Process model for data mining in health care sector

Diego Roa - María del Pilar Villamil Los Andes University Bogota,Colombia {df.roa34,mavillam}@uniandes.edu.co

ABSTRACT

This paper presents a process model to guide the data mining projects in the health care sector. The process model (PMH) is a specialization of CRISP-DM methodology and presents different issues associated to the data analysis and management. This proposal was validated in order to address real problems related to health care in Colombia. The results show that it is possible to establish new hypothesis about the clinical data, and revalidate these affirmations using the proposed process model.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: healthcare sector; D.2.8 [Data mining]: Process

Keywords

data mining, process model, healthcare, PMH

1. INTRODUCTION

The knowledge discovery process is one of the main factors to enhance competitiveness in organizations. The use of data mining techniques in this process is essential to analyse, understand and predict behaviours of an organization.

In the health care sector, there are many opportunities to apply data mining. Some of them are related to the improvement of the quality control in health care. In particular, analysis to detect and diagnose diseases, predict the responses of the organism to specific treatments and to identify epidemiological profiles, are relevant themes for the health care community.

There are many methodologies to tackle data mining opportunities such as CRISP-DM[2] or the virtuous cycle of data mining [8]. All of them are designed to improve the success of data mining projects. These methodologies are used in many sectors such as financial, pharmaceutical or health care industries. However, there are very specific characterisJuan Diego Arboleda Oracle Bogota,Colombia juan.arboleda.tabares@oracle.com

tics associated to these sectors that can be used to customize the process and to improve the quality and effectiveness of these kind of projects.

This paper presents a process model to guide the data mining process in the health care sector. This process model allows to reduce the costs and resources used in data mining projects. The process model was evaluated by analyzing 49,000,000 individual register of health care (RIPS) obtained from different sources, such as HMOs and the Minister of Social Protection in Colombia from 2003 to 2006. The evaluation goal was to compare treatments among Health Maintenance Organizations (HMOs), as well as verifying the compliance to standards of evidence gained from the scientific method (EBM- Evidence-based medicine); which will support the quality process of health services. The analysis was focused on hypertension diagnosis and allows to evidence similarities between the national guidelines and the health service. Specifically it was possible to identify the use of *captopril* an Angiotensin-Converter enzyme inhibitor medicament in the hypertension treatment. This medicament is cheaper according to other medicines of this type. One hypothesis is that it is pre-scripted for economical reasons. However, patients with this kind of treatment, returns to the healthcare institution with complications in the hypertension disease. This kind of complications increase the illness costs. Finally, validations about this process model enable new opportunities to establish public health policies in Colombia.

This paper is organized as follows. Section 2 describes problems related to the data mining process in healthcare. Section 3 presents Health care data management characteristics and the PMH the Process Model proposed for data mining in Healthcare sector. Section 4 exposes the validation method of the proposed process. Finally, Section 5 concludes the paper and presents other research issues.

2. DATA MINING AND HEALTH CARE

There are many studies that evidence the relevance of data mining techniques in the health care sector. These studies are associated to the treatment of patients and generally, to the identification of best practices in the treatments of specific diseases.

Some works such as [6, 16, 17, 15, 9] show different categories of problems related to the health care sector that are solved using data mining techniques. Some of them presents the use of association rules, sequential patterns or clustering in the prediction, analysis and monitoring of patient's treatments. On the other hand, there are studies such as [1, 7] that propose new algorithms to solve these issues.

Abidi and Stolba in [14, 10] describe the relevance of identifying clinical guides based on the individual registers of health care. These guides support medical tasks, increasing the quality of service of medical centers. However, these works are focuses in structuring clinical guidelines, and they do not emphasize in the methodology used to realize these projects.

Although some of the works such as [5] use the CRISP-DM methodology and others [11] the virtuous cycle data mining to improve the process quality, there are some characteristics associated to the specific domain that will be used to reduce the number of incidentals that may arise in a data mining project. These issues enhance the opportunity to use methodologies that explicitly include clinical concepts, problems related to the health care sector and moreover, that supports the selection process of data mining techniques based on the specific characteristics of this sector.

