Validity of Association Rules Extracted by Healthcare-Data-Mining

Hiroshi Takeuchi-*IEEE Senior Member* and Naoki Kodama

***Abstract*—A personal healthcare system used with cloud computing has been developed. It enables a daily time-series of personal health and lifestyle data to be stored in the cloud through mobile devices. The cloud automatically extracts personally useful information, such as rules and patterns concerning the user's lifestyle and health condition embedded in their personal big data, by using healthcare- data-mining. This study has verified that the extracted rules on the basis of a daily time-series data stored during a half- year by volunteer users of this system are valid.**

I. Introduction

Application of the Internet to healthcare such as m-health and p-health are current topics of interest [1, 2]. In particular, the demand for personalized healthcare systems to prevent diseases and improve health has been increasing recently [3]. Within this context we have developed a personal dynamic healthcare system (PDHS) using cloud computing [4]. It enables time-series daily-health and lifestyle data to be stored in a database utilizing mobile device and to extract personally useful information such as rules and patterns concerning lifestyles and health condition embedded in daily time-series personal health and lifestyle data. We call this 'healthcare-data-mining'.

In the healthcare-data-mining process [5], we first check the correlation between variations of the time-series health data and summations of the time-series lifestyle data. If the correlation coefficient is larger than a certain threshold value, the lifestyle is selected as an independent variable in the data-mining process relevant to the health condition. We have verified that association rules induced by healthcare-data-mining on the basis of time-series data stored during 2012/06/01 through 2012/11/30 by volunteer users of PDHS are valid.

II. Method

A. Volunteer Users (Subjects)

The volunteer users (subjects) of PDHS were three 22-year-old- students of Takasaki University of Health and Welfare. Two female students examined the effect of daily ingestion of a 'health- tea' (a food for specified health uses in Japan) and soybean milk, respectively, on their weight and body-fat. One male student examined the effect of daily smoking to his blood pressure at home.

B. Data Acquisition

Weight and body-fat percentage were measured every morning upon waking by using with a body composition meter (Tanita Inner Scan BC-521). Blood pressures (systolic and diastolic) were also measured every morning upon waking with an automatic blood pressure meter (Omron) using the oscillometric method. The data were taken three times and their average values were recorded. Energy expenditure due to exercise was measured with a wearable monitor (Omron) and energy supply was estimated from each day's breakfast, lunch and dinner contents. The amount of tea and soybean milk ingested a day were recorded every day. These data were input through mobile device and stored using a PDHS (Fig. 1).

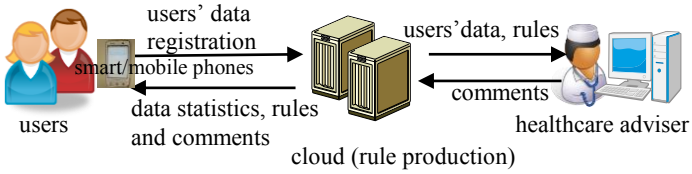

Figure 1. Personal dynamic healthcare system (PDHS)

C. Healthcare-Data-Mining Algorithm

1) Correlation check for determining independent variables

The time-series data analysis described here is based on the simple idea that the accumulation of the effects of lifestyle events, such as exercise and the ingestion of various foods, could affect personal health with some delay [5]. The delay may reflect complex bio-reactions such as those of the metabolism in a human body. In the analysis, the accumulation of the effects of lifestyle events is represented by a summation of lifestyle data, such as energy supply due to ingestion and amount of exercise. The accumulation of the effects may cause variation of health data, such as weight and body- fat percentage with some delay.

In the analysis, we examine the correlation coefficient, r, described as:

$$r(\Delta h_{nm}, e^{t}_{ij}) = \frac{Cov(\Delta h_{nm}, e^{t}_{ij})}{SD(\Delta h_{nm}) SD(e^{t}_{ij})} \quad (1)$$

Here,

$$\Delta h_{nm} = h_n - h_m \quad (2)$$

This work was supported in part 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).

is the difference of time-series health data h , representing the variation of health condition, and

$$e_{ij}^{t} = e_i + e_{i-1} + ... + e_j \qquad (3)$$

is the summation of time-series lifestyle data e during a certain period, representing the accumulation of the effects of lifestyle events. The delay is represented by retardation, $s = n - i \geq 1$ (Fig.2). In Eq.(1), $SD(\Delta h_{nm})$ and $SD(e^{t}_{ij})$ are the standard deviation of Δh_{nm} and e^{t}_{ij}, respectively, and $Cov(\Delta h_{nm}, e^{t}_{ij})$ is the covariance.

