Discovering Validation Rules from Microbiological Data

E. LAMMA and F. RIGUZZI and S. STORARI

Department of Engineering, University of Ferrara, Via Saragat 1, 44100 Ferrara, Italy

P. MELLO

DEIS, University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy

A. NANETTI

Clinical, Specialist and Experimental Medicine Department, Microbiology section, University of Bologna, Bologna, Italy

Received 31 Dec 2001

Abstract A huge amount of data is daily collected from clinical microbiology laboratories. These data concern the resistance or susceptibility of bacteria to tested antibiotics. Almost all microbiology laboratories follow standard antibiotic testing guidelines which suggest antibiotic test execution methods and result interpretation and validation (among them, those annually published by NCCLS [2)3)]). Guidelines basically specify, for each species, the antibiotics to be tested, how to interpret the results of tests and a list of exceptions regarding particular antibiotic test results. Even if these standards are quite assessed, they do not consider peculiar features of a given hospital laboratory, which possibly influence the antimicrobial test results, and the further validation process.

In order to improve and better tailor the validation process, we have applied knowledge discovery techniques, and data mining in particular, to microbiological data with the purpose of discovering new validation rules, not yet included in NCCLS guidelines, but considered plausible and

correct by interviewed experts. In particular, we applied the knowledge discovery process in order to find (association) rules relating to each other the susceptibility or resistance of a bacterium to different antibiotics. This approach is not antithetic, but complementary to that based on NCCLS rules: it proved very effective in validating some of them, and also in extending that compendium. In this respect, the new discovered knowledge has lead microbiologists to be aware of new correlations among some antimicrobial test results, which were previously unnoticed. Last but not least, the new discovered rules, taking into account the history of the considered laboratory, are better tailored to the hospital situation, and this is very important since some resistances to antibiotics are specific to particular, local hospital environments.

Keywords Knowledge Discovery and Data mining, Microbiology, Knowledge Based Systems, Knowledge Elicitation

§1 Microbiological Data validation

A huge amount of data regarding bacterial infections is daily collected from clinical microbiology laboratories. In particular, for microbiological analyses, data includes: information about the patient (sex, age, hospital unit where the patient has been admitted), the kind of material (specimen) to be analysed (e.g., blood, urine, saliva, pus, etc.) and its origin (the body part where the specimen was collected), the date when the specimen was collected (often substituted with the analysis request date), the species and the antibiogram of each identified bacterium.

For each isolated bacterium, the antibiogram represents its resistance to a series of antibiotics. The set of antibiotics used to test bacterial resistance can be defined by the user, and the antibiogram is a vector of couples (antibiotics, resistance), where four types of resistance are recorded: R when resistant, I when intermediate, S when susceptible, and null when unknown.

Almost all microbiology laboratories follow standard antibiotic testing guidelines which suggest antibiotic test execution methods and result interpretation. For instance, very often groups of antibiotics have similar answer when tested on the same bacterium species and when this accordance is not respected data cannot be validated.

NCCLS [2)], an international standard organization recognised by almost all laboratories as reference in routine work, writes an annual compendium, titled "Performance Standards for Antimicrobial Susceptibility Testing" [3)], regarding

testing guidelines for microbiological laboratory. NCCLS guidelines are basically composed, for each species, of a table that specifies the antibiotics to be tested, a table that specifies how to interpret the test of antibiotics and a list of exceptions regarding particular antibiotic test results. NCCLS guidelines are used for validating microbial results represented by antibiograms (e.g., automatically in systems like WHONET 5 [8], TheraTrac 2 [11] and ESMIS [4][5], or manually by microbiologists), possibly raising alarms when the results are not in accordance with each other.

We have applied knowledge discovery techniques, and data mining in particular, to microbiological data in order to discover new validation rules, not yet included in the NCCLS compendium, but eventually plausible and correct according to interviewed microbiology experts. In particular, we have exploited the WEKA system [6] in order to find (association) rules relating to each other the susceptibility or resistance to different antibiotics. The knowledge discovery approach here described is not antithetic, but complementary to that based on NCCLS rules: it proved very effective in validating some of them, and also in extending that compendium by "discovering" new rules not yet considered. The discovered knowledge has lead microbiologists to be aware of new correlations among some antimicrobial test results, which were previously unnoticed. Last but not least, new discovered rules, taking into account the history of the considered laboratory, are better tailored on the hospital situation, and this is very important since some resistances to antibiotics are specific to particular, local hospital environments.

