Method of Screening the Health of Persons with High Risk for Potential Lifestyle-related Diseases using LDA

Toward a Better Screening Method for Persons with High Health Risks

Keisuke Ogawa[1], Kazunori Matsumoto[1], Masayuki Hashimoto[1] and Ryoichi Nagatomi[2]
[1]*KDDI R&D Labs, Saitama, Japan*
[2]*Tohoku Graduate School of Biomedical Engineering, Miyagi, Japan*

Keywords: Latent Dirichlet Allocation, LDA, Metabolic Syndrome, Lifestyle-Related Disease.

Abstract: Recently, the number of patients with lifestyle-related diseases, such as diabetes mellitus, has increased dramatically. Lifestyle-related diseases are responsible for 60% of deaths in Japan. In order to screen persons at potentially high risk for these diseases, medical checkups for metabolic syndrome are used throughout Japan. Prediction and prevention of lifestyle-related diseases would yield a direct reduction in medical costs. However, many cases cannot be screened with a metabolic syndrome checkup. In this paper, we propose a new machine-learning-based screening method using medical checkup data and medical billings. By processing the medical data into a bag-of-words representation and classifying the health factors using latent Dirichlet allocation (LDA), the screening method achieves high accuracy. We evaluate the method by comparing the accuracy of predictions of the future incidence of the diseases. The results show that F-measure increases 0.17 compared with the conventional method. In addition, we confirmed that the proposed method classified persons with different health risk factors, such as a combination of metabolic disorders, hypertensive disorders, and mental disorders (stress).

1 INTRODUCTION

The number of patients with lifestyle-related diseases, such as diabetes mellitus, has increased dramatically. Lifestyle-related diseases are responsible for 60% of the deaths in Japan (Health, Labour and Welfare Statistics Association 2002). In general, the medical costs related to lifestyle-related diseases are expensive and occupy 30% of national medical expenditures (Ministry of Health, Labour and Welfare 2010). In Japan, to save medical costs, the health insurance society sponsors an annual medical checkup to screen persons with high health risks, such as metabolic syndrome. Metabolic syndrome is a combination of multiple metabolic disorders; specifically, a combination of obesity and hyperglycemia, hypertension, or dyslipidemia (WHO 1999; Ogushi 2007). Patients with metabolic syndrome have a greater possibility of acquiring lifestyle-related diseases such as diabetes mellitus or cardiovascular disease (K.G.M.M. Alberti et al. 2006). However, many cases cannot be screened with a metabolic syndrome checkup. In our investigation of a specific health insurance society, only 14% of people with newly diagnosed lifestyle-related diseases have metabolic syndrome. This fact indicates that there are other risk factors for lifestyle-related diseases (Maria D. Llorente et.al. 2006; H. Klar Yaggi 2006). As a result, many people at high risk for these diseases are missed. So the cost effect of annual medical checkups is limited. In this paper, to solve this problem, we propose the machine-learning-based screening method that uses not only annual medical checkup data, but also medical billings. They involve the personal history of diseases that infers the various health risk factors.

2 LIFESTYLE-RELATED DISEASES AND SCREENING

2.1 Lifestyle-related Diseases

The Ministry of Health, Labour and Welfare in Japan defined a list of lifestyle-related diseases (Mizushima research team of the Ministry of Health, Labour and Welfare 2007). To prevent these diseases, Japan developed the criteria to diagnose metabolic syndrome in the Japanese.

Ogawa K., Matsumoto K., Hashimoto M. and Nagatomi R..
Method of Screening the Health of Persons with High Risk for Potential Lifestyle-related Diseases using LDA - Toward a Better Screening Method for Persons with High Health Risks.
DOI: 10.5220/0005250905020507
In *Proceedings of the International Conference on Health Informatics* (HEALTHINF-2015), pages 502-507
ISBN: 978-989-758-068-0

2.2 Metabolic Syndrome

In Japan, the decision to diagnose metabolic syndrome is based on the following criteria: *waist circumference, blood sugar level, HbA1c, systolic blood pressure, diastolic blood pressure, triglycerides, HDL cholesterol, and LDL cholesterol,* and people are classified into the metabolic syndrome group, preliminary group, and normal group. The screening method is explained in Figure 1 (Wataru et al. 2008).

