Pattern Characterization in Multivariate Data Series using Fuzzy Logic
Applications to e-Health

W. Fajardo[1], M. Molina-Solana[1] and M. C. Valenza[2]
[1]*Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain*
[2]*Department of Physiotherapy, University of Granada, Granada, Spain*

Keywords: Health Care, Data Series, Fuzzy Logic.

Abstract: The application of classic models to represent and analyze time-series imposes strict restrictions to the data that do not usually fit well with real-case scenarios. This limitation is mainly due to the assumption that data are precise, not noisy. Therefore, classic models propose a preprocessing stage for noise removal and data conversion. However, there are real applications where this data preprocessing stage dramatically lowers the accuracy of the results, since these data being filtering out are of great relevance. In the case of the real problem we propose in this research, the diagnosis of cardiopulmonary pathologies by means of fitness tests, detailed fluctuations in the data (usually filtered out by preprocessing methods) are key components for characterizing a pathology.

We plan to model time-series data from fitness tests in order to characterize more precise and complete patterns than those being currently used for the diagnosis of cardiopulmonary pathologies. We will develop similarity measures and clustering algorithms for the automatic identification of novel, refined, types of diagnoses; classification algorithms for the automatic assignment of a diagnosis to a given test result.

1 INTRODUCTION

There are fields like Medicine (cardiopulmonary mechanics, in particular), where data series are of great importance but they are not fully used. The reason of that is the partial ignorance of methodologies for treating such information. Even more, as data in those domains are usually imprecise and contains lots of noise, it is a sensible choice to employ not only classic techniques from data mining but also fuzzy logic.

It is our aim to enrich the scope of tools for data series representation when they are multidimensional and the observations are imperfect. We aim to devise a simple representation model that takes into account that inherent imperfection.

The reason is that the data we foresee to employ come, not from rigorous analytics, but from observations of natural phenomena. During monitorization and direct observation of physical events, it is common to get imprecise data. That can be due to several factors such as imperfect measure tools, gathering errors or by the inherent nature of the phenomena. Depending on the particular problem, a preprocessing of the data might be harmful, as it could influence, in our examples, the medical diagnosis.

Several tools for managing data series are currently available, but they require such a preprocessing stage to remove imperfect information. That preprocessing removes some data from the series according to some criteria, leading to a loss in information that, in some cases, is not admissible.

The present research presents a novel approach, different from the classic one, which assumes the imperfection on the observations. Instead of pretending that data series are precise (by means of preprocessing and transformations), we assume that observations are imperfect and therefore we propose representation models according to that.

The team that supports this research is multidisciplinar and includes researchers from both the fields of Medicine and Computer Sciences (experts in Fuzzy technologies). This fact guarantees that the present research could be correctly completed.

Fajardo W., Molina-Solana M. and C. Valenza M..
Pattern Characterization in Multivariate Data Series using Fuzzy Logic - Applications to e-Health.
DOI: 10.5220/0004124701230128
In *Proceedings of the International Conference on Data Communication Networking, e-Business and Optical Communication Systems* (ICE-B-2012), pages 123-128
ISBN: 978-989-8565-23-5

2 BACKGROUND

Even though the world changes along time, the great majority of methods for pattern recognition (especially those that can be tagged as classic) are mostly dedicated to the task of detecting static patterns. Taking changes in time into consideration is, from a mathematical point of view, a mere addition of a new dimension to the original problem. However, the experience points out that the temporal component facilitates a great variety of new problems with great practical and scientific interest.

Despite what we have said in the former paragraph, it should be also acknowledged that Statistics have historically given a special attention to the study of space-temporal data. That fact has leaded to the development of an area within Statistics known as Time Series Analysis. This area is of great theoretical importance, but also has a huge practical interest, with application in fields such as Economy, Biology, Environmental Engineering or Medicine.

As said, the approach that the majority of statistical methods take for representing information in the context of time series is very precise regarding events and their timing. Because of that, those methods usually have limitations with respect to which phenomena they can deal with, and often, they impose restrictions that are rarely met in real problems, data and observations.

More on the contrary, the information in these cases is presented in an imperfect way, as it can be incomplete, imprecise, vague, fragmented or contradictory.

Due to that fact, researchers have worked on the representation of imperfect information in the context of time series with AI techniques. As those techniques are based on hypothesis less constrained that those from Statistics, they are better suited for real problems.

