บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน มีวัตถุประสงค์เพื่อจำแนกหมวดหมู่ข้อความของข้อคิดเห็นในแบบสอบถามปลายเปิด และทดสอบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด ด้วยวิธีเนอีฟเบย์และซัพพอร์ตเวกเตอร์แมชชีน มีเอกสารและงานวิจัยที่เกี่ยวข้องประกอบด้วยหัวข้อดังต่อไปนี้

2.1 แบบสอบถาม (Questionnaire)

2.2 กระบวนการเตรียมข้อมูล (Preprocessing)

2.3 การเรียนรู้และจำแนกหมวดหมู่

2.4 ซัพพอร์ตเวคเตอร์แมชชีน

2.5 เนอีฟเบย์

2.6 งานวิจัยที่เกี่ยวข้อง

2.1 แบบสอบถาม (Questionnaire)

แบบสอบถาม หมายถึง รูปแบบของคำถามเป็นชุด ๆ ที่ได้ถูกรวบรวมไว้อย่างมีหลักเกณฑ์และเป็นระบบ เพื่อใช้วัดสิ่งที่ผู้วิจัยต้องการจะวัดจากกลุ่มตัวอย่างหรือประชากรเป้าหมายให้ได้มาซึ่งข้อเท็จจริงทั้งในอดีต ปัจจุบันและการคาดคะเนเหตุการณ์ในอนาคต แบบสอบถามประกอบด้วยรายการคำถามที่สร้างอย่างประณีต เพื่อรวบรวมข้อมูลเกี่ยวกับความคิดเห็นหรือข้อเท็จจริง โดยส่งให้กลุ่มตัวอย่างตามความสมัครใจ การใช้แบบสอบถามเป็นเครื่องมือในการเก็บรวบรวมข้อมูลนั้น การสร้างคำถามเป็นงานที่สำคัญสำหรับผู้วิจัย เพราะว่าผู้วิจัยอาจไม่มีโอกาสได้พบปะกับผู้ตอบแบบสอบถามเพื่ออธิบายความหมายต่าง ๆ ของข้อคำถามที่ต้องการเก็บรวบรวม

แบบสอบถาม เป็นเครื่องมือวิจัยชนิดหนึ่งที่นิยมใช้กันมาก เพราะการเก็บรวบรวมข้อมูลสะดวกและสามารถใช้วัดได้อย่างกว้างขวาง การเก็บข้อมูลด้วยแบบสอบถามสามารถทำได้ด้วยการสัมภาษณ์หรือให้ผู้ตอบด้วยตนเอง

2.1.1 โครงสร้างของแบบสอบถาม (มารยาท โยทองยศ, 2554)

โครงสร้างของแบบสอบถาม ประกอบไปด้วย 3 ส่วนสำคัญ ดังนี้

2.1.1.1 หนังสือนำหรือคำชี้แจง โดยมากมักจะอยู่ส่วนแรกของแบบสอบถาม อาจมีจดหมายนำอยู่ด้านหน้าพร้อมคำขอบคุณ โดยคำชี้แจงมักจะระบุถึงจุดประสงค์ที่ให้ตอบแบบสอบถาม การนำคำตอบที่ได้ไปใช้ประโยชน์ คำอธิบายลักษณะของแบบสอบถาม วิธีการตอบแบบสอบถามพร้อมตัวอย่าง ชื่อ และที่อยู่ของผู้วิจัย ประเด็นที่สำคัญคือการแสดงข้อความที่ทำให้ผู้ตอบมีความมั่นใจว่า ข้อมูลที่จะตอบไปจะไม่ถูกเปิดเผยเป็นรายบุคคล จะไม่มีผลกระทบต่อผู้ตอบ และมีการพิทักษ์สิทธิของผู้ตอบด้วย

2.1.1.2 คำถามเกี่ยวกับข้อมูลส่วนตัว เช่น เพศ อายุ ระดับการศึกษา อาชีพ เป็นต้น การที่จะถามข้อมูลส่วนตัวอะไรบ้างนั้นขึ้นอยู่กับกรอบแนวความคิดในการวิจัย โดยดูว่าตัวแปรที่สนใจจะศึกษานั้นมีอะไรบ้างที่เกี่ยวกับข้อมูลส่วนตัว และควรถามเฉพาะข้อมูลที่จำเป็นในการวิจัยเท่านั้น

2.1.1.3 คำถามเกี่ยวกับคุณลักษณะหรือตัวแปรที่จะวัด เป็นความคิดเห็นของผู้ตอบในเรื่องของคุณลักษณะ หรือตัวแปรนั้น

2.1.2 ขั้นตอนการสร้างแบบสอบถาม (พิชิต ฤทธิ์จรูญ, 2544)