The afore mentioned issues motivate the realization of PMH, a specific process model for health care context, that will be presented in the section 3.

3. DATA MINING GUIDE FROM THE HEALTH CARE POINT OF VIEW

This section presents a brief description about health care context in subsection 3.1, with the purpose to highlight the opportunity to provide new guides associated to data mining applications to improve different kind of projects in health care. Subsection 3.2 describes a new process model to guide these projects.

3.1 Health care data management

The health care sector evidences challenges w.r.t. data management because of data characteristics such as volume, quality, availability, accessibility, and the relevance of the decisions involved during the process analysis.

Moreover, there are clinical guidelines that describe a set of steps to treat a specific disease. From these guides, it is possible to determine the service efficiency, the time involved in the treatment, and typical practices such as treatments and medications. This kind of information provides important elements to the decision's maker.

According to the volume of data analyzed, it is important to highlight that all clinical cases are relevant during process analysis. This is contrary to other sectors. In other domains a rule is meaningful when the support of the data is relatively high, whereas in the health care domain, analysis involving mortalities will be accounted for although the number of cases will not be significant, statistically speaking.

The ideas mentioned before motivates the development of the process model for healthcare (PMH), which is described in the next sections.

3.2 PMH overview

The knowledge in data mining projects frequently remains in few people like consultants and experts in specific domains. For this reason, the experiences and processes cannot be reusable and applied to similar projects in health care. Furthermore, it is necessary to know different guidelines, references and standards related to quality of service, with the purpose of understand the main characteristics of the health care domain.

As a result, this paper presents PMH (Process Model for the health care context). This process is an specialization of the CRISP-DM methodology proposed in Colombian health care context, based on the verification carried out in the pharmacological and non-pharmacological treatments in Colombia's health care institutions (IPS), through the application of data mining techniques on RIPS files. This process allows to reduce time and resources with respect to develop mining projects in this domain without a specific knowledge.

This guide proposes seven steps in an iterative way tacking into account health care context. The following paragraphs provide a description about the different steps involved in this process model. The numbers used in this description correspond to the number used in the figure 1.



Figure 1: Data mining methodology for health care sector

I. Scope Definition of the exercise. This step allows to define the business problem to be analysed. Several fundamental aspects must be clarified in this step: what, how, and why the assessment is done, as well as defining the criteria for the success of the exercise. Some of the typical questions proposed by experts are focused on problems to validate the effectiveness of pharmacological treatments in the emergency room according to IPS's best practices, others in the control and monitoring of chronic diseases according to the standards. In this step, these questions will be contextualized according to the service(hospitalizations, urgencies, procedures or medical appointments), and the diagnosis to be monitoring.

II. Selection of the reference guide. This step II consists on the selection of the reference guide(s) to evaluate the question proposed in the first step; for example the medical treatment used for a diagnosis. Currently, it is possible to use the expert advise to validate the quality of service in health care. On the other hand, there is specific manuals such as standards, protocols, and clinical guidelines proposed by governments and organizations that can be used as a reference guide.

The standards are evidence based on references used to evaluate the quality and performance of services, while protocols are documents that describe the rules of action depending on a specific circumstance [4]. Usually, protocols are specific documentation defined by each IPS. Also, the clinical guidelines are systematically developed statements to assist practitioners and patient decisions about appropriate health care for specific clinical circumstances [12].

About clinical guidelines, the World Health Organization (WHO) presents different guides related to the diagnosis, treatment and prevention of specific diseases such as obesity, malnutrition or diabetes. Furthermore, different countries have established national standards to treat a disease. For example, Colombia has the 412 resolution which suggests the set of activities and procedures that should be used in public health diseases [13].