From the specialized literature, we can see that some work have been done (with different results) in this direction (Mirikitani and Nikolaev, 2010), (Han et al., 2011). Even more, new techniques specially devised for this kind of problems have been proposed, such as Artificial Neural Networks, Soft Computing, Evolutionary Algorithms, Regressive Tress, etc. They all lie within the new area of "Intelligent Computing", which is currently very active.

However, direct management of imperfect data series is still an open problem. Only a few theoretical works can be found on this topic, and there is still a lack of studies and applications to concrete problems (including the solution of the specific problems of each domain).

In this research, we aim to study how to represent data series with imperfect observations and its applications from human ventilation and cardiopulmonary mechanics, with the goal of modeling and extracting information from these data. Theoretical studies will be validated with experimentation on real data.

From a computational point of view, the results of a fitness test can be described as a series of data. Each one of these data can be obtained from different sources. Therefore, a test result can be expressed as series of (imprecise) measures in several dimensions.

Regarding cardiopulmonary mechanics, monitoring the variables associated with the physical activity is an issue of great scientific and practical interest (Lötjönen, 2003), (Martin, 2000), (Korhonen, 2003). The development of new tools and systems capable of extracting more information from the aforementioned signals will be of special importance in several fields. For instance, in the prevention of working risks, by means of monitoring cardiopulmonary activity; as a therapeutic tool to check the evolution and suggest suitable training; or to identify injuries by the classification in several levels.

In particular, current diagnosis systems present the following limitations when dealing with effort tests:

- They only consider a few variables (which are the variables with diagnostic relevance), and only the observations at the beginning and end are taken into account. The evolution of those variables is fully ignored, and the diagnosis is given according to some reference tables. The treatment of those variables as time series will facilitate better classifications and will create new categories and subcategories of diagnosis.
- Scientists are starting to give clinical relevance to the other variables (variables with clinical relevance) that might influence the diagnosis. So far, those observations are collected but nothing has done with them as there is no reference value to compare with.

In the next section we will present some basic ideas about the tools and methodology that support the proposal of this research. It is not among the objectives of a presentation of antecedents, like this one, to offer a deeper study about time series and their applications. Having said that, it is appropriate to describe which tools are available and what other researchers have already done in the application of

Intelligent Computing techniques to Time Series Analysis.

2.1.1 Time Series and their Processing

In the following paragraphs we offer a brief introduction to time series and their processing. We do not aim here to make a deep review, but to establish the context in which the antecedents and objectives of this research do make sense. For a more detailed review on this topic, we point the interested reader to any specific text such as (Box and Jenkins, 2008) or (Han, Kamber and Pei, 2011).

In general terms, a time series is a series of values of a n-dimensional variable *x* that depends on time, that is *{x(t), t∈T}*. There are some differences between the management of unidimensional and multidimensional data series, that basically come from the possibility of interdependences among the components of *x*. Theoretically, *T* can be a continuous interval, but in practice, time is always considered as discrete. Therefore, the series can be seen as a list of observations in several time instants. The distance between those time points is fixed and it will be determined by the particular problem. Regarding variable *x*, *x(t)* could be the observation at a time *t* or the average value during the interval *[t-1,t]*. In any case, sampling the series is a crucial task that has to be performed at the very beginning (Han, Kamber y Pei, 2011). In the context of this research, we will suppose that the series have already been correctly sampled.

Time series are a special case of data series, that are series of observations over a variable (generally n-dimensional) indexed by the values of another unidimensional variable. As it is discussed in (Pyle, 1999), even though we often refer to time series -as they are the most common- everything about them can be almost directly applied to data series in general. In some cases, when the indexing variable is time, it is just an index and it is not playing any special role in the series, or inducing any dependencies.

2.1.2 Objetives of Data Series Processing

Historically, five goals can be identified in the analysis of data series. The first and the second are the most widely considered in literature due to their great practical interest. However, the borders among them are not well defined, and many times several are required to solve a particular problem. The aforementioned objectives are the following:

- Prediction of future values of the series.
- Classification of the series, globally or partially, in different categories.
- Description of the series according to a model.
- Description of the series according to the values of other series.
- Clustering and pattern discovery from time-series data.

From a formal view point, the prediction task can be seen as finding a function *F* which gives an estimation *x'(t+d)* of the values of *x* at time *t+d*.

That estimation is made from the last k values of *x* before *t* and other external factors, *d*. In other words, *x'(t+d)=F(x(t), x(t-1)...., x(t-k+1), d)*.

