บทที่ 5
การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ ด้วยวิธีวิเคราะห์การจำแนกประเภท

การวิเคราะห์การจำแนกประเภทถือเป็นวิธีการทางสถิติที่มีจุดประสงค์หลักเพื่อการจำแนกกลุ่มข้อมูลโดยที่ชุดข้อมูลที่นำมาวิเคราะห์จะมีตัวอย่างข้อมูล ที่มีการกำหนดกลุ่มข้อมูล (Categorical) และ มีตัวแปรเป็นค่าต่อเนื่อง (Continuous Variable) นอกจากนี้ จะต้องมีการอนุมานว่าค่าของตัวแปรที่จะนำมาวิเคราะห์มีการแจกแจงแบบปกติ และในการจำแนกกลุ่มข้อมูลหากข้อมูลมีสองกลุ่มจะเรียกการจำแนกกลุ่มข้อมูลแบบนี้ว่า การวิเคราะห์การจำแนกประเภทสองกลุ่ม (Two-Group Discriminant Analysis) ซึ่งวิธีนี้เป็นวิธีการวิเคราะห์จำแนกประเภทในระยะเริ่มต้นที่นักคณิตศาสตร์ชื่อ ฟิชเชอร์ (Fisher) กำหนดแนวคิดไว้ จึงอาจเรียกอีกอย่างว่า (Fisher's Discriminant Analysis) และหากข้อมูลมีมากกว่าสองกลุ่มจะเรียกว่า การวิเคราะห์การจำแนกประเภทเชิงพหุคูณ (Multiple Discriminant Analysis) ทั้งนี้หลักการวิเคราะห์จะอาศัยความสัมพันธ์ของความแตกต่างระหว่างข้อมูลกับค่าเฉลี่ยของข้อมูลที่อยู่ภายในกลุ่มเดียวกัน และ ความแตกต่างระหว่างข้อมูลกับค่าเฉลี่ยของข้อมูลระหว่างกลุ่มในการสร้างโมเดล หรือฟังก์ชันการจำแนกประเภท (Discriminant Function) โดยที่ฟังก์ชันการจำแนกประเภทดังกล่าว จะอยู่ในรูปของผลรวมเชิงเส้น (Linear Combination) ของตัวแปร ที่จะทำให้ความแตกต่างของค่าเฉลี่ยของข้อมูลที่อยู่ภายในกลุ่มข้อมูลแต่ละกลุ่มนั้น มีค่าน้อยที่สุด ขณะเดียวกันก็ทำให้ความแตกต่างของค่าเฉลี่ยของข้อมูลระหว่างกลุ่ม มีค่ามากที่สุด ซึ่งจำนวนของฟังก์ชันการจำแนกประเภททั้งหมด จะมีจำนวนเท่ากับจำนวนกลุ่มของข้อมูลลบออกหนึ่งเสมอ และ ฟังก์ชันการจำแนกประเภทที่ได้ จะถูกนำไปใช้ในการทำนายประเภทของข้อมูล

ในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ เพื่อจำแนกประเภทของข้อมูล จะอาศัยข้อมูลชุดเดียวกับที่ใช้ในบทที่ผ่านมา นั่นคือชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของมะเร็งลิวคีเมีย เพื่อที่จะจำแนกและทำนายกลุ่มของผู้ป่วยที่เป็นโรคมะเร็ง โดยใช้ ค่าการแสดงออกของยีนต่างๆ เป็นค่าของตัวแปรในการสร้างโมเดลจำแนกประเภท ทั้งนี้เนื่องจากกลุ่มของข้อมูลที่นำมาวิเคราะห์มี 2 กลุ่มข้อมูล วิธีการวิเคราะห์จำแนกประเภทที่จะนำเสนอในวิทยานิพนธ์ฉบับนี้จึงมุ่งเน้นไปที่ การวิเคราะห์การจำแนกประเภทสองกลุ่มเท่านั้น นอกจากนี้ เนื่องจากยีนที่เป็นตัวแปรต่างๆ นั้นมีจำนวนมาก การวิเคราะห์ข้อมูลจำเป็นต้องเลือกตัวแปรหรือสร้างตัวแปรใหม่ ให้มีจำนวนน้อยและเพียงพอที่จะเป็นตัวแปรที่มีความสำคัญกับการจำแนกกลุ่มข้อมูล (Discriminator Variables) ซึ่งในเทคนิคการวิเคราะห์จำแนกประเภทวิธีการเลือกตัวแปรสำหรับจำแนกประเภทก็ถือเป็นจุดประสงค์หนึ่งของการวิเคราะห์ แต่ใน

วิทยานิพนธ์ฉบับนี้ผู้วิจัยไม่ได้ลงในรายละเอียดและไม่ได้นำวิธีการดังกล่าวมาใช้ งานวิจัยนี้จึงเสนอวิธีการอื่นๆ จากการศึกษาในงานวิจัยต่างๆ และ จากทฤษฎีการวิเคราะห์ข้อมูลดังที่นำเสนอไป มาใช้แทน ซึ่งได้แก่ การเลือกตัวแปรโดยวิธีวิเคราะห์ค่าเอนโทรปี (Entropy) การเลือกตัวแปรโดยอาศัยการวิเคราะห์ค่าความแปรปรวน และการสร้างตัวแปรใหม่โดยวิธีวิเคราะห์องค์ประกอบหลัก

5.1 หลักการของวิธีวิเคราะห์การจำแนกประเภท

5.1.1 โมเดลการวิเคราะห์

กระบวนการวิเคราะห์การจำแนกประเภทข้อมูลแบ่งวิธีการวิเคราะห์เป็น 3 ขั้นตอนคือ การเลือกตัวแปรที่มีความสำคัญกับการจำแนกกลุ่มข้อมูล การหาฟังก์ชันเชิงเส้นสำหรับการจำแนกกลุ่มข้อมูล และ การจำแนกประเภทของข้อมูล ซึ่งจากขั้นตอนเหล่านี้ เมื่อพิจารณาที่วิธีการในรายละเอียด จะพบว่าในแต่ละขั้นตอนสามารถวิเคราะห์ แยกเป็นอิสระออกจากกันได้ ทำให้เราสามารถใช้วิธีการอื่นในการวิเคราะห์สำหรับบางขั้นตอน ซึ่งขั้นตอนที่เราให้ความสนใจ และตัดออกมาวิเคราะห์ต่างหากนั่นก็คือ การเลือกตัวแปรที่มีความสำคัญกับการจำแนกประเภท

ดังนั้นในการอธิบายเชิงทฤษฎีต่อไปนี้จะถือว่าตัวแปรที่นำมาวิเคราะห์เป็นตัวแปรที่มีความสำคัญกับการจำแนกประเภทเท่านั้น

จากเมตริกซ์ข้อมูล

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} \tag{61}$$

โดยที่ X เป็นเมตริกซ์ข้อมูล มี m เป็นจำนวนของตัวแปร ที่มีความสำคัญต่อการจำแนกประเภท และ n เป็นจำนวนของกลุ่มตัวอย่าง และ y เป็น เวกเตอร์ของกลุ่มข้อมูลที่ต้องการคำนายค่า ซึ่ง

$$y=[y_1, y_2, ..., y_i, ..., y_n]' \tag{62}$$

ฟังก์ชันการจำแนกประเภทซึ่งอยู่ในรูปของผลรวมเชิงเส้นของตัวแปรจะเขียนได้ดังสมการ

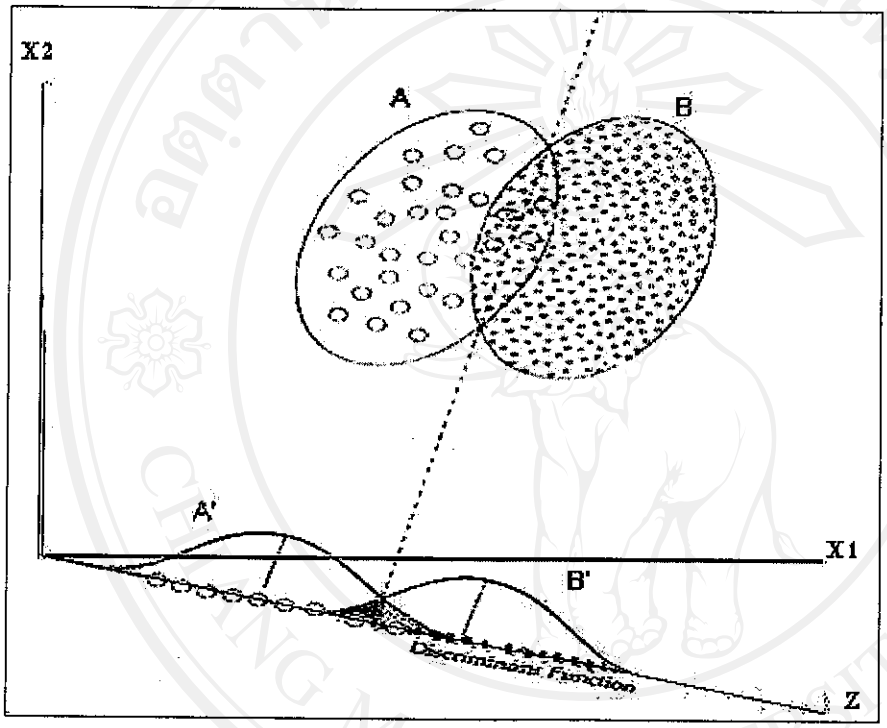
$$Z_{ik}=\omega_1 X_{i1}+\omega_2 X_{i2}+...+\omega_j X_{ij}+...+\omega_m X_{im} \tag{63}$$

โดยที่ Z_{ik} คือ ค่าคะแนนจำแนกประเภท (Discriminant Scores) ของฟังก์ชันการจำแนกประเภทตัวที่ k และตัวอย่างข้อมูลตัวที่ i

ω_j คือ น้ำหนักการจำแนกประเภท (Discriminant Weight) ของตัวแปรตัวที่ j

X_{ij} คือ ค่าของตัวแปรตัวที่ j และตัวอย่างข้อมูลตัวที่ i

จากสมการ (63) ฟังก์ชันการจำแนกประเภท สามารถอธิบายได้โดยอาศัยแผนภาพ ดังรูป 5.1

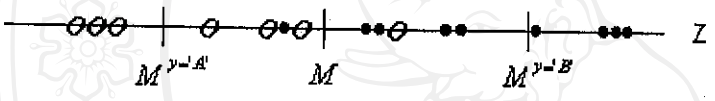


รูป 5.1 กราฟแสดงตัวอย่างของผลการวิเคราะห์การจำแนกประเภทกรณีที่ข้อมูลมีสองกลุ่ม

จากรูป 5.1 เป็นกราฟที่แสดงให้เห็นถึงการแจกแจงข้อมูลของตัวแปร X1 และ X2 และลักษณะของฟังก์ชันการจำแนกประเภท Z โดยจุดแต่ละจุดบนแกน X1 และ X2 แทนค่าของข้อมูล ซึ่งแบ่งออกเป็นสองกลุ่มโดยจุดใหญ่ๆ แสดงถึงกลุ่ม A ส่วน จุดเล็ก คือกลุ่มข้อมูล B ผลจากการวิเคราะห์การจำแนกประเภทจะทำให้ได้ฟังก์ชันการจำแนกประเภท Z ซึ่งค่า คะแนนการจำแนกประเภท จะเป็นค่าที่เกิดจากผลรวมเชิงเส้นของตัวแปร X1 และ X2 ดังนั้นกลุ่มของข้อมูลที่ได้ใหม่ A' และ B' จึงแสดงอยู่ในแกนของ Z ซึ่งเมื่อลากเส้นตั้งฉากตรงจุดกึ่งกลางบนแกน Z หรือจุดที่ใช้ในการแบ่งกลุ่มผ่านไปยังกลุ่มข้อมูลในแกนของตัวแปรเดิมทั้งสองตัวแปร เส้นดังกล่าวจะผ่านตรงส่วนของข้อมูล ณ ตำแหน่งที่แบ่งกลุ่มของข้อมูลได้ดีที่สุด ดังนั้นสิ่งนี้เองที่ทำให้เห็นประโยชน์ของฟังก์ชันการจำแนก

ประเภท ในการที่จะลดมิติของข้อมูลจากตัวแปรที่มีหลายๆ ตัวแปร ให้อยู่ในมิติเดียวในลักษณะของฟังก์ชันการจำแนกประเภท ซึ่งการที่ตัวแปรของข้อมูลมีจำนวนมากกว่า 2 ตัวแปร การพิจารณาจำแนกกลุ่มของข้อมูลโดยตรงนั้นจะไม่สามารถทำได้ แต่การลดมิติของข้อมูลให้เหลือเพียงมิติเดียวในลักษณะของฟังก์ชันการจำแนกประเภทจะสามารถทำได้ นอกจากนี้จากกราฟ จะเห็นได้ว่า การที่ฟังก์ชันการจำแนกประเภทนั้นจะแบ่งข้อมูลได้ดี จะต้องทำให้ ส่วนที่คาบเกี่ยวกัน(Overlap)ของข้อมูลบนแกน Z มีพื้นที่น้อยที่สุด นั่นคือจะต้องทำให้ความแตกต่างของค่าเฉลี่ยของข้อมูลบนแกนดังกล่าวที่อยู่คนละกลุ่มกันมีค่าสูงที่สุด ขณะเดียวกันก็ต้องทำให้ความแตกต่างของค่าเฉลี่ยของข้อมูลที่อยู่ในกลุ่มเดียวกันมีค่าน้อยที่สุดด้วย หลักการนี้ ถือเป็นหัวใจหลักของการวิเคราะห์การจำแนกประเภทที่ใช้สำหรับการประมาณค่าน้ำหนักการจำแนกประเภท (ω_j) เพื่อสร้างเป็นฟังก์ชัน ในสมการ(63)

ฟังก์ชันการจำแนกประเภท Z เมื่อนำค่าคะแนนการจำแนกประเภทในทุกๆตัวอย่างข้อมูล มาพล็อตลงในกราฟ จะแสดงได้ ดังรูป 5.2



รูป 5.2 ตัวอย่างกราฟคะแนนการจำแนกประเภทจากผลการวิเคราะห์จำแนกประเภท

จากรูป 5.2 เป็นกราฟแสดงการแจกแจงของค่าคะแนนการจำแนกประเภทในสองกลุ่มตัวอย่าง

โดยที่ M คือค่าเฉลี่ย (Mean) ของค่าคะแนนการจำแนกประเภทในกลุ่มตัวอย่างข้อมูลทั้งหมด

$M^{y='A'}$ คือค่าเฉลี่ยของค่าคะแนนการจำแนกประเภทในตัวอย่างข้อมูลที่อยู่กลุ่ม A และ

$M^{y='B'}$ คือค่าเฉลี่ยของค่าคะแนนการจำแนกประเภทในตัวอย่างข้อมูลที่อยู่กลุ่ม B

จากกราฟจะเห็นว่าข้อมูลจะแบ่งเป็นกลุ่มได้ดี กรณีที่ความแตกต่างของค่าเฉลี่ยของคะแนนการจำแนกประเภทกับค่าคะแนนการจำแนกประเภทในข้อมูลแต่ละตัว ซึ่งอยู่ในกลุ่มเดียวกันมีค่าน้อยนั่นหมายความว่าข้อมูลที่อยู่ในกลุ่มเดียวกันจะต้องอยู่ติดกันให้มากที่สุดนั่นเอง นอกจากนี้ ในขณะเดียวกันข้อมูลที่อยู่กันคนละกลุ่มจะต้องอยู่ห่างกันมากๆ นั่นคือ ค่าเฉลี่ยของคะแนนการจำแนกประเภทที่อยู่ในกลุ่มใดกลุ่มหนึ่ง จะต้องแตกต่างกับค่าเฉลี่ยของคะแนนการจำแนกประเภทที่อยู่อีกกลุ่มหนึ่งมากๆ ในการวัดค่าความแตกต่างของค่าเฉลี่ยของข้อมูลเหล่านี้จะทำให้อยู่ในรูปของ ผลรวมกำลังสอง(Sum of Square) เพื่อลดปัญหาเรื่องของเครื่องหมายที่แตกต่างกัน