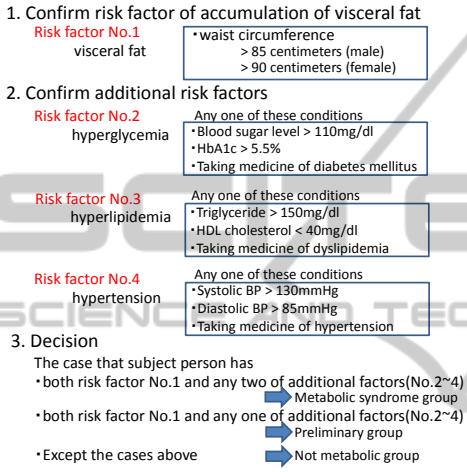

Figure 1: Screening for metabolic syndrome in Japan.

This screening checks only the combination of four risk factors: the combination of obesity and hyperglycemia, hypertension, or dyslipidemia. However, there are other risk factors reported, such as mental disorders (Maria D. Llorente et al. 2006; H. Klar Yaggi 2006). Therefore, many persons at high risk for these diseases cannot be screened by this method.

3 PROPOSED METHODS

To realize a screening method that can identify the various health risk factors, we propose a machine-learning-based screening method using medical checkup data and medical billings. In general, medical checkup data involve blood tests that indicate health status, and medical billings involve personal medical history that involves information about all diseases. Technical knowledge is needed to make the rules using medical billings for decisions like metabolic syndrome. Thus, we applied machine-learning techniques to handle the huge volume of data statistically. In this paper, we propose using latent Dirichlet allocation (LDA) (Blei et al. 2003). LDA is a kind of topic model where machine-learning techniques are mainly used for natural language processing. With LDA, we can model data (such as documents) as a mixture of multiple topics more precisely than the mixed Gaussian distribution such as k-means.

3.1 Latent Dirichlet Allocation

LDA has the advantage of easily modeling documents and is now applied to various data mining tasks, such as information retrieval and voice recognition (Ishiguro et al. 2012; Otsuka et al. 2012), data visualization, and image processing (Fei-Fei et al. 2005; Wang et al. 2009; Wang and Mori 2009; Niebles et al. 2008). LDA infers the topic of documents containing many words from a document set by assigning each word to a certain topic. In LDA, documents are handled as a bag-of-words representation, and these documents are analyzed according to a word-topic probability matrix (ϕ matrix) and topic-document probability matrix (θ matrix). Some approximation techniques estimate the parameters of LDA, such as variational Bayes (Blei et al. 2003), Gibbs sampling (Griffiths and Steyvers 2004), and collapsed variational Bayes (The et al. 2006). In this paper, we use the Gibbs (Griffiths and Steyvers 2004) technique. Now, we explain and review LDA following the notation of (Griffiths and Steyvers 2004). Let there be T topics and $w_1, \ldots w_D$ represent bag-of-words representations for each D document. (Document d becomes $w_d = \{w_{1,d}, \ldots w_{n,d}\}$ where N be number of all types of words.) Also, let z_i be the hidden topic from which w_i is generated, $\phi_j^{(w)} = p(w|z = j)$, and $\theta_j^{(d)} = p(z = j)$ for document d. LDA involves the following generative model:

$$\theta \sim \mathrm{Dir}(\alpha)$$
$$z_i|\theta^{(d_i)} \sim Mult(\theta^{(d_i)})$$
$$\phi \sim \mathrm{Dir}(\beta)$$
$$w_i|z_i, \phi \sim Mult(\phi_{z_i})$$

