J. Biomed. Eng. Res : 355-362, 2007

특징점 선택방법과 SVM 학습법을 이용한 당뇨병 데이터에서의 당뇨병성 신장합병증의 예측

조백환[1], 이종실[1], 지영준[1], 김광원[2], 김인영[1], 김선일[1]

[1]한양대학교 의용생체공학과
[2]성균관대학교 의과대학 내분비 대사 내과

(Received March 16, 2007. Accepted May 14, 2007)

Prediction of Diabetic Nephropathy from Diabetes Dataset Using Feature Selection Methods and SVM Learning

Baek Hwan Cho[1], Jong Shill Lee[1], Young Joon Chee[1], Kwang Won Kim[2], In Young Kim[1], Sun I. Kim[1]

[1]Department of Biomedical Engineering, Hanyang University
[2]Department of Endocrinology and Metabolism, Sungkyunkwan University

Abstract

Diabetes mellitus can cause devastating complications, which often result in disability and death, and diabetic nephropathy is a leading cause of death in people with diabetes. In this study, we tried to predict the onset of diabetic nephropathy from an irregular and unbalanced diabetic dataset. We collected clinical data from 292 patients with type 2 diabetes and performed preprocessing to extract 184 features to resolve the irregularity of the dataset. We compared several feature selection methods, such as ReliefF and sensitivity analysis, to remove redundant features and improve the classification performance. We also compared learning methods with support vector machine, such as equal cost learning and cost-sensitive learning to tackle the unbalanced problem in the dataset. The best classifier with the 39 selected features gave 0.969 of the area under the curve by receiver operation characteristics analysis, which represents that our method can predict diabetic nephropathy with high generalization performance from an irregular and unbalanced dataset, and physicians can benefit from it for predicting diabetic nephropathy.

Key words : diabetic nephropathy, feature selection, support vector machine, cost-sensitive learning

I. 서 론

당뇨병은 여러 가지 병인을 갖는 대사질환의 일종이며, 인슐린의 분비, 인슐린의 작용, 혹은 두 가지 모두에 의한 만성적인 고혈당을 특징으로 한다[1]. 최근 서구식 생활문화의 유입으로 인해 한국에서는 약 8.4%의 인구가 당뇨병을 앓고 있으며, 한국인의 질병부담 중에서 1위를 차지하고 있다[2]. 이러한 당뇨병은 심혈관질환, 신장질환, 하지절단, 시력감퇴를 포함한 여러 가지 합병증을 동반할 수 있는데, 심각한 경우 심한 장애나 사망에까지 이를 수 있다. 그 중에 당뇨병성 신장합병증은 대표적인 예라 할 수 있다.

본 논문은 과학기술부의 특정연구 개발 과제의 나노 바이오 기술 개발 바이오 전자 사업의 연구결과로 수행되었습니다. (2005-01249)

Corresponding Author : 김인영
서울 성동구 행당동 17 한양대학교 의용생체공학과
Tel : 02-2291-1713 / Fax : 02-2296-5943
E-mail : iykim@hanyang.ac.kr

현재까지 당뇨병성 신장합병증의 위험인자를 찾기 위한 많은 연구가 진행되어 왔다. 수년간의 지속적인 고혈당과 고혈압이 신장질환 및 심혈관질환에 주요 위험인자라고 알려졌으며[3, 4], 백혈구수의 증가, 혈소판 증가, 고지혈증 등도 당뇨병성 신장합병증과 관련되어 있다는 보고가 있었다[5-7]. 그러나 대부분의 선행연구들은 Student's t test나 Mann-Whitney U test와 같은 간단한 통계적인 기법들을 이용하여, 당뇨병성 신장합병증 환자군과 대조군에서 이들 위험인자의 평균치를 단순 비교한 결과이다. 신장기능의 퇴화는 개인마다 상당한 차이를 나타내며, 당뇨병성 신장합병증은 몇 가지의 결정적인 요인으로 설명되기 보다는, 오히려 여러 가지 복잡한 요인들의 종합적인 관계에 의하여 발병되는 것이라고 여겨진다. 따라서 단순한 통계기법을 통한 당뇨병성 신장합병증의 예측은 매우 어렵다.

이러한 문제점을 해결하기 위하여 최근 수십 년 동안 데이터 마이닝과 기계 학습 분야의 눈부신 발전으로 인해, 컴퓨터 보조진단, 전문가 시스템, 및 예후 연구 등 다양한 의학적 문제에 대한 접근이

이루어져 왔다 [8]. 그렇지만 이러한 기술들을 의학에 접목하기 위해서는 또한 많은 문제점들이 있으며, 그 중에 데이터의 획득에 있어서 가장 특징적인 문제점이 발견된다. 병원에서 전자의무기록 시스템을 사용한지 오래 되지 않았기 때문에, 당뇨병성 신장합병증의 예측과 같은 연구를 수행하기 위해 필요한, 많은 환자의 오랜 기간에 걸친 데이터를 구축하기가 아직까지는 힘들다. 또한, 대학병원과 같은 종합병원에서는 임상의가 다른 기관으로 빈번하게 이동할 수 있으므로, 임상의마다 실험실 검사 등의 환자에 대한 검사 및 치료 프로토콜이 부분적으로 상이할 수 있다. 뿐만 아니라, 환자 자신이 건강에 대한 자만이나 다른 이유로 인해 병원을 내원하는 시기가 불규칙적일 수 있다. 이러한 임상적 환경들 때문에 병원에서 구축되는 데이터는 매우 불규칙적이고 불완전한 것일 가능성이 매우 높고, 따라서 의미 있는 정보를 추출하기가 매우 어렵다.

