UNIVERSITY OF THE DISTRICT OF COLUMBIA

Agricultural Experiment Station (AES)

University of the District of Columbia

Community Outreach and Extension Services

September 2009

Developing Fuzzy-set-theory-based Data Mining Methodologies for Diabetes Data Analysis

Lily R. Liang, Ph.D.

Diabetes affects millions of people in the United States and is one of the leading causes of death. The overall aim of this interdisciplinary research is to develop a series of fuzzy-set-theory-based data mining approaches for finding genetic, environmental and behavioral factors associated with diabetes. In an early phase of this project, Professor Liang's research team developed an *X-test* family, a series of fuzzy-theory-based methodologies, to effectively measure the divergence of gene microarray data between diabetic and non-diabetic groups. Now this team is planning to further develop the *X*-test family to handle categorical and heterogeneous data which are collected clinically or through survey.

Much research has been conducted to investigate the genetic, environmental and behavioral causes of diabetes. Large amounts of data have been collected every year by Centers for Disease Control and Prevention, diabetes research centers and research scientists at various institutions. However, most of this collected data is currently analyzed with statistical methods, which do not handle issues of missing and noisy data and can only provide mathematical interpretation of the datasets without other forms of knowledge. Thus, there is a great need for exploring the diverse diabetes data with powerful classic computer science data mining techniques and to build new data mining tools for this purpose.

Some researchers have started applying data mining techniques of computer science, such as clustering and association rule mining, to analyze genetic diabetes data. However, other forms of data, such as clinical and survey data are still under-explored by computer science techniques. These data may not be numerical and may not be homogeneous; thus, posing larger challenges. Fuzzy-set-theory, as a well-known theory in the computer science field, was formalized by Professor Lofti Zadeh at the University of California in 1965 and has

Special points of interest:

- Diabetes affects millions of people in the United States and is one of the leading causes of death.
- The overall aim of this interdisciplinary research is to develop a series of fuzzy-set-theory-based data mining approaches for finding genetic, environmental and behavioral factors associated with diabetes.

(Continued from page 1)

been applied to many areas of sciences and engineering. In fuzzy-set-theory, variables are granulated into clusters with linguistic labels. The degree of belongingness of each variable is a degree between 0 (not-belong) and 1 (fully-belong). This theory models the way that humans perceive and quantifies the imprecision of such perception. In contrast to the conventional computation, which handles numerical values, fuzzy logic handles linguistic values, which makes it possible to compute with natural languages. We propose fuzzy-set-theory based approaches for diabetes data analysis because i) they can handle the qualitative variables; ii) they can handle noisy quantitative variables with granulation; and iii) they can reduce the impact of missing data by granulating the presented data.

ABOUT LILY R. LIANG, PH.D.

Dr. Lily R. Liang is an associate professor in the Department of Computer Science and Information Technology at the University of the District of Columbia. She received her doctorate degree in Computer Science and Engineering from the University of Nevada in 2004. Dr. Liang's research interests include bioinformatics, data mining, machine intelligence, fuzzy logic and digital image processing. Currently, she has several funded research projects and is collaborating with professionals in biology, health and nutrition to develop data mining techniques in these areas. Her most recent award is from the National Science Foundation for a project involving workforce development in information assurance. She has published a number of conference and journal papers and has made presentations at national and international conferences.

University of the District of Columbia
Agricultural Experiment Station
4200 Connecticut Avenue, N.W.
Intelsat Building, Suite 6L-21
Washington, DC 20008
(202) 274-7132 telephone
(202) 274-7119 fax