Temporal Data Classification of Diabetes Mellitus on Health Examination Data of Factory Employees

Kritsada Sriphaew, Somphop Pathomnop, and M. L. Kulthon Kasemsan

Abstract—**Diabetes mellitus is a chronic disease that reduces quality of life since it often causes other complications such as heart disease, stroke, high blood pressure, liver disease, kidney disease, neuropathy and the loss of some organs in the body. This work proposes a temporal features extraction model which extracts the features embedded in historical data such as health examination data for classification. The proposed model can be used with any promising classification methods such as Naïve Bayes, Logistic Regression, C4.5 (J48), Bagging and SVMs. The extended temporal features can improve the accuracy and F-measure of the classification. This work evaluates the proposed method on health examination data during 2004-2010 (7 years) of factory employees in Thailand. It consists of 43,523 employees in total where 28,808 employees have only one record and 14,715 employees is examined more than once. Features used for diabetes classification are categorized into three groups: Physical Examination, Urinalysis and Biochemistry. The experiments show that data with temporal features gives the 97.25% accuracy and 0.57 F-measure which is a lot higher than data without temporal features.**

Index Terms—**Temporal model, classification, diabetes, data mining, healthcare.**

I. Introduction

The Centers for Disease Control and Prevention of the United States, according to national diabetes fact sheet 2011 [1], discussed the situation of diabetes that it affects 25.8 million people (including children and adults) in the United States. Diabetes accounted for 8.3 % of the total population. Regarding this number, 18.8 million were diagnosed, while 7 million were not. About 1.9 million people aged 20 years or older were newly diagnosed in 2010. There is no doubt that why many researchers have aware of serious disease that causes loss of life or disability. Therefore, the idea that we should find a way to alarm the risk of diabetes becomes interesting. The concept is to find persons who are likely to have diabetes, surveillance then including monitoring those people who already have diabetes. This kind of alarm can help to reduce a huge cost of medical examination and treatment.

A number of studies have been advantage of data mining techniques in the diabetes domain. Several research [2, 3, 4, 5] are work on diabetes prediction using classification methods, such as Naïve Bayes, Logistic Regression, decision tree, Boosting and support vector machines (SVMs). B. A. Tama, et.al. [5] examined the factors that cause the diabetes. They investigated on medical records of patients from public hospitals in Indonesia during 2008-2009 where the patients are at least 10 years of age with the total of 435 patients, 347 patients (79.8%) are diabetes and 88 patients (20.2%) are not diabetes. The features used for investigation are gender, body mass index (BMI), blood pressure (BP), hyperlipidemia, fasting blood sugar (FBS), instant blood sugar, family history, diabetes gestational history, habitual smoker, plasma insulin and age. The results were summarized that the risk factors that affect diabetes are: 1) habitual smoker, 2) gestational history and 3) plasma insulin. There is no distinct accuracy of applying different classification methods. B. H. Cho, et.al. [1] Studied the risk factors that affect the incidence of diabetes in kidney, which is mostly the cause of death of the patients. The proposed method was to apply to find the best features set for diabetes prediction. SVMs provide accuracy of prediction better than the conventional statistic (t-test, x^2-test, variance). Data were derived from 10 years (1996-2005) clinical data of 292 patients with diabetic kidney. There are 184 features from both medical and clinical, such as physical examination and biochemistry. Their work can capture the important features that tend to be a risk factor for kidney disease in which 39 features get the highest ROC (Receiver Operating Characteristics) by means of SVMs for features selection to reduce the redundancy of features.

Instead of testing blood sugar level in plasma (glucose plasma) to diagnosis the diabetes, K. Takahashi et.al. [4], tested diabetes by means of hemoglobin (HbA1C) for their 4 years research. The results showed that hemoglobin (HbA1C) is useful in predicting diabetes. Additional features, i.e., Aminotransferase, γ-Glutamyl Transpeptidase are also useful for prediction. Another research work to predict the chance of diabetes patients in getting heart disease was presented in [7] by employing Naïve Bayes and finding a feature set that best constructs the prediction model. A set of such features are gender, age, genetic (family heredity), weight, blood pressure, fasting blood sugar level, post prandial blood glucose level (test blood sugar levels after eating) and HbA1C (glucose level of the hemoglobin a-C, 4months).