많은 경우에 있어서, 의학데이터는 양성데이터(암이나 심장병과 같은 심각한 질병을 가진 환자의 데이터)가 음성데이터(질병에 노출되지 않은 정상인의 데이터)보다 훨씬 적다. 이러한 불균형적인 데이터를 이용한 분류(classification) 혹은 예측(prediction) 연구의 경우에는 한쪽으로 치우친 결과(skewed result)대부분의 테스트 데이터에 대하여 majority 클래스로 판별할 가능성이 높다. 이런 경우 가장 간단한 방법으로, majority 클래스의 데이터의 개수를 minority 클래스의 데이터 개수와 동일하게 샘플링 하거나, minority 클래스의 데이터를 중복(duplicate)시켜 majority 클래스의 데이터 개수와 동일하게 하는 방법이 있다. 그러나 이런 방법으로 생성된 분류 규칙이나 예측 모델이 모집단을 반영한다고 보기 어렵고, 그 예측 성능 또한 보장하기 어렵다는 심각한 단점이 있다[9].

데이터 마이닝 기술의 의학적 접목에 있어서 또 다른 특징적인 점은 환자나 질병을 설명하는 특징점들이 매우 많을 수 있으며, 이는 유전체학과 같은 생물정보학 분야에서 쉽게 찾아 볼 수 있다. 예를 들어, 어떤 환자에 대해 질병의 유무를 DNA microarray를 통한 유전자 검사를 통해 진단할 때, 수 만개의 유전자를 한꺼번에 검사하며, 이 때의 특징점의 개수가 수 만개에 달하는 것이다. 특징점들이 많은 경우에는 예측 모델을 계산하는데 매우 많은 시간과 메모리가 소요되고, 불필요한 특징점들로 인한 성능 저하도 예상될 수 있다.

따라서 본 연구에서는 당뇨병을 가지고 있는 환자의 검사실 검사 및 각종 환자 정보를 이용하여 전처리를 거쳐 특징점들을 추출한 후, 여러 가지의 특징점 선택 알고리즘을 적용하여 불필요한 특징점들을 제거하고, 데이터의 불균형에 의한 문제점 해소를 위하여 support vector machine(SVM)을 이용한 cost-sensitive 학습법 등을 통하여 당뇨병성 신장합병증을 예측하고자 한다.

II. 재료 및 방법

A. 데이터 수집

삼성서울병원을 방문한 제 2형 당뇨병을 가지고 있는 외래환자의 데이터를 최장 10년간 (19962005) 연속적으로 수집하였다. 총 4321명의 당뇨병 환자가 표 1의 20가지의 검사항목 중 몇 가지를 검사 받아 그 결과를 저장하였고, 환자의 당뇨병성 신장합병증을 감별진단하기 위하여 다음과 같은 진단 기준으로 양성 환자를 검색하였다.

(1) 단백뇨 20~200 g/min

(2) 당뇨병 발병 시 미세단백뇨나 신장질환이 없음을 확인

(3) 이전에 당뇨병성 망막병증의 발병

서론에서 밝혔듯이 당뇨 환자가 불규칙적으로 외래방문을 하였고 방문할 때마다 항상 같은 검사항목을 수행하지 않았기 때문에, 구축된 데이터도 불규칙적인 데이터의 형태를 나타내었다. 따라서 이러한 데이터를 곧바로 이용하여 기계 학습과 같은 연구를 수행할 수 없으므로 전처리 과정을 수행하였다.

B. 전처리

각 실험실 검사 항목에서 최초 방문 시기와 최종 진단 직전 사이의 환자 상태의 추세를 나타내기 위하여, 각 검사 항목당 최소값, 최대값, 평균, 분산, 기울기, 추정값, 초기값, 최근값, K값 등 9개의 특징점을 추출하였다. 기울기는 주어진 기간 동안 동일한 검사항목의 데이터의 추세를 나타내는 값으로써 선형회귀분석법(linear regression analysis)을 적용하여 계산하였고, 추정값은 선형회귀분석을 이용하여 구한 회귀모델을 이용하여 최종 진단 시기의 검사 값을 추정하여 계산하였다. K값은 경제 분야에서 자주 이용되는 진동지표(stochastic oscillator)로써, 일정 기간 동안 지속적으로 측정된 어떤 변수 값들이 있을 때 가장 최근값이 이전의 값들에 비해 평균적으로 높은지 혹은 낮은지를 측정하는 지표이며 다음 식과 같이 계산하였다.

