ENABLING UBIQUITOUS DATA MINING IN INTENSIVE CARE
Features selection and data pre-processing

Manuel Santos, Filipe Portela
Departamento de Sistemas de Informação
Universidade do Minho
mfs@dsi.uminho.pt, cfp@dsi.uminho.pt

Keywords: Ubiquitous Data Mining, Real-time Intelligent Decision Support Systems, Organ failure prediction, Clinical Data Mining, Intensive Care Environment

Abstract: Ubiquitous Data Mining and Intelligent Decision Support Systems are gaining interest by both computer science researchers and intensive care doctors. Previous work contributed with Data Mining models to predict organ failure and outcome of patients in order to support and guide the clinical decision based on the notion of critical events and the data collected from monitors in real-time. This paper addresses the study of the impact of the Modified Early Warning Score, a simple physiological score that may allow improvements in the quality and safety of management provided to surgical ward patients, in the prediction sensibility. The feature selection and data pre-processing are also detailed. Results show that for some variables associated to this score the impact is minimal.

1 INTRODUCTION

The decision making process is a key factor in critical environments such as intensive care, either in prognosis or diagnosis of the patients' condition, for their lives may be at risk. According to some studies, medical error may be the eighth cause of death in industrialized countries (Kohn, Corrigan, & Donaldson, 2000). The development of Intelligent Decision Support Systems (IDSS) by means of Data Mining (DM) prediction models for mortality and organ failure may contribute to the reduction of medical error.

Presently, there is an IDSS for organ failure prediction by means of DM techniques. INTCare is being tested in the Intensive Care Unit (ICU) of Hospital Geral de Santo António (HGSA) (Santos et al., 2009; Vilas-Boas, Santos, Portela, Silva, & Rua, 2010; Villas Boas et al., 2010).

The high volume of clinical information that doctors have to deal with every day in ICU's explains the need to rely on Decision Support Systems (DSS). ICU's are a complex medical environment where patients' condition is critical and often their lives depend on the care provided in those units. Furthermore, there are so many data relating to a patient, that one person can't effectively process it (Donchin & Seagull, 2002). Safer, less expensive, and higher-quality health care can be achieved using clinical DSS (Menachemi & Brooks, 2006). Computer-based decision support tools are supposed to help practitioners to avoid errors, ensure quality and improve efficiency in healthcare. Health care providers can manually enter patient characteristics into the computer system, but ideally, information retrieval, registration and display should be as automated as possible.

Hence the interest by intensive care clinicians and researchers in the development of prediction models based on DM to support clinical decision. The emergence of network-based computing environments established a new dimension to this area, regarding distributed sources of data and computing. The growing use portable devices and communication networks is making ubiquitous access to large quantity of distributed data a reality and is changing our relationship with a real-time and distributed environment (Kargupta, 2001).

This paper presents the experiments regarding the INTCare system, an IDSS for ICU in real-time, and its pervasive and ubiquitous features. It focuses on the features selection and data pre-processing for mining intensive care data.

2 BACKGROUND AND RELATED WORK

Pervasive computing has been pointed as a new paradigm for the 21st century. Significant hardware developments, as well as advances in location sensors, wireless communications, and global networking, have advanced towards technical and economic viability (Saha & Mukherjee, 2003).

Ubiquitous computing, pervasive computing and ambient intelligence are concepts evolving in a growing number of applications in health care and are increasingly influencing it (Orwat, Graefe, & Faulwasser, 2008). However, the majority of these systems are in their prototype stages and there is a need for advanced research on their deployment, mainly clinical studies, economic and social analyses and user studies. Ubiquitous computing can transform some key features of user interaction: their location, the scope of the service, and its duration and frequency (Fano & Gershman, 2002).

3 PERVASIVENESS IN INTENSIVE CARE

3.1 Requirements of HealthCare environments and computer applications

To our knowledge, there is no IDSS that uses DM models and addresses the important aspects of critical environments, mainly the need for acting in real-time in a fast, reliable, secure and ubiquitous way.

To achieve that, it is essential to make clinical information available by ubiquitous devices (Varshney, 2007), making possible to avoid medical errors that occur due to lack of the correct information when and where it is required (Kohn, et al., 2000). The lack of correct and complete information may lead to decision errors in 50% (Bergs, Rutten, Tadros, Krijnen, & Schipper, 2005). Ubiquity in electronic medical records that contain detailed data about patients leverage their analysis by whom is authorized, anytime and anywhere (Varshney, 2009), allowing physicians and nurses to act in real-time and as fast as possible.