Although clinical guidelines vary in content, they have essentially the following structure:

Clinical guideline structure
0. Authors
1. Introduction
2. Disease detection
3. Diagnosis
4. Classification and Tracking
5. Disease evaluation
6. Non-pharmacological treatment
7. Pharmacological treatment
8. Disease complications
9. Disease special situations
10. Hospital treatment
11. Emergency treatment
12. Clinical guidelines future review recommendations
13. Bibliography

In the structure above, the interest lies (mainly) in points 4 to 11. The quality control proposed is based on the comparison between treatments with a specific admission diagnosis and an established clinical guide diagnosis.

The suggestion is to choose the reference guide that has been established in the national policies or regulations, because international clinical guidelines may not have validity in a specific country, or may not be applied in certain IPS because of socio-economic or epidemiological factors.

III. Identification of information sources. This step consists of the identification of useful information sources according to the scope of the project and the selected reference guides. This identification depends on the data quality and availability. These issues will be tacking into account to decide the use of these sources during the analysis process.

In the healthcare sector, there are different sources that can be obtained and used for the development of data mining projects.

Generally, countries have an individual healthcare register corresponding to every hospitalization, urgency or procedure associated to a patient. For example, the United States has the Electronic Health Record (EHR) which includes demographics, medical history, medication and allergies, immunization status and observations (among others). On the other hand, Colombia has the Individuals Registers of Health Care (RIPS in Spanish), that provides information related to the delivery of health services and demographic variables.

There are other kind of information sources related to national behavior such as national survey or naming standards. Some national surveys contains demographic and health information that can be used to support the data mining process. Furthermore, the WHO defines the CIE10 standard. This standard defines the classification and organization of diseases based on a unique code that represents the category and the specific affection.

Some naming standards are associated to specific health interventions available for each country. For example, the Australian Classification of Health Interventions (ACHI) contains all the procedures that are realized by HMOs in the country. The Unique Procedures Classification in healthcare (CUPS in Spanish) is the Colombian classification system for this information.

IV. Selection and preparation of healthcare information. It is necessary to have the support of health experts to select the information that is highly relevant to solve the proposed problem in step I. Each disease presents different characteristics. For example, there are diseases like prostate cancer or pre-eclampsia in which sex is not a determining factor. The first affects men, and the latter applies only for pregnant women. In chronic diseases like hypertension, time is a significant variable. Its treatment is based on monitoring the patient periodically with formulated procedures and medicines according to a specific order and to patient's evolution over time. On contrary, appendicitis treatment is considered relatively short. In this case the time variable is not relevant. These considerations must be taken into account in selection step.

The selected information follow a data cleaning process. Generally, health information has problems related to data management such as replication of records. Moreover, medicines data management proposed a new challenge to data mining experts. Usually, this data does not have a standard to be filled. For example, the medicine "amoxycillin" can be filled as "amox" or "amoxycilin". In these cases, the similarity word analysis can be used to solve the issue.

The corresponding discretization of continuous data and the standardization of information must be made, necessary procedures for the execution of mining algorithms, which should be in line with own business rules of the selected diagnosis. For example, in the case of Alzheimer's disease, age categories should be created after 40 years, being consistent with the characteristics of vulnerable populations, and the evolution of the disease over time. Different from Appendicitis disease, where age ranges should be used much broader, since it is a disease that can occur at any age. The complexity of both the discretization and the standardization of data may depend largely on the amount of selected information sources and the absence of the use of ontologies for the unification and standardization of medical terms and concepts.

An statistical analysis of this step is necessary for physicians because it's important to know the data percentage that must be cleaned and the problems that arise the datasets.

V. Information adjustment and preliminary analysis. In this step is important to analyse the resulting dataset of the previous phase. This analysis concerns to identify the main characteristics of the data and the discovering of new variables that are relevant for an specific situation. In healthcare, variables such as hospitalization window, total cost of treatment and patient satisfaction are relevant in many situations. For example, to solve questions like, what is the most expensive treatment?, which is the one with the lowest satisfaction?, which one represents a lower rate of days of staying? these variables are relevant in this context.

