Automatic Knowledge Discovery and Case Management: an Effective Way to Use Databases to Enhance Health Care Management

[1]Luciana SG Kobus, [2]Fabrício Enembreck, [1,2]Edson Emílio Scalabrin, [1]João da Silva Dias, [1]Sandra Honorato da Silva

Pontifical Catholic University of Paraná - Brazil

[1] Graduate Program on Health Technology
[2] Graduate Program on Applied Informatics

{kobus.l@pucpr.br, fabricio@ppgia.pucpr.br, scalabrin@ppgia.pucpr.br, jdias@voe.com.br, sandra.honorato@pucpr.br}

Abstract This paper presents a methodology based on automatic knowledge discovery that aims to identify and predict the possible causes that makes a patient to be considered of high cost. The experiments were conducted in two directions. The first was the identification of important relationships among variables that describe the health care events using an association rules discovery process. The second was the discovery of precise prediction models of high cost patients, using classification techniques. Results from both methods are discussed to show that the patterns generated could be useful to the development of a high cost patient eligibility protocol, which could contribute to an efficient case management model.

1 Introduction

The public and the supplementary Brazilian health program are strongly based on a health practice focused on curing. This fact leads to a high degree of complexity of procedures, to high costs of health care, to failure of the attendance of the real health clients' needs of health promotion and prevention of diseases, and difficulties on their access to health services. Such contradictory aspects interfere on the management of health organizations.

Case management is an opportunity of improving the population health condition. This model could be improved by the application of precise prediction models to identify a health system user that could become of high risk and high cost. The monitoring of this user could be the foundation to needed health care in the

right time, in an efficient way, in an attempt to avoid the development or worsening of a disease.

In this paper, two automatic knowledge discovery techniques were used on patient data. The first aimed to find rules that could show important relationships among variables that describe health care events. The second aimed to find precise prediction models of high risk and high cost patients. To improve the accuracy of the prediction model generated and diminish the negative impact obtained by sampling techniques, ensembles of classifiers were used. The use of these combining techniques was necessary because of the great amount of data. The combining of classifiers generated by samples of a same database could significantly improve the prediction's precision, even when the size of the samples is really small [10, 11]. Results from both methods are discussed, to show that the patterns generated could be useful to the development of a high cost patient eligibility protocol, as well as to the definition of an efficient and particular case management model for the population of the study. This paper is organized as follow. Section 2 shows the particularities of case management. Sections 3 and 4, respectively, show symbolical machine learning theoretical basis and methods. Finally, we present the conclusions from this study in Section 5.

2 Case Management

The case management primary goal is the search for benefits to the health system user and his family, as well for the health care providers and payers. This goals could be achieved by (a) the search for quality health care, in a way that the health care provided be appropriated and beneficial to the population; (b) the inpatient length of stay management; (c) the control of the utilization of support resources by the use of information systems based on protocols and decision support techniques; and (d) health care cost control assuring efficient results [6].

High risk patients correspond to almost 1% of a health care system population. However, this small part of users corresponds to 30% of the available resources utilization ([3, 4, 5]). The users which health care generates high costs to the health systems are those who present the most complex profiles by both clinical and psychosocial points of view. Near 45% of these patients have five or more diagnoses to describe their chronic condition, each one of these diagnostics could be the focus of a specific case management program [4]. However, the precise identification of a high cost patient is not a simple task, mainly if it is done without the help of suitable computational tools. This is why we chose to apply different symbolic machine learning techniques.

3 Symbolic Machine Learning and Meta-Learning

Symbolic learning systems are used in situations where the obtained model assumes a comprehensible shape. The induction of decision trees by ID3 system [7] and the production rules generation by decision trees [8] were important contributions by this field of Knowledge Discovery from Databases. More efficient versions of these algorithms were developed, like C4.5 and C5.0 [9].

Symbolic representation based on *association rules* are a powerful formalism that enables the discovery of items that occur simultaneously and frequently in a database. Each rule has a support. In [1] such support is defined as the relative number of cases in which the rules could be applied. In this study we won't explain how the methods work, because they are very well-known by the literature.

Some combining techniques allow that very precise prediction models could be built by combining classifiers generated by samples of training data sets. Such techniques usually use some heuristics to select examples and partition the dataset. Some studies show that the combining of classifiers generated from many database samples could significantly improve the precision of the prediction, even when the size of these samples is very small ([10, 11]). Two well-known ensemble techniques are *Bagging* e *Boosting*. The reader is invited to consult ([2, 10, 11]) for more details about them.

