การคัดแยกประเภทของมะเร็งเม็ดเลือดขาว โดยใช้วิธีการจัดอันดับ ร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน

Classification of Leukemia Data Using Ranking and Support Vector Machine

ภัทราวุฒิ แสงศิริ (Patharawut Saengsiri)* ดร.ศจีมาจ ณ วิเชียร (Dr.Sageemas Na Wichian)**
ดร.พยุง มีสัจ (Dr.Phayung Meesad)***

บทคัดย่อ

การค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนก เป็นปัญหาที่สำคัญสำหรับงานวิจัยทางด้านชีววิทยา เนื่องจากมีจำนวนยีนเพิ่มขึ้นเป็นจำนวนมาก ดังนั้นเทคนิคการลดมิติของข้อมูล จึงเป็นประโยชน์ในการช่วยค้นหากลุ่มย่อยของยีน สาเหตุเนื่องจากเมื่อข้อมูลมีจำนวนมิติหรือตัวแปรมาก ทำให้ข้อมูลเกิดการกระจาย (data sparse) และทำให้เกิดปัญหามิติข้อมูล (Curse of Dimensionality) งานวิจัยนี้จะนำเอาข้อมูลยีนของโรคมะเร็งเม็ดเลือดขาวแบบเฉียบพลัน (Acute Leukemia) ซึ่งมีจำนวนมิติของข้อมูล 7,129 มิติ แบ่งออกเป็น 2 กลุ่ม คือ ALL และ AML มาทำการทดลองและเพื่อเปรียบเทียบประสิทธิภาพของการลดมิติข้อมูล ระหว่างวิธี Correlation Based Feature Selection, Gain Ratio และ Information Gain โดยนำผลลัพธ์ที่ได้จากการลดมิติ มาเป็นข้อมูลอินพุตของซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เพื่อคัดแยกประเภทของโรคมะเร็ง ซึ่งผลการทดลองแสดงให้เห็นว่า การลดข้อมูลโดยวิธี Gain Ratio และ Information Gain มีความเหมาะสม คือ สามารถลดมิติของข้อมูลเหลือ 36 มิติ และเพิ่มความแม่นยำจากเดิม 73.53% เป็น 88.24%

ABSTRACT

Finding subset of informative gene is the main problem for biological processes. The numbers of gene grow up into the tens of thousands. Thus, dimension reduction of data is more benefit for finding subset of gene. When the data are many dimensions or variables that lead to data sparsity and cause the problem of curse of dimensionality. This research focuses on comparing efficiency of three dimension reduction techniques Correlation Based Feature Selection, Gain Ratio, and Information Gain using acute leukemia data that contains 7,129 dimensions, 2 groups that are ALL and AML. In this way, the output from each of the three dimension reduction techniques is used to input data of Support Vector Machine (SVM) for classifying data. The result of experiment shows that Gain Ratio and Information Gain

* *นักศึกษา หลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้า พระนครเหนือ*

** *อาจารย์ ภาควิชาวิทยาศาสตร์ประยุกต์และสังคม วิทยาลัยเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

*** *ผู้ช่วยศาสตราจารย์ ภาควิชาครุศาสตร์ไฟฟ้า คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

techniques are suitable for working with Support Vector Machine because it can reduce dimension of data into 36 dimensions and increase original corrected classification from 73.53% to 88.24%.

คำสำคัญ : การคัดแยกประเภท การจัดอันดับ ซัพพอร์ตเวกเตอร์แมชชีน
Key Words : Classification, Ranking Support, Vector Machine

บทนำ

การแยกประเภทของมะเร็งเม็ดเลือดขาวเป็นสิ่งที่จำเป็นสำหรับการตรวจวินิจฉัยอาการผู้ป่วย เพื่อที่จะทำให้การดำเนินดูแลการรักษาได้ตรงตามอาการของโรค มะลิกา (2549) กล่าวว่า ในทางการแพทย์การวินิจฉัยมะเร็งเม็ดเลือดขาวแต่ละชนิด มีความสำคัญอย่างยิ่ง เนื่องจากการตอบสนองต่อการรักษาด้วยยาหรือวิธีการรักษาของมะเร็งเม็ดเลือดขาวแต่ละชนิดแตกต่างกัน

แต่ในความเป็นจริงแล้ว การแยกประเภทของมะเร็งเม็ดเลือดขาว ซึ่งมียีนอยู่เป็นจำนวนมาก โดยอาศัยความสามารถของมนุษย์เพียงอย่างเดียวทำได้อย่างจำกัด Golub *et. al.* (1999) กล่าวว่า การดำเนินการเพื่อเพิ่มประสิทธิภาพในการแยกประเภทของมะเร็งให้เป็นแบบอัตโนมัติ จะทำให้เราสามารถทำนายประเภทของมะเร็งได้อย่างชัดเจน