표 1. 임상 검사 항목
Table 1. Clinical examination items

검사 항목	검사 항목
glycosylated hemoglobin (HbA1C, %)	platelet count (10㎕/L)
low-density lipoprotein cholesterol (LDL-C, mg/dL)	high-density lipoprotein cholesterol (HDL-C, mg/dL)
alkaline phosphatase (ALP, U/L)	cholesterol (mg/dL)
alanine aminotransferase (ALT, U/L)	Na+ (mmol/L)
aspartate aminotransferase (AST, U/L)	K+ (mmol/L)
creatinine (mg/dL)	uric acid (mg/dL)
blood urea nitrogen (BUN, mg/dL)	microalbumin (g/min)
triglyceride (mg/dL)	systolic blood pressure (sBP, mmHg)
white blood cell count (WBC, 10㎕/L)	diastolic blood pressure (dBP, mmHg)
hemoglobin (g/dL)	body mass index (BMI, kg/m2)

$$K = \frac{L - Min}{Max - Min} \quad (1)$$

L 은 가장 최근값을, Min은 최소값, Max는 최대값을 각각 나타낸다.

전처리를 거친 후, 진단 예측 시기의 나이, 당뇨병 발병 시 나이, 당뇨병 이환기간 및 성별을 포함하여 총 184개의 특징점을 각 환자 별로 추출하였고, 최종적으로 연구에 포함한 환자는 184개의 특징점을 모두 추출할 수 있는 환자로써 총 292명(양성 환자 33명, 음성 환자 259명)이었다.

C. 특징점 선택(Feature Selection) 방법

특징점 선택 방법은 분류 및 예측 연구에 있어서 성능향상에 불필요한 여분의 특징점들을 제거하고 가장 최적화된 특징점들만 선택하는 기계 학습 기법에서 사용되는 과정이다[10]. 일반적으로 filter, wrapper, embedded 방법으로 크게 나누어지는데, filter 방법은 전처리 과정의 일종으로 알려져 있으며 통계적인 수치에 의해 각 특징점들의 순위를 계산하여 높은 순위를 갖는 특징점을 선택하는 방법이다. Wrapper 방법은 분류기(classifier)를 black box로 여겨서 각 특징점이 예측에 어느 정도 영향을 미치는지 계산하여, 많은 영향을 미치는 특징점들을 선택하는 방법이며, embedded 방법은 wrapper 방법과 비슷하지만 분류기 설계와 보다 더 깊이 관여하여 각 특징점들의 영향력을 직접적으로 계산하는 방법이다.

표 2. ReliefF 알고리즘
Table 2. Algorithm of the ReliefF method

Line	Code
Inputs	All the instances and their class labels
1	Set all weights $W[f]=0$ 7
2	Set arbitrary iteration number a
3	for $i=1$ to a do
4	Randomly select an instance I_i
5	Find the k nearest hits H_j
6	Find the k nearest misses $M_j(C)$ for each class $C \neq class(I_i)$
7	for $f=1$ to the number of features do
8	$W[f] = W[f] - \sum_{j=1}^{k} \mathrm{diff}(f, I_i, H_j)/(m \cdot k) + \sum_{c \neq class(I_i)} \left[\frac{P(C)}{1 - P(class(I_i))} \sum_{j=1}^{k} \mathrm{diff}(f, I_i, M_j(c)) \right] / (m \cdot k)$
9	end for
10	end for
Outputs	Weight vector $W[f]$

ReliefF

ReliefF는 filter 방법의 일종으로써, 각 특징점이 클래스 간의 차이와 클래스 내의 유사도에 얼마나 영향을 미치는지를 평가하는 알고리즘이다[11]. 이 알고리즘은 전체 데이터에서 무작위로 선택된 데이터에 대하여 k개의 Euclidean 거리나 Manhattan 거리를 이용하여, 가장 가까운 거리에 있는 hit(선택된 데이터와 같은 클래스의 데이터)과 miss(선택된 데이터와 다른 클래스의 데이터)를 먼저 찾는다. 그 후에 hit과 miss의 각 특징점 값과 선택된 데이터의 특징점 값간 의 차이를 계산한다. 이러한 과정을 반복하여 각 특징점이 예측에 미치는 영향을 평가하는데 표 2에 ReliefF의 알고리즘을 간단하게 나타내었다.

함수 $\mathrm{diff}(f, I_i, I_j)$는 두 개의 데이터에서 각 특징점 값의 차이를 계산하는 함수이며, 따라서 가중치 벡터 $W[f]$는 선택된 데이터의 특징점 값과 miss의 특징점 값 $M_j(C)$의 차이가 크면 증가하고, 반대로 hit의 특징점 값 H_j의 차이가 크면 감소한다. 결국 가중치 벡터 $W[f]$의 각 원소가 각 특징점의 순위를 나타내는데, 값이 클수록 높은 순위를 나타낸다고 할 수 있으므로, 가중치 값이 낮은 특징점들을 제거해 나가는 방법으로 특징점 선택 방법을 적용한다.

Sensitivity Analysis

Sensitivity analysis는 wrapper 방법의 일종으로써, 표 3에 sensitivity 알고리즘을 간단하게 나타내었다[12]. Sensitivity analysis 알고리즘은 모든 특징점들을 특정 값으로 고정시킨 후에 하나씩 최소값에서 최대값으로 변화시키면서 출력의 변화량을 측정한다. 즉, 출력의 변화량이 클수록 더 중요한 특징점으로 판단할 수 있으며, ReliefF와 동일한 방법으로 낮은 순위의 특징점들을 제거해 나가는 방법으로 특징점을 선택한다.