ผลรวมกำลังสอง ของคะแนนการจำแนกประเภท จะแยกได้เป็น 3 กรณีดังนี้

(1) ผลรวมกำลังสองของค่าคะแนนการจำแนกประเภททั้งหมด (SS_t)

(2) ผลรวมกำลังสองของค่าคะแนนการจำแนกประเภทในกลุ่มเดียวกัน (SS_w)

(3) ผลรวมกำลังสองของค่าคะแนนการจำแนกประเภท ระหว่างกลุ่ม (SS_b)

สูตรการคำนวณแสดงได้ดังสมการต่อไปนี้

$$SS_t = \sum_{i=1}^{n} (Z_i - M)^2 \tag{64}$$

$$SS_w = \sum_{g=1}^{n_g} \sum_{i=1}^{n_{y=g}} (Z_i^{y=g} - M^{y=g})^2 \tag{65}$$

$$SS_b = \sum_{g=1}^{n_g} (M^{y=g} - M)^2 \tag{66}$$

จากค่าของผลรวมกำลังสองในสมการ (66)-(66) ฟังก์ชันการจำแนกประเภทที่จะจำแนกกลุ่มข้อมูลได้ดี จะต้องทำให้ ค่า SS_b มีค่ามากที่สุด และ ค่า SS_w มีค่าน้อยที่สุด ถ้ากำหนดให้ λ แทนอัตราส่วนของ SS_b และ SS_w นั่นคือ $\lambda = \frac{SS_b}{SS_w}$ ดังนั้นฟังก์ชันการจำแนกประเภทที่จะแยกกลุ่มข้อมูลได้ดีที่สุดนั้นจะต้องให้ค่า λ มากที่สุด

จากเวกเตอร์ของกลุ่มข้อมูล y ในสมการ (62) จะมีค่าเป็นกลุ่มข้อมูล G โดยที่ $G = 1, 2, ..., n_G$ ซึ่ง n_G เป็นจำนวนของกลุ่มข้อมูล และ $n_{y=G}$ เป็นจำนวนของตัวอย่างข้อมูลที่อยู่ในกลุ่ม G

และจากเมตริกซ์ข้อมูล X ในสมการ (61) เป็นเมตริกซ์ข้อมูลที่มีขนาด $n \times m$ โดยที่ n เป็นจำนวนตัวอย่างข้อมูลทั้งหมด ส่วน m เป็นจำนวนตัวแปร เขียนใหม่ให้อยู่ในรูปของ เวกเตอร์ ได้ตามสมการ (67)

$$X = [x_1^{y=G}, ..., x_i^{y=G}, ..., x_n^{y=G}]^T \tag{67}$$

โดยที่

$$x_i^{y=G} = [x_{i1}^{y=G}, ..., x_{i2}^{y=G}, ..., x_{im}^{y=G}]^T$$

การสร้างฟังก์ชันการจำแนกประเภท เพื่อที่จะทำให้ได้ ค่า λ มากที่สุดนั้น จะเริ่มจากการประมาณค่าน้ำหนักการจำแนกประเภทโดยอาศัยเวกเตอร์ของข้อมูล จากสมการ(67)

กำหนดให้ T เป็นเมตริกซ์ผลรวมกำลังสองของข้อมูลทั้งหมด (Total Sum of Square Matrix), B เป็นเมตริกซ์ผลรวมกำลังสองของค่าของข้อมูลระหว่างกลุ่ม (Between Group Sum of Square) และ W เมตริกซ์ผลรวมกำลังสองของค่าของข้อมูลภายในกลุ่ม (Within Group Sum of ทั้ง

3 เวกเตอร์จะมีขนาด $m \times m$ ซึ่งจากสมการ (63) กำหนดให้ ω เป็นเวกเตอร์ของค่าน้ำหนักการจำแนกประเภทขนาด $m \times 1$ และ Z เป็นเวกเตอร์ของคะแนนการจำแนกประเภทขนาด $n \times 1$
จะหา T, W และ B ได้ดังนี้

$$T = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$$

หรือ

$$T = X^T X \tag{68}$$

โดยที่

$$\bar{x} = (1/n) \sum_{i=1}^{n} x_i \tag{69}$$

จากสมการ(69) $\bar{x}$ เป็นเวกเตอร์ของค่าเฉลี่ย (Vector Mean) ของข้อมูลทั้งหมด (Grand Mean) ซึ่ง

$$\bar{x} = \left[\bar{x}_1, \quad \bar{x}_2, \quad \ldots, \quad \bar{x}_m \right]^T \tag{70}$$

นอกจากนี้

$$W = \sum_{G=1}^{n_G} S^{y=G} \tag{71}$$

โดยที่

$$S^{y=G} = \sum_{j=1}^{n_{y=G}} (x_j^{y=G} - \bar{x}^{y=G})(x_j^{y=G} - \bar{x}^{y=G})^T \tag{72}$$

และ

$$\bar{x}_i^{y=G} = (1/n_{y=G}) \sum_{j=1}^{n_{y=G}} x_{ji}^{y=G} \tag{73}$$

จากสมการ(73) $\bar{x}^{y=G}$ เป็นเวกเตอร์ของค่าเฉลี่ย (Vector Mean) ของข้อมูลกลุ่มที่ G ซึ่ง

$$\bar{x}^{y=G} = \left[\bar{x}_1^{y=G}, \quad \bar{x}_2^{y=G}, \quad \ldots, \quad \bar{x}_m^{y=G} \right]^T \tag{74}$$

สุดท้ายจะได้

$$B = \sum_{G=1}^{n_G} \sum_{j=1}^{n_{y=G}} (\bar{x}^{y=G} - \bar{x})(\bar{x}^{y=G} - \bar{x})^T = \sum_{G=1}^{n_G} n_{y=G} \bar{x}^{y=G} (\bar{x}^{y=G})^T - n\bar{x}\bar{x}^T \tag{75}$$

เมตริกซ์ผลรวมกำลังสองของค่าของข้อมูล ทั้ง 3 เมตริกซ์ นี้จะมีความสัมพันธ์กันดังสมการ

$$T = B + W \tag{76}$$

ต่อไปจะทำการหาเวกเตอร์ของค่าถ่วงน้ำหนัก ω เพื่อนำไปใช้ในการสร้างฟังก์ชันการจำแนกประเภท ดังต่อไปนี้

จากฟังก์ชันการจำแนกประเภท ในสมการ(63) นำมาเขียนให้อยู่ในรูปของผลคูณเมตริกซ์ดังนี้

$$Z = X\omega \tag{77}$$

และจาก ค่าผลรวมกำลังสองของค่าคะแนนจำแนกประเภททั้งหมด (SS_t) ในสมการ (64) นำมาเขียนให้อยู่ในรูปของผลคูณเมตริกซ์ดังนี้

$$\begin{aligned} SS_t &= Z^T Z = (X\omega)^T (X\omega) \\ &= \omega^T X^T X \omega \end{aligned} \tag{78}$$

จากสมการ(68) เนื่องจาก $T = X^T X$ ดังนั้น

$$SS_t = \omega^T T \omega$$

และจากความสัมพันธ์ในสมการ (76) จะได้

$$\begin{aligned} SS_t &= \omega^T (B + W)\omega \\ &= \omega^T B\omega + \omega^T W\omega \end{aligned} \tag{79}$$

จากสมการ (79) เมื่อเทียบกับสมการ (65) และ (66) จะเห็นว่า $\omega^T B\omega$ คือเมตริกซ์ SS_b และ $\omega^T W\omega$ คือเมตริกซ์ SS_w

จากเงื่อนไขที่กำหนดไว้ในตอนต้น กล่าวว่าฟังก์ชันการจำแนกประเภทที่จะแยกกลุ่มข้อมูลได้ดีที่สุดนั้นจะต้องให้ค่า λ มากที่สุดซึ่ง $\lambda = \frac{SS_b}{SS_w}$

ดังนั้น

$$\lambda = \frac{\omega^T B\omega}{\omega^T W\omega} \tag{80}$$

เพื่อที่จะประมาณค่า ω จึงต้องทำให้ λ มีค่ามากที่สุด ซึ่งจะทำได้โดย แก้สมการอนุพันธ์ของ λ เทียบ ω ให้มีค่าเท่ากับศูนย์ ดังนี้

$$\frac{\partial \lambda}{\partial \omega} = \frac{2(B\omega)(\omega^T W\omega) - 2(\omega^T B\omega)(W\omega)}{(\omega^T W\omega)^2} = 0 \qquad (81)$$

จากสมการ (81) หารด้วย $\omega W^T\omega$ จะทำให้

$$\frac{2(B\omega - \lambda W\omega)}{(\omega^T W\omega)^2} = 0$$

และ (82)

$$(W^{-1}B - \lambda I)\omega = 0$$

ต่อไปจะแก้สมการหา ω โดยที่คำตอบไม่เท่ากับศูนย์ (Nontrivial Solution) จากสมการ

$$|W^{-1}B\text{-}\lambda I| \neq 0 \qquad (83)$$

จากสมการ(83) จะได้ว่า **ค่าของ λ ที่กล่าวในตอนต้นนั้น คือค่า ไอเกนแวลยู ของเมตริกซ์ $W^{-1}B$ และ ω ก็คือไอเกนเวกเตอร์ที่ได้จากค่าของไอเกนแวลยูที่มีค่ามากที่สุด นั่นเอง**

จากฟังก์ชันการจำแนกประเภทในสมการ (63) ค่าน้ำหนักการจำแนกประเภท ω จะถูกนำไปแทนที่ในสมการ และผลจากผลรวมเชิงเส้นของตัวอย่างข้อมูลทั้งหมดจะทำให้ได้ เวกเตอร์ของของค่าคะแนนการจำแนกประเภท Z_i ที่กลุ่มข้อมูลมีการจำแนกได้ดีที่สุด นอกจากนี้ ค่าคะแนนการจำแนกประเภทดังกล่าวยังจะถูกนำไปใช้เป็นตัวตัดสิน ในการทำนายประเภทกลุ่มของข้อมูล ซึ่งถืออีกเป้าหมายหนึ่งของการวิเคราะห์

สำหรับวิธีการที่ใช้ในการทำนายกลุ่มของข้อมูลที่ไม่ทราบค่า เพื่อจำแนกกลุ่มข้อมูลนั้น มีวิธีการที่นำเสนอใหญ่ๆ อยู่ 2 วิธีการคือ วิธีตัดจากค่ากึ่งกลาง (Cutoff Value Method) และ วิธีการตัดสินใจทางสถิติ (Statistical Decision Theory) ในงานวิจัยนี้จะนำเสนอเพียง วิธีตัดจากค่ากึ่งกลางสำหรับการวิเคราะห์ ดังต่อไปนี้

วิธีการทำนายกลุ่มข้อมูลโดยวิธีตัดจากค่ากึ่งกลาง มีแนวคิดว่า เงื่อนไขของการจำแนกประเภทข้อมูลเพื่อให้จำแนกข้อมูลได้ดีที่สุดคือ จะต้องทำให้จำนวนของความผิดพลาดของการจำแนกกลุ่มข้อมูลเกิดขึ้นน้อยที่สุด(Minimum the Number of Incorrect Classification) ซึ่งมีวิธีการวิเคราะห์ดังนี้

1) หาค่าคะแนนการจำแนกประเภทของข้อมูลที่ทราบกลุ่ม (Training Data)

2) หาค่ากึ่งกลางข้อมูลเพื่อกำหนดให้เป็นจุดแบ่ง (Cutoff Value) จากคะแนนการจำแนกประเภท ดังสมการ

$$\text{Cut of Value} = \frac{n_A m^A + n_B m^B}{n_A + n_B} \tag{84}$$

จากสมการจะกำหนดให้

n_A, n_B คือจำนวนของข้อมูลในกลุ่ม A และ B ตามลำดับ

m^A, m^B คือค่าเฉลี่ยคะแนนการจำแนกประเภทของข้อมูลในกลุ่ม A และ B ตามลำดับ

3) หาค่าคะแนนการจำแนกประเภทในข้อมูลทดสอบ (Testing Data) Z^* จากสมการ (63) โดยใช้ค่าน้ำหนักการจำแนกประเภท จากการประมาณค่าในข้อมูลที่ทราบกลุ่มแล้ว ดังสมการ

$$Z^*_i = \omega_1 X_{i1} + \omega_2 X_{i2} + \ldots + \omega_m X_{im} \tag{85}$$

4) พิจารณา ค่าเฉลี่ยคะแนนการจำแนกประเภทของข้อมูล m^A, m^B

ทำนายกลุ่มข้อมูลดังเงื่อนไขต่อไปนี้

41) กรณีที่ $m^A > m^B$

จะจำแนกประเภทข้อมูลได้ว่าเป็นกลุ่ม A ถ้า Z^* > cut of value ในทางตรงกันข้ามจะเป็นกลุ่ม B

42) กรณีที่ $m^A < m^B$

จะจำแนกประเภทข้อมูลได้ว่าเป็นกลุ่ม B ถ้า Z^* > cut of value ในทางตรงกันข้ามจะเป็นกลุ่ม A

5.1.2 วิธีการเลือกตัวแปรและการสร้างตัวแปรที่มีความสำคัญต่อการจำแนกประเภท

การเลือกตัวแปรและการสร้างตัวแปรที่มีความสำคัญกับการจำแนกประเภท จะนำเสนอใน 3 วิธีการคือ การเลือกตัวแปรโดยวิธีวิเคราะห์ค่าเอนโทรปี (Entropy) การเลือกตัวแปรโดยการวิเคราะห์ค่าความแปรปรวน และ การสร้างตัวแปรใหม่โดยวิธีวิเคราะห์องค์ประกอบหลัก ซึ่งแต่ละวิธีการอธิบายได้ดังต่อไปนี้

- การเลือกตัวแปรโดยวิธีวิเคราะห์ค่าเอนโทรปี (Entropy)

ค่าเอนโทรปีของข้อมูลหมายถึงค่าความฟุ้งกระจายของข้อมูลในแต่ละตัวแปร ซึ่งตั้งสมมุติฐานไว้ว่า ข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีค่าไม่แตกต่างกันหรือมีค่าความฟุ้งกระจายของข้อมูลน้อย ดังนั้น หากตัวแปรหนึ่งๆ มีกลุ่มข้อมูลมากกว่า 1 กลุ่มแล้ว ผลรวมของค่าความฟุ้งกระจายของข้อมูลทุกๆ กลุ่มในตัวแปรนั้นจะต้องมีค่าน้อยตามไปด้วย ผลก็คือ การเลือกตัวแปรที่มีความสำคัญกับการจำแนกกลุ่มของข้อมูลนั้น จะสามารถพิจารณาจาก ค่าผลรวมของค่าเอนโทรปีนี้ได้ นั่นคือ หากค่าผลรวมของค่าของเอนโทรปีในตัวแปรดังกล่าวมีค่าน้อย ตัวแปรดังกล่าวก็จะใช้ในการจำแนกประเภทของข้อมูลได้ดี แต่หากค่าผลรวมของค่าของเอนโทรปีในตัวแปรดังกล่าวมีค่ามาก ตัวแปรดังกล่าวก็จะใช้ในการจำแนกประเภทของข้อมูลได้ไม่ดี

นอกจากนี้เนื่องจากจำนวนของข้อมูลที่อยู่ในแต่ละตัวกลุ่มนั้นอาจมีจำนวนไม่เท่ากัน ผลรวมของค่าความฟุ้งกระจายจึงต้องวัดออกมาเป็นค่าประมาณ จะเรียกค่าประมาณนี้ว่า ค่าประมาณของค่าเอนโทรปี การพิจารณาตัวแปรโดยอาศัยค่าผลรวมของค่าเอนโทรปีดังกล่าวจึงต้องใช้ค่าประมาณของค่าเอนโทรปีแทน สำหรับรายละเอียดในการวิเคราะห์ค่าเอนโทรปีอธิบายได้ดังต่อไปนี้

