MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

From Data Collection to Knowledge Data Discovery: A Medical Application of Data Mining

Alain Duhamel[a], Monique Picavet[b], Patrick Devos[a], Régis Beuscart[a]

[a] CERIM - Faculté de Médecine - 1, Place de Verdun – 59045 Lille Cedex - France
[b] University of Lille 1, Cité Scientifique- Bât M3 - 59655 Villeneuve d'Ascq Cedex, France

Abstract

Prison inmates are exposed to a variety of major risk factors (psychiatric disorders, suicide attempts, illicit drug use). From 1986 to 1996, the USA prison population more than doubled while in France, it increased from 35655 in 1980 to 51623 in 1995. In spite of these findings, very little information concerning the inmates population is available. At the present time, there is a desire to adopt a policy based on the prevention of recidivism, on adequate release planning and referrals to community-based services. The aim of the RAPPEL project was to build an information system for assessing the social and health status of prison inmates. The pilot project was set up at the prison of Loos and allowed the collection and analysis of nearly 15000 records. The aim of this paper is to present the extension of the project consisting in developing a regional network grouping 11 jails. Information locally available will serve as the basis for the information system of regional jails. Data mining techniques will provide solutions for the extraction of new information. Three data mining tools were experimented : association rules, classification trees and clustering. Further extension consists in a distributed approach allowing direct access to the information system by WEB tools.

Keywords:

Information system, DataWarehouse, KDD, Datamining, prisonners, public health.

Introduction

Prison inmates are exposed to a variety of major risk factors including suicide attempts, illicit drug use, excessive drinking or psychiatric disorders. From 1986 to 1996, the USA prison population more than doubled [1] while in France, it increased from 35655 in 1980 to 43913 in 1990 and to 51623 in 1995 [2]. More importantly, an average of 82820 persons per year (which represents approximately 1.5/1000 of the French population) was jailed between 1989 to 1994 with a mean length of stay of 6.9 months. In general, the majority of prisoners are released after short periods of stay within the prison environment. This flow of detainees going in and out of jails means that this population is characterized by a large interaction with the community. In recent times, there has been an evolution in terms of the profile of the population arriving in jails as illustrated by a rise in juvenile delinquency as well as a modification of the types of crimes committed, with in particular an increase in illicit drug related offences.

In spite of these findings, very little information concerning the inmates population is available. The primary function of the jail administration is to process the flow of arrivals into the prison. It has an obligation by law to provide care, to assure the continuity of care and to prepare inmates for reentry into the society. At the present time, there is a desire to adopt a policy based on the prevention of recidivism, on adequate release planning and referrals to community-based services. It was against this background that the "Recherche Action sur la Population Pénale Ecrouée à Loos" (RAPPEL) project was established in 1996. The aim of this project was to build an information system for assessing the social and health status of prison inmates. The pilot project set up at the prison of Loos now runs successfully [3]. Using the RAPPEL information system, the data of 14785 arrivals in the prison over a period of seven years (1989 to 1995) were analyzed [4].

Extending the project consists in applying the approach used in Loos to other jails of the Nord Pas de Calais region for the data collection. Data analysis is performed on the centralized site (Loos). Relevant information will be restituted to all sites using WEB tools (DataWeb). Information locally available will serve as the basis for the information system of regional jails. We underline the distributed aspect, the heterogeneity of the sources and the enormous increase in the quantity of data where data mining techniques provide solutions for the extraction of new information. Data mining concerns the discovery stage of the Knowledge Discovery in Databases (KDD) process. KDD can be defined as the process of non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data [5]. This approach is well fitted to the analysis of a vast amount of data where traditional statistical analysis based on the "hypothesis and test" paradigm becomes a time consuming process. Data mining has demonstrated its efficiency in some application field like marketing, and using this tool in medical information system is attractive.

The incorporation of regional and national data consisting in epidemiological, socioeconomic and penitentiary data makes it possible to work out and evaluate a regionally adapted policy on prison management.

In this paper we present the running pilot project and the approach retained for designing the regional information system for the management of prison inmates with in a network grouping the 11 jails of the region. Our approach relies on data mining for efficient knowledge discovery.

