Forecasting Potential Diabetes Complications

Yang Yang
Dept. of Computer Science
Tsinghua University
yangyang@keg.cs.tsinghua.edu.cn

Walter Luyten
Dept. of Biology
Katholieke Universiteit Leuven
Walter.Luyten@med.kuleuven.be

Lu Liu
EECS Dept.
Northwestern University
liulu26@gmail.com

Marie-Francine Moens
Dept. of Computer Science
Katholieke Universiteit Leuven
sien.moens@cs.kuleuven.be

Jie Tang
Dept. of Computer Science
Tsinghua University
jietang@keg.cs.tsinghua.edu.cn

Juanzi Li
Dept. of Computer Science
Tsinghua University
ljz@keg.cs.tsinghua.edu.cn

Abstract

Diabetes complications often afflict diabetes patients seriously: over 68% of diabetes-related mortality is caused by diabetes complications. In this paper, we study the problem of automatically diagnosing diabetes complications from patients' lab test results. The objective problem has two main challenges: 1) feature sparseness: a patient only takes 1.26% lab tests on average, and 65.5% types of lab tests are taken by less than 10 patients; 2) knowledge skewness: it lacks comprehensive detailed domain knowledge of association between diabetes complications and lab tests. To address these challenges, we propose a novel probabilistic model called Sparse Factor Graph Model (SparseFGM). SparseFGM projects sparse features onto a lower-dimensional latent space, which alleviates the problem of sparseness. SparseFGM is also able to capture the associations between complications and lab tests, which help handle the knowledge skewness. We evaluate the proposed model on a large collections of real medical records. SparseFGM outperforms (+20% by F1) baselines significantly and gives detailed associations between diabetes complications and lab tests.

Introduction

Diabetes mellitus, or simply diabetes, is a very common chronic disease, whose prevalence continues to increase, especially in the "Western" world. People with diabetes normally suffer from high blood glucose levels, which damages tissues over time and leads to life-threatening health complications, e.g., hypertension, coronary heart disease, hyperlipidemia, etc. Many of these complications seriously diminish the quality of life and even lead to premature death. Studies show that diabetes and its complications are major causes of early death in most countries (Roglic and Unwin 2010), and over 68% of diabetes-related mortality is caused by diabetes complications[1].

However, diabetes and its complications are often diagnosed late due to the long subclinical evolution, unclear pathogenesis, and poor medical facilities in some regions over the world. Although more than 471 billion USD were spent on healthcare for 371 million diabetes patients worldwide in 2012, still half of the people with diabetes remain undiagnosed, and 4.8 million people died in 2012 due to diabetes[2].

Challenges. In this paper, we study the problem of automatically diagnosing diabetes complications from lab test results. More specifically, given a series of lab test results of a potential diabetes patient, the goal of this study is to find particular complications that the patient may has. This can benefit a wide range of applications and studies, such as online diabetes complication diagnosis system, and studies of underlying patterns between diabetes complications and lab tests. The problem has several unique challenges:

Feature sparseness. There are 1945 different lab tests in our data set, while the clinical record of each patient only takes 24.43 different lab tests (1.26%) on average, which means only 1.26% of the features on average are used to represent each instance. At the same time, 65.5% types of lab tests are recoded in less than 10 clinical records among 181,933 records in total (0.0054%). The rare appearance of features will cause the ineffectiveness of the model learning and leads to poor performance.

Knowledge skewness. The performance of approaches to solve the objective problem highly relies on the domain knowledge of associations between diabetes complications and different types of lab tests. Previous work studied some of the associations. For example, (Zürbig et al. 2012) studied the association between tests on urinary proteome and diabetic nephropathy. (Voulgari, Papadogiannis, and Tentolouris 2010) studied the association between echocardiographic methods and diabetic cardiomyopathy. However, to the best of our knowledge, no study has provided a comprehensive detailed association. The skewness of the domain knowledge will cause an unbalanced performance.

Proposed Solution and Contributions. To address these challenges, we propose the Sparse Factor Graph Model (SparseFGM). For handling feature sparseness, SparseFGM projects the sparse features into a lower-dimensional latent space, which alleviates the sparseness issue. For handling knowledge skewness, SparseFGM models the associations between all diabetes complications and lab tests in our dataset by exponential-linear functions. More importantly, with the model, we can not only forecast a potential diabetes

[1] http://www.diabetes.org/

[2] http://www.idf.org/diabetesatlas/

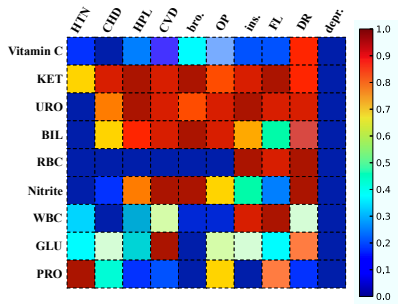

Figure 1: Association between 10 complications and 9 parameters of routine urine analysis. The parameters include: vitamin C, ketone (KET), urobilinogen (URO), bilirubin (BIL), red blood cell (RBC), nitrite, white blood cell (WBC), glucose (GLU), and protein (PRO). We refer to the experiment section for meanings of the complications' abbreviations.

complication, but also tell which types of lab test are most strongly associated with the diagnosed complication.

