Artificial Intelligence in Medicine (2008) **42**, 37—53

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods

Baek Hwan Cho[a], Hwanjo Yu[b], Kwang-Won Kim[c], Tae Hyun Kim[c], In Young Kim[a,*], Sun I. Kim[a]

[a] *Department of Biomedical Engineering, Hanyang University, Seoul, Republic of Korea*
[b] *Department of Computer Science, University of Iowa, Iowa City, IA, USA*
[c] *Division of Endocrinology and Metabolism, Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea*

Received 6 April 2007; received in revised form 21 September 2007; accepted 21 September 2007

KEYWORDS
Decision support systems;
Diabetic nephropathy;
Support vector machines;
Visualization;
Risk factor analysis;
Feature selection

Summary

Objective: Diabetic nephropathy is damage to the kidney caused by diabetes mellitus. It is a common complication and a leading cause of death in people with diabetes. However, the decline in kidney function varies considerably between patients and the determinants of diabetic nephropathy have not been clearly identified. Therefore, it is very difficult to predict the onset of diabetic nephropathy accurately with simple statistical approaches such as *t*-test or χ^2-test. To accurately predict the onset of diabetic nephropathy, we applied various machine learning techniques to irregular and unbalanced diabetes dataset, such as support vector machine (SVM) classification and feature selection methods. Visualization of the risk factors was another important objective to give physicians intuitive information on each patient's clinical pattern.
Methods and materials: We collected medical data from 292 patients with diabetes and performed preprocessing to extract 184 features from the irregular data. To predict the onset of diabetic nephropathy, we compared several classification methods such as logistic regression, SVM, and SVM with a cost sensitive learning method. We also applied several feature selection methods to remove redundant features and improve the classification performance. For risk factor analysis with SVM classifiers, we have developed a new visualization system which uses a nomogram approach.
Results: Linear SVM classifiers combined with wrapper or embedded feature selection methods showed the best results. Among the 184 features, the classifiers selected the

* Corresponding author at: Sungdong P.O. Box 55, Seoul 133-605, Republic of Korea. Tel.: +82 2 2291 1713; fax: +82 2 2296 5943.
E-mail address: iykim@hanyang.ac.kr (I.Y. Kim).

0933-3657/$ — see front matter
doi:10.1016/j.artmed.2007.09.005

same 39 features and gave 0.969 of the area under the curve by receiver operating characteristics analysis. The visualization tool was able to present the effect of each feature on the decision via graphical output.

Conclusions: Our proposed method can predict the onset of diabetic nephropathy about 2—3 months before the actual diagnosis with high prediction performance from an irregular and unbalanced dataset, which statistical methods such as *t*-test and logistic regression could not achieve. Additionally, the visualization system provides physicians with intuitive information for risk factor analysis. Therefore, physicians can benefit from the automatic early warning of each patient and visualize risk factors, which facilitate planning of effective and proper treatment strategies.

1. Introduction

Diabetes mellitus is a metabolic disorder characterized by chronic hyperglycemia (high blood sugar level) resulting from defects in insulin secretion, insulin action, or both [1]. A medical insurance cohort study that included 1.2 million subjects indicated that diabetes mellitus, in Korea, is the first leading cause of burden of disease, and 8.4% of the population suffers from this disease [2]. Diabetes can cause devastating complications including cardiovascular diseases, kidney failure, leg and foot amputations, and blindness, which often result in disability and death. Diabetic nephropathy is damage to the kidney because of diabetes; it is a common diabetic complication and a leading cause of death in people with diabetes [3].

Many researchers have been trying to determine risk factors of mortality in patients with diabetes via statistical approaches. There are also many studies on predictors of diabetic nephropathy. The long-term occurrence over the years of high blood glucose level and high blood pressure is highly indicative of the development of renal disease, other microvascular lesions, and macrovascular disease [4,5]. Clinical trials have demonstrated consistently that suppression of the glycosylated hemoglobin (HbA_{1C}) level is associated with decreased risk for clinical and structural manifestations of diabetic nephropathy in patients with types 1 and 2 diabetes [6,7]. Torffvit and Agardh showed that poor metabolic control and high blood pressure is associated with development of diabetic nephropathy in patients with type 2 diabetes [8]. A logistic regression analysis was adopted to generate prediction rules for identifying patients with diabetes at high risk of complications and for analyzing risk factors [9]. However, most of those prognostic studies compared mean values of independent predictors between patients with diabetic nephropathy and control groups, using simple statistical methods such as Student's *t*-test and the non-parametric Mann—Whitney *U*-test. The decline in kidney function varies considerably between patients, and determinants of the diabetic nephropathy have not been identified clearly. Therefore, it is difficult to predict diabetic nephropathy accurately using simple statistical approaches.

A number of studies have taken advantage of data mining techniques in the diabetes domain. A 1998 review provided evidence of the use of decision support systems to guide physicians through the clinical consultation [10]. The paper used time series methods and a causal probabilistic network, and proposed telemedicine and telecare for patients with diabetes. Others introduced various types of artificial neural networks (ANNs) or decision trees to predict the onset of diabetes and to identify risk factors among the data [11—14]. However, most of those papers attempted to predict the onset of diabetes itself, even though its complications are much more important for the quality of life and mortality.

Although many have tried to approach health problems through data mining techniques, there are many constraints and difficulties in this [15]. The main difficulty arises in data collection. Because hospitals have not used electronic medical record systems for long, it is very hard to collect a dataset large enough for such research. Moreover, in the university hospital setting, physicians and medical practitioners have occasionally transferred to other medical institutions, which might cause different patient care protocols, including physical examinations and interviews formats. Furthermore, because of over confidence in their health condition or for other personal reasons, some outpatients visit the hospital irregularly. These circumstances may lead to very irregular, incomplete, or missing data in the clinical setting, and thus make it difficult to extract meaningful information from the data.

Other difficult problems in medical data mining lie in the artificial intelligence algorithm itself. To diagnose whether patients have a disease, or to predict if they would have a disease in the future, several studies have attempted to apply learning algorithms such as decision trees, ANNs, and support vector machines (SVMs). Although they give high generalization performances, the users can barely understand the results by means of input variables, especially with ANNs and SVMs; that is why they are termed "black box" algorithms. Consequently, despite its relatively poor generalization performance, physicians and medical practitioners still make use of the logistic regression (LR) as a gold standard because it gives more information about the results: it provides not only the percentage of variance in the output variable (i.e., probabilistic output), but also the odds ratio of each input variable.

In information retrieval, there are various feature (i.e., variable) selection methods to rank the importance of input features and thereby enhance generalization performance by eliminating the disturbing features. Most of those methods are mainly concerned with the ranking of input features, rather than about the insights of each feature. However, in a practical clinical situation, physicians may wish to understand the actual effect of each feature on the results: an interpretation about how the prediction result would change if a feature's value were to change. Jakulin et al. introduced a nomogram approach for visualizing SVMs that can graphically expose its internal structure and visualize the effect of each feature by means of the log odds ratio, as with LR [16].

In this study, we aimed to predict the onset of diabetic nephropathy by learning SVM predictive model from an irregular and unbalanced diabetic dataset. We also attempted to identify some risk factors from clinical parameters using feature selection and nomogram visualization.

2. Classification methods

2.1. Logistic regression and the ridge estimator

LR is a popular method to generate a predictive model for dichotomous target variable. The relationship between the input variables and the response is not a linear function, but the logistic regression function:

$$P(\mathbf{x}) = \frac{1}{1 + e^{-(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_d x_d)}} \qquad (1)$$

where P is the probability that an event happens, α is the coefficient of the equation and β_i is the coefficient of the input variable. The log-likelihood function of the data is

$$L(\beta) = \sum_i (y_i \log P(\mathbf{x}_i) + (1 - y_i)\log(1 - P(\mathbf{x}_i))) \qquad (2)$$

The conventional LR optimizer finds the LR coefficients by maximizing $L(\beta)$ using the maximum likelihood estimate method. However, unstable parameter estimates may arise when the number of input variables is large or the input variables are highly correlated. Cessie and Houwelingen introduced ridge estimators to resolve this case [17]. They included a restriction term in the log-likelihood function thus

$$L^{\lambda}(\beta) = L(\beta) - \lambda||\beta||^2 \qquad (3)$$

where $L(\beta)$ is the unrestricted log-likelihood function and $||\beta||$ is the norm of the parameter vector β. The ridge parameter λ shrinks the norm of β, and the restriction term stabilizes the system to provide estimates with smaller variance.