Usually, *d=1*, but depending on the concrete application, a different value might be required. Almost all traditional methods for time series analysis and monitorization require *F* to be stationary, that is, *F* only depends on *t* by means of observations of *x*. In other words, *F* does not directly depend on the index variable *t*. From this point of view, the prediction is formally a problem of approximating functions; and it that case, a suitable technique from the static problem can be applied (see (Garbancho, 1994) and (Duda and Hart 1973)). As usual, the goodness of the fitting is measured by means of an error function in the form $E = \Sigma i=1,2,....N\ e(x'(t-i),x(t-i))$, where *e* is a function that measures the "difference" between the estimated value and the observed one, *x'(t-i)* y *x(t-i)* respectively. It is common for *e* to be a distance measure, but it could be of other type for particular applications (Dorffner, 1999).

So far, we have talked about determining the value of an observation from the former observations. A different problem is calculating the value of an observation at a time *t* from other observations (in other dimensions) at that same time point. That is objective 4. To do so in an effective way, it is mandatory that the different series are not independent. In fact, a casual relation should exist between them. The guessing of a value in a series from the values in other series (casual prediction) can be formally described, in its simplest version, as: given the series *x(t)*, *y(t)*, *z(t)*, ..., *t=1,2,...,T*, find *G* so that *x(t)=G(t, y(t), z(t),....., h)*.

Casual description of time series has been handled as a variant of the task of predicting multidimensional series. However, we consider that this description possesses specific problems that are quite interesting, and because of that, description will play an important role within this research. In particular, we believe that finding relations between series associated with the same phenomena is of great interest in a context of imperfection. This task

(goal 5 from the list above) has traditionally been called "pattern discovery" or "motif discovery" from time-series data, being typically tackled by applying clustering algorithms on these data (Warren Liao, 2005). However, the key component for these methods to succeed is the accuracy of the model representing the time-series and the function used to measure the similarity between two time-series being compared. Several distance/similarity measures have been proposed so far for full-sequence and sub-sequence matching (Warren Liao, 2005) (Fu, 2011).

On the other hand, second and third objectives are related with modeling and representing a data series, and its posterior classification in relation with others previously analyzed. The classification of data series is a logic extension of the classic classification of isolated values.

Modeling data series is implicitly contained within first, fourth and fifth objectives. In fact, *F* is a model of the series, and it could generate the series by means of using successive estimations as inputs. The difficult task is building models with a reduced number of parameters (degrees of freedom) that, in any case, must be lower than the number of available observations.

In order to model a data series, it is mandatory to employ a suitable representation and, in particular, one capable of handling all the particularities that differentiate data series from other kinds of data.

Our research has the aim (as we will see later) of representing data series, with the final goal of classifying those series according to their features and similarities with others. We also aim to identify relevant patterns in the particular study cases we propose. We foresee a representation capable of dealing with whole series of real data (especially if imperfect observations occur).

2.1.3 Imperfection in Data Series

The description about data series that we have made so far in this section implicitly assumes that a perfect model can be found to represent the data reducing the error as much as we pleased.

However, as we said in the introduction, when the researcher tries to model a data series (time series or other), he finds that that assumption is hardly the case. In fact, an exact description of the behavior of the data is never available. Because of that, a preprocessing stage is needed in order to transform the data to fit in a given model. As we previously said, two alternatives are possible to solve this issue (Motro, 1996).

The classic approach (and the one that most tools in literature take) consists on using only the part of the data that is precise (or assuming they are). In such a case, a preprocessing stage is often needed in order to transform those imperfect data to new values more precise. Once this stage is completed, classic algorithms and procedures described in the literature can be applied.

The second alternative, and the one we propose for this research, assumes form the beginning that the data are imperfect. This way, it is mandatory to devise representation models that do not impose a precise representation of the data. Even more, those models must deal with the intrinsic imperfection in a natural way, without ignoring it.

Nevertheless, and due to its historical relevance in the management of data series, we devote the next paragraphs to describe some of the sources of imperfection that are present when dealing with data series. We will also indicate how they have been solved in the literature. A deeper study on this issue may be found in (Han et al., 2011).

2.1.4 Management of Data Series Imperfection

As we have already said, most tools for data series management have opted to assume that data are precise. Because of that, they need a preanalysis and a preprocessing stage in order to guarantee an optimal result. The following are some of the sources of imperfection that motivate such a stage.

As a consequence of measure error or because of the influence of non-controllable factors, it must be assumed that even the most perfect model do have a residual error that cannot be deleted in any way. Usually, it is accepted that this error is a noise process, that is, it randomly comes from an unknown source. Historically, this error has been modeled as stationary Gaussian noise that follows a given normal distributions for each *t*. Under this hypothesis have been devised the most common methods for predicting future values of the series. They generically adjust to the autoregressive moving average model of order *k, q, ARMA[k,q]* (see (Box and Jenkins, 2008)).