표 3. Sensitivity Analysis 알고리즘
Table 3. Algorithm of the Sensitivity Analysis method

Line	Code
Inputs	A predictive model $F(x)$
1	Set all weights $W[f]=0$
2	Set $O[a]=0$
3	for $f=1$ to *the number of features* do
4	Initialize an instance $x=[x_1=mean(x_1), x_2=mean(x_2), \cdots, x_n=mean(x_n)]$
5	for $j=min(x-f)$ to $max(x_f)$ do
6	Set $x_f=j$
7	Set $O[j]=F(x)$
8	end for
9	$W[f]=\max(O)-\min(O)$
10	end for
Outputs	Weight vector $W[f]$

D. Support Vector Machine(SVM) 및 Cost-Sensitive 학습법

SVM은 분류 및 예측 기법에 사용되는 기계 학습법의 일종으로써, 두 그룹의 데이터를 구분시키는 초월평면(hyperplane)을 계산하는 방법이며, 이 때 마진(margin)을 최대화하는 방법이다[13-15]. 마진은 초월평면과 직교방향으로 가장 가까운 데이터 샘플과의 거리를 의미하며, 마진이 큰 초월평면을 찾는 이유는 기계학습법에 의해 두 그룹을 구분 짓는 분류기, 즉 초월평면을 계산한 후에 새로운 데이터인 테스트데이터로 분류할 때, 보다 더 정확한 예측을 할 수 있다는 가정 때문이다. 초월평면과 가장 가까운 데이터 샘플들을 support vector라고 하는데, SVM은 이 support vector들만을 이용하여 분류기를 모델링한다.

일반적으로 데이터가 선형분리 불가능하다고 판단하여, 학습시의 에러를 어느 정도 허용하는 soft margin 방법을 사용하는데 다음과 같은 최적화 문제로 해를 구한다.

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to } y_i(W\cdot X_i + b) \geq 1-\xi_i, \quad i=1,\ldots,n \\ &\qquad\qquad \xi_i \geq 0, \qquad i=1,\ldots,n \end{aligned} \tag{2}$$

Hard margin 방법에 비하여 C 값을 포함한 에러 부분이 추가된 것을 확인 할 수 있고, 위 식을 곧바로 이용하여 최적화를 한다면, 마진의 크기뿐만 아니라 에러 부분도 동시에 최소화 하도록 하기 때문에 계산상으로 매우 복잡해질 수 있다. 그러므로 dual problem으로 변경하면 다음과 같은 방법으로 쉽게 계산할 수 있다.

$$\begin{aligned} &\text{minimize } \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} y_i\, y_j\, \alpha_i\, \alpha_j K(x_i, x_j) \\ &\text{subject to } 0 \leq \alpha_i \leq C, \quad i=1,\ldots,n \\ &\qquad\qquad \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{3}$$

여기서 α는 Lagrange multiplier로써, 모든 학습데이터 마다 하나씩의 양수 값을 갖게 되는데, support vector 외에 다른 데이터는 모두 0의 값을 갖게 된다. 또한 K는 커널이라 불리며, 연구 목적과 데이터 분포에 따라 선형 커널뿐만 아니라 radial basis function(RBF)과 같은 여러 가지 비선형 커널도 많이 사용된다. 그리하여 최종적으로 다음 식과 같이 분류기의 식을 구할 수 있게 되는데, 일반적으로 테스트데이터를 이 식에 적용하였을 때, 양수값이 나오면 양성, 음수값이 나오면 음성으로 판별한다.

$$F(x) = \text{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K(x, x_i) + b\right) \tag{4}$$

그러나 본 연구에서와 같이 양성데이터가 음성데이터에 비해 상대적으로 매우 적을 때에는 위와 같은 방법을 바로 적용하면 좋은 성능을 보장하기 어렵다. 그 이유는, 생성된 초월평면은 상대적으로 개수가 적은 양성데이터들이 모여 있는 방향으로 과도하게 편향되어 위치하기 때문에, minority 클래스인 양성데이터가 테스트데이터로 적용되었을 때 음성으로 판별할 가능성이 매우 높다. 따라서 양성데이터와 음성데이터에 대한 에러의 cost(즉, penalty)를 각각 다르게 두고 학습하는 cost-sensitive 학습법을 사용하여 다음과 같이 최적화한다[16].

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|w\|^2 + C^+\sum_{\{i|y_i=+1\}}^{n}\xi_i + C^-\sum_{\{i|y_i=+1\}}^{n}\xi_i \\ &\text{subject to } y_i(W\cdot X_i + b) \geq 1-\xi_i, \quad i=1,\ldots,n \\ &\qquad\qquad \xi_i \geq 0, \qquad i=1,\ldots,n \end{aligned} \tag{5}$$

마찬가지로, dual problem으로 변경하면 다음과 같다.

$$\begin{aligned} &\text{minimize } \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} y_i\, y_j\, \alpha_i\, \alpha_j \langle x_i, x_j\rangle \\ &\text{subject to } 0 \leq \alpha_i \leq C^+, \quad \text{for } y_i = +1 \\ &\qquad\qquad 0 \leq \alpha_i \leq C^-, \quad \text{for } y_i = -1 \\ &\qquad\qquad \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{6}$$

식 (3)과 비교하였을 때, 값의 상위 한계점이 각각의 클래스마다 다르게 설정되는 것을 확인할 수 있다. 이러한 방법으로 학습하면 두 그룹을 분류하는 초월평면이 상대적으로 개수가 많은 majority 클래스의 학습데이터가 몰려있는 쪽으로 다소 이동하게 되어, minority 클래스의 테스트데이터를 더 정확하게 계산할 수 있는 확률이 높다.

