The 11th Khon Kaen University 2010 Graduate Research Conference
การประชุมทางวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษา ครั้งที่ 11

SDO2

การจำแนกข้อมูลโดยใช้วิธีซัพพอร์ตเวกเตอร์แมชชีน โดยปรับพารามิเตอร์ด้วยขั้นตอนวิธีเชิงพันธุกรรม

Data classification using Support Vector Machine Optimized by Genetic Algorithm

เดช ธรรมศิริ (Dech Thammasiri)* ดร.พยุง มีสัจ (Dr.Phayung Meesad)**

บทคัดย่อ

ในปัจจุบันยังไม่มีขั้นตอนวิธีที่ดีที่สุดสำหรับการคัดแยกข้อมูล ดังนั้นในงานวิจัยนี้จึงมีแนวคิดในการหาโมเดลที่เหมาะสมสำหรับเพิ่มประสิทธิภาพความแม่นยำในการจำแนกข้อมูล เปรียบเทียบประสิทธิภาพด้วย ฐานข้อมูล Austrian Credit , German Credit และ Bankruptcy Data โดยที่ได้นำเอาวิธีซัพพอร์ตเวกเตอร์แมชชีนสำหรับใช้ในการจำแนกข้อมูลร่วมกับวิธีการหาค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับวิธีซัพพอร์ตเวกเตอร์แมชชีนโดยใช้ขั้นตอนวิธีการทางพันธุกรรม เปรียบเทียบผลการวิจัยกับ วิธีการต้นไม้ช่วยตัดสินใจ วิธีการโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ การใช้ซัพพอร์ตเวกเตอร์แมชชีน ผลการวิจัยพบว่า การใช้ซัพพอร์ตเวกเตอร์แมชชีนและ วิธีการหาค่าพารามิเตอร์ที่เหมาะสมที่สุดสำหรับซัพพอร์ตเวกเตอร์แมชชีนโดยใช้ขั้นตอนวิธีการทางพันธุกรรม จะให้ค่าเฉลี่ยความแม่นยำสูงที่สุด

ABSTRACT

Recently, there have not best for the accuracy of data classification. In this paper, we try to find a suitable model to increase the accuracy of data classification. We use Austrian Credit, German Credit, and Bankruptcy Data as our testing data. We use the model which combines the support vector machine and optimize parameter by genetic algorithm. We compare the result from this model with the results from decision tree model, neuron network model and support vector machine model. The result indicates that the model which combines the support vector machine and optimize parameter by genetic algorithm provides the highest accuracy rate.

คำสำคัญ : ซัพพอร์ตเวกเตอร์แมชชีน วิธีการทางพันธุกรรม การจำแนกข้อมูล

Key words : Support Vector Machine, Genetic Algorithm, Classification

---

** นักศึกษาหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยี สารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

*** ผู้ช่วยศาสตราจารย์ ภาควิชาครุศาสตร์ไฟฟ้า คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ*

บทนำ

ปัจจุบันนักวิจัยหลายท่านได้ให้ความสำคัญกับปัญหาที่ต้องการวิธีการจำแนกข้อมูลที่แม่นยำ ดังเช่น การนำมาใช้ในการแก้ปัญหา การให้คะแนนสินเชื่อ (Yu, et. al., 2009) การวิเคราะห์รูปภาพ (Shaoning, et. al., 2005) พยากรณ์เกี่ยวกับคุณสมบัติของยา (Yang, et. al., 2009) การจำแนกเสียง (Chang and Soo, 2007) และอื่นๆ อีกมาก ผู้วิจัยแต่ละท่านมีการนำเสนอวิธีการที่หลากหลาย ไม่ว่าจะเป็นการใช้วิธีการทางสถิติเช่น การวิเคราะห์จำแนกประเภท (Discriminant Analysis) การวิเคราะห์ความถดถอยโลจีสติค (Logistic Regression) ได้ถูกนำมาใช้ในการสร้างโมเดลสำหรับการจำแนกข้อมูล(Banasik, et. al., 2001, Boyes, et. al., 1989, Sarlija, et.al., 2004) หรือ วิธีการที่นำเอาวิธีการทางด้านปัญญาประดิษฐ์ (Artificial Intelligence) มาประยุกต์ใช้ เช่น วิธีต้นไม้ช่วยตัดสินใจ (Decision Tree Models) (Davis, et. al, 1992) วิธีโครงข่ายประสาทเทียม (Artificial Neural Networks) (เดช และคณะ, 2551)วิธีทางพันธุกรรม (Genetic Algorithm) (Chen and Huang, 2003) นอกจากนั้นยังมีนักวิจัยหลายท่านพยายามหาโมเดลที่มีประสิทธิภาพเพิ่มขึ้น เช่น การเปรียบเทียบผลลัพธ์ที่ได้จากการวิจัย โดย Huanga, et. al. (2004) พบว่า วิธีซับพอร์ตเวกเตอร์แมชชีน (Support Vector Machine หรือ SVM) ให้ผลที่แม่นยำกว่าวิธีโครงข่ายประสาทเทียมอีกทั้ง ได้มีผู้วิจัยได้นำวิธีซัพพอร์ตเวกเตอร์แมชชีนมาจำแนกเกี่ยวกับการล้มละลาย (Fan & Palaniswami, 2000, Shin, at.el., 2005, Mina and Lee, 2005) ซึ่งผลที่ได้จากการวิจัยรายงานว่า วิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ผลที่แม่นยำกว่า วิธีการจำแนกข้อมูลแบบโครงข่ายประสาทเทียม และการวิเคราะห์จำแนกประเภท