The discovered association rules have been then transformed into alarm rules (through a syntactic transformation) to be used for data validation in the knowledge based system ESMIS.

The work here described is part of a wider project between University of Bologna and Dianoema S.p.A. (an Italian company developing software systems for the health care market), aiming at implementing the ESMIS knowledge based system [4][5] for the validation of microbiological data and the generation of alarms when critical situations occur. Within the project, part of ESMIS knowledge base has been defined by hand in accordance with NCCLS documents and with microbiologists' suggestions, and part has been automatically produced by discovering association rules (later on confirmed by experts).

In the rest of the paper we report about the knowledge discovery activity we performed, and show some results obtained from a microbiological database

containing the analysis performed along two years.

The paper is organized as follows. Section 2 presents preliminary information on association rules, and on the WEKA system in particular. Section 3 shows how alarm rules are generated from the discovered association rules. Section 4 describes what we have discovered by applying WEKA to a database containing the microbiological analyses done along two years by the microbiological laboratory of an Italian hospital. Related work is surveyed in section 5. We conclude and mention future work in section 6.

§2 Discovery of Association Rules

Association rules describe correlation of events and can be regarded as probabilistic rules. Events are "correlated" if they are frequently observed together. Good examples from real life are databases of sales transactions. In this case the aim is to find which items are usually bought together, information that is used for developing successful marketing strategies.

Given a table T, an *association rule* is a rule of the form

$\mathrm{A}_1{=}\mathrm{v}_{A_1}$, $\mathrm{A}_2{=}\mathrm{v}_{A_2}$,...,$\mathrm{A}_j{=}\mathrm{v}_{A_j} \Rightarrow \mathrm{B}_1{=}\mathrm{v}_{B_1}$, $\mathrm{B}_2{=}\ \mathrm{v}_{B_2}$,...,$\mathrm{B}_k{=}\mathrm{v}_{B_k}$

where A_1, A_2,..., A_j, B_1, B_2,...,B_k are attribute names of table T and v_{A_1}, v_{A_2},...,v_{A_j}, v_{B_1}, v_{B_2},...,v_{B_k} are values such that v_{A_l} (v_{B_h}) belongs to the domain of the attribute A_l (B_h). A record r of T satisfies a conjunction $\mathrm{A}_1{=}\mathrm{v}_{A_1}$, $\mathrm{A}_2{=}\mathrm{v}_{A_2}$,...,$\mathrm{A}_j{=}\mathrm{v}_{A_j}$ if all the equivalencies are true given the values of r.

The rule $X \Rightarrow Y$ holds with *confidence* c in table T, if and only if $c\%$ of records in T that satisfy X also satisfy Y.

The rule $X \Rightarrow Y$ has *support* s in table T, if and only if $s\%$ of records in T satisfies $X \cup Y$.

Given a table T, the task of mining association rules can be reformulated as finding all association rules with at least a minimum support (called *minsup*) and a minimum confidence (called *minconf*), where *minsup* and *minconf* are user-specified values. Of course, the higher is the support of a rule, the more general is the situation the rule represents. Moreover the higher is the confidence the fewer are the exceptions to the represented situations in the database.

In order to learn association rules for validating microbiological data, we have exploited the WEKA system [6], a collection of machine learning algorithms for solving real-world data mining problems. WEKA is written in Java and runs on almost any platform. It contains algorithms for performing classification, numeric prediction, clustering and learning association rules. As regards

association rule learning, WEKA employs the APRIORI algorithm [7].

§3 Generation of Alarm Rules

Generated association rules represent frequent patterns occurring in the database. Very often, the discovered rules are possibly related to each other according to the following generality relation.

