Advances in the Decision Making for Treatments of Chronic Patients Using Fuzzy Logic and Data Mining Techniques

M. Domínguez[1], J. Aroba[2], J. G. Enríquez[1], I. Ramos[1], J. M. Lucena-Soto[3] and M. J. Escalona[1]
[1]*Web Engineering and Early Testing (IWT2) Research Group, University of Seville, Av. Reina Mercedes s/n, Seville, Spain*
[2]*Departamento de Tecnologías de la Información, Escuela Técnica Superior de Ingeniería, University of Huelva, Carretera Huelva-Palos de la Frontera s/n, 21819 La Rábida-Palos de la Frontera, Huelva, Spain*
[3]*Hospital Universitario Virgen del Rocío, Avda. Manuel Siurot s/n - 41013 Seville, Spain*

Keywords: HIV, Virological Events, Prefurge, Unsupervised Learning.

Abstract: Virological events in HIV-infected patients can rise with no apparent reason. Therefore, when they appear, immunologists or medical doctors do not know whether they will produce other future virological events or they will entail relevant clinical consequences. This paper presents the results of applying Prefurge to HIV-infected patients' clinical data, with the aim of obtaining rules and information about this set of clinical trials data that will relate these kinds of virological events.

1 INTRODUCTION

Data mining (extracting hidden and predictable information from big databases) has a huge potential to help companies get meaningful information from their databases. Tools based on this technology allow predicting future tendencies and behaviours, focused on this stored knowledge. Automated prospective analyses, which are obtained through this technology, go further than the knowledge provided by classical support to decision making tools. These new tools can answer some questions that will take us long time to response, if they do not exist. Such tools explore the databases looking for hidden patterns, extracting predictable information that cannot be found by means of classical mechanisms (lessons learned or traditional tools) utilized by an expert.

Prefurge (Aroba, 2003) is a data mining tool dealing with fuzzy clustering, whose methodology is based on Sugeno and Yasukawa's (Sugeno and Yasukawa, 1993) proposal. Such proposal has been modified and adapted to be able to apply it to multi-parametric quantitative databases with more than a unique output variable. Therefore, Prefurge fits in the context of *Unsupervised Learning,* which means that it does not completely need a tag that determines a record class. This is an important detail, since an expert executes this process using experience-based criteria that highly influence the obtained rules.

In this paper, we apply Prefurge to a medical database of *Virgen del Rocío* University Hospital (composed of approximately 300,000 clinical trials from 5,000 patients who are being treated at this hospital) with the aim of reducing costs and the time spent on their treatments as well as improving their quality of life.

This paper has been structured as follows: in section 2 (Related Works), we show some works where Prefurge was successfully used in different knowledge areas; in section 3 (HIV Patients), we reflect on the motivation that make us elaborate this paper and we describe patients' parameters and the context where this paper is developed; in section 4 (Intelligent Data Analysis), we study the used database and its pre-processing in order to apply Prefurge, together with the analysis of the obtained results. Finally, section 5 states conclusions and future work to keep on working on this research line and applying Prefurge to different patients.

2 RELATED WORKS

Since it was developed as a result of a PH-thesis,

Domínguez M., Aroba J., G. Enríquez J., Ramos I., M. Lucena-Soto J. and J. Escalona M..
Advances in the Decision Making for Treatments of Chronic Patients Using Fuzzy Logic and Data Mining Techniques.
DOI: 10.5220/0004969503250330
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ICEIS-2014), pages 325-330
ISBN: 978-989-758-027-7

Prefurge (Aroba, 2003) has allowed discovering new and very relevant information (and sometimes unknown), using datasets coming from several and diverse contexts. All the obtained results have been validated by experts from different contexts who have confirmed the veracity and importance of the qualitative information generated by this tool. Some of the results obtained have led to publications in indexed journals of first level, such as (Aroba et al., 2008), (Gegúndez et al., 2008), (Grande et al., 2010a), (Grande et al., 2010b), (Francisco de Toro et al., 2011).

A segmented software cost estimation model based on fuzzy clustering is proposed in (Aroba et al., 2008). The use of fuzzy clustering allows obtaining different mathematical models for each cluster and also enables that the items of a project database can contribute with more than one cluster, while preserving constant time execution of the estimation process. The results of an evaluation of a concrete model using the ISBSG 8 project database are reported, yielding better figures of adjustment than its crisp counterpart.