E. 실험절차 및 평가방법

총 292개의 데이터를 이용하여 각 특징점 선택 방법의 성능을 비교하고, 전통적인 SVM 학습법과 cost-sensitive 학습법을 비교하기 위하여 receiver operating characteristics (ROC) 곡선의 곡선아래면적(area under curve, AUC)을 기본적으로 비교하였으며, 예측정확률(accuracy), 평균오차율(balanced error rate, BER), 민감도(sensitivity) 및 특이도(specificity)도 같이 비교하였다[17]. SVM의 학습 시에는 선형 커널과 비선형 커널인 RBF 커널을 사용하였으며, 초월 평면의 함수를 적용하여 테스트데이터를 판단할 때 sigmoid 함수를 이용하여 0에서 1사이의 확률 값을 이용하여 계산하였다[18]. 즉, 확률 값이 0.5 이상이면 양성데이터로 판단하고, 그 이하이면 음성데이터로 판단하였다.

표 4는 실제 진단과 예측된 결과를 비교할 때 사용되는 결정 매트릭스를 나타내고 있다. 여기서 계산할 수 있는 TP, FP, TN, FN을 이용하여 예측정확도, 평균오차율, 민감도 및 특이도를 다음과

표 4. 실제 진단과 예측된 결과를 판단하는 일반적인 결정 매트릭스
Table 4. A general decision matrix that contains information on actual diagnosis and prediction output.

Actual diagnosis	Predicted result	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

같이 계산할 수 있다.

$$\text{Accuracy}: \frac{TP+TN}{TP+FP+TN+FN} \tag{7}$$

$$\text{BER}: \frac{1}{2}\left(\frac{FN}{TP+FN}+\frac{FP}{FP+TN}\right) \tag{8}$$

$$\text{Sensitivity}: \frac{TP}{TP+FN} \tag{9}$$

$$\text{Specificity}: \frac{TN}{TN+FP} \tag{10}$$

본 연구에 사용된 데이터 개수가 비교적 많지 않기 때문에, leave one out cross validation (LOOCV) 방법으로 평가하였다. LOOCV 방법은 전체 데이터에서 하나만을 제외한 나머지 데이터를 이용하여 학습한 후, 제외된 하나의 데이터로 테스트하는 방법인데, 이러한 과정을 전체 데이터 개수만큼 반복한다. 이는 전체 데이터를 임의로 학습 데이터와 테스트 데이터로 구분하여 평가하는 hold out 방법에서 발생할 수 있는 바이어스(bias)를 최소화 할 수 있는 장점이 있는 반면, 대단위 데이터에서는 많은 계산 시간이 소요된다는 단점이 있다.

III. 결 과

A. 예비실험

먼저, 특징점 선택 방법을 적용하지 않고 184개의 모든 특징점을 이용하여 전통적인 SVM 학습법으로 실험한 결과를 표 5에 나타내었다. 예비실험에서는 SVM을 학습할 때 양성 데이터와 음성 데이터에 동일한 cost를 적용하였기 때문에, 표 5에서 나타난 양성 데이터에 대한 cost값인 C+가 1로 설정되어 있다.

실험결과 AUC값에 있어서 선형 커널이 RBF 커널보다 조금 더 나은 성능을 보였으나 모두 0.7 이하의 성능을 보여 좋은 결과라고 할 수 없다. 예측정확률은 두 커널 모두 0.887이라는 높은 결과 값을 나타내었으나, 평균오차율이 0.5이고 민감도가 0이었다. 이는 테스트데이터를 판별할 때 확률 값의 문턱치를 0.5로 설정하였기 때문에, 모든 테스트데이터를 음성으로 판별했기 때문이다.

표 5. 특징점 선택 방법과 cost-sensitive 학습법을 이용하지 않은 SVM 학습법을 이용한 예비실험 결과
Table 5. Results of SVM kernel methods without cost-sensitive learning or feature selection method

Parameter	SVM linear	SVM RBF
특징점 개수	184	184
C+[1]	1	1
AUC	0.675	0.644
예측정확률	0.887	0.887
평균오차율	0.5	0.5
민감도	0	0
특이도	1	1

[1] C+: 양성데이터에 대한 cost
* 예측정확률, 평균오차율, 민감도 및 특이도는 SVM의 최종 판단 시에 확률의 문턱치를 0.5로 하여 측정한 값임

B. ReliefF

ReliefF 방법에 의한 결과를 그림 1에 나타내었는데, 가장 순위가 낮은 특징점들을 하나씩 제거해 나가면서 AUC값을 계산한 결과이므로, 그래프를 오른쪽에서 왼쪽으로 관찰하는 것이 이해하기 쉽다. 선형 커널과 RBF 커널을 비교하였고 전통적인 학습법과 cost-sensitive 학습법을 비교하여 총 4가지의 방법을 비교하였는데, 가장 좋은 성능을 발휘하는 경우인 cost-sensitive 학습법을 사용한 선형 커널 방법과 RBF 커널 방법에서 특징점 개수가 약 10개 근처일 때 예비실험의 결과보다는 좋은 성능을 나타내었지만, AUC값이 0.8을 넘지 않았다.

