MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-491

Hidden Markov Model for Analyzing Time-Series Health Checkup Data

Ryouhei Kawamoto [a], Alwis Nazir [b], Atsuyuki Kameyama [c], Takashi Ichinomiya [c], Keiko Yamamoto [a], Satoshi Tamura [a,e], Mayumi Yamamoto [d] Satoru Hayamizu [a,e], Yasutomi Kinosada [c,e]

[a] *Graduate School of Engineering, Gifu University, Japan*
[b] *United Graduate School of Drug Discovery and Medical Information Sciences, Gifu University*
[c] *Biomedical Informatics, Gifu University Graduate School of Medicine, Japan*
[d] *Health Administration Center, Gifu University, Japan*
[e] *R&D Center for Human Medical Engineering, Gifu University, Japan*

Abstract

In this paper, we apply a Hidden Markov Model (HMM) to analyze time-series personal health checkup data. HMM is widely used for data having continuation and extensibility such as time-series health checkup data. Therefore, using HMM as probabilistic model to model the health checkup data is considered to be suitable, and HMM can express the process of health condition changes of a person. In this paper, a HMM with six states placed in a 2x3 matrix was prepared. We collected training features including the time-series health checkup data. Each feature consists of eight inspection parameters such as BMI, SBP, and TG. The HMM was then built using the training features. In the experiments, we built five HMMs for different gender and age conditions (e.g. male 50's) using thousands of training feature vectors, respectively. Investigating the HMMs we found that the HMMs can model three health risk levels. The models can also represent health transitions or changes, indicating the possibility of estimating the risk of lifestyle-related diseases.

Keywords:

Public health informatics, personal health records, hidden Markov model, health checkup, data mining, big data.

Introduction

In order to reduce medical expenses, personal health care has become increasingly important. In Japan, the Ministry of Health, Labour, and Welfare promotes the special health checkup and specific counseling guidance. In the field of medical information technology that supports medical care and hygiene, fundamental technologies for early detection of lifestyle-related diseases or prevention of the progress of disease are strongly recommended. Thus, personal health checkup results are very important. The results can also be used to find signs of developing adult diseases, such as Diabetes Mellitus (DM) and Hypertension (HT). Furthermore, by using the health checkup results obtained from numerous people as population data, it is possible to model distributions of the results and to analyze any trends. It is also possible to find patterns of health transitions from youth to old age.

Some studies that consider health checkup findings as one of the risk factors have tried to discover regularity and examined predictive models to diseases development using data mining methods such as regression analysis and decision trees [1][2]. In other works, a probabilistic model was applied to the field of medical diagnosis [3][4][5]; for example, a Bayesian network was employed to verify a relationship between the development of diseases and inspection results. However, since one-time health checkup results have no information on temporal change, it is sometimes impossible to judge recovery from a disease or health aggravation; to do so requires several years of data. To overcome this issue, it is essential to build a model that can indicate progressive changes in health, e.g. the process toward the current health condition. Such the model is also effective in predicting the future health conditions, and to estimating the risk declines in health.

This paper proposes a model to indicate the process of health condition changes, and investigates whether the model can correctly classify personal health checkup results as a healthy state (within the acceptable range) or a disease state (below or over the range), as well as whether the model can estimate a probability of transitioning from a healthy state to a disease state using derivatives of the health checkup results. To complete the model, the Hidden Markov Model (HMM) is therefore used in this paper. Since a health checkup result is generally obtained annually, a time-series health checkup sequence can be easily generated. The health checkup results of any year correspond to a feature vector and the sequence of exams corresponds to a feature sequence. By using feature sequences, we can build HMMs and estimate the probability of generating a given feature sequence. In addition, the sequence potentially has a temporal change; the derivative can be obtained by comparing the health checkup result in a certain year with that in the next year. This means we should be able to predict future results. Furthermore, the state transition can be extracted from the model. The transition probability indicates the risk of disease development or the possibility of recovery. Therefore, by investigating a probability using time-series health checkup features, we can calculate the risk for various diseases, such as DM and HT. As a result, we may be able to predict the risk of an individual of developing these diseases.

