Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Automated Healthcare Data Mining Based on a Personal Dynamic Healthcare System

Hiroshi Takeuchi, *Senior Member, IEEE*, Naoki Kodama, Takeshi Hashiguchi and Doubun Hayashi

***Abstract*—The automated healthcare-data-mining system reported 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 system enables users to input their daily data through a mobile phone and to transfer these data to a web-application server via the Internet. The web application server provides a data-mining service and uses mobile phones to inform users of important rules concerning their health and lifestyle data. Automated healthcare-data mining of the stored time-series data of volunteer users generated some useful rules correlating their lifestyles with body-fat index.**

I. Introduction

Applications of web technology to healthcare for remote patients are current topics of interest [1]-[3], and these advanced applications are also expected to be useful in preventive medicine. The demand for personal healthcare systems to prevent diseases and improve health has been increasing recently [4]. Although a system of periodical group health-checks (government sponsored) is now common in Japan, daily personal healthcare remains important in preventing diseases and improving overall health because disease risks differ from person to person owing to both genetic and environmental factors.

Within this context we have developed a personal dynamic healthcare system (PDHS) utilizing the Internet [5]. It enables time-series daily-health and lifestyle data to be stored in a database utilizing mobile phone and web technologies. Recent mobiles phones are not merely phones but are also wireless computers. The browser on a mobile phone can communicate with a web-application server via the Internet. Users can thus input their daily data through a mobiles phone and can transfer this data to a web-application server via the Internet. The web-application server provides a data mining service and, through a mobile phone, notifies users of important rules concerning their health and lifestyle data.

In a previous paper [6] we presented a concept of healthcare-data-mining, which extract rules correlating health and lifestyle data. Healthcare-data-mining was shown there to extract personally useful information such as rules and patterns concerning lifestyles and health conditions embedded in daily time-series personal health and lifestyle data. The present paper reports an automated rule induction method we studied in order to provide a data mining service for an unspecific number of users.

Manuscript received April 1, 2006. This work was supported by Grants-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Hiroshi Takeuchi is with the Department of Healthcare Informatics, Takasaki University of Health and Welfare, 37-1, Nakaorui-machi, Takasaki-shi, Gunma, 370-0033, Japan (phone: +81-27-352-1290; fax: +81-27-353-2055; e-mail: htakeuchi@takasaki-u.ac.jp).

Naoki Kodama is with the Department of Healthcare Informatics, Takasaki University of Health and Welfare, 37-1, Nakaorui-machi, Takasaki-shi, Gunma, 370-0033, Japan (e-mail: kodama@takasaki-u.ac.jp).

Takeshi Hashiguchi is with the Department of Translational Research for Healthcare and Clinical Science, Graduate School of Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyou-ku, Tokyo 113-8655, Japan (e-mail: hashiguchi-mi@umin.ac.jp).

Doubun Hayashi is with the Department of Translational Research for Healthcare and Clinical Science, Graduate School of Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyou-ku, Tokyo 113-8655, Japan (e-mail: hayashi-tky@umin.ac.jp).

II. Materials and Methods

A. System Configuration

The configuration for the personal dynamic healthcare system we developed is shown in Fig. 1. Each client has a mobile phone with a browser for communicating with a web-application server. The server side is a typical model-view-controller architecture. The web-application server (Tomcat) runs on a web server (Apache) and communicates with a database server (Hitachi HiRDB). Clients communicate with the web server by using the HTTPS protocol to encrypt personal data transferred via the Internet. Basic authentication is applied to the client side because clients are unspecific. An automated healthcare-data-mining program is installed on a data-mining server.

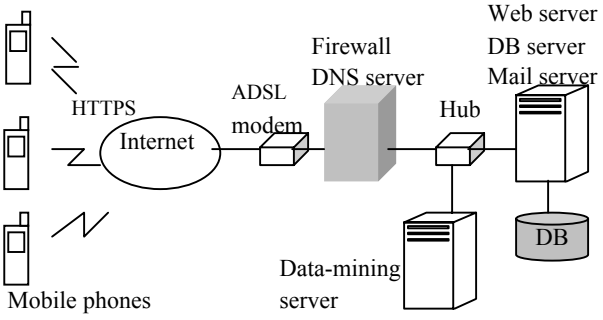

Fig. 1. Configuration of personal dynamic healthcare system.