Another problem that is well-known is the presence of values *x(t)* that does not really represent the usual behavior of the phenomena. Those values, known as *outliers*, are consequence of measure errors or erratic movements of the observed phenomenon. They should be removed from the series and, eventually, replaced by 'more normal' values if one wants the model to significative and

useful.

Finally, it is desirable to remove those systematic behaviors that could jeopardize the modeling process or hide other interesting behaviors. The detection of those problematic behaviors could provide initial valuable information. A typical example consists on removing the tendency (linear or not), that is, systematically increasing or decreasing the average value of the observation. A simple way or removing linear tendency consists on substituting the series *{x(t), t=1,2,...,T}* by series *{x'(t)=x(t)-x(t-1), t=2,...,T}*. In the same way, it is a good procedure to remove "stationarities" from the observations, that is, those patterns that are periodically repeated. Substituting *{x(t), t=1,2,...,T}*, by *{x'(t)=x(t)-x(t-s), t=s,...,T}* we can remove them.

When using traditional methods for time series analysis, removing repeating patterns depending on time is a task that must be done without exception. Should it not be performed, the series would not hold the stationary property that is assumed by those methods.

3 OBJETIVES

The main goal of this research is the development and application of a novel representation for time-series data that assumes the imperfection and noise inherent of these types of data. Therefore, instead of applying filters for making the data precise, we propose a model to capture and deal with this imprecision and noise.

The proposed abstract model for time-series data will be applied and tested on a real-case scenario: cardiopulmonary disease diagnosis, where our team members have extensive experience.

3.1 Previous Results

In the cardiorespiratory mechanical field, the fitness test result can be reduced to an ordered set of data, but health state of a patient is far from being a mere mechanical interpretation of this set. That is due to the fact that many factors interfere. They are again ordered and very closely related with the ordered set of data in the test result. Because of that, many problems related to the diagnosis in the cardiorespiratory mechanical can be easily managed as data series, but it presents the inconvenient, as it has already mentioned, that data require a preprocessing stage in order to remove those points that seem to be out of place or are noise (Pyle, 1996), (Cerrito and Cerrito, 2010), (Hans et al., 2011). This preprocessing could even completely remove the features that make a given piece to be genial. Therefore, the problem cannot be deal with in the classical way.

The previous experience of the team in the data series domain (Delgado et al., 2009) (Delgado et al., 2011) (Jiménez et al., 2011) supports the proposal of developing tools to enrich the theoretical field of representing data series. We propose a novel approach that assumes that data are imprecise and thus avoiding the preprocessing stage that in many cases is harmful and undesirable, as it could alter the integrity of the underlying information within the data series.

To do so, we will work with fuzzy techniques, in which several members of the team have an extensive experience (García et al., 2009), (García et al., 2010), (López et al.2008).

Also, thanks to the extensive experience of the team members in the cardiorespiratory field (Nistico et al., 2010), (Perez Riera et al., 2011), (Lamba et al., 2011), (Valenza, 2010), (Valenza, 2011), the case of study on monitoring fitness test results (*6 minutes walking test and shuttle test)* was identified as a real problem in which the representation of values with imprecise time-series models was required. In this domain (like in the music domain), data from the time-series cannot be aggressively filtered since this process flattens the data and hides the final diagnosis.

Furthermore, to establish a diagnosis from fitness test results, current reference models only take into account the values of some variables (called variables with diagnostic relevance) at the initial and ending stages of the fitness tests, ignoring the evolution of these variables during the whole test. The representation of the complete time-series in the model would lead to the identification of novel, refined, types of diagnoses, allowing personalized treatments. The scientific literature is also starting to highlight the role of other variables (called variables with clinical relevance) that are currently being monitored in fitness tests but not effectively being used yet in the models to establish a diagnosis. Therefore, there is a need for new time-series models which can inherently deal with uncertainty and incorporate these variables to effectively derive a diagnosis.

4 CONCLUSIONS

As we have seen along this paper, the representation

of information by means of data series capable of accommodating imperfection, presents advantages when employed in several fields.

In the case of cardiopulmonary diagnose, those tools make it possible to appropriately deal with some tasks that have been neglected to the date because of their complexity.

The accurate representation of imperfect data series enables a better management of that information, allowing its electronic sharing and treatment. Clinical diagnosis applications will greatly benefit from such a novel representation, which might lead to further understanding of illnesses and their automatic identification.