We evaluate the proposed model on a large collections of real medical records. Figure 1 shows the results obtained in our experiment to demonstrate the association between ten complications and nine parameters of routine urine analysis. The color in each square represents the association strength, discovered by our proposed model, between each urine test and complication. For example, protein (PRO) has strong association with hypertension (HTN) as chronic hypertension causes kidney damage, which in turn leads to the appearance of protein in the urine.

In all, our contributions of this paper are summarized in the following.

- We identify and formalize a new problem of diabetes complication diagnosing by a machine learning method. To the best of our knowledge, no previous work has extensively studied this problem.
- We propose a probabilistic model, SparseFGM, to solve the diabetes complication diagnosis problem, which integrates a feature dimensionality reduction process, association mining between complications and lab tests, and prediction into an unified model.
- We demonstrate the power of the proposed method using a large collection of real clinical records. Experimental results show that SparseFGM outperforms SVM, traditional factor graph models, and PCA based methods by $+20\%$ on average.

Organization. We first formally formulate the problem; we then introduce the proposed model and explain how to learn the model; next, we present the experimental results that validate the effectiveness of our model; we finally discuss related work and present the conclusions.

Problem Definition

We define several related concepts and formulate the diabetes complication forecasting problem.

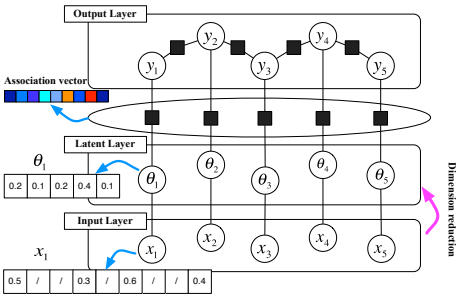

Figure 2: An illustration of the proposed model. Instances in the input layer are represented sparsely. The model projects a 9-dimensional feature space into a 5-dimensional latent space, which alleviates the sparseness issue. Vectors indicating associations between a particular complication and (groups of) lab tests are derivative from the feature factors, which bridge the latent layer and output layer of SparseFGM. Detailed dependencies between input layer and latent layer are omitted.

We first introduce the observed lab tests, which are performed on patients to evaluate the patient's health condition. Formally, we define a lab test record of a patient as follows:

Definition 1. **Lab Test Record.** We define a lab test record of a patient n as a set $\boldsymbol{u}_n = \{u_i\}$, where each u_i denotes a lab test performed on the patient n. We also use a tuple $u = (l, r)$ to represent each lab test, where l is the type of the lab test (e.g., WBC test in urine routine), and r is the result. Lab tests can have numerical or categorical types of results.

A patient can have several different lab test records as she may take lab tests more than once. For example, patients with serious diabetes may have to take tests once per week. Thus we further define a (lab test) *record sequence* for a particular patient to represent the order of the patient taking lab tests. Specifically, a record sequence of a patient is defined as a sequence $S = (s_i)$, where s_i is a lab test record. We say the test s_i is performed later than s_j if $i > j$.

We finally define the objective problem in this work. Generally, our goal is to diagnose diabetes complications from a set of lab test records of patients. Specifically, the problem can be defined as follows.

Definition 2. **Diabetes Complication Forecasting.** The inputs of diabetes complication forecasting include a diabetes complication C, a set of patients $V = \{v\}$, and a record sequence S_n for each patient n. The goal of diabetes forecasting is, for each patient n, and each lab test record $\boldsymbol{u}_n \in S_n$, determine if the patient n has the diabetes complication C, when the lab tests in $\boldsymbol{u}_n$ are performed on the patient n.

Model Framework

In this section, we explain the model we have developed for the diabetes complication forecasting problem. Table 1 sum-

Table 1: Notations in the proposed model.

