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

Time-Series Data Analyses for Healthcare-Data-Mining Based on a Personal Dynamic Healthcare System

Hiroshi Takeuchi[a], Yukari Ikeda[b], Naoki Kodama[a]

[a]*Department of Healthcare Informatics, Takasaki University of Health and Welfare, Gunma, Japan*
[b]*Ota General Hospital, Gunma, Japan*

Abstract

The healthcare-data-mining presented here extracts personally useful information, such as rules and patterns concerning lifestyles and health conditions, from daily time-series personal health and lifestyle data stored on a personal dynamic healthcare system by using mobile phone and web technologies. This paper focused on time-series data analyses in the healthcare-data-mining process. Interesting results concerning the relation between soybean protein ingestion and body-fat percentage were obtained.

Keywords:

personal healthcare, Internet, data mining,
time-series data

Introduction

We have developed a personal dynamic healthcare system (PDHS) utilizing the Internet. It enables time-series daily-health and lifestyle data to be stored in a database by using mobile phone and web technologies. Users of this system can input their daily data through a mobile phone and can transfer this data to a web-application server via the Internet. The web-application server provides a data mining service and notifies users, through a mobile phone, of important rules relevant to their health and lifestyle data. Here we report details of the time-series data analyses for the healthcare data mining and also present interesting results concerning the relations between lifestyle and health data for a volunteer user.

Methods

Healthcare-data-mining algorithm consists of two processes.

First process

We determined effective input fields by pre-examining the correlation between time-series lifestyle data and health data of interest as shown in Figure 1, where e is the given lifestyle data and h is the given health data. Here h_n is the value of h at date n and e_i is the value of e at date i. Assuming that lifestyle data affect health data with some delay, we defined a retarda-tion parameter s (=n-i).

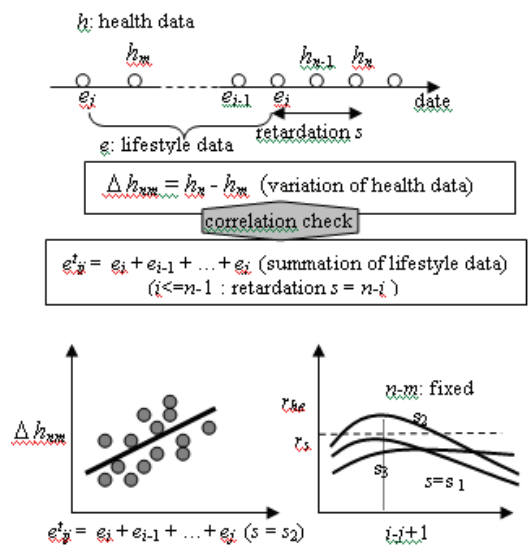

Figure 1 - Time-series data analysis

The correlation between Δh_{nm} and e^t_{ij} in time-series data was checked by changing n-m, i-j and s as parameters and calculating Pearson's product-moment correlation coefficient r_{he} for each (n-m, i-j, s) set. Here,

$$r_{he} = s_{he} / s_h \cdot s_e \tag{1}$$

where s_{he} is the covariance and s_h and s_e are respectively the standard deviation of Δh_{nm} and the standard deviation of e^t_{ij}. If, as shown in Figure 1, for some (n-m, i-j, s) set the r_{he} is larger than the threshold value r_s, then e^t_{ij} is selected as an input field. Actually, an input field is defined according to the values of i-j and s at which r_{he} becomes largest.

Second process

The second process is a rule induction from a number of time-

series data sets consisting of defined input fields and an output (target) field. An example of data set and a rule induction algorithm are shown in Figure 2.

Example of data set — input fields Y — output field X

date	Energy expenditure 5 (kcal)	Ingestion at meal 5 (kcal)	Alcohol intake	Body-fat index
10/06	1000	12000	much	higher
10/07	1200	11000	little	lower
10/08	800	13500	much	higher
……	……	……	……	……

Rule: If $Y = y$, then $X = x$ with probability p.
Y: input field, y: condition X: output field, x: value

Rules having higher-J-measure survive in the rule table

J-measure:

$$J(x|y) = p(y)\left(p(x|y)\log\frac{p(x|y)}{p(x)} + (1-p(x|y))\log\frac{(1-p(x|y))}{(1-p(x))}\right)$$

$p(y)$: probability of the rule's antecedent matching an example from the data set
$p(x)$: probability of the rule's consequent matching an example from the data set
$p(x|y)$: conditional probability of the rule's consequent conditioned on the antecedent

Figure 2 - Algorithm for rule induction

Input fields in the example of data set are defined by the first process. Energy expenditure 5, for example, is the summation of energy expenditure for 5 days. Output field (in this case, body-fat index) values are categorized into three classes, each having a symbolic value such as "higher", "moderate", or "lower". Border values used for this classification are determined so that the data frequency is similar in each class. In this rule induction algorithm, J-measures are estimated for all possible rules, and rules having higher-J-measure survive.