Pervasive healthcare addresses this challenge and, according to Varshney (Varshney, 2007), it may be defined as "healthcare to anyone, anytime, and anywhere by removing locational, time and other restraints". A pervasive IDSS may bridge most of this gaps and help to coordinate the various activities underlying patients care and survival (Scicluna, Murray, Xiao, & Mackenzie, 2008), relying on prediction and decision models that may be accessed anywhere. Based on the work done during the research for the INTCare project, and after a thorough analysis of the problem, some features were identified as being essential to the development of an IDSS. Within this context it was determined to be essential for an IDSS that uses DM and acts in real-time the following features:

a) **Online Learning**: The system should have the ability to operate in online mode, i.e. the DM models must be induced using online data, in contrast to the existing approaches that emphasize offline learning mode.
b) **Real-Time**: The system must be able to act in real-time. The process of acquisition and storage of data should occur immediately after the events take place;
c) **Adaptability**: The system must have the ability to automatically optimize the models with the latest data;
d) **Optimization**: Consists in optimizing the values obtained by the prevision models generated, analyzing all the results and selecting the best.
e) **Data Mining Models**: The success of this IDSS depends on the accuracy of DM models, i.e. the degree of reliability of these models should be high;
f) **Decision Models**: Achieving the best results is highly dependent on the decision models created. These models are based on various factors such as the decision and differentiation that are used in predicting models, so that they can help doctors choose the best decision during the process of decision making;
g) **Intelligent Agents**: This type of agents allows the system to work through autonomous actions that perform some essential tasks. These tasks support some system modules: Data Acquisition, Management of Knowledge, Inference and interface;
h) **Pervasive / Ubiquitous**: The technology of ubiquitous computing aims to improve the living and working environments and to build advanced devices, infrastructure and network operating systems.
i) **Safety**: This type of system that operates in critical environments must guarantee safety for both patients and users. By using ubiquity, some protocols need to be followed like access, authentication, communication (Langheinrich, 2001; Nigel, 2002; Piramuthu, 2003)

4 UBIQUITOUS DATA MINING IN INTENSIVE CARE

Advanced analysis of distributed data for extracting useful knowledge is the next expected step in the increasingly connected world of ubiquitous computing (Neaga & Harding, 2005). In this context, UDM is attracting a great amount of attention. Much of the DM tasks which are being done currently focus on databases or data warehouses whose data is physically located in one place. However, the scenario arises where information may be located in different physical locations. Therefore, the goal is to effectively mine distributed data streams which are located in heterogeneous sites (Hsu, 2002). The challenge and the novelty of this approach is to bring ubiquity to DM, which is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships that provide a clear and useful result to the data analyst (Giudici, 2003). The abundance of mobile devices, such as PDA's and cellular phones, coupled with the progress made in wireless communication, has made possible to perform mobile DM (Horovitz, Gaber, & Krishnaswamy, 2005). Consequently, considerable effort has focused on research and development in this area.

The main functions performed by these systems are: (1) data acquisition of biological sensors, (2) data analysis, to detect some abnormal situations, and if they are detected, then (3) generation alarm and notifying doctors so they can make the right decision (Goñi et al., 2009). Next, we will describe the experimental settings we worked on regarding the data description, data pre-processing, features selection, DM techniques used and results. We developed prediction models for dysfunction/failure of five organic systems - cardiovascular, respiratory, renal, coagulation and liver - as well as the outcome.

4.1 Data description and pre-processing

The data were gathered in the ICU of HGSA and were collected in the first five days of stay of thirty two patients.
They can be divided in five groups, described below.

1) **SOFA scores:**

In intensive care, there are some scores to assess severity of illness, like the Sequential Organ Failure Assessment (SOFA), which is commonly used in ICU on a daily basis to score the degree of dysfunction/failure of six organic systems – Cardiovascular, Respiratory, Renal, Liver, Coagulation and Neurological (Vincent et al., 1996). SOFA is scored in a scale from 0 (normality) to 4 (failure) for each organic system. In this experiment, we transformed the SOFA scores in binary variables, where 0 describes normality and 1 describes dysfunction/failure and comprises the original SOFA.

$$SOFA_{Cardio,\ Resp,\ Renal,\ Liver,\ Coagulat,\ neuro} = \{0,1\}$$

The variables required to calculate de SOFA scores derive from heterogeneous sources, with different frequencies, as shown in table 1.