SYMBOL	DESCRIPTION
K	The number of latent variables;
L	The number of different lab tests;
C	A particular diabetes complication the model aims to forecast;
x_{nl}	The l-th value in the n-th instance node;
y_n	A label node indicate whether the patient corresponding to the n-th instance has the diabetes complication C or not;
$\boldsymbol{\theta}_n$	The parameter of the multinomial distribution over latent variables specific to the n-th instance node;
μ_{kl}, δ_k	The parameters of the Gaussian distribution used to sample x_{nl}, which has a numerical value and, is specific to the latent variable k;
$\phi_{klx_{nl}}$	The parameter of the multinomial distribution over x_{nl}, which has a categorical value, specific to the latent variable k;
α, β	Parameters used to define the feature factor and the correlation factor respectively.

marizes the notations used in the proposed model.

For each patient's lab test record $\mathbf{u}_n$, we create an instance node $\boldsymbol{x}_n$. Assume that there are in total L available (different) lab tests, we define $\boldsymbol{x}_n$ as a L-dimensional vector. For each lab test $u_{ni} = (l, r) \in \mathbf{u}_n$, we set $x_{nl} = r$. We also set $x_{nl'} = /$ to denote that there is no lab test with type l' in $\mathbf{u}_n$. We associate a variable label y_n for $\mathbf{x}_n$'s corresponding patient, to denote whether the patient has the given complication C or not. Particularly, we use $y_n = 1$ to denote a positive result, and use $y_n = -1$ to denote a negative result. Based on this formulation, we can build a classification model to map the input instance to the target label, i.e., $f(\boldsymbol{x}_n) \rightarrow y_n$. With this formulation, more-or-less standard technologies can be employed for learning and inference, for example SVM (Chang and Lin 2011). However, SVM treats all instances independently and cannot capture the dependencies between instances (e.g., the dependencies between records in the record sequence). Factor graph model (Kschischang, Frey, and Loeliger 2001) can be leveraged to model the dependencies. However, the traditional factor graph model still cannot deal with the challenges in our problem: feature sparseness and knowledge skewness.

To this end, we propose a Sparse Factor Graph Model (SparseFGM). The basic idea here is that we first project instances from the original sparse space onto a lower dimensional latent space. The latent space is used to capture the correlation between different lab tests. In this way, the model alleviates the feature sparseness problem. Moreover, the graphical structure of the SparseFGM model is designed to model the correlation between different labels. Figure 2 shows a simple example of an SparseFGM. Instances in the input layer (represented as 9-dimensional vector) are sparse. The model projects all instances onto a 5-dimensional latent space, which alleviates the sparseness issue. Further the factor function (indicated as black rectangles) between labels is defined to model the dependencies between instances. Please notice that SparseFGM is not limited in diabetes complication forecasting problem. In general, it can be used into other similar scenarios.

According to the model, we can write joint distribution of a given set of instances $\mathbf{X}$ over Y as

$$P(y_n|\boldsymbol{\theta}_n, \boldsymbol{x}_n) = P(y_n|\boldsymbol{\theta}_n) \prod_l (\sum_{k=1}^{K} \theta_{nk} \cdot \Omega_{x_{nl}k}) \tag{1}$$

where l is an index of $\mathbf{x}_n$ which satisfies $x_{nl} \neq /$; K is the number of latent variables in SparseFGM; $\boldsymbol{\theta}_n$ is the parameter of a multinomial distribution which assigns latent variables to the n-th instance; $\Omega_{x_{nl}k}$ is the parameter of a distribution which generates values of x_{nl} when it is assigned with latent variable k. In the diabetes complication diagnosis problem, there are two kinds of lab test results: numerical and categorical. Similar with (Liu et al. 2013), we assume the numerical values are drawn from some Gaussian distributions, and categorical values are drawn from multinomial distributions. Thus $\Omega_{x_{nl}k}$ is defined as

$$\Omega_{x_{nl}k} = \begin{cases} N(x_{nl}|\mu_{kl}, \delta_k) & x_{nl} \text{ is numerical} \\ \phi_{klx_{nl}} & x_{nl} \text{ is categorical} \end{cases} \tag{2}$$

where μ_{kl} is the mean of the Gaussian distribution w.r.t. latent variable k and the category of patient symptom l; we assume all Gaussian distributions corresponding to the same latent variable k share the same standard deviation δ_k; $\phi_{klx_{nl}}$ is the probability that the l-th dimension of an instance variable assigned with the latent variable k has the value x_{nl}.

Regarding the feature factor $P(y_n|\boldsymbol{\theta}_n)$, we define it as

$$P(y_n|\boldsymbol{\theta}_n) = \frac{1}{Z_1} \exp\{\boldsymbol{\alpha} \cdot f(\boldsymbol{\theta}_n, y_n)\} \tag{3}$$

where $\boldsymbol{\alpha}$ is a parameter of SparseFGM, Z_1 is a normalization factor, and $f(\boldsymbol{x}_n, y_n) = y_n \cdot \boldsymbol{x}_n$.