สำหรับในงานทางด้านเหมืองข้อมูล หรือ Data Mining บางกรณีเมื่อข้อมูลมีจำนวนมิติ หรือตัวแปรมาก และมิติข้อมูลไม่ได้ไปในทิศทางเดียวกัน ทำให้ข้อมูลเกิดการกระจาย (data sparse) ซึ่งจะทำให้บางจุดอาจจะไม่มีข้อมูลอยู่เลยซึ่งสิ่งนี้ถูกเรียกว่าเป็น ปัญหาของมิติข้อมูล ดังนั้นการขจัดปัญหาของมิติข้อมูลนั้นนอกจากจะช่วยทำให้สามารถลดการใช้ทรัพยากรต่าง ๆ แล้ว ยังเพิ่มความสามารถในการแยกประเภทข้อมูลได้แม่นยำขึ้นอีกด้วย ซึ่งเทคนิคการลดมิติ เป็นขั้นตอนหนึ่งของการเตรียมข้อมูล (data preprocessing) สามารถค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนกได้อย่างมีระสิทธิภาพ

ปัจจุบันมีหลายเทคนิคสำหรับการค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนก อาทิ วิธีการวิเคราะห์องค์ประกอบ (Principle Component Analysis: PCA) Yijuan (2008) ได้ชี้ให้เห็นว่าวิธีนี้มีประสิทธิภาพต่ำลงเมื่อทดลองค้นหากลุ่มย่อยของยีส ที่มีมิติจำนวนมาก เช่น 2,324 ยีน Yun (2009) กล่าวว่าข้อเสียของ PCA คือไม่สามารถวิเคราะห์ตัวแปรและตัวอย่างไปพร้อม ๆ กัน ภัทราวุฒิ และคณะ (2009) ทำการทดลองเปรียบเทียบวิธีการลดมิติกลุ่มข้อมูลโรคมะเร็งระหว่างเทคนิคการคัดเลือกตัวแปรแบบถอยหลังทีละขั้น (Backward Stepwise Feature Selection: BSFS) กับวิธี PCA เพื่อนำมาเป็นข้อมูลเข้าโครงข่ายประสาทเทียม (ANN) พบว่าวิธี PCA มีประสิทธิภาพในการลดมิติของข้อมูลต่ำลงมาก เมื่อข้อมูลมีการกระจายตัวสูง ซึ่งต่างกับวิธี BSFS ที่มีประสิทธิภาพลดต่ำลงเพียงเล็กน้อย

นอกจากนี้ในงานวิจัยเกี่ยวกับการค้นหากลุ่มย่อยของยีนส่วนใหญ่ เช่น Yijuan (2008) Mitsubayashi *et. al.* (2008) Laura *et. al.* (2008) และ Jin and Sung (2008) มักนิยมใช้วิธีการจัดอันดับการแสดงออกของยีน (Ranking Gene Expression) อาทิ T-test, Difference of Mean, Correlation based Feature Selection หรือ Chi-square ผลที่ได้จากการทดลองคือข้อมูลการแสดงออกของยีนในระดับที่แตกต่างกัน ทั้งนี้มีความเชื่อว่าระดับการแสดงออกที่สูงของยีนในตัวอย่างทดสอบใด ๆ ย่อมแสดงว่ายีนนั้นมีอิทธิพลต่อตัวอย่างนั้น ๆ อย่างไรก็ตามวิธีการจัดอันดับยีนเพียงอย่างเดียวก็ยังไม่เพียงพอต่อการคัดแยกประเภทของมะเร็งเม็ดเลือดขาว

เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector machine : SVM) ซึ่งเป็นวิธีการหนึ่งของเรียนรู้แบบมีผู้สอน (Supervised Learning) และมีประสิทธิภาพสำหรับการคัดแยกประเภทกลุ่มข้อมูล อย่างไรก็ตาม Kamal *et. al.* (2009) กล่าวว่า

จากผลลัพธ์การทดลองแยกประเภทยีนด้วย ซัพพอร์ตเวกเตอร์แมชชีน พบว่า มีความไวมากต่อจำนวนของยีนที่ถูกคัดเลือก และ Cheng-San *et. al.* (2008) กล่าวว่า ข้อมูลการแสดงออกของยีนที่มีมิติที่มากและมีลักษณะขนาดเล็ก ทำให้การฝึกฝนและทดสอบของวิธีทั่วไปในการคัดแยกประเภททำได้ยาก