An identification method for a class of dynamic system, known as piecewise affine system, is presented in (Gegúndez et al., 2008). Such systems are composed of a set of affine maps which relate inputs and outputs. The aim of the proposed method is to obtain a model of the system from a set of input-output data. The method uses a process of fuzzy clustering in order to obtain a subset of representatives from the original data set, so reducing the amount of information to be processed, while retaining the significant information from the original data and minimizing the effect of noise on such data.

A qualitative fuzzy model is proposed in (Grande et al., 2010a) to help characterizing and interpreting the behaviour of arsenic in a complex system submitted to acid mine drainage processes (Tinto river). The conclusions present that the factors that most directly control the presence of total dissolved arsenic are temperature and rainfall, and therefore, pH.

The main objective of (Grande et al., 2010b) is to characterize the behaviour of arsenic conditioned by the presence of other elements which are characteristic of acid mine drainage pollution, both in generating milieu and receiving milieu of the Tinto river. The paper analyzes the different behaviours arsenic assumes at both sites, as well as its relationships with the rest of the elements.

Fuzzy logic tools have been also used in others sanitary contexts, such e.g. Computer-Aided Diagnosis (CAD) of the paroxysmal atrial fibrillation (Francisco de Toro et al., 2011), where this fuzzy logic tool helped other evolutionary methodologies, generating rules that can be easily used by medical specialists.

3 HIV PATIENTS

Antiretroviral treatments (ART) against HIV have reached a high success rate in infected patients. The goal of ART is to eliminate the viral particle in plasma limiting its infectivity (viral suppression). HVI RNA, quantified by real time RT-PCR, in plasma (viral loads) and CD4+ T cell count in peripheral blood are the most important response marks to ART.

Periodic monitoring is necessary, once viral suppression is achieved. In fact, positive viral loads events are detectable. Depending on the duration and viral load concentration of these events they may have clinical implications. The most common events are blips, generally defined as single, low-level viral load measurements that are followed by a return to virologic suppression. Obviously, the increased sensitivity of the technique has led to an increased frequency of blips. Nowadays, RT-PCR real time technique is able to detect 20 copies/ml (COBAS® AmpliPrep/COBAS® TaqMan® HIV-1 Test). The causes and clinical implications of blips have been debated. Firstly, some blips may be associated to biological variability and test variability (± 0.5 $\log_{10}$) (Bryan et al., 2011). Although some studies associate blips occurrences with a higher frequency of relapse (Greub et al., 2002), most studies have confirmed that blips do not result in increased risk of virologic or clinical failure (García-Gasco et al., 2008), (Nettles et al., 2005).

Since we have a large number of samples, it would be useful to establish whether blips relate or not to subsequent clinical changes in patients.

4 INTELLIGENT DATA ANALYSIS

4.1 Source Data

Virgen del Rocío University Hospital has provided us with the data collection that we are going to process. Such data sampling contains around

300,000 clinical trials of more than 5,000 patients that are being treated at this hospital.

These data are recorded in a spreadsheet whose rows include the following type of data:

- Date of the clinical trial.
- Clinical History Identification (CHI), a number that uniquely identifies each patient.
- Age of the patient.
- Clinical Trial Request (CTR), a number that exclusively identifies each request for one or more clinical trials.
- Clinical Trial Code (CTC), a number that individually identifies each clinical trial.
- Clinical Trial Description, which can be one of the following: Viral Load, Lymphocyte Count, CD3+T-cells, CD4+T-cells or CD8+T-cells.
- Clinical trial result.
- Comments on the results.

Every single clinical trial and its results have been previously anonymised to protect the health privacy of each patient.

Data Preparation

The given data collection needs an early processing with the aim of correcting each wrong, void or out-ranged detail, so we removed each invalid row and setting a maximum threshold of detectable viral load in 20 copies/mL, although DHHS defines virologic suppression below the limit of assay detection between 20 and 75 copies/mL, in order to maximize the number of detected blips. (Aroba, 2003) These corrections make a right input for Prefurge, and reduce the data sampling to 250,000 clinical trials and 2,500 distinct patients.

Data Generation

There are some data that are susceptible of being sets of antecedents or consequences and therefore, they can be processed by Prefurge.