Data classification is typically based on the data recorded at the same time or at any time. However, the research by R. Peter et.al. [6] Uses the data in the present analysis to predict what will happen in the future. The research worked on weather forecasting using meteorological data from Texas Commission of Environmental Quality and strains of influenza from the Google Flu Trends. Regarding weather forecasting, results showed that using all 40 features of 886 instances, the accuracy of data classification using temporal data was improved where either SVMs or ID3 was used as

Manuscript received February 29, 2012; revised April 30, 2012.

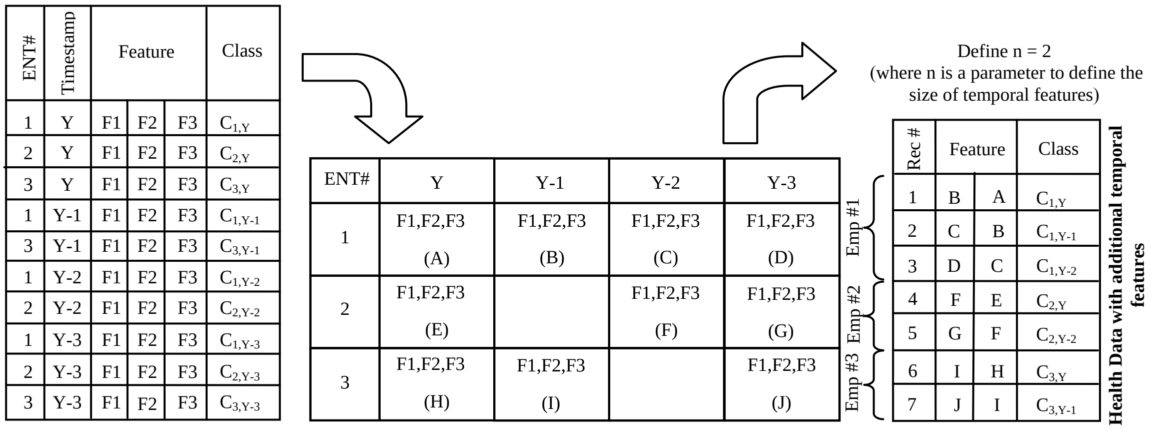
Authors are with the Faculty of Information Technology, Rangsit University, Pathumthani, Thailand (e-mail: s.kritsada@it.rsu.ac.th, tel.: +66-2-9972200 ext. 4068; fax: +66-2-9972200 ext.4076).

classification methods. They concluded that the classification with temporal data model provides more accurate results than the typical data.

In this work, we present an empirical study of classification of diabetes mellitus on health examination data of factory employees in Thailand. Section 2 described our proposed a temporal feature extraction model. A design of experiments to test the proposed method is given in section 3. Experimental results are shown in section 4 and conclusion is presented in section 5.

II. Temporal Features Extraction Model

Temporal data is the data inherently with time. An entity in the data may have different values over the time. An entity is an abstract concept of object such as person, animate object or thing. Health examination data is also a temporal data. It collects a historical data which is stored from the past to the present. A person may examine their health once or twice a year. Especially in some businesses such as factory, health check-up is periodically provided for employees. It is well-known that this historical data can help in medical diagnosis. Therefore, we propose a model that employs both current and historical data for classification called temporal feature extraction model. Although this work focuses on the domain of health examination data the proposed method can be applied to any temporal data.



ENT#	Timestamp	Feature			Class
1	Y	F1	F2	F3	$C_{1,Y}$
2	Y	F1	F2	F3	$C_{2,Y}$
3	Y	F1	F2	F3	$C_{3,Y}$
1	Y-1	F1	F2	F3	$C_{1,Y-1}$
3	Y-1	F1	F2	F3	$C_{3,Y-1}$
1	Y-2	F1	F2	F3	$C_{1,Y-2}$
2	Y-2	F1	F2	F3	$C_{2,Y-2}$
1	Y-3	F1	F2	F3	$C_{1,Y-3}$
2	Y-3	F1	F2	F3	$C_{2,Y-3}$
3	Y-3	F1	F2	F3	$C_{3,Y-3}$

a) Original relational table.