Dir and Mult mean the Dirichlet distribution and multinomial distribution, respectively. α and β are hyperparameters for the document-topic and topic-word Dirichlet distributions, respectively. Here we assume α and β are scalars resulting in symmetric Dirichlet priors. Given observed words, we have to infer the hidden topics. To approximate this posterior, we resort to a Markov chain Monte Carlo (MCMC) sampling scheme, specifically a collapsed Gibbs sampling:

$$P(z_i = k | z_{-i}, w, \alpha, \beta) \propto \left(\frac{n_{-i,k}^{(d)} + \alpha}{\sum_u^T (n_{-i,u}^{(d)} + \alpha)} \right) \left(\frac{n_{-i,k}^{(w_i)} + \beta}{\sum_{w'}^N (n_{-i,k}^{(w')} + \beta)} \right)$$

where $n_{-i,v}^{(d)}$ is the number of times topic k is used in document d, and $n_{-i,k}^{(w_i)}$ is the number of times word w_i is generated by topic k. The $-i$ notation signifies that the counts omit the value of z_i. Here, we regard the document as a person, words as the frequency of word counts in medical billings and discretized medical checkup data, and the topic as a health factor, respectively.

3.2 Feature Extraction

In this subsection, we show the feature extraction from medical checkup and medical billing data. Medical billing is a receipt for medical activity that the health insurance society uses to pay the medical cost to the medical institution. It involves the name of the disease, medicine, and inspection as decided by physicians. We use the frequency of the disease, medicine, and inspection from medical billings in one year as a feature vector of LDA. The maximum frequency is 12 because medical billings are generated once a month. Table 1 shows examples of the features of medical billings.

Table 1: Features of medical billings.

Disease		Medicine		Inspection	
Name	Frequency	Name	Frequency	Name	Frequency
Diabetes Mellitus	8	Mucodyne 100mg	8	Albumin	3
Allergic rhinitis	5	Zyloric	6	Blood urea nitrogen	3
Hyperuricemia	3	Seibule	6	Creatinie	3
…	…	…	…	…	…

For medical checkups, blood tests, measurements, and questionnaires about lifestyle and medicines are involved. In this paper, we use only blood test and measurement data (Table 2).

Table 2: Medical checkup data.

Measurement data	Blood test data
Height	triglyceride
Weight	HDL cholesterol
BMI	LDL cholesterol
Waist circumference	GOT
Diastolic pressure	GPT
Systolic pressure	γ-GTP
	Fasting blood sugar level
	HbA1c

Also, because medical checkup data are continuous data different from medical billings, we transform the data into a bag-of-words representation that can be used in the LDA (Figure 2).

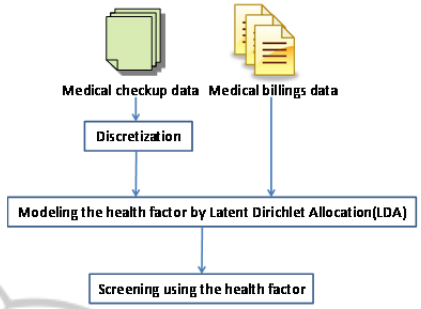

Figure 2: Data preparation in proposed method.

The blood test and measurement data are not always in a normal distribution, so we make the data discrete using a cumulative distribution. Also, in general, blood test and measurement data have multiple medical meanings according to the value, so we represent the data as a mixture of the frequency of three words: high, normal, and low. Specifically, we divide the checkup data into 13 classes and assign each pair of words in Figure 3.

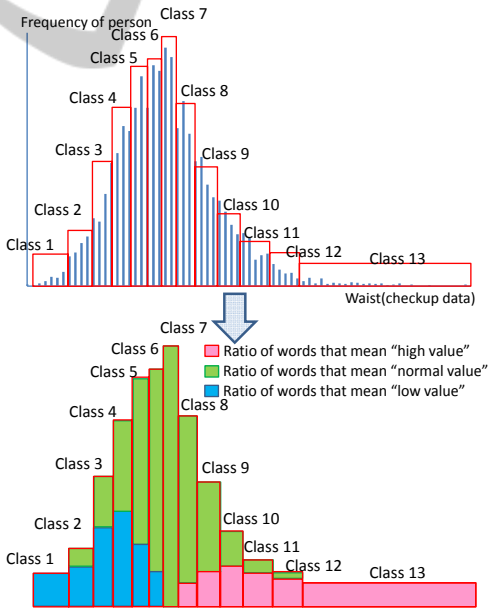

Figure 3: Bag-of-words representation of medical checkup data.