However, they are not the only data we want to process, but we are also interested in the connection among these virology events (U.S. Dep., www.hhs.gov):

- Virologic Suppression: Two or more confirmed HIV RNA levels (copies/mL) in a row below the limit of assay detection.
- Virologic Blip or Blip: After a virologic suppression, an isolated detectable HIV RNA level (copies/mL) that is followed by a return to a virologic suppression.
- Incomplete Virologic Response: Two consecutive plasma HIV RNA (copies/mL) confirmed positive levels.
- Persistent low-level viremia: Three or more consecutive plasma HIV RNA (copies/mL) confirmed positive levels.
- Virologic failure: A high detectable HIV RNA level (over 1,000 copies/mL in this study).
- Time between each trial: Number of days spent from a clinical trial to the next one.

These data have been calculated from the Viral Load (VL) results for each patient, and added to the initial data set, in order to be processed by Prefurge in the same way that the other data.

It seems that calculated data are not going to provide new information, as they are calculated in terms of the frequency of changes on viral load values. Therefore, it is not direct related information.

4.2 Application of Data Mining Techniques

Prefurge (Predictive Fuzzy Rules Generator)

It is important to note that the graphic output provided by Prefurge enables an easy interpretation of the fuzzy rules in a natural language. As an example, Figure 1 shows a rule generated by Prefurge. In rule (a) of Figure 1, the fuzzy set assigned to each parameter is represented by a polyhedron. The parameter values are represented on the *x* axis of each fuzzy set, and the value of membership to a cluster on the *y* axis. This fuzzy rule would be interpreted as follows: "IF *Age* is *small* and *VL* is *bigger or equal to the average* THEN *Blips* are *very small*"

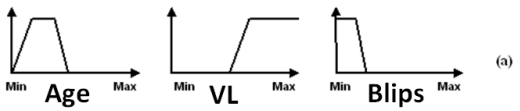

Figure 1: Example rule.

Due to the discrete nature of some parameters that we will process, fuzzy sets generated in some rules, represent discrete ranges of values, instead quantitative values such as *big* or *small*.

Experiments Design

Data input must have a specific structure to be able to be processed by Prefurge. This concrete data input has been structured by creating an initial spreadsheet whose columns contain an antecedent or a consequence.

Differentiating useful data is the first step involving ignoring the useless data; thus, we will only use clinical related data:

- Date. It is necessary to calculate the days spent from one to the subsequent virologic event.
- Age. It is a clinical parameter.
- Result. It refers to the main clinical parameters results on the clinical trials:
  - Viral Load.
  - Lymphocyte Count.
  - CD4+T-cells
- Virologic blip or 'blip'.
- Incomplete virologic response.
- Persistent low-level viremia.
- Virologic failure.
- Time between each trial (TBT), which can be another useful parameter.

Some data can take very different ranges in their values so they have been normalized, when needed.

Experiments with New Parameters

There are a lot of parameters that we can use as input to Prefurge, but which of them are really interesting? Which of them should we use? To answer these questions we must start with the most basic parameter, Viral Load, since it constitutes the source of any forward virologic event. We can use Viral Load as the first and main antecedent of any rule.

Moreover, a blip directly depends on Viral Load, so we can use it as the first consequence of a rule. Furthermore, we can add any clinical parameter as an antecedent, since we do not know if it will show any extra result previously.

We can design the first experiment with this information:

Experiment #1

- Antecedents: Age, TBT, Viral Load.
- Consequents: Blips.

Following the same way of thinking, we notice that an incomplete virologic response needs a previous blip to get raised; therefore we can use the blips as antecedents of the consequent incomplete virologic response.

Experiment #2

- Antecedents: Age, TBT, Viral Load, Blips.
- Consequences: Incomplete Virologic Response.

Persistent low-level viremia appears when a blip raises becoming an incomplete virologic response. Thus, we can consider it as an antecedent of a Persistent low-level viremia for a new experiment.

Experiment #3

- Antecedents: Age, TBT, Viral Load, Blips, Incomplete Virologic Response.
- Consequents: Persistent Low-level Viremia.

These events may indicate that the treatment is failing (Aldous and Haubrich, 2009), the virus is mutating or the expression of proviral DNA integrated in lately infected cells (García-Gasco et al., 2008), whence a virologic failure can appear. In this case, a virologic failure could be treated as a consequence of the rest of the virologic events.

Experiment #4

- Antecedents: Age, TBT, Viral Load, Blips, Incomplete Virologic Response, Persistent low-level viremia.
- Consequences: Virologic Failure.

