A System for Monitoring Nosocomial Infections

E. Lamma[1], M. Manservigi[2], P. Mello[1], F. Riguzzi[3], R. Serra[4], S. Storari[1]

Abstract. In this work, we describe a project, jointly started by DEIS University of Bologna and Dianoema S.p.A., in order to build a system which is able to monitor nosocomial infections. To this purpose, the system computes various statistics that are based on the count of patient infections over a period of time. The precise count of patient infections needs a precise definition of bacterial strains. In order to find bacterial strains, clustering has been applied on the microbiological data collected along two years in an Italian hospital.

1 MICROBIOLOGICAL DATA ANALYSIS

A very important problem that arises in hospitals is the monitoring and detection of nosocomial infections. A hospital-acquired or nosocomial infection is a disease that develops after the admission to the hospital, and is the consequence of a treatment, not necessarily a surgical one, or work by the hospital staff. Usually, a disease is considered a nosocomial infection if it develops 72 hours after the admission to the hospital. In Italy, this problem is very serious: actually almost the 15% of patients admitted to hospitals develop a nosocomial infection. In order to monitor nosocomial infections, the results of microbiological analyses must be carefully collected and analysed.

In Italy, a great number of hospitals manages analysis results by means of a software system named Italab C/S, developed by Dianoema S.p.A. Italab C/S is a Laboratory Information System based on a Client/Server architecture, which manages all the activities of the various analysis laboratories of the hospital. Italab C/S stores all the information concerning patients, the analysis requests, and the analysis results. In particular, for bacterial infections 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);
- for every different bacterium identified, its species and its antibiogram.

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 possibly recorded: R when resistant, I when intermediate, S when sensitive, and null when unknown.

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 bacteria of the same species may have evolved differently and have developed different resistances to antibiotics. Bacteria with similar antibiograms are grouped into "strains".

From these data, infections are now monitored by means of a Italab C/S module called "Epidemiological Observatory" that periodically generates reports on the number of infections detected in the hospital. These reports are configurable and show the number of found infections with respect to other data such as specimen characteristics (material and origin) and patient characteristics (hospital unit, sex, age, etc.). Examples of such reports are:

- for every species, for every material and for every origin, show the number of infections found;
- for every antibiotics and for every species, show the number of found bacteria that are resistant (sensitive or intermediate) to the antibiotics.

In order to count the number of infections, the "Epidemiological Observatory" analyses the data regarding the positive culture results of a particular time period (3 or 6 months). Every identified bacterium compared with the other bacteria found on the same patient in the previous N days (usually N is 30). The bacterium is counted as an infection provided that:

1. its species is different from that of the others;
2. its strain is different from that of bacteria of the same species previously found on the patient.

This is because, in case the strain is the same, the new bacterium is considered as a mutation of the previous one rather than a new infection.

In order to detect when two bacteria belong to the same strain, Italab C/S uses a very simple difference function for computing the percentage of antibiotics in the antibiogram having different values for the two bacteria. If this percentage is below a user defined threshold (usually 30%), then they belong to the same strain.

However, this approach for detecting when two bacteria belong

[1] D.E.I.S., Università di Bologna, Italy, e-mail: pmello@deis.unibo.it
[2] DIANOEMA S.p.A. Via Carracci 93, 40100 Bologna, Italy
[3] Dipartimento di Ingegneria, Università di Ferrara, Italy
e-mail: {mpiccardi, friguzzi}@ing.unife.it
[4] Ospedale Molinette (San Giovanni Battista), Corso Bramante 88/90, 10134 Torino, Italy

to the same strain is quite rough: it is not universally accepted by microbiologist and does not seem to work in all possible situations (different hospitals, different units within a hospital).

In order to improve the accuracy of the system in recognising strain membership, we defined, helped by microbiologists, a new strain membership criteria.

The first step consists in identifying all existing strains in a target hospital. In some cases, strain descriptions can be provided by the microbiologist, in other cases this is not possible and clustering is applied to all the antibiograms found in the past for every bacterium species. Each cluster found is considered as a strain and its description is stored by the system.

A new bacterium is considered as a new infection provided that no bacterium of the same species and strain is found in the same patient in the previous N days. The new bacteria is classified as belonging to a strain by using a membership function that depends on the strain description used.