เหตุผลของการนำเอาซัพพอร์ตเวกเตอร์แมชชีน เป็นตัวแทนของการเรียนรู้คือ วิธีซัพพอร์ตเวกเตอร์แมชชีนไม่ต้องการข้อสันนิฐานก่อนหน้าเกี่ยวกับ ข้อมูลที่เป็นข้อมูลนำเข้า (Input Data) เช่น ข้อมูลที่เป็นรูปแบบโค้งปกติ และ ข้อมูลที่ต่อเนื่องกัน ซึ่งเป็นความแตกต่างจากโมเดลทางสถิติโดยทั่วไป วิธีซัพพอร์ตเวกเตอร์แมชชีนสามารถกระทำการเปลี่ยนโครงสร้างของข้อมูล (Mapping) ข้อมูลที่ไม่ได้อยู่ในรูปแบบที่สามารถแบ่งด้วยสมการเส้นตรงได้ (Nonlinear) จากข้อมูลนำเข้าต้นฉบับไปสู่ลักษณะที่มีมิติสูง (High Dimensional Feature Space) ซึ่งทำให้วิธีซัพพอร์ตเวกเตอร์แมชชีนสามารถใช้ฟังก์ชันสำหรับจำแนกข้อมูลได้ ลักษณะเด่นเฉพาะอีกประการหนึ่งคือ วิธีซัพพอร์ตเวกเตอร์แมชชีนพยายามเรียนรู้ เส้นที่แบ่งกลุ่มข้อมูล ให้ได้ระยะห่างสูงสุด ซึ่งความสามารถสามารถนี้ช่วยให้วิธีซัพพอร์ตเวกเตอร์แมชชีนหลุดพ้นออกจากปัญหาคำตอบที่เหมาะสมที่สุดแบบวงแคบเฉพาะที่ได้ (Local Optimum) ซึ่งปัญหานี้พบได้บ่อยในการเรียนรู้ของวิธีโครงข่ายประสาทเทียม หากแต่ปัญหาที่พบในวิธีซัพพอร์ตเวกเตอร์แมชชีน คือ การกำหนดค่าพารามิเตอร์หากกำหนดค่าไม่เหมาะสมจะมีผลทำให้โมเดลที่ได้มีประสิทธิภาพไม่ดีนัก

จากปัญหาข้างต้นได้มีผู้วิจัยนำเสนอแนวทางในการกำหนดค่าพารามิเตอร์ที่เหมาะสมสำหรับวิธีซัพพอร์ตเวกเตอร์แมชชีน ดังเช่น เดชและคณะ (2552) ได้นำเสนอการใช้วิธีการค้นหาแบบกริชพื่อปรับปรุงประสิทธิภาพของวิธีซัพพอร์ตเวกเตอร์แมชชีน จากหลายงานวิจัยที่ผ่านมาพบหากมีการกำหนดค่าพารามิเตอร์ที่เหมาะสมให้กับวิธีซัพพอร์ตเวกเตอร์แมชชีนจะสามารถทำให้โมเดลที่ได้มีประสิทธิภาพความแม่นยำในการจำแนกข้อมูลที่ดีขึ้น

วัตถุประสงค์ของการวิจัย

1. เพื่อนำเสนอขั้นตอนวิธีการจำแนกข้อมูลโดยพื้นฐานการเรียนรู้แบบผสมผสาน
2. เพื่อหาประสิทธิภาพความแม่นยำสำหรับการจำแนกข้อมูลที่นำเสนอ

ทฤษฎีที่เกี่ยวข้อง

วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เป็นวิธีการแก้ปัญหาด้วยการลดข้อผิดพลาด

จากประสบการณ์ต่ำสุด (Minimized Empirical Error or Risk) และเพิ่มระยะแยกแยะมากสุด (Maximized Margin) เป็นแนวทางที่ซัพพอร์ตเวกเตอร์ใช้ ในการแก้ปัญหา วิธีซัพพอร์ตเวกเตอร์แมชชีนพัฒนาจากหลักการของ SRM (Structure Risk Minimization) ซึ่งให้ผลการเรียนรู้ โดยรวมที่ดีกว่าเมื่อเทียบกับวิธีพื้นฐาน ERM (Empirical Risk Minimization) ที่นิยมใช้ในระบบ โครงข่ายประสาทเทียมทั่วไป

วิธีซัพพอร์ตเวกเตอร์แมชชีน (SVM) นั้นเป็นเครื่องมือสำหรับการเรียนรู้แบบใหม่บนพื้นฐานของการเรียนรู้ทฤษฎีทางสถิติซึ่งให้ประสิทธิภาพที่ดีกว่าวิธีการโดยทั่วไปสำหรับปัญหาในการจำแนกข้อมูล SVM นั้นได้รับการนำเสนอครั้งแรกโดย Vapnik (1995) และแนวคิดพื้นฐานของทฤษฎีสามารถอธิบายได้โดยสรุปดังต่อไปนี้

ให้กลุ่มข้อมูลทดลอง $D = \{x_i, y_i; i = 1, 2, ..., n\}$ ในขณะที่ $x_i = (x_{i1}, x_{i2}, ..., x_{in}) \in R^n$ *นั้นเป็นข้อมูลนำเข้า และ* $y_i \in \{+1, -1\}$ (+1="ข้อมูล Class 1", -1=" ข้อมูล Class 0") ซึ่งเป็นการกำหนดกลุ่มเป้าหมาย ให้ SVM โดยที่ SVM นั้นมุ่งเป้าเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแยกแยะค่าที่ไม่ทราบได้โดยใช้สมการที่ (1)

