journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Contrasting temporal trend discovery for large healthcare databases

Goran Hrovat[a,*], Gregor Stiglic[a,b], Peter Kokol[a,b], Milan Ojsteršek[a]

[a] Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova ulica 17, 2000 Maribor, Slovenia

[b] Faculty of Health Sciences, University of Maribor, Žitna ulica 15, 2000 Maribor, Slovenia

ARTICLE INFO

Article history:
Received 23 April 2013
Received in revised form
4 August 2013
Accepted 9 September 2013

Keywords:
Data mining
Decision support
Trend discovery

ABSTRACT

With the increased acceptance of electronic health records, we can observe the increasing interest in the application of data mining approaches within this field. This study introduces a novel approach for exploring and comparing temporal trends within different in-patient subgroups, which is based on associated rule mining using Apriori algorithm and linear model-based recursive partitioning. The Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality was used to evaluate the proposed approach. This study presents a novel approach where visual analytics on big data is used for trend discovery in form of a regression tree with scatter plots in the leaves of the tree. The trend lines are used for directly comparing linear trends within a specified time frame. Our results demonstrate the existence of opposite trends in relation to age and sex based subgroups that would be impossible to discover using traditional trend-tracking techniques. Such an approach can be employed regarding decision support applications for policy makers when organizing campaigns or by hospital management for observing trends that cannot be directly discovered using traditional analytical techniques.

1. Introduction

Applications where trend-discovery is applied have become important tools for helping companies analyze information and providing support when making decisions mainly for the purpose of reducing business costs. In addition, more and more data is stored by healthcare organizations and available within immense data repositories on a daily basis. Extensive information is buried within large amounts of available clinical data that could be used to enhance new knowledge [1–3]. Hence, new techniques need to be applied within knowledge discovery systems in order to take advantage of such an amount of data [4]. This paper presents a novel approach that exploits larger medical datasets in order to discover trends in healthcare. Whilst fully featured medical datasets are difficult to obtain due to privacy issues, we were able to use datasets without attributes that could identify individual patients. During this study we consulted the largest dataset in this area – i.e. the Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality. The dataset contains hospital discharge records for a stratified sample of approximately 20% of US community hospitals. In this study, we used data from the years 2000 to 2009, where each year includes approximately 7 million discharge records.

The aim of our work was to identify significant subgroups (e.g. males aged between 18 and 50) from these large datasets, and to comprehensibly show trends for each significant subgroup. The subgroups were structured within a tree structure

* *Corresponding author*. Tel.: +386 41 947746.
E-mail address: goran.hrovat@uni-mb.si (G. Hrovat).

0169-2607/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2013.09.005

where, within the leaves, graphs were generated representing the trends. From such trees the trends for rules can be easily employed by policymakers and insurance companies to target their campaigns [5,6] or activities toward different age or gender groups. The meaning of a term subgroup here is different from that in pattern mining subgroup discovery, where aim is to find subgroup for which the distribution of a single target variable varies from its distribution in the whole database [7].

2. Hospital discharge dataset

The NIS dataset [8] contains hospital discharge records for a stratified sample of approximately 20% of US hospitals. This new dataset is available every year and consists of approximately 7 million discharge records. Each record contains up to 126 attributes comprising the personal characteristics of the patient including, age, gender, race, patient's county of residence; administrative information including admission month, year, admission type; and medical information including up to 15 diagnoses and up to 15 procedures. During our experiment, records from 10 consecutive years were used, namely 2000 to 2009, where we selected only those patients aged 18 or over, resulting in a final dataset of 65,308,185 discharge records. Diagnoses and procedures were coded according to the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes. An ICD-9-CM contains codes for diagnoses, where taxonomy containing five-digit codes is applied. The first three digits represent the general diagnosis and are followed by two additional digits describing a more specific subgroup of the general diagnosis. Additionally, the ICD-9-CM contains codes for procedures, where taxonomy of four-digit codes is applied and the first two digits represent the general procedure, followed by two digits describing a more specific subgroup of the general procedure.

3. Methods

3.1. Association rule mining

Association rules were used for discovering interesting relations between diagnoses within NIS datasets. For mining association rules we used one of the more popular algorithms called Apriori [9,10]. It was developed by Agrawal and Srikant for mining rules on large datasets consisting of transactions, where each transaction contains a set of items. In our case the discharge records were treated as transactions and attributes (e.g. age, gender, admission month, diagnoses) were considered as items. During the first stage Apriori finds frequent itemsets where the user specifies a minimum support threshold. Generated and pre-defined minimum confidence was considered from these itemsets' rules. The discovered rules followed the form $X \Rightarrow Y$, where X consists of one or more items and Y of only one item. $X \cap Y = \varnothing$. X is called an antecedent or the left-hand side of the rule (LHS) whilst Y is called the consequent or right-hand side of the rule (RHS). For each rule, measures of interest are calculated, to allow for the ranking of often huge sets of discovered rules [11]. The most common measures of interest included support [9], confidence [9], and lift [12], however many other interesting measures could be applied (e.g. all-confidence [13], χ^2 [14]). Support of a rule is defined as $supp(X \Rightarrow Y) = P(X \cup Y)$, the fraction of transactions within the dataset that contain the itemset $X \cup Y$. Confidence in a rule is defined as $conf(X \Rightarrow Y) = supp(X \Rightarrow Y)/supp(X)$. Confidence can also be estimated using $conf(X \Rightarrow Y) = P(Y|X)$. The lift of a rule is defined as $lift(X \Rightarrow Y) = P(X \cup Y)/(P(X) \cdot P(Y)) = conf(X \Rightarrow Y)/supp(Y) = conf(Y \Rightarrow X)/supp(X)$. Lift measures the dependence between X and Y. X depends on the absence of Y when lift is below 1. X and Y depend on each other when lift is greater than 1 and X and Y are independent when lift is 1. The χ^2 measure of interest regarding a rule is defined as $\chi^2 = (lift - 1)^2 \cdot supp \cdot conf/((conf - supp) \cdot (lift - conf))$.

We performed association rule mining in statistical open-source language R [15] using package arules [16] where all the functionalities of the original Apriori algorithm were implemented.

3.2. Model-based recursive partitioning

Model-based recursive partitioning, as introduced by Zeileis et al. [17], was used for building a regression tree based on linear regression and splitting with respect to partitioning variables. The regression tree was built in the following order:

1. A linear model was estimated for all observations.
2. Those parameters, estimated using the objective function, were tested for instabilities within all partitioning variables. The most significant partitioning variable, with a p value less than the chosen alpha (0.05), was selected for further computation of the split points. The empirical fluctuation process [18] was calculated, in order to assess parameter instability. In the case of numerical variables, the maximum of the squared L2 norm of the empirical fluctuation process scaled by its variance was performed, whereas χ^2 statistics was used for categorical variables [19].
3. A split point was computed and a binary node, pointing toward two child-nodes, was created.
4. Steps 1–3 were repeated for all the child nodes. Recursion stopped once the p value became greater than the alpha or the number of observations within each node was lower than the user-selected value–this step is also called a pre-pruning of the tree. From the comprehensibility point of view, it is important to choose an appropriate alpha value that directly controls the size of a tree. The same applies to the argument min-split that controls the minimal number of samples within a node.

Model-based recursive partitioning in R was applied using the party package [20].

3.3. Discovery of temporal trends

In this subsection we introduce a novel approach for discovering temporal trends. We show that association rule mining in combination with model-based recursive partitioning can be used to discover trends in healthcare.

In our experiment only age, gender, admission month, year, and (i.e. in all up to 15) diagnoses were used from the