กำหนดให้ X เป็นเวกเตอร์ของข้อมูล ซึ่งข้อมูลที่เป็นไปได้คือ D ทั้งนี้จำนวนของ D ที่เป็นไปได้คือ n_D และ p(X= D) เป็นความน่าจะเป็นที่ X จะมีค่าเท่ากับ D โดยที่

$$X= [0.63, 0.63, 0.63, 0.63, 0.63, 0.63] \quad (86)$$

จะเห็นว่าข้อมูลในตัวแปร X มีค่าเดียวคือ 0.63

ดังนั้น p(X=0.63) =1 นั่นแสดงว่าค่าของข้อมูลในตัวแปร X นั้นมีค่าที่แน่นอน คือ 0.63 ไม่สามารถเป็นค่าอื่นได้ หรืออีกความหมายหนึ่งคือ ความไม่แน่นอนของข้อมูลในตัวแปร X นั้นมีค่าเป็น 0 นั่นเอง ความไม่แน่นอนของข้อมูลนี้ เรียกอีกอย่างหนึ่งว่า ความฟุ้งกระจายของข้อมูล
เมื่อกำหนดค่าของข้อมูลให้ตัวแปร X ใหม่ดังนี้

$$X= [0.63, 0.42, 0.63, 0.42, 0.42, 0.63] \quad (87)$$

จะเห็นว่าจากข้อมูลนี้มี p(X=0.63) = 0.5 และ p(X=0.42) = 0.5 ดังนั้น จากความน่าจะเป็นที่ได้ แสดงให้เห็นว่าข้อมูลมีความไม่แน่นอนที่จะเป็น 0.63 หรือ 0.42 กันแน่ นั่นคือ ตัวแปร X มีความฟุ้งกระจายของข้อมูลสูง

การวัดค่าความฟุ้งกระจายของข้อมูล ($I(X)$) ในตัวแปร X สามารถวัดได้จากสมการ (88)

$$I(X) = -\sum_{i=1}^{n_D} p(X = D_i) \log_2 \left(p(X = D) \right) \quad (88)$$

จากสมการ (86) คำนวณหาค่า $I(X)$ ได้ดังนี้

$$\begin{aligned} I(X) &= -p(X=0.63)\log_2(p(0.63)) \\ &=1\log_2(1) \\ &=0 \end{aligned} \tag{89}$$

จากสมการ (89) แสดงให้เห็นว่าข้อมูลในสมการ (86) มีค่าความฟุ้งกระจายของข้อมูลเป็น 0

จากสมการ (87) คำนวณหาค่า $I(X)$ ได้ดังนี้

$$\begin{aligned} I(X) &= -(p(X=0.63)\log_2(p(0.63)) + p(X=0.42)\log_2(p(0.42))) \\ &=0.5\log_2(0.5)+ 0.5\log_2(0.5) \\ &=1 \end{aligned} \tag{90}$$

จากสมการ (90) แสดงให้เห็นว่าข้อมูลในสมการ (87) มีค่าความฟุ้งกระจายของข้อมูลเป็น1

ผลที่ได้จากสมการ (89) แสดงให้เห็นกรณีที่ข้อมูลมีค่าไม่แตกกันมาก จะทำให้ค่าความฟุ้งกระจายของข้อมูลน้อย ส่วนในสมการ (90) แสดงให้เห็นว่า เมื่อข้อมูลมีความแตกต่างกันอยู่ภายในมาก ค่าความฟุ้งกระจายของข้อมูลก็จะมีค่ามากตามไปด้วย ด้วยเหตุนี้หากพิจารณาข้อมูลที่อยู่ภายในกลุ่มเดียวกัน ข้อมูลดังกล่าวจะมีค่าไม่แตกกันมาก ซึ่งจะทำให้ค่าความฟุ้งกระจายของข้อมูลที่อยู่ภายในกลุ่มเดียวกันมีค่าไม่แตกต่างกันด้วย

ดังนั้น ในการเลือกตัวแปรที่มีความสำคัญกับการจำแนกกลุ่มข้อมูล จะต้องพิจารณาค่าความฟุ้งกระจายของข้อมูลในแต่ละกลุ่มด้วย นั่นคือ ตัวแปรที่จะจำแนกกลุ่มข้อมูลได้ดีนั้น จะต้องมีค่าความฟุ้งกระจายของข้อมูลในทุกๆ กลุ่มน้อยที่สุด ซึ่งจะทำให้ผลรวมของค่าความฟุ้งกระจายของข้อมูลทุกๆ กลุ่มในตัวแปรนั้น มีค่าน้อยตามไปด้วย ทั้งนี้เนื่องจากจำนวนของข้อมูลในแต่ละกลุ่มนั้นมีไม่เท่ากัน การคำนวณค่าผลรวมจึงต้องมีการปรับน้ำหนักจากจำนวนของข้อมูลในแต่ละกลุ่มด้วย ซึ่งค่าผลรวมของค่าความฟุ้งกระจายดังกล่าวหลังจากปรับค่าน้ำหนักแล้ว ในที่นี้จะเรียกว่า ค่าประมาณของค่าเอนโทรปี (Expectation Entropy) ของตัวแปร X ให้สัญลักษณ์เป็น $E(X)$ ซึ่งคำนวณได้ดังนี้

$$E(X) = \sum_{i=1}^{n_C}\left(\left(\frac{n_{X\in C_i}}{n}\right).I(X_{X\in C_i}) \right) \tag{91}$$

จากสมการ กำหนดให้ $n_{X\in C_i}$ คือจำนวนของข้อมูลในตัวแปร X เป็นสมาชิกของกลุ่ม C_i ส่วน n_C คือจำนวนของกลุ่มข้อมูล และ $I(X_{X\in C_i})$ คือค่าความฟุ้งกระจายของข้อมูลหรือค่าเอนโทรปี ของข้อมูลที่อยู่ในตัวแปร X และเป็นสมาชิกของกลุ่ม C_i

ผลจากสมการ กรณีที่ตัวแปร X มีหลายตัวแปร เช่นอยู่ในรูปของเมตริกซ์หลายๆ ตัวแปร ดังสมการ (61) การเลือกตัวแปรที่มีผลต่อการจำแนกประเภทนั้นก็จะอาศัย การเรียงลำดับค่า $E(X)$ แล้วเลือกตัวแปรที่มีค่าน้อยๆ ทั้งนี้จำนวนตัวแปรที่เลือกนั้นควรจะเป็นกี่ตัว ขึ้นกับ วิจารณญาณของผู้วิจัย

จากวิธีการวิเคราะห์ค่าเอนโทรปีสำหรับเลือกตัวแปร ที่นำเสนอไปนี้ จะพบว่าค่าของข้อมูลเป็นค่าที่สามารถระบุได้แน่นอนเนื่องจากมีจำนวนน้อย ทำให้สามารถบอกเป็นค่าแบบไม่ต่อเนื่องได้ (Discrete) จึงสามารถที่จะหาค่าความน่าจะเป็นของข้อมูลได้ง่าย แต่เนื่องจากลักษณะของข้อมูลที่นำมาวิเคราะห์คือ ข้อมูลดีเอ็นเอไมโครอาร์เรย์ซึ่งค่าของข้อมูลเป็นค่าต่อเนื่อง (Continuous) การหาค่าความน่าจะเป็นของข้อมูลนั้นทำได้ยาก การวิเคราะห์ค่าเอนโทรปี ต้องปรับวิธีการให้เหมาะสม โดยข้อมูลเริ่มต้นก่อนที่นำมาวิเคราะห์จะต้องแปลงให้อยู่ในรูปของค่าไม่ต่อเนื่องก่อน ซึ่งจะทำให้วัดค่าความน่าจะเป็นของข้อมูลได้ ทั้งนี้วิธีการปรับข้อมูลเริ่มต้นนั้น ก็คือ การแบ่งช่วงค่าของข้อมูลเป็นช่วงๆ แทนค่าของข้อมูลตรงๆ จากนั้น ก็วัดความถี่ของข้อมูลที่อยู่ในช่วงนั้นๆ สำหรับหาค่าความน่าจะเป็นของข้อมูล

(Han and Kamber, 2001)

- การเลือกตัวแปรโดยการวิเคราะห์ค่าความแปรปรวน

เนื่องจากความแปรปรวนของข้อมูลจะใช้ในการอธิบาย ความแตกต่างของข้อมูลในตัวแปรหนึ่งๆ ซึ่งถ้าข้อมูลมีความแปรปรวนสูง แสดงว่าข้อมูลมีความแตกต่างกันสูง และถ้าข้อมูลมีลักษณะที่คล้ายๆ กัน หรือมีคุณสมบัติเหมือนกันๆ ค่าความแปรปรวนของข้อมูลนั้นย่อมมีค่าต่ำ ซึ่งผลจากตรงนี้เอง จึงตั้งสมมุติฐานในการเลือกตัวแปรที่สำคัญต่อการจำแนกประเภท นั่นคือ ตัวแปรที่มีความสำคัญกับการจำแนกประเภทคือตัวแปร ที่ให้ค่าความแปรปรวนของข้อมูลที่อยู่ภายในกลุ่มเดียวกันต่ำ ในทุกๆ กลุ่มของข้อมูลในตัวแปรนั้นๆ ซึ่งสามารถวัดออกมาได้โดยการพิจารณาจากสัดส่วนของค่าความแปรปรวนของข้อมูลทั้งหมดในตัวแปรนั้นๆ กับผลรวมค่าความแปรปรวนของข้อมูลในกลุ่มต่างๆ ดังสมการ

$$V(X) = \frac{Var(X)}{\sum_{i=1}^{n_D} Var(X_{D_i})} \qquad (92)$$

จากสมการ V(X) เป็นค่าสัดส่วนความแปรปรวนของข้อมูลในตัวแปร X ส่วน $Var(X)$ เป็นความแปรปรวนของข้อมูลทั้งหมดในตัวแปร X และ $Var(X_{D_i})$ คือความแปรปรวนของข้อมูลในตัวแปร X ซึ่งข้อมูลอยู่ในกลุ่ม D_i จากค่าสัดส่วนของความแปรปรวนที่ได้เมื่อตัวแปร X มีหลายตัวแปร จะเลือกตัวแปร X ที่มีค่า V(X) มากๆ เป็นตัวแปรที่เหมาะสมกับการจำแนกประเภทข้อมูล

(Liu, 2004)

- การสร้างตัวแปรใหม่โดยวิธีวิเคราะห์องค์ประกอบหลัก

จากข้อมูลซึ่งมีหลายๆตัวแปร จะสร้างค่าของข้อมูลชุดใหม่ ให้อยู่ในรูปขององค์ประกอบหลักแทนตัวแปรเหล่านี้ ซึ่งค่าของข้อมูลนี้ก็คือ ค่าคะแนนองค์ประกอบหลักนั่นเอง ผลจากการวิเคราะห์องค์ประกอบหลักจะทำให้ ตัวแปรของข้อมูลใหม่คือองค์ประกอบหลัก เป็นตัวแปรที่ มีมิติของข้อมูลน้อยกว่าตัวแปรเดิม นอกจากนี้การใช้องค์ประกอบหลักเป็นตัวแปร จะทำให้ข้อมูลที่นำไปใช้ในการวิเคราะห์เพื่อจำแนกประเภทนั้นยังเป็นข้อมูลชุดเดิมที่มีความแปรปรวนสูง นั่นหมายความว่า ตัวแปรของข้อมูลเดิมที่กำหนดมาตอนต้นนั้น ตัวแปรทุกๆ ตัวมีความสำคัญต่อการจำแนกประเภทของข้อมูลเหมือนๆ กัน

จากข้อมูลที่ใช้สำหรับการวิเคราะห์การจำแนกประเภทจะมีอยู่ 2 ชุดคือชุดข้อมูลที่จะใช้ในการสร้างโมเดล (Training Data) และชุดข้อมูลสำหรับทดสอบ (Testing Data) ดังนั้นการสร้างตัวแปรใหม่ จะต้องทำทั้งสองชุดข้อมูล ดังนี้

1) วิเคราะห์องค์ประกอบหลัก และหาคะแนนองค์ประกอบหลัก ของชุดข้อมูลที่จะใช้ในการสร้างโมเดล สำหรับการจำแนกประเภทข้อมูล
2) นำค่าพารามิเตอร์จากการวิเคราะห์วิเคราะห์องค์ประกอบหลักในข้อ 1.) เช่น ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน และ ค่าสัมประสิทธิ์ขององค์ประกอบหลัก มาใช้ในการสร้าง คะแนนขององค์ประกอบหลักในชุดข้อมูลทดสอบ โดยใช้ สมการ (27) ในการคำนวณ ผลก็คือจะทำให้ชุดข้อมูลทดสอบมีตัวแปรของข้อมูล เป็นองค์ประกอบหลัก เช่นเดียวกับ ชุดข้อมูลสำหรับสร้างโมเดล

5.2 การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์

แสดงการประยุกต์ การวิเคราะห์การจำแนกประเภท กับข้อมูลดีเอ็นเอไมโครอาร์เรย์ ได้ดังนี้

5.2.1 แหล่งข้อมูลและลักษณะข้อมูล

ข้อมูลที่นำมาใช้ เป็น กรณีศึกษา คือ ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของมะเร็งชนิดลิวคีเมีย ซึ่งได้อธิบายถึงแหล่งที่มาและลักษณะของข้อมูลแล้วในบทที่ 3 หัวข้อ 3.1.2 และจากตาราง 3.14 แสดงให้เห็นถึงตัวอย่างของข้อมูลที่จะใช้ในการสร้างโมเดลการจำแนกประเภท ซึ่งมีทั้งหมด 38 ตัวอย่างข้อมูล โดยแบ่งออกเป็น 2 กลุ่มได้แก่ กลุ่ม เอแอลแอล (ALL) 27 ตัวอย่าง และกลุ่มเอเอ็มแอล (AML) 11 ตัวอย่าง สำหรับข้อมูลที่ใช้ในการทดสอบผลการวิเคราะห์ จะตัดออกมาจากชุดข้อมูลทั้งหมดซึ่งแสดงไว้ในตาราง 3.13 โดยมีตัวอย่างของข้อมูลที่จะใช้ทดสอบผลการวิเคราะห์จำนวน 34 ตัวอย่าง ซึ่ง

ประเภทของมะเร็งในชุดข้อมูลนี้จะประกอบไปด้วย กลุ่ม เอแอลแอล 20 ตัวอย่าง และกลุ่มเอเอ็มแอล 14 ตัวอย่าง โดยแสดงลักษณะของข้อมูลชุดนี้ในดังตาราง 5.1

ตาราง 5.1 ข้อมูลการแสดงออกของยีนจากดีเอ็นเอไมโครอาร์เรย์ของมะเร็งชนิดลิวคีเมีย สำหรับทดสอบผลการวิเคราะห์

Gene ID (X_j)	Sample 39	Sample 40	Sample 41	...	Sample 71	Sample 72
M55150_at	2.816	3.108	3.109		3.316	3.244
U50136_rna1_at	3.051	3.026	3.146		3.405	3.359
Y12670_at	2.778	2.528	2.759		3.027	3.404
U46499_at	2.000	2.164	2.041		2.877	2.761
M77142_at	2.471	2.215	2.501	⋮	2.000	2.000
X95735_at	2.474	2.487	2.490		3.417	3.223
M80254_at	2.000	2.037	2.000		2.827	2.487
M23197_at	2.417	2.004	2.490		2.611	2.439
J04615_at	3.326	3.609	3.571		2.000	2.565
...	...	...	...		...	...
Type of Leukemia Cancer	ALL	ALL	ALL	...	AML	AML

5.2.2 วิธีการวิเคราะห์

1) วิเคราะห์ข้อมูลโดยใช้ฟังก์ชั่นการวิเคราะห์และพัฒนาขึ้นเอง จากโปรแกรม ภาษา R เวอร์ชัน 2.3.1 ในการคำนวณ

2) เตรียมข้อมูลโดยใช้วิธีการเดียวกับการวิเคราะห์ข้อมูลในบทที่ 3 หัวข้อ 3.2.2 ซึ่งผลการวิเคราะห์จะได้ยีนจำนวน 3,051 ยีน ในกลุ่มข้อมูลตัวอย่าง 72 กลุ่มข้อมูล ที่ประกอบไปด้วย ตัวอย่างข้อมูลสำหรับสร้างโมเดล 38 ข้อมูล และสำหรับทดสอบ 34 ข้อมูล เป็นข้อมูลตั้งต้น

