

**A COMPARATIVE STUDY ON VARIABLE SELECTION IN MICROARRAY CLASSIFICATION**

**SIRIKUL LAOSRIVICHIT 5537223 EGTI/M**

**M.Sc. (TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)**

**THESIS ADVISORY COMMITTEE: SOTARAT THAMMABOOSADEE, Ph.D.,  
SUPAPORNKIATTISIN, Ph.D., WARANYUWONGSEREE, Ph.D.**

**ABSTRACT**

This thesis proposes using the shrinkage method for logistic regression. The normal exponential gamma (NEG) distribution is a Bayesian-inspired method. This method uses the normal-exponential-gamma prior for coefficients of model. The typical problems of microarray data have been solved by the LASSO and the elastic net. The comparative study of the two famous methods, NEG distribution and double exponential (DE) distribution were measured by the number of selected variables, predictive accuracy, deviance and computational time. The paired sample t-test was used to analyze the mean difference between two values. The results showed that the NEG distribution was more efficient than the LASSO, the elastic net and the DE distribution.

**KEY WORDS: NORMAL EXPONENTIAL GAMMA DISTRIBUTION / LASSO /  
ELASTIC NET /DOUBLE EXPONENTIAL DISTRIBUTION/ DNA  
MICROARRAY**

74pages

การศึกษาเปรียบเทียบการคัดเลือกตัวแปรในการคัดแยกข้อมูลชนิดไมโครอะเรย์  
A COMPARATIVE STUDY ON VARIABLE SELECTION IN MICROARRAY  
CLASSIFICATION

สิริกฤต เหล่าศรีวิจิตร 5537223 EGTI/M

วท.ม. (เทคโนโลยีการจัดการระบบสารสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: โยพาศรัตตธรรมบุษดี, Ph.D., สุภาภรณ์ เกียรติสิน, Ph.D.,  
วรัญญู วงษ์เสรี, Ph.D.

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการลดขนาดสัมประสิทธิ์ของแบบจำลองการถดถอยโลจิสติก โดยอิงแนวคิดจากทฤษฎีของเบย์และใช้การแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติเป็นความน่าจะเป็นก่อนหน้าของสัมประสิทธิ์ของแบบจำลอง ในปัญหาการศึกษาดีเอ็นเอไมโครอะเรย์ การศึกษาเปรียบเทียบประสิทธิภาพวิธีที่นำเสนอกับวิธีที่นิยมใช้กันมากในปัจจุบัน คือวิธีแลชโซ วิธีช่ายยัดหยุ่น และวิธีที่มีการอิงแนวคิดจากทฤษฎีของเบย์และใช้การแจกแจงแบบเลขชี้กำลังคู่ ในการศึกษาประสิทธิภาพจะใช้จำนวนยีนที่ถูกเลือก ค่าความถูกต้องของการทำนาย ค่าความผิดพลาดของการทำนาย และเวลาที่ใช้ในการสร้างแบบจำลองการจำแนก และใช้หลักการทางสถิตินำมาวิเคราะห์หาความแตกต่างของค่าเฉลี่ยสองค่า (paired t-test) จากผลการศึกษาพบว่าการใช้การแจกแจงแบบแกมมาเลขชี้กำลังแบบปรกติมีประสิทธิภาพในการทำนายและความสามารถในการคัดเลือกยีนสูงกว่าทั้งวิธีแลชโซ วิธีช่ายยัดหยุ่น และการใช้การแจกแจงแบบเลขชี้กำลังคู่

74หน้า