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

$$\phi(x) = [\phi_1(x), \phi_2(x), ..., \phi_n(x)]^T \qquad (2)$$

จากตัวอย่างสมการที่ (2) เป็นกลุ่มข้อมูล x ซึ่งไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงแบ่งแยกได้ โดยที่ $\phi()$ แทนฟังก์ชันสำหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่อยู่ในรูปแบบที่สมการเชิงเส้นสามารถแยกแยะได้ w_j แทนค่าน้ำหนัก (Weighting) ที่เชื่อมโยงจาก Feature Space ไปสู่ Output Space และ b นั้นเป็นค่าโน้มเอียง (Bias หรือ Threshold) ที่ตั้งไว้ทำได้ให้สมการที่ (3) สำหรับการจำแนกข้อมูล

$$f(x) = \sum_{j=1}^{n} w_j \phi_j(x) + b \qquad (3)$$

ฟังก์ชันที่เรานั้นสร้างการแบ่งประเภทของข้อมูลเชิงเส้นและได้ฟังก์ชันการแบ่ง $f(x)$ ถ้าข้อมูลเซต D นั้นสามารถแยกโดยใช้สมการแบบเส้นตรง ก็แสดงว่ามีการแบ่งแยกประเภทที่สมบูรณ์กรณีการจัดแบ่งประเภทสำหรับค่า y ถูกกำหนดให้มีค่า เป็น -1 และ +1 ดังนั้นในสมการ (4) สามารถแทนด้วยสมการ (5) และสามารถเขียนรวมกันได้ในสมการ (6)

$$w^T \cdot x + b \geq +1 \;\; ; \; y_i = +1 \qquad (4)$$

$$w^T \cdot x + b \leq -1 \;\; ; \; y_i = -1 \qquad (5)$$

$$y_i(w^T \cdot x + b) - 1 \geq 0 \;\; ; \; \forall i \qquad (6)$$

ดังนั้นสมการที่เป็นเชิงเส้นและสามารถแยกแยะข้อมูลได้เกิดจากการเพิ่มขอบเขต เส้นแบ่งที่มีความกว้างเท่ากับ $2 / \|w\|^2$

หากแต่บางครั้งไม่สามารถแยกแยะข้อมูลได้ถูกต้องทั้งหมด ทำให้ต้องมีการกำหนดตัวแปรเพื่อยอมรับค่าความผิดพลาด โดยทำการเพิ่มตัวแปร ξ (Slack Variable) โดยเขียนแสดงรายละเอียดได้ดังสมการที่ (7) และ (8)

$$w^T \cdot x + b \geq +1 - \xi_i \;\; ; \; y_i = +1 \qquad (7)$$

$$w^T \cdot x + b \leq -1 + \xi_i \;\; ; \; y_i = -1 \qquad (8)$$

โดยที่ $\xi_i > 0$

ทำให้ได้โครงสร้าง ของ SVM ซึ่งประกอบด้วย 2 ส่วนหลัก คือ การเพิ่มระยะแยกแยะมากสุด และ การแก้ปัญหาด้วยการลดข้อผิดพลาดให้ต่ำที่สุด ดังแสดงในสมการที่ (9)

$$\underset{w,b,\xi}{Minimize} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \qquad (9)$$

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

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

โดยมีค่า C ซึ่งเป็นค่าตัวแปรที่ผู้ใช้สามารถกำหนดค่าได้เองเพื่อปรับความสมดุลระหว่างการให้ความสำคัญของระยะแยกแยะสูงสุด หรือให้ความสำคัญกับค่าความผิดพลาดที่ต้องการให้ต่ำที่สุด โดยปกติค่า C จะกำหนดให้มีค่ามากกว่า 1 จากนั้นทำการแก้ปัญหาด้วยฟังก์ชันลากรองจ์ (Lagrangian) ด้วยการกำหนดค่าตัวแปรแบบเซตคู่ (Dual Sets) เพิ่มเติม แล้วทำการแก้ปัญหาจากการกำหนดข้อจำกัดที่ดีที่สุด (Constrained Optimization) ทำให้ได้ผลดังสมการที่ (10)

$$\underset{\alpha}{Minimize} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i \qquad (10)$$

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

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

โดยที่ α_i เป็น Lagrange Multipliers เพื่อที่จะใช้สำหรับการแก้ปัญหาที่ดีที่สุด α_i^* เพื่อใช้ในการจำแนกข้อมูลที่ไม่ได้เป็นฟังชันการจำแนกข้อมูลแบบเชิงเส้นดังสมการที่ (11)

$$f(x) = \sum_{j=1}^{N} w_j \alpha_j^* K(x, x_j) + b \qquad (11)$$

ในขณะที่ $K(x, x_j) = \phi^T(x)\phi(x_j)$ นั้นเป็น Kernel Function และมี Kernel Function ที่พบได้บ่อย 3 แบบด้วยกัน ดังแสดงในสมการที่ (12) (13) และ (14)ได้แก่

Polynomial Kernel:

$$k(x_i, x_j) = (1 + x_i \cdot x_j)^d \qquad (12)$$

Radial Basis Function Kernel :

$$k(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \qquad (13)$$

Sigmoid Kernel :

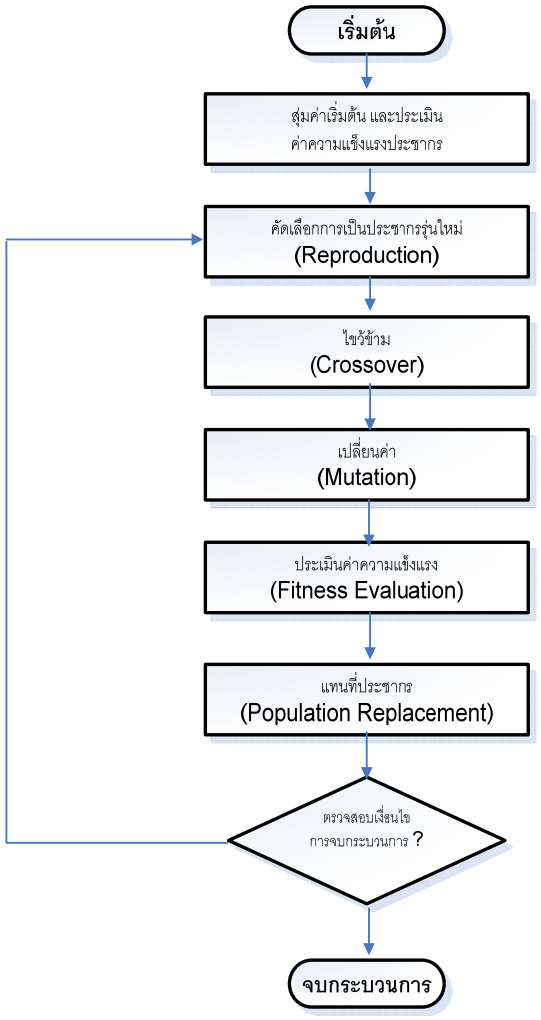
$$k(x_i, x_j) = \tanh(kx_i \cdot x_j) - \delta) \qquad (14)$$