The RAPPEL project is encouraged by the regional administration of Nord Pas de Calais and several ministries in France (the ministries of Health, of Justice and of internal affairs).

Materials and methods

The local penitentiary database

The database was initially developed in the Loos prison which houses about 1150 inmates (1060 males and 90 females) and has an average number of 2200 arrivals a year.

Information concerning the inmate population is of great complexity, heterogeneous (socioeconomic data, medical data, penitentiary data) and confidential. A lot of this information corresponds to self-reported data and could be skewed. Our aim was to identify the most informative variables for building an information system for assessing the health status of prison inmates.

Firstly, a questionnaire gathering information regarding demographic data, level of education and professional status, arrest charge, social and family history, lifestyle, medical and psychiatric history, suicidal ideation, illicit drug use or alcohol abuse was built.

From 1988 to 1995, all new prisoners were interviewed on their arrival into the prison using this standard questionnaire. All questionnaires were anonymously recorded by a company specializing in data collection.

Statistical analysis was then performed using the SAS software [6] to determine the characteristics of the prison population. A database was designed using the questionnaire and the results of the statistical analysis. The data were finally transferred into the RAPPEL database. From 1998, the questionnaires of new prisoners are recorded using the database interface.

Regional network

The first extension of the project consisted in developing the regional network grouping 11 jails of the region. Figure 1 shows the geographical distribution of this network. The information concerning each new arrival in any of the 11 prisons is recorded in the local database.

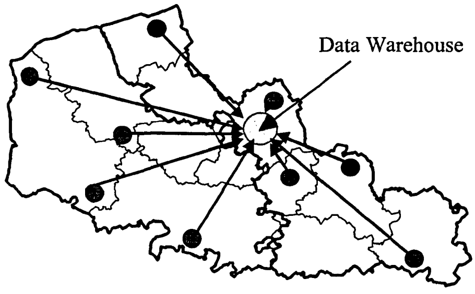

Figure 1 : the regional network. Centralization of regional prisoners data in the Data Warehouse of Loos.

The purpose of data warehousing is basically to ensure that users will be able to find correct and directly comprehensive information when it is needed. The major proposal is to build a Data Warehouse (DW) separated from the transaction database (the RAPPEL database). The DW was designed using the dimensional modeling methodology. In this structure, a set of indicators can be described according a set of dimensions such as geographical or time level. The dimensions and the indicators of the DW were identified using the results of statistical analysis performed on the database. An executive information system was designed and connected to the DW. This system makes it possible to obtain a comprehensive summary of the jail population in real time by using standard reports and graphics (curves, charts, ...), drill-down and roll-up on hierarchical dimensions, and analysis of time trends of indicators.

In the current state of the system, data acquisition is realized locally in each prison. These standardized local data are transferred and then processed in the central site. The DW is designed from this centralized inmate database. Regional and national data sources (epidemiological and socioeconomic data, penitentiary data) have been added to the information system in order to perform comparisons (Figure 2).

The extension of the project aims to design a reporting system allowing each site to obtain standard reports on its own data and direct access to specific views of the centralized DW by WEB in real time. This requires the introduction of a component interfacing a WEB browser and the DW.

Statistical analysis

The database were first analyzed using classical statistical procedures : univariate descriptive analysis, crosstabulation tables and tests for comparisons between inmate characteristics and the general population, hierarchical classification and factorial analysis to identify different subgroups among the inmates. An analysis of time trends of indicators was performed by means of the Cochran-Armitage test. Theses analyses allowed to assess the major characteristics of arrivals in jails [4].

