ICNSE-915
Cluster-based Classification of Diabetic Nephropathy among Type 2 Diabetic Patients

Guan-Mau Huang
Department of Computer Science and Engineering, Yuan Ze University, Chung-Li,Taiwan
E-mail address: s1016017@mail.yzu.edu.tw,

Yu-Chun Lee
Department of Computer Science and Engineering, Yuan Ze University, Chung-Li,Taiwan
E-mail address:s1026031@mail.yzu.edu.tw

Julia Tzu-Ya Weng[a,*]
Department of Computer Science and Engineering, Yuan Ze University, Chung-Li,Taiwan
E-mail address:julweng@saturn.yzu.edu.tw

Yi-Cheng Chen
Department of Computer Science and Information Engineering, Tamkang University, New Taipei city, Taiwan
E-mail address: ycchen@mail.tku.edu.tw

Lawrence Shih-Hsin Wu
[c]Institute of Medical Sciences, Tzu Chi University, Hualien City, Taiwan
E-mail address: lshwu@hotmail.com

Abstract

The prevalence of type 2 diabetes is increasing at an alarming rate. Various complications are associated with type 2 diabetes, with diabetic nephropathy being the leading cause of renal failure among diabetics. Often, when patients are diagnosed with diabetic nephropathy, their renal functions have already been significantly damaged, speeding up the progression towards end stage renal disease. Therefore, a risk prediction tool may be beneficial for the implementation of early treatment and prevention. In the present study, we propose to develop a prediction model integrating clustering and classification approaches for the identification of diabetic nephropathy among type 2 diabetes patients. Clinical and genotyping data are obtained from 345 type 2 diabetic patients(160 with non-diabetic nephropathy and 185 with diabetic nephropathy). The performance of using clinical features alone for cluster-based classification is compared with that of utilizing a combination of clinical and genetic attributes. We find that the inclusion of genetic features yield better prediction results. Further refinement of the proposed approach has the potential to facilitate the accurate identification of diabetic nephropathy and the development of better treatment in a clinical setting.

Keyword: type 2 diabetes, diabetic nephropathy, classification, clustering, genetic

1. Introduction

Diabetes is a metabolic disorder contributed by multiple factors, including diet, lifestyle, and genes. Despite the advancement in health care, the prevalence of diabetes is still on the rise, with more than 150 million people worldwide being affected with this debilitating disease [1].

Diabetes can result invarious complications, damaging the heart, blood vessels, eyes, kidneys, and nerves. Currently, more than 1.5 million people are affected with T2D in Taiwan.

Previously, we conducted a candidate gene analysis on 345 T2D patients, analyzing the association of 20 candidate genes with T2D, as well as the related complications such as obesity and diabetic nephropathy (DN).Through the generalized multi-dimensional reduction approach [2], we identified various gene-gene interactions that may represent genetic susceptibilities to T2D, obesity, and DN[3]. Our findings suggest that T2D and DN may be attributable to multiple factors that include the interactions between genes and the environment.

As the major contributor of end-stage renal disease (ESRD), DN is one of the most fatal complications of diabetes [4]. Compared to non-diabetics, the likelihood of dying from renal disease is 17 times greater for diabetics [5]. DN affects about 30% of the people with type 1 diabetes, and 25-40% of the people with T2D [6, 7]. Unfortunately, the exact mechanisms underlying the progression from DN to ESRD is not yet fully understood.

Generally, DN symptoms are not obvious. In type 1 diabetic patients, DN develops over an initial 10 to 15 years of diabetes; whereas in type 2 diabetic patients, the onset is less clearly defined [8]. When clinical indices for renal functions (e.g. urinary protein level) become abnormal, the kidneys have already been significantly damaged and prevention against ESRD may already be too late at this stage [9]. Thus, it would be beneficial to develop a prediction model utilizing more indirect clinical measures, such as gender, history of diabetes, body mass index (BMI), etc. Moreover, since existing studies suggest that genes play a role in diabetes and DN risk [3, 10], integrating genetic information in the prediction of DN susceptibility may lead to more accurate identification.