Rule R1 is *more general* than rule R2 if they have the same consequent, but the conditions in R2's antecedent are a superset of those in R1's antecedent. For instance among the four rules below:

1. Amoxicillin+ClavulanicAcid=S, Clindamycin=S ==> Oxacillin=S
2. Amoxicillin+ClavulanicAcid=S, Clindamycin=S
Trimethoprim+Sulfamethoxazole=S ==> Oxacillin=S
3. Amoxicillin+ClavulanicAcid=S, Clindamycin=S, Penicillin=R ==> Oxacillin=S
4. Amoxicillin+ClavulanicAcid=S, Clindamycin=S, Penicillin=R
Trimethoprim+Sulfamethoxazole=S ==> Oxacillin=S

(for the sake of simplicity, we have omitted support and confidence in the reported alarm rules).

Rule 4 is the most specific, rule 1 is the most general, and rule 2 and 3 are intermediate (and not comparable with each other).

Most general association rules represent normal, minimal patterns which occur frequently in the database.

To the most general rules, we have then applied syntactic transformations in order to produce alarm rules, to be used in ESMIS for data validation. Abnormality, which we want to capture in the data validation process, is represented by antibiograms which do not *satisfy* some discovered rule, i.e., satisfy the rule antecedent but not its consequent (the implication is not true). Therefore, if an association rule of the kind: $X \Rightarrow Y$ represents the *regular* (and usually quite frequent) situation, the rule: $X, not(Y) \Rightarrow alarm(Y)$ (where the consequent is complemented and moved to the antecedent) represents an *abnormal* situation. When X and $not(Y)$ simultaneously occur, an alarm has to be raised because the usual value for Y should be true instead of false, when X is true.

The condition $not(Y)$ has been obtained in the following way: when Y is a singleton condition, we have considered the result for an antibiotic in an antibiogram as two-valued, where R is the complementary value of S and vice-versa. For instance, the alarm rule produced from rule 1 is:

1'. Amoxicillin+ClavulanicAcid=S, Clindamycin=S, Oxacillin=R

==> alarm(Oxacillin=S)
Otherwise, when Y is a composed condition, e.g.:
537. Oxacillin=R ==>Amoxicillin+ClavulanicAcid=R,Penicillin=R
we just move its negation to the body of the alarm rule. In this case, for the sample rule above n. 537, we obtain the following alarm rule:
537'. Oxacillin=R, not([Amoxicillin+ClavulanicAcid=R, Penicillin=R])]
==>alarm([Amoxicillin+ClavulanicAcid=R, Penicillin=R])

§4 Knowledge Discovery and Interpretation

We have applied WEKA to an Italab C/S database [*1] containing data about bacterial antibiograms which where collected over a 24-month period at the Clinical, Specialist and Experimental Medicine Department - Microbiology Section of the University of Bologna.

Each record describes a single isolated bacterium with values for the following attributes: organism name, gram stain/morphology, date of collection, nosocomial status, source of isolate (e.g., sputum, blood, urine), location of patient in the hospital (e.g., Internal Unit or External Unit, for the time being), and test results: Resistant (R), Intermediate (I), Susceptible (S) or Unknown (U) for each member of a set of antimicrobials.

We have discovered alarm rules pertaining to the bacterium species Staphilococcus Aureus, Escherichia Coli and to a bacterium family including four species belonging to Enterobacteriaceae.

Since microbiologists are much more interested in discovering new insights concerning bacterial resistance to antimicrobials rather than concerning their susceptibility, the available data have been filtered and by removing records where the bacterium is susceptible to all the tested antimicrobials.

After the rules are discovered, they have been analyzed one by one by the authors in order to identify rules already present in ESMIS and they have been analyzed by microbiologists in order to identify new relationships in the data. Microbiologists looked for two kinds of rules: general rules describing the expected behaviour of bacteria and exceptional rules describing rare or novel events. This latter kind of rules can be considered as a chance as defined in [?)] since they represent a risk that must be taken into account by microbiologists.

[*1] Italab C/S is a Laboratory Information System, developed by Dianoema S.p.A. [1)], used in most Italian hospitals to manage all the activities of their laboratories. Italab C/S stores all the information concerning patients, the analysis requests and the analysis results.

Usually general rules have high support and confidence, while specific rules have low support and confidence.

4.1 Staphilococcus Aureus

The considered dataset for Staphilococcus Aureus contains 7009 records having, among the attributes, 41 different antibiotics. The filtering process reduced the number of record to 3734.