Table 1: Data sources for sofa score calculation

SOFA	***Variables***	***Source***	***Frequency***
Cardiovascular	Blood Pressure	BM	Minute
	Dopamine, dobutamine, noradrenaline	EHR	Day
Respiratory	PaO_2/FiO_2	NR	Day
Renal	Creatinine	EHR	Day
Liver	Bilirubin	EHR	Day
Coagulation	Blood platets	EHR	Day
Neurological	Glasgow Coma Score	NR	Hour

In the data collecting phase, we have realized that the Glasgow Come Score is rarely registered, so it was not possible to generate models for the neurological system due to the missing data regarding its evaluation. As we can see in table 1, not only variables are registered in different sources, but also they have different frequencies. E.g. Blood Pressure is registered every minute, whereas creatinine is only measure and registered once a day. Taking this into account, for the construction of the final dataset, we transformed every SOFA score to an hourly register, for we are making predictions in an hourly basis.

2) **Critical events:**

Despite being commonly used, the SOFA score is controversial, for it can't predict the precisely individual outcome (Vincent et al., 1998). As a consequence, research has evolved towards the use of intermediate outcomes such as Critical Events (CE) (Silva, Cortez, Santos, Gomes, & Neves, 2008). CE was defined by a panel of experts and relates to four physiological variables – Blood Pressure (BP), Heart Rate (HR), Urine Output (UR) and Oxygen Saturation (O2). Whenever these parameters are out of normal ranges for 10 minutes, it is considered a CE for the corresponding parameter. They were calculated hourly and, subsequently, it was derived a new variable – Accumulated Critical Events (ACE) – to reflect the patients' clinical evolution/severity of illness.

3) **Case Mix =** {Age, Admission type, Admission from}

4) **Ratios =** {ACE_{BP}, ACE_{SO2}, ACE_{HR}, ACE_{Ur} }

5) **MEWS =** {MS_{BP}, MS_{HR}, MS_{RR}, MS_{Temp}, MS_{AVPU}}

The Modified Early Warning Score (MEWS) is a simple, physiological score that may allow improvements in the quality and safety of management provided to surgical ward patients (Gardner-Thorpe, Love, Wrightson, Walsh, & Keeling, 2006). This score uses data derived from four physiological parameters - Systolic Blood Pressure, Heart Rate, Respiratory Rate and Body Temperature - and levels of consciousness - Alert, Voice, Pain, Unresponsive (AVPU) - as shown in table 2. The MEWS (MS) uses a scale from 0 (normal) to 3 (extreme risk) to classify the danger of each score.

With the data collected in the ICU, we were able to calculate the score for only two variables for the MEWS: Blood Pressure and Heart Rate.

Table 2: variables and data sources for the mews calculation

Score	**Scale**	
	Source	***Frequency***
Blood Pressure	BM	Minute
Heart Rate	BM	Minute
Respiratory Rate	BM	Minute
Body temperature	BM	Minute
AVPU	NR	Daily

To use it in final dataset, with hourly data, it was calculated the number of occurrences for each score in each hour, i.e., how many minutes the values were out of normal range (danger) and its score.

In conclusion, the data required to calculate the variables derive from heterogeneous sources - Bedside Monitors (BM), Nursing Records (NR) and Electronic Health Records (EHR) - as shown in Table 2. By exploring different and incremental scenarios we intend to attest the importance of new variables, other than the critical events. The inclusion of ratios points to the severity of the patient's clinical evolution and the inclusion of SOFA scores relates to the multi-organ failure perspective.

4.2 Features - Incremental approach

For the prediction of each organic system and outcome, it was explored five scenarios regarding the inclusion of the variables mentioned before – M1, M2, M3, M4 and M5. Previous research (Villas Boas, et al., 2010) has shown that the inclusion of ratios and the SOFA scores of other organic system for the prediction of dysfunction/failure of some systems presented better results when comparing to the approach on the CE alone. Accordingly, we have added a new variable – MS - as input and compared the results to conclude if its inclusion leads to better results. Also, we have tried to use exclusively the MS to predict organ failure/dysfunction of the cardiovascular system. We used the MS only for the prediction of this system because the only available data relates to. Data for the other organic systems, as shown in table 3, was not registered.

M1 = {Case Mix, CE}
M2 = {Case Mix, CE, Ratios}
M3 = {Case Mix, CE, Ratios, SOFA}.
M4 = {Case Mix, CE, Ratios, SOFA, MEWS}
M5 = {Case Mix, MEWS}

4.2.1 Modelling

The DM techniques used for the experiments were Artificial Neural Networks (ANN), Decision Trees (DT), Regression and Ensembles. To avoid overfitting due to their biased distribution, we performed normalization by using the logarithmic function. We used fully connected multilayer perceptron's with 5 hidden neurons and logistic activation function. The training technique applied was the backpropagation algorithm. To assure statistical significance, 30 runs were applied to all tests. For the DT, we used Classification and Regression Tree (CART) algorithm. The splitting method used for partitioning the data was the Gini reduction. The default algorithm splits a node into 2 branches and, to avoid overfitting, the maximum number of branches from a node was set to 10 and the splits were evaluated as a reduction in impurity (Gini index). We used logistic regression since the targets have a binary measurement. The selection method used was the Stepwise. The ensemble method used a combined mode of the ANN, DT and Regression with the Mean probability function.