In the present study, we propose to employ a cluster-based classification approach to identify DN among T2D patients. Compared to using clinical features alone for the classification, we find that the integration of clinical and genetic features yields a higher performance in distinguishing DN from non-DN diabetic patients.

2. Materials and Methods

2.1 T2D data

The T2D data consists of 345 Taiwanese patients who were recruited form the Tri-Service General Hospital in Taipei in 2002[3, 11].The case group comprises 185 T2D patients with DN, and the control group comprised 160 T2D without DN. Table 1 provides an overview of the demographic and clinical data of these participants.

2.2 Clinical and Genetic Feature selection

In our previous studies [3, 11], statistical analyses have already identified clinical and genetic attributes that show significant differences between the control and case groups. These statistically significant features are selected as the clinical and genetic features for the present study and are shown in Tables 1 and 2, respectively.

Table 1. Demographic and clinical characteristics of the study subjects

Parameter	DN(case) n = 185	T2D without DN(control) n = 160	P-value
Age	58±9.9	56.6±8.9	0.14
Gender(M/F)	102/83	68/92	0.0327
History(year)	15.2±7.4	13±6.6	0.001
BMI	25.7±3.8	24.8±4	0.02
Fasting plasma glucose(FPG; mg/dl)	177.2±66.6	166.4±43.6	0.06
Hemoglobin A1c (HbA1c; %)	8.9±2	8.6±1.4	0.12
Serum total cholesterol(STC; mg/dl)	205.9±50	197±34.9	0.063
LDL (low-density lipoprotein; mg/dl)	111.5±33.6	112.6±27.9	0.728
HDL (high-density lipoprotein; mg/dl)	38.9±12.6	44.1±11.4	<<0.0001
Serum triglycerol (mg/dl)	223.8±171.3	145.2±95.3	<<0.0001
SGOT (serum glutamic-oxaloacetic transaminase; mg/dl)	22.9±16.4	21.9±12.1	0.503
SGPT (serum glutamic-pyruvic transaminase; mg/dl)	21±20.3	23.8±16.7	0.21

Table 2.Genotype distributions of the significant candidate genes in T2D patients with and without DN.

Gene	SNP	Alleles	Genotype	T2D with DN	T2D without DN
PPAR	rs1801282	C/G	CC/CG/GG	170/13/2	142/12/6
UCP1	-3826 A > G	A/G	AA/AG/GG	46/84/55	31/83/46
UCP3	rs1800849	A/G	AA/AG/GG	15/72/98	15/60/85
PLIN	rs894160	A/G	AA/AG/GG	42/78/65	29/78/53
ADIPOQ	rs266729	C/G	CC/CG/GG	118/59/8	86/62/12

2.3 System Flow

The performances of classifying DN based on clinical and genetic features separately are compared with that of cluster-based classification integrating clinical and genetic features. There are too many factors that can lead to diabetes and its associated complications.We believe that the cluster-based classification approach [12] approach should increase the accuracy of DN identification. Experiments are conducted in WEKA 3.6.5 tool[13]. Weka (Waikato Environment for Knowledge Analysis) is a JAVA based platform for data mining and data analysis. We used clustering package inside. We used the K means clustering package for clustering.Clustering is the process of grouping similar elements andthe K means algorithm is a method of vector quantization originally derived for signal processing and later evolved into a popular clustering approach in data mining[14].Then, we used the C4.5 decision tree in WEKA. This is a well-known classification technique in decision tree

induction [15]. The parameter setting are c=0.5 and M=2 for the confidence factor and minimum number of objects, respectively.