ดังนั้นเพื่อที่จะค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนกและเพิ่มประสิทธิภาพในการแยกประเภทของมะเร็งเม็ดเลือดขาว งานวิจัยนี้จึงได้ทำการเปรียบเทียบวิธีการจัดอันดับจำนวน 3 วิธีคือ Correlation Based Feature Selection (CFS), Gain Ratio (GR) และ Information Gain (IG) เพื่อใช้เป็นตัวลดมิติของข้อมูลยีนที่เหมาะสม และนำมาเป็นอินพุตของซัพพอร์ตเวกเตอร์แมชชีนเพื่อการคัดแยกประเภทมะเร็งเม็ดเลือดขาว

ในหัวข้อต่อไป จะกล่าวถึงอุปกรณ์และวิธีการวิจัย หัวข้อที่ 3 จะแสดงผลการวิจัยและการอภิปรายผล สรุปอภิปรายผลในหัวข้อที่ 4 และข้อเสนอแนะในหัวข้อที่ 5

อุปกรณ์และวิธีการวิจัย

ถ้ามิติของข้อมูลเข้ามีขนาดใหญ่มาก การลดขนาดของมิติข้อมูลจะช่วยให้เพิ่มความแม่นยำของการวิเคราะห์ข้อมูล สำหรับจุดมุ่งหมายในการลดมิติของข้อมูล Cunningham (2007) กล่าวสรุปไว้ดังนี้

(ก) เพื่อเจาะจงกลุ่มของตัวแปรที่ต้องการลด ซึ่งทำให้ผลลัพธ์ของการพยากรณ์สามารถนำไปใช้ประโยชน์ได้เต็มที่

(ข) อัลกอริธึมที่ใช้เรียนรู้หลายๆ อย่างใช้เวลาในการฝึกฝน หรือการแยกประเภทกลุ่มข้อมูลเพิ่มขึ้นตามจำนวนของตัวแปร

(ค) สิ่งรบกวนหรือตัวแปรที่ไม่เกี่ยวข้องสามารถมีอิทธิพลต่อการแยกประเภทกลุ่มข้อมูลซึ่งมักจะส่งผลต่อความแม่นยำ

(ง) บางครั้งจำนวนของตัวแปรข้อมูลเข้าที่มาก อาจใช้อธิบายได้พอๆ กับค่าเฉลี่ยของตัวแปรข้อมูลเข้าทั้งหมด

สำหรับเทคนิคการคัดกลุ่มย่อยของยีนที่มีอำนาจจำแนก ซึ่งจะใช้วิธีลดตัวมิติของข้อมูล มีหลายวิธีด้วยกัน แต่ในงานวิจัยนี้จะใช้วิธีการจัดอันดับโดยนำเสนอ 3 วิธี คือในหัวข้อ 2.1 2.2 และ 2.3

1. เทคนิคการเลือกมิติข้อมูลโดยใช้ความสัมพันธ์ (Correlation Based Feature Selection : CFS)

Mark (1999) ได้กล่าวถึง CFS ว่าคืออัลกอริธึมการกรองที่ง่ายๆ โดย CFS จะจัดอันดับกลุ่มย่อยของมิติข้อมูล ตามความสัมพันธ์ที่อยู่บนพื้นฐานของฟังก์ชันการประมาณแบบ heuristic ซึ่งกลุ่มย่อยของมิติข้อมูลจะมีความสัมพันธ์กันสูงกับคลาสและไม่มีความสัมพันธ์กับคลาสอื่นๆ สำหรับมิติข้อมูลที่ไม่เกี่ยวข้องอาจจะถูกละทิ้ง เพราะมิติข้อมูลเหล่านี้อาจจะมีความสัมพันธ์ต่ำกับคลาส มิติข้อมูลที่ซ้ำซ้อนอาจจะถูกขจัดออกไปจากกลุ่มมิติข้อมูลที่มีความสัมพันธ์สูง สมการประเมินกลุ่มย่อยของมิติข้อมูลแบบ CFS แสดงในสมการที่ (1)

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (1)$$

โดยที่ Ms คือ ค่าที่ค้นหาได้ของมิติข้อมูลกลุ่มย่อย S ซึ่งประกอบด้วยมิติข้อมูล k

$\overline{r_{cf}}$ คือ ค่าเฉลี่ยความสัมพันธ์ของตัวแปรกับคลาส $(f \in S)$

$\overline{r_{ff}}$ คือ ค่าเฉลี่ยความสัมพันธ์ระหว่างมิติของข้อมูล

2. เทคนิคการวัด Information Gain (IG)

