# Knowledge Discovery from Microbiology Data: Many-Sided Analysis of Antibiotic Resistance in Nosocomial Infections

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

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

**Abstract.** Nosocomial infections and antimicrobial resistance (AR) are highly important problems that impact the morbidity and mortality of hospitalized patients as well as their cost of care. The goal of this paper is to demonstrate our analysis of AR by applying a number of various data mining (DM) techniques to real hospital data. The data for the analysis includes instances of sensitivity of nosocomial infections to antibiotics collected in a hospital over three years 2002-2004. The results of our study show that DM makes it easy for experts to inspect patterns that might otherwise be missed by usual (manual) infection control. However, the clinical relevance and utility of these findings await the results of prospective studies. We see our main contribution in this paper in introducing and applying a many-sided analysis approach to real-world data. The application of diversified DM techniques, which are not necessarily accurate and do not best suit to the present problem in the usual sense, still offers a possibility to analyze and understand the problem from different perspectives.

## 1 Introduction

Nosocomial infections and antimicrobial resistance (AR) are highly important problems that impact the morbidity and mortality of hospitalized patients as well as their cost of care. 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 a serious infectious disease as meningitis is often the result of nosocomial infection.

*Antibiotics* are 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 bacteria that cause hospital-acquired infections are resistant to at least one of antibiotics most commonly used to treat infections.

K.-D. Althoff et al. (Eds.): WM 2005, LNAI 3782, pp. 360 – 372, 2005.

Analysis of 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 AR and to avoid the negative consequences of this phenomenon. Traditional hospital infection control surveillance and localization of hospital infection relies on the manual review of suspected cases of nosocomial infections and the tabulation of basic summary statistics; and AR surveillance consists of the construction of annual or semi-annual, hospital-wide antibiogram summaries. Since such manual activities require considerable time and resources, produced measures and patterns are often not up-to-date and certainly many of potentially useful patterns remain undiscovered. Such relatively inefficient analysis results in coping with the increasing complexity of AR, and proves the importance of introducing computer-based surveillance. Recent reports have described effective computer applications for infection control [9].

It has been widely recognized recently that sophisticated, active, and timely intra-hospital surveillance is needed. Computer-assisted infection control surveillance research has focused on identifying high-risk patients, the use of expert systems to identify possible cases of nosocomial infection, and the detection of deviations in the occurrence of predefined events [1].

Numerous DM algorithms have recently been developed to extract knowledge (previously unknown and potentially interesting patterns and relations) from large databases. In this paper we apply a number of different commonly used DM techniques to real clinical data trying to evaluate possibilities to reveal some interesting patterns of AR and to construct data models that would help in the prediction of AR and in understanding its development.

Many real-world DM studies are focused on a few techniques, which have the best performance in a certain sense (generalization accuracy, explanation power, simplicity, speed etc.), even if the accomplishment of a many-sided analysis might be available from the time and computational resources points of view. Furthermore, in the cases when a number of DM techniques are applied the goal is usually to find a single technique (e.g., a classifier) that has the best performance and to disregard the others. However, in this paper we emphasize the data and problem understanding perspective. We see the main contribution of this paper in demonstrating how the many-sided analysis of a real-world problem has been performed and how the application of diversified DM techniques has helped us to enhance our understanding of the nature of the data that represents the problem.

In our experimental study we apply Naïve Bayesian classification (NB), C4.5 decision trees, *k*- nearest neighbor classification (*k*NN), and a rule-based classifier (JRip) to build antibiotic sensitivity prediction models. Besides, we apply the principle of natural clustering, i.e. grouping data into partitions related to certain pathogen and/or antibiotic types. Different filter-based and wrapper-based feature selection techniques are applied in this study to analyze the importance of the features for predicting the sensitivity of pathogens to antibiotics. As the AR concept is rather unstable and changes over time, we try to track possible changes of the concept, applying three different strategies for this purpose, in order to better understand the patterns of AR process development.

The paper is organized as follows: in Section 2 we consider the phenomenon of nosocomial infection and the problem of AR, in Section 3 data collection and

organization are described, in Section 4 we present the results of our many-sided data analysis, and in Section 5 we conclude with a brief summary and further research directions.