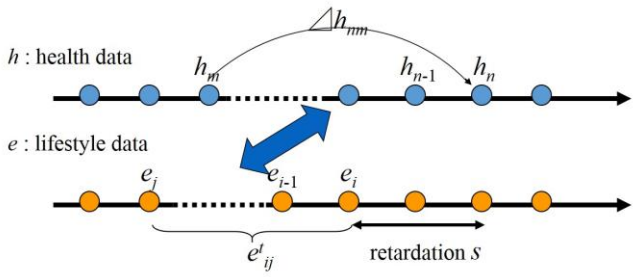

Figure 2. Reference figure for checking time-series correlation between variation of health condition and accumulation of effects of lifestyle events.

The correlation coefficient, r is estimated for the time-series health and lifestyle data in a certain period by changing $n - m$, $i - j$, and s as parameters. If the maximum value of r is larger than a certain threshold value, the lifestyle is selected as an independent variable in the data-mining process relevant to the health condition.

2) Data-mining process

To extract association rules, we used the ITRULE algorithm that is based on information theory [6]. This algorithm produces a simple rule:

$$\text{If } \boldsymbol{Y} = y, \text{ then } \boldsymbol{X} = x \text{ with probability } p \qquad (4)$$

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,

$$\text{If } \boldsymbol{Y} = y \text{ and } \boldsymbol{Z} = z, \text{ then } \boldsymbol{X} = x \text{ with probability } p. \qquad (5)$$

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. This measure provides a method for ranking competing rules (rules having a higher-J- measure survive). The J-measure is defined by

$$J(x \mid y) = p(y)\left(p(x \mid y)\log\frac{p(x \mid y)}{p(x)} + (1 - p(x \mid y))\log\frac{(1 - p(x \mid y))}{(1 - p(x))} \right) \qquad (6)$$

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.

III. Validity of Association Rules

A. Rule for Health-tea Ingestion

The subject was a 22-year-old female. The PDHS automatically induced a rule 'If average energy supply by meal ingestion a day during four days is larger than 1369.5 kcal and average amount of health-tea ingestion a day during two days is less than 177.5 ml, then body-fat percentage is higher after two days' [confidence: 77.8% support: 10.5%].

To ascertain whether this rule is valid, we first showed the scatter plots of body-fat variation vs. average energy supply by meal ingestion a day during four days in Fig. 3. A significant positive correlation was observed. It was shown that all the data of body-fat variation were positive when energy supply by ingestion a day was larger than 1350 kcal.

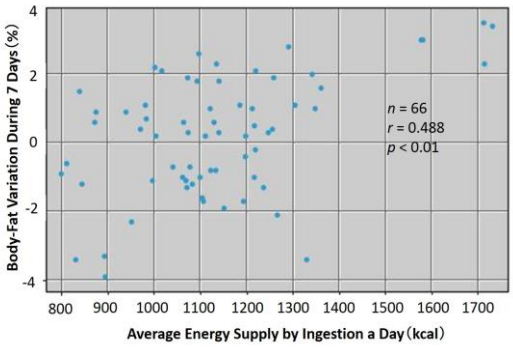

Figure 3. Scatter plots of body-fat variation vs. avearge energy supply by meal ingestion a day during four days.

Second, we present the scatter plots of body-fat variation vs. average amount of health-tea ingested a day during two days (Fig. 4).

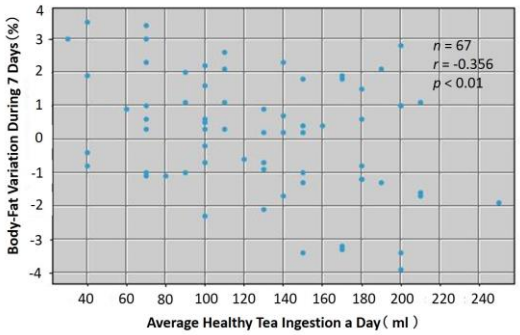

Figure 4. Scatter plots of body-fat variation vs. average amount of health-tea ingested a day during two days.

A significant negative correlation was observed and it was shown that the probability of positive was higher than that of negative in the body-fat variation when the average amount of health-tea ingestion was less than 180 ml. In this way, the induced rule was supported by the relevant scatter plots.

B. Rule for Soybean Milk Ingestion

The subject was also a 22-year-old female. She ingested a different amount of soybean milk each day for 3 months and measured her weight and body-fat percentage almost every day under the same conditions. Soybean protein is well known to lower blood cholesterol. Soybean β-conglycinin is thought to lower triglyceride levels and decrease body fat.

However, the PDHS automatically induced this rule: 'If the average amount of soybean milk ingestion a day during ten days is larger than 550 ml, then weight is higher after three days' [confidence: 100% support: 15.0%]. Relevant scatter plots are shown in Fig. 5. A positive correlation was indeed observed between variations of weight and the average amount of soybean milk ingestion, and almost all weight data increased when the average amount of soybean milk ingestion a day during ten days was larger than 540 ml.