ENT#	Y	Y-1	Y-2	Y-3
1	F1,F2,F3 (A)	F1,F2,F3 (B)	F1,F2,F3 (C)	F1,F2,F3 (D)
2	F1,F2,F3 (E)		F1,F2,F3 (F)	F1,F2,F3 (G)
3	F1,F2,F3 (H)	F1,F2,F3 (I)		F1,F2,F3 (J)

b) Temporal view of table regarding entities.

Define n = 2
(where n is a parameter to define the size of temporal features)

	Rec #	Feature		Class
Emp #1	1	B	A	$C_{1,Y}$
	2	C	B	$C_{1,Y-1}$
	3	D	C	$C_{1,Y-2}$
Emp #2	4	F	E	$C_{2,Y}$
	5	G	F	$C_{2,Y-2}$
Emp #3	6	I	H	$C_{3,Y}$
	7	J	I	$C_{3,Y-1}$

Health Data with additional temporal features

c) Extracted temporal features

Fig. 1. Temporal feature extraction model.

A model to extract temporal features starts by taking original relational table with entity number and timestamp of data as input. Assume that the relational table is as shown in figure 1a), one instance has three features (F1, F2, F3) and one predefined class. One entity (ENT#) can have different values over the time which reflected as several records in the table. For example, ENT#1 occurs in timestamp Y, Y-1, Y-2 and Y-3. The data can be seen in temporal view as in figure 1b). Note that it is not necessary that an entity will be the data on every timestamp. Therefore, we handle this missing by shifting the value when we extract as temporal feature node as in figure 1c). It is noted that we need to define parameter n which is a user-defined value for extracting n-consecutive temporal data for classification. The n parameter is studied in section 4.

III. Experimental Setting

A. Dataset

The dataset used in this research is health examination data during 2004 - 2010 (7 years) of factory employees. Total amount of data consists of 43,523 employees: 28,808 employees have only one record of health examination and 14,715 employees are examined more than once in which the total number of records in this case is 41,186. Features collected from health examination can be categorized into three groups: physical examination (F1), urinalysis (F2) and biochemistry (F3). The details of each feature group are given in table 1.

TABLE I: Three Groups of Features For Health Examination Data and Their Discretization.

Group of Features	Feature Name	Nominal Value	Group Of Features	Feature Name	Nominal Value
F1	Age *	< 45, 45 – 49, >= 50	F2	UPr	Negative, Trace, 1+, 2+, 3+, 4+
	Sex	Male, Female		USu	Negative, Trace, 1+, 2+, 3+, 4+
	Weight *	<= 50, 51 – 99, >= 100	F3	CRE * (Creatinine)	Normal, High – Normal, High (<=1.50) (1.51-3.90)(>=3.91)
	Height *	< 150, 151 – 169, >= 170		GPT *	Normal, High – Normal, High (<=45) (46-89) (>=90)
	BMI ** (Body Mass Index)	Thin, Normal, Obesity (<=18) (19-25) (>=26)		CHO * (Cholesterol)	Normal, High – Normal, High (<=200) (201-240) (>=241)
	BP_S *** (Systolic Blood Pressure)	Normal, High – Normal, Hypertension (<130) (130-139) (>=140)		TG * (Triglyceride)	Normal, High – Normal, High (<=170) (171-400) (>=401)
	BP_D *** (Diastolic Blood Pressure)	Normal, High – Normal, Hypertension (<85) (85-89) (>=90)	Class	FBS_dm	Non-Diabetes, Diabetes
	Pulse *	<= 60, 61 – 79, >= 80			
	PE_Nor	Normal, Abnormal			

Note: The star marks given after each features inform the source for discretization:

* means the discretized value of such feature is defined by expert,

** means the discretized value of such feature is guided by WHO BMI classification [8],

*** means the discretized value of such feature is guided by WHO-ISH guideline 2003 [9].

Since some features are numeric but some classification methods (i.e., Naïve Bayes) could not support numeric features, we employ the discretized values by using the WHO [8, 9] guideline and asking the experts. The features and their discretized values including their sources are given in table 1.

However, there is an unstable conclusion in the research works about which feature groups are the best for diabetes classification. Therefore, we take all of these feature groups into account for studying the effect of each feature groups to the performance of diabetes classification.