C. Sensitivity Analysis

그림 2에 sensitivity analysis에 의한 결과를 나타내었으며, ReliefF 방법과 마찬가지로 각각 4가지의 방법을 비교하였다.

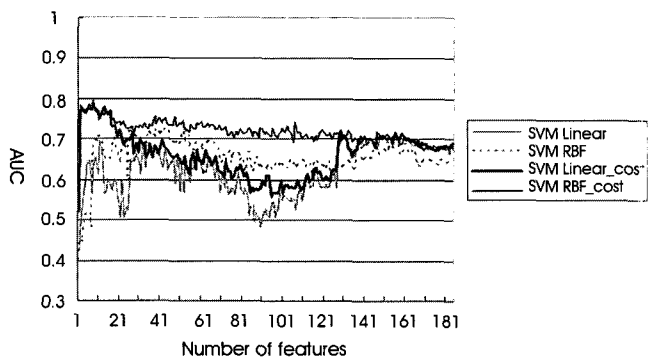

그림 1. ReliefF 방법을 사용하여 특징점을 하나씩 제거하였을 때, 커브아래면적(AUC)의 변화. 가 포함된 SVM 분류 모델은 cost-sensitive 학습법을 이용한 경우를 나타낸다.
Fig. 1. Performance variation of each classifier using the ReliefF method. The SVM classifiers that include 'cost' in the legend show the best AUCs for cost-sensitive learning.

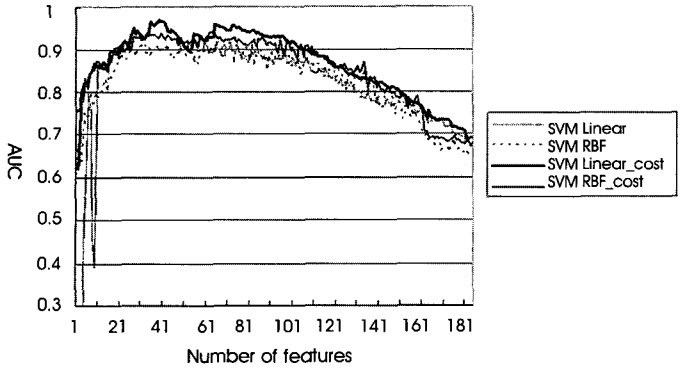

그림 2. Sensitivity analysis 방법을 사용하여 특징점을 하나씩 제거하였을 때, 커브아래면적(AUC)의 변화가 포함된 SVM 분류 모델은 cost-sensitive 학습법을 이용한 경우를 나타낸다.

Fig. 2. Performance variation of each classifier using the sensitivity analysis method. The SVM classifiers that include 'cost' in the legend show the best AUCs for cost-sensitive learning.

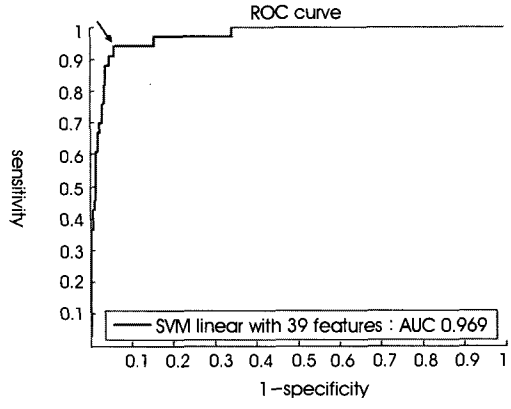

그림 3. 가장 좋은 성능을 나타내는 분류기의 ROC 커브. 화살표가 위치한 곳에서 민감도 0.97, 특이도 0.822를 각각 나타낸다.

Fig. 3. Receiver operation characteristic (ROC) curve of the best classifier. The arrow indicates the optimum point of the classifier, giving a sensitivity of 0.97 and a specificity of 0.822.

ReliefF 방법과는 다르게, 모든 경우에 있어서 특징점의 개수가 특정 지점까지 줄어들 때까지 성능이 계속 향상되다가 그 이후에 다시 감소하는 경향을 공통적으로 나타내었다. 전반적으로 선형 커널과 cost-sensitive 학습법을 포함한 방법(SVM Linear')이 가장 좋은 성능을 나타내었다.

표 5에서는 각 4가지 방법 중 가장 좋은 경우의 결과를 비교하였는데, 선형 커널과 39개의 특징점을 선택하였을 때 0.969의 AUC 값을 나타내었다. 이 때, 가장 좋은 성능을 나타내었던 두 개의 분류기(SVM linear와 SVM linear')는 모두 양성데이터에 대한 cost가 1이었다. 이것은 cost-sensitive 학습법을 이용한 결과보다는 모든 클래스의 데이터에 동일한 cost를 적용하는 전통적인 학습법을 적용하였을 때가 가장 좋은 결과를 나타내었다. 또한, 예비실험에서와 마찬가지로 테스트데이터의 판별 시에 확률 값의 문턱치를 0.5로 설정하였을 때, 예측정확률은 0.9 이상이지만 평균오차율이 높고, 특이도가 높은 반면민감도가 매우 낮았다.