In order to find bacterial strain, the clustering algorithm is executed on data regarding the positive cultures (only bacterial specie and relative antibiogram) of a large period of time (ex. 12 months) that have been found at the hospital where the system will be installed.

Applying clustering to find bacterial strain is useful also because it can be useful for giving the microbiologist new insights about the hospital population of bacteria and their resistance to antibiotics.

In order to test this approach for strain identification, we have performed a number of prototypical clustering experiments on data from various bacterial species. In this experimental phase we have used Intelligent Miner by IBM [3] for its free availability to academic institutions and its powerful graphical interface. However, clustering in final system will be performed by special purpose code.

2 THE DEMOGRAPHIC CLUSTERING ALGORITHM

The demographic clustering algorithm that is enclosed in Intelligent Miner [1] builds the clusters by comparing each record with all clusters previously created and by assigning the record to the cluster that maximizes a similarity score. New clusters can be created throughout this process.

The similarity score of two records is based on a voting principle, called Condorset [1]. The distance is computed by comparing the values of each field, assigning it a vote and then summing up the votes over all the fields. For categorical attributes, the votes are computed in this way: if the two records have the same value for the attribute, it gets a vote of +1, otherwise it gets a vote of –1. For numerical attributes, a tolerance interval is established and the vote is now continuous and varies from -1 to 1: -1 indicates values far apart, 1 indicates identical values and 0 indicates that the values are separated exactly by the tolerance interval. The overall score is computed as the sum of the score for each attribute.

In order to assign a record to a cluster, its similarity score with all the clusters is computed. To this purpose, the distribution of values of each field for the records in the cluster is calculated and recorded. The similarity between a record and a cluster is then computed by comparing the field values of the record with the value distribution of the cluster. In this way, it is not necessary to compare the record with each record in the cluster.

The algorithm assigns the record to the cluster with the highest similarity score. In case the score is negative for all clusters, then the record is a candidate for forming a new cluster. In this way, the number of clusters does not have to be known in advance but can be found during the computation.

Table 1. Modal values of the resistance for each cluster.

Cluster ®	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Dimension ®	339	1266	65	276	2	2	5	3	2
AMIKACINA	S	R	R	S	S	R	S	S	R
AMOXI_A_ CLAVULANIC	S	R	S	R	S	R	S	R	R
AMOXICILLINA	R	R	R	R	R	R	R	R	R
CEFAZOLINA	S	R	S	R	S	R	S	R	S
CEFOTAXIME	S	R	S	R	S	R	S	R	R
CEFUROXIME_ PARENTE	S	R	S	R	S	R	S	R	S
CIPROFLOXACINA	S	R	R	S	R	R	R	S	I
CLINDAMICINA	S	R	R	S	S	S	R	S	S
COTRIMOXAZOLO	S	R	R	S	R	S	S	R	R
DOXICICLINA	S	S	S	S	S	S	S	R	S
ERITROMICINA	S	R	R	S	R	R	R	S	R
GENTAMICINA	S	R	R	S	I	R	S	I	R
IMIPENEM	S	R	S	R	S	R	S	R	S
MEZLOCILLINA	R	R	R	R	-	R	S	-	-
NETILMICINA	S	R	R	S	S	R	S	S	R
OFLOXACINA	S	R	R	S	R	R	R	S	R
OXACILLINA	S	R	S	R	S	S	S	R	S
PEFLOXACINA	S	R	R	S	R	R	R	S	R
PENICILLINA_G	R	R	R	R	R	R	R	R	R
RIFAMPICINA	S	S	S	S	R	R	R	S	S
TEICOPLANINA	S	S	S	S	S	S	S	S	S
TIAMFENICOLO	S	S	S	S	S	S	S	R	S
VANCOMICINA	S	S	S	S	S	S	S	S	S
Resistance level	14,7	69,4	44,8	44,2	20,8	50,9	32,7	51,4	47,9

This process is repeated a fixed number of times ("phases") and clusters are updated until either the maximum number of phases is reached or the maximum number of clusters is achieved or the clusters centres do not change significantly as measured by a user-determined margin.

3 RESULTS

We have considered all the bacteria belonging to the species Staphilococcus Epidermidis. The dataset contains 1961 records having the attributes described in section 1. They have been collected from the 5th of March 1997 to the 20th of November 1999 at Le Molinette Hospital in Turin, Italy.

