A System for Monitoring Nosocomial Infections*

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

[1] D.E.I.S., Università di Bologna,
Viale Risorgimento 2, 40136 Bologna, Italy
{elamma,pmello}@deis.unibo.it, sstorari@iol.it
[2] DIANOEMA S.p.A. Via Carracci 93, 40100 Bologna, Italy
mmanservig@dianoema.it
[3] Dipartimento di Ingegneria, Università di Ferrara,
Via Saragat 1, 44100 Ferrara, Italy
friguzzi@ing.unife.it
[4]Ospedale Molinette (San Giovanni Battista),
Corso Bramante 88/90, 10134 Torino, Italy

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 that is found by applying clustering to data on past infections. Moreover, the system is able to identify critical situations for a single patient (e.g., unexpected antibiotic resistance of a bacterium) or for hospital units (e.g., contagion events) and alarm the microbiologist.

1 Introduction

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. A community infection, instead, is an infection acquired by the patient before the admission to the hospital. Usually, a disease is considered a nosocomial infection if its symptoms appear more than 48 hours after the admission to the hospital.

Nosocomial infections are much more dangerous than community infections because they are caused by bacteria that are much more resistant to antibiotics. Usually nosocomial infections are resistant to more than one antibiotic, while community infections are resistant to a single or very few antibiotics. As a consequence, the cure of a community infection normally does not pose problems while it may prove difficult to cure nosocomial ones. In Italy, this problem is very serious: actually almost 15% of the patients admitted to hospitals develop a

* Appears in the proceedings of the International Symposium on Medical Data Analysis, ISMDA 2000

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. Analysis results are collected from automatic analyzers connected to the system or from manual input. The results are then checked for validity and stored in a relational database. In the past, we have designed and implemented an expert system for the validation of clinical analysis [4].

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. In particular, on the basis of the microbiological analysis results stored in the Italab C/S relational database, the system must:

1. provide statistics about the number of nosocomial infections in the various areas of the hospital;
2. identify critical situations for a single patient (e.g., unexpected antibiotic resistance of a bacterium) or for hospital units (e.g., contagion events) and alarm the microbiologist.

Statistics about infection frequencies are useful from two points of view: on one hand they can be used in order to monitor the diffusion of nosocomial infections over time in the hospital, on the other hand they are a valid help for clinicians to perform a first diagnosis. In this task it is very important to correctly count of the number of infections in each area of the hospital over a period of time. In order to avoid counting a mutation of a bacterium as a new infection, data mining techniques (and clustering, in particular) have been applied in order to find group of similar bacteria (strains). In sections 2, 3 and 4, we briefly report about the first results we have obtained by applying clustering on the microbiological data collected along two years in an Italian hospital.

In order to address the latter task, instead, we will adopt a knowledge-base approach and build a surveillance expert that is able to identify critical situations and correspondingly generate alarms (see section 5).

2 Microbiological Data Analysis

Italab C/S stores all the information concerning patients, the analysis requests, and the analysis results. In particular, data for bacterial infections 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 (antibiotic, 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 most frequent infections detected in the hospital. In particular, for each area of the hospital, the system must show the n (where n is configurable by the user, usually 10) bacterium species most frequently infecting the area. For each species, the system must show:

- the name of the species,
- the percentage of isolated bacteria belonging to that species over the total number of isolated bacteria,
- the resistance of the bacteria to antibiotics, i.e., for each antibiotic, the percentage of found bacteria that are resistant, intermediate and sensitive.

This information is useful because it allows to monitor the behavior of nosocomial infections over time. Moreover, it allows the clinician to give the patient an "empirical therapy": whenever a patient shows the symptoms of an infections, the clinician is able to give him immediately an antibiotic cure by considering the symptoms and the frequency of the candidate infections in the area of the hospital, without waiting for microbiological analysis results that usually take two to three days to come back from the laboratory. Then, when the analysis results become available, he will be able to provide the patient a more accurate cure in case the empirical therapy had not the desired effect.