การสร้างแบบสอบถามประกอบไปด้วยขั้นตอนสำคัญ ดังนี้

ขั้นที่ 1 ศึกษาคุณลักษณะที่จะวัด

การศึกษาคุณลักษณะอาจดูได้จาก วัตถุประสงค์ของการวิจัย กรอบแนวความคิดหรือสมมติฐานการวิจัย จากนั้นจึงศึกษาคุณลักษณะ หรือตัวแปรที่จะวัดให้เข้าใจอย่างละเอียดทั้งเชิงทฤษฎีและนิยามเชิงปฏิบัติการ

ขั้นที่ 2 กำหนดประเภทของข้อคำถาม

ข้อคำถามในแบบสอบถามอาจแบ่งได้เป็น 2 ประเภท คือ

1. คำถามปลายเปิด (Open Ended Question) เป็นคำถามที่เปิดโอกาสให้ผู้ตอบสามารถตอบได้อย่างไม่กำจัดวงคำตอบ ซึ่งคาดว่าน่าจะได้คำตอบที่แน่นอน สมบูรณ์ ตรงกับสภาพความเป็นจริงได้มากกว่าคำตอบที่จำกัดวงให้ตอบ คำถามปลายเปิดจะนิยมใช้กันมากในกรณีที่ผู้วิจัยไม่สามารถคาดเดาได้ล่วงหน้าว่าคำตอบจะเป็นอย่างไร หรือใช้คำถามปลายเปิดในกรณีที่ต้องการได้คำตอบเพื่อนำมาเป็นแนวทางในการสร้างคำถามปลายปิด แบบสอบถามแบบนี้มีข้อเสียคือ มักจะถามได้ไม่มากนัก การรวบรวมความคิดเห็นและการแปลผลมักจะมีความยุ่งยาก

2. คำถามปลายปิด (Closed Question) เป็นคำถามที่ผู้วิจัยมีแนวคำตอบไว้ให้

ผู้ตอบเลือกตอบจากคำตอบที่กำหนดไว้เท่านั้น คำตอบที่ผู้วิจัยกำหนดไว้ล่วงหน้ามักได้มาจากการทดลองใช้คำถามในลักษณะที่เป็นคำถามปลายเปิด หรือการศึกษากรอบแนวความคิดสมมติฐานการวิจัย และนิยามเชิงปฏิบัติการ คำถามปลายเปิดมีวิธีการเขียนได้หลาย ๆ แบบ เช่น

แบบให้เลือกตอบอย่างใดอย่างหนึ่ง แบบให้เลือกคำตอบที่ถูกต้องเพียงคำตอบเดียว แบบผู้ตอบจัดลำดับความสำคัญหรือแบบให้เลือกคำตอบหายคำตอบ

ขั้นที่ 3 การร่างแบบสอบถาม

เมื่อผู้วิจัยทราบถึงคุณลักษณะหรือประเด็นที่จะวัด และกำหนดประเภทของข้อคำถามที่จะมีอยู่ในแบบสอบถามเรียบร้อยแล้ว ผู้วิจัยจึงลงมือเขียนข้อคำถามให้ครอบคลุมทุกคุณลักษณะหรือประเด็นที่จะวัด โดยเขียนตามโครงสร้างของแบบสอบถามที่ได้กล่าวไว้แล้ว และหลักการในการสร้างแบบสอบถาม ดังนี้

1. ต้องมีจุดมุ่งหมายที่แน่นอนว่าต้องการจะถามอะไรบ้าง โดยจุดมุ่งหมายนั้นจะต้องสอดคล้องกับวัตถุประสงค์ของงานวิจัยที่จะทำ

2. ต้องสร้างคำถามให้ตรงตามจุดมุ่งหมายที่ตั้งไว้ เพื่อป้องกันการมีข้อคำถามนอกประเด็นและมีข้อคำถามจำนวนมาก

3. ต้องถามให้ครอบคลุมเรื่องที่จะวัด โดยมีจำนวนข้อคำถามที่พอเหมาะ ไม่มากหรือน้อยเกินไป แต่จะมากหรือน้อยเท่าใดนั้นขึ้นอยู่กับพฤติกรรมที่จะวัด ซึ่งตามปกติพฤติกรรมหรือเรื่องที่จะวัดเรื่องหนึ่งๆ นั้นควรมีข้อคำถาม 25-60 ข้อ

4. การเรียงลำดับข้อคำถาม ควรเรียงลำดับให้ต่อเนื่องสัมพันธ์กัน และแบ่งตามพฤติกรรมย่อยๆ ไว้เพื่อให้ผู้ตอบเห็นชัดเจนและง่ายต่อการตอบ นอกจากนั้นต้องเรียงคำถามง่ายๆ ไว้เป็นข้อแรกๆ เพื่อชักจูงให้ผู้ตอบอยากตอบคำถามต่อ ส่วนคำถามสำคัญๆ ไม่ควรเรียงไว้ตอนท้ายของแบบสอบถาม เพราะความสนใจในการตอบของผู้ตอบอาจจะน้อยลง ทำให้ตอบอย่างไม่ตั้งใจ ซึ่งจะส่งผลเสียต่อการวิจัยมาก