2.2. Cost sensitive learning in SVMs

SVMs, an emerging classification technique, have been intensively benchmarked against a variety of techniques; it is one of the best-known classification techniques with computational advantages and good generalization performance. The main idea of SVMs is to maximize the margin, which is defined as the distance from the separating hyperplane to the closest training samples (support vectors) [18].

However, for an unbalanced dataset that has far more positives than negatives or vice versa, the general classifiers may produce poor generalization performance as the hyperplane may be moved far away to the minority training samples. For this reason, Veropoulos et al. used a cost-sensitive learning approach for SVM; their key point is to give different cost (penalty) to the errors of each class [19]. Accordingly, the optimization strategy of SVM is as follows:

$$\text{minimize} \quad \frac{1}{2}||\mathbf{w}||^2 + C^{+}\sum_{\{i|y_i=+1\}}^{n} \xi_i + C^{-}\sum_{\{i|y_i=-1\}}^{n} \xi_i \qquad (4)$$

$$\begin{aligned} \text{subject to} \quad & y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, & i = 1,\ldots,n \\ & \xi_i \geq 0, & i = 1,\ldots,n \end{aligned} \qquad (5)$$

where $\mathbf{w}$ is a weight vector of the separating hyperplane, ξ_i is a slack variable that allow the margin constraints to be violated and C is a user parameter to be tuned. The first term of the objective function is about the margin maximization, and the second and third part is for controlling the penalties for positive and negative training samples. Geometrically, when positive data are in the minority, it gives more weight to the positive support vectors (C^{+} is larger than C^{-}), and pushes the hyperplane towards where the negative (majority) data exist.

Table 1 Algorithm of the ReliefF method

```
Inputs: All the instances and their class labels
Set all weights $\mathbf{W}[f] = 0$
Set arbitrary iteration number $a$
for $i = 1$ to $a$ do
    Randomly select an instance $I_i$
    Find the $k$ nearest hits $H_j$
    Find the $k$ nearest misses $M_j(C)$ for each class $C \neq \text{class}(I_i)$
    for $f = 1$ to the number of features do
        $\mathbf{W}[f] = \mathbf{W}[f] - \sum_{j=1}^{k} \text{diff}(f, I_i, H_j)/(m \cdot k) + \sum_{C \neq \text{class}(I_i)} \left[ \frac{P(C)}{1 - P(\text{class}(I_i))} \sum_{j=1}^{k} \text{diff}(f, I_i, M_j(C)) \right] /(m \cdot k)$
    end for
end for
Outputs: Weight vector $\mathbf{W}[f]$
```

3. Feature selection methods

Feature selection is a machine learning process that selects a feature subset from the whole feature set and removes redundant features that do not contribute to the performance. Feature selection methods have been introduced to avoid the "curse of dimensionality", which means the required number of calculations becomes huge as the number of dimensions increases, while retaining or even enhancing the performance.

There are three main approaches in feature selection: filter, wrapper, and embedded methods [20]. Filter methods select high ranked features based on a statistical score as a preprocessing step; ReliefF is a popular filter algorithm in microarray classification problems because of its simplicity. Wrapper performs selection taking into account the classifier as a black box and ranking the subset of features by their predictive power. Because a full search requires 2^n different evaluations, forward selection or backward elimination methods are used. Sensitivity analysis could be adopted to calculate the importance of each feature with any classifier. Embedded methods, in contrast to wrapper approaches, select features considering the classifier design at the same time.

3.1. ReliefF algorithm

A key idea in ReliefF is to evaluate the contribution of each feature to inter-class difference and intra-class similarity [21]. With a randomly selected data, the algorithm looks for the k nearest hits (those with the same class label) and misses (those with a different class label). After that, it updates the quality of the contribution of features with respect to the difference between the feature values of the selected data and nearest ones. The pseudocode is shown in Table 1. Function diff(f, I_i, I_j) calculates the difference between the feature values of two instances. Thus, the weight vector $\mathbf{W}[f]$ increases when the feature value of the selected instance is different from that of the nearest miss $M_j(C)$. On the other hand, it decreases when there is a difference between the feature values of the selected instance and the nearest hit $H_{j.}$ Finally, according to the weight values of $\mathbf{W}[f]$, one can identify the ranking of each feature and perform feature selection by eliminating features with smallest weight values.

3.2. Sensitivity analysis with SVM

Sensitivity analysis is another method that has been widely used to rank input features in terms of their contribution to the deviation of the outputs [22]. It involves varying every input feature over a reasonable range with the others fixed, and observing the relative changes in the outputs. As a result, features that produce larger deviation in the output are considered more important and one can select features by the same method that is adopted in the ReliefF method above. Table 2 shows the pseudocode of the sensitivity analysis method. The algorithm calculates the difference between maximum and minimum output of the predictive model when a feature value varies from its possible minimum to maximum with the other features fixed to their mean values.

Table 2 Algorithm of the sensitivity analysis method

```
Line     Code
Inputs   A predictive model $F(\mathbf{x})$
         Set all weights $\mathbf{W}[f] = 0$
         Set $\boldsymbol{O}[a] = 0$
         for $f = 1$ to the number of features do
             Initialize an instance $\boldsymbol{x} = [x_1 = \mathrm{mean}(x_1), x_2 = \mathrm{mean}(x_2), \ldots, x_n = \mathrm{mean}(x_n)]$
             for $j = \min(x_f)$ to $\max(x_f)$ do
                 Set $x_f = j$
                 Set $\boldsymbol{O}[j] = F(\mathbf{x})$
             end for
             $\mathbf{W}[f] = \max(\boldsymbol{O}) - \min(\boldsymbol{O})$
         end for
Outputs  Weight vector $\mathbf{W}[f]$
```

3.3. Recursive feature elimination with SVM (SVM—RFE)

SVM—RFE is an example of the embedded method and a similar approach that removes less important features recursively, except for using the weight magnitude as a ranking criterion [23]. The outline of the algorithm is presented in Table 3. In the iterative training process in SVM—RFE, one can find the best subset of features that provides the highest performance. However, this algorithm is limited theoretically to the linear kernel in SVM, because it is difficult to calculate the weight vector for a non-linear kernel because of the kernel characteristics of the implicit mapping. The authors of the paper [23] mentioned the non-linear kernel version of SVM—RFE that is computationally more expensive.

4. Risk factor analysis with nomogram visualization of SVM

4.1. General concept of the nomogram in predictive models

The nomogram gives the insight of a model by visualizing the effect of each feature on the prediction. Fig. 1 shows an imaginary example of the nomogram visualizing a predictive model. Suppose that one is trying to predict whether a patient will have diabetic nephropathy by considering two features. One is whether the patient smokes or not; the other is systolic blood pressure level. In the figure, log odds ratio (Log OR) scores of the individual features are summed up using the topmost axis of the nomogram and used to estimate the probability of having diabetic nephropathy (the bottommost

Table 3 Algorithm of the SVM—RFE method

```
Line     Code
Inputs   Training instances $\mathbf{x}_o$ and their class labels $\mathbf{y}$
         Initialize subset of surviving features $\boldsymbol{s} = [1,2,\ldots,n]$
         Initialize feature ranking list $\boldsymbol{r} = [\,]$
         while ( $\boldsymbol{s} \neq [\,]$ )
             Restrict training instances to the subset of surviving features $\boldsymbol{X} = \boldsymbol{X}_0(:,\boldsymbol{s})$
             Train the SVM with the restricted instances and their class labels
             Compute the weight vector of the SVM model $\mathbf{w} = \sum_i^{sv} y_i \alpha_i \mathbf{x}_i$
             Compute the ranking criteria $C_k = (\mathbf{w}_k)^2$, for all $k$
             Update the feature ranking list according to the criteria
             Eliminate features with the lowest ranking
         end while
Outputs  Feature ranking list $\boldsymbol{r}$
```

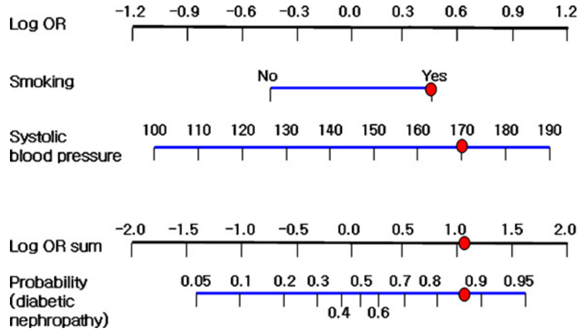


Figure 1 An imaginary nomogram example of a support vector machine (SVM) model that predicts the probability of having diabetic nephropathy within 1 year. The nomogram gives the insight of a model by presenting the Log OR score for each feature, which denotes the effect of the individual feature on the prediction (the higher the Log OR score, the higher the risk).

axis of the nomogram). In this example, the Log OR score of smoking is 0.46 and that of systolic blood pressure level of 170 is 0.63. Thus, the sum of the Log OR scores becomes 0.46 + 0.63 = 1.09, which corresponds to the probability of 0.86 that the patient will have diabetic nephropathy within a year.