Note that HMM is used as a tool for genome analysis and speech recognition; HMM is also utilized in machine health monitoring with the aim of avoiding unexpected failure [6]. HMM is suitable for modeling data having continuation and extensibility [3][4]. As mentioned, health checkup data have continuity of this kind. The data also have the needed extensibility because the duration of state transition depends on each person. Accordingly, it is reasonable to employ HMM for health checkup data.

Materials and Methods

1. Summary

In this paper, we built an HMM using training data that consist of health checkup features that each includes continuous feature vectors. After training, we categorized states in the HMM into three risk levels by comparing the standard value (the acceptable range) and its mean vector of Gaussian distribution on the state. We evaluated the HMM by investigating the transitions in the HMM using the training data.

2. HMM training using time-series health checkup data

We collected health checkup records provided by the medical center in Gifu prefecture from 2002 to 2007. From the data, we used the health checkup features that each had continuous four-year, five-year and six-year inspection data. Each record consisted of questionnaires, somatometry, and inspection parameters, e.g. blood chemical analysis and liver-function test [7]. There were roughly 20 inspection parameters, and we carefully chose the following eight parameters:

- Body mass index (BMI)
- Systolic blood pressure (SBP)
- Hematocrit (Ht)
- Platelet (PLT)
- Glucose oxidase test (GOT)
- Total cholesterol (T.Chol)
- Neutral fat (TG)
- Casual blood glucose (CBG)

We chose these parameters for three reasons. First, they were rarely missing. Second, they indicate the health of a biological system, e.g. cyclic (blood), liver, and lipid. Finally, a physician usually uses about 10-20 parameters for diagnosis. Parameters for the same function are strongly correlated, and it is recommended to use uncorrelated parameters for HMM. So we chose one or two parameters for each system. An example of health checkup record is shown in Table 1. In the table, each column vectors indicate eight inspection data points from individuals 51 years old to 55 years old, respectively.

We built different HMMs depending on gender and age (30's, 40's, and 50's) because the likelihood and distribution of health conditions is different depending on gender and age; most people in their 30's are healthy, in contrast, the risk of lifestyle-related disease increases in the 40's and relatively higher rates of people in their 50's have lifestyle-related diseases. In this paper, in order to build HMMs for males, we used 4,164, 5,733, 7,604 features for 30's 40's, and 50's, respectively. For female HMMs, 2,480 and 3,481 features were used for 40's and 50's, respectively. Note that a 30's female HMM could not be built due to lack of data. Hidden Markov Model Toolkit (HTK) was used to train the HMMs [8]. HTK is a tool for training and recognition, developed by Cambridge University.

Table 1 – An example of health checkup record (male 50's)

	age				
	51	**52**	**53**	**54**	**55**
BMI (kg/m^2)	22.1	21.7	21.4	21.2	20.8
SBP (mmHg)	130	132	128	128	138
Ht (%)	44.8	42.5	44.1	43.7	42.0
PLT (10^4/ml)	16.7	16.9	16.9	16.6	19.9
GOT (IU/l)	21	18	19	20	18
T.Chol (mg/dl)	266	263	276	264	242
TG (mg/dl)	148	169	164	135	174
CBG (mg/dl)	119	176	140	186	225

3. Structure of HMM

In accordance with our preliminary experiments, the structure of the HMM was created as shown in Figure 1. The HMM has six states, structured in a 2x3 matrix. In the first column (state[1], state[3] and state[5] in Figure 1) all the transitions are allowed, as they are in the second column (state[2], state[4] and state[6]). Transitions between each row (e.g. from state[1] to state[2]) are also possible. The HMM is designed to express three risk levels: e.g. state[1] and state[2] indicate healthy states, and state[5] and state[6] indicate high risk. Regarding output probability, the number of Gaussian distributions is one for each state; a state has one Gaussian pdf. Model training was conducted by applying the Baum-Welch algorithm, which is available in HTK.