## 2 Nosocomial Infections and the Problem of Antibiotic Resistance

Infections acquired during a hospital stay are called nosocomial infections. Formally, they are defined as infections arising after 48 hours of hospital admission. Infections arising earlier are assumed having 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 arising nosocomial infection is the colonization of skin and mucous tunic by hospital microorganism cultures. The peculiarity of these cultures is the acquisition of unpredictable AR according to the policy of the use of antimicrobial medications in the present department or institution.

Multiple investigations, conducted in different institutions, have shown the possibility of reduction of the number of nosocomial infections by one third only, 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 antibiotics used.

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 this 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 vary significantly for bacteria of the same species. This is due to the fact that the same bacteria of the same species may have evolved differently and have developed different resistances to antibiotics. However, very often groups of antibiotics have similar resistance when tested on a given bacterium species, despite its strains [6].

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 bacteria that should be controlled by antibiotics 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* (AR), also known as antimicrobial resistance or drug resistance [11].

AR is an especially difficult problem for nosocomial infections in hospitals because they attack 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[11].

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 [11]. In short, antimicrobial resistance is driving up health care costs, increasing the severity of disease, and increasing the death rates of some infections.

## 3 Data Collection and Organization

Data for our analysis were collected in the Hospital of N.N. Burdenko Institute of Neurosurgery, Moscow, Russia, using the analyzer *Vitek-60* (developed by *bioMérieux*, www.biomerieux.com) over the years 1997-2003 and the information systems *Microbiologist* (developed by the Medical Informatics Lab of the institute) and *Microbe* (developed by the 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 bacterium 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 the National Committee for Clinical Laboratory Standards (NCCLS) [3]. 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.). These features are summarized in Table 1.

**Table 1.** Dataset characteristics

| Name | Type |
|---|---|
| **Patient and hospitalization related** | |
| sex | {Male, Female} |
| age | [0;72], mean 29.8 |
| recurring stay | {True,False} |
| days of stay in NSI before test | [0;317], mean 87.5 |
| days of stay in ICU | [0;237], mean 34 |
| days of stay in NSI before specimen was received | [0;169] mean 31.6 |
| bacterium is isolated when patient is in ICU | {True,False} |
| main department | {0,…,9} |
| department of stay | {0,…,11} |
| **Pathogen and pathogen groups** | |
| pathogen name | {Pat_name1, …, Pat_name17} |
| group1 | {True,False} |
| … | … |
| group15 | {True,False} |
| **Antibiotic and antibiotic groups** | |
| antibiotic name | {Ant_name1, …, Ant_name39} |
| group1 | {True,False} |
| … | … |
| group15 | {True,False} |
| **sensitivity** | {Sensitive, Intermediate, Resistant} |

Each bacterium in a sensitivity test in the database is isolated from a single specimen that may be blood, 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 4430 instances of sensitivity tests related to the meningitis cases of the period January 2002 – August 2004.

We introduced 5 grouping binary features for pathogens and 15 binary features for antibiotics. These binary features represent hierarchical grouping of pathogens and antibiotics into 5 and 15 categories respectively. Thus, each instance in the dataset had 34 features that included information corresponding to a single sensitivity test augmented with the data concerning the used antibiotic, the isolated pathogen, the sensitivity test result and clinical features of the patient and his/her demographics.

## 4 Data Analysis

In this section we report both the results achieved by different classification models and our experimental many-sided analysis approach including diversified DM techniques. In our experimental studies we have used various data-mining techniques available in the machine learning library with Java implementation WEKA 3.4.2 [14].

### 4.1 Basic Classifiers

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 the data about the antibiotic, the isolated pathogen, and the demographic and clinical features of the patient (Table 1 above). The classification problem was tried to be solved using six classifiers (columns in Table 2): Naïve Bayes (NB), Bayesian Network (BN), three nearest neighbor-based classifiers (1NN, 3NN, and 15NN with weights), and a decision tree classifier (C4.5). As the seventh classifier the rule-based Jrip (Figure 1) was used.