REFERENCES

Box, G. E.; Jenkins, G. M. and Reinsel, G. C. (eds). *2008. Time Series Analysis: Forecasting and Control, 4th edn. Published by John Wiley and Sons, New York.*

Cerrito, P.; Cerrito, J. 2010. *Clinical Data Mining for Physician Decision Making and Investigating Health Outcomes*: Methods for Prediction and Analysis. IGI Global.

Delgado, M; Fajardo, W.; Molina-Solana, M. 2009. *Inmamusys: Intelligent multiagent music system*, Expert Systems with Applications, 36(3), pp. 4574-4580, ISSN 0957-4174

Delgado, M; Fajardo, W.; Molina-Solana, M. 2011. *A state of the art on computational music performance*, Expert Systems with Applications, 38(1), pp. 155-160, ISSN 0957-4174

Dorffner, G. 1999. *Neural Networks for Time Series Processing*" Tech. Rep., Medical Cybernetics and A.I., Univ. Viena, Viena.

Fu T C. 2011. *A* review *on time-series data mining*. Engineering Applications of Artificial Intelligence; 24(1); 164-181

García, F.; López, F J.; Cano, C.; Blanco, A. 2009. *FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral.* BMC Bioinformatics. 10(224).

García, F.; Blanco, A.; Shepherd, J. 2010. *An intuitionistic approach to scoring DNA sequences against transcription factor binding site motifs.* BMC Bioinformatics. 11(551) pp. 1-13.

Garbancho, A.G. *Estadística Elemental Moderna.* 16. 1994. Economía.

Han, J.; Kamber, M; Pei, J. 2011. *DataMining Concepts and Techniques.* Morgan Kaufmann.

Jiménez, A; *Molina*-Solana, M.; Berzal, F.; Fajardo, W. 2009. *Mining transposed motifs in music.* Journal of Intelligence Information Systems, 36 (1), pp. 99-115.

Korhonen, I., Parkka, J. 2003. *Health monitoring in the home of the future* Engineering in Medicine.

Lamba, J; Simpson, C. S.; Redfearn, D. P.; Michael, K. A.; Fitzpatrick, M.; Baranchuk, A. 2011. *Cardiac Resynchronization Therapy for the Treatment of Sleep Apnea: A Meta-analysis.* Europace (in press)

Warren Liao T. *2005*. *Clustering of time-series data - a survey*. Pattern Recognition; 38, 1857-1874

López, J.; Blanco, A.; García, F.; Cano, C.; Marín, A. 2008. *Extracting Knowledge from Heterogenous Data Biological by Fuzzy Association Rules*. BMC Bioinformatics. 9 (107) pp. 1-18.

Lötjönen, J., Korhonen, I., Hirvonen, K., Eskelinen, S. 2003. A*utomatic sleep-wake and nap analysis with a new wrist worn online activity monitoring device Vivago Wristcare* - Sleep.

Martin, T., Jovanov E. 2000. *Issues in wearable computing for medical monitoring applications: a case study of a wearable ECG monitoring device*. Wearable Computers.

Mirikitani, D. T.; Nikolaev, N. 2010. "Recursive Bayesian Recurrent Neural Networks for Time-Series Modeling," *Neural* Networks*, IEEE Transactions on*, 21 (2), pp.262-274.

Motro, A. 1996. *Sources of uncertainty, imprecision, and incosistency in information systems.* Vol. Uncertainty Management in Information Systems: From Needs to Solutions, in *Uncertainty Management in Information Systems: From Needs to Solutions.*, edited by A. and Smets, P. Motro, 9-34. Kluwer Academic Publishers,

Nistico, A; Iliescu, E. A.; Fitzpatrick, M. F; White C. A.. 2010. *Polycythemia due to obstructive sleep apnea in a patient on* hemodialysis. Hemodialysis International. 14: 333-6. Wiley

Perez Riera, A. R.; Ferreira, M.; Hopman, W. M.; McIntyre, W. F.; Baranchuk, A. 2011. *Electrovectorcardiographic Characterization of The Type-1 Brugada ECG Pattern*. Europace; P1214: 139.

Pyle, D. 1999. *Data Preparation for Data Mining.* San Francisco: *Morgan* Kaufmann Publishers Inc..

Valenza-Peña, M. C. 2010. *Patterns and attitudes of spanish Healt students who smoke* European Repiratory Journal. pp. 780-781. ISBN 1399-3003.

Valenza-Peña, M. C. 2011. *Analysis of the main sleep-respiratory parameters in patients with fibromyalgia*. European Respiratory Journal. pp. 805-806. ISBN 1399-3003.