In order to count the number of bacteria for each species, the "Epidemiological Observatory" analyses the data regarding the positive culture results of a particular time period (3 or 6 months). Every identified bacterium is compared with the other bacteria found on the same patient in the previous N days (usually N is 10). The bacterium is counted provided that its strain is different from that of bacteria of the same species found on the patient in the previous N days. 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 similarity function that computes 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 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 give 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.

3 The Demographic Clustering Algorithm

The demographic clustering algorithm that is enclosed in Intelligent Miner [2] 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 [2]. 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.

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.

4 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.

Table 1. Resistance to antibiotics of Staphilococcus Epidermidis on all the dataset

Antibiotics	*S*	*R*	*I*	*null*
AMIKACINA	35,1%	61,4%	0,9%	2,6%
AMOXI/A.CLAVULANICO	20,2%	74,2%	-	5,6%
AMOXICILLINA	0,9%	98,9%	-	0,2%
CEFAZOLINA	20,7%	79,0%	0,1%	0,2%
CEFOTAXIME	16,5%	60,1%	0,3%	23,1%
CEFUROXIME PARENTERALE	20,7%	79%	-	0,3%
CIPROFLOXACINA	29,6%	66,2%	3,9%	0,2%
CLINDAMICINA	49,7%	48,8%	1%	0,5%
COTRIMOXAZOLO	48,4%	50,9%	0,5%	0,2%
DOXICICLINA	85,7%	13,8%	0,4%	0,1%
ERITROMICINA	29,1%	69,1%	1,5%	0,2%
GENTAMICINA	32,1%	65%	2,8%	0,1%
GENTAMICINA_HL	-	-	-	100%
IMIPENEM	20,8%	78,8%	0,2%	0,2%
MEZLOCILLINA	0,7%	22,1%	-	77,2%
NETILMICINA	34,1%	61,7%	1,7%	2,5%
NITROFURANTOINA	-	0,1%	-	99,9%
NORFLOXACINA	-	-	-	100%
OFLOXACINA	33%	65,6%	1,1%	0,3%
OXACILLINA	21,1%	78,8%	-	0,1%
PEFLOXACINA	28,6%	69,3%	1,9%	0,2%
PENICILLINA_G	1%	98,8%	-	0,2%
PIPERACILLINA	0,1%	-	-	99,9%
RIFAMPICINA	54,6%	21,9%	0,5%	22,9%
STREPTOMICINA_HL	0,1%	-	-	99,9%
TEICOPLANINA	99,3%	0,2%	0,4%	0,1%
TIAMFENICOLO	86,7%	30,2%	0,9%	0,2%
VANCOMICINA	99,8%	-	-	0,2

We first report the distribution of attribute values relative to the complete dataset. In table 1 the percentage of R, S, I and null values are reported for each of the 23 antibiotics that have been tested.

As in the PTAH system [1], 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). Figure 1 shows the distribution of the resistance level in the dataset: the X axis reports the level of resistance while the Y axis reports the percentage of records in the dataset having that resistance level.

An experiment was performed where the number of phases was set to 3. In this case, 9 clusters were found with a global Condorset value of 0.843. The clustering has been performed by considering only the record fields relative to antibiotics resistance.

Fig. 1. Distribution of bacteria resistance level (GRADORESISTENZA in Italian)