For example, in class 1, the bag-of-words representation becomes {high-value-words of waist, normal-value-words of waist, low-value-words of waist} = {0,0,12}, in class 2, the bag-of-words representation becomes {0,1,11} and so on.

4 EXPERIMENT AND RESULTS

4.1 Arrangement of Experiments

To compare the prediction accuracy for the future incidence of the diseases by the proposed method with that by conventional metabolic syndrome decisions, we perform experiments. We use the medical checkup data and medical billings of a specific common insurance society with about 30,000 members. We used data on men in 2010 (4404 members) for training LDA. To evaluate the methods, we measured precision and recall for persons who did not have medical billings for lifestyle-related diseases in 2009 but who developed lifestyle diseases in 2011–2013. Also, we set the hyperparameter of LDA in Table 3.

Table 3: Hyperparameters of LDA.

α(Hyperparameter of dirichlet prior dist.)	β(Hyperparameter of dirichlet prior dist.)	L(Learning count in Gibbs Sampler)
0.1	0.01	40000

4.2 Comparison with Conventional Metabolic Syndrome Decision

We show the result of LDA (topic number is 3) and metabolic syndrome in Table 4-1 and Table 4-2. Persons are classified into 3 groups by metabolic syndrome decision, so we use the same 3 topics(groups) in LDA. In LDA, we assign the topic of the highest probability to each person. In Table 4-1 and Table 4-2, extraction rate of new cases of lifestyle-related diseases in 2010-2013(ratio of numbers for each group to number of new cases of lifestyle-related diseases) is shown as precision. (We exclude people who already had medical billings for lifestyle-related diseases in 2009 (already developed).)

Table 4-1: Conventional screening result.

Group(Metabo decision)	Number	Number of outbreaks†	Number of excluded cases‡	Precision	Recall	F-measure
Normal	2533	759	576	0.30	0.66	0.41
Preliminary	521	216	209	0.41	0.19	0.26
Metabolic	279	174	289	0.62	0.15	0.24

Table 4-2: Screening result of the proposed method.

Group(Main Topic Number)	Number	Number of outbreaks†	Number of excluded cases‡	Precision	Recall	F-measure
1	1803	484	261	0.27	0.42	0.33
2	66	36	379	0.55	0.03	0.06
3	1464	629	431	0.42	0.55	0.48

†Number of new cases of lifestyle-related diseases in 2010-2013.
‡Number of lifestyle-related disease patients in 2009.

As a result, topic number 3 for LDA has the highest F-measure (topic number 3). In this result, we think topic number 3 is a high-risk cluster that involves many persons at potentially high risk for lifestyle-related diseases. In addition to this result, we can see the health risk factors. Now, we look into the ϕ matrix of these topics. In Tables 5-1 and 5-2, the major terms for disease in the ϕ matrix and average frequency for a person and averaged medical checkup data for persons in each topic are shown.

Table 5-1: Major disease terms in ϕ matrix and averaged frequency for a person (topic number 3).

Major words of disease in Φ matrix	Averaged Frequency for a person		
	Topic 1	Topic 2	Topic 3
Allergic rhinitis	0.39	1.40	0.40
Myopic astigmatism	0.19	0.55	0.13
Allergic conjunctivitis	0.13	0.28	0.14
…	…	…	…
Hypertension	0.13	3.1	0.42
Diabetes Mellitus	0.09	2.5	0.19
Hyperuricemia	0.09	1.7	0.18
…	…	…	…
Acute bronchitis	0.13	0.31	0.15
Acute upper respiratory inflammation	0.10	0.22	0.12

Table 5-2: Averaged medical checkup data for persons in each topic (topic number 3).