We run the system with *minsup* equal to 0.1 and with decreasing *minconf*. The number of most general rules found was around 6500. Out of these, 10 are among the 27 rules that are present in the NCCLS report (and in the ESMIS knowledge base) regarding Staphilococcus aureus. In particular, we have discovered those rules which relate to each other the results of two classes of antibiotics, i.e., Oxacillin and Penicillin (when a bacterium is resistant to Oxacillin it must also be resistant to any kind of Penicillin), and the resistance result for Oxacillin and Penicillin with β-lactamase inhibition (when a bacterium is resistant to Oxacillin it must also be resistant to any Penicillin with β-lactamase inhibition). For instance, the following two instances of these general rules were found (the first has confidence equal to 1, the second equal to 0.99):
537. Oxacillin=R,not([Amoxicillin+ClavulanicAcid=R, Penicillin=R])
==>alarm([Amoxicillin+ClavulanicAcid=R, Penicillin=R]
2071. Clarithromycin =R, Oxacillin=R,
not([Amoxicillin+ClavulanicAcid=R, Ceftriaxone=R, Penicillin=R]
==> alarm([Amoxicillin+ClavulanicAcid=R,ceftriaxone=R, Penicillin=R])

Furthermore, among the discovered rules, microbiologists have identified two rules which were not present in ESMIS knowledge base, but were considered relevant for ESMIS. The identified rules were (the first with confidence equal to 1, the second equal to 0.99):
1080'. Teicoplanin =S, Vancomycin =R ==>alarm(Vancomycin =S)
1539'. Vancomycin =S, Teicoplanin =R ==>alarm(Teicoplanin =S)
which relate to each other the results of two (last-generation) antibiotics (i.e., Teicoplanin and Vancomycin).

4.2 Escherichia Coli

The dataset for Escherichia Coli contains 7165 records having, among the attributes, 25 different antibiotics.

The filtering process reduced the number of record to 3285. From these

data, with a minimum support equal to 0.8 (and minimum confidence equal to 1) around 3500 most general rules were discovered. Among these, we did not find any of the 7 NCCLS rules present in ESMIS for this microorganism. However, two rules not present in ESMIS were considered relevant for the system by microbiologists:

Cefotaxime=S, Ceftazidime =R ==>alarm(Ceftazidime =S)

Ceftazidime =S , Cefotaxime=R ==>alarm(Cefotaxime=S)

This couple of rules relates to each other the results of two classes of antibiotics, i.e., Cefotaxime and Ceftazidime (when a bacterium is susceptible to Cefotaxidime it must also be susceptible to Ceftazidime, and vice-versa).

With lower support, but with minimum confidence still equal to 1, we have also discovered a rule already considered in ESMIS in accordance with the NCCLS compendium, the one relating the resistance to Piperacillin with the resistance to Ampicillin when the bacterium was isolated from the urinary tract.

4.3 Enterobacteriaceae

We have also done further experiments by considering four different bacteria belonging to the Enterobacteriaceae family: Enterobacter Cloacae, Klebsiella Oxytoca, Klebsiella Pneumoniae and Proteus Mirabilis. The considered dataset contains 3387 records having, among the attributes, 28 different antibiotics.

The filtering process reduced the number of record to 2656. From this data, with minimum support equal to 0.68 (and minimum confidence equal to 1) we have discovered around 2500 most general rules. Among the discovered rules, we did not find any of the 7 NCCLS rules present in ESMIS for these microorganims. However, we have found two rules not present in ESMIS that were considered relevant for the system by microbiologists. These are the same as those found for the Escherichia Coli: the rules relating to each other the results of Cefotaxime and Ceftazidime.

With lower support, but with minimum confidence still equal to 1, we have also discovered one rule already considered in ESMIS in accordance with the NCCLS compendium, the one relating the resistance to Cefotaxime with the resistance to Cephalotin.

4.4 Trivial Rules

As shown above, the number of discovered rules is in the order of several

thousand, while the interesting rules are usually in order of a few tens: the number of uninteresting or trivial rules is thus quite high and the process of identifying interesting rules among the trivial ones is quite long.