ผู้วิจัยได้เลือกใช้ Radial Basis Function เป็น Kernel Function สำหรับงานวิจัยครั้งนี้ เนื่องจากได้มีผู้วิจัยหลายท่านได้ทดลองแล้วพบว่า ให้ผลลัพธ์ที่ดี

วิธีการทางพันธุกรรม (Genetic Algorithm) เป็นเทคนิคสำหรับค้นหาผลเฉลย (Solutions) หรือคำตอบโดยประมาณของปัญหา โดยอาศัยหลักการจากทฤษฎีวิวัฒนาการจากชีววิทยา และ การคัดเลือกตามธรรมชาติ (Natural Selection) นั่นคือ สิ่งมีชีวิตที่เหมาะสมที่สุดจึงจะอยู่รอด กระบวนการคัดเลือกได้เปลี่ยนแปลงสิ่งมีชีวิตให้เหมาะสมยิ่งขึ้น ด้วยตัวปฏิบัติการทางพันธุกรรม (Genetic Operator) (Holland, 1962) เช่น การถ่ายทอดพันธุกรรม (Inheritance) การคัดเลือกคู่ (Selection) การกลายพันธุ์ (Mutation) และการสลับสายพันธ์ (Crossover) เป็นต้น เนื่องจากเป็นการหาคำตอบที่เหมาะสม ดังนั้นคำตอบที่ได้จึงไม่ใช่คำตอบที่ดีที่สุด เสมอไป โดยทั่วไปแล้ว ขั้นตอนวิธีเชิงพันธุกรรมอย่างง่าย(Holland, 1992) มีขั้นตอนในการหาคำตอบดังนี้

1. สร้างขึ้นใหม่ (Reproduction) หรือหาค่าความแข็งแรง(Fitness Value)
2. สลับสายพันธ์ (Crossover)
3. กลายพันธ์ (Mutation)

ซึ่งขั้นตอนวิธีเชิงพันธุกรรมดังกล่าวข้างต้นได้ถูกนำมาขยายความเป็นกระบวนการการทำงานของโปรแกรมเพื่อที่จะวิเคราะห์หาคำตอบโดยการกำหนดค่าเริ่มต้นของปัญหาขึ้นมาซึ่งเรียกว่า สายพันธ์เริ่มต้น จากนั้นผ่านกระบวนการหาคำตอบจากกระบวนการ ทั้งสามข้างต้น ก็จะได้สายพันธุ์ที่สอง จากนั้นตรวจสอบว่าคำตอบที่ได้ เป็นคำตอบที่พอใจหรือยัง หากยังไม่พอใจก็ให้วนไปยังสามกระบวนการข้างต้นอีกครั้ง เกิดเป็นสายพันธุ์ที่สาม ที่สี่ และต่อไปเรื่อยๆ จนกว่าจะได้คำตอบที่พอใจ ขั้นตอนกระบวนการทำงานดังภาพที่ 1



ภาพที่ 1 ขั้นตอนกระบวนการวิธีพันธุกรรม

วิธีการดำเนินการวิจัย

ข้อมูลที่ใช้ในการวิจัย

เพื่อที่จะวัดประสิทธิภาพของโมเดล งานวิจัยครั้งนี้ได้ใช้กลุ่มข้อมูล ทั้งสิ้น 3 กลุ่มข้อมูลซึ่งประกอบด้วยข้อมูลจาก UCI ซึ่งเป็นข้อมูลเปิดสำหรับนำไปทดสอบประสิทธิภาพอัลกอลิทึมทางด้าน AI ได้แก่ฐานข้อมูล Austrain Credit (Anonymous, 2009a) และ German Credit (Anonymous, 2009b) ข้อมูลอีกส่วนหนึ่งเป็นข้อมูลที่ได้จากการเก็บข้อมูลจริงในงานวิจัยของนักศึกษาระดับปริญญาเอก ได้แก่ข้อมูล bankruptcy data (Pietruszkiewicz, 2004) รายละเอียดดังตารางที่ 1

ตารางที่ 1 รายละเอียดข้อมูลที่ใช้

	ตัวอย่าง	แอตทริบิว	คลาส +1	คลาส -1
Austrain Credit	690	14	370	383
German Credit	1000	24	300	700
bankruptcy data	240	30	128	112

เครื่องมือที่ใช้ในการทดลอง

ในงานวิจัยครั้งนี้ ผู้วิจัยได้ใช้เครื่องคอมพิวเตอร์ส่วนบุคคล ซีพียู Intel Celeron M ความเร็ว 1.87 GHz. หน่วยความจำหลัก 1.5 Gb. สำหรับโปรแกรมที่ใช้ในการพัฒนาได้ใช้โปรแกรม Matlab และใช้ LibSVM (Chang & Lin , 2009) ซึ่งเป็น Open Source สำหรับใช้ทดสอบวิธีซัพพอร์ตเว็กเตอร์แมชชีน ส่วน Neuron Network และ Decision Tree ใช้ Toolbox ใน Matlab เป็นตัวทดสอบ