4 Method

The population of the study was the data from the users of Curitiba Health Institute (ICS), which is responsible for the health care of the Curitiba (Paraná, Brazil) City Hall employees and their families[1]. Initially, data related to the period of 2001 to 2005 were analyzed, in a total of 55.814 users and 1.168.983 entries. Considering that the ICS epidemiological profile is congruent with the one of Curitiba area, and the operational need to decrease the number of entries from the original database, two criteria of data selection were defined: (i) users with age equal or above 40 years old; and (ii) users that had, in their health care entries, at least one registration related to the cardiovascular diseases group of the international code of diseases version 10 (ICD-10).

After we applied the criteria, the initial sample decreased to 401.041 entries, referring to 8.457 users and 1.799 diseases' codes. These two last values will define, respectively, the numbers of lines and columns of the database that will be used ahead for the generation of association rules and prediction models.

The *first phase* aims to discover associations among procedures that generate a pattern which is indicative of high cost and high complexity and learn with this association to detect similar cases in the future. Apriori [1] algorithm was applied to discover such association rules.

1 Data utilization was authorized by both the Institute and the Pontifical Catholic University of Paraná Ethics on Research Committee, register number 924.

The amount of relevant rules after both analysis process (subjective and trust analysis) is of 18. Table 1 shows only 2 of them. The high association of cardiovascular procedures with emergency consultations could be related or to the serious condition of the user's health, or to the poor monitoring of the users with cardiovascular problems, once few procedures or exams are requested in a frequent way to the cardiovascular diagnostic.

The fact of an "emergency consultation" event is associated to a heart procedure shows the importance of establishing a monitoring protocol to users that were submitted to cardiovascular procedures. The outpatient following after heart procedures should be a priority not only for the user's health, but also to the proper management of health care providers and payers.

Table 1: Relevant rules after subjective and trust analysis of the significant events.

ID	Association Rules
R01	10,5% of the users presented in their historic the procedure MYOCARDIC REVASCULARIZATION; among these, 75% have the probability to present association with an EMERGENCY CONSULTATION
R10	11,8% of the users presented in their historic the fact of being from the male sex associated to the procedure of GLYCATED HAEMOGLOBIN; among these, 100% have the probability to present association with referential values of MYOCARDIC REVASCULARIZATION.

The rule R10 was considered relevant because it indicates a relationship between male users, the GLYCATED HAEMOGLOBIN procedure, and the MYOCARDIC REVASCULARIZATION procedure. Accordingly to the same specialists, this is important information to establish educational and preventive programs to users of the male sex. These 18 rules compose the first part of the decision support system to help improve the characterization of a user of the health system that should be managed.

The *second phase*: the training database preparation for obtaining the prediction models uses as entry the same data that were used in the first analysis and the same selection criteria. However, some attributes were removed by a filter process and other attributes were included by derivation (Table 2). Next section shows the application of different machine learning techniques to obtain high cost patients prediction models.

Table 2. Attributes added to training database.

Attribute	Description
procedure	Sequential number indicating the 1^{st}, 2^{nd}, …, n^{th} procedure of a patient.
nprocedure	Number of procedures of a patient.
distance1	Distance, in number of months, between the date of the first procedure, and the date of the n^{th} procedure of a patient.
distance2	Distance, in number of months, between the date of the last event and the date of the first event of a patient.
age	Age, in years, of a patient.
sex	Sex of the patient: {F,M}
class	Each class (CC1, CC2, CC3, CC4, CC5) represents a numeric interval where the values of the class $CC_i < CC_{i+1}$. The class CC_5 represents patients of higher cost. The determination of interval boundaries was done by an expert of the field.

Prediction Models Discovery: Using the prepared database accordingly to the method described in the last section, we considered that an important characteristic would be the comprehensiveness of the models. So, we opted to use as basis an algorithm that generates decision trees like C4.5 and J48 [12].

Samples of different sizes (5% and 10%) were extracted from the database (almost 400.000 entries). The sampling procedure used is the stratification with replacement (similar to that one used in Bagging). Next, the models produced by J48 algorithm, Bagging and AdaBoosting were compared among each other, using the cross validation procedure. Both Bagging and AdaBoosting algorithms implemented in Weka platform used the J48 algorithm as basis, and Bagging was configured to generate 10 classifiers.

Table 3: General Results with Cross Validation.