In order to overcome this problem, we are developing a filtering program which can guide and help the microbiological experts in focusing on certain rules, allowing the selection of rules that satisfy certain conditions. A condition is a obtained as a combination of patterns by means of the logical operators AND, OR and NOT. A pattern is a conjunction of equations <attribute>=<value> that must be in the right or left side. In the equivalences <attribute>=<value>, in place of <attribute> it is possible to specify a set of attributes, e.g., it is possible to specify an equivalence of the form <family of antibiotics>=<value>.

This filtering system allows to eliminate uninteresting rules not on the basis of support and confidence but on the basis of microbiologists' interests. This is very important since the interesting rules may represent rare events and thus have low support and confidence.

§5 Related Work

During the last few years, many surveillance systems have been developed in order to monitor microbiological analysis results and to early identify infection and epidemiological events. Some of these systems encompassed also data validation. For instance, WHONET 5 [8)] is a software for the management of microbiology laboratory test results; GermWatcher [9)] is an expert system which applies both local and international criteria for detecting potential nosocomial infections, according to criteria developed by CDC's NNIS [10)] (Center for Disease Control, National Nosocomial Infection Surveillance); TheraTrac 2 [11)] is a system for microbiological data validation and real-time alarming, deeply integrated with Vitek, an expert system for test results validation, that is integrated, in its turn, with some analysis instruments. All these systems use international standard guidelines in order to define controls to be executed on laboratory test results, but in none of them knowledge discovery has been applied, as we have done for the ESMIS knowledge base (see [4)5)]).

It is worth mentioning some work done in the microbiological and medical field that apply inductive and data mining techniques. Previous work on the detection of data inconsistencies in patient records has been done by applying inductive learning to a database of atherosclerotic coronary heart disease patients [12)]. In particular, confirmation rules for the detection of outliers are discovered

by exploiting inductive methods. The authors also consider the application of descriptor based classifiers.

In [14], data mining techniques are applied to patient data from several hospitals and along three years in order to discover associations, e.g., within diagnoses and medical treatment, with the purpose of enhancing medical quality management.

In [15], the system PTAH is presented that was developed for the analysis of antibiogram data in order to help medical doctors in the prescription of antibiotics for the treatment of nosocomial infections. PTAH performs four types of analysis: resistance level over time, hierarchical clustering of antibiograms, similarity of antibiograms and effectiveness of antimicrobials over time.

In [16], the demographic clustering algorithm that is enclosed in Intelligent Miner [17] is applied in order find interesting clusters of antibiograms.

A related work which is also worth mentioning is reported in [18][19], where the system DMSS is described. In DMSS the mining of association rules is applied to microbiological data in order to automatically identify new, unexpected, and potentially interesting patterns in hospital infection control and public health surveillance data. To this purpose, a bias (possibly subject to be changed in a certain number of iterations of the mining process) is given by the experts in order to drive the search for rules. For instance, the bias may specify rules that contain a given bacterium in its left-hand part and resistance to some antimicrobial in its right-hand part. Data are partitioned into disjoint monthly sets prior to analysis, so that the temporal trend of the confidence of rules discovered in successive months can be computed. In this way, the system uses data mining techniques to identify new, unexpected and interesting temporal patterns for surveillance, being able to identify an increment, for instance, in the number of resistance results to a given antimicrobial for a given species.

§6 Conclusions and Future Works

In this paper we have described the application of data mining techniques in order to automatically discover association rules from microbiological data, and obtain from them alarm rules for data validation. The work here described is part of a wider project, jointly started by the University of Bologna and Dianoema S.p.A., aiming at implementing a knowledge based system (named ESMIS [4][5]) for validating microbiological data.

The knowledge discovery approach we have followed proved very effective

in validating part of ESMIS knowledge based (which is written in accordance with NCCLS guidelines and with microbiology experts), and also in extending it with new validation rules, confirmed by interviewed microbiologists, and peculiar to the considered hospital laboratory.

The contribution of this work is therefore twofold: on one side it provides a way for making human experts aware of new correlations among some antimicrobial test results, previously unnoticed. On the other side, it can be considered an automatic method for validating and possibly extending the knowledge base of an expert system in the microbiological domain.