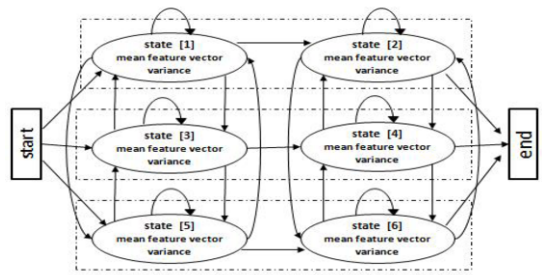

Figure 1 – An HMM having 6 states (like a 2x3 matrix)

Results

1. Mean vectors and transition probabilities in HMMs

Mean feature vectors and state transition probabilities in trained HMMs are shown in the appendix (Appendix Table1, 2, 3, 4 and 5). In each table, there are relatively large differences between states in the BMI, SBP and TG elements. Compared with these six states, it is found that the elements in the state[1] and state[2] have small values and are in the acceptable range, meaning these states indicate a healthy state.

We further analyzed the result for male HMMs. In the 30's, a BMI value in state[4] is greater than those in the healthy states, and values in state[3], state[5] and state[6] are greater than the maximum value of the optimal range (25). A TG value exceeds the treatment threshold (250) in state[6]. GOT values in state[5] and state[6] are significantly high. In the 40's, a BMI value reaches almost the upper limit in state[4] and state[6], and the values in state[3] and state[5] are over the limit. SBP values are almost over the upper limit (140) in state[5] and state[6]. Furthermore, TG and GOT values in both of these states are higher than those of the others. In the 50's, BMI values between state[3] and state[6] are greater than those in state[1] and state[2]. Regarding TG values, state[3] and state[4] are obviously higher than state[1] and state[2], and moreover, state[5] and state[6] are over 200. SBP values are almost over the upper limit on state [5] and state [6]. In addition, all the values in the 40's are generally greater than those in the 30's, as well as the 50's. According to BMI, SBP and TG, states can be categorized into three groups: state[1] and state[2] (low values), state[3] and state[4] (middle values), state[5] and state[6] (high values). GOT, T.Chol and CBG have roughly the same likelihood.

As for the results of women in their 40's, BMI values are almost over the upper limit in state[5] and state[6]. TG is over

150 in state[6]. T.Chol values in state[4], state[5] and state[6] are also over 200. In the 50's, there are few differences in BMI and SBP, however, a TG value in state[6] is over the threshold (200). Female HMMs do not have significant differences between states, but have almost the same results as male HMMs.

Next, transition probabilities were evaluated. Self-loop transitions and transitions within the same row are likely to have the largest probabilities. In contrast, the probability of transitions among different rows is less than 0.1 in most cases.

2. Comparison with health risk examination

In Japan, health risk examination is conducted based on health checkup results: e.g. 'A' for almost no risk, 'B' for small risk, and 'D1' and 'D2' for recommendation of treatment. The difference between 'D1' and 'D2' is that more detailed inspection is required for 'D2' as compared to 'D1'. In the following comparison and discussion, we treat 'D1' and 'D2' as 'D' since the difference is not crucial in this case. In addition, 'A' and 'B' are grouped into 'AB,' meaning healthy. Using an HMM, we can obtain a state transition sequence that corresponds to health checkup features by applying the Viterbi algorithm [3]: e.g. state[1] → state[1] → state[3] → state[3] → state[4] → end. We then compared the health risk examination results and the transition results. Note that we referred to the guidelines of the Japan society of Ningen dock [9] to conduct the health risk examination.

Table 2 – Comparison of health risk examination results with HMM state transition results

			health risk examination result				
			A	**B**	**C**	**D1**	**D2**
30's	state	1	34.0%	12.0%	41.4%	0.0%	12.0%
		2	33.8%	17.3%	37.7%	0.1%	11.0%
		3	20.1%	7.5%	49.0%	0.1%	23.2%
		4	12.2%	5.9%	49.0%	0.4%	32.5%
		5	1.0%	1.2%	25.6%	1.7%	70.4%
		6	2.1%	1.3%	29.8%	4.6%	62.2%
			A	**B**	**C**	**D1**	**D2**
40's	state	1	19.1%	6.5%	47.8%	0.1%	26.4%
		2	20.6%	10.6%	50.2%	0.1%	18.5%
		3	10.1%	4.7%	52.5%	0.5%	32.1%
		4	9.9%	5.9%	51.0%	1.8%	31.5%
		5	1.9%	0.6%	31.5%	2.5%	63.5%
		6	0.9%	0.8%	18.9%	7.6%	71.8%
			A	**B**	**C**	**D1**	**D2**
50's	state	1	10.6%	4.7%	61.2%	0.3%	23.2%
		2	5.0%	3.6%	55.1%	2.6%	33.7%
		3	6.8%	3.3%	55.9%	0.7%	33.4%
		4	4.3%	2.9%	47.8%	3.0%	42.1%
		5	0.9%	0.3%	22.9%	4.8%	71.0%
		6	1.2%	0.7%	22.1%	5.7%	70.3%