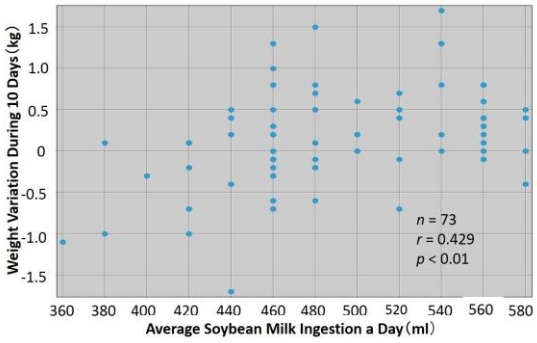

Figure 5. Scatter plots of weight variation vs. average amount of soybean milk ingestion a day during ten days.

It is noted, however, that weight decreased when the average amount of soybean milk ingestion a day was less than 420 ml. In addition, body-fat decreased when the average amount of soybean milk ingestion a day was less than 480 ml whereas it increased when it was larger than 500 ml, as shown in Fig. 6. That is, there is a strong non-linear correlation between body-fat variation and amount of soybean ingestion.

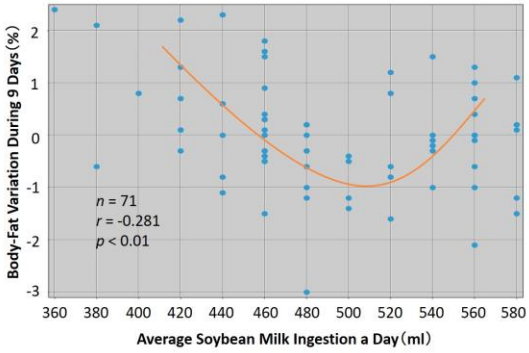

Figure 6. Scatter plots of body-fat variation vs. average amount of soybean milk ingestion a day during ten days.

Thus, this subject ingested soybean milk so that the amount was less than 400 ml a day for another 3 months. The PDHS automatically induced the following rule: 'If the average amount of soybean milk ingested a day during ten days is larger than 305 ml, then weight is lower the next day' [confidence: 86%, support: 8.6%]. Scatter plots of weight and body-fat vs. average amount of soybean milk ingestion a day are shown respectively in Figure 7 and 8.

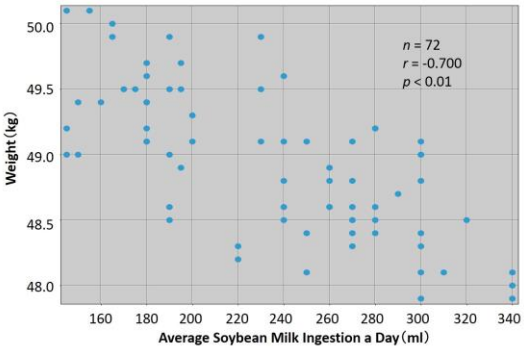

Figure 7. Scatter plots of weight vs. average soybean ingestion a day during ten days.

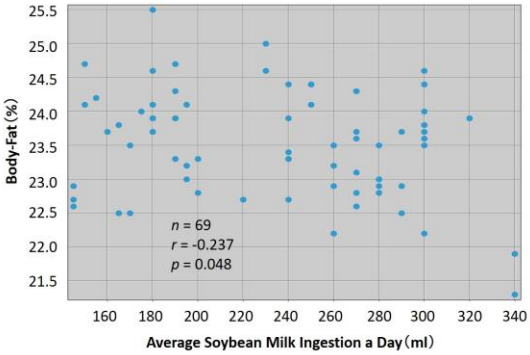

Figure 8. Scatter plots of body-fat vs. average soybean ingestion a day during ten days.

The induced rule was supported by the significant negative correlations observed in these scatter plots although the negative correlation between body-fat and average amount of soybean milk ingestion was at a significant level of 5% (p = 0.048).

C. Rule for Smoking

The subject was a 22-year-old male who is a heavy smoker and anxious about its effect on his blood pressure. The PDHS automatically induced a rule 'If the number of cigarettes smoked a day is larger than 8.5, then both systolic and diastolic blood pressures are higher next day' [confidence: 83.8% support: 20%].

To verify whether this rule was valid, we show scatter plots of systolic and diastolic blood pressures vs. the number of cigarettes smoked a day, respectively in Fig. 9 and 10. Significant positive correlations were observed in these scatter plots, and variations of blood pressures were found to be almost positive when the number of cigarettes smoked a day was larger than 8.5.