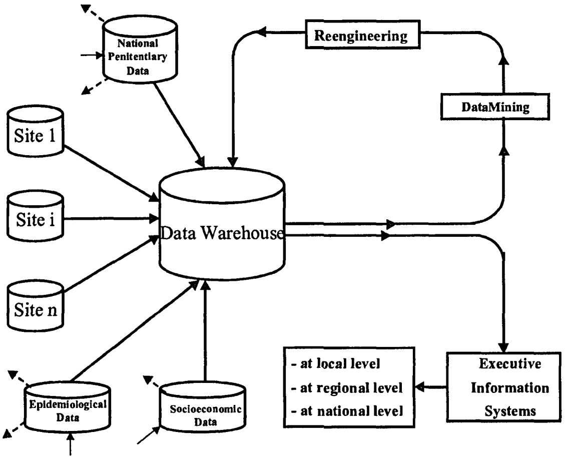

Figure 2 : the regional information system for inmates

Data Mining

The usual statistical methods were well fitted to the structure of the local penitentiary data : standardized data, small number of parameters and observations. Additional data sources (Figure 2) complicate the extraction of useful information for decision support by these methods : increase in the number of dimensions and indicators, multiplication of the tests to be performed, requirement of large numbers of specific models, ... In these situations, methods for efficient knowledge discovery in databases referred to data mining become essential [7]. Data mining techniques correspond to the step of knowledge extraction in the whole KDD process which includes the following steps : data collection, data preparation (removing errors and inconsistencies), search for pattern (data mining), validation and interpretation of results, and visualization.

Data mining refers to a set of techniques for an automatic analyze of data in order to identify relationships among variables, to identify clusters, to determine classification rules or to build specific models [7]. For the project, we choose to use the following data mining tools : association rules, classification trees and clustering using the k-means algorithm. These methods are presented in the next section.

Association rules

The aim of this technique is the discovery of associations among variables in the database. These associations are often referred to correlations between variables. In the prison population, several major problems have been identified (illicit drug use, psychiatric disorders, suicide attempts, ...). These disorders rarely occur in an isolated pattern and it would be interesting to identify the different associations existing between them. If this mining tool shows proves effective, we will use it for more general problems such as the identification of unexpected and interesting trends.

An association rule is a rule of the form "A and B → C" where A, B and C are values of variables of the database. The algorithm allows for the finding of all associations in the database which satisfy the user specified minimum support and minimum confidence constraints [8]. A rule is characterized by two parameters : the support s and the confidence c, defined by $s = P(A \cap B \cap C)$ and $c = P(C/A \cap B)$ (P stands for Probability).

Association rule mining will be performed using the CBA software [9] and the results will be validated by means of statistical tests and expert opinion.

Classification trees

An important problem in the management of prison populations is to identify factors related to public health problems such as illicit drug use, excessive drinking or psychiatric disorders. Our aim is to identify profiles at risk related to these problems in order to provide comprehensive decision rules. For example : {"drug abuse" and "psychiatric disorder" then probability{"suicidal ideation"}= 60%. In this situation, there is a variable which needs be explained (dependent variable). We therefore opt for the use of a supervised classification method. Classification tree is a non-linear discrimination method, which uses a set of independent variables to split a sample into progressively smaller subgroups. The procedure is iterative : at each branch in the tree, it selects the independent variable that has the strongest association with the dependent variable according to a specific criterion. The Chaid method, based on the chi-square test of association, was used in this study. In the Chaid algorithm, the "splits" computation stops when no more variables having a significant association with the dependent variable can be identified. This Chaid method naturally deals with interactions between the independent variables that are directly available from the examination of the tree. The final nodes identify subgroups defined by different sets of independent variables. In the prison population context, these subgroups correspond to inmates with high or, on the contrary, weak risk with regard to the dependent variable. Classification tree analysis will be performed using the Sipina software [10].

Clustering

The prison population is a heterogeneous one. The identification of clusters of individuals, i.e subgroups of inmates requiring the same type or level of care would be of great interest to determine an adequate policy. Our aim will be to determine which groups of inmates must be helped and why. It then becomes possible to establish an order of priority, to adapt social and medical care and to use resources optimally.

For this purpose, cluster analysis using the k-means clustering procedure will be performed using the SAS software [6]. Cluster analysis is a multivariate procedure allowing to classify the patients in different groups or clusters relating to different profiles. These clusters are not defined a priori and are such that individuals in a given cluster are close to each other in the sense of the Euclidean distances and individuals in different clusters tend to be dissimilar. Because the Euclidean distance is used, the variables must be numerical. If it is not the case, factorial analysis such as multiple correspondence analysis can be performed before clustering to transform the data into numerical scores.