The main accuracy results are presented as rows in Table 2. From the first row of Table 2 (*34 attributes*) it can be seen that the performance of classifiers in terms of accuracies differs much one to another. The Bayesian approaches performed poorly in

**Table 2.** Classification accuracy results for different feature subsets and different natural clusters

| | | NB | BN | 1NN | 3NN | 15NN | C45 |
|---|---|---|---|---|---|---|---|
| 34attributes | | .660 | .665 | .820 | .782 | **.841** | .809 |
| 32attrWOant&pat | | .642 | .660 | .727 | .723 | **.757** | .771 |
| 10attrWOanyAnt&pat | | .593 | .630 | .561 | .658 | **.670** | **.670** |
| 11attrFSbyC45 | | .674 | .686 | .824 | .802 | **.845** | .818 |
| 5 attrFSby1NN | | .735 | .728 | .814 | .773 | **.841** | .794 |
| gram - | 2 134 | .670 | .685 | .831 | .774 | **.840** | .809 |
| gram + | 2 296 | .756 | .768 | .803 | .785 | .830 | **.844** |
| b_lactam | 435 | .682 | .684 | .822 | .736 | **.831** | .819 |
| c_penem | 116 | .875 | .880 | .877 | .883 | **.907** | .858 |
| ceph | 183 | .626 | .615 | .775 | .665 | **.788** | .759 |
| pen | 136 | .726 | .752 | .882 | .875 | **.894** | .865 |

the case when all the features were used. A possible reason for that is that the features are redundant and highly correlated, with complex inter-feature dependencies.

Interestingly, 1NN was significantly better than the 3NN classifier. Maybe the 34 dimensions result in a situation where the space is not dense enough and the second and the third neighbours are actually quite distant instances. This is partly supported by our finding that when weighing was used with weights equal to *1/distance*, 15NN produced much better results.

The C4.5 decision tree classifier, and also the rule-based JRip classifier performed much better than the Bayesian approaches, yet their accuracy was less than the accuracy achieved by the best NN approaches (1NN and 15NN with weights). When we analyzed the rules selected by rule or tree based classification, we noticed that quite often the Intermediate sensitivity class was completely ignored by those rules. This can be explained by looking at the class distribution of the instances. In our data, classes with instances related to the sensitive and resistant cases of pathogens are dominating and nearly balanced (44.4% and 50.7% correspondingly), and easier to predict. On the contrary, there were very few instances of sensitivity tests where the pathogen sensitivity was intermediate (4.9%). Therefore the behavior of rule-based classifiers becomes clearer – instances of the Intermediate sensitivity class are treated as noise without a loss in generalization accuracy. It would be interesting to see in the future research whether this Intermediate class can be recognized with a reasonable accuracy separately at all, or otherwise it might be reasonable to leave this class out of consideration. At least with the considered in this paper approaches the confusion matrix shows that this class has generalization accuracy no better than random.

### 4.2 Feature Selection

In order to analyze the importance of features for predicting sensitivity we applied commonly used automatic feature selection techniques. One group of these techniques is related to the so-called filter approach that assumes evaluation of individual features or feature subsets independently from the learning algorithm. We applied ranking procedures based on the *Relief* and *Chi-square* measures. Both of the methods select *antibiotic_short_name, years_old, total_days_in_icu, patogen_short_name, days before_test, dept_of_stay,* and *main_dept* among the top seven features. The feature ranking techniques show that most information is concentrated in the features related to antibiotics, much 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.

The other group of feature selection techniques corresponds to the wrapper approach that assumes evaluation of feature subsets according to the accuracy of predictive model built on these feature subsets. In our experiments we used the same classifiers for the wrapper-based feature selection. Although feature subsets selected with the wrapper approach were different, *days_before_test, patogen_short_name,* and *antibiotic_short_name* have always been selected. The fourth row of Table 2 (*11attrFSbyC45*) includes accuracy results for basic classifiers when the C4.5 decision tree was used with the wrapper approach to select 11 features from the 32 features (patient name and antibiotic name were excluded). The resulting accuracies are slightly higher for all the basic classifiers than using all the features. Because this

difference is highest with 3NN it supports our suspect that the multidimensionality of the original space has a negative effect.

It is interesting that the best feature subsets for the 1NN, 3NN, 15NN with weights and C4.5 decision tree classifiers were selected when C4.5 was used as the wrapper learning algorithm. We also tried to use the 1NN classifier in the wrapper approach to select 5 features from the 32 features (patient name and antibiotic name were excluded) and the corresponding results are presented in the fifth row of Table 2 (*5attrFSby1NN*). So, surprisingly, e.g., 1NN classifier performed better (.824 vs. .814) when the C4.5 decision tree (not 1NN itself) was used with the wrapper approach. It is interesting to note that the Bayesian basic classifiers give their highest accuracies with this very limited feature set, which is perhaps due to smaller interdependency of features.