4.3 Results

Previous research has concluded that, for intensive care, is preferable to have models that favor sensitivity (Vilas-Boas, et al., 2010). Hence, the assessment of the models, presented in table 3, is made in terms of sensitivity, derived from the confusion matrices. For each system, we present the DM technique that had the best results for each scenario, in terms of sensitivity.

As we explained before, MEWS could only be used to relate to the cardiovascular system, so there are no results for the other systems regarding M4 and M5. Table 4 sums the results and presents, for each organic system, the best scenario and the technique with the best results, in terms of sensitivity.

Table 3: Results by scenario and technique for each target

M	Target in terms of sensivity (%)					
	Cardio	*Respirat*	*Renal*	*Liver*	*Coag*	*Out*
M1	DT 92.4	ANN 89.2	ANN 97.7	DT 94.3	ANN 94.2	ANN 98.3
M2	ANN 91.3	ANN 96.2	ANN 98	ANN 95.4	ANN 97.5	ANN 98
M3	ANN 93.4	DT 95	ANN 98.1	ANN 98.3	ANN 95.6	ANN 97.6
M4	ANN 73.2	-	-	-	-	-
M5	DT 90.0	-	-	-	-	-

Table 4: Best results for each target by scenario and technique

Target	Scenario	Sens	Technique
Cardio	M3	93.4	ANN
Respirat	M2	96.2	ANN
Renal	M3	98.1	ANN
Liver	M3	98.3	ANN
Coag	M2	97.5	ANN
Out	M1	98.3	ANN

5 DISCUSSION

This experiment focused on features selection and data pre-processing for the construction of the final dataset for building the DM models for prediction of organ dysfunction/failure. The experiments about the features selection had the objective of analyzing the impact on models' sensitivity of adding more variables. Accordingly, we conclude that adding the score MEWS not only did not result in better models, but actually worsened their sensitivity for all systems and outcome. E.g. prediction of the cardiovascular system has better results using CE, ratios and SOFA scores (93.4%), however, when we added the MEWS scores, sensitivity decayed to 73.2%. We can't fairly terminate that one shouldn't use MEWS to predict organ dysfunction/failure.

We also concluded that, regarding features selection, the variables to be included are highly dependent on the organic system; there isn't a scenario that is better for all systems, as shown in table 3. E.g. for the prediction of the liver system, it should be used CE, ratios and the SOFA of the other systems (M3), whereas for the respiratory system, the best results are achieved by not including the SOFA (M2). The models have achieved great results in terms of sensitivity and the best technique for all systems and outcome was the ANN.

The INTCare system is still being tested, however, in the future, the tasks regarding data preprocessing will be automatically performed by the application, with minimal human intervention and awareness, with an Electronic Nursing Record (ENR).

6 CONCLUSION AND FUTURE WORK

In this paper we presented the work undergoing the INTCare system, a real-time IDSS for intensive care. We developed DM models for prediction of organ dysfunction/failure and outcome and we presented the scenarios and techniques we experimented.

As shown in table 3, the inclusion of the new variable MEWS did not contribute to a better performance of the models in terms of sensitivity. However, one mustn't discard this line of investigation because in this experiment we only used two out of six scores for MEWS. It would be interesting to, in future research, rebuild the models with all the six variables of MEWS and compare the results. Also in future work, we will continue towards ubiquity and pervasiveness of the system, with the features presented in section 3 regarding DM and decision models and intelligent agents and some problems related to web traffic need to be studied and predicted (Piramuthu, 2003).

ACKNOWLEDGEMENTS

The authors would like to thank FCT (Foundation of Science and Technology, Portugal) for the financial support through the contract PTDC/EIA/72819/2006.

REFERENCES

Bergs, E. A. G., Rutten, F. L. P. A., Tadros, T., Krijnen, P., & Schipper, I. B. (2005). Communication during trauma resuscitation: do we know what is happening? *Injury*, *36*(8), 905-911.

Donchin, Y., & Seagull, F. J. (2002). The hostile environment of the intensive care unit. *Current opinion in critical care*, *8*(4), 316.