To model the correlations between different labels, we define the correlation factor. Intuitively, a patient's health condition will not change quickly. Thus, we create a correlation factor between two labels y_n and $y_{n'}$ corresponding to two adjacent lab test records in the same patient's *record sequence*. Specifically, we define the correlation factor as

$$P(y_n, y_{n'}) = \frac{1}{Z_2} \exp\{\beta \cdot g(y_n, y_{n'})\} \tag{4}$$

where β is real number, Z_2 is a normalization factor, and $g(y_n, y_{n'})$ is defined as a vector of indicator functions.

By combining all the factors together, we can obtain the log-likelihood objective function for SparseFGM:

$$\begin{aligned} O(\lambda) &= \sum_n \log P(y_n|\boldsymbol{\theta}_n, \boldsymbol{x}_n) + \sum_c \log P(y_{c_1}, y_{c_2}) \\ &= \sum_n \boldsymbol{\alpha} f(\boldsymbol{\theta}_n, y_n) + \sum_c \beta g(y_{c_1}, y_{c_2}) \\ &+ \sum_n \sum_l \log \sum_k \theta_{nk} \boldsymbol{\Omega}_{k,l,x_{nl}} - \log Z \end{aligned} \tag{5}$$

where c is an indicator for 2-cliques among label nodes $\mathbf{Y}$, and $Z = Z_1 Z_2$.

Input: a feature matrix X, learning rate η
Output: estimated parameters λ

Initialize $\boldsymbol{\alpha}, \beta, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\phi}$ randomly;
Initialize $\boldsymbol{\delta} \leftarrow \mathbf{1}$;
repeat
  Calculate $P(k_{nl}|\boldsymbol{x}_n, \lambda_n)$ according to Eq. 8;
  Update $\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\phi}$ according to Eq. 9-12;
  Call LBP to calculate $\boldsymbol{E}[\sum_n f(\boldsymbol{\theta}_n, y_n)]$ and $\boldsymbol{E}[\sum_c g(y_{c_1}, y_{c_2})]$;
  Call LBP to calculate $\boldsymbol{E}_{P_\alpha(\boldsymbol{y}|\boldsymbol{\theta})}[\sum_n f(\boldsymbol{\theta}_n, y_n)]$ and $\boldsymbol{E}_{P_\beta(\boldsymbol{y})}[\sum_c g(y_{c_1}, y_{c_2})]$;
  Calculate $\frac{\partial O(\boldsymbol{\alpha},\beta)}{\partial \boldsymbol{\alpha}}$ and $\frac{\partial O(\boldsymbol{\alpha},\beta)}{\partial \beta}$ according to Eq. 14;
  $\boldsymbol{\alpha}_{\text{new}} = \boldsymbol{\alpha}_{\text{old}} + \eta \frac{\partial O(\boldsymbol{\alpha},\beta)}{\partial \boldsymbol{\alpha}}$;
  $\beta_{\text{new}} = \beta_{\text{old}} + \eta \frac{\partial O(\boldsymbol{\alpha},\beta)}{\partial \beta}$
until *Convergence* ;

Algorithm 1: Learning algorithm of SparseFGM.

Model Learning

We introduce the learning algorithm to estimate the optimal parameter configuration $\lambda = \{\boldsymbol{\alpha}, \beta, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\delta}\}$ for the proposed SparseFGM, i.e., finding parameters to maximize the log-likelihood, i.e.,

$$\begin{aligned} \lambda^\star &= \arg\max_\lambda O(\lambda) \\ s.t. \ &\sum_k \theta_{nk} = 1, \ \sum_x \phi_{klx} = 1 \end{aligned} \tag{6}$$

Unfortunately Eq. 6 does not have a closed-form solution. Thus, we propose an EM-like approximation learning algorithm to estimate the parameters.

We first introduce the general idea of the learning algorithm. By regarding labels inferred by the last updated model configuration, α and β as fixed, the remaining part of the model could be learned as a mixture generative model with parameters $\boldsymbol{\theta}$ and Ω. On the other hand, by fixing $\boldsymbol{\theta}$ and Ω, the remaining part of the model is similar to FGM and we are able to utilize a gradient based method to estimate $\boldsymbol{\alpha}$ and β. Next, we introduce how we update these parameters in detail, and give a framework of the learning algorithm.