Han and Kamber (2006) ได้กล่าวว่าอัลกอริธึม IG ใช้ในการเลือกมิติของข้อมูล เพื่อใช้ในการแบ่งแยกข้อมูล อัลกอริทึม IG จะคำนวณค่า Gain สำหรับแต่ละมิติข้อมูล ซึ่งถ้ามิติข้อมูลใดมีค่า Gain สูงสุดจะถูกเลือกให้เป็นกลุ่มย่อยของมิติข้อมูลที่มีอำนาจจำแนก สมการที่ 2 แสดงการคำนวณค่า Entropy และสมการที่ 2 แสดงการคำนวณค่า Information Gain

$$Entropy(t) = -\sum_i p(j \mid t)\log_2 p(j \mid t) \qquad (2)$$

โดยที่ $-\sum_i$ คือผลรวมของความน่าจะเป็นของค่า j ที่เกิดในคลาส t

$$GAIN = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right) \quad (3)$$

โดยที่ $Entropy(p)$ คือ ค่า Entropy ของตัว Root

$\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$ คือ ค่า Entropy ในแต่ละโหนดย่อย

3. เทคนิคการวัดแบบ Gain Ratio (GR)

Mark (1999) ได้แสดงวิธีของ GR ว่าเป็นการประเมินความน่าเชื่อถือของมิติข้อมูลโดยการวัด Gain Ratio ในแต่ละคลาส การคำนวณ GR โดยนำค่า Information Gain มา สมการที่ 4 แสดงการคำนวณวัด Gain Ratio

$$GainR(Class, Attribute) = \frac{(H(Class) - H(Class \mid Attribute))}{H(Attribute)} \quad (4)$$

4. เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM)

Vapnik (1995) ได้นำเสนอเทคนิคซัพพอร์ตเวกเตอร์แมชชีน วัตถุประสงค์ที่สำคัญของวิธีการนี้ คือ พยายามที่จะทำการลดความผิดพลาดจากการทำนาย (Minimize error) พร้อมกับเพิ่มระยะแยกแยะให้มากที่สุด (Maximized Margin) ซึ่งต่างจากเทคนิคโดยทั่วไปเช่นโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ที่มุ่งเพียงทำให้ความผิดพลาดจากการทำนายให้ต่ำที่สุดเพียงอย่างเดียว

กำหนดลักษณะข้อมูลทดลอง $D = \{x_i, y_i; I = 1,2,...,n\}$ โดยที่ $xi = (x_{i1}, x_{i2},...,x_{in}) \in R^n$ เป็นข้อมูลนำเข้า และ $y_i \in (+1, -1)$ ซึ่ง y_i จะใช้แสดงตัวแปรของคลาส เช่น คลาส ALL มีค่าเท่ากับ 1 และคลาส AML มีค่าเท่ากับ -1 เป็นต้น ทั้งนี้ SVM จะใช้ฟังก์ชั่นการตัดสินใจที่มีความสามารถในการแยกแยะค่าดังสมการที่ (5)

$$y = sign\left\{ \sum_{j=1}^{n} w_j \phi_j(x) + b \right\} \quad (5)$$

$$\phi(x) = [\phi_1(x), \phi_2(x), ..., \phi_n(x)]^T \quad (6)$$

การแบ่งกลุ่มข้อมูล x ซึ่งไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรง จะต้องอาศัยสมการที่ (6) โดยที่ ϕ (x) จะใช้แทนฟังก์ชันสำหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่อยู่ในรูปแบบสมการเชิงเส้น และสามารถนำมาคัดแยกประเภทได้ W_j แทนค่าน้ำหนัก (Weighting) ที่เชื่อมโยงจากปริภูมิลักษณะ (feature space) ไปสู่ปริภูมิผลลัพธ์ (output space) และ b นั้นเป็นค่าโน้มเอียง (Bias หรือ threshold) ดังที่กำหนดไว้ดังสมการที่ (7)

$$f(x) = \sum_{j=1}^{n} w_j \phi_j(x) + b \quad (7)$$

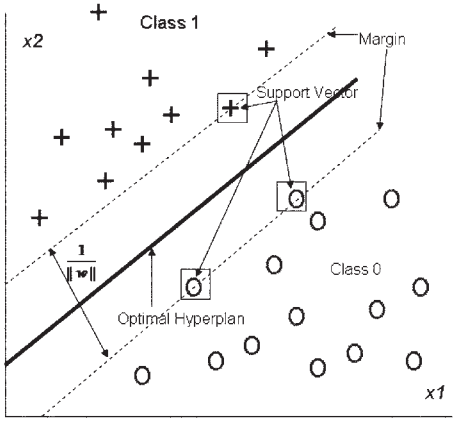
ดังนั้นจากสมการดังกล่าว สามารถแบ่งประเภทของข้อมูลได้ในลักษณะเชิงเส้น จากฟังก์ชันการแบ่ง $f(x)$ และถ้าข้อมูลในกลุ่ม D สามารถแยกโดยใช้สมการเส้นตรงได้แล้ว ก็แสดงว่ามีการแบ่งแยกประเภทที่สมบูรณ์ ในกรณีที่การจัดแบ่งประเภทสำหรับค่า y ถูกกำหนดให้มีค่าของคลาสเป็น -1 และ +1 เมื่อเป็นเช่นนี้แล้วสมการที่ (8) และสมการ (9) และนำมารวมกันดังสมการที่ (10)