Experiment #1 Results

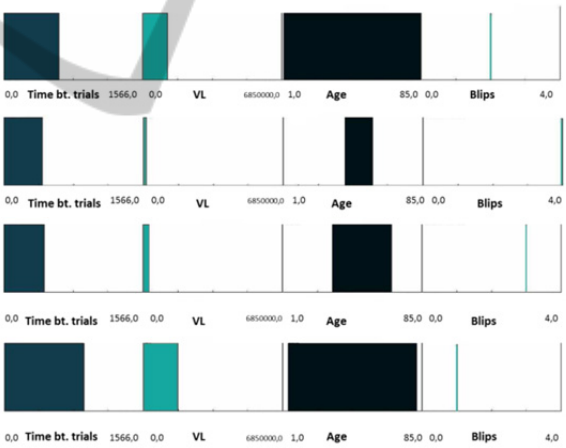

Figure 2: Experiment #1 Results: From top to bottom, Rules #1, #2, #3, and #4.

Experiment #2 Results

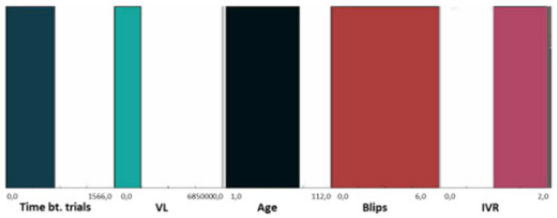

Figure 3: Experiment #2 Results Rule #1.

Experiment #3 Results

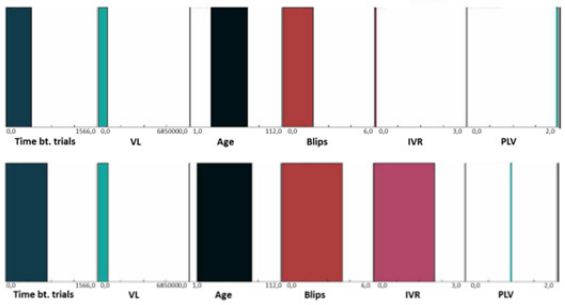

Figure 4: Experiment #3 Results: From top to bottom, Rules #1 and #2.

Experiment #4 Results

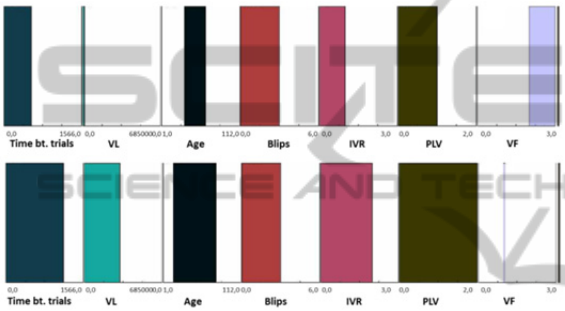

Figure 5: Experiment #4 Results: From top to bottom, Rules #1 and #2.

4.3 Analysis

The first step in the experiments results is to remove the rules whose antecedents do not discriminate any range of values. In that case, any antecedent value can produce the consequence, so it does not provide us with any useful information.

Experiment #1 Analysis

In Exp #1, Rule #2, Rule #3 and Rule #4 show a link between Age and Blips, as the lowest age range seems to be linked to more blips, but unfortunately this relation is caused by the number of records in that age range are the most in the sample; therefore the number of blips detected is higher.

Experiment #2 Analysis

Rule #1 antecedent blips take all the possible values in its range, so this rule does not reveal any useful information.

Experiment #3 Analysis

Rule #1 and Rule #2 antecedents, such as Time between trials, Viral Load, Age or Blips do not provide any useful information because their values have very similar ranges.

Nevertheless, Rule #1 shows how the absence of any Incomplete Virologic Response antecedent is related to the presence of Persistent low-level viremia. Rule #2 shows how the presence of Incomplete Virologic Response antecedent is also associated with that presence of Persistent low-level viremia. This demonstrates that Incomplete Virologic Response is completely independent of Persistent low-level viremia, as it can appear without any related previous Incomplete Virologic Response.

Experiment #4 Analysis

Rule #1 and Rule #2 show that higher Viral Load values relate to less Virologic Failures, but a deeper analysis of data shows that Rule #1 concerns to patients who have null or very low values on their Viral Load for a long time, that is showed as lower Viral Load in Rule #1.

4.4 Learned Lessons

These results can show us that we must not stop working on a set of data, even if it has been processed many times in many ways; because when we use non-classical statistical methods (such Prefurge) we can take out important hidden new information from the data set.