3) เลือกตัวแปร สำหรับการจำแนกประเภท โดย วิธีการวิเคราะห์ค่าเอนโทรปี วิธีการวิเคราะห์ค่าความแปรปรวน และวิธีวิเคราะห์องค์ประกอบหลัก สำหรับวิธีการวิเคราะห์องค์ประกอบหลักจะทำการสร้างข้อมูลชุดใหม่ให้อยู่ในรูปของคะแนนองค์ประกอบหลัก ทั้งข้อมูลที่ใช้ในการสร้างโมเดล และข้อมูลที่จะทดสอบ

4) วิเคราะห์การจำแนกประเภท เพื่อสร้างโมเดลของการจำแนกประเภท โดยอาศัยตัวแปรที่ได้จากขั้นตอนที่ 3 ซึ่งจะทำให้ แบ่งการวิเคราะห์ข้อมูลออก เป็น 3 การทดลอง หลักๆ ตามตัวแปรที่ได้ ทั้ง 3 วิธี

5) ทำนาย กลุ่มของข้อมูล ในตัวอย่างข้อมูลทดสอบ และเปรียบเทียบผลการทำนาย กับ กลุ่มของข้อมูลเดิม

5.2.3 ผลการทดลอง

แสดงผลการทดลองใน 3 การทดลอง ตามวิธีการเลือกตัวแปรสำหรับจำแนกประเภทดังนี้

- **การทดลองที่ 1**

การทดลองนี้ ใช้วิธีการวิเคราะห์ค่าเอนโทรปีในการเลือกตัวแปรที่มีความสำคัญต่อการจำแนกประเภท ซึ่งในที่นี้ก็คือ ยีน โดยจากการวิเคราะห์จะเลือกยีนที่จำนวน 100 ตัว มีค่าประมาณของค่าเอนโทรปี E(X) น้อยที่สุด มาวิเคราะห์ ทั้งนี้จะเปรียบเทียบผลกับกลุ่มยีน 100 ตัวที่มีค่าประมาณของค่าเอนโทรปีมากที่สุด ยีนและค่าเอนโทรปีในยีน 100 ที่มีค่าประมาณของค่าเอนโทรปี มากและน้อยที่สุดนี้ จะแสดงได้ดังตาราง 5.2

ตาราง 5.2 ยีนและค่าประมาณของค่าเอนโทรปีของยีนที่มีค่าน้อยและมากที่สุด 100 ยีน

No.	ยีนที่มีค่า E(X) น้อยที่สุด	E(X)	ยีนที่มีค่า E(X) มากที่สุด	E(X)
1	M55150_at	0.083	X53795_at	0.600
2	U50136_rna1_at	0.142	U07223_at	0.599
3	Y12670_at	0.147	U73379_at	0.599
4	U46499_at	0.160	L16896_at	0.597
5	M77142_at	0.191	L42324_at	0.596
6	X95735_at	0.203	Y08302_at	0.595
7	M80254_at	0.208	U41387_at	0.594
8	M23197_at	0.220	HG2279-HT2375_at	0.593
9	J04615_at	0.221	U79255_at	0.593
10	U82759_at	0.227	U31384_at	0.593
11	M91432_at	0.229	U28368_at	0.593
12	M92287_at	0.229	L02950_at	0.593
13	M16038_at	0.237	HG511-HT511_at	0.593
14	M21551_rna1_at	0.241	HG4263-HT4533_at	0.593
15	U32944_at	0.262	D90070_s_at	0.592
16	U22376_cds2_s_at	0.262	X93996_rna1_at	0.592
17	J03801_f_at	0.271	X57351_at	0.592
18	M19045_f_at	0.271	U92014_at	0.592
19	X14008_rna1_f_at	0.271	J04162_at	0.592
20	HG3454-HT3647_at	0.277	U79266_at	0.592
21	Y00787_s_at	0.278	AF003743_at	0.592
22	X85116_rna1_s_at	0.280	L11066_at	0.591
23	L47738_at	0.281	L10910_at	0.591
24	U62136_at	0.283	Z72499_at	0.591
25	L09209_s_at	0.288	M85276_at	0.591
26	X59417_at	0.290	D42087_at	0.591
27	M22960_at	0.295	X89750_at	0.590
28	D10495_at	0.296	U91932_at	0.590

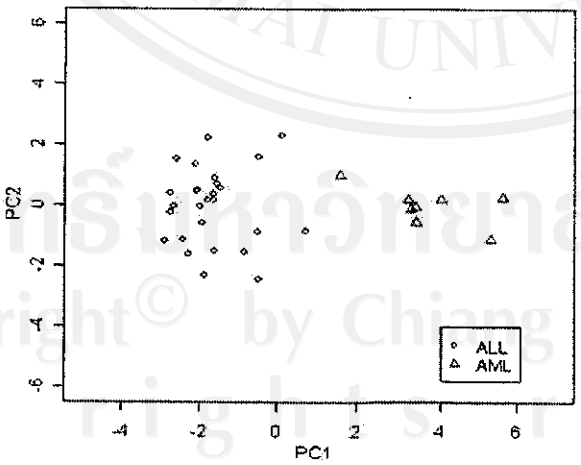
ตาราง 5.2(ต่อ) ยีนและค่าประมาณของค่าเอนโทรปีของยีนที่มีค่าน้อยและมากที่สุด 100 ยีน

No.	ยีนที่มีค่า E(X) น้อยที่สุด	E(X)	ยีนที่มีค่า E(X) มากที่สุด	E(X)
29	M83652_s_at	0.300	D78012_at	0.590
30	M63138_at	0.303	HG3355-HT3532_at	0.590
31	AFFX-HUMTFRR/M11507_M_at	0.303	M97676_at	0.590
32	M27891_at	0.303	X92972_at	0.590
33	M83667_rna1_s_at	0.303	M81181_s_at	0.589
40	M31523_at	0.319	HG3925-HT4195_s_at	0.588
41	M95678_at	0.319	X05196_at	0.588
42	S82470_at	0.322	U23803_at	0.588
43	M57710_at	0.322	L00022_s_at	0.588
44	M12959_s_at	0.322	U64105_at	0.587
45	M96326_rna1_at	0.323	U37251_at	0.587
46	M13452_s_at	0.323	L22454_at	0.587
47	X58431_rna2_s_at	0.325	L76568_xpt3_f_at	0.587
48	X70297_at	0.329	U30827_s_at	0.587
49	Y00339_s_at	0.329	S77763_at	0.586
50	X16546_at	0.330	M91585_at	0.586
51	HG620-HT620_at	0.332	M97287_at	0.586
52	U43519_at	0.333	U82275_at	0.586
53	X16832_at	0.334	D17716_at	0.586
54	D26308_at	0.335	L19161_at	0.586
55	M91036_rna1_at	0.339	M36429_s_at	0.586
56	X61587_at	0.339	U01147_at	0.586
57	U41767_s_at	0.339	L42374_s_at	0.585
58	M13485_at	0.340	L78833_cds4_at	0.585
59	M81695_s_at	0.340	HG3148-HT3324_s_at	0.585
60	U65928_at	0.344	U85245_at	0.585
61	L08246_at	0.344	X91247_at	0.585
62	M28130_rna1_s_at	0.344	U79252_at	0.584
63	J04027_at	0.346	D11428_at	0.584
64	U00802_s_at	0.346	D63813_at	0.584
65	J05243_at	0.346	X57398_at	0.584
66	L20941_at	0.346	X00274_at	0.584
67	M32304_s_at	0.346	U90916_at	0.584
68	M65085_at	0.346	U03642_at	0.584
69	M62762_at	0.347	M90656_at	0.584
70	X64364_at	0.347	U22526_at	0.584
71	D26579_at	0.347	U77718_at	0.584
72	M58603_at	0.347	X96969_at	0.583
73	L41870_at	0.348	L76224_at	0.583
74	Z69881_at	0.348	U11875_s_at	0.583
75	M92843_s_at	0.350	Z19002_at	0.583
76	Z30644_at	0.350	U76421_at	0.583
77	X64072_s_at	0.350	U57911_at	0.583

ตาราง 5.2(ต่อ) ยีนและค่าประมาณของค่าเอนโทรปีของยีนที่มีค่าน้อยและมากที่สุด 100 ยีน

No.	ยีนที่มีค่า E(X) น้อยที่สุด	E(X)	ยีนที่มีค่า E(X) มากที่สุด	E(X)
78	D26156_s_at	0.351	M93426_at	0.583
79	U73737_at	0.351	M32053_at	0.583
80	Y07604_at	0.352	L76465_at	0.583
81	X04085_rna1_at	0.353	M11717_rna1_at	0.583
82	M98399_s_at	0.354	J04430_s_at	0.583
83	X62654_rna1_at	0.354	X57351_s_at	0.583
84	Z32765_at	0.354	U44799_s_at	0.583
85	X52056_at	0.354	X83416_s_at	0.583
86	U26173_s_at	0.354	U70321_at	0.583
87	U80457_at	0.354	U52426_at	0.583
88	M29610_s_at	0.355	V00572_at	0.582
89	J03589_at	0.356	D13900_at	0.582
90	M28209_at	0.358	X77094_at	0.582
91	M31303_rna1_at	0.358	U83171_at	0.582
92	M14636_at	0.359	L19605_at	0.582
93	X80907_at	0.359	AF007875_at	0.582
94	D14874_at	0.360	HG2247-HT2332_at	0.582
95	X06985_at	0.360	M61906_at	0.582
96	HG2855-HT2995_at	0.361	X83127_at	0.582
97	U73960_at	0.362	X17651_at	0.582
98	D13641_at	0.362	L11285_at	0.582
99	M22324_at	0.362	D64158_at	0.582
100	U05259_rna1_at	0.362	D50663_at	0.581

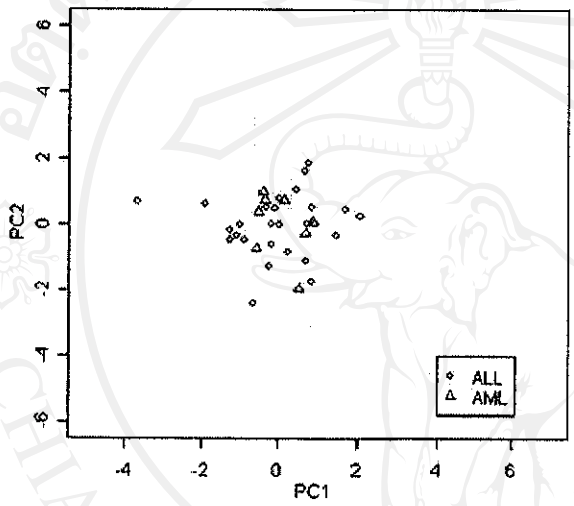
ผลจากตาราง 5.2 เมื่อนำข้อมูลจากตาราง 3.14 และคัดเลือกยีนจากตาราง 5.2 ที่มีค่าประมาณของค่าเอนโทรปีน้อยสุด มาใช้เป็นตัวแปรสำหรับการวิเคราะห์องค์ประกอบหลัก ตามวิธีการวิเคราะห์องค์ประกอบหลัก แสดงลักษณะการกระจายตัวของข้อมูลนี้ใน 2 องค์ประกอบหลักแรก ได้ดังกราฟ



รูป 5.3 กราฟผลของการวิเคราะห์องค์ประกอบหลักจากการคัดเลือกยีนที่มีค่า E(X) น้อยที่สุด100 ยีน

ผลจากรูป 5.3 แสดงให้เห็นว่ายีนที่มีค่าประมาณของค่าเอนโทรปี E(X) น้อยที่สุด เมื่อนำมาใช้เป็นตัวแปร และวิเคราะห์ข้อมูลดังกล่าวโดยการวิเคราะห์องค์ประกอบหลักซึ่งผลจากการวิเคราะห์จะแสดงข้อมูลในกราฟ 2 มิติ พบว่าตัวอย่างข้อมูลที่อยู่ในกลุ่ม ที่แตกต่างกัน จะแยกจากกันอย่างชัดเจน ซึ่งแสดงให้เห็นว่า ยีนเหล่านี้เป็นยีนที่เหมาะสมกับการใช้เป็นตัวแปรในการวิเคราะห์ เพื่อที่จำแนกประเภทของข้อมูล

และเพื่อที่จะยืนยันผลดังกล่าว เมื่อนำกลุ่มยีนมีค่าประมาณของค่าเอนโทรปี มากที่สุด 100 ยีน มาวิเคราะห์องค์ประกอบหลักจะแสดงผลได้ดังรูป 5.4



รูป 5.4 ผลของการวิเคราะห์องค์ประกอบหลักจากการคัดเลือกยีนที่มีค่า E(X) มากที่สุด100 ยีน

จากรูป 5.4 แสดงให้เห็นผลที่ตรงข้ามกับการคัดเลือกยีนที่มีค่าประมาณของค่าเอนโทรปีมากที่สุด นั่นคือ กลุ่มยีนเหล่านี้ไม่เหมาะที่จะใช้เป็นตัวแปรในการจำแนกประเภท เนื่องจากลักษณะของข้อมูลที่อยู่ในยีนเหล่านี้ เป็นข้อมูลที่ไม่ได้แยกข้อแตกต่าง ของกลุ่มตัวอย่างข้อมูลได้

ผลจากการวิเคราะห์ในตอนต้น จะเลือกยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด 100 ยีน มาใช้เป็นตัวแปรในการวิเคราะห์การจำแนกประเภท ซึ่งผลจากการวิเคราะห์จะทำให้ได้ พารามิเตอร์1 ตัว คือค่าน้ำหนักการจำแนกประเภท (Discriminant Weight: ω_j) มาใช้หาค่าคะแนนการจำแนกประเภท (Z) ตามสมการ (63) และจะแสดงค่าพารามิเตอร์ดังกล่าวได้ดังตาราง 5.3

ตาราง 5.3 ค่าน้ำหนักการจำแนกประเภทในยีนที่เลือกตามค่า E(X) 100 ตัว

Gene Id (X_j)	Discriminant Weight (ω_j)	Gene Id (X_j)	Discriminant Weight (ω_j)
M55150_at	1.073	HG620-HT620_at	0.326
U50136_rna1_at	0.444	U43519_at	0.448
Y12670_at	-0.542	X16832_at	-0.173
U46499_at	0.434	D26308_at	0.181
M77142_at	0.324	M91036_rna1_at	-0.192
X95735_at	0.318	X61587_at	0.042
M80254_at	0.100	U41767_s_at	0.200
M23197_at	0.335	M13485_at	-0.247
J04615_at	0.001	M81695_s_at	0.429
U82759_at	1.113	U65928_at	-0.859
M91432_at	0.213	L08246_at	0.330
M92287_at	-0.602	M28130_rna1_s_at	0.161
M16038_at	0.162	J04027_at	-0.292
M21551_rna1_at	0.618	U00802_s_at	-0.094
U32944_at	-0.214	J05243_at	-0.125
U22376_cds2_s_at	-0.416	L20941_at	-0.166
J03801_f_at	0.062	M32304_s_at	0.231
M19045_f_at	-0.208	M65085_at	-0.338
X14008_rna1_f_at	0.157	M62762_at	0.040
HG3454-HT3647_at	-0.112	X64364_at	-0.516
Y00787_s_at	0.030	D26579_at	0.426
X85116_rna1_s_at	0.514	M58603_at	0.083
L47738_at	0.457	L41870_at	-0.082
U62136_at	-0.420	Z69881_at	-0.300
L09209_s_at	0.305	M92843_s_at	-0.063
X59417_at	-0.042	Z30644_at	0.850
M22960_at	0.450	X64072_s_at	0.218
D10495_at	0.574	D26156_s_at	0.096
M83652_s_at	-0.317	U73737_at	-0.419
M63138_at	-0.228	Y07604_at	0.360
AFFX-HUMTFRR/M11507_M_at	0.091	X04085_rna1_at	-0.205
M27891_at	0.345	M98399_s_at	0.506
M83667_rna1_s_at	-0.095	X62654_rna1_at	-0.007
X17042_at	0.210	Z32765_at	0.093
U67963_at	-0.468	X52056_at	0.380
M11147_at	0.083	U26173_s_at	0.254
U46751_at	0.249	U80457_at	-0.209
D38073_at	0.099	M29610_s_at	0.276
X74262_at	0.018	J03589_at	-0.082
M31523_at	0.245	M28209_at	-0.302
M95678_at	-0.448	M31303_rna1_at	-0.158
S82470_at	0.143	M14636_at	0.004
M57710_at	-0.168	X80907_at	-0.170