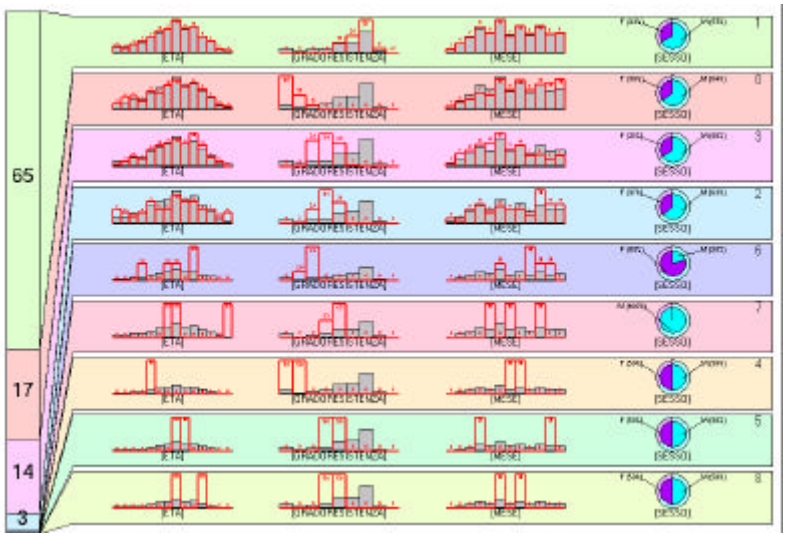

Fig. 2. Value distribution in the clusters

Figure 2 shows a graphic representation of all the clusters, where the number on the left represents the percentage of records in each cluster and the number on the right is the cluster identifier. Within each cluster, the distribution of values for a variable is

depicted by a sub-chart, either a histogram for numeric variables or a pie-chart in the case of categorical variables. The sub-charts have an overlay structure that enables them to depict the distribution of the associated variable within both the individual cluster and the overall population for comparison purposes. As regards the histograms, the filled histogram is the distribution of values over all the data, while the empty histogram is the distribution of values in the cluster. As regards the pie-charts, the internal pie is referred to the cluster, while the external ring is referred to the overall dataset.

The variables whose distribution is depicted in Figure 2 are: ETÀ (age), GRADORESISTENZA (resistance level), MESE (month) and SESSO (sex).

Table 2 shows the modal values of antibiotic reactions in the 9 clusters. The second row shows the number of elements of each 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 3 shows the resistance level to antibiotics in cluster 1. From this figure 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.

Table 2. 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

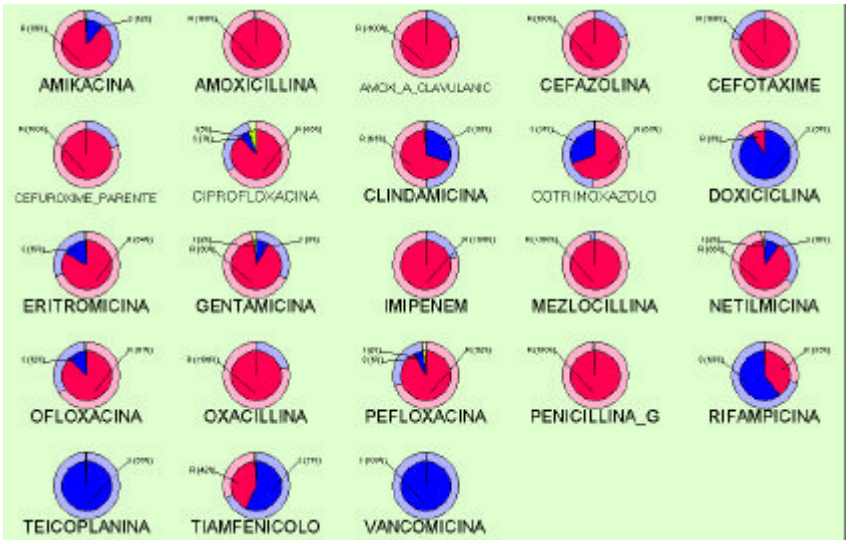

Fig. 3. Resistance to antibiotics in cluster 1

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.

5 Surveillance Expert System

Given a newly isolated bacterium, the system must provide the following alarms:
1. unexpected resistance to an antibiotic,
2. contagion by a multi-resistant bacterium,
3. ineffective therapy.

The first alarm is given if the new isolated bacterium is resistant to an antibiotic to which it is normally sensitive. This is important in order to timely alert the microbiologist about the appearance of an unexpected resistance. To this purpose, the system should store, for each species, the list of antibiotics to be monitored.

The second alarm is given if a bacterium belonging to the same strain was previously found on another patient in the same hospital area within a given period X

(usually 30 days). This case indicates that there has been a contagion among the two patients and therefore it must be signaled so that future contagia are avoided. The alarm must be given only if the isolated bacterium is resistant to more than two antibiotics because otherwise it is considered not dangerous.