As in the PTAH system [2], an additional feature was computed for each record: the level of resistance, that represents the percentage of antibiotics for which the bacterium was resistant over the total number of antibiotics whose resistance was known (R, S, I).

In this experiment, we have set the maximum number of phases to 3 and we have found 9 clusters with a global Condorset value of 0.843. The clustering has been performed by considering only the record fields relative to antibiotics resistance. Table 1 shows the modal values of antibiotics reaction in the 9 clusters. The second row shows the number of elements of the cluster and the last the average resistance level of the cluster.

Cluster 1 is the biggest and is the one with the highest level of resistance (average of 69.4 %).

Figure 1 shows the resistance level to antibiotics of cluster 1: in each pie-chart, the internal pie is referred to the cluster, while the external ring is referred to the overall dataset. From figure 1 we can observe that in cluster 1 the percentage of resistant bacteria is higher for all antibiotics with respect to the complete dataset except for Doxiciclina for which the percentage of sensitive bacteria is higher. Cluster 0 has the same behaviour with R substituted for S: Doxiciclina is the only antibiotics for which the percentage of resistant bacteria is higher.

Clusters 3 and 2 are characterised by values of the resistance level that are intermediate between those of cluster 1 and 0. Clusters 4, 5, 6, 7 and 8 contain few elements and this means that some antibiograms are significantly different from all the others.

On the basis of these results, some comments can be made. We expected that the majority of bacteria from the same species had similar behaviour and that, more rarely, we could find "abnormal" bacteria that had become more resistant. On the contrary, by clustering the Staphylococcus Epidermidis bacteria, we have found that the majority of bacteria is highly resistant and that rarer cases are characterised by a higher sensitivity to antibiotics. This is probably due to the nature of this bacterium. In fact, another clustering experiment performed over *Escherichia Coli* bacteria has shown that bigger clusters have a lower resistance level and smaller cluster have a higher resistance.

Clustering of the antibiograms was performed as well in the PTAH system [1]. In PTAH, the clusters are hierarchically organised: low level clusters are grouped into higher level cluster and so on, up to the root cluster that contains all the data. The hierarchy enables the user to study the clusters at different levels of granularity. In this way it is possible to discover the different types of resistance vectors and to evaluate their frequency.

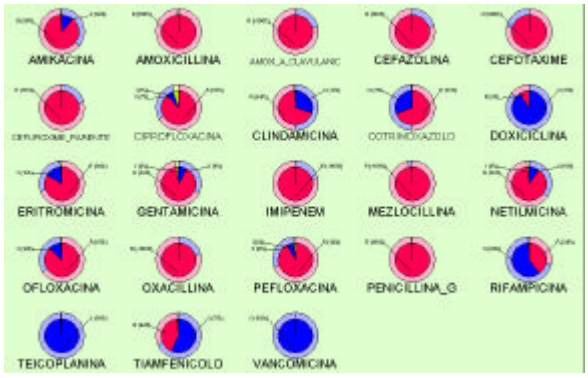

Figure 1. resistance to antibiotics in cluster 1.

We owe to PTAH a number of inspiring ideas, first of all the introduction of the resistance level variable for a bacterium that is very useful for providing an indication of the dangerousness of bacteria, and also the clustering of bacteria. However, we do not use hierarchical clustering as PTAH does: this is due to the fact that the results here presented are obtained from a first study. In the future we plan to adopt as well a hierarchical clustering algorithm because we think that the results will probably be easier to be interpreted by a medical doctor.

ACKNOWLEDGEMENTS

We are grateful to Dr. Furlini (S.Orsola Malpighi Hospital, Bologna) and Dr. Andollina (Officine Ortopediche Rizzoli, Bologna) for helpful discussions. This work has been partially supported by DIANOEMA S.p.A., Bologna, Italy and by the MURST project ``Intelligent Agents: Interaction and Knowledge Acquisition".

4 BIBLIOGRAPHY

[1] Cabena, Hadjinian, Stadler, Verhees, Zanasi, "Discovering Data Mining – from concept to implementation", Prentice Hall – IBM

[2] 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

[3] Intelligent Miner, http://www.software.ibm.com/data/iminer/fordata