Results

In the automated healthcare-data-mining, the two processes described above are executed successively. The first process itself (time-series data analysis), however, produces interesting results. Here, we will report effects of soybean protein ingestion for an example.

Soybean protein is well known to lower blood cholesterol. Soybean β -conglycinin is thought to lower triglyceride levels and decrease body fat. One volunteer user (female, 22 years old) ingested different amount of soybean milk each day for about 6 months and measured her body-fat percentage almost every day under the same conditions.

The correlation coefficient r_{he} between body-fat percentage and the summation of soybean milk ingestion is, for various values of the retardation parameter s, shown in Figure 3 as a function of summation-day number (i-j+1). Negative correlation coefficients mean that ingestion of soybean milk was associated with a lower body-fat percentage (Figure 4). It should be noted that the r_{he} between body-fat percentage and the sum- mation of soybean milk ingestion strongly depends on (i-j+1) and s and that its absolute value is greatest when i-j+1 =12 and s = 4.

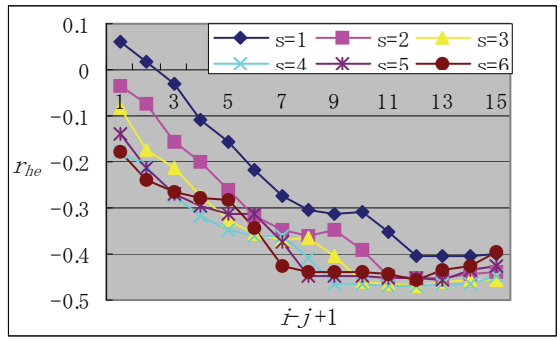

Figure 3 - Correlation coefficient r_{he} between body-fat percentage and summation of soybean milk ingestion as a function of summation-day number (i-j +1) for various values of retardation parameter s

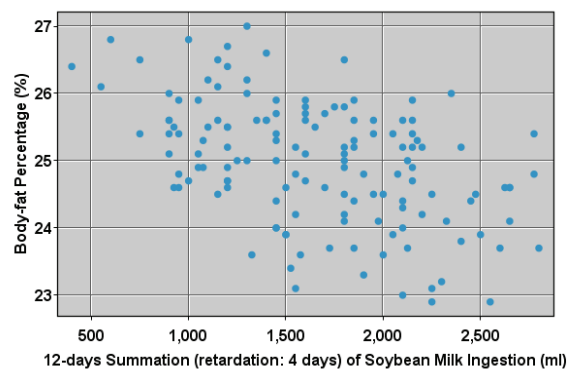

Figure 4 - Typical scatterplot of body-fat percentage (n:139, max:27.0 min:22.9, med:25.0, SD:0.906) versus soybean milk ingestion (correlation coefficient: -0.470)

Discussion

Body-fat percentage lowering may be due to an action of soybean β -conglysinin, which is reported to act on genes related to fat metabolism and thereby decrease triglyceride level. If so, Figure 3 suggests that in the human body the effect of β -conglysinin is gradual: for this volunteer user, the absolute value of the correlation coefficient r_{he} was maximum at i-j +1 = 12 (twelve days successive ingestion) and s = 4 (four days delay). These values of i-j and s at which correlation coefficient becomes largest may differ from person to person.

In the automated healthcare-data-miming process, the correlation between Δh_{nm} and e^t_{ij} in time-series data is checked by changing n-m, i-j, and s as parameters (n-m = 1-10, i-j = 0-9, s = 1-3). Time-series data analyses described here, however, suggest that ranges of i-j and s values should be extended more.

Conclusion

Time-series data analyses for healthcare-data-mining process in the personal dynamic healthcare system produced interesting results concerning the relations between soybean protein ingestion and body-fat percentage for a volunteer user of our PDHS.