	Topic 1	Topic 2	Topic 3
Height	170.7	169.9	171.7
Weight	63.5	71.7	76.0
BMI	21.8	24.8	25.8
Waist	78.4	86.3	89.2
Diastolic pressure	117.6	124.0	126.5
Systolic pressure	75.0	80.6	82.8
triglyceride	99.7	146.4	166.5
HDL cholesterol	61.2	55.3	51.6
LDL cholesterol	116.3	117.0	130.4
GOT	19.9	25.9	26.6
GPT	19.9	31.7	36.3
γ-GTP	39.1	61.6	66.8
Fasting blood sugar level	90.2	105.5	96.6
HbA1c	5.0	5.6	5.3

For example, in topic 2, the probability of lifestyle diseases, such as hypertension, diabetes mellitus, and hyperuricemia, is the highest but other topics are not. Looking into Table 5-2, almost all of the medical checkup data (except height, fasting blood sugar level, and HDL cholesterol) in topic 3 are higher than topic 1. So, from these results, we can approximate that topic number 1 is the healthy people topic, topic number 2 is the lifestyle disease patient topic, and topic number 3 is the unhealthy people topic.

4.3 Proposal with More Topic Numbers

In the proposed method, we can determine the topic numbers at will. Here, we show the result of LDA where the topic number increases. In Table 6, the result of topic number 10 is shown. Here, also, we exclude people who already had medical billings for lifestyle-related diseases in 2009 (already developed).

Table 6: Screening result of proposed method (number of topics: 10).

Main Topic Number	Number	Number of outbreaks	Number of excluded cases	Precision	Recall	F-measure
1	550	139	62	0.25	0.12	0.17
2	301	87	56	0.28	0.08	0.12
3	1079	478	370	0.44	0.42	0.43
4	136	46	19	0.34	0.04	0.07
5	79	33	76	0.42	0.03	0.06
6	43	24	145	0.56	0.02	0.04
7	337	100	44	0.30	0.09	0.14
8	732	214	135	0.29	0.19	0.23
9	7	6	183	0.86	0.01	0.02
10	69	22	121	0.31	0.02	0.04

In the result, topics 3, 4, 5, 6, and 9 reflect high precision rates. Looking into major terms for disease in the ϕ matrix and average frequency for a person (in Table 7-1) and averaged medical checkup data (in Table 7-2), we can see that the risk factors are separated more precisely.

Table 7-1: Major disease terms in ϕ matrix and averaged frequency for a person (topic number 10).

Topic 3		Topic 4		Topic 5	
Words	Frequency	Words	Frequency	Words	Frequency
Allergic rhinitis	0.43	Rib fracture	0.01	Insomnia	2.67
・・・		・・・		Melancholy	2.51
				Gastritis	0.92
				・・・	

Topic 6		Topic 9	
Words	Frequency	Words	Frequency
Stomach ulcer :	1.59	Hypertension	4.7
Diabetes mellitus :	1.95	Hyperlipidemia	4.81
Liver dysfunction	0.81	Diabetes mellitus	3.47
・・・		・・・	

In topics 3 and 4, there are no diseases with high probability. Topic 5 has insomnia, melancholy, and the digestive disease gastritis (Table 9-1). On the other hand, the checkup data of topic 3 are much higher than that of others, especially the metabolic disorder related items. Also, in topic 4, only blood pressure is higher (Table 9-2). So, it seems that the risk factors for topics 3 and 4 are abnormal checkup values (combination of metabolic disorders and hypertensive disorders, respectively), and the risk factor for topic 5 is stress (mental disorders).

Table 7-2: Averaged medical checkup data for persons in each topic (topic number 10).

	Topic 3	Topic 4	Topic 5	Topic 6	Topic 9
Height	172.5	168.3	170.2	170.3	169.2
Weight	77.9	65.2	67.2	69.2	72.4
BMI	26.2	23.0	23.2	23.8	25.3
Waist	90.6	80.5	82.1	83.8	87.5
Diastolic pressure	129.4	126.5	117.7	120.0	125.1
Systolic pressure	85.1	81.0	76.6	77.7	80.8
triglyceride	168.3	81.5	127.7	136.8	159.2
HDL cholesterol	51.4	65.1	57.4	55.8	55.2
LDL cholesterol	130.6	97.1	121.3	115.6	114.0
GOT	27.3	20.9	22.2	23.4	26.8
GPT	38.2	21.0	24.5	25.7	33.5
γ-GTP	67.9	43.0	48.7	50.7	65.9
Fasting blood sugar level	98.3	86.8	89.1	102.1	110.9
HbA1c	5.3	4.9	5.0	5.4	5.9