Table 2 shows the comparison of results for male data. In the 30's, in state[1] and state[2] the percentages of 'A' and 'B' (='AB') as well as 'C' are 40-50% respectively; 'Dl' and 'D2' (='D') are about l0%. In state[3] and state[4], 'AB' decreases to roughly 20%, 'C' increases to 50%, and 'D' increases to 30%. In state[5] and state[6], 'AB' becomes 1-2%, 'C 'and 'D' become approximately 30% and 70% respectively. In the 40's, 'AB' is 25-30%, 'C' is 50%, and 'D 'is 20 - 25% in state[1] and state[2]. In state[3] and state[4], 'AB' decreases whereas 'D' increases to 30%. Almost the same results are obtained in state[5] and state[6]. Finally in the 50's, 'AB', 'C' and 'D' are 10-15%, 55-60%, 25-35% respectively in state[1] and state[2]. The same tendencies are observed in the other states as in the 30's and 40's.

It is found that the three state groups have the same ratios. As compared with the result in Appendix Table 1-3, it is also observed that if BMI and SBP increase, the ratio of 'AB' decreases and that of 'D' increases.

Discussion

1. HMM parameters in detail

We investigated mean vectors of states in HMMs. According to the results reported in the previous section, it is clear that we can classify the six states into three groups by BMI, SBP and TG values:

- state[1] and state[2] – healthy states
  In these states, mean vectors have low values within the acceptable range. In these states most data have examination results of 'A' and 'B'. Thus these states indicate 'healthy' or 'normal' conditions.
- state[3] and state[4] – low-grade unhealthy states
  Mean vectors have relatively high values within the acceptable range, or sometimes exceed the upper limit. In the health risk examination results, half of the data in these states have a 'C' grade. So in these cases, it is not required to treat diseases immediately, but successive monitoring is essential.
- state[5] and state[6] – high-grade unhealthy states
  Mean vectors are out of the acceptable range. The examination results indicate a 'D' grade. In these conditions, it is strongly recommended to treat lifestyle-related diseases.

Both the mean vector investigation and the health risk examination results show the same classification tendency. From these discussions, it is concluded that our HMMs can model health conditions and health risk. It is remarkable that the HMMs had the same initialized parameters at the beginning; all the states had the same values. Therefore, only training data made the differences. This indicates that HMM can be utilized as a data mining method for the health checkup data. the results show that the most important risk factors for hypertension are SBP and TG, followed by BMI [2]. So it is natural for physicians to use these parameters to decide whether a patient has hypertension or hyperlipidemia. According to this previous work, we believe our results are reasonable.

We also further explored transition probabilities. As mentioned, the probabilities of transition by self-loop within a state or transition to the next state in the same row are larger than the other transitions. Because there are few drastic changes in only one year, and in most cases the inspection result is similar to the prior year, the above results are reasonable. It would be remarkable and informative if a state transition indicates a change of health risk level, since such a transition rarely occurs.

2. Possibility to risk estimation of life-style diseases

BMI and SBP are important inspection parameters for obesity and high blood pressure diseases. TG is also important for dyslipidemia disease [10][11]. Since these parameters strongly affected HMM parameters, our HMMs might estimate the risk of these diseases. Other inspection parameters could not have obvious differences compared with the acceptable ranges. Nevertheless, focusing on CBG in male results the following maximum values were obtained: 142.7 in the 30's, 161.8 in the 40's, and 158.2 in the 50's. Similarly, the minimum values were observed as: 89.9 in the 30's, 95.7 in the 40's, and 97.9 in the 50's. Since there are significant differences in these maximum and minimum values, if we employ FBG (Fasting Plasma Glucose) which could not use in this paper, further risk estimation may be available.

