefore, most probably, have nosocomial infection. These rules allowed us to discover local patterns of antibiotic sensitivity, peculiar to a certain group of patients only.

For example, the first two rules shown in Figure 1 represent associations found for children. Departments 1 and 2 in NSI are children departments. These rules show that bacteria isolated in young patients are quite often sensitive (significantly more often than in random) to antibiotics in general irrespective of the pathogen and antibiotic types. The third rule is a pattern inherent to male patients, who have stayed in ICU less than 21 days, who have their first visit to NSI, and in whom strains of *Pseudomonas aeruginosa* were isolated. These patients are quite sensitive to antibiotics too. The fourth rule describes a typical dependency that vancomycin, which is normally applied to gram positive bacteria, has very high effectiveness. The last rule in the figure was unexpected to experts. It describes a relation between the antibiotic "ti-

carcillin/clavulanic acid" and the pathogen *Acinetobacter calcoacedicus baumannii complex*, and shows that the present pathogen is sensitive to the given antibiotic for all the six cases of sensitivity tests considered in the data. This rule, as well as rules 1-3 need further validation with more data and sound statistical tests in order to be able to find the area of their applicability.

```
1: (7.2 < years_old <= 14.4) & (main_dept = 1)  => pat_ab_sens = S (81/24)
2: (days_fefore_test < 16) & (main_dept = 2)  => pat_ab_sens = S (47/7)
3: (pathogen_short_name = p_aeruginosa) & (recurring = FALSE) and (sex = M) &
(days_in_ICU < 21) pat_ab_sens = S (82/14)
4: (antibiotic_short_name = vancomycin) => pat_ab_sens = S (44/1)
5: (antibiotic_short_name = tic_clav) & (pathogen_short_name = a_calc_baumannii) =>
pat_ab_sens = S (6/0)
```

Fig. 1 – Examples of classification rules produced by JRip. The numbers in brackets denote the number of instances satisfying to the left part of the rule (support) and the number of exceptions found for this rule

Evaluation and interpretation of results is an important stage of knowledge discovery process. In this context data visualization plays an important role. In our study we have used Linear Discriminant Analysis (LDA), and data projection techniques: Principal Component Analysis (PCA) and random projection for data visualization purposes – visual identification of different groups, and better understanding of data. LDA was the most useful technique in terms of visualization allowing us to identify interesting sub groups of instances with different behaviour, because it is based on the use of class variable and maximizing the between-class variance of the data on the extracted features, and it extracts 2 features (number of classes minus one), which is very convenient for visualization.

6 Directions of Further Research

In this paper we have presented the results of our pilot experimental study. We found rather high classification accuracy results that look quite promising and optimistic. Because the results are achieved using only the postoperative meningitis data of one year we plan to continue analysis of microbiology data of the whole database covering all the years 1997-2004 to have additional support to discovered patterns and relations.

Beside the modeling approaches applied so far, there are also other interesting directions to continue. One of these might be to enhance the model to a more fine-grained level considering different interesting sub contexts of the whole domain area separately (for example, different pairs antibiotic-pathogen). These sub contexts can be formed for example from the interestingness and/or periodicity points of view.

Another important interesting further research direction is to consider the antibiotic sensitivity as a concept drift problem [7]. Tracking the concept drift might be applied at several different levels: for example, at the level of one hospital, at the level of some time interval(s), of covering several hospitals in one or several countries. In any case recognition of the concept drift as early as possible hopefully helps to start the necessary counter actions in time.

Acknowledgments: This research is partly supported by the Academy of Finland, and the Graduate School COMAS of the University of Jyväskylä, Finland. This material is based upon works supported by the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111.

References

1. Brossette S.E., Sprague A.P., Hardin J.M., Waites K.B., Jones W.T., Moser S.A. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc 1998, 5(4): 373 – 381.
2. Cohen W. 1995. Fast effective rule induction. In: A. Prieditis and S. Russell (Eds.) Proceedings of 12th International Conference on Machine Learning (ICML-95), pp. 115 – 123, Tahoe City, CA, Morgan Kaufman.
3. Ferraro M.J., et al. Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria that Grow Aerobically: Approved Standard: Sixth Edition & Performance Standards for Antimicrobial Susceptibility Testing. Wayne, PA: National Committee for Clinical Laboratory Standarts, NCCLS, 2004. (Documents M7-A6 and M100-S14, www.nccls.org).
4. Gaynes R.P. Surveillance of nosocomial infections: a fundamental ingredient for quality. Infect Control Hosp Epidemiol 1997, 18(7): 475– 478.
5. Lamma E., Manservigi M., Mello P., Nanetti A., Riguzzi F., Storari S., The automatic discovery of alarm rules for the validation of microbiological data, 6th Int. Workshop on Intelligent Data Analysis in Medicine and Pharmacology, IDAMAP 2001, London, UK, 2001.
6. The Problem of Antibiotic Resistance, NIAID Fact Sheet. National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, U.S. Department of Health and Human Services, USA (available at www.niaid.nih.gov/factsheets/antimicro.htm)
7. Tsymbal A. The problem of concept drift: definitions and related work, Technical Report TCD-CS-2004-15, Department of Computer Science, Trinity College Dublin, Ireland, 2004.
8. Witten I., Frank E. 2000. Data Mining: Practical machine learning tools with Java implementations", Morgan Kaufmann, San Francisco.