The preliminary analysis is based on descriptive statistics. The objective is to review the statistical data distribution in order to avoid biased results. In this step can be identified how many men or women are involved in the dataset, how is the age distribution or if the treatment is based significantly on drugs rather than procedures. The health experts evaluate the results and if necessary, this step is review again.

VI. Selection and implementation of data mining algorithms. This step consists in the identification of data mining algorithms to achieve the objectives proposed. In healthcare, there are probable classifications of the mining problems. The first, is related to the analysis of treatments, which is feasible to predict the organism response against specific procedures. Next, the monitoring and evolution of patients in infectious and chronic disease. The latter, is based on verify the proper provision of health services. For example, for the last classification there situations in which is appropriate to know the percentage of compliance of treatments with respect to a clinical guidelines or which are the implications of meeting/failing the clinical guidelines in terms of costs. The data mining algorithms that are proposed are association rules, sequential patterns and clustering techniques to solve specific problems in healthcare sector.

Association Rules: this technique identifies the cause- effect relations between variables. It is possible to characterize an specific group using clustering techniques, and determine the behaviour of an specific cluster using association rules.

In healthcare, it's interesting to discover relations among events. For example, if a patient's disease evolves to a chronic phase, the probability to have a decease based on specific characteristics can be found. Moreover, In the context of quality control, the association rules allow us to discover rules that may or may not correspond to established treatments in clinical guidelines It is recommendable to use this technique in the diagnosis of acute illness. This kind of illness usually are applied during a patient's admission to the IPS and does not require periodic monitoring to ensure new procedures or drugs depending on the patient progress. Furthermore, association rules could be used in problems that analyze the treatment's behaviour.

The possible way to do in quality control in the health care it is as follows:

Two data sets are created, one for drug treatment information and one for non-drug information for each patient with the same admission diagnosis, and with the same method of admission. It is important that the information of the admission method is not mixed, since the in each treatments can be different. Thus, given an X diagnosis in hospitalization, we have the following data set of medications:

Diagnosis =X method of admission = Hospitalization			
ID item			
Patient ID + Admission ID	medications		
1C	M30, M90		
2A	M90, M70		
35	M40		
4T	M30, M90, M70		
5M	M90		
6P	M30, M90, M70		

Figure 2: medication dataset

For this example, an association rule would be: "In the pharmacological treatment of a patient with an X diagnostic, if medicines M30 and M90 are provided, then M70 medicine will also be provided."

This technique required some parameters to be customized: The support and the confidence. The first one is the proportion of transactions in the data set which contain the itemset:

 $supp(A) = occurrence(A) \div size(dataset).$

The second one is defined as the conditional probability (P(B|A)): the occurrence of A, given the occurrence of B:

$$conf(A \Rightarrow B) = supp(A \bigcup B) \div supp(A).$$

Sequential Patterns: This technique detects cause-effect relations considering the time periods in which transactions occurred. In health care context, there are many situations that involves periodic controls and patient's monitoring. As mentioned before, standards, clinical guidelines and protocols specify sequences of treatments that should be applied in a explicit situation. For this reason, it is feasible to compare the patient's procedures versus reference guides using sequential patterns. Chronic diseases like hypertension or diabetes requires that patients return to the IPS several times with the same diagnosis. This kind of problems involves many variables in each time period. The recommendation is to use sequential patterns to analyze the evolution of patient's health based on the treatments applied.

Based on the concept that, at any given time, a clinical event is the formulation of one or more drugs or procedures we propose the creation of two data sets, drugs and procedures datasets. In this case the records should be grouped by patient, ordering clinical events in ascending order by date and time.