B. Data-mining Process

1) Concept of Healthcare Data Mining: The concept

behind healthcare data mining is shown in Fig. 2 [6]. Lifestyle data are independent (input) variables and health data are target (output) variables. That is, lifestyle data are involved in antecedents, and health data are involved in consequents. Rules relating health and lifestyle data are expressed as follows: If lifestyle data_1 fulfills condition 1 and lifestyle data_2 fulfills condition 2, then health data has a certain value for the case of those two antecedents. These rules are extracted from daily time-series personal health and lifestyle data for each user. Such personal information will be useful for a user's tailor-made health maintenance.

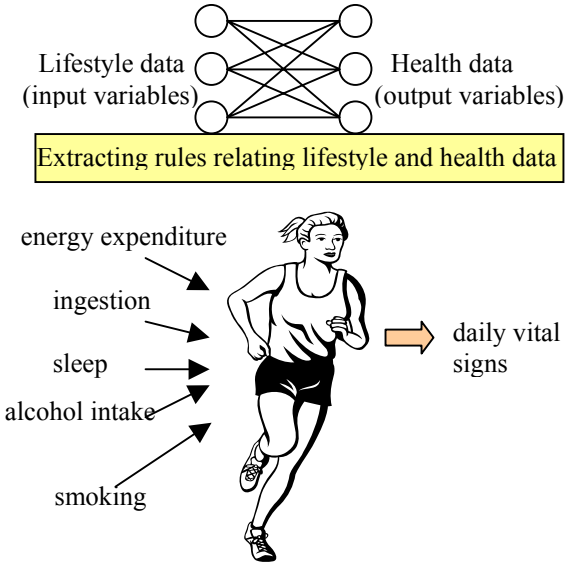

Fig. 2. Concept of healthcare data mining [6].

2) Data-mining Method: Because the extracted rules and patterns should appear clear and reasonable to each user, an association rule mining and classification might be promising techniques for the healthcare data mining [7],[8]. Rule induction should be done automatically because on-line acquisitions from large amounts of user data are required. We therefore explored the use of a generalized rule induction algorithm, which is more general and simple than classification would be. We used the ITRULE algorithm based on information theory [9],[10].

The ITRULE algorithm produces a simple rule:

$$\text{If } \boldsymbol{Y} = y, \text{ then } \boldsymbol{X} = x \text{ with probability } p \qquad (1)$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two fields (attributes) and x and y are values for those fields. The consequent is restricted to being a single value assignment expression while the antecedent may be a conjunction of such expressions. For example,

$$\text{If } \boldsymbol{Y} = y \text{ and } \boldsymbol{Z} = z, \text{ then } \boldsymbol{X} = x \text{ with probability } p. \qquad (2)$$

The complexity of a rule is defined as the number of conjuncts appearing in the rule's antecedent. In our healthcare data mining, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are lifestyle-data items and $\boldsymbol{X}$ is a health-data item.

The ITRULE algorithm uses the J-measure to generate rules to summarize patterns in the stored data [10]. This measure provides a method for ranking competing rules (rules having higher-J- measure survive). The J-measure is defined by

$$J(x \mid y) = p(y)\left(p(x \mid y)\log\frac{p(x \mid y)}{p(x)} + (1-p(x \mid y))\log\frac{(1-p(x \mid y))}{(1-p(x))} \right) \qquad (3)$$

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

The ITRULE algorithm automatically generates rules through the following steps.

a) Process each interesting output field X_i (health-data item) in turn. Derive all rules for the current output field before moving on to the next.

b) For each output field, select each possible output field value x_k. For example, body-fat percentage $X_i = x_k$: higher. All rules predicting the current output field value (x_k: higher) are generated before the next output field value is considered.

c) For each output field value (x_k: higher), select each input field Y_m (lifestyle-data item).

d) For each input field, select each possible condition y_q. The conditions depend on the type of the input field.

- For symbolic fields (for example, alcohol intake), each value for the field represents a possible condition.