Algorithm	**Sample 5%**	**Sample 10%**
J48	73.78 ±1.02	83.51 ±0.64
Bagging	77.54 ±0.92	87.15 ±0.57
Boosting	81.56 ±0.79	91.30 ±0.59

Results from the Discovered Classification Models: Table 3 illustrates the accuracy of the algorithms. On can notice that the standard deviation is quite small. This shows that the distribution of the samples has a statistical correspondence to the original data. With the results showed in Table 3 we can observe that Bagging and Boosting methods improved the percentage of properly classified examples, and the last one is the best algorithm for the experiments done. From Table 3, we can state with statistical significance that Bagging is better than J48 algorithm and that Boosting is better than Bagging for the present problem. This happens to both samples. It was also possible to observe that as larger the sample is better is the prediction rate. So, we conclude that ensemble of classifiers techniques could be effective in situations where available data correspond only to a small part of the *tuple* space.

Subjective Evaluation of Interesting Patterns: The subjective evaluation of the obtained patterns is part of the quality evaluation of the obtained rules, accordingly to the specialist's points of view. Table 4 presents the first three rules obtained, to make the explanation easier due to the limitations of space in this paper. By now, we can observe that all rules pointed out that a patient considered as of high cost (CC5) is the ones that had a high number of procedures utilization within a period of 30 months.

Table 4. Sample of obtained rules.

R1:	procedure > 47 & distance2 <= 30	&	nprocedure > 160: CC5 (3257.0)
R2:	nprocedure > 273 & distance2 > 31 &		distance2 <= 32 & age > 40: CC5 (940.0)
R3:	distance2 <= 31 & procedure > 50	&	nprocedure > 133 & age > 61: CC5 (530.0/1.0)

Analysis shows coherence with results that arose from the rules. This can be

confirmed by the first rule (R1), where the patient of high cost presents a high procedure utilization within a short period of time. In general, we could observe that the frequent utilization of the health care services leads to high cost. This could be easily observed, because the number of procedures was high in all 20 first rules. The lower absolute number of procedures was 78 (R18), which is still a high number. This number is worsened if we consider the short time in which these procedures occur. In the case of rule 18, the period is less than 20 months. This set of rules is the second part of the decision support system to identify the health system user to be managed and it help to predict if a patient will become of high cost or not.

5 Conclusions

It is fundamental that health care providers include intervention protocols and case management in their practice. Then, health care personnel, responsible for the patients' orientation could do their jobs in the most rental manner as possible, focusing on quality aspects of health care for that user already identified with some degree of risk. In this paper, machine learning and data mining techniques were used to help this task. It was observed that ensemble of classifiers could increase the trustiness of the prediction model generated, and decrease the negative impact obtained with the use of sampling techniques, generating a high cost patient's prediction model with accuracy of 90%. This model is very useful for the development of an eligibility protocol of high cost patients, as well for the improvement of an efficient and individualized management model for the population. By the other hand, the discovered associations between procedures and pathologies allow that management protocols could improve health care and direct resources for prevention and education, decreasing the amount of high cost patients in a medium and long period of time.

References

1. Agrawal R, Imielinski T, Swami A, Mining association rules between sets of items in large databases, In Proceedings ACM International Conference on Management of Data (SIGMOD), 1993, pp. 207-216.
2. Breiman L, Bagging Predictors, Machine Learning, n. 2, vol. 24, pp. 123-140, 1996.
3. Crooks P, Managing high-risk, high cost patients: the Southern California Kaiser permanent experience in the Medicare ESRD Demonstration Project, The Permanent Journal. v.9, n.2, 2005.
4. Forman S, Targeting the Highest-Risk Population to Complement Disease Management, Health Management Technology. v. 25, n. 7, Jull, 2004.
5. Knabel T, Louwers J, Intervenability: another measure of health risk, Health Management Technology. v.25, n.7, July, 2004.

6. May CA, Schraeder C, Britt T, Managed care and case management: roles for Professional nursing, Washington: American Nurses Publishing; 1996.
7. Quinlan JR, Induction of decision trees, Machine Learning vol. 1, Kluwer Academic Publishers, pg. 81-106, Netherlands, 1986.
8. Quinlan JR, Generating Production Rules from Decision Trees, In Proceedings International Joint Conference on Artificial Intelligence (IJCAI), pp. 304-307, 1987.
9. Quinlan JR, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, 1993.
10. Shapire RE, The Boosting Approach to Machine Learning: An Overview, MSRI Workshop on Nonlinear Estimation and Classification, 2002.
11. Ting KM, Witten IH, Stacking Bagged and Dagged Models, Proceedings International Conference on Machine Learning (ICML), 1997: 367-375.
12. Witten IH, E. Frank, Data Mining, Morgan Kauffman Publishers, San Francisco, USA, 2000.