Besides the commonly used automatic feature selection techniques, we applied expert (manual) feature selection. Semantically, the sensitivity concept is related first of all to the pathogen and antibiotic features. Therefore it was interesting for us to see how good accuracy can be achieved if the information about the pathogen and antibiotic is excluded from the models. First we excluded the *pathogen name* and *antibiotic name* attributes but left all the grouping features. The corresponding results for the basic classifiers are presented in the second row of Table 2 (*32attrWOant&pat)*. The accuracies are smaller than using all the 34 features but the accuracies are still much higher than 50% and we can assume that groupings of antibiotics and pathogens into categories were appropriate and the grouping features contained relevant information.

Next, we excluded all the attributes related to the pathogen and antibiotic using only the 10 patient and hospitalization related features. The corresponding results for the basic classifiers are presented in the third row of Table 2 (*10attrWOanyAnt&pat*). The accuracies are smaller than those above but they are still much higher than 50% (the accuracy of naïve prediction of the majority class). This fact indicates that in our data exist some interesting patterns independent from antibiotics and pathogens and they are related to the demographics and hospital stay information only. We applied the rule-based classifier JRip on this feature subset in order to find an explanation.

### 4.3 Classification Rules

JRip is Weka's implementation of the RIPPER rule learner [2] that efficiently produces easy interpretable *if-then* rules. Association and classification rules are considered to be quite useful in healthcare as they offer the possibility to extract invaluable information and build important knowledge bases quickly and automatically. In general, association and classification rules mining is a common approach in microbiological data that helps to discover new knowledge about the phenomenon or to find support for already known relations between concepts and their features [1].

In collaboration with medical experts we tried to mine interesting patterns by means of classification rules construction. Beside many interesting rules, several expected relationships between pathogens and antibiotics were found during the expert evaluation of discovered rules. There were some interesting rules discovered that determined associations between sex and AR, age and AR, location of a patient in

the hospital and AR. However, their clinical relevance and utility await the results of prospective clinical studies currently under investigation. Five of the found rules in our study are presented as an example in Figure 1. The first rule e.g. shows that young patients who have not stayed long in ICU are contracted with bacteria that are quite often sensitive to antibiotics in general irrespective of the pathogen and antibiotic types, and thus the problem of AR concerns young patients to a smaller degree.

1: (total_days_in_icu <= 6) & (years_old >= 2) & (years_old <= 14) => pat_ab_sens=S (420/92)
2: (7 < years_old <= 14) & (main_dept = 1) => pat_ab_sens = S (81/24)
3: (days_fefore_test < 16) & (main_dept = 2) => pat_ab_sens = S (47/7)
4: (pathogen_short_name = p_aeruginosa) & (recurring = FALSE) and (sex = M) & (days_in_ICU < 21) pat_ab_sens = S (82/14)
5: (antibiotic_short_name = vancomycin) => pat_ab_sens = S (44/1)

**Fig. 1.** Classification rule examples produced by JRip on a feature subset that includes information about the patients and their hospitalization only. The numbers in the brackets denote the number of instances satisfying to the left part of the rule (support, 420) and the number of exceptions found for this rule (92).

Based on the discussion with infection control practitioners, we have found that although not all classification and association rules were interesting, some suggested potential nosocomial outbreaks and changes of patterns in microbial resistance. It would be interesting to investigate whether the found classification rules correspond to some of the ones known by medical experts as general dependencies (as it might be the case with the example rules) or are these rules just reflecting the peculiarities of the distribution of our data over the selected attributes (which is often referred to as overfitting).