Using the Log OR line, one can easily see how much each feature influences the target probability. When the Log OR score of a feature is high, it gives more positive effect on the probability (the higher the Log OR score, the higher the risk that the patient will have diabetic nephropathy). Moreover, longer features on the nomogram will have a wider range of Log OR score and thus have stronger effects on the target prediction probability. For example, the systolic blood pressure line is longer than the smoking line, which implies that the probability of diabetic nephropathy is more strongly associated with systolic blood pressure.

4.2. How to draw a nomogram with SVM

Jakulin et al. [16] introduced a nomogram method for SVM predictive models, in which the distance from a data sample ($\mathbf{x}$, y) to the separating hyperplane of SVM is considered an independent variable and denoted as $\delta(\mathbf{x})$. Given a kernel function $K(\mathbf{x}, \mathbf{z})$, the distance can be replaced by the decision function in SVM as follows:

$$\delta(\mathbf{x}) \cong b + \sum_{j=1}^{N} y_j \alpha_j K(\mathbf{x}, \mathbf{z}_j) \tag{6}$$

where b is the bias, α expresses the coefficients of support vectors $\mathbf{z}$ in SVM and N is the number of support vectors. When the kernel is linearly decomposable with respect to each feature, the distance becomes:

$$\delta(\mathbf{x}) \cong b + \sum_{k=1}^{M} [\mathbf{w}]_k \tag{7}$$

and

$$[\mathbf{w}]_k = \sum_{j=1}^{N} y_j \alpha_j K(\mathbf{x}_k, \mathbf{z}_{j,k}) \tag{8}$$

where M is the number of features, $\mathbf{x}_k$ is the kth feature of data vector $\mathbf{x}$ and $\mathbf{z}_{j,k}$ is the kth feature of the jth support vector.

Considering the class label y as a dependent variable, the probability that the sample belongs to the positive group (in binary classification problem) is denoted as

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(A + B \times \delta(\mathbf{x}))}} \tag{9}$$

The parameters A and B can be calculated by optimizing the log-likelihood function, as is done in LR. A cross-validation is performed internally to prevent overfitting [26]. After optimizing the parameters A and B, one can revise (9) as

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{k=1}^{M} [\beta]_k\right)}} \tag{10}$$

where $\beta_0 = A + B \times b$ and $[\beta]_k = B \times [\mathbf{w}]_k$. β_0 is an intercept, a constant delineating the prior probability in the absence of any features, and $[\beta]_k$ is the effect vector that maps the value of the kth feature into a point score, which finally becomes a line of the Log OR for the feature in the nomogram as seen in Fig. 1. Henceforth, when the summation of the Log OR scores of each feature becomes high, the probability that y is equal to one (the probability that the sample belongs to the positive class) becomes also high. Note that linear kernel is feasible for decomposing itself by each feature, whereas neither polynomial kernel nor radial basis function (RBF) kernel is decomposable linearly.

4.3. Nomogram-based recursive feature elimination (nomogram-RFE)

Because the prediction output is mainly associated with the effect vector (i.e., Log OR), one can deduce that a feature is more important when the length of the line in the nomogram is longer, as described above. Consequently, a feature selection method based on the nomogram determines more important features according to the lengths of the lines.

From an SVM model trained at each iterative round, the nomogram-RFE method calculates the lengths of lines that correspond to their features in nomograms. As with SVM—RFE, nomogram-RFE can remove features recursively that have low effects on prediction output (i.e., short length of the line). During the iterative training process, one can find a subset of features that provides the best predictive performance.

5. Experimental set-up

5.1. Data preparation

Data of all patients with diabetes who attended the outpatient clinic in Samsung Medical Center, Seoul, Korea, have been collected consecutively for up to 10 years (1996—2005). A total of 4321 adult patients with type 2 diabetes mellitus have taken several physical examinations out of the following 20 items at a visit: glycosylated hemoglobin (HbA_{1C}, %), cholesterol (mg/dL), alkaline phosphatase (ALP, U/L), alanine aminotransferase (ALT, U/L), aspartate aminotransferase (AST, U/L), creatinine (mg/dL), blood urea nitrogen (BUN, mg/dL), triglyceride (mg/dL), white blood cell count (WBC, $\times 10^3\ \mu L^{-1}$), hemoglobin (g/dL), platelet count ($\times 10^3\ \mu L^{-1}$), high-density lipoprotein cholesterol (HDL-C, mg/dL), low-density lipoprotein cholesterol (LDL-C, mg/dL), Na^+ (mmol/L), K^+ (mmol/L), uric acid (mg/dL), microalbumin (μg/min), systolic blood pressure (sBP, mmHg), diastolic blood pressure (dBP, mmHg) and body mass index (BMI, kg/m^2). For the differential diagnosis of diabetic nephropathy, we selected the positive patients from the whole dataset when the following criteria were fulfilled: (1) 20—200 μg/min in urinary albumin (microalbumin), (2) no evidence of microalbumin or renal failure at the time of diabetes diagnosis, and (3) prior evidence of diabetic retinopathy, as the development of renal disease is strongly associated with the occurrence of retinopathy [24,25].

As mentioned above, there was some difficulty in data collection: the patients have not always taken all the tests at the visit, and their visits were very irregular. Some patients visited the hospital less than once a year. Fig. 2 illustrates an imaginary example for a patient that could exist in the dataset. This patient's first visit to this hospital was in the early part of 2000 and only two parameters were measured (HbA_{1C} and cholesterol) at that time. The patient took five tests 8 months later; took four tests another 7 months later, and so on. As shown, this sequential dataset does not always have the same items at every visit and the time gaps are irregular. Thus, one cannot directly use such irregular and incomplete data in SVM classification, which led us to preprocess this data.

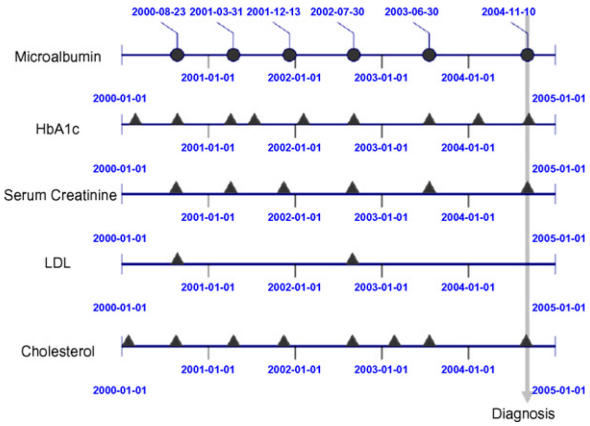


Figure 2 A virtual example of irregular and incomplete data in the clinical setting. This patient was diagnosed with diabetic nephropathy on 10 November 2004 by measuring microalbumin secretion.

5.2. Feature extraction using the quantitative temporal abstraction

Over the past decade, much work has been done to extract relevant features from the time-stamped longitudinal data and the temporal abstraction (TA) is one of the most interesting approaches for that purpose [26—28]. In the clinical domains, the goal of TA task is to evaluate and summarize the state of the patient over a period. Several researches in the diabetes mellitus domain have also incorporated the TA framework [29,30]. Those researches are dealing with high-frequency domains where the clinical parameters have been measured at least a few times a day such as monitoring in the intensive care unit, and focusing mainly on the knowledge-based TA approach which requires the clinician's domain knowledge and derives meta features that form symbolic descriptions of the data. Verduijn et al. compared the qualitative (knowledge-based) TA procedure with the quantitative (data-driven) TA that extracts meta features from the data using statistical summaries (such as mean, variance, the slope coefficient) with minimum use of domain knowledge [31]. They concluded that the qualitative TA procedure is preferable to the quantitative TA in the case study about the prediction from intensive care monitoring data. Thus, we tried to extract the features from the sparse time-stamped data (each variable has been measured a few times per year) by simple data-driven abstraction rather than the knowledge-based TA method.

Using the quantitative TA, we derived the following nine meta features from each laboratory examination dataset to represent the trend of the sequential data over the period between the first visit and the latest visit before the diagnosis. These were the minimum, maximum, mean, variance, slope, estimated value on the date of prediction (EST), initial value, the latest value before the prediction, and K value (see below).

