Poster D-1

Development of prediction system in TYPE 2 DIABETES using molecular epidemiological data

Authors:
Ju-young Lee (*nih,korea*)
Seung-Woo Shin (*nih,korea*)
Seol-hee Bae (*nih,korea*)
Keun-Joon Park (*nih,korea*)
In Ae Hur (*nih,korea*)
Jeoung Ho Cha (*nih,korea*)
Haesook Min (*nih,korea*)
Eun Ji Hwang (*nih,korea*)
Keun yong Park (*nih,korea*)
Hyun-Woo Han (*nih,korea*)
Hang-Suk Choi (*nih,korea*)

Short Abstract: Finding high risk factors related Type 2 Diabetes(T2DM), we tried to optimizing model ; SVM and decision tree. Population–based 1145 cases and 1122 controls are selected. Using epidemic and genetic data, they are analyzed their association related T2DM. We suggest the genetic information in molecular epidemiological data might be useful for the classification of high risk of T2DM.

Long Abstract:
It is widely hoped that the information in human genome will provide a means of elucidating the genetic function of complex disease. Type 2 Diabetes a complex genetic disease including gene-environment interaction. Various association studies between SNP(Single nucleotide polymorphisms) and Type 2 Diabets have been researched.The aim of this study is to find the relationship between SNP(Single nucleotide polymorphisms) and disease and find high risk factor. In addition to, we also tried to analyze the impact of this correlation on the T2DM (Type 2 Diabetes)
In order to find early warning system on Type 2 Diabetes disease, we compared the new model with existing models in terms of performance of error rate(i.e. misclassification rate). And then we compared the optimized model for Type 2 diabetes, which will provided the patients with solution for early detection and care.Comparing the performance of error rate(misclassification rate) between system we developed and others we would like to approach optimizing the model and classification rule using SVM(Support vector machine), Decision tree in Machine Learning algorithm.
For the next step, it is a way to make cross validation what the most useful method of a data set for the learning and validation is.
For the current study, two statistical and machine learning models, SVM(Support vector machine) and decision trees were applied to the integrated epidemiological and genetical data to classify normal controls and cases. We use SVM to classify patients(cases) from normals(controls), because its performance for classification is better than other machine learning algorithms. Therefore, we make support vectors using the SNP data but the SVM only can use numerical data as input form. Converting categorical to numerical data, we use

two basic ideas: One approach is to use frequency of each SNP allele type which consists of column. Another approach is to convert the SNP allele type into numerical data to make support vector. We compared the performance of two approaches later. It also assumes independence of each for the features(SNPs).

In the entire data set this is not allowed missing values imputation of molecular epidemiological data.

Using various approaches, we have developed a new system for prediction related Type 2 Diabetes with our population based data.

It focused on developing the new projection system which minimized misclassification rate for Type 2 diabetes.The majority of our pilot study was to assess classification rules for the cases(Type 2 diabetes) and controls (Type 2 diabetes).

This study was supported by community based cohort study Ansung(5018) and Ansan(5020), Korean Genome and Epidemiology Study (KoGES), from Korea Institute of Health, Korea.For data related Type 2 Diabetes-it is one of major chronic disease in Korea, they in community based cohort study Ansung and Ansan, Korean Genome and Epidemiology Study (KoGES), were selected age-matched population–based 1122 cases and 1145 controls. Using questionnaire, we surveyed Single nucleotide polymorphisms (SNP) of 15 genes, which are founded 85loci, and various items for clinical, insulin resistance, behavior and so on.

First, for their integrated data which are epidemical and genetical, we have performed chi-square test for independence among allele type and analyzed the model including their association related Type 2 Diabetes.

In 2 by 2 tables, we focused that their frequencies depend on SNPs. Hence, we have selected the significant variables for the optimizing model and summarized the associations in allele types.

Two types of data from the epidemiological and genetical were used in the analysis: categorical measurements of 85 SNPs and quantitative and categorical measurements of clinical. Several SNPs in INSR, IRS1, IRS2, CAP, CBL, RAPGEF, PRKCD, PRKCZ, PDK, AKT, PTPN, PIK3CA, RHOQ, TGFBI, PERC are significantly different distributed between cases and control.In our experiments, the accuracy of classification performance is 70 ~ 50%. The results of two approaches of converting the SNP allele type into numerical data are similar each other.

And the results of statistical methods and machine learning methods don’t show large differences.

Two types of data from the epidemiological and genetical were used in the analysis: categorical measurements of 85 SNPs and quantitative and categorical measurements of clinical.

We suggest the genetical information in molecular epidemiological data might be useful for the classification of individuals into groups of high or low risk of Type 2 Diabetes.