Step 1: Update $\{\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\delta}\}$. When fix $Y, \{\boldsymbol{\alpha}, \beta\}$, SparseFGM could be regarded as a mixture generative model with the generative process as follows: first, each instance variable has a latent variable distribution parameterized by θ_n. Next, for each instance variable x_n, a latent variable k is sampled according to $k \sim Mult(\theta_n)$. After that, x_{nl} is sampled according to $x_{nl} \sim \Omega_{kl}$. Finally, each label variable y_n is sampled according to $y_n \sim P(y_n|\boldsymbol{\theta}_n)$.

With the generative process above, we can define the mixture generative model's log-likelihood, which could also been seen as the combination of related terms from Eq. 5:

$$\begin{aligned} O(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\delta}) = &\sum_n \sum_{l_1} \log \sum_k \theta_{nk} \frac{\exp\{\frac{-(x_{nl_1} - \mu_{kl})^2}{2\delta_k^2}\}}{\delta_k \sqrt{2\pi}} \\ &+ \sum_n \sum_{l_2} \log \sum_k \theta_{nk} \phi_{kl_2 x_{nl_2}} \end{aligned} \tag{7}$$

By Jensen's inequality, we obtain the lower bound of Eq. 7. We derivate the lower bound with respect to each parameter and set them to zero, we have

$$P(k_{nl}|\boldsymbol{x}_n, \lambda_n) = \frac{\theta_{nk}\Omega_{klr}}{\sum_k \theta_{nk}\Omega_{klr}} \tag{8}$$

$$\theta_{nk} = \frac{\sum_l P(k_{nl}|\boldsymbol{x}_n, \lambda_n)}{\sum_l \sum_{k_{nl}} P(k_{nl}|\boldsymbol{x}_n, \lambda_n)} + \alpha_k y_n \tag{9}$$

$$\mu_{kl} = \frac{\sum_n P(k_{nl}|\boldsymbol{x}_n, \lambda_n) x_{nl}}{\sum_n P(k_{nl}|\boldsymbol{x}_n, \lambda_n)} \tag{10}$$

$$\delta_k^2 = \frac{\sum_n \sum_l (x_{nl} - \mu_{kl})^2}{N \times L_1} \tag{11}$$

$$\phi_{klr} = \frac{\sum_n P(k_{nl}|\boldsymbol{x}_n, \lambda_n)}{\sum_n \sum_r P(k_{nl}|\boldsymbol{x}_n, \lambda_n)} \tag{12}$$

where L_1 is the number of numerical features, and N is the total number of lab test records.

Step 2: Update $\{\boldsymbol{\alpha}, \beta\}$. When fix $\boldsymbol{\theta}$ and Ω, the remaining part of SparseFGM is similar with FGM, which has the log-likelihood as

$$O(\boldsymbol{\alpha}, \beta) = \sum_n \boldsymbol{\alpha} f(\boldsymbol{\theta}_n, y_n) + \sum_c \beta g(y_{c_1}, y_{c_2}) - \log Z \tag{13}$$

We then adopt a gradient descent method to optimize Eq. 13. The gradient for each parameter is calculated as

$$\begin{aligned} \frac{\partial O(\boldsymbol{\alpha}, \beta)}{\partial \boldsymbol{\alpha}} &= \boldsymbol{E}[\sum_n f(\boldsymbol{\theta}_n, y_n)] - \boldsymbol{E}_{P_\alpha(\boldsymbol{y}|\boldsymbol{\theta})}[\sum_n f(\boldsymbol{\theta}_n, y_n)] \\ \frac{\partial O(\boldsymbol{\alpha}, \beta)}{\partial \beta} &= \boldsymbol{E}[\sum_c g(y_{c_1}, y_{c_2})] - \boldsymbol{E}_{P_\beta(\boldsymbol{y})}[\sum_c g(y_{c_1}, y_{c_2})] \end{aligned} \tag{14}$$

We adopt Loopy Belief Propagation (LBP) (Murphy, Weiss, and Jordan 1999) to compute the marginal probability of $\boldsymbol{y}$ and thus compute the two expectations. We then update α and β with a learning rate η with the gradient. Algorithm 1 summarizes the learning algorithm for SparseFGM.

Experiments

In this section, we present experimental results to demonstrate the effectiveness of the proposed approach. All codes used in the paper will be publicly available.

Experimental Setup

We use a collection of real medical records from a famous geriatric hospital. The dataset spans over one year, containing 181,933 medical records corresponding to 35,525 unique patients and $1,945$ kinds of lab tests in total. On average each clinical record contains 24.43 different lab tests (1.25% of all lab tests), which indicates the feature sparsity problem is serious for this problem.

We view each medical record as an instance, and aim to infer whether the corresponding patient has a particular diabetes complication or not from the lab test results. Nine complications are taken into account in our experiments:

Table 2: Performance of diabetes complication forecasting.