B. Classification Methods and Tools

Several classification methods [2, 4, 5, 7] have been applied to the task of diabetes classification, but there is no benchmark to show which method is the most effective. Therefore, we investigate the promising classification methods which were applied in several works on diabetes domain, i.e., Naïve Bayes [7], Logistic Regression [4], C4.5 (or J48 in WEKA Tools) [5], Bagging [5], and SVMs [2]. In this work, we use WEKA [10] as a tool, and all experiments are conducted with 10-fold cross validation.

IV. Results and Discussions

Classification accuracy and F-measure are used for performance study. Table 2 shows the among three features groups including their combinations. Table 3 shows the comparison of data with and without temporal features where n is a parameter of temporal feature extraction model. For example, T3 is the case where n is equal to 3, i.e., three consecutive health examination records of a person are encoded as one instance in temporal model. Figure 3 shows the area under the curve (AUC) of ROC in each case of n parameter.

TABLE II: Comparison of Data With (T2-T7) and Without (Muti-All) Temporal Features.

Method	Evaluation	F1	F2	F3	F1+F2	F1+F3	F2+F3	F1+F2+F3
N.Bayes	Accuracy	94.42	97.17	95.85	96.48	92.70	97.13	94.82
	F-measure	0.13	0.51	0.01	0.46	0.21	0.50	0.41
Logistic R.	Accuracy	95.91	97.11	95.91	97.17	95.90	97.15	97.15
	F-measure	0.00	0.49	0.00	0.50	0.01	0.50	0.50
C4.5(J48)	Accuracy	95.91	97.17	95.91	97.16	95.91	97.17	97.16
	F-measure	0.00	0.51	0.00	0.51	0.00	0.51	0.50
Bagging	Accuracy	95.91	97.16	95.91	97.16	95.90	97.16	97.16
	F-measure	0.00	0.50	0.00	0.50	0.01	0.51	0.50
SVMs	Accuracy	95.91	97.06	95.91	96.90	95.91	96.97	96.97
	F-measure	0.00	0.48	0.00	0.43	0.00	0.45	0.45

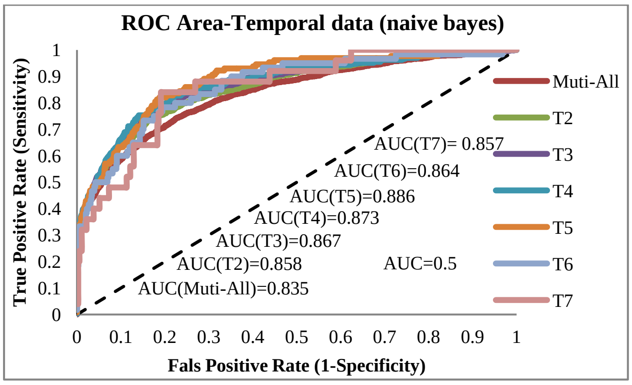


Fig. 2. ROC and AUC using naïve bayes.

TABLE III: Compare to Evaluate Data Classification with Temporal.

Algorithm	Evaluation	Muti-All	T2	T3	T4	T5	T6	T7	Avg.
N. Bayes	Accuracy	94.41	89.42	86.27	85.85	85.60	86.86	86.40	**86.73**
	F-Measure	0.41	0.34	0.33	0.34	0.36	0.36	0.33	**0.34**
Logistic R.	Accuracy	96.96	97.19	97.01	97.01	95.27	89.87	87.41	**93.96**
	F-Measure	0.50	0.56	0.59	0.61	0.55	0.37	0.22	**0.48**
C4.5 (J48)	Accuracy	96.94	97.16	96.99	97.08	96.91	96.79	94.71	**96.61**
	F-Measure	0.50	0.56	0.59	0.61	0.66	0.67	0.43	**0.59**
Bagging	Accuracy	96.96	97.25	97.01	97.19	96.68	96.39	94.96	**96.58**
	F-Measure	0.50	0.57	0.58	0.62	0.63	0.60	0.41	**0.57**
SVMs	Accuracy	96.81	97.20	97.12	97.14	96.59	96.59	92.44	**96.18**
	F-Measure	0.47	0.55	0.59	0.61	0.62	0.65	0.38	**0.56**

TABLE IV: The Number of Instances with (T2-T7) and Without (Muti-All) Temporal Features.