표 6. 특징점 선택 방법과 cost-sensitive 학습법을 이용하지 않은 SVM 학습법을 이용한 예비실험 결과

Table 6. Results of SVM kernel methods without cost-sensitive learning or feature selection method

Parameter	SVM linear	SVM RBF	SVM linear[1]	SVM RBF[1]
특징점 개수	39	75	39	63
C+[3]	1	1	1	7
AUC	0.969	0.918	0.969	0.943
예측정확률	0.921	0.894	0.921	0.897
평균오차율	0.322	0.430	0.322	0.428
민감도	0.364	0.152	0.364	0.152
특이도	0.992	0.988	0.992	0.992

[1] : Cost-sensitive 학습법을 포함하여 적용한 분류 모델 중에 가장 좋은 성능을 나타내는 모델

[2] C+: 양성데이터에 대한 cost

* 예측정확률, 평균오차율, 민감도 및 특이도는 SVM의 최종 판단 시에 확률의 문턱치를 0.5로 하여 측정한 값임

D. ROC 커브 분석

그림 3에 선형 커널과 sensitivity analysis 방법에 의한 예측 모델 중 39개의 특징점을 이용한 가장 좋은 성능을 나타내는 분류기의 ROC 커브를 나타내었다. 표 5에서 나타낸 바와 같이 커브아래면적인 AUC값이 0.969임을 확인할 수 있다. 실선으로 표시된 화살표에 해당하는 지점은 테스트데이터에 대한 확률 값의 문턱치가 0.13을 나타내며 이 때의 민감도는 0.97, 특이도는 0.822, 예측정확도는 0.839, 평균오차율은 0.104를 나타내었다. 이는 확률 값의 문턱치를 0.5로 설정했던 표 5의 결과에 비하여 예측정확도와 특이도가 소폭 감소하였으나, 민감도와 평균오차율이 매우 향상되었음을 알 수 있다.

V. 토의 및 결론

당뇨병성 신장합병증을 예측하기 위하여 4321명의 당뇨환자로부터 10년간의 임상데이터를 전산화한 데이터를 구축하였으나, 매우 기계 학습법을 곧바로 이용하기에 그 특징이 매우 불규칙적이었다. 그래서 전처리를 통하여 총 184개의 특징점을 추출하였으며, 이로 인해 실제 연구에 사용된 데이터는 292개로 감소되었다. 그러나 특정 병원에서 특정 기간에 내원한 모든 환자들 중에서 184개의 특징점을 모두 추출할 수 있는 데이터를 동일한 기준으로 모두 연구에 포함하였을 뿐만 아니라, 나이, 성별, 당뇨병 발병 시기, 당뇨병 이환 기간에 있어서 두 그룹 간에 통계적으로 유의미한 차이가 없었기 때문에, 연구 데이터 감소로 인한 바이어스는 없었

다고 판단된다.

예비실험 결과에서 나타났듯이, 전처리를 통하여 추출된 184개의 특징점을 모두 이용하여 SVM을 적용한 경우에는 매우 낮은 성능을 보였다. 이는 특징점의 개수가 데이터 개수에 비해 매우 많았고 추출된 특징점들 중에서 예측에 불필요한 특징점이 포함됐기 때문이며, 이를 해결하기 위하여 본 연구에서 두 가지의 특징점 선택 방법을 적용하였다. 그 중에서 wrapper 방법의 일종인 sensitivity analysis 방법을 이용한 경우가 ReliefF 방법을 이용한 경우보다 높은 성능을 나타내었다. 이것은 filter 방법의 일종인 ReliefF는 분류 학습기와 직접적인 관련이 없이 데이터만으로 특징점의 순위를 판단하기 때문에 계산상의 이점이 있지만, wrapper나 embedded 방법은 더 많은 계산량을 요구한다는 단점에도 불구하고 분류 학습기와 밀접하게 관련하여 특징점을 선택하므로 더 좋은 성능을 나타낼 가능성이 높다는 것을 나타낸다.

연구에 사용된 데이터는 양성데이터의 개수가 음성데이터보다 상대적으로 작았기 때문에 cost-sensitive 학습법을 사용하였으나, 실험결과 양쪽 클래스의 데이터에 같은 cost를 적용하는 전통적인 학습법을 사용한 경우가 가장 좋은 성능을 나타내었다. Cost-sensitive 학습법은 특징점 공간(feature space)에서 초월평면의 위치를 상대적으로 개수가 많은 클래스의 데이터 쪽으로 평행 이동시킨 것으로 볼 수 있으므로, 테스트데이터를 분류 모델에 적용하여 계산된 결과 값으로부터 최종 판단할 때의 문턱치를 변화시키는 방법과 비슷하다. 따라서 이러한 방법은 본 연구에서 사용된 성능비교 방법인 ROC 곡선의 곡선아래면적에는 영향을 미치지 않기 때문에 위와 같은 결과가 도출되었다고 판단된다. 그러므로 곡선아래면적은 각 클래스데이터에 대한 cost보다는 다른 파라미터를 조정하였을 때 변화가 더 많이 일어난다고 볼 수 있다. ROC 곡선을 분석한 결과 테스트데이터에 대한 확률 값의 문턱치를 0.5에서 0.13으로 낮추었을 때가 가장 최적화된 결과를 도출함을 알 수 있었다. 이러한 결과는 양성데이터개수와 음성데이터 개수의 불균형 때문에 발생할 수 있는 것이라 생각된다.