We calculated the slope using linear regression analysis with consecutive values of the laboratory examination data over a given period. We could also estimate the EST, applying the date of prediction to the generated regression model. K value is similar to the stochastic oscillator in the financial domain, which computes the location of the latest feature value relative to its range over a given period, as follows:

$$K = \frac{L - \mathrm{Min}}{\mathrm{Max} - \mathrm{Min}} \tag{11}$$

where L is the latest value, Min is the minimum and Max is the maximum.

After preprocessing, we had 292 records (33 positives and 259 negatives) comprising 184 features for each instance, including the demographic features of the patients: age on the date of prediction, age of onset of diabetes, diabetes duration, and sex.

Table 4 The extracted features after the preprocessing

Feature index	Feature[a]
1–4	Onset age, diabetes duration, age, sex
5–13	White blood cell count (WBC)
14–22	Hemoglobin
23–31	Platelet count
32–40	Serum cholesterol level
41–49	Serum aspartate aminotransferase (AST) level
50–58	Serum alanine aminotransferase (ALT) level
59–67	Serum alkaline phosphatase (ALP) level
68–76	Blood urea nitrogen (BUN)
77–85	Creatinine
86–94	Uric acid
95–103	Na^+
104–112	K^+
113–121	Serum triglycerides level
122–130	High density lipoprotein cholesterol (HDL-C) level
131–139	Low density lipoprotein cholesterol (LDL-C) level
140–148	Glycosylated hemoglobin (HbA_{1C})
149–157	Microalbumin
158–166	Systolic blood pressure (sBP)
167–175	Diastolic blood pressure (dBP)
176–184	Body mass index (BMI)

[a] Each feature set except for 1–4 has 11 features: slope, mean, variance, maximum, minimum, K, EST (estimated value on the date of prediction), initial value and latest value.

Table 4 summarizes the features extracted with the preprocessing step.

5.3. Performance evaluation

Several classification algorithms were compared with each other. Especially for SVMs, the effect of cost-sensitive learning was examined, compared with equal cost learning. More importantly, we evaluated the effects of the various feature selection methods: statistical feature selection, ReliefF, sensitivity analysis, SVM–RFE, and nomogram-RFE.

Throughout the experimental process, we took advantage of leave-one-out cross-validation (LOOCV) to evaluate the performance of each prediction method. The LOOCV is equivalent to k-fold cross validation, where k is the number of data objects. Because the sigmoid parameters should be obtained to use the nomogram approach, we used the probabilistic outputs of SVM classification in all the experiments [32]. Assuming equal loss for misclassified negative and misclassified positive, the optimal threshold for the probabilistic output is $P(y = 1|f) = 0.5$. Instead, an alternative threshold was also applied in the test phase. We used Weka [33] and LIBSVM [34] for implementations of ReliefF and SVM.

Because the dataset in this study had an unbalanced distribution of positives and negatives, the classification accuracy (the rate of correctly classified test samples) was not sufficient as a performance measurement of the predictive models. Suppose that one has a dataset including 10 positives and 90 negatives. A simple decision rule that classifies all the instances as negative would represent 90% accuracy, whereas it could not correctly predict any positive instance. From a medical point of view, a misclassified negative is the most critical decision, because the patient could not have appropriate medical care in that case. However, one also needs to reduce the number of misclassified positives, which leads to unnecessary additional physical examination or treatment.

Accordingly, sensitivity and specificity analysis is common in medicine. Sensitivity stands for the percentage of patients correctly recognized by the classification whereas specificity means the percentage of healthy subjects recognized by the classification. However, there is a trade-off between sensitivity and specificity, meaning it is necessary to calculate other stable evaluation metrics. Therefore, we used the area under the curve (AUC) of a receiver operating characteristic (ROC) curve as a target performance metric. The ROC curve typically plots false positive rate (1 − specificity) versus true positive rate (sensitivity) while a decision threshold is being varied. The

AUC is a convenient way of comparing classifiers where a random classifier has an area of 0.5, and an ideal classifier has an area of 1.0.

We also calculated other judging criteria to evaluate a classifier, such as the balanced error rate (BER) and the harmonic mean of sensitivity and specificity (HMSS). The definition of BER is

$$\mathrm{BER} = \frac{1}{2}\left(\frac{\text{false negative}}{\text{pos}} + \frac{\text{false positive}}{\text{neg}}\right) \tag{12}$$

In an example mentioned above, the BER would be 50% because the first term of (12) is 10/10 but the second would be 0/90. The HMSS is analogous to the *F* measure in information retrieval. The general definition of *F* measure for a non-negative real value α is

$$F_\alpha = \frac{(1+\alpha) \times \text{precision} \times \text{recall}}{\alpha \times \text{precision} + \text{recall}} \tag{13}$$

where the precision is the fraction of predicted positives that are the actual positives while the recall is the fraction of the actual positives that are predicted by a classifier, which is identical to sensitivity. The F_1 measure, where α is one, is the traditional *F* measure that is indeed the harmonic mean of precision and recall, i.e.,

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{14}$$

However, the *F* measure does not take into account performance on the negative class because it is hard to estimate the true negative in information retrieval. To overcome the drawback of the *F* measure, we used HMSS, defined as

$$\mathrm{HMSS} = \frac{2 \times \text{sensitivity} \times \text{specificity}}{\text{sensitivity} + \text{specificity}} \tag{15}$$

This gives equal weight to both sensitivity and specificity, as the F_1 measure does for precision and recall.

6. Results

6.1. Baseline performances with all the features

First, we performed conventional LR without any feature selection method, and it failed to optimize the solution. Therefore, we used ridge LR as a replacement for the conventional method. Table 5 compares the ridge LR, SVM with linear kernel, and SVM with RBF kernel using all 184 features. We applied the equal weights to both positives and negatives for the baseline SVM experiments. For the SVM methods, the prediction accuracies were as high as 0.887 for both kernels when the decision threshold was 0.5. However, the HMSSs gave zeros and the BERs were about 0.5, which were clearly because of the imbalanced dataset and the low sensitivity of the classifiers. Although the SVM with linear kernel showed the best performance in terms of AUC, all three methods showed poor performance (all were lower than 0.7), suggesting the need for a search for feature selection method.

6.2. Statistical feature selection method

We applied the χ^2-test for the feature *sex* and Student's *t*-test for the remaining 183 features. Among the 184 features, Table 6 summarizes significantly different features between the two groups: those who developed (positive) and those who did not develop diabetic nephropathy (negative). For the positive patients, the initial value of

Table 5 Comparison of the logistic regression and SVM kernel methods without cost-sensitive learning or feature selection method[a]

Parameter	Ridge LR[b]	SVM with linear kernel	SVM with RBF kernel
# Features	184	184	184
C+[c]	n/a[d]	1	1
AUC	0.571	0.675	0.644
Accuracy	0.774	0.887	0.887
HMSS	0.559	0	0
BER	0.379	0.5	0.5
Sensitivity	0.424	0	0
Specificity	0.819	1	1

[a] Note that accuracy, HMSS, BER, sensitivity and specificity are measures using a threshold of 0.5 in probabilistic output.
[b] Ridge LR: logistic regression with a ridge estimator.
[c] C+: cost weight for positive instances.
[d] n/a: not available.

Table 6 Results of statistical analysis for each feature between the two groups (only these 17 features showed significant differences from the 184 features tested)

Feature index	Feature name	Positive (n = 33)	Negative (n = 259)	Significance
		Mean ± S.D.	Mean ± S.D.	
12	WBC (initial)	7.26 ± 1.70	6.63 ± 1.61	*
63	ALP (min)	72.76 ± 25.83	63.07 ± 18.22	*
117	Triglyceride (min)	100.82 ± 44.91	82.56 ± 34.59	**
123	HDL-C (mean)	44.76 ± 10.20	49.50 ± 11.31	*
125	HDL-C (max)	53.15 ± 12.59	60.08 ± 14.33	**
130	HDL-C (latest)	41.54 ± 9.93	46.81 ± 13.60	*
141	HbA_{1C} (mean)	8.10 ± 1.12	7.58 ± 1.17	*
143	HbA_{1C} (max)	10.42 ± 1.61	9.75 ± 1.86	*
144	HbA_{1C} (min)	6.46 ± 0.95	6.05 ± 0.94	*
148	HbA_{1C} (latest)	8.10 ± 1.31	7.53 ± 1.57	**
150	Microalbumin (mean)	10.16 ± 4.08	6.68 ± 3.08	***
151	Microalbumin (variance)	2.86 ± 1.87	1.72 ± 1.39	**
152	Microalbumin (max)	13.31 ± 4.31	8.65 ± 4.03	***
153	Microalbumin (min)	7.24 ± 4.75	4.85 ± 2.77	**
156	Microalbumin (initial)	10.23 ± 5.30	6.72 ± 3.70	***
157	Microalbumin (latest)	9.92 ± 5.50	6.79 ± 3.96	**
172	dBP (K)	0.45 ± 0.21	0.53 ± 0.27	*

[*] $P < 0.05$.
[**] $P < 0.01$.
[***] $P < 0.001$.