We are currently applying the knowledge discovery process to a new species (Pseudomonas) and extending ESMIS knowledge base with rules regarding that bacterium.

Future work will also be devoted to investigate whether knowledge discovery (still based on the mining of association rules) might help microbiologists in the early identification of nosocomial infections and of resistance increments as done, for instance, in the DMSS system [18)19)].

Moreover, we plan to complete the development of the filtering system described in section 4.4 in order to be able to use it for the identification of interesting rules among the discovered ones.

§7 Acknowledgements

This work has been partially supported by Dianoema S.p.A. under MURST (Ministero dell'Università e della Ricerca Scientifica e Tecnologica) Project n. 23204/DSPAR/99. The authors would like to thank Marco Manservigi and Giovanni Pizzi of Dianoema S.p.A. Authors are in debt with Massimo Perelli for his help in doing the experiments.

References

1) Dianoema S.p.A., see the web site: `http://www.dianoema.it`, 9 July 2001.

2) NCCLS, National Committee for Clinical Laboratory Standards, see the web site: `http://www.nccls.org`.

3) Mary Jane Ferraro et. al., *Performance Standards for Antimicrobial Susceptibility Testing; Eleventh Informational Supplement*, NCCLS document M100-S11 Vol. 21 No. 1, 2001.

4) E. Lamma, P. Mello, A. Nanetti, G. Poli, F. Riguzzi, S.Storari, "An Expert System for Microbiological Data Validation and Surveillance" to appear in *Proceedings of ISMDA 2001*, Lecture Notes in Computer Science, Springer-Verlag.

5) E. Lamma, L. Maestrami, P. Mello, F. Riguzzi, S. Storari, "Rule-based Programming for Building Expert Systems: a Comparison in the Microbiological Data Validation and Surveillance Domain", to appear in *Electronic Notes in Theoretical Computer Science 59, 4*, Elsevier Science Publishers, 2001.

6) I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, October 1999.

7) Agrawal Rakesh, Srikant Ramakrishnan, "Fast Algorithms for Mining Association Rules", in *Proceedings of the 20th International Conference on Very Large Databases*, 1994.

8) World Health Organization Department of communicable disease surveillance and response, *WHONET 5 - Microbiology Laboratory Database Software*, WHO/CDS/CSR/DRS/99.1

9) M.G.Kahn, S.A.Steib, V.J.Fraser, W.C.Dunghan, *An Expert System for Culture-Based Infection control Surveillance*, Washington University, 1992.

10) Center for Disease Control National Nosocomial Infection Surveillance, CDC NNIS, see the web site:`www.cdc.org`, 9 July 2001.

11) Theratrac, Biomerieux, see the web site: `http://www.theratrac.com`, 9 July 2001.

12) D. Gamberger,N. Lavrac, G. Krstacic, T. Smuc, "Inconsistency tests for patient records in a coronary heart disease database", in *Proceedings of IDAMAP2000*, 2000.

13) Y. Ohsawa "Chance Discoveries for Making Decisions in Complex Real World", New Generation Computing, Vol. 20 No. 2, February 2002.

14) W. Stuhlinger, O. Hogl, H. Stoyan, M. Muller, "Intelligent data mining for medical quality management", in *Proceedings of IDAMAP2000*, 2000.

15) M. Bohanec, M. Rems, S. Slavec, B. Urh, "PTAH: A system for supporting nosocomial infection therapy", in N. Lavrac, E. Keravnou, B. Zupan (eds) *Intelligent Data Analysis in Medicine and Pharmacology*, Kluwer Academic Publishers, 1997.

16) E.Lamma, M.Manservigi, P.Mello, F.Riguzzi, R.Serra, S.Storari, "A System for Monitoring Nosocomial Infections", in *Proceedings of IDAMAP2000*, 2000.

17) Intelligent Miner, see the web site: `http://www.software.ibm.com/data/iminer/fordata`, 9 July 2001.

18) P.A.Hymel, S.E.Brossette "Data Mining-Enhanced Infection Control Surveillance: Sensitivity and Specificity", *Poster at SHEA 2001*, 2001.

19) DMSS, see the web site: `http://www.medmined.com`, 9 July 2001.