$$w^T \cdot x + b \geq +1 \;\; ; \; y_i = +1 \quad (9)$$

$$w^T \cdot x + b \leq -1 \;\; ; \; y_i = -1 \quad (10)$$

$$y_i(w^T \cdot x + b) - 1 \geq 0 \;\; ; \; \forall i \quad (11)$$

จากสมการดังกล่าวที่เป็นเชิงเส้น และมีความสามารถในการแยกแยะข้อมูลทำให้เกิดการเพิ่มระยะของเส้นแบ่งเขตที่มีความกว้างเท่ากับ $2/\|w\|^2$ ดังแสดงในภาพที่ 1



ภาพที่ 1 การแยกประเภทกลุ่มข้อมูลด้วยเทคนิค SVM

อย่างไรก็ตามเนื่องจากในบางกรณีการคัดแยกประเภท ไม่สามารถทำได้ถูกต้องโดยสมบูรณ์ ดังนั้นจึงต้องมีการกำหนดตัวแปรสำหรับยอมรับค่าความผิดพลาด โดยการเพิ่มตัวแปร ξ (Slack Variable) โดยเพิ่มตัวแปรนี้เข้าไปในสมการที่ (10) และ (11) ทำให้เกิดสมการที่ (12) และ (13) ดังนี้

$$w^T \cdot x + b \geq +1 - \xi_i \; ; \; y_i = +1 \qquad (12)$$

$$w^T \cdot x + b \leq -1 + \xi_i \; ; \; y_i = -1 \qquad (13)$$

จากการกำหนดค่า $\xi_i > 0$ ทำให้โครงสร้างของ SVM บรรลุวัตถุประสงค์ใน 2 ส่วนคือ การเพิ่มระยะแยกแยะให้มากที่สุดและลดข้อผิดพลาดในการทำนายให้ต่ำที่สุด ดังแสดงในสมการที่ (14)

$$\underset{w,b,\xi}{Minimize} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \qquad (14)$$

โดยที่ : $y_i(w^T\phi(x_i) + b) + \xi_i - 1 \geq 0$

$\xi_i \geq 0, i = 1, 2, ..., N$

จากสมการดังกล่าวจะสังเกตได้ว่ามีค่า C ซึ่งเป็นตัวแปรที่ใช้กำหนดความสมดุลของระหว่างระยะแยกแยะที่สูงสุด หรือลดค่าความผิดพลาดในการทำนายให้ต่ำที่สุด ซึ่งค่า C จะถูกผู้ใช้งานกำหนดขึ้น ทั้งนี้โดยปรกติค่า C มักกำหนดให้มีค่ามากกว่า 1 และนำมาแก้ปัญหาค่าที่เหมาะสมที่สุดด้วยฟังก์ชั่นลากรองจ์ (Lagrangian) ด้วยการกำหนดค่าตัวแปรแบบเซตคู่ (Dual Sets) เพิ่มเติม แล้วทำการแก้ปัญหาจากการกำหนดข้อจำกัดที่ดีที่สุด (Constrained Optimization) ทำให้ได้ผลดังสมการที่ (15)

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{N}\alpha_i \qquad (15)$$

โดยที่ : $\sum_{i=1}^{N} y_i \alpha_i = 0$

$0 \leq \alpha_i \leq C, i = 1, 2, ..., N$

โดยที่ α_i เป็น Lagrange multipliers เพื่อที่จะใช้สำหรับการแก้ปัญหาที่ดีที่สุด α_i^* เพื่อใช้ในการจำแนกข้อมูลที่ไม่ได้เป็นฟังชันการจำแนกข้อมูลแบบเชิงเส้นดังสมการที่ (16)

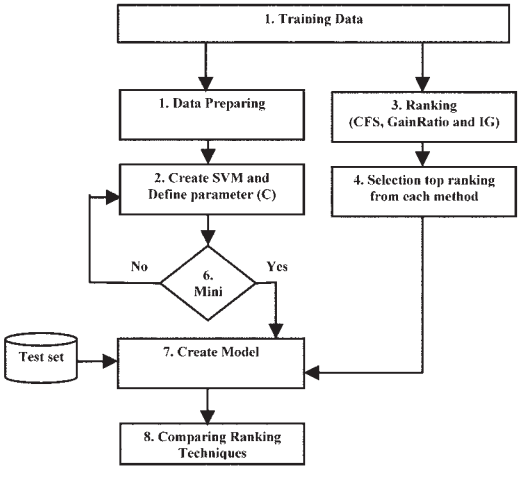
$$f(x) = \sum_{j=1}^{N} w_j \alpha_j^* K(x, x_j) + b \qquad (16)$$