ตาราง 5.3 (ต่อ) ค่าน้ำหนักการจำแนกประเภทในยีนที่เลือกตามค่า E(X) 100 ตัว

M12959_s_at	0.272	D14874_at	-0.423
M96326_rna1_at	0.027	X06985_at	-0.067
M13452_s_at	-0.461	HG2855-HT2995_at	-0.281
X58431_rna2_s_at	1.442	U73960_at	-0.008
X70297_at	0.632	D13641_at	0.282
Y00339_s_at	0.559	M22324_at	0.373
X16546_at	0.225	U05259_rna1_at	0.117

จากตาราง 5.3 พารามิเตอร์ที่ได้จะถูกนำไปใช้ในการหาค่าคะแนนการจำแนกประเภทในข้อมูลทดสอบ ซึ่งมีจำนวน 34 ตัวอย่าง และ ทำนายกลุ่มของข้อมูลจากค่าคะแนนการจำแนกประเภทดังกล่าว ซึ่งได้ผลการทำนายข้อมูล ดังตาราง 5.4

ตาราง 5.4 ผลการจำแนกประเภทของข้อมูล จากการวิเคราะห์การจำแนกประเภท โดยใช้ยีนที่มีค่า E(X) น้อยที่สุด 100 ยีน เป็นตัวแปร

Samples	Original Class	Results of LDA	Samples	Original Class	Results of LDA
39	ALL	ALL	56	ALL	ALL
40	ALL	ALL	57	AML	AML
41	ALL	ALL	59	AML	AML
42	ALL	ALL	59	ALL	ALL
43	ALL	ALL	*60	AML	ALL
44	ALL	ALL	*61	AML	ALL
45	ALL	ALL	62	AML	AML
46	ALL	ALL	63	AML	AML
47	ALL	ALL	64	AML	AML
48	ALL	ALL	65	AML	AML
49	ALL	ALL	*66	AML	ALL
50	AML	AML	67	ALL	ALL
51	AML	AML	68	ALL	ALL
52	AML	AML	69	ALL	ALL
53	AML	AML	70	ALL	ALL
*54	AML	ALL	71	ALL	ALL
55	ALL	ALL	72	ALL	ALL
			Correct: 88.24 %		
			Error: 11.76 %		

จากตาราง 5.4 แสดงให้เห็นตัวอย่างข้อมูล 34 ตัวอย่าง ซึ่งมีการระบุกลุ่มข้อมูลเริ่มต้น (Original Class) และ แสดงผลของการวิเคราะห์การจำแนกประเภท (Result of LDA) จากผลการทำนายกลุ่มที่ได้ พบว่ากลุ่มข้อมูลที่ได้จากการทำนาย ซึ่งตรงกับกลุ่มเริ่มต้น (Correct) คิดเป็น 88.24 เปอร์เซ็นต์ และ กรณีที่ทำนายผิดพลาด (Error) คิดเป็น 11.76 เปอร์เซ็นต์

จากผลการวิเคราะห์ที่ได้ เป็นการใช้ยีนทั้ง 100 ตัวในการวิเคราะห์ ดังนั้น หากมีการเลือกยีนโดยกำหนดจำนวนยีนที่มีค่าประมาณของค่าเอ็นโทรปีน้อยๆ และ ยีนที่มีค่าประมาณของค่าเอ็นโทรปีมากๆ ในลักษณะของช่วงยีน เช่น ยีนตัวที่ 1-2 ยีนตัวที่ 1-3 มาใช้เป็นตัวแปรในการวิเคราะห์ จะแสดงผลของการทำนายที่ได้จากกลุ่มยีนดังกล่าวดังตาราง 5.5

ตาราง 5.5 ผลสรุปของการวิเคราะห์การจำแนกประเภทของกลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุดและมากที่สุด โดยแบ่งเป็นช่วงยีน 99 ช่วงยีน

	ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่าเอนโทรปีน้อยที่สุดเป็นตัวแปร		ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่าเอนโทรปีมากที่สุดเป็นตัวแปร	
ช่วงของยีน (Gene Range)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)
1-2	79.41	20.59	58.82	41.18
1-3	79.41	20.59	58.82	41.18
1-4	79.41	20.59	58.82	41.18
1-5	85.29	14.71	61.76	38.24
1-6	88.24	11.76	58.82	41.18
1-7	85.29	14.71	58.82	41.18
1-8	91.18	8.82	61.76	38.24
1-9	91.18	8.82	61.76	38.24
1-10	91.18	8.82	64.71	35.29
1-11	91.18	8.82	64.71	35.29
1-12	94.12	5.88	67.65	32.35
1-13	91.18	8.82	61.76	38.24
1-14	91.18	8.82	67.65	32.35
1-15	88.24	11.76	76.47	23.53
1-16	82.35	17.65	70.59	29.41
1-17	82.35	17.65	67.65	32.35
1-18	82.35	17.65	70.59	29.41
1-19	82.35	17.65	64.71	35.29
1-20	85.29	14.71	64.71	35.29
1-21	82.35	17.65	61.76	38.24
1-22	82.35	17.65	61.76	38.24
1-23	88.24	11.76	64.71	35.29
1-24	88.24	11.76	64.71	35.29
1-25	88.24	11.76	58.82	41.18
1-26	85.29	14.71	52.94	47.06
1-27	85.29	14.71	58.82	41.18
1-28	79.41	20.59	50.00	50.00
1-29	76.47	23.53	50.00	50.00
1-30	94.12	5.88	52.94	47.06

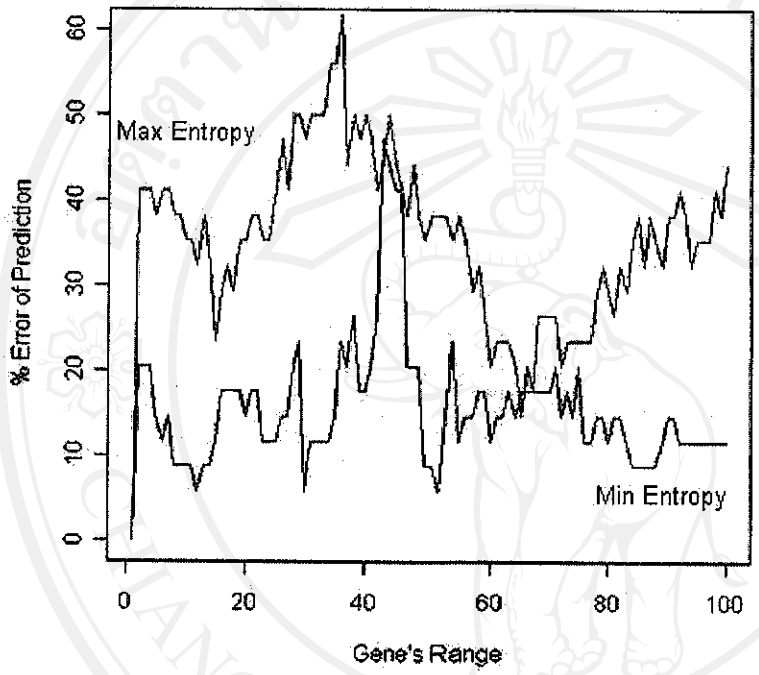
ตาราง 5.5 (ต่อ) ผลสรุปของการวิเคราะห์การจำแนกประเภทของกลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด และมากที่สุด โดยแบ่งเป็นช่วงยีน 99 ช่วงยีน

	ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่าเอนโทรปีน้อยที่สุดเป็นตัวแปร		ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่าเอนโทรปีมากที่สุดเป็นตัวแปร	
ช่วงของยีน (Gene Range)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)
1-31	88.24	11.76	50.00	50.00
1-32	88.24	11.76	50.00	50.00
1-33	88.24	11.76	50.00	50.00
1-34	88.24	11.76	44.12	55.88
1-35	85.29	14.71	44.12	55.88
1-36	76.47	23.53	38.24	61.76
1-37	79.41	20.59	55.88	44.12
1-38	73.53	26.47	50.00	50.00
1-39	82.35	17.65	52.94	47.06
1-40	82.35	17.65	50.00	50.00
1-41	79.41	20.59	52.94	47.06
1-42	73.53	26.47	58.82	41.18
1-43	52.94	47.06	55.88	44.12
1-44	55.88	44.12	50.00	50.00
1-45	58.82	41.18	55.88	44.12
1-46	58.82	41.18	58.82	41.18
1-47	79.41	20.59	61.76	38.24
1-48	79.41	20.59	55.88	44.12
1-49	79.41	20.59	61.76	38.24
1-50	91.18	8.82	64.71	35.29
1-51	91.18	8.82	61.76	38.24
1-52	94.12	5.88	61.76	38.24
1-53	85.29	14.71	61.76	38.24
1-54	76.47	23.53	64.71	35.29
1-55	88.24	11.76	61.76	38.24
1-56	85.29	14.71	64.71	35.29
1-57	85.29	14.71	70.59	29.41
1-58	82.35	17.65	67.65	32.35
1-59	82.35	17.65	73.53	26.47
1-60	88.24	11.76	79.41	20.59
1-61	85.29	14.71	76.47	23.53
1-62	85.29	14.71	76.47	23.53
1-63	82.35	17.65	76.47	23.53
1-64	85.29	14.71	79.41	20.59
1-65	82.35	17.65	85.29	14.71

ตาราง 5.5 (ต่อ) ผลสรุปของการวิเคราะห์การจำแนกประเภทของกลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด และมากที่สุด โดยแบ่งเป็นช่วงยีน 99 ช่วงยีน

	ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่าเอนโทรปีน้อยที่สุดเป็นตัวแปร		ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่าเอนโทรปีมากที่สุดเป็นตัวแปร	
ช่วงของยีน (Gene Range)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)
1-66	82.35	17.65	79.41	20.59
1-67	82.35	17.65	82.35	17.65
1-68	82.35	17.65	73.53	26.47
1-69	82.35	17.65	73.53	26.47
1-70	82.35	17.65	73.53	26.47
1-71	79.41	20.59	73.53	26.47
1-72	85.29	14.71	79.41	20.59
1-73	82.35	17.65	76.47	23.53
1-74	85.29	14.71	76.47	23.53
1-75	79.41	20.59	76.47	23.53
1-76	88.24	11.76	76.47	23.53
1-77	88.24	11.76	76.47	23.53
1-78	85.29	14.71	70.59	29.41
1-79	85.29	14.71	67.65	32.35
1-80	88.24	11.76	70.59	29.41
1-81	85.29	14.71	73.53	26.47
1-82	85.29	14.71	67.65	32.35
1-83	88.24	11.76	70.59	29.41
1-84	91.18	8.82	64.71	35.29
1-85	91.18	8.82	61.76	38.24
1-86	91.18	8.82	67.65	32.35
1-87	91.18	8.82	61.76	38.24
1-88	91.18	8.82	64.71	35.29
1-89	88.24	11.76	67.65	32.35
1-90	85.29	14.71	61.76	38.24
1-91	85.29	14.71	61.76	38.24
1-92	88.24	11.76	58.82	41.18
1-93	88.24	11.76	61.76	38.24
1-94	88.24	11.76	67.65	32.35
1-95	88.24	11.76	64.71	35.29
1-96	88.24	11.76	64.71	35.29
1-97	88.24	11.76	64.71	35.29
1-98	88.24	11.76	58.82	41.18
1-99	88.24	11.76	61.76	38.24
1-100	88.24	11.76	55.88	44.12

ผลจากตาราง 5.5 พบว่ากลุ่มยีนที่มีค่าประมาณของเอนโทรปีน้อยที่สุด เมื่อนำไปใช้เป็นตัวแปรจำแนกประเภทโดยการวิเคราะห์การจำแนกประเภท จะให้ผลการทำนายที่เป็นไปได้ว่า ดีกว่ากลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีมากที่สุด แต่จำนวนยีนจำนวนเท่าใดจึงจะเหมาะสมเป็นสิ่งที่ต้องศึกษาต่อไป และผลที่ได้จากตารางเมื่อนำค่าเปอร์เซ็นต์ความผิดพลาดจากการทำนาย (Error of Prediction) มาพล็อตในกราฟ จะแสดงได้ดังรูป 5.5



รูป 5.5 กราฟผลสรุปของการวิเคราะห์การจำแนกประเภทของกลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด และมากที่สุด โดยแบ่งเป็นช่วงยีน 99 ช่วงยีน

จากรูป 5.5 จะเห็นว่าช่วงของยีนที่มีค่าประมาณเอนโทรปีน้อย มีแนวโน้มในสัดส่วนที่สูง ที่เปอร์เซ็นต์ความผิดพลาดจากการทำนายจะต่ำกว่ากลุ่มยีนที่มีค่าประมาณของเอนโทรปีมาก แต่ทั้งนี้เนื่องจากกราฟมีลักษณะขึ้นลง ๆ ไม่แน่นอน และยีนบางช่วงที่มีค่าประมาณของเอนโทรน้อย ก็มีเปอร์เซ็นต์ความผิดพลาดสูงกว่า ช่วงของยีนที่มีค่าประมาณของเอนโทรปีสูง กระบวนการคัดเลือกยีนโดยวิธีวิเคราะห์ค่าเอนโทรปี เพื่อให้ได้ ยีน และจำนวนยีนที่เหมาะสม จริงๆ กับการจำแนกประเภทจำเป็นที่ต้องมีศึกษา และปรับปรุงวิธีการให้มีประสิทธิภาพต่อไป

- **การทดลองที่ 2**

การทดลองนี้ ใช้วิธีการวิเคราะห์ค่าสัดส่วนของค่าความแปรปรวน V(X) ในการเลือกตัวแปรที่มีความสำคัญต่อการจำแนกประเภท ซึ่งในที่นี้ก็คือ ยีน โดยจากการวิเคราะห์จะเลือกยีนที่จำนวน 100 ตัว ที่มีค่าสัดส่วนของค่าความแปรปรวน มากที่สุดมาวิเคราะห์ ทั้งนี้จะเปรียบเทียบผลกับกลุ่มยีน 100 ตัวที่มีค่าสัดส่วนของค่าความแปรปรวนน้อยที่สุด ยีนและค่าสัดส่วนของค่าความแปรปรวน ในยีน 100 ที่มีค่ามากที่สุดและน้อยที่สุด จะแสดงได้ดังตาราง 5.6

ตาราง 5.6 ยีนและค่าสัดส่วนของค่าความแปรปรวน V(X) มากและน้อยที่สุด 100 ยีน

No.	ยีนที่มีค่า V(X)มากที่สุด	V(X)	ยีนที่มีค่า V(X) น้อยที่สุด	V(X)
1	M27891_at	1.690	HG1496-HT1496_s_at	0.289
2	X95735_at	1.604	X65977_at	0.289
3	U50136_rna1_at	1.445	X13334_at	0.289
4	U22376_cds2_s_at	1.296	M22612_f_at	0.289
5	Y12670_at	1.247	Z46632_r_at	0.289
6	M23197_at	1.238	X99076_rna1_at	0.289
7	M55150_at	1.214	L77701_at	0.289
8	M31523_at	1.205	L24564_at	0.289
9	M16038_at	1.168	U06155_s_at	0.291
10	X74262_at	1.149	X53961_at	0.291
11	U82759_at	1.121	X06256_at	0.298
12	L09209_s_at	1.108	X96584_at	0.303
13	L47738_at	1.096	D83657_at	0.304
14	Y00787_s_at	1.090	J02973_rna1_at	0.305
15	J05243_at	1.074	Z49269_at	0.305
16	M31211_s_at	1.069	M87860_at	0.310
17	X85116_rna1_s_at	1.056	X13955_s_at	0.312
18	Z15115_at	1.049	AB000584_at	0.313
19	X04085_rna1_at	1.046	X71345_f_at	0.315
20	U46499_at	1.044	L32866_at	0.319
21	M91432_at	1.027	J05272_at	0.322
22	X70297_at	1.025	L06419_at	0.323
23	X17042_at	1.024	M93221_at	0.323
24	U05259_rna1_at	1.021	D14826_s_at	0.326
25	D88422_at	1.018	U64998_at	0.327
26	L41870_at	1.005	U52518_at	0.331
27	S50223_at	1.005	S71824_at	0.332
28	U62136_at	1.004	L44140_cds4_s_at	0.332
29	M89957_at	0.995	D10923_at	0.335
30	M28130_rna1_s_at	0.991	X04602_s_at	0.335
31	U29175_at	0.991	Y00081_s_at	0.336
32	X15949_at	0.991	U21931_at	0.336