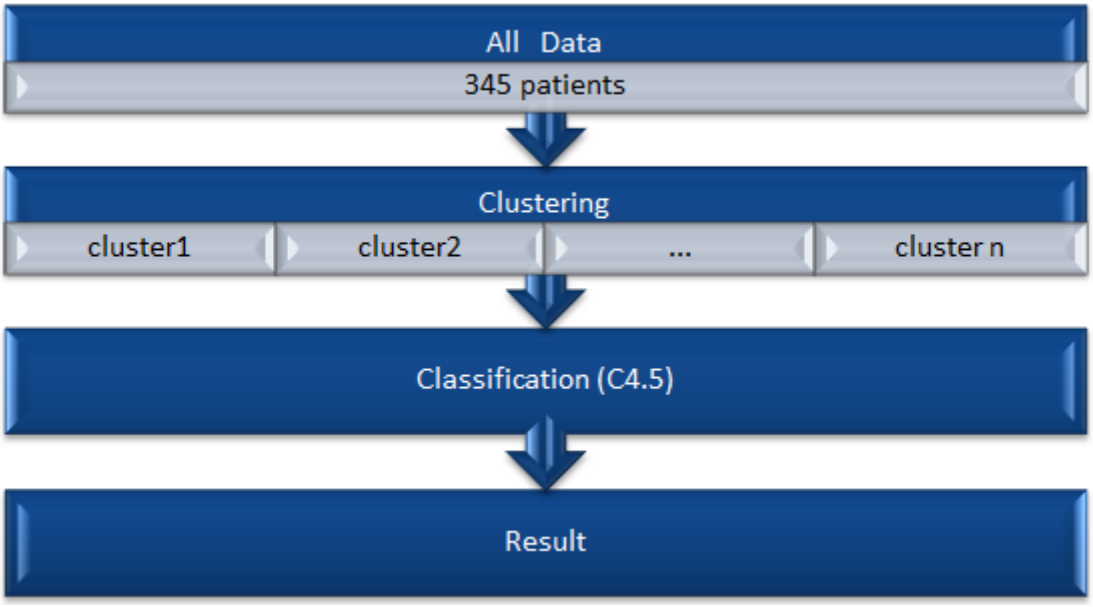

Fig. 1 System flow of the proposed cluster-based classification approach for the identification of DN among T2D patients.

Finally, we use a 10-fold cross validation of the data to evaluate the performance of the proposed approach in accuracy, sensitivity, specificity.For the measurement of accuracy, sensitivity and specificity, the following mathematical model is used, where TP represents the correctly identifiedtrue positive; FP indicates the incorrectly identified false positive; TN is the correctly rejected true negative; FN stands for the incorrectly rejected falsenegative.Table 3 describes each performance measures in the context of DN classification.

$$Accuracy = \frac{TP + TN}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

$$Sensitivity = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

$$Specificity = \frac{TN}{\mathrm{TN} + \mathrm{FP}}$$

Table 3.True Positive, True Negative, False Positive and False Negative confusion matrix for DN classification.

Predict	Classified as DN	Classified as not DN
T2D with DN	TP	FN
T2D without DN	FP	TN

3. Results and Discussions

The performances of using clinical features or genetic features alone for DN classification are shown in Tables 4 and 5, respectively. The classification accuracy measures seem to be low. Perhaps these approaches ignore the possible interactions among genes and clinical traits that underlie the disease mechanisms of DN. Thus, we decide to employ a cluster-based

classification approach, employing a combination of clinical and genetic attributes. Performance of the proposed approach is shown in Table 6. The proposed approach outperforms the classification methods based on clinical or genetic features separately.

Table 4.Accuracy measures for DN classification based on clinical features

Parameter	Accuracy
LDL	53.62%
HDL	56.36%
Serum triglycerol	62.61%

Table 5.Accuracy measures for DN classification based on genetic features

Parameter	Accuracy
PPAR	51.62%
UCP1	53.62%
UCP3	53.47%

Our present finding supports the previous study that the mechanisms underlying DN may involve the interactions between genes and the environment. Each genetic or clinical attributes may contribute marginal effects to DN. Therefore, when predicting DN among T2D patients, it may be important to consider both genetic and clinical factors. Using a cluster-based classification approach, we also provide support for the notion that complex diseases like T2D and its associated DN complication are genetically heterogeneous. The same genetic features that may help predict DN susceptibility for one individual may not apply for everyone. However, grouping individuals based on clinical similarities and classify individuals with DN may represent a logical solution.