WBC count ($P < 0.05$), minimum value of ALP ($P < 0.05$), and minimum value of triglyceride ($P < 0.01$) were significantly higher than those of negative patients. For HDL-C mean ($P < 0.05$), maximum ($P < 0.01$), and the latest value ($P < 0.05$) of the positive patients were significantly lower than those of negatives. The positive patients had significantly higher values in mean ($P < 0.05$), maximum ($P < 0.05$), minimum ($P < 0.05$), and the latest value ($P < 0.01$) of HbA_{1C} than negatives. Lastly, for the microalbumin of the positive patients, mean ($P < 0.001$), variance ($P < 0.01$), maximum ($P < 0.001$), minimum ($P < 0.01$), the initial value ($P < 0.01$), and the latest value ($P < 0.01$) were much higher.

Using only these 17 statistically different features, we performed five different combinational approaches. Table 7 shows the results, where an SVM classifier using the RBF kernel and cost-sensitive learning was found to give the best AUC (0.807) and SVM classifiers with equal cost learning gave the worst AUC. The HMSSs and BERs of all the combinational methods were poor because of the low sensitivity of the classifiers, which means that the classifiers predicted—most of the test data as negatives with the threshold of 0.5.

Table 7 Comparison of logistic regression and SVM kernel methods with statistically selected features (accuracy, HMSS, BER, sensitivity, and specificity were measured using a threshold of 0.5 in the probabilistic output)

Parameter	Ridge LR[a]	SVM linear	SVM RBF	SVM linear*[b]	SVM RBF*[b]
# Features	17	17	17	17	17
C+[c]	n/a[d]	1	1	16	13
AUC	0.776	0.276	0.679	0.793	0.807
Accuracy	0.880	0.887	0.890	0.894	0.890
HMSS	0.348	0	0.059	0.167	0.167
BER	0.411	0.5	0.485	0.456	0.458
Sensitivity	0.212	0	0.030	0.091	0.091
Specificity	0.965	1	1	0.996	0.992

[a] Ridge LR: logistic regression with a ridge estimator.
[b] (*) SVM with cost sensitive learning including equal cost learning.
[c] C+: cost weight for positive instances.
[d] n/a: not available.

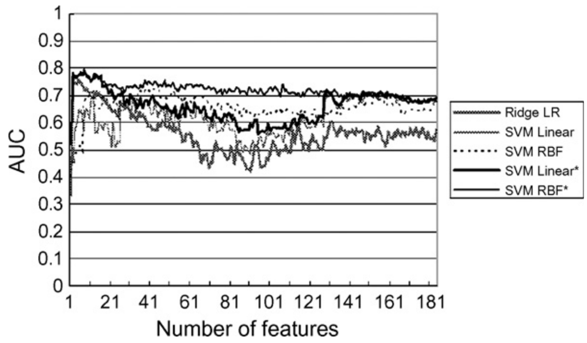


Figure 3 Performance variation of each classifier using the ReliefF method. Ridge LR stands for logistic regression with a ridge estimator and the SVM classifiers that include an asterisk shows the best area under the curve (AUC) values of the receiver operating characteristic curve for cost-sensitive learning and the equal cost learning methods.

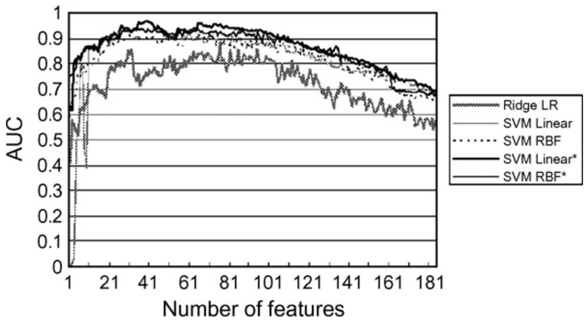


Figure 4 Performance variation of each classifier using the sensitivity analysis method. Ridge LR stands for logistic regression with a ridge estimator and the SVM classifiers with asterisks shows the best AUCs among the cost-sensitive learning and the equal cost learning method.

6.3. ReliefF method

To evaluate the ranking of the features in terms of their effects on the discriminating ability, we applied the ReliefF method to the whole dataset as a preprocessing step. Based on this feature ranking, we eliminated low-ranked features one by one and trained the classifiers again with the remaining features. Fig. 3 plots the performance variation based on the number of features selected by ReliefF. Note that it is easier to understand the graph by examining it from the right to the left, because we used the backward elimination. Even after the classifiers were trained based on the feature ranking by ReliefF, no classifier showed any significant enhancement of performance compared with the classifier with statistical feature selection method. In Table 8, which shows the performance of the classifiers in the best cases, the results are poorer than those from the statistical feature selection method. The HMSSs and BERs were 0 and around 0.5, respectively.

6.4. Sensitivity analysis method

Unlike with the ReliefF method, we trained a classifier first and then eliminated the lowest-ranked feature by estimating the feature ranking based on sensitivity analysis. Fig. 4 illustrates the performance variation versus the number of selected features by sensitivity analysis. Unlike the results with the ReliefF method, the performance variation had a common tendency in that it increased until the number of the features decreased to some point, and then it started to degrade as the number reached one. In most cases, the linear kernel-based SVM classifier trained by the cost-sensitive learning method (SVM linear*) represented the best performance, and LR was the worst with respect to AUC.

Table 9 compares the five classifiers in the best cases with each classification method. Among these, the linear kernel based SVM classifier with 39

Table 8 Comparison of logistic regression and SVM kernel methods with the ReliefF method in the best cases (accuracy, HMSS, BER, sensitivity and specificity were measured using a threshold of 0.5 in the probabilistic output)

Parameter	Ridge LR[a]	SVM linear	SVM RBF	SVM linear*[b]	SVM RBF*[b]
# Features	2	159	44	8	8
C+[c]	n/a[d]	1	1	7	7
AUC	0.758	0.714	0.742	0.784	0.796
Accuracy	0.877	0.887	0.886	0.884	0.886
HMSS	0	0	0	0	0
BER	0.506	0.5	0.5	0.502	0.5
Sensitivity	0	0	0	0	0
Specificity	0.988	1	1	0.996	1

[a] Ridge LR: logistic regression with a ridge estimator.

[b] (*) The best SVM classifier that includes the cost sensitive learning method.

[c] C+: cost weight for positive instances.

[d] n/a: not available.

Table 9 Comparison of logistic regression and SVM kernel methods with sensitivity analysis method in the best cases (accuracy, HMSS, BER, sensitivity and specificity were measured using a threshold of 0.5 in the probabilistic output)

Parameter	Ridge LR[a]	SVM linear	SVM RBF	SVM linear*[b]	SVM RBF*[b]
# Features	76	39	75	39	63
C+[c]	n/a[d]	1	1	1	7
AUC	0.879	0.969	0.918	0.969	0.943
Accuracy	0.911	0.921	0.894	0.921	0.897
HMSS	0.867	0.532	0.263	0.532	0.264
BER	0.130	0.322	0.430	0.322	0.428
Sensitivity	0.818	0.364	0.152	0.364	0.152
Specificity	0.848	0.992	0.988	0.992	0.992

[a] Ridge LR: logistic regression with a ridge estimator.
[b] (*) The best SVM classifier that includes the cost sensitive learning method.
[c] C+: cost weight for positive instances.
[d] n/a: not available.

selected features showed the highest AUC (0.969). Although the HMSS and the BER outcomes of LR were better than those of SVM classifiers, SVM classifiers had higher AUCs. Notably, the best classifier (SVM linear) used equal cost learning, rather than cost-sensitive learning.

6.5. SVM—RFE and nomogram-RFE

Fig. 5 shows that the embedded feature selection methods (SVM—RFE and nomogram-RFE) had exactly the same pattern in performance variation as the number of selected features decreased to some point (approximately 10), meaning that they eliminated the least-ranked features with the same order. Classifiers with cost-sensitive learning seemed superior to those with equal cost learning for most cases. However, both had the same peak points of performance with the same number of features. The results of SVM—RFE and nomogram-RFE (both using linear kernel) had exactly the same pattern as the SVM classifier using linear kernel and sensitivity analysis discussed in Section 6.4.

