MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-1072

Restricted Versus Unrestricted Search Space: Experience from Mining a Large Japanese Database

Hendrik Nienhoff[a], Ursula Huebner[a], Andreas Frey[b], Mareike Przysucha[a], Michio Kimura[c]

[a] *Health Informatics Research Group, Osnabrueck University of Applied Sciences, Germany*
[b] *Nuertingen-Geislingen University, Germany*
[c] *Hamamatsu University School of Medicine, Japan*

Abstract

The aim of this study was to investigate whether standard Big Data mining methods lead to clinically useful results. An association analysis was performed using the apriori algorithm to discover associations among co-morbidities of diabetes patients. Selected data were further analyzed by using k-means clustering with age, long-term blood sugar and cholesterol values. The association analysis led to a multitude of trivial rules. Cluster analysis detected clusters of well and badly managed diabetes patients both belonging to different age groups. The study suggests the usage of cluster analysis on a restricted space to come to meaningful results.

Keywords:

diabetes mellitus; data mining; association analysis; cluster analysis.

Introduction

Data mining methods have become the core of discovering knowledge in large datasets (Big Data) [1] and have been applied in medicine [2] [3]. The aim of this study was to evaluate the usefulness of different standard data mining methods for finding useful patterns in a dataset of diabetes patients.

Methods

Based on a clinical data warehouse at the Hamamatsu University hospital, 26,890 patients were analyzed to extract patients diagnosed with diabetes type 2 (ICD10-Code 'E11'). After extensive data transformation and cleansing, association and cluster analyses were performed (using the WEKA Tool) to detect assocations of co-morbidities within the group of diabetes patients and to investigate if diabetes patients could be clustered by age, blood sugar and cholesterol values into different sub-groups. Both analyses belong to the unsupervised and describing methods of data mining [1]. Whereas the association analysis was performed on a search space of 73 variables, the cluster analysis was restricted to three variables.

Results

Using similar minimum support and confidence values as [2], 45,314 association rules were identified out of which the five strongest are shown in Table 1, ranked by confidence. Clustering with k=4 clusters showed that high levels of blood sugar went along with high cholesterol levels, each for younger and for elderly patients (Tab. 2).

Table 1– Five strongest association rules identified (n=1,339)

Rule (X→Y)	n	Supp.(X)	Conf. X→Y
H26 → H52	311	0.23	0.95
I20 → I10	320	0.24	0.84
I50 → I10	458	0.34	0.74
I50 → E78	458	0.34	0.68
E78 → I10	735	0.55	0.66

E78: Hypercholesterolaemia; H26: Other cataract; H52: Disorders of refraction; I10: Essential (primary) hypertension; I20: Angina pectoris; I50: Heart failure

Table 2– 4-Cluster centres for low-density lipoprotein

		Age		HbA_{1c}		LDL	
Clstr	n	mean	σ	mean	σ	mean	σ
1	135	68.7	7.5	8.6	1.2	85.4	25.8
2	273	75.1	5.2	6.6	0.7	60.4	17.4
3	191	56.2	8.1	6.3	0.7	66.6	19.0
4	63	39.9	9.3	8.6	1.8	89.2	26.5
total	**662**	**65.0**	**13.3**	**7.1**	**1.4**	**70.1**	**23.6**

Discussion

The association rules identified are similar to the results of Kim et.al. [2], who also detected primarily trivial rules. The cluster analysis showed that patients could be separated into two groups of diabetes patients with well or badly managed longterm blood sugar values. Well managed patients also had lower LDL values. These findings were independent of age. In summary, restricting the search space by knowledge yielded more useful clinical patterns. Data analysis was only possible after elaborate and time-consuming preprocessing.

References

[1] Maimon O and Rokach L. Data mining and knowledge discovery handbook. New York: Springer, 2005.

[2] Kim H, Shin A, Kim A and Kim Y. Comorbidity Study on type 2 diabetes mellitus using data mining. Korean Journal of internal medicine 2011; 27(2): 197-202.

[3] Valent F, Tillati S and Zanier L. Prevalence and comorbidities of known diabetes in northeastern Italy. Journal of diabetes investigation, 2013; 4(4): 355-360.

Address for correspondence

Hendrik Nienhoff, Rheiner Landstr. 35, 49078 Osnabrueck, Germany
Tel: +491786885188 E-mail: hendrik@nienhoff.com