- For numeric fields (for example, energy expenditure), values are sorted and each value is tested as a binary split boundary. For each potential split, the J-measure is calculated and the split with the highest J value is selected as the split for the rule. There are then two possible conditions: greater than the split value, and less than or equal to the split value.

e) For the rule $Y_m = y_q \Rightarrow X_i = x_k$, compute the J-measure.

f) If the value of J is greater than the highest J for any rule in the table predicting the same outcome ($X_i = x_k$), or if the number of rules in the table is less than the maximum number of rules in the table, and if the minimum support and confidence criteria are met, insert the rule in the table (replacing the lower-J rule if necessary) and calculate J_s. Otherwise, proceed to the next input field value. A system administrator determines the maximum number of rules and the minimum support and confidence criteria, and J_s is defined [10] by

$$J_s = \max\left(p(y)p(x \mid y)\log\left(\frac{1}{p(x)}\right), p(y)(1-p(x \mid y))\log\left(\frac{1}{1-p(x)}\right) \right) \qquad (4)$$

g) If $J_s > J_{min}$, specialize the rule. Here J_{min} is the smallest J for rules in the table predicting the same outcome. The rule is specialized by adding conditions to the antecedent, in the same manner used for the original unspecialized (single antecedent) rules. Input fields that have already been

evaluated as antecedents for the current outcome are not considered as potential specializing conditions. Each specialized rule is evaluated by testing its J value against those of other rules in the table with the same outcome, and if its value exceeds $J_{\min}$ the specialized rule replaces that minimum-J rule in the table.

h) Repeat until all input field values, input fields, output field values, and output fields have been considered.

3) Input Fields: The most important process in healthcare data mining in the personal dynamic healthcare system is its automated determination of input fields. Too many (or ineffective) input fields result in excessive consumption of CPU-time. We determined effective input fields by pre-examining the correlation between time-series lifestyle data and health data of interest as shown in Fig. 3, where e is the given lifestyle data and h is the given health data. Here h_n is the value of h at the n-th date, e_i is the value of e at the i-th date, and s is a retardation parameter. The quantities Δh_{nm} and e^t_{ij} are defined as follows:

$$\Delta h_{nm} = h_n - h_m \tag{5}$$

$$e^t_{ij} = e_i + e_{i-1} + \ldots\ldots + e_j \tag{6}$$

The correlation between Δh_{nm} and e^t_{ij} in terms of time-series data was examined by changing n-m, i-j, and s as parameters (n-m = 1-10, i-j = 0-9, and s = 1-3) and calculating Pearson's product-moment correlation coefficient r for each (n-m, i-j, s) set. We assume that lifestyle data before the n-th date affect the health data at the n-th date ($i \leq n$-1). If more than one Pearson's correlation coefficient is larger than a threshold value R_s, we suppose that e is a candidate input variable. In this case, the input field for e is defined according to the values of $(i\text{-}j)_{\max}$ and $s_{\max}$ at which r becomes largest. For example, the input field for e is defined by $e_i + e_{i-1} + e_{i-2}$ ($i = n$-2) in the case that $(i\text{-}j)_{\max} = 2$ and $s_{\max} = 2$. A large value of $(i\text{-}j)_{\max}$ implies that long-term lifestyle data affect the present health data, and a large value of $s_{\max}$ implies that lifestyle data affect health data with a large retardation.

When values of e are not numeric but symbolic, correlation of e with h in time-series data is examined by transforming symbolic values to numeric ones. The symbolic values "much," "moderate," and "little," for example, are respectively transformed to e values 3, 2, and 1. Thus, $e_i + e_{i-1} = 3 + 1 = 4$ when e_i = "much" and e_{i-1} = "little." In the data mining process, however, the input field for e has a symbolic value. An example of an input field value in case $(i\text{-}j)_{\max} = 1$ is ("much" | "little").

4) Output Fields: The output (target) variables in the healthcare data-mining process are interesting health data h. When calculating Δh_{nm} for h values that are symbolic, the symbolic h values are transformed into numeric ones in a manner similar to that used to transform symbolic e values. In the data-mining process, however, the output field values are symbolic. While values of h are numeric, time-series data 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 in each class is a similar number.