5 CONCLUSIONS AND FUTURE WORK

One of the most important results rated by the expert team has been meeting the independence between the Incomplete Virologic Response and Persistent low-level viremia. It allows researchers to focus their efforts on other parameters or setting new conditions to their experiments, trying to get different results from these found right now. This situation may even cause that other related research projects using the parameters that have been processed could be developed more quickly.

Analysing the remaining results, we have realized that we have not found any new relation between some parameters studied and the knowledge that we have of them up to now. The causes of this gap perhaps may be explained by this discrete nature of most parameters, since Prefurge is a data mining tool designed for continuous and numerical sets of data. However, we could even

process discrete attributes (e.g. Blips) generating some meaningful information, although its interpretation is somehow different: antecedents and consequences are no longer *big* or *small* values, but discrete values as, for instance, zero blips, one blip, two blips, and so on.

Despite these difficulties, this work lays the foundations for developing a research line focused on the use of Prefurge with HIV chronic patients, which provides health researchers with tools and methodologies to process this type of biohazardous data. This possible solution will help scientists and millions of people all over the world.

One of our most priorities regarding future work will be testing all these experiments with another algorithm to check how Prefurge really behaves. As soon as we totally confirm that Prefurge is our best tool, we can improve it with new functionalities according to our main needs.

ACKNOWLEDGEMENTS

This research has been supported by MeGUS of the Ministerio de Ciencia e Innovación and by NDTQ-Framework project (TIC-5789) of Junta de Andalucía, Spain.

REFERENCES

Aroba, J. *Advances in the decision making in software development projects.* PhD thesis, University of Sevilla, Spain, 2003.

Aroba,J., Cuadrado, J., Sicilia, M., Ramos, I., García, E. 2008. Segmented software Cost Estimation Models based on fuzzy clustering. In *Journal of Software and Systems*

Gegúndez, M., Aroba, J., Bravo, J. 2008. Identification of piecewise affine systems by means of fuzzy clustering and competitive learning. In *Journal of Engineering Applications of Artificial Intelligence*

Grande, J., Andújar, J., Aroba, J., de la Torre, M. 2010a. Presence of As in the fluvial network due to AMD processes in the Riotinto mining area (SW Spain): A fuzzy logic qualitative model. In *Journal of Hazardous Materials*

Grande, J., Andújar, J., Aroba, J., Belrán, R., de la Torre, M., Cerón, J., Gómez, T. 2010b. Fuzzy Modeling of the Spatial Evolution of the Chemistry in the Tinto River (SW Spain). In *Journal of Water Resource Management*

U.S. Department of Health and Human Services http://www.hhs.gov/

Aldous, J., & Haubrich, R. (2009). Defining Treatment Failure in Resource-Rich Settings. Curr Opin HIV AIDS, 459-466.

García-Gasco, P et al. Episodes of low-level viral rebound in HIV-infected patients on antiretroviral therapy: frequency, predictors and outcome. *Journal of Antimicrobial Chemotherapy* (2008) 61, 699–704

M. Sugeno, T. Yasukawa. A Fuzzy-Logic Based approach to qualitative Modeling. IEEE Transactions on Fuzzy Systems, Vol.1.Pp.: 7-31, 1993.

Bryan R. Cobb, Jeffrey E. Vaks, Tri Do, Regis A. Vilchez-Evolution in the sensitivity of quantitative HIV-1 viral load tests. *Journal of Clinical Virology* 52S (2011) S77– S82

Greub G, Cozzi-Lepri A, Ledergerber B, Staszewski S, Perrin L, Miller V, Francioli P, Furrer H, Battegay M, Vernazza P, Bernasconi E, Günthard HF, Hirschel B,Phillips AN, Telenti A; Frankfurt HIV Clinic Cohort and the Swiss HIV Cohort Study. Intermittent and sustained low-level HIV viral rebound in patients receiving potent antiretroviral therapy. AIDS: Volume 16(14), 27 September 2002, pp 1967-1969

Nettles et al. Intermittent HIV-1 Viremia (Blips) and Drug Resistance in Patients Receiving HAART. JAMA, February 16, 2005—Vol 293, No. 7

Francisco de Toro, Javier Aroba & Eduardo Ros (2011): Computer-Aided Diagnosis of the Paroxysmal Atrial Fibrillation: A Fuzzy-Evolutionary Approach. Applied Artificial Intelligence, 25:7, 590-608