### 4.4 Natural Clustering

We continued our experimental study, applying so-called natural clustering of our data. The main reason to do that is to try to find out the possibly different semantics for each group of pathogens and antibiotics. Therefore we were interested in how the accuracy of classification models varies from one cluster to another and whether it is possible to achieve better accuracy applying a classifier locally in each cluster instead of the use of global classification. The basic accuracy results are presented in the lower parts of Table 2. Rows 6 and 7 (*gram*– and *gram*+) include the accuracies of base classifiers for the pathogen clusters 'gram positive' and 'gram negative'. The base classifiers are developed using all the other features than the feature *gram*+ which is used to build the clusters. The number of instances in each cluster is presented after the name of the cluster. The clusters *gram*– and *gram*+ are approximately of the same size. In Figure 2 the achieved accuracies for each base classifier are presented for both clusters separately, the average of accuracies within those two clusters, and the global accuracy achieved using all the features. For the Bayesian base classifiers the differences of accuracies between the clusters seem to be quite big and at the same time the accuracies for the clusters are always higher than the global accuracies. The NN-classifiers do not manage to achieve higher accuracies in average with clustering compared to the global accuracies. C4.5 base classifiers for

clusters result in at least as high accuracies for clusters as globally and inside the cluster *gram+* even higher than 15NN.

Row 8 (*b_lactam*) of Table 2 includes the accuracies of the base classifiers for the antibiotic group, whose subgroups are included in the three following rows (*c_penem*, *ceph*, and *pen*). The base classifiers are produced leaving the excluded antibiotic group features away. The results are presented in Figure 2 (right), subgroups *c_penem*, *ceph*, and *pen* are labeled *ant1*, *ant2* and *ant3* respectively. Inside the clustering related to *b_lactam* the global base classifiers are no worse than the *ceph*-cluster classifiers for every type of base classifiers. On the contrary for the two other clusters (*c_penem* and *pen*) the cluster classifiers outperform the global classifiers for every type of base classifiers. It can also be seen from the figure that the average accuracy of classifiers (except C4.5) is higher when they are applied locally within each cluster comparing to the global classifiers' accuracy.

In [8] different dimensionality reduction techniques have been applied locally in produced pathogen clusters. The results of our experiments show that the proper selection of a local DR technique can lead to a significant increase of predictive accuracy comparing to the global classification with or without DR. The amount of features extracted or selected locally is always smaller than that in the global space that also shows the usefulness of natural clustering in coping with data heterogeneity.

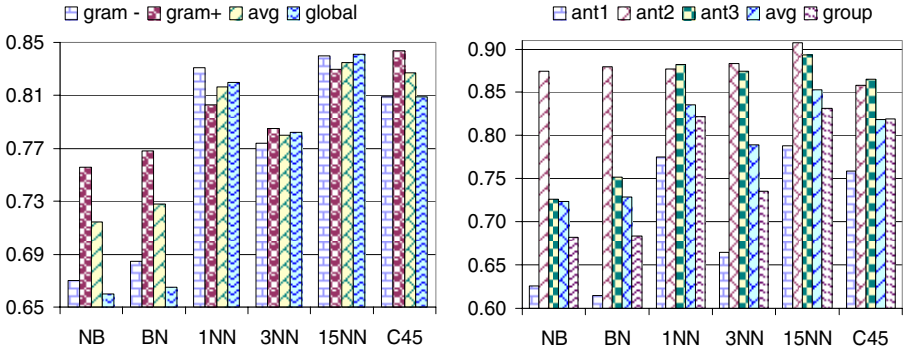


**Fig. 2.** Classification accuracies for two main pathogen clusters (left) and for b_lactam antibiotics clusters (right)

### 4.5 Tracking Concept Drift

Most DM algorithms assume that data is a random sample from some stationary distribution, while real data in clinical institutions are gathered over long time periods (months or years) and therefore naturally violate this assumption. Kukar [5] states that even in most strictly controlled environments some unexpected changes may happen due to fail and/or replacement of some medical equipment, or due to changes in personnel. Clearly, the sooner some change is discovered the better, since corresponding actions can be applied to prevent the undesirable effects.

Often the cause of change is hidden, not known a priori, making the learning task more complicated. Changes in the hidden context can induce more or less radical changes in the target concept, which is generally known as the problem of concept

drift [12, 13]. An effective learner should be able to track such changes and quickly adapt to them.

It is commonly acknowledged that the sensitivity to antibiotics of many bacteria may change over time significantly due to the antibiotic resistance phenomenon described above causing a significant level of concept drift in the AR domain. Antibiotic resistance is a typical example of concept drift in the medical context. New pathogen strains develop their resistance to antibiotics which were previously effective through mutation and natural selection.