Complication	Method	Precision	Recall	F1
HTN	SVM	0.3804	0.4789	0.4241
	FGM	0.5666	0.4959	0.5075
	FGM+PCA	0.5741	0.3284	0.4178
	SparseFGM	0.4714	0.6319	**0.5400**
CHD	SVM	0.2132	0.0636	0.0980
	FGM	0.6264	0.1369	0.2247
	FGM+PCA	0.2425	0.8367	0.3761
	SparseFGM	0.2522	0.7972	**0.3832**
HPL	SVM	0.2208	0.0460	0.0761
	FGM	0.6557	0.0591	0.1084
	FGM+PCA	0.2047	0.8035	0.3262
	SparseFGM	0.2796	0.8396	**0.4195**

hypertension (HTN), coronary heart disease (CHD), hyperlipidemia (HPL), cerebrovascular disease (CVD), bronchitis (bro.), osteoporosis (OP), insomnia (ins.), fatty liver (FL), diabetic retinopathy (DR), and depression (depr.). In the experiments, we randomly picked 60% of the medical records as training data and the rest for testing.

Evaluation Aspects. We evaluate our method on the following three aspects:

- **Forecasting Performance.** We evaluate the proposed model in terms of Precision, Recall, F1-Measure, and compare with baseline methods to validate its effectiveness.
- **Association pattern illustration.** We use the discovered association patterns between diabetes complications and lab tests as anecdotal evidence to further demonstrate the effectiveness of our method.

We compare the following methods for forecasting diabetes complications.

SVM. Lab test results are treated as features and LIBSVM (Chang and Lin 2011) is employed as the classification model for complication forecasting.

FGM. A traditional factor graph is used as the classification model. We employ a gradient descent algorithm to learn the parameters in FGM (Tang, Zhuang, and Tang 2011; Hopcroft, Lou, and Tang 2011), and set the learning rate parameter as 0.1.

FGM+PCA. To solve the feature sparsity problem in FGM, PCA (Jolliffe 1986) is employed to convert the features into a set of principle components, which is used as the input of FGM then.

SparseFGM. SparseFGM is our proposed model. In all experiments, we empirically set the number of latent variables in SparseFGM as 100, and set $\eta = 0.1$.

All algorithms were implemented in C++, and all experiments were performed on a Mac running Mac OS X with Intel Core i7 2.66 GHz and 4 GB memory.

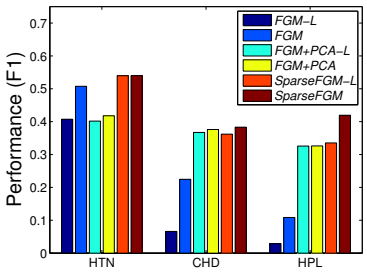

Figure 3: Correlation factor analysis. FGM-L, FGM+PCA-L, and SparseFGM-L denote the methods not considering the correlation factor.

Forecasting Performance

Table 2 shows the performance of different methods on diabetes complication forecasting task. Due to the space limitation, we only present the results of three examples for most common diabetes complications: hypertension (HTN), coronary heart disease (CHD), and hyperlipidemia (HPL). From the table, we can see that SparseFGM achieves the highest F1-score in all three tasks compared with other methods on average. Generally, SparseFGM offers the possibility to model the dependencies and constraints together with modeling latent class information. The results show that recall is increased without hurting precision substantially, or in some cases while even improving precision.

SVM and traditional factor graph based methods (FGM) suffer from the feature sparsity particularly in terms of recall, averagely 59.9% decrease compared with SparseFGM in all tasks. By a careful investigation, we find that most parameters in FGM tend to be zero due to the serious feature sparsity problem, which cause the false diagnose results of the models.

With the effect of dimension reduction, FGM+PCA based methods improve the performance a lot. For example, FGM+PCA increase 40.3% in terms of recall compared with FGM in CHD the forecasting task. However, FGM+PCA separates sparse coding and classification process into two steps, while SparseFGM integrate them into a uniform model. Thus SparseFGM estimates the parameters better and outperforms FGM+PCA by 13.5% in terms of F1 on average.

Factor Contribution Analysis. Figure 3 demonstrate how the correlation factor helps in this problem. Generally, models considering the correlations between labels increase 16.29% in terms of F1 compared with their counterpart models without considering the correlation factor. Intuitively, correlation factors improve the performance by bringing the prior knowledge that "a patient's health condition will be similar in adjacent time stamps".

Association pattern illustration

Based on SparseFGM, We can discover the underlying patterns between diabetes complications and lab tests at three granularities: micro-level, meso-level, and macro-level.