วิธีการประเมินประสิทธิภาพ

สำหรับงานวิจัยนี้ใช้ 5-Fold Cross-Validation ซึ่งเป็นโมเดลที่เหมาะสมสำหรับนำมาประเมินผลความวัดความเที่ยงตรงของโมเดลเพื่อลดปัญหาการไบแอสของข้อมูลที่เลือกมาสำหรับเรียนรู้และทดสอบ

โมเดลที่ใช้เปรียบเทียบเพื่อหาประสิทธิภาพ

ในงานวิจัยนี้ได้ผู้วิจัยทดลองความแม่นยำในการให้คะแนนเครดิตกับโมเดล 4 รูปแบบด้วยกัน ดังตารางที่ 2 สำหรับทุกโมเดลใช้วิธีการ 5-Fold Cross-Validation เพื่อวัดความเที่ยงตรงของโมเดล สำหรับโมเดลใดที่ใช้ วิธีซัพพอร์ตเวกเตอร์แมชชีน เป็นเครื่องมือสำหรับจำแนกข้อมูลและใช้ Radial Basis Function เป็น Kernel Function รายละเอียดแต่ละโมเดลดังตารางที่ 2

ตารางที่ 2 รายละเอียดโมเดลที่ทดลอง

โมเดล	รายละเอียดโมเดลที่ทดลอง	ชื่อย่อ
1	Decision Tree	DT
2	Neuron Network	NN

3	SVM	SVM
4	GA+SVM	GSVM

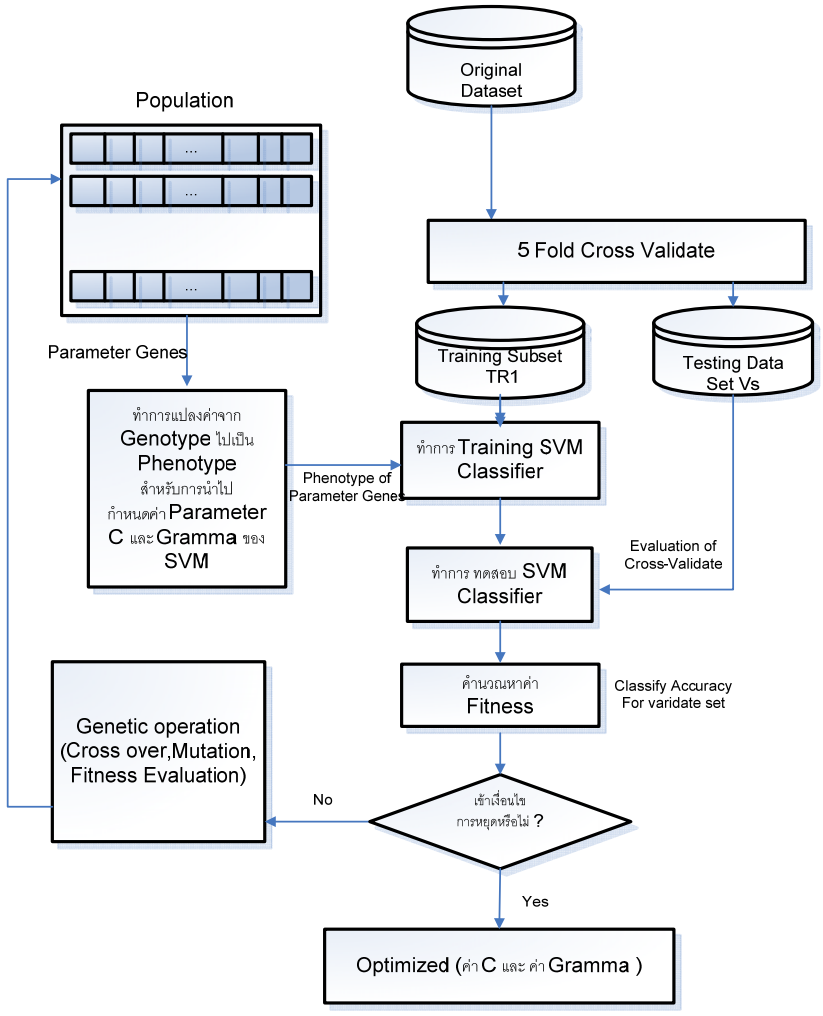
ซึ่งในแต่ละโมเดล มีรายละเอียดดังนี้

1. โมเดลต้นไม้ช่วยตัดสินใจ (Decision Tree) เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปของ ต้นไม้ช่วยตัดสินใจ จากนั้นนำโมเดลที่ได้มาทดสอบกับข้อมูลทดสอบเพื่อประเมินหาประสิทธิภาพความแม่นยำของโมเดล

2. โครงข่ายประสาทเทียม(Neuron Network)มีพื้นฐานมาจากการจำลองการทำงานของสมองมนุษย์ด้วยโปรแกรมคอมพิวเตอร์ จุดมุงหมายของโครงข่ายประสาทเทียมคือต้องการให้คอมพิวเตอร์มีความชาญฉลาดในการเรียนรู้เหมือนที่ มนุษย์มีการเรียนรู้ สามารถฝึกฝนได้ และสามารถนำความรู้และทักษะ โดยในงานวิจัยนี้ได้ใช้ โครงข่ายประสาทเทียมที่มีการกำหนดค่า ฟังก์ชันถ่ายโอน และโครงสร้างเครือข่ายที่แตกต่างกัน 6 รูปแบบเพื่อประเมินหาผลลัพธ์ของโมเดลที่ดีที่สุดสำหรับนำมาเปรียบเทียบประสิทธิภาพ