ในขณะที่ $K(x, x_j) = \phi^T(x)\phi(x_j)$ นั้นเป็น kernel function

สำหรับในงานวิจัยนี้ผู้วิจัยได้เลือกใช้ kernel function แบบ Radial Basis Function เนื่องจาก Cheng (2008) และ เดช และคณะ (2009) กล่าวว่าผลการทดลองที่ผ่านมาพบว่าเป็น kernel function ที่ให้ผลดี

5. วิธีการวิจัย

งานวิจัยนี้จะนำเอาข้อมูลยีนของโรคมะเร็งเม็ดเลือดขาวแบบเฉียบพลัน (Acute Leukemia) หรือ Leukemia Dataset จาก Pablo de Olavide University of Seville *(http://www.upo.es/eps/bigs /datasets.html)* ซึ่งข้อมูลชุดนี้ประกอบด้วย 2 คลาสคือ ALL และ AML ซึ่งมีจำนวนมิติของข้อมูล 7,129 มิติ จำนวน 76 ตัวอย่าง



ภาพที่ 2 ขั้นตอนการทำงานของโมเดล SVM ร่วมกับวิธีการจัดอันดับในรูปแบบต่าง ๆ

5.1 นำข้อมูลมาผ่านกระบวนการเตรียมข้อมูล (Data Preprocessing) โดยทำการ Normalize ให้มีค่าระหว่าง –1 ถึง +1 จากนั้นแบ่งข้อมูลออกเป็นชุดฝึกฝน (Training Data) และชุดทดสอบ (Testing Data) จำนวนอย่างละ 38 ตัวอย่างเท่า ๆ กัน

5.2 ทดลองใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เพื่อแยกประเภทของคลาส โดยสังเกตจากการเปลี่ยนค่าพารามิตเตอร์ (C) ที่แสดงค่าเฉลี่ยความผิดพลาดที่ต่ำสุดว่าค่าใดที่มีความเหมาะสมที่สุด

5.3 ทดลองใช้วิธีการจัดอันดับข้อมูล 3 วิธีคือ CFS, GR และ IG และคัดเลือกกลุ่มย่อยของยีนที่มีอำนาจจำแนกจากการเลือกยีนที่มีค่าอันดับจากมากไปน้อยในแต่ละวิธี

5.4 นำผลลัพธ์ที่ได้จากการหาค่าพารามิเตอร์ที่เหมาะสมของของซัพพอร์ตเวกเตอร์แมชชีน มาสร้างโมเดล เพื่อร่วมกับกลุ่มย่อยของยีนที่ได้จากแต่ละวิธีการจัดอันดับในขั้นตอน 2.5.3

5.5 การวัดประสิทธิภาพของวิธีการจัดอันดับว่าวิธีใดจะเหมาะสมในการทำงานร่วมกับซัพพอร์ตเวกเตอร์แมชชีน ในการลดมิติข้อมูลเข้า โดยใช้วิธีการเปรียบเทียบค่าความแม่นยำ (precision) ค่าความระลึก (Recall) และการวัดประสิทธิภาพโดยรวม (F-measure) ที่ได้จากเทคนิคซัพพอร์ตเวกเตอร์แมชชีน ดังสมการที่ (17) (18) (19) ตามลำดับ

$$\Pr ecision = \frac{TP}{TP + FP} \quad (17)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (18)$$

$$F - measure = \frac{2x(\mathrm{Re}callx \Pr ecision)}{\mathrm{Re}call + \Pr ecision} \quad (19)$$

ผลการวิจัยและการอภิปรายผล

1. ผลลัพธ์การหาค่าพารามิตเตอร์ *(C)*

เพื่อที่จะหาค่าพารามิเตอร์ (C) ที่เหมาะสมสำหรับเทคนิคซัพเวกเตอร์แมชชีน จึงได้ทำการทดลองการแยกประเภทข้อมูล และโดยเปลี่ยนค่าพารามิตเตอร์และสังเกตค่าต่าง ๆ ซึ่งพบว่าเมื่อกำหนดค่า C เท่ากับ 10 จะมีค่า MSE ต่ำสุดคือ 0.5145 ค่าความถูกต้องจากการแยกประเภท 73.53% มีค่าความแม่นยำ 34.6% ค่าความระลึก 58.8% และการวัดประสิทธิภาพโดยรวม 43.6%