Conclusion

In this paper, we proposed the introduction of HMM to health checkup data analysis. Time-series health checkup features having eight inspection parameters were used, and HMMs with six states placed in a 2x3 matrix were employed. According to the experimental results that compared mean vectors in HMMs with the inspections' acceptable ranges and the health risk examination results, the model built is suitable for categorizing three risk levels using primarily BMI, SBP and TG values. The model also shows the possibility of estimating the risk of lifestyle-related diseases. The advantage of HMM is that it can model not only one-time health checkup data but also model time-series changes as state transitions. Note that the number of states and possible transitions in an HMM should be optimized for the data and the task. For example, taking the states and transitions of health risk levels in each age into account, it is necessary to determine a proper HMM topology for each case. Our future work includes the following: investigation of current risk estimation of adult diseases, estimation of future health risk by applying proposed HMMs, and building HMMs using health checkup data obtained from multiple medical centers or hospitals. In the third work, we will investigate the influence of the data on HMM parameters and health risk examination, as compared with HMMs built using the data obtained from only one medical center. We would like to address the above issues in our future research and its practical application.

Acknowledgments

This research was done with the support of Regional Innovation Strategy Support Program (City Area Program) founded by Ministry of Edition, Culture, Sports, Science in Japan.

References

[1] Akdag B, Camdeviren H, Degirmencioglu S, Fenkci S, Rota S, Sermez Y. "Determination of risk factors for hypertension through the classification tree method," Adv. Ther.. 2006; 23: 855-891.

[2] Chang C, Jiang BC, Wang C. "Using data mining techniques for multi-disease prediction modeling of hypertention and hyperlipidemia by common risk factors," ESWA. 2011; 38: 5507-5513.

[3] Motomura Y, Sato T. "Networks for uncertainty modeling: uncertain modeling," Journal of JSAI. 2000; 15: 575-582 (in Japanese).

[4] Arita M. "Bayesian network and bioinformatics," Journal of JSAI. 2002; 539-545 (in Japanese).

[5] Suyari H. "Introduction to Bayesian network (1)," Medical Imaging Technology. 2003; 21: 315-318 (in Japanese).

[6] Yu J. "Health condition monitoring of machines based on hidden Markov model and contribution analysis," IEEE T INSTRUM MEAS. 2012; 61: 2200-2211.

[7] Hayamizu S, Kinosada Y, Tamura S, Yamamoto K,. "Visual Analysis of Health Checkup Data Using Multidimensional Scaling," JACIII. 2012; 16: 1: 26-32.

[8] HTK Speech Recognition Toolkit. Available from: http://htk.eng.cam.ac.uk

[9] Japan Society of Ningen Dock. Available from: http://www.ningen-dock.jp/

[10] Nara M, Yamakado M. "Ningen dock follow-up guide," Bunkodo 2009.

[11] Ministry of Health, Labour Welfare. http://www.mhlw.co.jp/topics/bukyoku/kenkou/seikatu/

[12] Lin W, Orgun MA, Williams GJ. "Mining temporal patterns from health care data," Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science. 2002; 2454: 222-231.

Address for correspondence

Ryouhei Kawamoto
Graduate School of Engineering
Gifu University, Japan

Appendix Table 1 – An HMM parameter set for male 30's

mean values							
		state					
		1	**2**	**3**	**4**	**5**	**6**
BMI		21.9	21.1	26.1	24.7	28.5	28.0
SBP		125.5	123.0	124.7	122.5	137.5	132.5
Ht		45.7	45.5	43.7	42.3	47.4	46.3
PLT		25.0	24.2	26.1	25.4	25.2	25.9
GOT		22.3	19.7	18.3	18.5	35.4	26.4
T.Chol		191.2	196.1	148.3	177.5	213.1	237.9
TG		114.3	121.4	130.7	112.4	220.7	254.6
CBG		100.0	89.9	142.7	100.0	103.7	99.5
transition probabilities							
		state (to)					
		1	**2**	**3**	**4**	**5**	**6**
state (from)	**1**	0.71	0.09	0.20	-	0.00	-
	2	-	0.99	-	0.00	-	0.00
	3	0.00	-	0.59	0.33	0.08	-
	4	-	0.00	-	0.77	-	0.00
	5	0.00	-	0.02	-	0.54	0.44
	6	-	0.00	-	0.00	-	0.70