Results

The information system provides the decision-makers with synthetic and strategic information for the improvement of penitentiary policies. Data mining tools allows a rapid and exhaustive exploration of the database. In this section, we present few meaningful results.

Classical statistical analysis (summary of results)

A total of 14785 questionnaires were analyzed. Of the study population, 56% had had no professional qualification and 62% was unemployed.Thirty-one percent of prisoners had psychiatric disorders and almost half of them (43%) had already been hospitalized in psychiatric unit. Nearly 16% of detainees admitted having made at least one suicide attempt. More than fifth of prisoners were alcohol dependent while nearly 40% used illicit drug. Amongst the arrivals reporting psychiatric disorders or illicit drug abuse or both (59%), 70% had not received health assistance and 53% had neither received educational support nor health assistance.

Data mining (preliminary results)

Association rules mining was performed on a subset of 12 variables including arrest charges, health and educative assistance, psychiatric disorders, alcohol and illicit drug abuse. Support and confidence were fixed to 10% and 40% respectively. Using CBA software, 2763 rules were identified. Because data were strongly correlated, numerous rules were useless. It was then necessary to detect the relevant rules. Three of them are presented below :

R1 :{unemployment} AND {live alone} THEN {heroin use}; s=38%, c = 44%, i.e "Among the inmate which are unemployed and live alone, 44% use heroin".

R2 : {suicide attempt} AND {psychiatric disorder} THEN {no health assistance}; s=10%, c=60%;

R3 : {drug offence} AND {live alone} THEN {educative assistance}; s=12%, c=85%;

Statistical analysis had shown that the rate of Suicide Attempt (SA) was 15.6% (n=2310). Obviously this rate was not homogeneous according the health and/or socioeconomic status of inmates. A classification tree analysis using the Chaid method was performed to analyze relationships between SA and a set of risk factors (Figure 3).

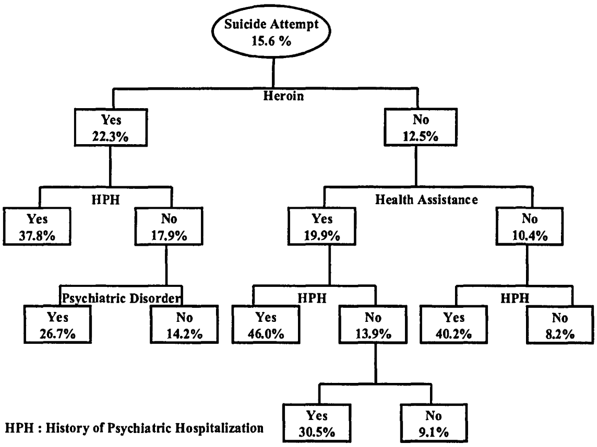

Figure 3 : Classification tree to explain suicide attempt

Several interesting patterns can be deduced from the tree. For example, among the inmates without heroin addiction, but with health assistance and history of psychiatric hospitalization, the rate of suicide attempt was 46% vs 15.6% in the database.

Discussion and conclusion

In this paper, we have presented the RAPPEL Information System (IS) designed for the management of the inmates population. This IS presently runs in Loos. Since 1998, an average of 2200 arrivals a year are recorded in the system. Using the RAPPEL information system, statistical analysis performed on nearly 15000 arrivals have provided a state of characteristics of the prison population and demonstrated that there is a considerable need of multidisciplinary assistance in this population [4].

The aim of an information system for inmates is to provide decision makers with information on the health of their detainees and their social and economical status in order to orient the choice of efficient policies, to allow for subsequent analysis of such policies and their adjustment on the basis of trends observed. On the basis of the results obtained in Loos, an extension of the project to the entire region was decided and approved by the department of justice as well as the regional administration. It is this first extension to 11 jails that has been reported here. Presently, the characteristics of each arrival into any one of these 11 prisons are recorded locally using the RAPPEL database. These data will regularly feed the centralized Data Warehouse (DW).