3. ซัพพอร์ตเวกเตอร์แมชชีน ได้ทำการกำหนดค่าพารามิเตอร์โดยระบุ ค่าพารามิเตอร์ C มีค่าเท่ากับ 100 และค่า พารามิเตอร์แกรมม่าสำหรับ RBF kernel function กำหนดค่าไว้ที่ 1

4. วิธีขั้นตอนเชิงพันธุกรรมร่วมกับซัพพอร์ตเวกเตอร์แมชชีนโดยนำข้อมูลสำหรับสอนมาสอนโมเดล ร่วมกับซัพพอร์ตเวกเตอร์แมชชีนและใช้วิธีการเชิงพันธุกรรมเพื่อหาค่าพารามิเตอร์ (C,γ) ที่เหมาะสมที่สุด โดยมีการกำหนดค่าดังนี้ ออกแบบยีนสำหรับแทนค่า C โดยค่า C ใช้ยีน 10 ตัว ส่วนค่าแกรมม่าใช้ยีน 14 ตัวในการเก็บข้อมูลโดยข้อมูลเป็นแบบไบนารี่ ทำการกำหนด โครโมโซมเริ่มต้นที่ 30 โครโมโซม กำหนดจำนวนเจนเนอเรชั่น (Generation) ไว้ที่ 50 เจนเนอเรชั่น การไขว้ข้าม (Cross over) ใช้การไขว้ข้ามแบบจุดเดียว กำหนดอัตราไว้ที่ 0.6 การเปลี่ยนค่า (Mutation) ใช้วิธีการ แบบสุ่ม 1 จุด กำหนดอัตราไว้ที่ 0.1 การประเมินค่าความแข็ง (Fitness Evaluation) ใช้ค่า Mean Absolute Percent Error (MAPE) ซึ่งเป็นผลลัพธ์ที่ได้จากวิธีซัพพอร์ตเวกเตอร์แมชชีน สำหรับการ คัดเลือกเป็นประชากรรุ่นใหม่ ใช้วิธี การคัดเลือกแบบวง ล้อรูเล็ต (Roulette Wheel Selection) ร่วมด้วยวิธี การคัดเลือกชั้นยอด (Elitism Selection) กำหนดการคัดเลือกชั้นยอด (Elitism Selection) ไว้ที่ 5 โครโมโซม และ การคัดเลือกแบบวง ล้อรูเล็ต (Roulette Wheel Selection) ไว้ที่ 25 โครโมโซม จากนั้นจึงนำโมเดลไปทดสอบความแม่นยำในการจำแนกข้อมูลดังภาพที่ 3



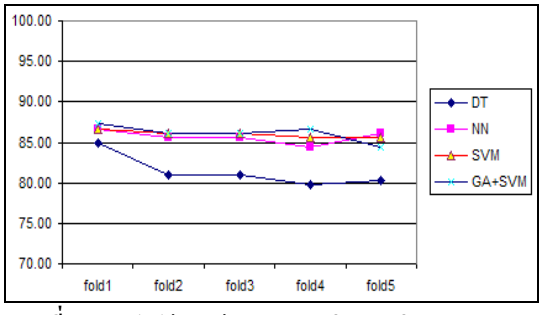
ภาพที่ 3 โมเดลวิธีพันธุกรรมร่วมกับซัพพอร์ตเวกเตอร์แมชชีน

ผลการวิจัยและการอภิปรายผล

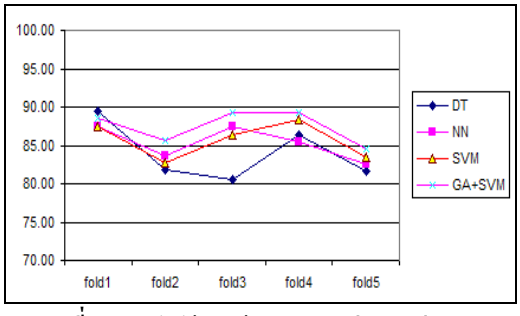
ผลการทดลองที่ได้จากงานวิจัยสำหรับฐานข้อมูล Austrian Credit พบว่า วิธีการซัพพอร์ตเวกเตอร์แมชชีนร่วมกับวิธีการทางพันธุกรรม ให้ผลเฉลี่ยความแม่นยำที่สูงที่สุดเมื่อนำมาทดสอบกับข้อมูล Test Set คือ 87.28% รองลงมาเป็นวิธีซัพพอร์ตเวกเตอร์แมชชีนเพียงอย่างเดียว และวิธีโครงข่ายประสาทเทียม ให้ผลความแม่นยำ 86.71% วิธีต้นไม้ตัดสินใจให้ความแม่นยำเฉลี่ยที่ 84.97% ตามลำดับ ดังตารางที่ 3 และแสดงผลเป็นกราฟดังภาพที่ 4 และ 5

ตารางที่ 3 เปรียบเทียบค่าเฉลี่ยความแม่นยำแต่ละโมเดล

	fold1		fold2		fold3		fold4		fold5	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
DT	89.42	84.97	81.73	80.92	80.58	80.92	86.41	79.77	81.55	80.35
NN	87.50	86.71	83.65	85.55	87.38	85.55	85.44	84.39	82.52	86.13
SVM	87.42	86.71	82.69	86.13	86.41	86.13	88.35	85.55	83.50	85.55
GA+SVM	88.46	**87.28***	85.58	86.13	89.32	86.13	89.32	86.71	84.47	84.39