	Diagnosis =Y method of admission = Ambulatory			
	Patient ID	Date and Hour of the clinical event	Procedures	
s1	1	JUN 25 10:00 A.M.	P30	
	1	JUN 25 11:30 A.M.	P90	
s2	2	JUN 24 05:10 P.M.	P10, P20	
	2	JUN 25 08:20 A.M.	P30	
	2	JUN 25 08:55 A.M.	P40, P60, P70	
s3	3	JUL 07 3:00 A.M.	P30, P50, P70	
s4	4	JUL 12 6:45 A.M.	P30	
	4	JUL 12 6:55 A.M.	P40, P70	
	4	JUL12 7:10 A.M.	P90	
s5	5	AGO 08 01:35 A.M.:	P90	

Figure 3: structure of a sequential pattern dataset

All clinical events from a patient arranged in order, can be seen together as a sequence, where each event corresponds to a set of drugs or procedures. Figure 4 shows the sequences found in previous patients.

Patient Id	Sequences of the Patient	
1	{{P30} {P90}}	
2	{{P10 P20} {P30} {P40 P60 P70}}	
3	{{P30 P50 P70}}	
4	{{P30} {P40 P70} {P90}}	
5	{{P90}}	

Figure 4: sequence of patients procedures

Clustering: to found groups of elements with similar characteristics. In healthcare, is very common to analyze populations based on specific characteristics. Nevertheless, it's possible to use this technique in the validation of right treatments to right people. The reference guides define specific treatments for people with Specific demographic characteristics. In this case, the use of clustering determines subsets based on procedures and demographic information.

In addition, there are situations in which data quality is poor. For this reason, its impossible to analyze particular issues in the dataset because of the confidence of data. The suggestion is to use clustering techniques to generalize the main characteristics of an specific group. To perform these analysis, a data set must be created which includes patient information such as gender, age, marital status and race, (among others), as well as drugs and/or procedures provided, grouped into treatments found in the previous section, as shown in figure 5:

As in previous algorithms, it is important to analyze the number of people supporting each cluster, before making any conclusions. It is also essential to understand, that a cluster represents a very small percentage of the population

	Profile	Pharmacological Treatment	Non Pharmacological Treatment
	Sex	Mx	Px
	Age		
	Marital		
	Status		
p1	F, 61, C	T3={M80, M20	P320, P150}
p2	M, 15, S	T2={M70, M10	P420}
рЗ	M, 68, D	T1={M20, M70	P150}
P4	F, 70, C	T3={M80, M20	P320, P150}
P5	F, 14, S	T2={M70, M10	P4 20}
P6	M, 55, C	T1={M20, M70	P150}
:			
<u>pn</u>	M, 17, S	M70, M10	P420

Figure 5: dataset for clustering technique

does not necessarily implies that should be discarded. It all depends on the clinical context that is being evaluated and the criteria of the medical expert.

VII. Result Validation and Impact Analysis. This final steps concerns to tunning up the data mining model based on the health expert feedback and the results achieved. A detailed study of the results is made by a board of medical experts in the area supported by a technical group of people to determine whether or not to repeat some of the above steps, possibly with a change in strategy or range, or by a refinement of the data used, to get specific conclusions.

4. VALIDATION

This proposal is validated based on the PMH process and the product obtained by applying this steps in the solution of a real problem in Colombia's health care sector.

Our proposal is a specialization of the CRISP-DM methodology. The CRISP-DM process has been validated in several domains and modified during more than a decade, based on the application of the process in many data mining projects. In that sense, we refined the CRISP phases in order to improve the knowledge and special issues of the health care sector, reducing the time and resources that have to be used for understand this particular domain.

The PMH process incorporates the diagnosis and reference guide selection in the business understanding phase of CRISP-DM, and presents the selection criteria for these steps. Furthermore, in the other phases is possible to understand the principal problems associated with data quality in health care, and to determine which algorithms should be used to solve several problems in this sector.

The next sections are focused on the product validation. It applies the PMH process in the quality control of Colombia's health care sector and validate the results based on the experts criteria. The development of this exercise consists in two iterations of the process. The specifics steps of PMH are described below.