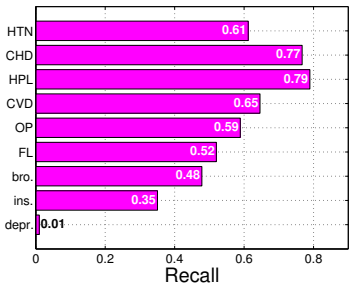

Figure 4: Association analysis at macro-level. X-axis denotes the performance of the proposed model on diagnosing each diabetes complication in terms of recall.

Micro-level. At micro-level, we calculate the association score $AS(c, e)$ of a complication c with a lab test type e as

$$AS(c, e) = \sum_k \alpha_k^c \theta_{ek} \tag{15}$$

where k is the latent variables in SparseFGM, and $\boldsymbol{\alpha}^c$ is the feature factor parameter of SparseFGM for diagnosing complication c.

The example which illustrates the association between the lab tests of urine routine and 10 diabetes complications is shown in Figure 1. We see that urinary glucose (GLU) is strongly associated with diabetic retinopathy (DR), because DR typically occurs in diabetic patients whose blood glucose levels are not well controlled. To our surprise, we find that insomnia (ins.) is associated with most urine routine measures, such as white blood cells (WBC). This association has an interesting explanation: WBC in the urine typically is found in urinary tract infections which cause frequent voiding, so people who have to go to the toilet frequently can hardly get a good sleep at night.

Meso-level. At meso-level, we study how different groups of lab tests associated with diabetes complications. The grouping information could be extracted from the proposed model naturally as SparseFGM clusters features into K groups implicitly. The details of the experimental results can be found in our supplemental materials.

Macro-level. At macro-level, we study the different strength of diabetes complications forecasting depending on lab test results. We say a complication is more diagnosable from lab test results if the proposed model is able to target more positive instances correctly among all positive instances. Thus the strength can be estimated by the measurement recall. Figure 4 shows the results, from which we see that hyperlipidemia (HPL) can be diagnosed based on lab tests precisely, while depression (depr.) is usually recognized from psychological investigation instead of physiological lab tests such as blood test. In this experiment, to avoid the label balance issue affecting on the analysis results, we set the ratio of positive to negative instances for each complication as $1 : 5$. Thus the recall results are different from the comparison experimental results, in which all the testing samples in the dataset are used.

We also conduct an experiment to demonstrate how the parameters of the proposed model influence the performance. See details in our supplemental materials.

Related Work

Recent years, utilizing health care data to study diabetes, which is a common chronic disease, attracts the interests of researchers. Liu et al. (Liu et al. 2013) propose a Symptom-Diagnosis-Treatment model to mine the diabetes complication patterns and symptoms from electronic medical records, which is the same data set employed in this paper. Besides, Neuvirth et al. (Neuvirth et al. 2011) studied the personalized care management of diabetes patients at risk. For the medical work, Groot et al. (De Groot et al. 2001) examine the strength and consistency of the relationship between depression and diabetes complications in studies of type 1 and type 2 adult patients with diabetes. (Katon et al. 2003) and (Kanaya, Grady, and Barrett-Connor 2002) investigate two major diabetes complications: depression and coronary heart disease respectively, which are used as sample labels in our experiments. However, to the best of our knowledge, none of the previous work investigated how to predict diabetes complication by machine learning methodologies.

Sparse modeling is a component in many machine learning tasks. Yoshida et al. (Yoshida and West 2010) also propose a class of sparse latent factor models and relevant learning algorithms. Their model focuses on Gaussian distribution while our proposed model brings in a multinomial distribution to better model categorical features. Krishnapuram et al. (Krishnapuram et al. 2005) and Zhong et al. (Zhong and Wang 2008) learn models for a sparse Multinomial Logistic Regression and a sparse Conditional Random Field respectively with Laplacian prior, which is different from our method fundamentally. Tan et al. (Tan et al. 2012) study the problem of large-scale sparse coding and subset selection problem by proposing a convex matching pursuit scheme. Zhang et al. (Zhang et al. 2011) try to solve the sparse coding problem by proposing a feature selection framework. Mairal et al. (Mairal et al. 2010) propose a stochastic online learning algorithm for learning dictionaries adapted to sparse coding tasks, and proven its convergence. Lee et al. (Lee et al. 2006) propose efficient sparse coding algorithms to solve L1-regularized least squares problem and an L2-constrained least squares problem.