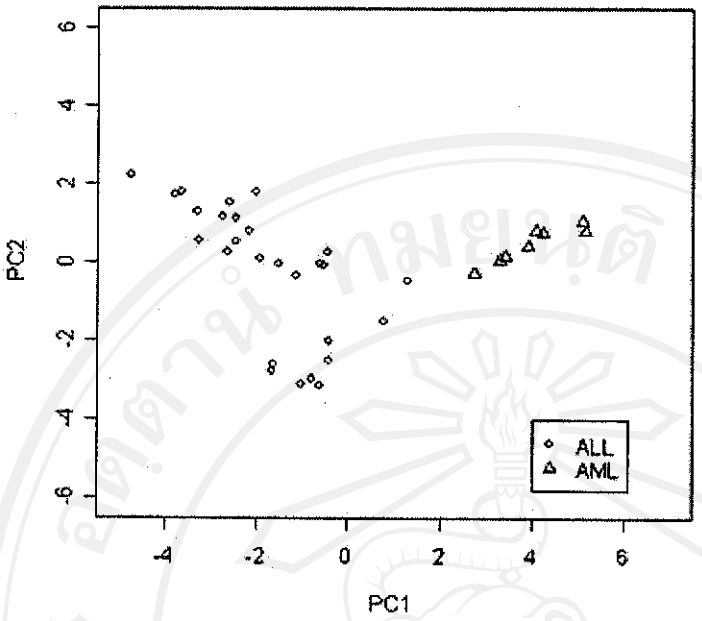
ตาราง 5.6 (ต่อ) ยีนและค่าสัดส่วนของค่าความแปรปรวน V(X) มากและน้อยที่สุด 100 ยีน

No.	ยีนที่มีค่า V(X)มากที่สุด	V(X)	ยีนที่มีค่า V(X) น้อยที่สุด	V(X)
33	X62654_rna1_at	0.988	L40387_at	0.338
34	X82240_rna1_at	0.985	U40434_at	0.338
35	X07743_at	0.980	J03040_at	0.338
39	U72936_s_at	0.974	M25897_at	0.342
40	Z69881_at	0.974	M25077_at	0.343
41	U32944_at	0.970	L35594_at	0.343
42	D88270_at	0.968	D10202_at	0.344
43	M63138_at	0.967	D26129_at	0.347
44	M11147_at	0.964	X52213_s_at	0.347
45	X59350_at	0.964	X78706_at	0.347
46	U73737_at	0.960	M19159_at	0.349
47	M94633_at	0.957	X13839_at	0.349
48	M11722_at	0.953	U21049_at	0.352
49	X74801_at	0.952	S68271_s_at	0.353
50	X63469_at	0.949	M61764_at	0.355
51	M12959_s_at	0.947	L11239_at	0.356
52	D38073_at	0.945	X14787_at	0.356
53	U49020_cds2_s_at	0.936	D38535_at	0.357
54	U49844_at	0.935	U78095_at	0.358
55	U50928_at	0.935	M14660_at	0.359
56	X61587_at	0.935	HG3995-HT4265_at	0.360
57	M96326_rna1_at	0.932	AFFX-HUMISGF3A/M97935_MA_at	0.361
58	M37435_at	0.932	X52773_at	0.362
59	U20998_at	0.930	D83542_at	0.363
60	L33930_s_at	0.927	M13755_at	0.365
61	U27460_at	0.927	X05409_at	0.366
62	J03801_f_at	0.926	M63582_at	0.366
63	D87078_at	0.925	L08010_at	0.371
64	M60527_at	0.924	D43949_at	0.372
65	M83233_at	0.923	X63131_s_at	0.372
66	M77142_at	0.923	M85247_at	0.372
67	M21551_rna1_at	0.922	M26602_at	0.373
68	L08246_at	0.920	X69910_at	0.373
69	M22960_at	0.918	HG3495-HT3689_at	0.373
70	M29696_at	0.913	S77812_at	0.374
71	X14008_rna1_f_at	0.911	M31551_s_at	0.374
72	U31556_at	0.909	U03090_at	0.377
73	M62762_at	0.904	U70451_at	0.377
74	X76648_at	0.900	U27325_s_at	0.378
75	M27783_s_at	0.899	U42408_at	0.378
76	D14874_at	0.899	U87408_at	0.379
77	U79285_at	0.898	M91029_cds2_at	0.380
78	M92287_at	0.893	AB002382_at	0.381

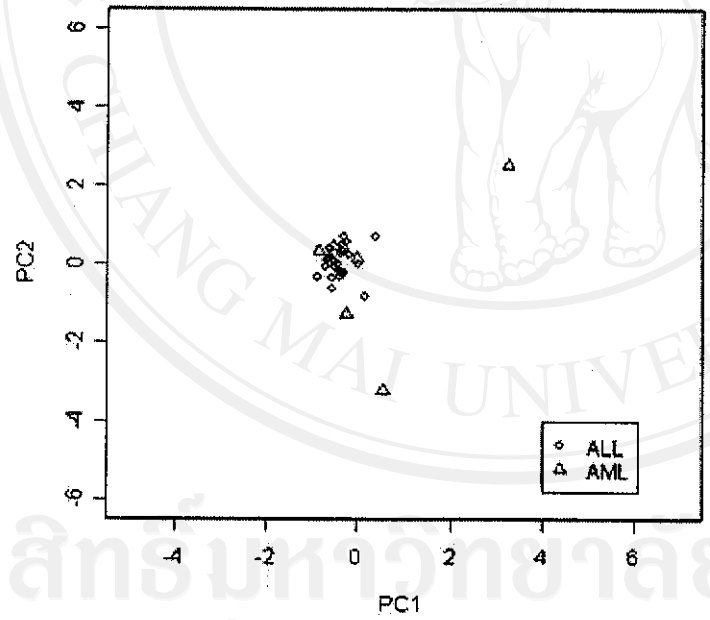
ตาราง 5.6 (ต่อ) ยีนและค่าสัดส่วนของค่าความแปรปรวน V(X) มากและน้อยที่สุด 100 ยีน

No.	ยีนที่มีค่า V(X)มากที่สุด	V(X)	ยีนที่มีค่า V(X) น้อยที่สุด	V(X)
79	HG4582-HT4987_at	0.892	U07664_at	0.381
80	Y07604_at	0.889	D84110_at	0.382
81	U65928_at	0.885	X87211_at	0.382
82	D21262_at	0.884	U43916_s_at	0.383
83	U38846_at	0.882	HG1205-HT1205_at	0.383
84	D86967_at	0.880	L22342_at	0.383
85	M63438_s_at	0.879	M13241_at	0.383
86	L07956_at	0.878	D13897_rna2_at	0.384
87	X58431_rna2_s_at	0.876	L11329_at	0.384
88	M13792_at	0.875	X06482_at	0.384
89	M95678_at	0.875	X55668_at	0.384
90	U84388_at	0.872	U52513_at	0.385
91	Y00433_at	0.871	L01664_at	0.385
92	X68560_at	0.870	X14329_at	0.385
93	U16954_at	0.870	U70735_at	0.386
94	S76965_at	0.867	X53800_s_at	0.386
95	L05148_at	0.867	V00594_at	0.387
96	X59871_at	0.864	L11370_at	0.387
97	D63391_at	0.864	M63262_at	0.387
98	Z68747_at	0.862	U47677_at	0.388
99	X74301_s_at	0.859	U70663_at	0.388
100	M81695_s_at	0.859	U79261_s_at	0.388

จากตาราง 5.6 เลือกยีน 100 ยีน ที่มีสัดส่วนความแปรปรวนของข้อมูล V(X) มากและน้อยที่สุด ตามลำดับ มาวิเคราะห์องค์ประกอบหลัก เพื่อแสดงลักษณะการกระจายตัวของข้อมูลใน 2 มิติ แสดงผลการวิเคราะห์ได้ดังกราฟ ในรูป 5.6 และรูป 5.7



รูป 5.6 ผลของการวิเคราะห์องค์ประกอบหลักจากการคัดเลือกยีนที่มีค่า V(X) มากที่สุด100 ยีน



รูป 5.7 ผลของการวิเคราะห์องค์ประกอบหลักจากการคัดเลือกยีนที่มีค่า V(X) น้อยที่สุด100 ยีน

จากกราฟในรูป 5.6 และ รูป 5.7 แสดงให้เห็นข้อแตกต่างของ ลักษณะการกระจายตัวของกลุ่มข้อมูลในยีนที่มีสัดส่วนความแปรปรวนของข้อมูลมากที่สุด และน้อยที่สุด ตามลำดับ ซึ่งกรณีแรกจะ

เห็นว่ากลุ่มยีนที่เลือกมา มีลักษณะของข้อมูลที่แยกข้อแตกต่างของกลุ่มข้อมูลตัวอย่างได้ดีกว่า ซึ่งให้ผลคล้ายกับการเลือกยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุด

จากการวิเคราะห์จำแนก โดยอาศัยยีนที่มีค่าสัดส่วนของความแปรปรวนของข้อมูลมากที่สุด 100 ยีนมาใช้เป็นตัวแปรจำแนกประเภท จะแสดง ค่าน้ำหนักการจำแนกประเภท จากผลการวิเคราะห์ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ สำหรับกลุ่มตัวอย่างข้อมูลที่ใช้สร้างโมเดล ได้ดังตาราง 5.7

ตาราง 5.7 ค่าน้ำหนักการจำแนกประเภทในยีนที่เลือกตามค่า V(X) 100 ตัว

Gene Id (X_j)	Discriminant Weight (ω_j)	Gene Id (X_j)	Discriminant Weight (ω_j)
M55150_at	1.073	HG620-HT620_at	0.326
U50136_rna1_at	0.444	U43519_at	0.448
Y12670_at	-0.542	X16832_at	-0.173
U46499_at	0.434	D26308_at	0.181
M77142_at	0.324	M91036_rna1_at	-0.192
X95735_at	0.318	X61587_at	0.042
M80254_at	0.100	U41767_s_at	0.200
M23197_at	0.335	M13485_at	-0.247
J04615_at	0.001	M81695_s_at	0.429
U82759_at	1.113	U65928_at	-0.859
M91432_at	0.213	L08246_at	0.330
M92287_at	-0.602	M28130_rna1_s_at	0.161
M16038_at	0.162	J04027_at	-0.292
M21551_rna1_at	0.618	U00802_s_at	-0.094
U32944_at	-0.214	J05243_at	-0.125
U22376_cds2_s_at	-0.416	L20941_at	-0.166
J03801_f_at	0.062	M32304_s_at	0.231
M19045_f_at	-0.208	M65085_at	-0.338
X14008_rna1_f_at	0.157	M62762_at	0.040
HG3454-HT3647_at	-0.112	X64364_at	-0.516
Y00787_s_at	0.030	D26579_at	0.426
X85116_rna1_s_at	0.514	M58603_at	0.083
L47738_at	0.457	L41870_at	-0.082
U62136_at	-0.420	Z69881_at	-0.300
L09209_s_at	0.305	M92843_s_at	-0.063
X59417_at	-0.042	Z30644_at	0.850
M22960_at	0.450	X64072_s_at	0.218
D10495_at	0.574	D26156_s_at	0.096
M83652_s_at	-0.317	U73737_at	-0.419
M63138_at	-0.228	Y07604_at	0.360
AFFX-HUMTFRR/M11507_M_at	0.091	X04085_rna1_at	-0.205
M27891_at	0.345	M98399_s_at	0.506
M83667_rna1_s_at	-0.095	X62654_rna1_at	-0.007
X17042_at	0.210	Z32765_at	0.093

ตาราง 5.7 (ต่อ) ค่าน้ำหนักการจำแนกประเภทในยีนที่เลือกตามค่า V(X) 100 ตัว

Gene Id (X_j)	Discriminant Weight (ω_j)	Gene Id (X_j)	Discriminant Weight (ω_j)
U67963_at	-0.468	X52056_at	0.380
M11147_at	0.083	U26173_s_at	0.254
U46751_at	0.249	U80457_at	-0.209
D38073_at	0.099	M29610_s_at	0.276
X74262_at	0.018	J03589_at	-0.082
M31523_at	0.245	M28209_at	-0.302
M95678_at	-0.448	M31303_rna1_at	-0.158
S82470_at	0.143	M14636_at	0.004
M57710_at	-0.168	X80907_at	-0.170
M12959_s_at	0.272	D14874_at	-0.423
M96326_rna1_at	0.027	X06985_at	-0.067
M13452_s_at	-0.461	HG2855-HT2995_at	-0.281
X58431_rna2_s_at	1.442	U73960_at	-0.008
X70297_at	0.632	D13641_at	0.282
Y00339_s_at	0.559	M22324_at	0.373
X16546_at	0.225	U05259_rna1_at	0.117

ผลจากการทำนาย ในชุดข้อมูลทดสอบ โดยใช้ค่าน้ำหนักการจำแนกประเภทจากตาราง 5.7 จะแสดงได้ดังตาราง 5.8

ตาราง 5.8 ผลการจำแนกประเภทของข้อมูล จากการวิเคราะห์การจำแนกประเภท โดยใช้ยีนที่มีค่า V(X) มากที่สุด 100 ยีน เป็นตัวแปร

Samples	Original Class	Results of LDA	Samples	Original Class	Results of LDA
39	ALL	ALL	56	ALL	ALL
40	ALL	ALL	57	AML	AML
41	ALL	ALL	59	AML	AML
42	ALL	ALL	59	ALL	ALL
43	ALL	ALL	60	AML	AML
44	ALL	ALL	61	AML	AML
45	ALL	ALL	62	AML	AML
46	ALL	ALL	63	AML	AML
47	ALL	ALL	64	AML	AML
48	ALL	ALL	65	AML	AML
49	ALL	ALL	*66	AML	ALL
50	AML	AML	67	ALL	ALL
51	AML	AML	68	ALL	ALL
52	AML	AML	69	ALL	ALL
53	AML	AML	70	ALL	ALL
54	AML	AML	71	ALL	ALL
55	ALL	ALL	72	ALL	ALL
			Correct: 97.06 %		
			Error: 2.94 %		

จากตาราง 5.8 แสดงให้เห็นตัวอย่างข้อมูล 34 ตัวอย่าง ซึ่งมีการระบุกลุ่มข้อมูลเริ่มต้น (Original Class) และ แสดงผลของการวิเคราะห์การจำแนกประเภท (Result of LDA) จากผลการทำนายกลุ่มที่ได้ พบว่ากลุ่มข้อมูลที่ได้จากการทำนาย ซึ่งตรงกับกลุ่มเริ่มต้น คิดเป็น 97.06 เปอร์เซ็นต์ และ กรณีที่ทำนายผิดพลาดคิดเป็น 2.94 เปอร์เซ็นต์

จากผลการวิเคราะห์ที่ได้ เป็นการใช้ยีน 100 ตัวที่มีค่าสัดส่วนความแปรปรวนมาก ในการวิเคราะห์ จะเห็นว่าให้ผลการทำนายที่น่าพอใจ และเมื่อพิจารณาทุกๆ ช่วงยีน จะแสดงผลของการทำนายที่ได้จากช่วงยีนต่างๆ ดังตาราง 5.9

ตาราง 5.9 ผลสรุปของการวิเคราะห์การจำแนกประเภทของกลุ่มยีนที่มีค่าสัดส่วนของความแปรปรวนมากที่สุดและน้อยที่สุด โดยแบ่งเป็นช่วงยีน 99 ช่วงยีน

	ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่า V(X) มากที่สุด		ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่า V(X) น้อยที่สุด	
ช่วงของยีน (Gene Range)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)
1-2	94.12	5.88	55.88	44.12
1-3	91.18	8.82	70.59	29.41
1-4	85.29	14.71	70.59	29.41
1-5	85.29	14.71	52.94	47.06
1-6	88.24	11.76	52.94	47.06
1-7	85.29	14.71	61.76	38.24
1-8	88.24	11.76	61.76	38.24
1-9	88.24	11.76	61.76	38.24
1-10	88.24	11.76	61.76	38.24
1-11	91.18	8.82	55.88	44.12
1-12	91.18	8.82	55.88	44.12
1-13	91.18	8.82	58.82	41.18
1-14	91.18	8.82	58.82	41.18
1-15	94.12	5.88	58.82	41.18
1-16	94.12	5.88	58.82	41.18
1-17	94.12	5.88	52.94	47.06
1-18	94.12	5.88	50.00	50.00
1-19	94.12	5.88	52.94	47.06
1-20	94.12	5.88	55.88	44.12
1-21	94.12	5.88	55.88	44.12
1-22	91.18	8.82	55.88	44.12
1-23	91.18	8.82	50.00	50.00
1-24	88.24	11.76	67.65	32.35
1-25	85.29	14.71	73.53	26.47
1-26	82.35	17.65	52.94	47.06

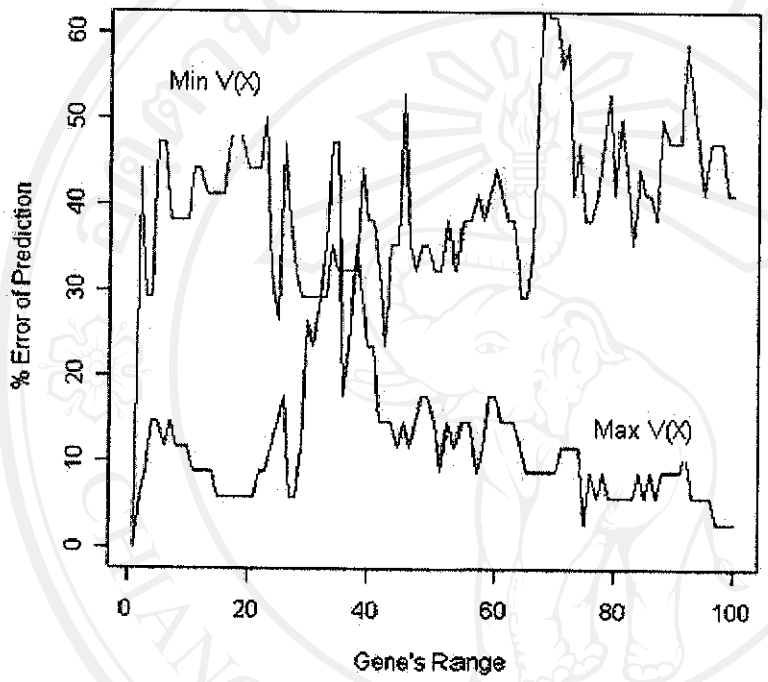
ตาราง 5.9 (ต่อ) ผลสรุปของการวิเคราะห์การจำแนกประเภทของกลุ่มยีนที่มีค่าสัดส่วนของความแปรปรวนมากที่สุดและน้อยที่สุด โดยแบ่งเป็นช่วงยีน 99 ช่วงยีน

	ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่า V(X) มากที่สุด		ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่า V(X) น้อยที่สุด	
ช่วงของยีน (Gene Range)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)
1-27	94.12	5.88	61.76	38.24
1-28	94.12	5.88	67.65	32.35
1-29	88.24	11.76	70.59	29.41
1-30	73.53	26.47	70.59	29.41
1-31	76.47	23.53	70.59	29.41
1-32	70.59	29.41	70.59	29.41
1-33	64.71	35.29	70.59	29.41
1-34	52.94	47.06	64.71	35.29
1-35	52.94	47.06	67.65	32.35
1-36	82.35	17.65	67.65	32.35
1-37	76.47	23.53	67.65	32.35
1-38	64.71	35.29	67.65	32.35
1-39	70.59	29.41	55.88	44.12
1-40	76.47	23.53	61.76	38.24
1-41	76.47	23.53	61.76	38.24
1-42	85.29	14.71	67.65	32.35
1-43	85.29	14.71	76.47	23.53
1-44	85.29	14.71	64.71	35.29
1-45	88.24	11.76	64.71	35.29
1-46	85.29	14.71	47.06	52.94
1-47	88.24	11.76	64.71	35.29
1-48	85.29	14.71	67.65	32.35
1-49	82.35	17.65	64.71	35.29
1-50	82.35	17.65	64.71	35.29
1-51	85.29	14.71	67.65	32.35
1-52	91.18	8.82	67.65	32.35
1-53	85.29	14.71	61.76	38.24
1-54	88.24	11.76	67.65	32.35
1-55	85.29	14.71	61.76	38.24
1-56	85.29	14.71	61.76	38.24
1-57	91.18	8.82	58.82	41.18
1-58	88.24	11.76	61.76	38.24
1-59	82.35	17.65	58.82	41.18
1-60	82.35	17.65	55.88	44.12
1-61	85.29	14.71	58.82	41.18
1-62	85.29	14.71	61.76	38.24
1-63	85.29	14.71	61.76	38.24

ตาราง 5.9 (ต่อ) ผลสรุปของการวิเคราะห์การจำแนกประเภทของกลุ่มยีนที่มีค่าสัดส่วนของความแปรปรวนมากที่สุดและน้อยที่สุด โดยแบ่งเป็นช่วงยีน 99 ช่วงยีน

	ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่า V(X) มากที่สุด		ผลการวิเคราะห์โดยใช้กลุ่มยีนที่มีค่า V(X) น้อยที่สุด	
ช่วงของยีน (Gene Range)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)	เปอร์เซ็นต์การทำนายผิด (Error)
1-64	88.24	11.76	70.59	29.41
1-65	91.18	8.82	70.59	29.41
1-66	91.18	8.82	64.71	35.29
1-67	91.18	8.82	47.06	52.94
1-68	91.18	8.82	35.29	64.71
1-69	91.18	8.82	38.24	61.76
1-70	91.18	8.82	38.24	61.76
1-71	88.24	11.76	44.12	55.88
1-72	88.24	11.76	41.18	58.82
1-73	88.24	11.76	58.82	41.18
1-74	88.24	11.76	52.94	47.06
1-75	97.06	2.94	61.76	38.24
1-76	91.18	8.82	61.76	38.24
1-77	94.12	5.88	58.82	41.18
1-78	91.18	8.82	52.94	47.06
1-79	94.12	5.88	47.06	52.94
1-80	94.12	5.88	58.82	41.18
1-81	94.12	5.88	50.00	50.00
1-82	94.12	5.88	55.88	44.12
1-83	94.12	5.88	64.71	35.29
1-84	91.18	8.82	55.88	44.12
1-85	94.12	5.88	58.82	41.18
1-86	91.18	8.82	58.82	41.18
1-87	94.12	5.88	61.76	38.24
1-88	91.18	8.82	50.00	50.00
1-89	91.18	8.82	52.94	47.06
1-90	91.18	8.82	52.94	47.06
1-91	91.18	8.82	52.94	47.06
1-92	88.24	11.76	41.18	58.82
1-93	94.12	5.88	47.06	52.94
1-94	94.12	5.88	52.94	47.06
1-95	94.12	5.88	58.82	41.18
1-96	94.12	5.88	52.94	47.06
1-97	97.06	2.94	52.94	47.06
1-98	97.06	2.94	52.94	47.06
1-99	97.06	2.94	58.82	41.18
1-100	97.06	2.94	58.82	41.18

ผลจากตาราง 5.9 พบว่ากลุ่มยีนที่มีค่าสัดส่วนของความแปรปรวนของข้อมูลมากที่สุด เมื่อนำไปใช้เป็นตัวแปรจำแนกประเภทโดยการวิเคราะห์การจำแนกประเภท จะให้ผลการทำนายที่ดีกว่ากลุ่มยีนที่มีค่าสัดส่วนของความแปรปรวนของข้อมูลน้อยที่สุด แต่จำนวนยีนจำนวนเท่าใดจึงจะเหมาะสมเป็นสิ่งที่ต้องศึกษาต่อไป และผลที่ได้จากตาราง เมื่อนำค่าเปอร์เซ็นต์ความผิดพลาดจากการทำนาย (Error of Prediction) มาพล็อตในกราฟ จะแสดงได้ดังรูป 5.8



รูป 5.8 กราฟผลสรุปของการวิเคราะห์การจำแนกประเภทของกลุ่มยีนที่มีค่าสัดส่วนของความแปรปรวนของข้อมูลมากที่สุด และน้อยที่สุด โดยแบ่งเป็นช่วงยีน 99 ช่วงยีน

จากรูป 5.8 จะเห็นว่าช่วงของยีนที่มีค่าสัดส่วนของความแปรปรวนของข้อมูลมาก มีแนวโน้มในสัดส่วนที่สูงที่เปอร์เซ็นต์ความผิดพลาดจากการทำนาย จะต่ำกว่ากลุ่มยีนที่มีค่าสัดส่วนของความแปรปรวนของข้อมูลน้อย แต่ทั้งนี้เนื่องจากกราฟมีลักษณะขึ้นลง ๆ ไม่แน่นอน และยีนบางช่วงที่มีค่าสัดส่วนของความแปรปรวนของข้อมูลต่ำ กลับมีเปอร์เซ็นต์ความผิดพลาดสูงกว่า ช่วงของยีนที่มีค่าสัดส่วนของความแปรปรวนของข้อมูลสูง กระบวนการคัดเลือกยีน โดยวิธีวิเคราะห์ค่าสัดส่วนความแปรปรวนของข้อมูล เพื่อให้ได้ ยีน และจำนวนยีนที่เหมาะสมจริงๆ กับการจำแนกประเภทจำเป็นที่ต้องมีศึกษา และปรับปรุงวิธีการให้มีประสิทธิภาพต่อไป เช่นเดียวกับการวิเคราะห์ค่าเอนโทรปี

- **การทดลองที่ 3**

โดยวิธีนี้ ตัวแปรที่ใช้สำหรับการจำแนกประเภทจะต่างกับสองวิธีที่ผ่านมา ซึ่งใช้ยีนเป็นตัวแปร แต่ในการทดลองนี้ ตัวแปรที่ใช้ในการวิเคราะห์ คือ องค์ประกอบหลักของข้อมูล ทั้งนี้ องค์ประกอบหลักของข้อมูล ถือว่าเป็นตัวแปรที่สร้างขึ้นใหม่แทนตัวแปร ยีน อย่างเดิม และการสร้างองค์ประกอบหลัก จะอาศัยวิธีการวิเคราะห์องค์ประกอบหลัก ในการหาองค์ประกอบหลักของยีนทั้ง 3,051 ยีน นอกจากนี้ ชุดข้อมูลที่จะใช้ในการสร้างองค์ประกอบหลัก จะเป็นชุดข้อมูลที่ใช้ในการสร้างโมเดลการวิเคราะห์การจำแนกประเภท ซึ่งมีจำนวน 38 ตัวอย่าง

ผลจากการวิเคราะห์องค์ประกอบหลัก ทำให้ได้องค์ประกอบหลักที่เป็นตัวแทนของยีนทั้งหมดด้วยค่าความแปรปรวนต่างๆ แสดงได้ดังตาราง 5.10

ตาราง 5.10 องค์ประกอบหลักของยีน และความแปรปรวน ของแต่ละองค์ประกอบ จำนวน 36 องค์ประกอบหลัก

องค์ประกอบหลักที่ (PC)	% ความแปรปรวน	% ความแปรปรวนสะสม	องค์ประกอบหลักที่ (PC)	% ความแปรปรวน	% ความแปรปรวนสะสม
1	16.19	16.19	19	1.44	80.41
2	13.71	29.89	20	1.39	81.80
3	7.87	37.77	21	1.37	83.17
4	5.97	43.74	22	1.33	84.50
5	4.75	48.49	23	1.28	85.78
6	3.84	52.33	24	1.23	87.01
7	3.34	55.67	25	1.21	88.22
8	3.19	58.85	26	1.17	89.39
9	2.68	61.54	27	1.14	90.53
10	2.59	64.13	28	1.10	91.64
11	2.29	66.41	29	1.08	92.72
12	2.15	68.56	30	1.07	93.78
13	1.90	70.47	31	1.01	94.80
14	1.87	72.33	32	0.93	95.73
15	1.79	74.13	33	0.92	96.65
16	1.66	75.79	34	0.88	97.53
17	1.63	77.42	35	0.86	98.39
18	1.55	78.97	36	0.84	99.23

จากตาราง 5.10 จะเห็นว่าค่าความแปรปรวนขององค์ประกอบหลักในระดับแรกๆ จะมีค่าสูง ซึ่งถือได้ว่า องค์ประกอบหลักเหล่านี้จะเป็นตัวแทนของข้อมูล หรือตัวแปร ของข้อมูล ได้ดี ทั้งนี้ในการวิเคราะห์การจำแนกประเภทในการทดลองนี้จะใช้องค์ประกอบหลักของข้อมูลที่ 36 องค์ประกอบหลัก

แรก ซึ่งมีความแปรปรวนของข้อมูลทั้งหมดถึง 99.23 เปอร์เซ็นต์ มาใช้เป็นตัวแปรสำหรับจำแนกประเภท

เมตริกซ์ข้อมูลของคะแนนองค์ประกอบหลัก ที่ได้จากองค์ประกอบทั้ง 36 องค์ประกอบหลักในชุดข้อมูลสำหรับสร้างโมเดลการจำแนกประเภท จะแสดงได้ดังตาราง 5.11 และเมตริกซ์ข้อมูลของคะแนนองค์ประกอบหลักในชุดข้อมูลทดสอบจะแสดงได้ดังตาราง 5.12

ตาราง 5.11 เมตริกซ์คะแนนองค์ประกอบหลักของชุดข้อมูลสำหรับสร้างโมเดลการจำแนกประเภท

Sample	PC1	PC2	....	PC36
1	-106.82	57.18		-67.16
2	-76.02	-11.512		-5.69
3	-222.49	124.42		54.11
4	-46.39	19.62	⋮	17.27
5	-113.04	-230.09		-16.60
...	...	...		...
37	59.17	211.44		-11.73
38	41.81	173.44		1.13

ตาราง 5.12 เมตริกซ์คะแนนองค์ประกอบหลักของชุดข้อมูลสำหรับทดสอบ

Sample	PC1	PC2	....	PC36
39	68.75	41.77		18.65
40	81.67	21.76		39.88
41	15.80	-288.21		40.49
42	10.41	17.91	⋮	24.85
43	118.31	-92.06		-1.03
...	...	...		...
71	65.52	-53.63		56.14
72	-56.09	-77.44		61.52

ผลจากตาราง 5.11 และ ตาราง 5.12 แสดงให้เห็นว่าข้อมูลการแสดงออกของยีนชุดเดิมนั้น มีการแปลงค่าให้อยู่ในรูปของคะแนนองค์ประกอบหลัก ในมิติที่น้อยลง ซึ่งจะเป็นผลดีต่อการสร้างโมเดลการจำแนกประเภท และ การทำนายกลุ่มของตัวอย่างข้อมูล

จากการวิเคราะห์การจำแนกประเภท โดยอาศัยคะแนนองค์ประกอบหลักของชุดข้อมูลที่แสดงในตาราง 5.11 จะทำให้ค่าพารามิเตอร์ที่สำคัญหรือ ค่าน้ำหนักการจำแนกประเภท แสดงได้ดังตาราง 5.13

ตาราง 5.13 ค่าน้ำหนักการจำแนกประเภทขององค์ประกอบหลัก 36 องค์ประกอบหลักที่มีค่าความแปรปรวนสูงที่สุด