4.1 **Problem description**

Hypertension is a chronic disease that affects 20% of people in the world. This is considered the first cause of morbility and the most representative disease related to cardiovascular affections. For this reason, the objective of this exercise was to evaluate the pharmacological and non-pharmacological treatment for this disease in Colombia.

4.2 First iteration

Step I consists on the description of the problem according to the context, related to the diagnosis and type of disease that is going to be tackled. In this case, it is relevant to analyze the characteristics of hypertension. This is a chronic disease, generally asymptomatic and it requires continuous medical assistance. The typical complications of hypertension are related to cardiac failures. This complications implies hospitalizations, urgencies and complex procedures.

Based on the health experts support, there are different reference guides related to hypertension, but its treatment may differ from one country to another. For this reason, the selection of the reference guide was focused on the "Clinical guideline for hypertension disease" proposed by the Association of Faculties of Medicine of Colombia [3].

As mentioned in section III, Colombia has the Individual Registers of Health Care. The objective of this data is related to bill the delivery of health services made by the Colombian's IPS. The information was collected from HMOs and the Minister of Social Protection from years 2003 to 2006 (49,000,000 of individual register of health care). Furthermore, the RIPS data are filled based on the CIE10 and CUPS standards. For this reason, these information is also taken into account.

According to the expert opinion, the relevant information for this analysis is described in table 2:

Variables
Principal diagnosis
IPS Identification
Date
Sex
Department
Type of medical service
Medicine
Procedures
Length of stay
decease diagnosis
related diagnosis

The statistical analysis results for the dataset shows that a 77% of the registers correspond to procedures, 17% are medical appointments and a 6% of the data to medicines prescriptions. Moreover, a 67% correspond to men and a 33% to women. On the other hand, a 16% of the patients return to the IPS for health controls.

In the pharmacological treatment of hypertension, we want to analyze the most relevant medication sequences. In this case, the time variable is highly relevant because we want to trace the prescription of medicines for a specific disease. For this reason, we used sequential patterns, according to the ideas presented in subsection 3.2. Our model was implemented using the IBM Intelligent Miner 8.1 data mining tool.

The figure 6 shows the results of the mining model.

SEQUENTIAL PATTERNS - "medicamentos-hiper-vuelve"			
Elle Sort	Eiter		
Sequence			
Support	Itemsets	-	
49.074	[MEDICAMENTO: CAPTOPRIL] [MEDICAMENTO: CAPTOPRIL]	>	
22.222	[MEDICAMENTO: CLORURO SODIO] [MEDICAMENTO: CLORURO SODIO]		

Figure 6: Sequential patterns results

The main result is the use of *captopril*. More than 40% of the patients were medicated with this medicine. Based on health experts opinions, the *captopril* is a medicine used in mono-therapy treatments and is prescripted for economical reasons. Other important conclusion is related to data quality of medicines; a specific naming standarisation was used to resolve the problem. In the same way, several manual process was realised to mitigate replication of records problem.

4.3 Second iteration

Based on the results of the first iteration, a second iteration of the steps suggested in PMH was performed. Therefore, the health experts proposed to analyze if the health system may incur in higher costs because of the prescription of *captopril* to the patients. Using the RIPS data, we want to determine the complications related to these patients.

According to the PMH, it is necessary to prepare the data that will be used by the model. For this reason, the data presents demographic information and relevant aspects about the evolution of a patient's disease.

In this case, new variables have to be include based on the health expert recommendation. This variables are associated to the evolution to a chronic phase in the clinical history of patients. For this reason, were introduced the date of the patient's complication, the associated diagnosis, the number of hospitalizations or procedures before and after the complication of hypertension. The next step of the methodology, proposes a preliminary analysis of the information. In this case, the information consists on all the records of the patients that have suffered this disease.

To describe the complication of a patient, the sequential clustering technique was used. As described in section 3,with this technique it is possible to find clusters of patients with similar sequences and characteristics.

The results of the second iteration shows that patients with similar sequences associated to the use of *captopril* have complications such as chronic cardiac failures, hypertensive crisis or heart attacks.

