MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Towards Identification of Prognostic Subgroups in Rheumatology by of Modeling Expert Knowledge in Knowledge Discovery Procedures

Örjan Dahlström[a, b], Toomas Timpka[c, d], Ursula Hass[e], Thomas Skogh[f], Ingrid Thyberg[f]

[a] *The Swedish Institute for Disability Research, Linköping and Örebro University, Sweden*
[b] *Department of Behavioural Sciences, Linköping University, Sweden*
[c] *Section of Social Medicine, Department of Health and Society, Linköping University, Sweden*
[d] *Department of Computer and Information Science, Linköping University, Sweden*
[e] *Centre for Medical Technology Assessment, Department of Health and Society Linköping University, Sweden*
[f] *Division of Rheumatology, Department of Molecular and Clinical Medicine, Linköping University Sweden*

Abstract

The aim of this study is to examine whether or not heuristic modeling of physicians' expert knowledge can improve standardized data mining methods in prognostic assessments of patients with rheumatoid arthritis (RA). Five consultants in rheumatology were interviewed and their experiences of clinical test values for RA patients, as distinguished from corresponding values for healthy reference populations, were modeled in a knowledge engineering step. The results were used in K-mean-clustering to determine prognostic subgroups. Modeling experts' knowledge was found promising for semi-automatic data mining in the chronic disease setting. Further studies using categorical baseline data and prospective outcome variables are warranted.

Keywords:

knowledge engineering, medical decision making,
knowledge discovery in databases,
clinical decision support systems,
semi-automated data mining, rheumatoid arthritis

Introduction

Chronic diseases constitute a large part of the disease burden in western countries. Early identification of patients with chronic disease at risk of progression is essential, especially when early intervention can induce remission or retard the disease process. Identifying subgroups of patients with a less favorable prognosis at an early stage is therefore both a key clinical and public health priority. A specific type of decision support is assistance with knowledge discovery in databases (KDD) [1]. The aim of this study is to examine whether or not heuristic modeling of physicians' expert knowledge can improve standardized data mining methods in prognostic assessments of patients with RA. The starting point for the heuristic modeling is that the clinical test values for patients with a specific chronic disease differ from those of healthy reference populations.

Materials and methods

The research was performed in three steps. In a knowledge engineering step, physicians' experiences from the clinical use of four variables included in the evaluation of RA patients were elicited. In the second step, these data were analyzed and represented in a model that was implemented in an algorithm for the determination of prognostic groups. For comparison, the same algorithm was also used to develop groups based on the crude data set. In the final step, prognostic group validation, the model was validated by comparing physicians' global assessment of disease activity scores (PGA (scores 0–4, where 0 corresponds to no activity and 4 represents high activity)) for the patient groups identified with and without the model.

The knowledge engineering was based on data from a prospective multi-center study (TIRA) where 320 patients diagnosed with RA were diagnosed early and included during 1996 – 1998 [].The interval-/ratio variables in the database were especially interesting for this study, because they could be used in the standard method of K-mean clustering. The set of variables used here included the erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), the 'number of swollen joints' and the 'number of tender joints'.

To find different subgroups of patients, the method of K-mean clustering was used on crude data, and then with two kinds of transformed data:

- Model A – Using the transformed data.
- Model B – Using the transformed data but with a predefined group of patients with values below the lower limit values (x_0) for both ESR and CRP.

To keep the results on a clinically comprehensive and not too complex level, the number of prognostic subgroups was limited to a maximum of four groups. The case with two subgroups was found trivial. Further cluster analyses were therefore restricted to three and four subgroups.

Results

Determination of prognostic groups

Running K-mean-clustering using three clusters resulted in a 'low value group' (low values for both ESR and CRP), a 'medium value group' (medium values for both ESR and CRP), and a 'high value group' (high values for both ESR and CRP).

Running K-mean-clustering using four clusters made the situation more complex. As in the case of three clusters, a 'low value group' and a 'high value group' appeared. There were also two 'medium value groups'. Despite minor differences in sizes and medium values for the clusters in the different model(s), the results were quite similar in this respect.

Prognostic group validation

The three patient subgroups identified using model B on the basis of ESR and CRP differed significantly with regard to PGA scores (Table 1). In comparison, the analyses based on the crude data identified only significant differences between the two extreme groups. In the case of four clusters, no significant differences were found in the crude data model with regard to PGA. Model A and Model B, however, both allowed the identification of significantly different groups based on ESR and CRP (Model A: 1 significant difference, Model B: 3 significant differences).

Table 1 – Validation of derived models A and B against crude data using physicians' global assessment scores (PGA) as proxy for patient outcome. Mean values (95% confidence interval) are displayed for each cluster (three and four clusters, respectively)

	Crude data		Model A		Model B	
3 Clusters	N	PGA	N	PGA	N	PGA
1	188	1.89 (1.78-2.01)	154	1.84 (1.71-1.96)	59	1.56 (1.37-1.76)
2	70	2.10 (1.93-2.27)	72	2.04 (1.87-2.21)	149	1.99 (1.87-2.11)
3	21	2.33 (1.89-2.77)	53	2.30 (2.08-2.53)	71	2.30 (2.12-2.48)
4 Clusters						
1	171	1.88 (1.76-2.00)	146	1.82 (1.69-1.95)	59	1.56 (1.37-1.76)
2	36	2.19 (1.98-2.41)	31	2.16 (1.95-2.38)	110	2.00 (1.86-2.14)
3	51	2.00 (1.77-2.23)	62	2.05 (1.86-2.23)	58	2.07 (1.87-2.27)
4	21	2.33 (1.89-2.77)	40	2.30 (2.02-2.58)	52	2.31 (2.08-2.54)

Discussion

A knowledge engineering approach where heuristic expert knowledge was used to transform data showed to be more efficient for identifying prognostic subgroups than K-mean clustering of crude data. The resulting subgroups passed the first 'objective gateway,' which means that they differed significantly when using basic statistical criteria for cross-sectional evaluation, compared to the analysis of crude data. The PGA used for evaluation is the variable available in the TIRA database that was the closest proxy to patient outcomes from a clinical point of view. Within this limitation, we found that the model representing physicians' expert knowledge in chronic disease was more efficient in identifying prognostic subgroups than standard procedures based on crude data. Further studies using categorical baseline data and prospective outcome variables are warranted. This effort will be implemented in the Swedish TIRA project using more comprehensive clinical data sets.

References

[1] Kurgan LA, and Musilek P. A survey of Knowledge Discovery and Data Mining process models. *Knowledge Engineering Review* 2006;21(1):1-24

[2] Thyberg I, Hass UA, Nordenskiöld U, and Skogh T. Survey of the use and effect of assistive devices in patients with early rheumatoid arthritis: a two-year followup of women and men. *Arthritis Rheum* 2004;51(3):413-21.

Address for correspondence

Örjan Dahlström, Linköping University, Department of Behavioural Sciences, SE-581 83 Linköping, Sweden.
orjda@ibv.liu.se