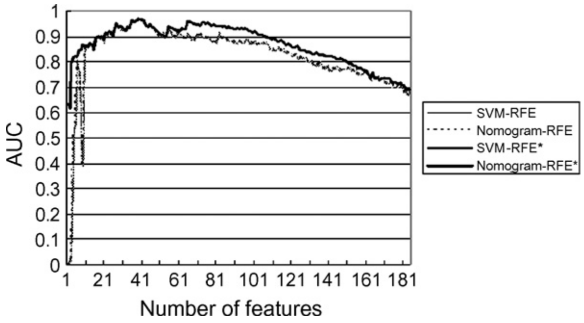


Figure 5 Performance variation of each classifier using the SVM—RFE and nomogram-RFE methods. Ridge LR stands for logistic regression with a ridge estimator, and the SVM classifier that includes an asterisk shows the best AUCs among the cost-sensitive learning and the equal cost learning methods.

For the best cases, all four approaches gave the identical highest AUC (0.969), and they took advantages of 39 features. Although the accuracies of the classifiers were greater than 0.9, the HMSSs and the BERs were still poor, with a threshold of 0.5.

6.6. The best classifier

From the results of the various methods above, SVM classifier with a linear kernel and the 39 selected features showed the best performance in terms of AUC. However, it needed an alternative threshold to enhance the HMSS and the BER. We thus tried to use as an alternative threshold the rate of positives over the total records in the training dataset (the probability threshold varied between 0.12 and 0.13 with respect to the training dataset). When the alternative threshold was applied (Table 10), the HMSS increased and the BER decreased, although accuracy became lower. These changes imply that SVM approaches may need an alternative probability decision threshold instead of 0.5 for imbalanced datasets.

Table 10 Comparison of the threshold variations in the best prediction

Parameter	SVM Linear with 39 features SVM RBF*[a]	
Threshold	0.5	Ratio[b]
# features	39	
C+[a]	1	
AUC	0.969	
Accuracy	0.921	0.839
HMSS	0.532	0.890
BER	0.322	0.104
Sensitivity	0.364	0.970
Specificity	0.992	0.822

[a] C+: cost weight for positive instances.
[b] Ratio: The rate of positives over the total data records.

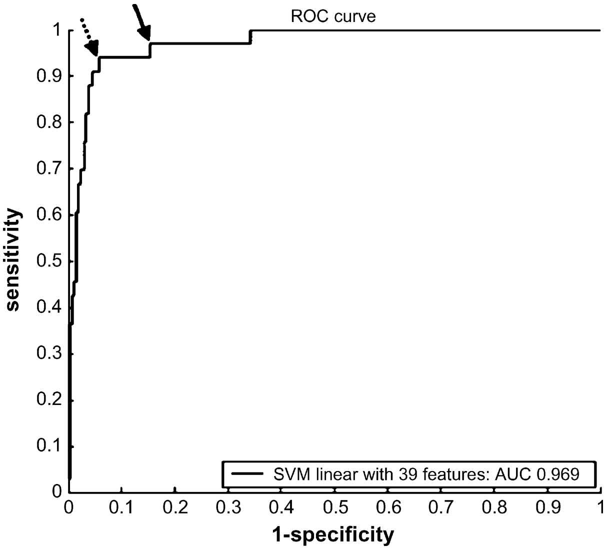
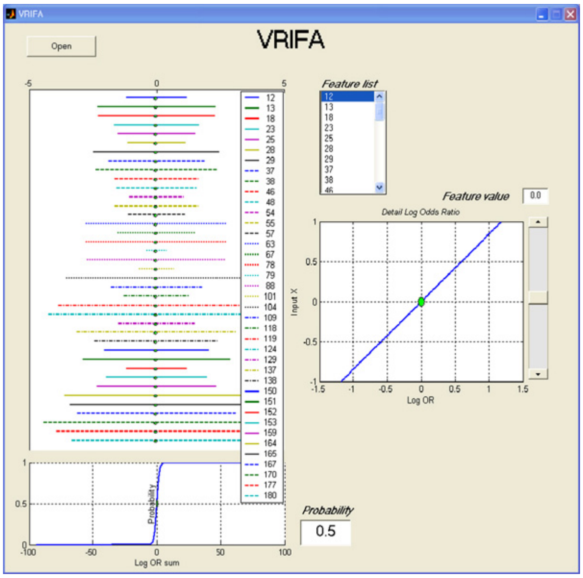


Figure 6 Receiver operating characteristic (ROC) curve of the best classifier. The AUC of the classifier is 0.969. The solid arrow indicates an optimum threshold point (probability of 0.13) for the classifier, which gives a sensitivity of 97% and a specificity of 85%. The dashed arrow indicates another threshold point (probability of 0.25) for the classifier, which gives 94% sensitivity and 95% specificity.

Fig. 6 shows the ROC curve of the best classifier. The solid arrow indicates the optimum point of the classifier, giving a sensitivity of 0.97 and a specificity of 0.85as a threshold is 0.13. The dashed arrow indicates another optimum option that gives a sensitivity of 0.94 and a specificity of 0.95 at a threshold of 0.25.

6.7. Prediction time gap

We applied another statistical test related to the time gap, that is, the interval between the date of the latest examination and the date that we were trying to predict. As shown in Table 11, there was no significant difference of prediction gap between positives and negatives: both groups had about 0.2 years (2.5 months) of prediction gap. For the microalbumin measurements, there was also no significant difference. All patients had around 2 years to take another examination of microalbumin for the final diagnosis of diabetic nephropathy. The total follow-up period was also not significantly different: approximately 5—6 years for both groups.



Feature value 0.0

Probability 0.5

Figure 7 The VRIFA system applied by the linear-kernel-based SVM classifier with 39 selected features, as described in Section 6.6.

6.8. Interpretation of the results using the VRIFA system

Based on the best classifier (nomogram-RFE with the 39 features) in Section 6.6, we have plotted the effects of the selected features using the nomogram visualization tool for risk factor analysis (VRIFA) in Fig. 7. The upper left part of the figure shows the ranges of effect values of the 39 features. In that part, features 124 (HDL-C variance) and 170 (dBP maximum) seem to have had high effects on the prediction because they show wider ranges than others do. The detailed Log OR (effect value) of each feature at a specific input value can be seen in the right part of VRIFA. By applying all the weight values and the intercept to the expression in Eq. (10), we show an estimate of the probability of having diabetic nephropathy at the bottom of the figure.

Table 11 The results of Student's *t*-test applied to the means of the prediction gap and time of follow-up between the two groups (all tests were non-significant)

Time	Mean ± S.D.	
	Positive (n = 33)	Negative (n = 259)
Prediction gap (year)	0.21 ± 0.24	0.19 ± 0.12
Microalbumin gap (year)	2.05 ± 0.99	1.87 ± 0.91
Follow-up period (year)	5.37 ± 1.98	5.86 ± 1.99

Fig. 8 shows the detailed Log OR for all the selected features. When the Log OR increases to the upper right end as the feature value increases as for the initial WBC, it shows that the higher the feature value, the higher is the probability that the patients would develop diabetic nephropathy. By contrast, when the line goes down as the feature value increases (as for minimum Hemoglobin value), the higher the value of the lower is the probability that a patient might develop diabetic nephropathy.

Examining each feature closely, some of the features selected statistically in Section 6.2 have also been included in this selected feature set (initial WBC; minimum ALP; mean, maximum and minimum microalbumin, and variance in microalbumin), and they show similar tendencies: thus a higher value for each feature indicates positive (bad) effects of having diabetic nephropathy. However, some other features have inconsistent effects compared with the commonly believed facts. For example, the effect values decrease as the mean, EST and initial values of systolic blood pressure increase. In addition, the effect values decrease as the mean and minimum values of BMI increase. However, these features showed no statistically significant difference between groups.