ภาพที่ 4 กราฟเปรียบเทียบความแม่นยำแต่ละโมเดลสำหรับข้อมูล Test

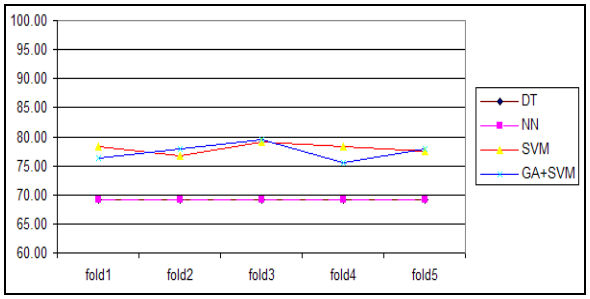


ภาพที่ 5 กราฟเปรียบเทียบความแม่นยำแต่ละโมเดลสำหรับข้อมูล Train

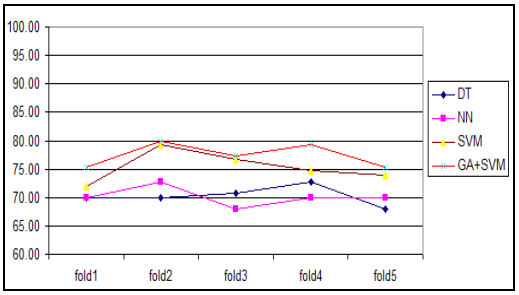
ผลการทดลองที่ได้จากงานวิจัยสำหรับฐานข้อมูล German Credit พบว่า วิธีการซัพพอร์ตเวกเตอร์แมชชีนร่วมกับวิธีการทางพันธุกรรม ให้ผลเฉลี่ยความแม่นยำที่สูงที่สุดให้ผลเฉลี่ยความแม่นยำที่สูงที่สุดเมื่อนำมาทดสอบกับข้อมูล Test Set คือ 79.60% รองลงมาเป็นวิธีซัพพอร์ตเวกเตอร์แมชชีนเพียงอย่างเดียวให้ผลความแม่นยำ 79.20% วิธีโครงข่ายประสาทเทียม และ วิธีต้นไม้ตัดสินใจให้ผลความแม่นยำที่ 69.20% ตามลำดับ ดังตารางที่ 4 และแสดงผลเป็นกราฟดังภาพที่ 6 และ 7 ตามลำดับ

ตารางที่ 4 เปรียบเทียบค่าเฉลี่ยความแม่นยำแต่ละโมเดล

	fold1		fold2		fold3		fold4		fold5	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
DT	70.00	69.20	70.00	69.20	70.67	69.20	72.67	69.20	68.00	69.20
NN	70.00	69.20	72.67	69.20	68.00	69.20	70.00	69.20	70.00	69.20
SVM	72.00	78.40	79.33	76.80	76.67	79.20	74.67	78.40	74.00	77.60
GA+SVM	75.33	76.40	80.00	78.00	77.33	**79.60***	79.33	75.60	75.33	78.00



ภาพที่ 6 กราฟเปรียบเทียบความแม่นยำแต่ละโมเดลสำหรับข้อมูล Test

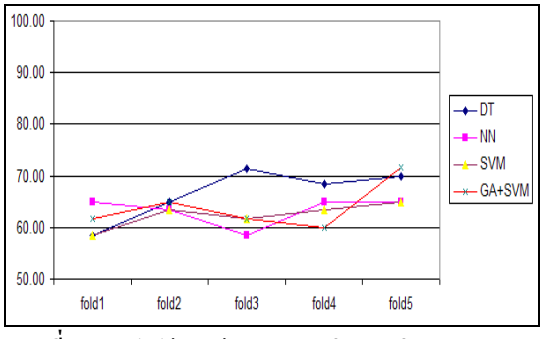


ภาพที่ 7 กราฟเปรียบเทียบความแม่นยำแต่ละโมเดลสำหรับข้อมูล Train

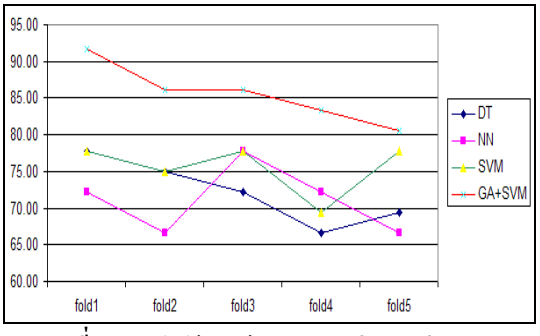
ผลการทดลองที่ได้จากงานวิจัยสำหรับฐานข้อมูล bankruptcy data พบว่า วิธีการซัพพอร์ตเวกเตอร์แมชชีนร่วมกับวิธีการทางพันธุกรรม ให้ผลเฉลี่ยความแม่นยำที่สูงที่สุด เมื่อนำมาทดสอบกับข้อมูล Test Set คือ 71.67% รองลงมาเป็นวิธีต้นไม้ตัดสินใจให้ผลความแม่นยำ 71.33% วิธีซัพพอร์ตเว็กเตอร์แมชชีน และ วิธีโครงข่ายประสาทเทียม ให้ผลความแม่นยำที่ 65.0% ตามลำดับ ดังตารางที่ 5 และแสดงผลเป็นกราฟดังภาพที่ 8 และ 9

ตารางที่ 5 เปรียบเทียบค่าเฉลี่ยความแม่นยำแต่ละโมเดล

	fold1		fold2		fold3		fold4		fold5	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
DT	77.78	58.33	75.00	65.00	72.22	71.33	66.67	68.33	69.44	70.00
NN	72.22	65.00	66.67	63.33	77.78	58.33	72.22	65.00	66.67	65.00
SVM	77.78	58.33	75.00	63.33	77.78	61.67	69.44	63.33	77.78	65.00
GA+SVM	91.67	61.67	86.11	65.00	86.11	61.67	83.33	60.00	80.56	**71.67***