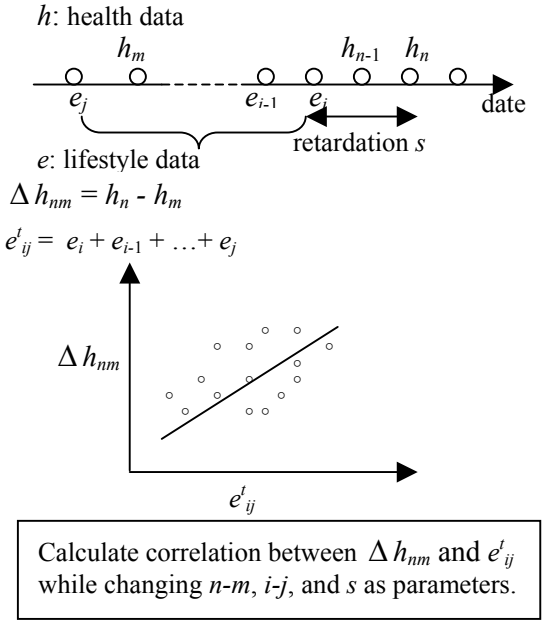

Fig. 3. Algorithm for defining the input field automatically [6].

III. Results

Registered data items and their data types for a volunteer user are listed in Table I. Data for these items were gathered almost every day for a year (about 330 records).Body-fat percentage was measured by body-impedance method. The measurements were done regularly in the morning under the same conditions. Energy expenditure (kcal) due to exercise was estimated from gym-training records and wearable-walking-monitor records. Ingestion (kcal) was estimated from each day's breakfast, lunch, and dinner contents. The symbolic values assigned for alcohol intake were "too much," "much," "moderate," "little," and "very little".

TABLE I
REGISTERED DATA ITEMS OF A VOLUNTEER USER

Health data	Lifestyle data
Body-fat percentage (numeric)	Energy expenditure (numeric) Ingestion (numeric) Alcohol intake (symbolic)

A. Automated Healthcare Data Mining Flow

1) Data Check: This process checks how much lifestyle health data has been stored for each user. The data mining processes for each user automatically start when, in the first three months, the number of lifestyle and health data sets becomes larger than N_s. We set N_s = 80 based on the

assumption that about 10% data missing is allowable.

2) Correlation Check: This process calculates Pearson's product-moment correlation coefficients r between Δh_{nm} and e^{t}_{ij} in terms of time-series data by changing n-m, i-j, and s as parameters (n-m = 1-10, i-j = 0-9, and s = 1-3). If the maximum r is larger than R_s, the corresponding e^{t}_{ij} is automatically selected as an input field. We set $R_s = 0.3$ based on the assumption that this value is the criteria determining if "correlation" is significant or not.

Once a year the correlation check process is executed on the fresh (most recent three months) data for each user. That is, defined input fields are assumed to be effective for a year.

3) Rule Induction: This process generates useful rules correlating lifestyle-data with health-data. Since complicated rules may confuse users, we make the output easier to use for personal healthcare by restricting the maximum rule order to 2 (*i.e.*, we restricted the number of antecedents to 2).

The minimum support and confidence criteria of rules in the rule table are adjustable. We set the minimum support S_s and the minimum confidence C_s so that the number of the rules in the table is less than 10.

B. Automatically Defined Input Fields

Correlations between lifestyle-data items and body-fat percentage were calculated. The calculation results and defined input fields are summarized in Table II (A). The first N_s (80) time-series data sets for each data item were used in this calculation. In Table II (A), r is Pearson's product-moment correlation coefficient when $(n\text{-}m)=(n\text{-}m)_{max}$, $(i\text{-}j)=(i\text{-}j)_{max}$, and $s=s_{max}$.

TABLE II

(A) CALCULATED CORRELATION COEFFICIENTS AND DEFINED INPUT FIELDS

Input variable	r	Defined input field
Energy expenditure	-0.369	$e_i + e_{i-1} + \ldots + e_{i-4}$ ($i = n$-2)
Ingestion	0.321	$e_i + e_{i-1} + \ldots + e_{i-4}$ ($i = n$-2)
Alcohol intake	0.608	e_i ($i = n$-2)

(B) CATEGORIZATION OF TIME-SERIES DATA

Field value	Data range	Data frequency
higher	≥ 18.0 %	115
moderate	17.3 -17.9 %	112
lower	≤ 17.2%	104

C. Automatically Defined Output Fields

The output variable was body-fat percentage. Time-series body-fat percentage data were categorized into three classes, each having a symbolic value as higher, moderate, or lower. This categorization is summarized in Table II (B). Border values used for this categorization were automatically determined so that the data frequency was similar in each class.