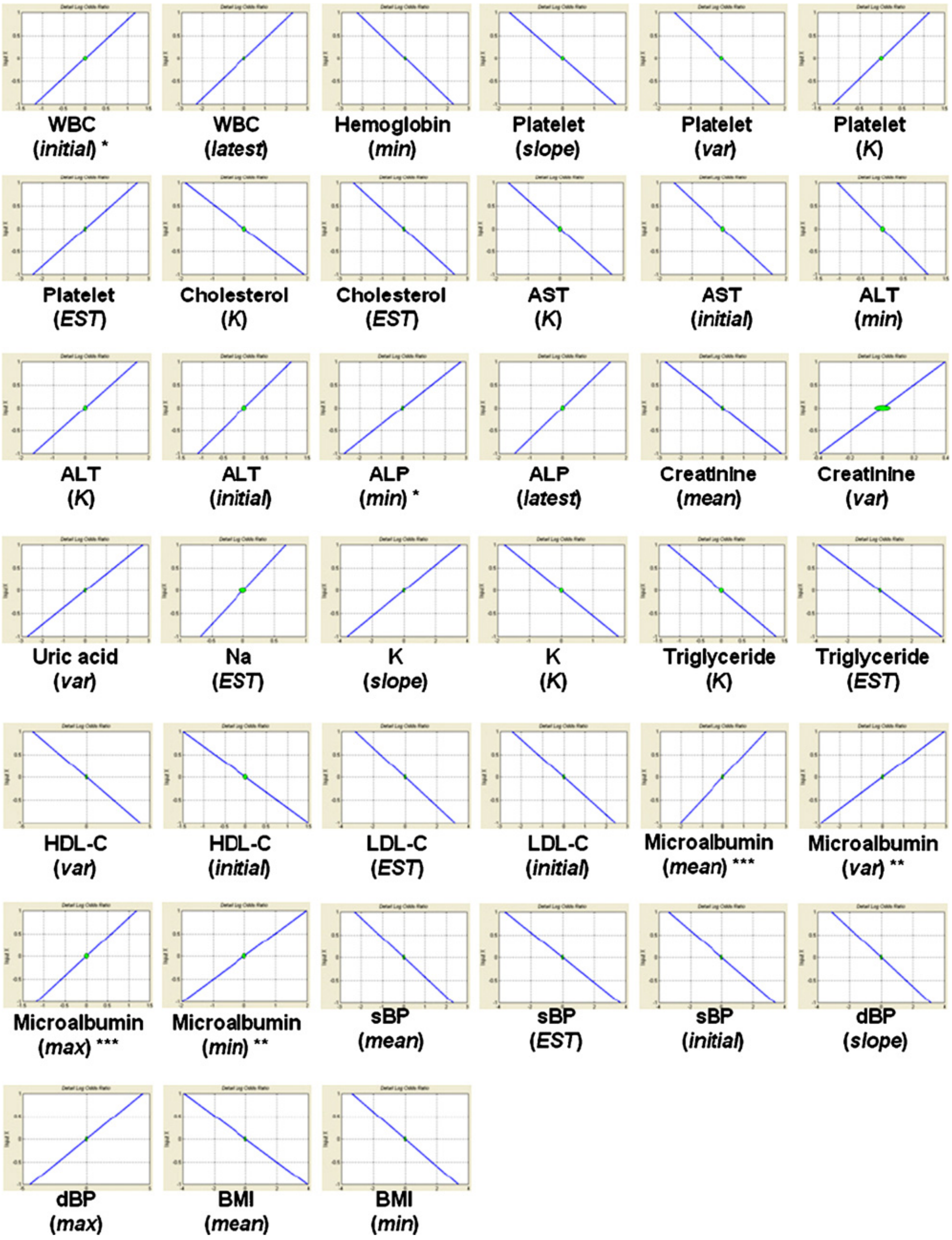

WBC (*initial*) *
WBC (*latest*)
Hemoglobin (*min*)
Platelet (*slope*)
Platelet (*var*)
Platelet (*K*)
Platelet (*EST*)
Cholesterol (*K*)
Cholesterol (*EST*)
AST (*K*)
AST (*initial*)
ALT (*min*)
ALT (*K*)
ALT (*initial*)
ALP (*min*) *
ALP (*latest*)
Creatinine (*mean*)
Creatinine (*var*)
Uric acid (*var*)
Na (*EST*)
K (*slope*)
K (*K*)
Triglyceride (*K*)
Triglyceride (*EST*)
HDL-C (*var*)
HDL-C (*initial*)
LDL-C (*EST*)
LDL-C (*initial*)
Microalbumin (*mean*) ***
Microalbumin (*var*) **
Microalbumin (*max*) ***
Microalbumin (*min*) **
sBP (*mean*)
sBP (*EST*)
sBP (*initial*)
dBP (*slope*)
dBP (*max*)
BMI (*mean*)
BMI (*min*)

Figure 8 The detailed effect values (log odds ratios) of the selected features. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

7. Discussion

7.1. Data preparation and feature extraction

It was difficult to acquire such a large electronic medical dataset, with 4321 patient follow-up records aged up to 10 years. Moreover, the original data were irregular and incomplete sequences, which made this study challenging. Therefore, we extracted 184 meta features for preprocessing, shrinking the number of records to 292; this step was essential for the machine learning application. To represent a record for each physical and laboratory examination (20 items), we employed the qualitative TA approach that calculates several fundamental statistical features such as minimum, maximum, mean and variance; we also estimated the trend of each item. These extracted features sufficed to describe the data and to distinguish the two groups of patients who did or did not develop diabetic nephropathy, because the classification performance with several feature selection methods showed promising results. However, our future work will compare the simple statistical abstraction with more complex abstraction with the knowledge-based TA procedure.

Despite this reduction in data, there was unlikely to be any bias in sampling. In the preprocessing step, we selected all data records under the same conditions. Thus, they all needed the full 184 features, which means that the patients should have had at least two examinations for all 20 items before the final diagnosis. Moreover, there were no significant differences in age, sex, age at the onset of diabetes, or the duration of diabetes between the two groups.

7.2. Feature selection methods

There were too many features to be trained compared with the number of records; this raises the issue of "curse of dimensionality", which refers to the exponential growth of hypervolume as a function of dimensionality. To deal with this, we adopted several feature selection methods; some of which showed better results than the classifier using all 184 features. This fulfils the dictum of Occam's razor in that the fewer assumptions are made to explain a phenomenon, the better it is.

Using statistical analysis, we clarified that some features differed significantly different between the two groups. However, those features were not sufficient to distinguish the two groups because the classification performance using those features was poor compared with those with other feature selection methods. Statistical methods, such as Student's *t*-test, compare the means of groups and indicate if there is any probable difference between them. However, such statistical differences could not always guarantee linear or non-linear separable values for each group of patients, so it was hard to predict diabetic nephropathy using a statistical feature selection.

Filter methods such ReliefF have advantages in computation because they do not interact with classifiers, which is why they have been used for simple approaches. By contrast, the wrapper and embedded methods consider classifier design and are thus computationally expensive. However, these tactics — sensitivity analysis, SVM—RFE and nomogram-RFE — showed better performances than ReliefF.

7.3. Classification methods

The failure of conventional LR here could have arisen from the large number of features. The preprocessing step might be another source of this failure, as we extracted these features from irregular sequential datasets and they must have had high correlations with each other (i.e., multicolinearity). With every feature selection method used in this study, SVM classification was superior to ridge LR in terms of the AUC. However, SVM classification using cost-sensitive learning gave almost the same results as the equal cost learning method. Cost-sensitive learning moves the separating hyperplane away to the minority class data; this is very similar to the bias variation scheme, in which the user can alter the decision threshold using bias. Moreover, threshold variation does not related with the AUC of the ROC curve. Therefore, the AUC may depend mostly on the tuning of other parameters, rather than the penalty to the error of the minority class data.

7.4. Considerations of the VRIFA system

LR is used widely in medicine because it is able to describe feature ranking by an odds ratio and it can give intuitive information to medical practitioners. However, it is very difficult to interpret the results using odds ratio when independent variables (input features) have continuous values. Therefore, ridge LR made little use of this advantage, because the features used in this study all had continuous values except for one feature: sex.

VRIFA, a visualization system using a nomogram in SVM, has been developed as a prototype. It gives intuitive visualization of the effect of each feature and a probabilistic output. Physicians can benefit from it for predicting diabetic nephropathy and can analyze the effects of the features. Therefore, it may be very useful to help develop effective treatment strategies and guide patients in choosing a life style. In addition to its graphical output, we took advantage of VRIFA for a feature selection method (nomogram-RFE) using the dynamic ranges of features. In this application, nomogram-RFE produced a value of 0.969 of the AUC using 39 features, as did the other wrapper and embedded methods.

7.5. Clinical considerations

The features selected by our feature selection methods were not always concordant with statistically significant features. Some of the selected features in the VRIFA showed unexpected tendencies that are contrary to common medical thoughts: for example, that high blood pressure and a high BMI might be associated with diabetic nephropathy. However, these features were not significantly different between the two groups of patients. A possible explanation of the phenomenon is that the best classifiers used linear functions and thus there exist linear relationships among the features. For example, feature A could be important only if it is used with feature B in a linear SVM classifier. These relationships cannot be identified by *t*-test, as it does not take into account of the linear relationship.