2. ผลลัพธ์จากเทคนิคการจัดอันดับ

การจัดอันดับข้อมูล ได้นำเสนอไว้ในตารางที่ 1

ตารางที่ 1 จำนวนข้อมูลที่ได้จากการจัดอันดับข้อมูล

Ranking Method	Number of Attributes
1. CFS	36
2. Gain Ratio	915
3. IG	915

มีหลายงานวิจัยหลายงานอาทิ Kamal *et. al.* (2009), Shen (2009) และ Golub *et. al.* (1999) ที่แนะนำว่าการใช้มิติของข้อมูลยีนระหว่าง 30-50 มิติ เป็นขนาดที่สมเหตุผลสำหรับการระบุหน้าที่ของยีน ดังนั้นงานวิจัยนี้จึงเลือกมิติข้อมูลจำนวน 36 มิติโดยทำการคัดเลือกมิติข้อมูลจากการจัดอันดับ จากมากไปน้อย

3. ผลลัพธ์จากการทดลองแยกประเภทข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ข้อมูลนำเข้าที่ได้จากวิธีการจัดอันดับ

ในส่วนของผลลัพธ์จากการทดลองจากการลดมิติข้อมูลด้วยวิธีการจัดอันดับ เพื่อนำเป็นข้อมูลอินพุตของการแยกประเภทด้วยซัพพอร์ตเวกเตอร์แมชชีน แสดงให้เห็นว่าวิธี GR และ IG มีค่าความแม่นยำ ความระลึก และการวัดประสิทธิภาพโดยรวมเท่ากัน คือ 90.2% 88.2% และ 87.8% ตามลำดับ ซึ่งสูงกว่าวิธี CFS คือ 86.4% 82.4% และ 82.4% นำเสนอในตารางที่ 2

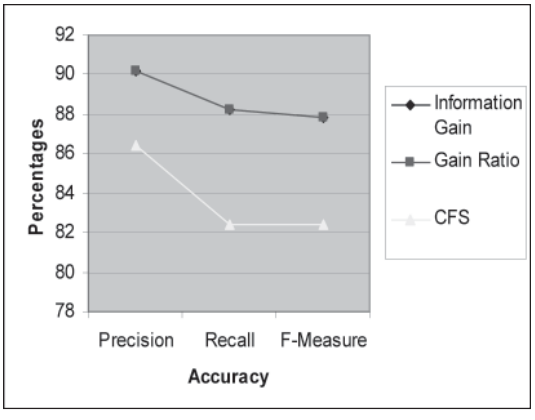
นอกจากนี้วิธี GR และ IG ยังเพิ่มค่าความถูกต้องจากการแยกประเภทเป็น 88.24% ได้นำเสนอในตาราง 3 และภาพที่ 3 แสดงการวัดประสิทธิภาพในรูปของกราฟ

ตารางที่ 2 ผลการวัดประสิทธิภาพของวิธีการจัดอันดับ

	Precision (%)	Recall (%)	F-Measure (%)
CFS	86.4	82.4	82.4
GR	90.2	88.2	87.8
IG	90.2	88.2	87.8

ตารางที่ 3 ผลการวัดค่าความผิดพลาดเฉลี่ยต่ำสุด และค่าความถูกต้องจากการแยกประเภท

Ranking Method	MSE	Correct Classify (%)
CFS	0.4201	82.35 %
GR	0.343	88.24 %
IG	0.343	88.24 %



ภาพที่ 3 กราฟแสดงประสิทธิภาพของวิธีการจัดอันดับ

สรุปผลการวิจัย

เมื่อมิติของข้อมูลมีเป็นจำนวนมาก จะทำให้เกิดปัญหามิติของข้อมูล และส่งผลให้การแยกประเภทข้อมูลมักเกิดความผิดพลาด รวมถึงการสิ้นเปลืองทรัพยากรต่าง ๆ ดังนั้นงานวิจัยครั้งนี้ จึงมีวัตถุประสงค์เพื่อที่จะค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนกและเพิ่มประสิทธิภาพในการแยกประเภทของมะเร็งเม็ดเลือดขาว โดยใช้วิธีการจัดอันดับเพื่อลดมิติของข้อมูล ร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน เพื่อคัดแยกประเภทมะเร็ง ผลลัพธ์แสดงให้เห็นถึงประสิทธิภาพของวิธีลดมิติข้อมูลด้วยวิธี Gain Ratio และ Information Gain ร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีนมีค่าความแม่นยำ ความระลึก และการวัดประสิทธิภาพโดยรวมเท่ากัน คือ 90.2% 88.2% และ 87.8% ตามลำดับ ซึ่งสูงกว่าวิธี CFS ร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีนคือ 86.4% 82.4% และ 82.4% ตามลำดับ

งานวิจัยนี้ยังได้แสดงให้เห็นว่าเมื่อใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน แยกประเภทข้อมูลเพียงอย่างเดียว โดยใช้จำนวนมิติทั้งหมดคือ 7,129 มิติ จะได้ค่าความถูกต้องจากการแยกประเภท 73.53% ซึ่งต่ำกว่าเมื่อใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แม-ชชีน ทำให้ลดจำนวนมิติของข้อมูลเหลือ 36 มิติ ซึ่งจะให้ค่าความถูกต้องจากการแยกประเภท เพิ่มขึ้นเป็น 88.24%