In the case of the third alarm, the system looks if another bacterium was isolated for the same patient in a previous period Y (usually 30 days, more for some chronic infections). In case a previous bacterium was found, its antibiogram is compared with the one of the current bacterium and an alarm is generated if the number of resistances is increased. The system should show the species and antibiograms of the two bacteria. Moreover, the system should compare the species of the two bacteria and produce one of the following two alarms:

1. if the species are different, then the system should alert the user that the therapy was ineffective because the patient was infected by two different bacteria but it was cured for only one,
2. if the species are the same, then the system should indicate that the therapy has failed.

6 Related Works

PTAH [1] is a decision support system that was designed for helping medical doctors in the prescription of antibiotics for the cure of nosocomial infections. It has been evaluated in the General Hospital Jesenice in Slovenia on data collected since 1994. PTAH can perform four type of analyses:

- resistance level over time,
- hierarchical clustering of antibiograms,
- similarity of antibiograms,
- effectiveness of antibiotics over time

Each analysis is performed over data regarding a single bacterium species at a time.

The resistance level for a bacterium is plotted over time in order to identify occasional and partly periodic appearances of highly resistant bacteria. This allow medical doctors to identify possible inefficiencies in the antibiotic therapy that have caused the appearance of highly resistant bacteria.

Besides identifying time trends of resistance level, PTAH performs clustering of antibiograms. The clusters are hierarchically organized: 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 level of granularity. In this way it is possible to discover the different types of resistance vectors and to evaluate their frequency.

The third type of analysis is aimed at showing the similarities of antibiograms in order to identify possible infection diffusion in the hospital. To this purpose, a two dimensional graph is plot where each horizontal line corresponds to a resistance vector and the X axis is associated to time: each point in the graph represents an antibiogram at a given moment in time. Antibiograms are ordered on the Y axis in a de-crescent way according to their resistance level. Two points are connected by a line if the antibiograms differ for at most one element and have a time difference

smaller than a given threshold. In this way, it is possible to identify moments in time where the same bacterium was found with similar characteristics at several patients thus leading to the identification of a possible epidemic: groups of point vertically aligned indicate the presence of a set of similar bacteria at the same time and the higher they are in the graph the more dangerous they are.

The antibiotic effectiveness is analyzed by plotting over time the cumulative or moving average percentage of antibiograms that are resistant to a given antibiotic. In this way it is possible to identify when the effectiveness of an antibiotic becomes too low so that it can be dismissed for medical and financial reasons.

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. Differently from PTAH, we do not cluster only resistance vectors but also data regarding the patient: this data turned out to be important since the most significant variables for the clustering are age, resistance level, month and sex. Moreover, 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.

7 Conclusions

This paper describes a project, jointly started by DEIS University of Bologna and Dianoema S.p.A., in order to build a system for monitoring nosocomial infections. This function is really important because of the high risks and costs associated to nosocomial infections.

Computing the frequency of infections in the various areas of the hospital is really important for monitoring nosocomial infections. The behavior over time of this frequency may highlight possible hygienic problems. Moreover, it can be used for an early diagnosis and therapy.

In order to correctly compute infection frequencies, it is important not to count twice bacteria belonging to the same strain. To this purpose, we have adopted clustering for performing the identification of the bacterial strains. We report on a first experiment in using clustering on data regarding the Staphilococcus Epidermidis bacteria.

Another important aspect of infection monitoring consists in generating alarms regarding newly identified bacteria. Alarms are raised in case a unexpectedly resistant bacterium is found, in case a contagion among patients of a unit is detected or in case the therapy is found to be ineffective.

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".

References

1. 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
2. Cabena, Hadjinian, Stadler, Verhees, Zanasi, "Discovering Data Mining – from concept to implementation", Prentice Hall – IBM
3. Intelligent Miner, http://www.software.ibm.com/ data/iminer/fordata
4. Evelina Lamma, Paola Mello, Sara Monesi, Sergio Storari: "An expert system approach for clinical analysis result validation", Proceedings of International conference on Artificial Intelligence, IC-AI'2000, 26-29 giugno 2000