Conclusion

In this paper, we study the problem of diabetes complication forecasting. We propose a novel probabilistic model, Sparse Factor Graph Model (SparseFGM), which integrates sparse modeling and complication diagnosing into an uniform model. By using this model, we cannot only forecast diabetes complications but also discover the underlying association between complications and lab test types. We validate the model on a large collection of real medical records. Experimental results demonstrate that the proposed model outperforms baseline methods significantly. We also demonstrate how to discover the association between complications and lab test types at three granularities.

Acknowledgements. The work is supported by the co-funding of Tsinghua University and KU Leuven, and Natural Science Foundation of China (No. 61103065).

References

Chang, C.-C., and Lin, C.-J. 2011. Libsvm: a library for support vector machines. *TIST* 2(3):27.

De Groot, M.; Anderson, R.; Freedland, K. E.; Clouse, R. E.; and Lustman, P. J. 2001. Association of depression and diabetes complications: a meta-analysis. *Psychosom Med* 63(4):619–630.

Hopcroft, J.; Lou, T.; and Tang, J. 2011. Who will follow you back? reciprocal relationship prediction. In *CIKM'11*, 1137–1146. ACM.

Jolliffe, I. 1986. Principal component analysis. *Springer Series in Statistics, Berlin: Springer, 1986* 1.

Kanaya, A. M.; Grady, D.; and Barrett-Connor, E. 2002. Explaining the sex difference in coronary heart disease mortality among patients with type 2 diabetes mellitus: a meta-analysis. *Archives of Internal Medicine* 162(15):1737.

Katon, W.; Von Korff, M.; Lin, E.; Simon, G.; Ludman, E.; Bush, T.; Walker, E.; Ciechanowski, P.; and Rutter, C. 2003. Improving primary care treatment of depression among patients with diabetes mellitus: the design of the pathways study. *Gen Hosp Psychiatry* 25(3):158–168.

Krishnapuram, B.; Carin, L.; Figueiredo, M. A.; and Hartemink, A. J. 2005. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *TPAMI* 27(6):957–968.

Kschischang, F. R.; Frey, B. J.; and Loeliger, H. A. 2001. Factor graphs and the sum-product algorithm. *IEEE TOIT* 47:498–519.

Lee, H.; Battle, A.; Raina, R.; and Ng, A. 2006. Efficient sparse coding algorithms. In *NIPS'06*, 801–808.

Liu, L.; Tang, J.; Cheng, Y.; Agrawal, A.; Liao, W.-k.; and Choudhary, A. 2013. Mining diabetes complication and treatment patterns for clinical decision support. In *CIKM'13*, 279–288. ACM.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2010. Online learning for matrix factorization and sparse coding. *JMLR* 11:19–60.

Murphy, K. P.; Weiss, Y.; and Jordan, M. I. 1999. Loopy belief propagation for approximate inference: An empirical study. In *UAI'99*, 467–475.

Neuvirth, H.; Ozery-Flato, M.; Hu, J.; Laserson, J.; Kohn, M. S.; Ebadollahi, S.; and Rosen-Zvi, M. 2011. Toward personalized care management of patients at risk: the diabetes case study. In *SIGKDD'11*, 395–403. ACM.

Roglic, G., and Unwin, N. 2010. Mortality attributable to diabetes: estimates for the year 2010. *Diabetes Res Clin PR* 87(1):15–19.

Tan, M.; Tsang, I. W.; Wang, L.; and Zhang, X. 2012. Convex matching pursuit for large-scale sparse coding and subset selection. In *AAAI'12*.

Tang, W.; Zhuang, H.; and Tang, J. 2011. Learning to infer social ties in large networks. In *ECML/PKDD'11*, 381–397. Springer-Verlag.

Voulgari, C.; Papadogiannis, D.; and Tentolouris, N. 2010. Diabetic cardiomyopathy: from the pathophysiology of the cardiac myocytes to current diagnosis and management strategies. *Vasc Health Risk Manag* 6:883.

Yoshida, R., and West, M. 2010. Bayesian learning in sparse graphical factor models via variational mean-field annealing. *JMLR* 99:1771–1798.

Zhang, X.; Yu, Y.; White, M.; Huang, R.; and Schuurmans, D. 2011. Convex sparse coding, subspace learning, and semi-supervised extensions. In *AAAI'11*.

Zhong, P., and Wang, R. 2008. Learning sparse crfs for feature selection and classification of hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* 46(12):4186–4197.

Zürbig, P.; Jerums, G.; Hovind, P.; MacIsaac, R. J.; Mischak, H.; Nielsen, S. E.; Panagiotopoulos, S.; Persson, F.; and Rossing, P. 2012. Urinary proteomics for early diagnosis in diabetic nephropathy. *Diabetes* 61(12):3304–3313.