ข้อเสนอแนะ

การลดมิติของข้อมูล อาจจำเป็นที่จะต้องทำการทดสอบกับข้อมูลให้หลากหลายชุดข้อมูล เพื่อดูประสิทธิภาพการทำงานที่แท้จริง ซึ่งจะเป็นส่วนหนึ่งของงานในอนาคตที่น่าสนใจ

เอกสารอ้างอิง

เดช ธรรมศิริ, ณรงค์ โพธิ, วาทินี นุ้ยเพียน, ภัทราวุฒิ แสงศิริ, ภรัณยา อำมฤครัตน์ และ พยุง มีสัจ. 2009. การให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยเทคนิคซัพพอร์ตเวกเตอร์แมชชีนรวมทั้งการเลือกใช้ลักษณะที่เหมาะสมร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีค้นหาแบบกริช. in The 5th National Conference on Computing and Information Technology. (5)2009 : 249-257.

ภัทราวุฒิ แสงศิริ ศจีมาจ ณ วิเชียร และพยุง มีสัจ. 2009. การเปรียบเทียบประสิทธิภาพการลดตัวแปรข้อมูลเข้าที่เหมาะสมสำหรับโครงข่ายประสาทเทียมระหว่างเทคนิคการเลือกตัวแปรแบบถอยหลังทีละขั้น และการวิเคราะห์องค์ประกอบ เพื่อพยากรณ์กลุ่มข้อมูลโรคมะเร็ง in The 5th National Conference on Computing and Information Technology. (5)2009 : 851-858.

มะลิกา วัฒนะ. 2549. การศึกษาคุณลักษณะทางกายภาพของมะเร็งเม็ดเลือดขาวเฉียบพลัน. วิทยานิพนธ์ปริญญาวิทยาศาสตร์มหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยขอนแก่น.

Cheng-San, Y., Li-Yeh, C.,*et al.* 2008. A hybrid approach for selecting gene subsets using gene expression data. Soft Computing in Industrial Applications, SMCia '08. IEEE Conference on. 159-164.

Cheng-San, Y., Li-Yeh, C., Chao-Hsuan, K., and Cheng-Hong, Y. 2008. A hybrid approach for selecting gene subsets using gene expression data. in Soft Computing in Industrial Applications, SMCia '08. IEEE Conference. 159-164.

Golub, TR., Slonim, DK., *et al.* 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286(5439): 531-537.

Jiawei Han and Micheline Kamber. 2006. Data Mining Concepts and Techniques. San Francisco,CA 94111. Morgan Kaufmann Publishers.

Jin-Hyuk, H., and Sung-Bae, C. 2008. Cancer classification with incremental gene selection based on DNA microarray data. in Computational ntelligence in Bioinformatics and Computational Biology,CIBCB '08. IEEE Symposium. 70-74.

Kamal, A., Zhu, X., *et al.* 2009. The Impact of Gene Selection on Imbalanced Microarray Expression Data. Bioinformatics and Computational Biology. 259-269.

Laura, LE., Sanna, F., Riitta, L., and Tero, A. 2008. Reproducibility-Optimized Test Statistic for Ranking Genes in Microarray Studies. IEEE/ACM Trans. Comput. Biol. Bioinformatics. 1545-5963, vol. 5: 423-431.

Mark, A., Hall. 1999. Correlation-based Feature Selection for Machine Learning. Doctor of Philosphy. Department of Computer Science, The University of Waikato Newzealand.

Mitsubayashi Hikaru, Seiichiro Aso, Tomomasa Nagashima and Yoshifumi Okada. 2008. Accurate and robust gene selection for disease classification using a simple statistic. ISSN 0973-2063 (online) 0973-2063 (print), Bioinformation. 3(2): 68-71.

Pádraig Cunningham. 2007. Dimension Reduction. Technical Report UCD-CSI-2007-7. *August 8, 2007*: 1-4.

Shen, Q-m., Shi, W., *et al.* 2009. New gene selection method for multiclass tumor classification by class centroid. Journal of Biomedical Informatics. 42(1): 59-65.

Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer-Verlag. New York.

Yijuan, L., Qi, T., *et al.* 2008. Adaptive discriminant analysis for microarray-based classification. ACM Trans. Knowl. Discov. Data. 1556-4681 2(1): 1-20.

Yun, L., and Chongyi, Y. 2009. Feature Selection for Density-Based Clustering. Intelligent Ubiquitous Computing and Education, International Symposium. 226-229.