PC (X_j)	Discriminant Weight (ω_j)	PC (X_j)	Discriminant Weight (ω_j)	PC (X_j)	Discriminant Weight (ω_j)
1	0.089	13	-0.114	25	-0.042
2	0.311	14	0.006	26	0.070
3	0.356	15	0.134	27	-0.153
4	0.037	16	-0.008	28	-0.009
5	-0.012	17	-0.049	29	-0.059
6	-0.096	18	0.017	30	0.052
7	0.143	19	-0.028	31	0.026
8	-0.178	20	-0.080	32	0.033
9	-0.182	21	-0.035	33	-0.016
10	0.155	22	-0.158	34	0.027
11	0.184	23	-0.106	35	-0.128
12	-0.141	24	-0.107	36	0.049

ผลจากการทำนายกลุ่มของตัวอย่างข้อมูลใน 34 ตัวอย่างในชุดข้อมูลทดสอบ โดยใช้องค์ประกอบหลัก 36 องค์ประกอบหลักเป็นตัวแปร จะแสดงได้ดังตาราง 5.14

ตาราง 5.14 ผลการจำแนกประเภทข้อมูล จากการวิเคราะห์การจำแนกประเภทโดยอาศัย 36 องค์ประกอบหลักแรก เป็นตัวแปร

Samples	Original Class	Results of LDA	Samples	Original Class	Results of LDA
39	ALL	ALL	56	ALL	ALL
40	ALL	ALL	57	AML	AML
41	ALL	ALL	59	AML	AML
42	ALL	ALL	59	ALL	ALL
43	ALL	ALL	60	AML	AML
44	ALL	ALL	61	AML	AML
45	ALL	ALL	62	AML	AML
46	ALL	ALL	63	AML	AML
47	ALL	ALL	64	AML	AML
48	ALL	ALL	65	AML	AML
49	ALL	ALL	*66	AML	ALL
50	AML	AML	67	ALL	ALL
51	AML	AML	68	ALL	ALL
52	AML	AML	69	ALL	ALL
53	AML	AML	70	ALL	ALL
54	AML	AML	71	ALL	ALL
55	ALL	ALL	72	ALL	ALL
			Correct: 97.06 %		
			Error: 2.94 %		

จากตาราง 5.14 แสดงให้เห็นผลของการทำนายกลุ่มตัวอย่างข้อมูล พบว่า กลุ่มข้อมูลที่ทำนายนั้นทำนายผิดจากกลุ่มที่กำหนดมาตอนต้นเพียง 1 ตัวอย่างข้อมูล หรือคิดเป็น 2.94 เปอร์เซ็นต์

และจากจำนวนองค์ประกอบหลักที่พิจารณาที่ 36 องค์ประกอบแรกนั้น เมื่อทดลองจำแนกประเภทข้อมูล โดยเลือกองค์ประกอบของข้อมูลใหม่ เป็นช่วงขององค์ประกอบหลักต่างๆ มาใช้เป็นตัวแปร ผลของการทำนายจะคิดเป็นเปอร์เซ็นต์การทำนายกลุ่มข้อมูลที่ผิดและถูก เมื่อเทียบกับกลุ่มของตัวอย่างข้อมูลในเริ่มต้น แสดงได้ดังตาราง 5.15

ตาราง 5.15 ผลสรุปของการวิเคราะห์การจำแนกประเภทในชุดข้อมูลโดยอาศัยองค์ประกอบหลักเป็นตัวแปรในช่วงองค์ประกอบหลักต่างๆ 35 ช่วง

ช่วงขององค์ประกอบหลัก (PC Range)	% ความแปรปรวนสะสม	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)
1-2	29.89	14.71	85.29
1-3	37.77	8.82	91.18
1-4	43.74	8.82	91.18
1-5	48.49	8.82	91.18
1-6	52.33	8.82	91.18
1-7	55.67	2.94	97.06
1-8	58.85	8.82	91.18
1-9	61.54	5.88	94.12
1-10	64.13	2.94	97.06
1-11	66.41	2.94	97.06
1-12	68.56	2.94	97.06
1-13	70.47	2.94	97.06
1-14	72.33	2.94	97.06
1-15	74.13	5.88	94.12
1-16	75.79	5.88	94.12
1-17	77.42	5.88	94.12
1-18	78.97	5.88	94.12
1-19	80.41	5.88	94.12
1-20	81.80	5.88	94.12
1-21	83.17	5.88	94.12
1-22	84.50	5.88	94.12
1-23	85.78	5.88	94.12
1-24	87.01	5.88	94.12
1-25	88.22	5.88	94.12
1-26	89.39	5.88	94.12
1-27	90.53	2.94	97.06
1-28	91.64	2.94	97.06
1-29	92.72	2.94	97.06

ตาราง 5.15(ต่อ) ผลสรุปของการวิเคราะห์การจำแนกประเภทในชุดข้อมูลโดยอาศัยองค์ประกอบหลักแรกๆ เป็นตัวแปรในช่วงองค์ประกอบหลักต่างๆ 35 ช่วง

ช่วงขององค์ประกอบหลัก (PC Range)	% ความแปรปรวนสะสม	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)
1-30	93.78	2.94	97.06
1-31	94.80	2.94	97.06
1-32	95.73	2.94	97.06
1-33	96.65	2.94	97.06
1-34	97.53	2.94	97.06
1-35	98.39	2.94	97.06
1-36	99.23	2.94	97.06

จากตาราง 5.15 แสดงผลการทำนายการวิเคราะห์การจำแนกประเภท โดยใช้ช่วงขององค์ประกอบหลักแรกๆ เป็นตัวแปร พบว่าให้ผลการทำนายที่มีเปอร์เซ็นต์การทำนายถูกตามกลุ่มที่กำหนดมาตอนต้นในระดับที่สูง เช่น ช่วงองค์ประกอบหลัก 1-27 ให้เปอร์เซ็นต์การทำนายที่ถูกถึง 97.06 เปอร์เซ็นต์ เป็นต้น แม้ว่าที่ 2 องค์ประกอบหลักแรกจะให้ผลการทำนายที่มีเปอร์เซ็นต์การทำนายผิดที่สูงอยู่ก็ตาม ซึ่งสาเหตุอาจจะเป็นเพราะว่าความแปรปรวนของข้อมูลในองค์ประกอบหลักทั้ง 2 องค์ประกอบหลักนั้น มีค่าที่น้อยเพียง 29.89 เปอร์เซ็นต์ ซึ่งมีผลทำให้องค์ประกอบหลักทั้ง 2 เป็นตัวแทนของข้อมูลทั้งหมดไม่ดีนัก การใช้องค์ประกอบหลักนี้ ในการวิเคราะห์จึงให้ผลการทำนายที่ผิดพลาดมากดังกล่าว

นอกจากนี้ ผลจากการทำนายกลุ่มของตัวอย่างข้อมูล โดยใช้ช่วงขององค์ประกอบหลัก ที่มีค่าความแปรปรวนของข้อมูลน้อยๆ เป็นตัวแปร จะแสดงผลได้ดัง ตาราง 5.16

ตาราง 5.16 ผลสรุปของการวิเคราะห์การจำแนกประเภทในชุดข้อมูลโดยอาศัยองค์ประกอบหลักที่มีค่าความแปรปรวนน้อยๆ เป็นตัวแปรในช่วงองค์ประกอบหลักต่างๆ 19 ช่วง

ช่วงขององค์ประกอบหลัก (PC Range)	% ความแปรปรวนสะสม	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)
35-36	1.70	41.18	58.82
34-36	2.58	41.18	58.82
33-36	3.50	41.18	58.82
32-36	4.43	41.18	58.82
31-36	5.44	41.18	58.82
30-36	6.51	41.18	58.82

ตาราง 5.16(ต่อ) ผลสรุปของการวิเคราะห์การจำแนกประเภทในชุดข้อมูลโดยอาศัยองค์ประกอบหลักที่มีค่าความแปรปรวนน้อยๆ เป็นตัวแปรในช่วงองค์ประกอบหลักต่างๆ 19 ช่วง

ช่วงขององค์ประกอบหลัก (PC Range)	% ความแปรปรวนสะสม	เปอร์เซ็นต์การทำนายผิด (Error)	เปอร์เซ็นต์การทำนายถูก (Correct)
29-36	7.59	41.18	58.82
28-36	8.69	41.18	58.82
27-36	9.83	41.18	58.82
26-36	11.00	41.18	58.82
25-36	12.21	41.18	58.82
24-36	13.44	41.18	58.82
23-36	14.72	41.18	58.82
22-36	16.05	41.18	58.82
21-36	17.42	41.18	58.82
20-36	18.81	41.18	58.82
19-36	20.25	41.18	58.82
18-36	21.80	41.18	58.82
17-36	23.43	41.18	58.82

จากตาราง 5.16 แสดงให้เห็นผลการทำนายกลุ่มข้อมูล ที่มีเปอร์เซ็นต์การทำนายที่ผิดจากข้อมูลเดิมอยู่ในระดับที่สูง

ข้อสรุปจากตาราง 5.15 และ 5.16 สรุปได้ว่า องค์ประกอบหลัก ในระดับแรกๆ ซึ่งมีค่าความแปรปรวนของข้อมูลสูงมากๆ เป็นองค์ประกอบหลักที่ใช้เป็นตัวแปร ในการวิเคราะห์การจำแนกประเภทได้ดี กว่าองค์ประกอบหลักที่มีความแปรปรวนของข้อมูลต่ำมากๆ อย่างไรก็ตาม จากตาราง 5.15 จะพบว่ามีบางช่วงองค์ประกอบ ที่แม้จะให้ค่าความแปรปรวนของข้อมูลสูง แต่เปอร์เซ็นต์การทำนายผิด ก็ยังมากกว่า ช่วงองค์ประกอบอื่นที่มีค่าความแปรปรวนสะสมน้อยกว่า การหาว่า จำนวนองค์ประกอบจำนวนเท่าใดเหมาะสมเป็นตัวแปรทำนายการจำแนกประเภทนั้นเป็นสิ่งที่ต้องศึกษาต่อไป

5.3 วิจารณ์และสรุปผล

แนวทางการประยุกต์วิธีวิเคราะห์การจำแนกประเภทกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ในงานวิจัยนี้คือ ใช้การวิเคราะห์การจำแนกประเภท สำหรับวิเคราะห์ข้อมูลดีเอ็นเอไมโครเรย์ของมะเร็งลิวคีเมีย เพื่อจำแนกและทำนายประเภทโรคมะเร็งของตัวอย่างข้อมูลในชุดข้อมูลทดสอบโดยใช้ยีนเป็นตัวแปร ทั้งนี้แยกการวิเคราะห์ข้อมูลออกเป็น 3 การทดลอง ตามวิธีการเลือกตัวแปรจำแนกประเภทหรือการคัดเลือกยีนที่จำเป็นสำหรับการจำแนกประเภท ได้แก่ วิธีการวิเคราะห์ค่าเอนโทรปี วิธีวิเคราะห์ค่าความแปรปรวนและวิธีวิเคราะห์องค์ประกอบหลัก ผลการวิเคราะห์ที่ได้นำไปเปรียบเทียบผลการทำนายกับ ประเภทของตัวอย่างข้อมูล ในชุดข้อมูลทดสอบ ที่กำหนดมาตอนต้น

ผลการวิเคราะห์ที่ได้ทั้ง 3 การทดลอง สรุปได้ดังนี้

1) ผลการวิเคราะห์การจำแนกประเภท โดยใช้วิธีการวิเคราะห์ค่าเอนโทรปีในการเลือกยีน พบว่า การวิเคราะห์การจำแนกประเภทจะให้ผลการทำนายกลุ่มข้อมูลได้ดี เมื่อยีนที่นำมาใช้เป็นตัวแปรจำแนกประเภทเป็นยีนที่มีค่าประมาณของเอนโทรปีต่ำมากๆ ซึ่งผลการวิเคราะห์ ที่ได้เมื่อใช้ยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยที่สุดจำนวน 100 ยีน เป็นตัวแปรจำแนกประเภท พบว่าผลการทำนายกลุ่มข้อมูลมีความถูกต้องถึง 88.24 เปอร์เซ็นต์ แต่ข้อจำกัดของวิธีการนี้ก็คือ ไม่สามารถระบุจำนวนของยีนที่เหมาะสม นอกจากนี้ จากผลการเปรียบเทียบผลการทำนาย โดยใช้กลุ่มของยีนในหลายๆช่วงยีนเป็นตัวแปร ไม่สามารถสรุปได้ว่ากลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีน้อยๆ จะให้ผลการวิเคราะห์ได้ดีกว่ากลุ่มยีนที่มีค่าประมาณของค่าเอนโทรปีมากๆ เสมอไป

2) ผลการวิเคราะห์การจำแนกประเภท โดยใช้วิธีการวิเคราะห์ค่าความแปรปรวนในการเลือกยีน พบว่า การวิเคราะห์การจำแนกประเภทจะให้ผลการทำนายกลุ่มข้อมูลได้ดีเมื่อยีนที่นำมาใช้เป็นตัวแปรจำแนกประเภทเป็นยีนที่มีค่าสัดส่วนของค่าความแปรปรวนมากที่สุด ซึ่งผลการวิเคราะห์ที่ได้ เมื่อใช้ยีนที่มีค่าสัดส่วนของค่าความแปรปรวนมากที่สุดจำนวน 100 ยีน พบว่าผลการทำนายกลุ่มข้อมูลมีความถูกต้องถึง 97.06 เปอร์เซ็นต์ แต่ข้อจำกัดของวิธีการนี้ก็คือ ไม่สามารถระบุจำนวนของยีนที่เหมาะสมได้ นอกจากนี้ จากผลการเปรียบเทียบผลการทำนาย โดยใช้กลุ่มของยีนในหลายๆช่วงยีนเป็นตัวแปร ยังพบว่ากลุ่มยีนที่มีค่าสัดส่วนของค่าความแปรปรวนมากไม่ได้ให้ผลการวิเคราะห์ดีกว่ากลุ่มยีนที่มีค่าสัดส่วนของค่าความแปรปรวนที่น้อยกว่า เสมอไป

3) ผลการวิเคราะห์การจำแนกประเภทโดยตัวแปรที่สร้างขึ้นด้วยวิธีการวิเคราะห์องค์ประกอบหลัก พบว่า องค์ประกอบหลักที่มีค่าความแปรปรวนมากๆ จะสามารถใช้เป็นตัวแปรจำแนกประเภทได้ ซึ่งผลการวิเคราะห์โดยใช้องค์ประกอบหลักดังกล่าว จะให้ผลการวิเคราะห์ที่มีความถูกต้องสูง ซึ่งจากการวิเคราะห์ที่ 36 องค์ประกอบหลักซึ่งมีความแปรปรวน 99.23 เปอร์เซ็นต์จะให้ผลการทำนายกลุ่มข้อมูลที่ถูกต้องถึง 97.06 เปอร์เซ็นต์ ทั้งนี้ข้อดีของการวิเคราะห์องค์ประกอบหลักก็คือตัวแปรที่ใช้ในการจำแนกประเภทของข้อมูลนั้น ถือได้ว่าเป็นตัวแทนของตัวแปรในข้อมูลทั้งหมด ด้วยค่าความแปรปรวนที่สูง ปัญหาในเรื่องของการเลือกจำนวนยีน หรือจำนวนตัวแปรที่เหมาะสม นั้นจึงไม่มี

จากผลการวิเคราะห์จำแนกประเภททั้ง 3 การทดลองพบว่า เทคนิคการวิเคราะห์จำแนกประเภทให้ผลการวิเคราะห์ที่ดีหรือไม่นั้นขึ้นกับตัวแปรจำแนกประเภท ซึ่งถ้าก่อนการวิเคราะห์มีการเลือกตัวแปรจำแนกประเภทที่ดีแล้ว ผลการทำนายกลุ่มของข้อมูลย่อมให้ผลการทำนายที่ดีตามไปด้วย ทั้งนี้ นอกจากเทคนิควิธีการที่ใช้ในการเลือกตัวแปรจำแนกประเภทดังที่นำเสนอ ยังมีหลายวิธีการที่นำเสนอในผลงานวิจัยต่างๆ ที่จำเป็นจะต้องมีการวิเคราะห์เปรียบเทียบผลต่อไป