Data analysis of the regional Information System will require advanced tools, adapted for the analysis of huge quantities of data. Three data mining techniques have been selected for the following reasons :

(unsupervised classification) and identification of complex correlations (association rules)

(2) they are suitable in the context of medical data : mixture of different variable types, missing data

(3) the results can be explained by a logical representation understandable by physicians and are close to diagnosis reasoning

(4) computing tools are available to support these techniques.

Other data mining techniques such as logistic regression or genetic algorithm could be also used for our study.

Several limitations must be indicated concerning the data mining tools used. Firstly, about the association rules : the strength of this technique is that it can efficiently discover the complete set of associations that exists in data. However, this strength comes with a major drawback : the number of discovered association rules can be huge [11]. Choosing a high level of s can result in the elimination of significant rules. An important research actually concerns pruning, i.e. a process which could remove insignificant associations [11]. Major improvements of algorithms of associations rules discovery are expected in the future.

Secondly, concerning classification trees, the major drawback is the effect of overfitting which results in very large tree. In case of overfit, the tree doesn't only reflect the characteristics of the whole population but also the particularities, the noise of the training set. To reduce this effect, pruning algorithms have been developed [12]. In tree pruning, the unreliable branches of the tree are eliminated. The subgroups corresponding to the final nodes are then more robust.

Third, an important limitation for using k-means algorithm for clustering is the need to determine a priori the number of clusters. Different methods have been proposed to determine the "optimal" number of clusters but none of them is really accurate.

A lot of research is presently concentrated on the improvement of data mining tools. One option consists in using several methods competitively for solving the same problem and then chooses the "majority" result (boosting methods)[13]. In any case, results must be validated by medical experts [24].

(1) they are adapted to our intended objectives of improving the management of the prison population : identification of subgroups linked with pathologies (supervised classification), identification of homogeneous subgroups

References

[1] Veysey BM, Steadman HJ, Morrissey JP, Johnsen M. In search of the missing linkages : continuity of care in U.S. jails. *Behav.Sci.Law* 1997;15:383-97.

[2] Guillonneau M, Kensey A. La santé en milieu carcéral. Eléments d'analyse démographique. *Revue Française des affaires sociales* 1997;1:41-59.

[3] Duhamel A, Archer E, Devos P, Nuttrens MC and Beuscart R. A prototype of an information system for assessing the health status of prison inmates. *Stud Health Technol Inform* 1999;68:37-41.

[4] Duhamel A, Renard JM, Nuttens MC. Social and health status of arrivals in a French prison: a consecutive case study from 1989 to 1995. *Rev. Epidém. et Santé Publ.* In Press.

[5] Adriaans P and Zantinge. Data Mining. Edinburgh : Addison Wesley, 1996.

[6] The SAS system. SAS Institute INC., Cary, NC 27513.

[7] Lavrac N. Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine*, 1999;16:3-23.

[8] Agrawal R, Imielinski T., Swami A, Mining associations rules betweens sets of items in large databases. SIGMOD- 1993;207-216.

[9] Liu B, Hsu W., Ma Y,, CBA Version 2.0.1999, National University of Singapor. Site : www.comp.nus.edu.sg/dm2/

[10] Zhighed DA and Rakotomalala R. Graphes d'induction : Apprentissage et Data Mining. Paris : Hermès, 2000.

[11] Liu B, Hsu W., Ma Y,, Pruning and Summarizing the Discovered Associations. In ACM SIGKDD 1999; San Diego, CA, USA.

[12] Breiman L, Friedman JH, Ohlsen RA and Stone CJ. Classification and Regression trees. Belmont (CA) : Wadsworth, 1984

[13] Breiman L. Combining predictors, Technical report, Statistics Department, Berkeley, 1998.

Address for correspondence:

Alain Duhamel, CERIM, Faculté de Médecine, 1 Place de Verdun, 59045 LILLE Cedex, FRANCE. E-mail: alain.duhamel@univ-lille2.fr.