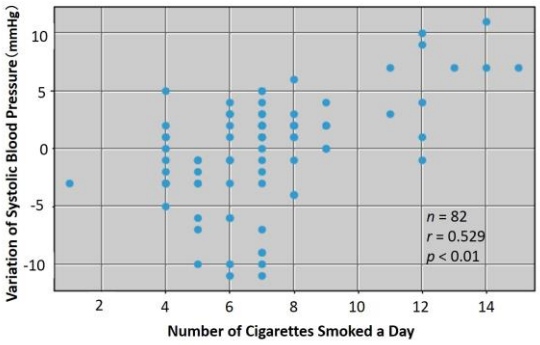

Figure 9. Scatter plots of systolic blood pressure vs. number of cigarettes smoked a day.

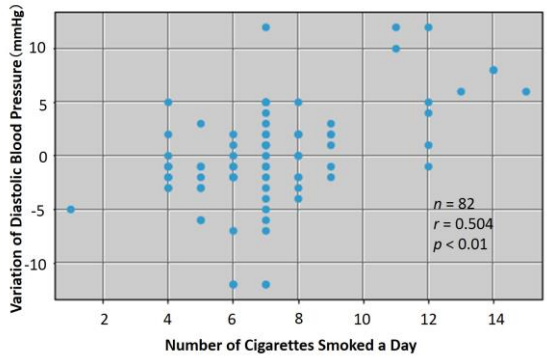

Figure 10. Scatter plots of diastolic blood pressure vs. number of cigarettes smoked a day.

The automatically induced rule was clearly supported by these relevant scatter plots.

IV. Discussion

The validity of association rules automatically extracted by the healthcare-data-mining for three subjects were checked by examining relevant scatter plots from which rules were induced. Of the three subjects, the rule for smoking was most definitely supported by the relevant scatter plots. A rule should be a cause-and-effect relationship rather than correlation. In this context, it can be seen from Figs. 9 and 10, that positive and negative probabilities of blood pressure variations were almost the same when the number of cigarettes smoked a day was less than 8.5 whereas a positive variation was definitely dominant when it was larger than 8.5. The data-mining algorithm thus chose 8.5 as the critical number of cigarettes smoked a day to induce a rule for blood pressure.

The rule for health-tea ingested was also clearly supported by the relevant scatter plots. Energy supply by meal ingestion during four days and amount of health-tea ingested during two days were selected as independent variables in the data-mining process relevant to body-fat variation. Positive and negative probabilities of body-fat variations were almost the same when the average energy supply by ingestion a day was less than 1350 kcal whereas a positive variation was definitely dominant when it was larger than 1350 kcal as shown in Figure 3. It can be also seen from Fig. 4 that a positive body-fat variation was dominant when the amount of health-tea ingested a day was less than 180 ml compared to when it was larger than 180 ml.

The first induced rule for soybean milk ingestion was not incorrect but could be misunderstood. The fact is that an excess ingestion of soybean milk during ten days increased weight of the subject. However, the induced rule may state that 'If the average amount of soybean ingestion a day during ten days is larger, then the weight is higher after three days'. This rule may confuse the subject since she expects a decrease in both weight and body-fat as a result ingested soybean milk. The algorithm of healthcare-data-mining may produce such confusing rules when there is a strong non-linear correlation between health data (target variable) and lifestyle data (independent variable). We show that a proper rule (the second rule for soybean milk ingestion) could be automatically induced by a priori setting limits in the values of the independent variable.

V. Conclusion

The algorithm of healthcare-data-mining automatically produced proper rules between health conditions and lifestyles for two of three volunteer users (subjects) of our PDHS. For one of three subjects, the algorithm induced a confusing (but not incorrect) rule because of a strong non-linearity in the scatter plots of target variables vs. independent variables. In such a case, there needs to be a priori limits on the values of independent variable.

Acknowledgment

We express our sincere thanks to volunteer users (students of Takasaki University of Health and Welfare) of our PDHS.

References

[1] R. S. H. Istepanian, A. Sungoor, and K. A. Earle, "Technical and compliance considerations for mobile health self-monitoring of glucose and blood pressure for patients with diabetes," *Proc. 31st Annual International Conference of the IEEE EMBS*, 2009, pp. 5130-5133..

[2] H. Kumpusch, D. Hayn, K. Kreiner, M. Falgenhauer, J. Mor, and G. Schreier, "A mobile phone based telemonitoring concept for the simultaneous acquisition of biosignals and physiological parameters, *Proc. 13trd World Congress on Medical and Health Informatics*, 2010, pp.1344-1348.

[3] I. Korhonen, E. Mattila, A. Ahtinen, J. Salminen, L. Hopsu, R. Lappalainen, and T. Leino, "Personal health promotion through personalized health technologies – Nuadu experience," *Proc. 31st Annual International Conference of the IEEE EMBS*, 2009, pp. 316-319.

[4] 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.

[5] H. Takeuchi, N. Kodama, T. Hashiguchi, and D. Hayashi, "Automated healthcare data mining based on a personal dynamic healthcare system," *Proc. 28th IEEE EMBS Annual Int. Conf.* 2006, pp.3604 -3607.

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