On the other hand, the statistically significant features showed the same tendencies as the visualization output. The patients with diabetic nephropathy already had higher values than the unaffected patients in WBC counts and microalbumin values, which is consistent with previous researches that are described in Section 1. In particular, frequent high microalbumin values before the outbreak of diabetic nephropathy suggests a possible demand for lowering of the diagnostic threshold for microalbumin for such patients.

The start of diabetic nephropathy does not depend on some particular features, but on a complex interrelationship involving many. However, physicians could use this VRIFA system to estimate a patient's overall probability of having diabetic nephropathy and easily find the most important risk factors. Early stage of diabetic nephropathy is reversible, meaning that a patient who does not develop overt albuminuria (over 200 μg/min) could revert to normal kidney function if treated appropriately. Therefore, early diagnosis of this disease is very important and this study could be an immediate option for this purpose. When analyzing the time interval of prediction, our proposed method used average 5-year follow-up data and predicted diabetic nephropathy about 2—3 months before the actual diagnosis.

The limitation of this study is that it did not deal with medication information, for example the use of insulin, oral agents, or antihypertensive drugs. There are still difficulties in including such information. Therefore, the concept of prediction is becoming increasingly difficult to pursue because many patients are treated with anti-hypertensive drugs and other types of interventions when microalbuminuria is diagnosed; such measures often return the patient's albumin excretion to normal [35,36]. This problem may require another challenging data mining task, for example ontology and text mining. Future work will include the medication information and more data records to support the existing models.

8. Conclusions

The goal of any prognostic study with machine learning methods is to support rather than replace clinical judgment. Our model scores are not sufficiently accurate or complete to supplant decision-making by physicians. In this aspect, the main objective of this study has suggested a new way to give information to physicians to plan efficient and proper treatment strategies.

In this study, we tried to predict diabetic nephropathy and determine its risk factors. We performed a preprocessing step to deal with the incomplete and unbalanced practical dataset, and assessed various machine-learning techniques, such as SVMs, cost-sensitive learning and feature selection methods. The proposed method predicted the onset of diabetic nephropathy with promising performance and provided a graphical tool for risk factor analysis, which may generate new hypothesis that motivates further investigation and research. In addition, we were able to detect high microalbumin values for the patients before they developed diabetic nephropathy.

The most significant aspect of this study is that, to our knowledge, it is the first trial to apply data mining technology for predicting a diabetic complication. The number of hospitals that adopted an electronic medical record system has increased geometrically during the past decade. This will probably lead to the more efficient collection of data and make it easier to study diseases such as diabetes and hypertension via machine learning or artificial intelligence technology. Thus we believe that this study will have wide application.

Acknowledgement

This study was supported by a Nano-bio Technology Development Project, Ministry of Science and Technology, Republic of Korea (2005-01249).

References

[1] Alberti KG, Zimmet PZ. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1. Diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. Diabetic Med 1998;17: 539—53.

[2] Jo HS, Sung JH, Choi JS, Hwang MS, Jeong HJ, Bae SC. Quality control of diagnostic coding in the Korean Burden of Disease Project. In: Schellekens W, editor. Proceedings of the international conference of the international society for quality in health care. Oxford: Oxford University Press; 2004 p. 181.

[3] Gross JL, De Azevedo MJ, Silveiro SP, Canani LH, Caramori ML, Zelmanovitz T. Diabetic nephropathy: diagnosis, prevention, and treatment. Diabetes Care 2005;28:164—76.

[4] Mogensen CE. Microalbuminuria, blood pressure and diabetic renal disease: origin and development of ideas. Diabetologia 1999;42:263—85.

[5] Rossing P, Hougaard P, Borch-Johnsen K, Parving H. Predictors of mortality in insulin dependent diabetes: 10 year observational follow up study. Br Med J 1996;313:779—84.

[6] Shichiri M, Kishikawa H, Ohkubo Y, Wake N. Long-term results of the Kumamoto study on optimal diabetes control in type 2 diabetic patients. Diabetes Care 2000;28(Suppl. 2):B21—9.

[7] UKPDS. Intensive blood—glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes. Lancet 1998;352:837—53.

[8] Agardh C, Agardh E, Torffvit O. The prognostic value of albuminuria for the development of cardiovascular disease and retinopathy: a 5-year follow-up of 451 patients with type 2 diabetes mellitus. Diabetes Res Clin Pract 1996;32:35—44.

[9] Selby JV, Ferrara A, Karter AJ, Liu J, Ackerson LM. Developing a prediction rule from automated clinical databases to identify high-risk patients in a large population with diabetes. Diabetes Care 2001;24:1547—55.

[10] Carson ER. Decision support systems in diabetes: a systems perspective. Comput Methods Progr Biomed 1998;56:77—91.

[11] Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artif Intell Med 2002;26:37—54.

[12] Park J, Edington DW. A sequential neural network model for diabetes prediction. Artif Intell Med 2001;23:277—93.

[13] Polak S, Mendyk A. Artificial intelligence technology as a tool for initial GDM screening. Expert Syst Appl 2004;26:455—60.

[14] Shanker MS. Using neural networks to predict the onset of diabetes mellitus. J Chem Inform Comput Sci 1996;36:35—41.

[15] Cios KJ, Moore GW. Uniqueness of medical data mining. Artif Intell Med 2002;26:1—24.

[16] Jakulin A, Mozina M, Demsar J, Bratko I, Zupan B. Nomograms for visualizing support vector machines. In: Grossman R, Bayardo R, Bennett K, editors. Proceedings of the 17th international conference on knowledge discovery and data mining (KDD '05). New York: ACM Press; 2005. p. 108—17.

[17] Cessie S, Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc Ser C: Appl Stat 1992;41:191—201.

[18] Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.

[19] Veropoulos K, Cristianini N, Campbell C. Controlling the sensitivity of support vector machines. In: Aiello LC, Dean T, Kollerbaur A, editors. Proceedings of the 16th international joint conference on artificial intelligence (IJCAI '99), workshop ML3. San Francisco: Morgan Kaufmann; 1999. p. 55—60.

[20] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157—82.

[21] Kononenko I. Estimating attributes: analysis and extensions of relief. In: Bergadano F, Raedt LD, editors. Proceedings of 7th European conference on machine learning (ECML'94). Berlin: Springer; 1994. p. 171—82.

[22] Stevensen M, Winter R, Widrow B. Sensitivity of feed forward neural networks to weight errors. IEEE Trans Neural Networks 1990;1:71—80.

[23] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389—422.

[24] Gilbert RE, Tsalamandris C, Allen TJ, Colville D, Jerums G. Early nephropathy predicts vision-threatening retinal disease in patients with type 1 diabetes mellitus. J Am Soc Nephrol 1998;9:85—9.

[25] Mogensen CE, Vigstrup J, Ehlers N. Microalbuminuria predicts proliferative diabetic retinopathy. Lancet 1985;1: 1512—3.

[26] Shahar Y, Tu SW, Musen MA. Knowledge acquisition for temporal-abstraction mechanisms. Knowl Acquis 1992;4: 217—36.

[27] Shahar Y, Musen MA. Knowledge-based temporal abstraction in clinical domains. Artif Intell Med 1996;8:267—98.

[28] Stacey M, McGregor C. Temporal abstraction in intelligent clinical data analysis: a survey. Artif Intell Med 2007;39:1—24.

[29] Bellazzi R, Larizza C, Magni P, Montani S, Stefanelli M. Intelligent analysis of clinical time series: an application to diabetic patients monitoring. Artif Intell Med 2000;20: 37—57.

[30] Seyfang A, Miksch S, Marcos M. Combining diagnosis and treatment using Asbru. Int J Med Inform 2002;68:49—57.

[31] Verduijn M, Sacchi L, Peek N, Bellazzi R, de Jonge E, de Mol BAJM. Temporal abstraction for feature extraction: a comparative case study in prediction from intensive care monitoring data. Artif Intell Med 2007;41:1—12.

[32] Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smolar AJ, Bartlett P, Schoelkopf B, Schuurmans D, editors. Advances in large margin classifiers. Cambridge: MIT Press; 1999. p. 61—74.

[33] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. San Francisco: Morgan Kaufmann; 2005.

[34] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm; 2001 [accessed September 4, 2007].

[35] Christensen CK, Krussel LR, Mogensen CE. Increased blood pressure in diabetes: essential hypertension or diabetic nephropathy? Scand J Clin Lab Invest 1987;47:363—70.

[36] Pedersen EB, Mogensen CE. Effect of antihypertensive treatment on urinary albumin excretion, glomerular filtration rate and renal plasma flow in patients with essential hypertension. Scand J Clin Lab Invest 1976;36:231—7.