Data	Number of instances	%
Muti-All	41,186	100.00
T2	26,471	64.27
T3	11,756	28.54
T4	5,378	13.06
T5	2,201	5.34
T6	997	2.42
T7	397	0.96

The results show that urinalysis (F2) gives the highest accuracy and F-measure. Biochemistry (F3) is more efficient than physical examination (F1) in every method except naïve bayes (table 2).

The result in table 3 shows that the data with temporal features (T2-T7) has higher accuracy than non-temporal data (Muti-All) with the highest accuracy at T2. Logistic Regression, C4.5 (J48), Bagging, SVMs have the accuracy of 97.19%, 97.16%, 97.25%, 97.20%, respectively. Naïve Bayes does not found any advantage on temporal data.

In table 3, the experiments are done on data derived from different features sets. Muti-All is the data without temporal features. In the case of Muit-All, we model one health examination record of a person at a time as one instance. In the case of T2, we model two consecutive health examination records as one instance. We repeatedly model this temporal information to T7 which is the maximum health examination records (7 years) of a person that we have. The number of instances for data with regard to each temporal feature case is shown in table 4. Table 3 also shows that extending to be using temporal features provides have higher F-measure than data without temporal features (Muti-All). The F-measure of Logistic Regression with T4 = 0.61, C4.5 (J48) with T6 = 0.67, Bagging with T5 = 0.63 and SVMs with T6 = 0.65. The average F-measures of C4.5 (J48), Bagging, Svms are 0.59, 0.57, 0.56, respectively.

Fig. 2 shows the ROC and AUC of data with and without temporal features using naïve bayes. Data with temporal features has higher AUC than non-temporal for T2 to T7 cases. AUC is the highest at T5 although the number of instances for training the classification model is only 5.34 % of the non-temporal data.

V. Conclusion

A group of features that is useful for diabetes classification is urinalysis, i.e., protein in urine, sugar in urine. C4.5 (J48), Logistic Regression, Bagging and SVMs provide no significant difference of classification accuracy and F-measure, but Naïve Bayes is not the case. Temporal feature extraction model is very useful for diabetes classification, i.e., higher accuracy and F-measure can be achieved when employing temporal features. It is possible to apply temporal feature extraction model to the other fields that contain features on the collection of history data such as medical data, weather forecasting, and business failure forecasting.

References

[1] National Center for Chronic Disease Prevention and Health Promotion. National Diabetes Fact Sheet. [Online]. Available: http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf, 2011.

[2] B. H. Cho, H. Yu, K. Kim, T. H. Kim, I. Y. Kim, and S. I. Kim, "Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods," *Journal Artificial Intelligence in Medicine.* 2008, vol. 42, pp. 37-53.

[3] H. N. A. Pham and E. Triantaphyllou, "Prediction of Diabetes by Employing a New Data Mining Approach Which Balance Fitting and Generalization," *Computer and Information Science.* 2008, vol. 131, pp. 11-26.

[4] K. Takahashi, H. Uchiyama, S. Yanagisawa, and I. Kamae, "The Logistic Regression and ROC Analysis of Group-based Screening for Predicting Diabetes Incidence in Four Years," *The Kobe journal of medical science.* 2006, vol. 52, no. 6, pp. 171-180.

[5] B. A. Tama, F. S. Rodiyatul, and Hermansyah, "An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital," in *Proceeding of The International Conference on Informatics, Cybernetic, and Computer Applications.* Bangalore. 2010, vol. 9, no. 2, pp. 287-294.

[6] R. Peter and T. Thomas, "Temporal Data Classification using Linear Classifiers," *Journal Information Systems.* 2011, vol. 36, no. 1, pp. 30-41.

[7] G. Parthiban, A. Rajesh, and S. K. Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naïve Bayes Method," *International Journal of Computer Applications.* 2011, vol. 24, pp. 7-11.

[8] World Health Organization. BMI Classification. [Online]. Available: http://apps.who.int/bmi/index.jsp?introPage=intro_3.html, 2011.

[9] World Health Organization. 2003 World Health Organization (WHO)/International Society of Hypertension (ISH) statement on management of hypertension. [Online]. Available: http://www.who.int/cardiovascular_diseases/guidelines/hypertension/en/, 2011.

[10] I. H. Witten and E. Frank, *Data mining: Practical Machine Learning Tools and Techniques*, 3rd Edition. San Francisco: Morgan Kaufmann, 2011.