Appendix Table 2 – An HMM parameter set for male 40's

mean values							
		state					
		1	2	3	4	5	6
BMI		20.4	20.5	25.0	24.8	26.5	24.8
SBP		122.4	126.0	129.2	132.7	141.1	138.9
Ht		43.7	45.7	45.3	45.1	46.8	44.4
PLT		25.3	24.5	24.8	26.9	23.4	24.9
GOT		20.6	20.5	22.6	25.0	32.1	47.3
T.Chol		186.4	223.7	203.2	220.0	219.3	210
TG		97.4	120.0	167.9	217.8	245.8	349.5
CBG		98.1	130.1	99.1	95.7	161.8	124.2
transition probabilities							
		state (to)					
		1	2	3	4	5	6
state (from)	1	0.83	0.17	0.00	-	0.00	-
	2	-	0.29	-	0.00	-	0.00
	3	0.00	-	0.74	0.20	0.06	-
	4	-	0.04	-	0.61	-	0.03
	5	0.00	-	0.00	-	0.77	0.23
	6	-	0.00	-	0.01	-	0.05

Appendix Table 3 – An HMM parameter set for male 50's

mean values							
		state					
		1	2	3	4	5	6
BMI		20.6	21.1	25.5	25.2	22.4	22.5
SBP		133.9	127.7	137.2	139.1	142.9	142.6
Ht		43.7	43.2	45.6	44.7	42.6	41.7
PLT		24.7	24.6	24.6	24.2	23.9	24.4
GOT		23.1	20.2	25.7	24.5	53.8	40.3
T.Chol		202.2	206.0	216.3	206.2	199.0	197.8
TG		125.4	110.8	183.6	196.7	285.8	200.5
CBG		109.5	97.9	102.8	142.6	158.2	110.2
transition probabilities							
		state (to)					
		1	2	3	4	5	6
state (from)	1	0.59	0.27	0.14	-	0.00	-
	2	-	0.76	-	0.00	-	0.00
	3	0.00	-	0.75	0.25	0.00	-
	4	-	0.02	-	0.53	-	0.02
	5	0.00	-	0.00	-	0.65	0.35
	6	-	0.00	-	0.04	-	0.72

Appendix Table 4 – An HMM parameter set for female 40's

mean values							
		state					
		1	2	3	4	5	6
BMI		19.9	20.8	21.3	21.6	25.9	27.7
SBP		118.5	119.2	121.6	122.9	127.3	135.9
Ht		38.3	38.3	33.9	37.9	39.6	38.9
PLT		25.4	23.2	30.2	25.0	28.6	29.6
GOT		24.4	19.4	16.9	17.1	20.2	21.0
T.Chol		188.4	197.5	187.8	200.0	215.9	238.4
TG		92.4	65.3	93.0	118.4	94.3	154.8
CBG		107.9	95.9	97.1	93.5	88.9	93.7
transition probabilities							
		state (to)					
		1	2	3	4	5	6
state (from)	1	0.48	0.46	0.06	-	0.00	-
	2	-	0.76	-	0.06	-	0.00
	3	0.73	-	0.26		0.01	-
	4	-	0.00	-	0.55	-	0.00
	5	0.00	-	0.00	-	0.74	0.26
	6	-	0.00	-	0.00	-	0.63

Appendix Table 5 – An HMM parameter set for female 50's

mean values							
		state					
		1	2	3	4	5	6
BMI		21.8	20.3	21.1	23.5	23.5	24.7
SBP		134.5	124.6	121.4	127.5	128.8	135.1
Ht		37.1	38.7	40.1	40.4	38.9	40.4
PLT		25.9	23.6	23.3	24.8	25.1	27.1
GOT		20.5	21.0	23.4	36.1	20.1	21.0
T.Chol		208.4	202.8	247.5	234.0	220.1	233.9
TG		113.3	96.1	86.8	127.4	141.2	205.6
CBG		142.2	105.7	91.9	117.3	98.1	98.5
transition probabilities							
		state (to)					
		1	2	3	4	5	6
state (from)	1	0.45	0.49	0.06	-	0.00	-
	2	-	0.78	-	0.00	-	0.00
	3	0.00	-	0.75	0.25	0.00	-
	4	-	0.02	-	0.43	-	0.15
	5	0.00	-	0.00	-	0.74	0.26
	6	-	0.00	-	0.06	-	0.58