Fano, A., & Gershman, A. (2002). The future of business services in the age of ubiquitous computing. *Communications of the ACM*, *45*(12), 87.

Gardner-Thorpe, J., Love, N., Wrightson, J., Walsh, S., & Keeling, N. (2006). The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study. *Annals of The Royal College of Surgeons of England*, *88*(6), 571.

Giudici, P. (2003). *Applied data mining: Statistical methods for business and industry*: John Wiley & Sons, Inc.

Goñi, A., Burgos, A., Dranca, L., Rodríguez, J., Illarramendi, A., & Bermúdez, J. (2009). Architecture, cost-model and customization of real-time monitoring systems based on mobile biological sensor data-streams. *Computer Methods and Programs in Biomedicine, 96*(2), 141-157.

Horovitz, O., Gaber, M. M., & Krishnaswamy, S. (2005). Making sense of ubiquitous data streams - A fuzzy logic approach. In R. Khosla, R. J. Howlett & L. C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Pt 2, Proceedings* (Vol. 3682, pp. 922-928). Berlin: Springer-Verlag Berlin.

Hsu, J. (2002, 2002). *Data mining trends and developments: The key data mining technologies and applications for the 21st century*.

Kargupta, H. (2001, 2001). *CAREER: Ubiquitous Distributed Knowledge Discovery from Heterogeneous Data*.

Kohn, L. T., Corrigan, J., & Donaldson, M. S. (2000). *To Err Is Human: Building a Safer Health System*: National Academy Press.

Langheinrich, M. (2001). *Privacy by Design - Principles of Privacy-Aware Ubiquitous Systems.* Paper presented at the Proceedings of UbiComp 2001.

Menachemi, N., & Brooks, R. G. (2006). Reviewing the benefits and costs of electronic health records and associated patient safety technologies. *Journal of Medical Systems, 30*(3), 159-168.

Neaga, E. I., & Harding, J. A. (2005). An enterprise modeling and integration framework based on knowledge discovery and data mining. *International Journal of Production Research, 43*(6), 1089-1108.

Nigel, D. (2002). Beyond Prototypes: Challenges in Deploying Ubiquitous Systems, *1*, 26-35.

Orwat, C., Graefe, A., & Faulwasser, T. (2008). Towards pervasive computing in health care–A literature review. *BMC Medical Informatics and Decision Making, 8*(1), 26.

Piramuthu, S. (2003). On learning to predict web traffic. *Decision Support Systems, 35*(2), 213-229.

Saha, D., & Mukherjee, A. (2003). Pervasive computing: a paradigm for the 21st century. *Computer*, 25-31.

Santos, M. F., Portela, F., Vilas-Boas, M., Machado, J., Abelha, A., Neves, J., et al. (2009). Information Modeling for Real-Time Decision Support in Intensive Medicine. In S. Y. Chen & Q. Li (Eds.), *Proceedings of the 8th Wseas International Conference on Applied Computer and Applied Computational Science - Applied Computer and Applied Computational Science* (pp. 360-365). Athens: World Scientific and Engineering Acad and Soc.

Scicluna, P., Murray, A., Xiao, Y., & Mackenzie, C. F. (2008). Challenges to Real-Time Decision Support in Health Care. *Agency for Healthcare Research and Quality*.

Silva, Á., Cortez, P., Santos, M. F., Gomes, L., & Neves, J. (2008). Rating organ failure via adverse events using data mining in the intensive care unit. *Arificial Intelligence in Medicine, 43*(3), 179-193.

Varshney, U. (2007). Pervasive healthcare and wireless health monitoring. *Mobile Networks and Applications, 12*(2), 113-127.

Varshney, U. (2009). *Pervasive Healthcare Computing: EMR/EHR, Wireless and Health Monitoring*: Springer-Verlag New York Inc.

Vilas-Boas, M., Santos, M. F., Portela, F., Silva, Á., & Rua, F. (2010). *Hourly prediction of organ failure and outcome in intensive care based on data mining techniques*. Paper presented at the 12th International Conference on Enterprise Information Systems.

Villas Boas, M., Gago, P., Portela, F., Rua, F., Silva, Á., & Santos, M. F. (2010). *Distributed and real time Data Mining in the Intensive Care Unit*. Paper presented at the 19th European Conference on Artificial Intelligence - ECAI 2010.

Vincent, J. L., de Mendonca, A., Cantraine, F., Moreno, R., Takala, J., Suter, P. M., et al. (1998). Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. *Critical care medicine, 26*(11), 1793.

Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonca, A., Bruining, H., et al. (1996). The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive care medicine, 22*(7), 707-710.