To our knowledge, as the disease mechanisms of DN are complex and the symptoms are not obvious, there are no accurate diagnostic methods for DN. We believe the proposed approach, with further refinement in parameter settings and testing in a bigger sample size, may have the potential to facilitate the early identification of individuals with DN susceptibility and therefore, early prevention or treatment for DN.

Table6.Cluster-based classification integrating clinical and genetic attributes

Cluster	Classification		10 fold cross validation
Female, BMI < 24	Serum triglycerol LDL	PLIN ADIPOQ	**Accuracy** = 84.48% **Sensitivity** = 61.11% **Specificity** = 95%
Female, BMI > 24	HDL	ADIPOQ	**Accuracy** = 64% **Sensitivity** = 87.69% **Specificity** = 34.61%
Male, BMI < 24	Serum triglycerol LDL HDL	UCP1 UCP3	**Accuracy** = 68.53% **Sensitivity** = 88.89% **Specificity** = 37.14%
Male, BMI > 24	Serum triglycerol	UCP1	**Accuracy** = 71.60% **Sensitivity** = 85.42% **Specificity** = 51.52%

4. Acknowledgments and Legal Responsibility

We thank Dr. S.L. Wu for providing detailed demographic, clinical, and genetic data for the T2D patients analyzed in the present study.

5. References

[1] X. Ren, Type 2 diabetes mellitus associated with increased risk for colorectal cancer: evidence from an international ecological study and population-based risk analysis in China,*Public Health*,2009,pp.540-544.

[2] G.B. Chen, Practical and theoretical considerations in study design for detecting gene-gene interactions using MDR and GMDR approaches,PLoS One,2011,e16981.

[3] L.S. Wu, Association and interaction analyses of genetic variants in ADIPOQ, ENPP1, GHSR, PPARgamma and TCF7L2 genes for diabetic nephropathy in a Taiwanese population with type 2 diabetes,Nephrol Dial Transplant,2009,pp.3360-3366.

[4] J.E. Lee, Risk of ESRD and all cause mortality in type 2 diabetes according to circulating levels of FGF-23 and TNFR1,PLoS One,2013,e58007.

[5] C.Y. Hong, K.S. Chia, Markers of diabetic nephropathy. Journal of diabetes and its complications, 1998, 12(1), pp.43-60.

[6] J.Y. Hsiao, The relationship between diabetic autonomic neuropathy and diabetic risk factors in a Taiwanese population,JInt Med Res,2011,pp.1155-1162.

[7] D.M. Good, Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease,Mol Cell Proteomics,2010,pp.2424-2437.

[8] A. Ntemka, F. Iliadis, N.A. Papanikolaou,D.Grekas, Network-centric Analysis of Genetic Predisposition in Diabetic Nephropathy. Hippokratia 2011, 15(3), pp.232-237.

[9] F. Sharifiaghdas, Evaluating percutaneous nephrolithotomy-induced kidney damage by measuring urinary concentrations of beta2-microglobulin,Urol J,2011,pp.277-282.

[10] P.M. Thorsby, Comparison of genetic risk in three candidate genes (TCF7L2, PPARG, KCNJ11) with traditional risk factors for type 2 diabetes in a population-based study--the HUNT study,Scand J Clin Lab Invest,2009,pp.282-287.

[11] E. Lin, Gene-gene interactions among genetic variants from obesity candidate genes for nonobese and obese populations in type 2 diabetes,Genet Test Mol Biomarkers,2009,pp.485-493.

[12] A.K. Banerjee, Classification and clustering analysis of pyruvate dehydrogenase enzyme based on their physicochemical properties,Bioinformation,2010,pp.456-462.
[13] F. Firouzi, A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software,Eur J Gastroenterol Hepatol,2007,pp.1075-1081.
[14] B. Chen, Using hybrid hierarchical K-means (HHK) clustering algorithm for protein sequence motif super-rule-tree (SRT) structure construction,Int J Data Min Bioinform,2010,pp.316-330.
[15] M.U. Khan, Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare,ConfProc IEEE Eng Med Biol Soc,2008,pp.5148-5151.