ภาพที่ 8 กราฟเปรียบเทียบความแม่นยำแต่ละโมเดลสำหรับข้อมูล Test



ภาพที่ 9 กราฟเปรียบเทียบความแม่นยำแต่ละโมเดลสำหรับข้อมูล Train

สรุปผลการวิจัย

การวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อสร้างโมเดลที่มีประสิทธิภาพ สำหรับการจำแนกข้อมูล ผลที่ได้แสดงให้เห็นว่าการนำเอาเทคนิคซัพพอร์ตเวกเตอร์แมชชีนสำหรับการจำแนกข้อมูล ร่วมกับวิธีทางพันธุกรรมสำหรับช่วยในการกำหนดค่าพารามิเตอร์ที่เหมาะสม พบว่ามีความแม่นยำมากกว่า วิธีการโครงข่ายประสาทเทียม วิธีการต้นไม้ตัดสินใจรวมทั้งการใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนเพียงอย่างเดียว หากแต่ปัญหาที่พบในงานวิจัยนี้คือ การใช้งานร่วมกันระหว่างวิธีการซัพพอร์ตเวกเตอร์แมชชีนและวิธีทางพันธุกรรมจะใช้เวลาสำหรับการประมวลผลนาน อีกทั้งยังต้องการใช้หน่วยความจำสำหรับการประมวลผลที่สูง

ข้อเสนอแนะ

ในอนาคตผู้วิจัยมีแนวคิดในการนำเอาวิธีการอื่นมาประยุกต์ใช้เพื่อให้เกิดความแม่นยำในการจำแนกข้อมูลที่เพิ่มขึ้น รวมทั้งลดระยะเวลาในการประมวลผลโดยอาจนำเอาวิธีการประมวลผลแบบขนานเข้ามาเพื่อช่วยแก้ปัญหาเกี่ยวกับเรื่องระยะเวลาในการประมวลผลซึ่งใช้เวลานาน

เอกสารอ้างอิง

เดช ธรรมศิริ วาทินี นุ้ยเพียร ภัทราวุฒิ แสงศิริ ภรัณยา อำมฤครัตน์ ฉรงค์ โพธิ และ พยุง มีสัจ. 2551. การให้คะแนนสินเชื่อโดยวิธีการทำเหมืองข้อมูลด้วยเทคนิคโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ. 2nd National Conference on Information Technology 2008.

เดช ธรรมศิริ และ พยุง มีสัจ. 2552. การให้คะแนนสินเชื่อด้วยเทคนิคซับพอร์ตเวกเตอร์แมชชีนและการลดมิติ ข้อมูลด้วยวิธีพีซีเอร่วมกับการหาค่าพารามิเตอร์ที่เหมาะสมด้วยวิธีค้นหาแบบกริช. การประชุมวิชาการระดับชาติด้านวิทยาศาสตร์และเทคโนโลยี ครั้งที่ 2.

Australian Credit Approval Data set. Retrieved February,10, 2008,From http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/australian/.

Banasik, J., Crook, J., and Thomas, L. 2001. Scoring by Usage**.** Journal of the Operational Research Society. 52 : 997–1006.

Boyes, W. J., Hoffman, D. L., and Low, S. A. 1989. An econometric analysis of the bank credit scoring problem. Journal of Econometrics. 40 : 3–14.

Chang-Hoon, M., and Soo-Young. L., 2007. Noise-Robust Speech Recognition Using Top-Down Selective Attention With an HMM Classifier. IEEE Signal processing letters. 14 : 489-491.

Chang, C.,and Lin, C.J., Retrieved August,8, 2008, from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Chen, M. C., and Huang, S. H. 2003. Credit scoring and Rejected instances reassigning through evolutionary computation techniques. Expert Systems with Applications. 24 : 433–441.

Davis, D. B., and Gammerman, A. J., 1992. Machine learning algorithms for credit-card Applications. Journal of Mathematics Applied in Business and Industry. 4 : 43-51.

Fan, A., and Palaniswami, M. 2000. Selecting Bankruptcy predictors using a support vector machine Approach. Proceedings of the International Joint Conference on Neural Networks.

German Credit Data set. Retrieved February,10, 2008, from http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/.

Holland, J. H., 1962. Outline for a logical theory of Adaptive systems. Journal of the Association for Computing Machinery. 3 : 297-314.

Holland, J. H., 1992. Adaptation in Natural and Artificial Systems An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. The MIT Press.

Huanga, Z., Chena, Hsinchun., Hsua, C. J., Chenb W. H., Wuc, S. 2004. Credit rating analysis with support vector machines and neural networks : a market comparative study.

Mina, J. H., and Lee, Y.C. 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications. 28 : 603-614.

Pietruszkiewicz, W. 2004. Application of discrete Predicting structures in an earlywarning expert system for financial distress. Ph.D. Thesis, Szczecin Technical University, Szczecin, 2004.

Sarlija, N., Bensic, M., and Bohacek, Z. 2004. Multinomial Model in consumer credit scoring.10th International Conference on Operational Research Trogir: Croatia.

Shaoning, P., Daijin, K., and Sung Yang, B. 2005. Face membership authentication using SVM classification tree generated by membership-based LLE data partition. Neural Networks. IEEE Transactions. 16 : 436-446.

Shin, K.S., Lee, T.K., and Kim, H.J. 2005. An application of Support vector machines in bankruptcy prediction Model. Expert Systems with Applications. 28 : 127-135.

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

Yang, S.-Y., Huang, Q., Li , L.-L., Ma, C.-Y., Zhang, H., Bai ,R., Teng, Q.-Z., Xiang , M.-L., and Wei, Y.-Q.2009. An integrated scheme for feature selection and parameter setting in the support vector machine modeling and its application to the prediction of pharmacokinetic properties of drugs. Artificial Intelligence in Medicine, 46 : 155-163.

Yu, L., Yue, W., Wang, S., and Lai,K. K. 2009. Support vector machine based multiagent ensemble learning for credit risk evaluation. Expert Systems with Applications, vol. In Press, Corrected Proof.