Beside that, a significant level of virtual concept drift is pertinent to the AR domain. Hidden changes in context may not only be a cause of a change of target concept, but may also cause a change of the underlying data distribution. Even if the target concept remains the same, and it is only the data distribution that changes, this may often lead to the necessity of revising the current model, as the model's error may no longer be acceptable with the new data distribution.

In this paper we suggest a window-based approach to analyze the underlying concept drift. We divide the data into blocks of equal size (or equal time intervals) and use these blocks as train and test sets sequentially in order to understand the underlying concept and data distribution changes. We suggest three strategies for that which use different blocks as train and test sets: (I) build a model for each of the blocks sequentially, besides the last one, and test it on the last block; (II) build a model for each of the blocks sequentially, besides the last one, and test it on the next block; and (III) build a model on the first block and test it on each of the next blocks.

These strategies, applied together, and with their results depicted in a single graph, help to better understand underlying changes in the concept and in data distribution. We have constructed such graphs for different time intervals and different learners, and discussed them with medical experts. Discussion of the experimental results with medical experts revealed a few interesting, expected and unexpected, dependencies.

Expectedly enough, the most interesting graphs are built with 15NN, which is the strongest learner in this domain. Most of the graphs constructed confirm our expectations that a significant level of concept and data distribution drift is pertinent to the AR domain. In Figure 3, two graphs are shown for the three strategies with the 15NN learner and 3-month time intervals, corresponding to the seasons of year in Russia (and many European countries), starting with March 2002 (Spring), and December 2002 (Winter).

A few interesting dependencies can be seen in this figure. First, clear periodicity in behaviour is seen (Strategy III). The model for Spring 2002 (left) has its highest accuracy in Springs 2003 and 2004. The same is with the Winter 2002 model (right), which works well again in Winter 2003, and worse in all the other seasons. This is not true, however, with the Summer and Autumn models (not shown here). There is no periodicity with them. After discussion of this phenomenon with medical experts, a hypothesis has been raised that this happens due to the typical bacteria outbreaks which always happen in Winter and in Spring. In terms of these outbreaks, Winter and Spring are more similar seasons to each other, in contrast to Autumn and Summer, which can be drastically different year to year, depending much on the weather. This hypothesis needs to be checked with other data, in other contexts (different hospitals, countries, bacteria, diseases, etc), but it was accepted after

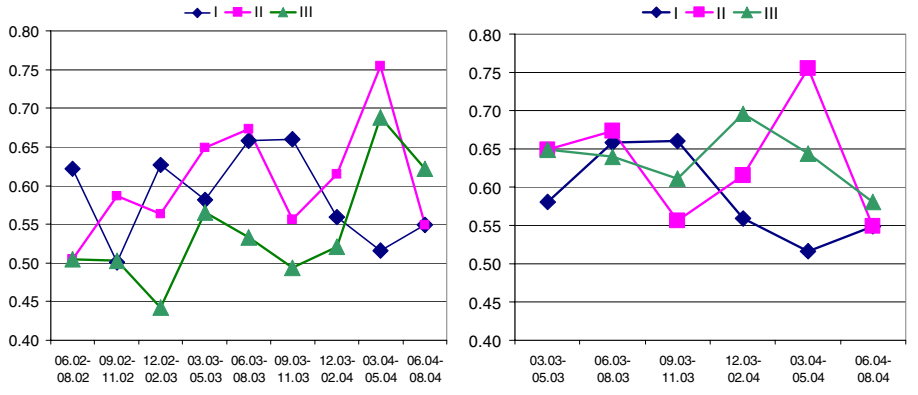


**Fig. 3.** Tracking CD with the 3 strategies. Accuracy results for Spring 2002 (left) and Winter 2002 (right)

analysis of the tracking concept drift graphs by the experts as very interesting and promising hypothesis.

Among the seasonal models, the Winter models are always much more stable and accurate, with accuracy not dropping much on the other seasons (see Figure 3, right, Strategy III). This behaviour was identified by experts as interesting and a possible explanation for it is that the winter period is characterized by a much bigger variability of bacteria than all the other seasons. In Winter, it is often so that the same bacteria strains appear, common for the other seasons, besides those common for Winter only. This hypothesis needs further research as well and experimental check in other contexts.