D. Rule Induction

The 4 rules generated when we set $S_s = 0.04$ and $C_s = 0.60$ are summarized in Fig. 4. All are simple and reasonable. Both ingestion and energy expenditure are key lifestyle data affecting the daily body-fat percentage of the volunteer user in question. It should be pointed out that long-term (5 days) lifestyle data affect the present body-fat percentage. Rule 2, with the highest confidence (0.72), is a specialization of Rule 1. It suggests that ingesting more than 2000 kcal ingestion a day when the daily energy expenditure is less than 280 kcal increases the body-fat percentage of this volunteer user. Rules 3 and 4 are also very informative. The body-fat percentage of this user is indicated to be decreased by increasing energy expenditure to more than 350 kcal a day or decreasing energy intake to less than 1650 kcal a day.

Alcohol intake for day before yesterday was selected as a input field as shown in Table II (A), but this field was not involved in the antecedent of rules generated under the present rule induction conditions.

Output Field: Body-fat Percentage (B.F.P.)
total number of records: 331
higher: >= 18.0 % I : Instances
lower: <= 17.2 % S: Support
C: Confidence
rule induction conditions: maximum rule order = 2, S >= 0.04, C >= 0.60

	antecedent	consequent	I	S	C
1	[Ingestion5 > 9925 kcal	] [B.F.P.=higher]	37	0.112	0.62
2	[Ingestion5 > 9925 kcal, Expend.5 <1417 kcal	] [B.F.P.=higher]	29	0.088	0.72
3	[Expend.5 > 1747 kcal	] [B.F.P.=lower]	21	0.063	0.62
4	[Ingestion5 < 8225 kcal	] [B.F.P.=lower]	14	0.042	0.71

・Ingestion5: ingestion for 5 days
・Expend.5: energy expenditure for 5 days

Fig.4. Automatically generated rules for body-fat percentage.

REFERENCE

[1] N. H. Lovell, F. Magrabi, B. G. Celler, K. Huynh, and H. Garsden, "Web-based acquisition, storage, and retrieval of biomedical signals," *IEEE Eng. Medicine and Biology*, vol. 20, no. 3, 2001, pp. 38–44.

[2] J. Cai, S. Johnson, and G. Hripcsak, "Generic data modeling for home telemonitoring of chronically ill patients," *Proc. AMIA Symp*. 2000, pp. 116–20.

[3] C. Mazzi, P. Ganguly, and M. Kidd, "Healthcare application based on software agents," *Medinfo 2001 Proceedings*, 2001, pp. 136–140.

[4] T. Hashiguchi, H. Takeuchi, and A. Uemura, "Highly advanced healthcare support services for the 21st century," *Hitachi Review*, vol. 50, no. 1, 2001, pp. 2–7.

[5] H. Takeuchi, T. Hashiguchi, and T. Shintani, "Personal dynamic healthcare system utilizing mobile phone and web technologies," *Proc. 2nd Int. Conf. on Advances in Biomedical Signal and Information Processing*, 2004, pp. 304–307.

[6] H. Takeuchi, N. Kodama, T. Hashiguchi, and N. Mitsui, "Healthcare data mining based on a personal dynamic healthcare system," *Proc. 2nd Int. Conf. on Computational Intelligence in Medicine and healthcare,* 2005, pp. 37-43.

[7] M. F. Usama, P. S. Gregory, P. Smyth, and U. Ramasamy, *Advances in Knowledge Discovery and Data Mining*, The AAAI Press, 1996.

[8] M. J. A. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, Inc., 1997.

[9] R. M. Goodman and P. Smyth, "An information-theoretic model for rule based expert systems," presented at the 1988 Int. Symp. on Information Theory, Kobe, Japan, 1988.

[10] P. Smyth and R. M. Goodman, "An information theoretic approach to rule induction from databases," *IEEE Trans. Knowledge and Data Engineering,* vol. 4, no. 4, 1992, pp. 301–316.