본 연구에서 고려될 수 있는 가장 큰 문제점은 당뇨환자들이 사용했을 가능성이 매우 높은 인슐린이나 혈압강하와 관련된 약물정보를 포함하지 않았다는 것이다. 이러한 정보를 이용하기 위해서는 향후 텍스트 마이닝이나 온톨로지 등 여러 가지 다른 데이터 마이닝 기법을 동시에 적용해야 할 필요성이 있다.

본 연구에서 이용한 기계 학습방법으로 당뇨환자로부터 당뇨병성 합병증을 예측하여 ROC 곡선의 곡선아래면적이 0.969라는 높은 예측 성능을 나타내었다. 그러나 몇 가지의 특징점 선택 방법을 통하여 중요한 39개의 특징점들만으로 이러한 예측 성능을 나타내었지만, 실제 임상 의사들이 본 연구를 통해 얻을 수 있는 정보는 단지 특정 환자가 당뇨병성 합병증이 발병할 확률이 어느 정도 되는지에 대한 수치이다. 기계 학습법을 비롯한 여러 가지 데이터 마이닝 방법을 의학에 적용시킬 때 그 최종 목적이 임상 의사를 대체하고자 하는 것이 아닌 임상 의사를 보조하고자 하는 것임을 상기할 때, 이러한 정보는 충분치 못하다고 할 수 있다. 따라서 향후에는 더 나은 특징점 선택 방법과 함께, 시각화 방법 등을 통하여 위험 인자에 대한 정보를 임상의사에게 직접 제공할 필요가 있다.

본 연구의 가장 중요한 의의는 당뇨병성 합병증을 예측하기 위하여 기계 학습법과 같은 공학적 접근법을 이용한 최초의 사례이다. 현재에도 많은 병원들이 전자의무기록과 관련된 시스템을 도입하는 추세에 있고 이러한 수치는 앞으로 전 세계적으로 기하급수적으로 늘어날 것이며, 또한 본 연구에서 사용된 임상데이터와 비슷한 대량의 전자임상데이터를 매우 쉽게 획득할 수 있을 것으로 생각된다. 그러므로 향후 이와 관련된 다양한 연구를 진행할 수 있을 것으로 판단된다.

참고문헌

[1] K. G. Alberti and P. Z. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation," *Diabet Med*, vol. 17, pp. 539-553, 1998.

[2] H. S. Jo, J. H. Sung, J. S. Choi, M. S. Hwang, H. J. Jeong, and S. C. Bae, "Quality control of diagnostic coding in the Korean Burden of Disease Project," presented at Int society for quality in health care's Int Conf, Amsterdam, Netherlands, October 2004.

[3] C. E. Mogensen, J. Vigstrup, and N. Ehlers, "Microalbuminuria predicts proliferative diabetic retinopathy," *Lancet*, vol. 1, pp. 1512-1513, 1985.

[4] P. Rossing, P. Hougaard, K. Borch-Johnsen, and H. Parving, "Predictors of mortality in insulin dependent diabetes: 10 year observational follow up study," *BMJ*, vol. 313, pp. 779-784, 1996.

[5] T. Furuta, T. Saito, T. Ootaka, J. Soma, K. Obara, K. Abe, and K. Yshinaga, "The role of macrophages in diabetic glomerulosclerosis," *American Journal of Kidney Diseases*, vol. 21, pp. 480-485, 1993.

[6] G. Sterner, J. Carlson, and G. Ekberg, "Raised platelet levels in diabetes mellitus complicated with nephropathy," *Journal of Internal Medicine*, vol. 244, pp. 437-441, 1998.

[7] T. Onuma, T. Kikuch, M. Tsutsui, S. Shimura, J. Matsui, A. Boku, and K. Takebe, "High incidence of diabetic nephropathy in non-insulin-dependent diabetic patients with heterozygous familial hypercholesterolemia," *Current therapeutic research*, vol. 55, pp. 532-536, 1994.

[8] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artf Intell Med*, vol. 26, pp. 1-24, 2002.

[9] C. Elkan, "The foundations of cost-sensitive learning," in *Proc.17th Int Joint Conf Artif Intell*, Seattle, WA, August 2001.

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157-1182, 2003.

[11] I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Proc. ECML*'94, Catania, Italy, April 1994.

[12] M. Stevensen, R. Winter, and B. Widrow, "Sensitivity of feed

forward neural networks to weight errors," *IEEE Trans. Neural Networks*, vol. 1, pp. 71-80, 1990.

[13] V. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer, 1995.

[14] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

[15] A. B. Magil and A. H. Cohen, "Monocytes and focal glomerulosclerosis," *Laboratory Investigation*, vol. 61, pp. 404-409, 1989.

[16] K. Veropoulos, N. Cristianini, and C. Campbell, "Controlling the sensitivity of support vector machines," in *Proc.the Int Joint Conf Artif Intell*, Stockholm, Sweden, August 1999.

[17] H. S. Choi, Y. H. Cho, B. H. Cho, W. K. Moon, J. G. Im, I. Y. Kim, and S. I. Kim, "A study on the multi-view based computer aided diagnosis in digital mammography," *Journal of Biomedical Engineering Research*, vol. 28, pp. 162-168, 2007.

[18] J. C. Platt, *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, in Advances* in Large Margin Classifiers: MIT Press, 1999.