## 5 Conclusions and Directions of Further Research

Surveillance of AR and nosocomial infections is one of the most important functions of a hospital infection control program.

In this paper we have presented the results of our experimental study in AR prediction for nosocomial infections. We have achieved rather high generalization accuracy (84.5%) that is quite promising in terms of better understanding the problem and patterns of AR. The results were achieved using data with patients having meningitis over the last three years only, and we plan to continue our analysis of the whole NSI database of nosocomial infections including older data collected since 1997. Beside good accuracy results achieved, we emphasized the data and problem understanding perspective.

Besides, in this paper we demonstrated how complex many-sided analysis of a real-world problem could have been performed, how application of different DM techniques has helped us to understand better the nature of the data that represents the problem.

DM makes it easy for experts to inspect patterns that might otherwise be missed by usual (manual) infection control surveillance methods. However, the clinical relevance and utility of these findings await the results of prospective studies.

Besides the modeling approaches applied so far, there are also other interesting promising directions to continue. One of these might be to enhance the models to a more fine-grained level considering different sub contexts of interest of the whole domain area separately. These sub contexts can be formed for example from the interestingness or periodicity points of view. For example, specific antibiotic-pathogen pairs could be isolated, certain time intervals chosen, and the peculiarity of behaviour of AR for these clusters analyzed.

Promising area not included in the present study is feature extraction. Our preliminary experiments show that this domain includes many correlated and redundant features, and the inherent dimensionality of the data is relatively low (10 features extracted with PCA is enough to obtain the accuracy of 80% with *k*NN). Feature extraction might be yet another perspective helping to understand better the problem with our many-sided analysis.

Another important direction of further research is to analyze further antibiotic sensitivity as a concept drift problem. Likewise, infection control systems require or will require tools that recognize trends in nosocomial infection and AR in an efficient and timely manner. Tracking and handling concept drift, when applied at the levels of one hospital or many hospitals in one or several countries, helps to start necessary counter actions in time. We are currently developing a technique based on an ensemble of classifiers with dynamic integration to handle concept drift in the AR data.

Besides the issues of data collecting and cleaning (in a timely fashion) for AR analysis, the information on how results of the analysis will be used to make proactive decisions (that possibly would change current practice in a hospital) is of high importance.

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

## References

1. Brossette SE, Sprague AP, Jones WT, Moser SA A data mining system for infection control surveillance, Methods of Information in Medicine 39(4-5), 2000, pp. 303-310
2. Cohen W. 1995. Fast effective rule induction. In: Proc. of 12th International Conference on Machine Learning (ICML-95), pp. 115-123, 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. Kukar M. Drifting concepts as hidden factors in clinical studies. In: Proc. 9th Conf. on Artifi-cial Intelligence in Medicine in Europe, AIME 2003, Springer, LNCS, 2003, 355-364.

6. 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, UK, 2001.
7. Ma L, Tsui FC, Hogan WR, Wagner MM, Ma H. A Framework for Infection Control Surveillance Using Association Rules. In Proc American Medical Informatics Association Annual Fall Symposium, Omni Press CD, pp. 410-414, 2003
8. Pechenizkiy M., Tsymbal A., Puuronen S. 2005. Local Dimensionality Reduction within Natural Clusters for Medical Data Analysis, (to appear) In Proc. 18th IEEE Int. Symp. on Computer-Based Medical Systems CBMS'2005, IEEE CS Press.
9. Samore M, Lichtenberg D, Saubermann L, et al. A clinical data repository enhances hospital infection control. In Proc American Medical Informatics Association Annual Fall Symposium, 1997; 56–60.
10. Streed SA, Sheretz RJ, Reagan DR: Computers in hospital epidemiology. In: Mayhall CG (ed), Hospital Epidemiology and Infection Control, Baltimore:Williams & Wilkins, Chapter8, pp. 115-122.
11. 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)
12. 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.
13. Widmer G., Kubat M., Effective learning in dynamic environments by explicit context tracking, Proc. 6th European Conf. on Machine Learning ECML-1993, Springer-Verlag, Lecture Notes in Computer Science 667, 1993, 227-243.
14. Witten I., Frank E. 2000. Data Mining: Practical machine learning tools with Java implementations", Morgan Kaufmann, San Francisco.