Also, topics 6 and 9 have different diseases. In topic 6, the probability of digestive disease and liver disease is higher, and in topic 9, the probability of cardiovascular disease is higher. So, from these results, we assume that topics 3, 4, and 5 are unhealthy people topics with different health risk factors (combination of metabolic disorders, hypertensive disorders, and mental disorders) and that topics 6 and 9 are lifestyle disease patient topics with different lifestyle-related diseases (digestive, liver and, cardiovascular diseases). Thus, we think the topics are separated by different health risk factors.

5 CONCLUSIONS

In this paper, we propose a new screening method using latent Dirichlet allocation (LDA). By making the medical checkup data and medical billings into a bag-of-words representation, the screening method achieves high accuracy. We evaluate the method by comparing the accuracy of predictions for the future incidence of the diseases. The result shows that F-measure increases 0.17 compared with the conventional method. In addition, by increasing the topic numbers, we confirmed that the proposed method classified high risk persons into each of the different health factors of combination of metabolic disorders, hypertensive disorders, and mental disorders. Thus, we think the topics are separated by different health risk factors.

REFERENCES

Mizushima Research Team of the Ministry of Health,

Labour and Welfare, Report of Lifestyle-Related Disease Administration Using Medical Checkup and Billing Data 2007.

Health, Labour and Welfare Statistics Association, Trend of National Health, *Journal of Health and Welfare Statistics*: 449-453, 2002.

Ministry of Health, Labour and Welfare, Overview of national medical cost in 2010, http://www.mhlw.go.jp/toukei/saikin/hw/k-iryohi/10/, 2010.

K. G. M. M. Alberti, P. Zimmet and J. Shaw, Metabolic syndrome—a new world-wide definition. A Consensus Statement from the International Diabetes Federation, Diabetic Medicine, 23:469-480, 2006.

Maria D. Llorente and Victoria Urrutia, Diabetes, Psychiatric Disorders, and the Metabolic Effects of Antipsychotic Medications, Journal of Clinical Diabetes, Vol.24 No.1:18-24, 2006.

H. Klar Yaggi, Andre B. Araujo and John B. McKinlay, Sleep Duration as a Risk Factor for the Development of Type 2 Diabetes, Journal of Diabetes Care, Vol.29 No.3:657-661,2006.

World Health Organization (WHO), Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications, Report of a WHO consultation, 1999.

Ogushi, Metabo no wana, Kadokawa shinsho, 2007.

Wataru Sakamoto, Naoki Isogawa, and Masashi Goto, Statistical problem of Japanese metabolic syndrome criteria, The Behaviometric Society of Japan, 69:177-192, 2008.

David M. Blei. Ng, A. Y. and Jordan, M. I., Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993-1022, 2003.

Fei-Fei.L. and Perona P., A Bayesian Hierarchical Model for Learning Natural Scene Categories, CVPR 2005. IEEE Computer Society Conference, 2005.

Chong Wang, David M. Blei, and Li Fei-Fei, Simultaneous image classification and annotation, CVPR 2009, IEEE Conference, 2009.

Yang Wang, and Greg Mori, Human Action Recognition by Semi-latent Topic Models, Pattern Analysis and Machine Intelligence, IEEE Transactions, 2009.

Niebles et al., Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, *International Journal of Computer Vision*, Vol. 79:299-318,2008.

Griffiths T.L. and Steyvers M., Finding scientific topics, proceedings of the National Academy of Science, 101:5228-5235, 2004.

Yee Whye Teh et al., A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation, *NIPS*, Vol. 19, 2006.

Ishiguro et al., Probabilistic Speaker Diarization with Bag-of-Words Representations of Speaker Angle Information, IEEE Transcription, Vol. 20(2):447-460, 2012.

Otsuka et al., Bayesian Unification of Sound Source Localization and Separation with Permutation Resolution, Proc. AAAI, 2012.