In general, it is important to analyze that the use of *cap-topril* is pre-scripted for economical reasons. This strategy is useful in a short term period, but in a long term, we can observe that patients with this kind of treatment, returns

to the healthcare institution with complications in the hypertension disease. This kind of complications increase the illness costs.

5. CONCLUSIONS AND FUTURE WORK

This paper proposes a process model to guide the data mining process in the health care sector. It suggests a set of iterative and facultative steps to improve the results of the mining process. This process model was evaluated using the analysis of quality of service for the treatment of hypertension in Colombia. The results shows that it is possible to establish new hypothesis about the datasets, and revalidate this affirmations using the proposed process model. At the same way, these results evidence some facilities provided to the data mining expert to guide their process, specially associated to the knowledge about healthcare context such as data sources, reference guides and data mining techniques.

An exhaustive validation of the process model is considered as future work, in terms of a formal comparison between the use of CRISP-DM and PMH. At the same time, new kind of question from the expert point of view will be interesting to resolve using this process model. In particular, the identification of an epidemiological profile for Colombian population.

6. ACKNOWLEDGMENT

The authors appreciate the support of Jose Abasolo, professor at Los Andes University, which provides the initial ideas and suggestions for the development of this article.

7. REFERENCES

- E. A.M. and H. Fu. Privacy preserving distributed learning clustering of healthcare data using cryptography protocols. In *Computer Software and Applications Conference Workshops (COMPSACW)*, 2010 IEEE 34th Annual, pages 140 –145, july 2010.
- [2] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000.
- [3] A. colombiana de Facultades de Medicina. Guía clínica para la hipertensión arterial. http://www.redsalud.gov.cl/archivos/guiasges/ hipertension_arterial_primaria.pdf.
- [4] eHow Health. Definition of clinical protocol.
- [5] X. Fang. Are you becoming a diabetic? a data mining approach. In Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on, volume 5, pages 18 -22, august 2009.
- [6] K. Harleen and S. K. Wasan. Empirical study on application of data mining techniques in health care. *Journal of computer science 2*, pages 194–200, 2006.
- [7] A. Li, S. Wang, H. Zheng, L. Ji, and J. Wu. A novel abnormal ecg beats detection method. In *Computer* and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, volume 1, pages 47 –51, february 2010.
- [8] B. M.J.A. and L. G.S. Mastering data mining. 2000.
- [9] L. N. Predicting the risk of future hospitalization. In Database and Expert Systems Applications (DEXA), 2010 Workshop on, pages 120 –124, september 2010.

- [10] S. N. and T. A. M. The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making, 2002.
- [11] L. Nevine and M. Malek. Data mining for cancer management in egypt case study: Childhood acute lymphoblastic leukemia. 2005.
- [12] N. I. of Health. About clinical practice guidelines.
- M. of Social Protection. Resolución 412. http://mps.minproteccionsocial.gov.co/pars/cajaherram/documentos/Biblioteca/CompendioNormativo/ resolucion_412_00.pdf, 2002.
- [14] A. S. S. Raza and A. S. Raza. A case for supplementing evidence base medicine with inductive clinical knowledge: Towards a technology-enriched integrated clinical evidence system. In *Proceedings of the Fourteenth IEEE Symposium on Computer-Based Medical Systems*, CBMS '01, pages 5–, Washington, DC, USA, 2001. IEEE Computer Society.
- [15] N. R. T. and P. Jian. Introduction to the special issue on data mining for health informatics. *SIGKDD Explor. Newsl.*, 9:1–2, June 2007.
- [16] van Driel M. A., C. K., K. P. P., L. J. A., and B. H. G. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet*, 11(1):57–63+, 2003.
- [17] C. Yu, P. L. Henning, C. W. W., and O. Jorn. Drug exposure side effects from mining pregnancy data. *SIGKDD Explor. Newsl.*, 9:22–29, June 2007.