5. ลักษณะของข้อความที่ดี ข้อคำถามที่ดีของแบบสอบถามนั้น ควรมีลักษณะดังนี้

1) ข้อคำถามไม่ควรยาวจนเกินไป ควรใช้ข้อความสั้น กะทัดรัด ตรงกับวัตถุประสงค์และสองคล้องกับเรื่อง

2) ข้อความ หรือภาษาที่ใช้ในข้อความต้องชัดเจน เข้าใจง่าย

3) ค่าเฉลี่ยในการตอบแบบสอบถามไม่ควรเกินหนึ่งชั่วโมง ข้อคำถามไม่ควรมากเกินไปจนทำให้ผู้ตอบเบื่อหน่ายหรือเหนื่อยล้า

4) ไม่ถามเรื่องที่เป็นความลับเพราะจะทำให้ได้คำตอบที่ไม่ตรงกับข้อเท็จจริง

5) ไม่ควรใช้ข้อความที่มีความหมายกำกวมหรือข้อความที่ทำให้ผู้ตอบแต่ละคนเข้าใจความหมายของข้อความไม่เหมือนกัน

6) ไม่ถามในเรื่องที่รู้แล้ว หรือถามในสิ่งที่วัดได้ด้วยวิธีอื่น

7) ข้อคำถามต้องเหมาะสมกับกลุ่มตัวอย่าง คือ ต้องคำนึงถึงระดับการศึกษา ความ

สนใจ สภาพเศรษฐกิจ ฯลฯ

8) ข้อคำถามหนึ่งๆ ควรถามเพียงประเด็นเดียว เพื่อให้ได้คำตอบที่ชัดเจนและตรงจุดซึ่งจะง่ายต่อการนำมาวิเคราะห์ข้อมูล

9) คำตอบหรือตัวเลือกในข้อคำถามควรมีมากพอ หรือให้เหมาะสมกับข้อคำถามนั้น แต่ถ้าไม่สามารถระบุได้หมดก็ให้ใช้ว่า อื่นๆ โปรดระบุ .....

10) ควรหลีกเลี่ยงคำถามที่เกี่ยวกับค่านิยมที่จะทำให้ผู้ตอบไม่ตอบตามความเป็นจริง

11) คำตอบที่ได้จากแบบสอบถาม ต้องสามารถนำมาแปลงออกมาในรูปของปริมาณและใช้สถิติอธิบายข้อเท็จจริงได้ เพราะปัจจุบันนี้นิยมใช้คอมพิวเตอร์ในการวิเคราะห์ข้อมูล ดังนั้นแบบสอบถามควรคำนึงถึงวิธีการประมวลข้อมูลและวิเคราะห์ข้อมูลด้วยโปรแกรมคอมพิวเตอร์ด้วย

ขั้นที่ 4 การปรับปรุงแบบสอบถาม

หลังจากที่สร้างแบบสอบถามเสร็จแล้ว ผู้วิจัยควรนำแบบสอบถามนั้นมาพิจารณาทบทวนอีกครั้งเพื่อหาข้อบกพร่องที่ควรปรับปรุงแก้ไข และควรให้ผู้เชี่ยวชาญได้ตรวจสอบแบบสอบถามนั้นด้วยเพื่อที่จะได้นำข้อเสนอแนะและข้อวิพากษ์วิจารณ์ของผู้เชี่ยวชาญมาปรับปรุงแก้ไขให้ดียิ่งขึ้น

ขั้นที่ 5 วิเคราะห์คุณภาพแบบสอบถาม

เป็นการนำแบบสอบถามที่ได้ปรับปรุงแล้วไปทดลองใช้กับกลุ่มตัวอย่างเล็กๆ เพื่อนำผลมาตรวจสอบคุณภาพของแบบสอบถาม ซึ่งการวิเคราะห์หรือตรวจสอบคุณภาพของแบบสอบถามทำได้หลายวิธี แต่ที่สำคัญมี 2 วิธี ได้แก่

1. ความตรง (Validity) หมายถึง เครื่องมือที่สามารถวัดได้ในสิ่งที่ต้องการวัด โดยแบ่งออกได้เป็น 3 ประเภท คือ

1) ความตรงตามเนื้อหา (Content Validity) คือ การที่แบบสอบถามมีความครอบคลุมวัตถุประสงค์หรือพฤติกรรมที่ต้องการวัดหรือไม่ ค่าสถิติที่ใช้ในการหาคุณภาพ คือ ค่าความสอดคล้องระหว่างข้อคำถามกับวัตถุประสงค์ หรือเนื้อหา (IOC: Index of item Objective Congruence) หรือดัชนีความเหมาะสม โดยให้ผู้เชี่ยวชาญ ประเมินเนื้อหาของข้อถามเป็นรายข้อ

2) ความตรงตามเกณฑ์ (Criterion-related Validity) หมายถึง ความสามารถของแบบวัดที่สามารถวัดได้ตรงตามสภาพความเป็นจริง แบ่งออกได้เป็นความเที่ยงตรงเชิงพยากรณ์และความเที่ยงตรงตามสภาพ สถิติที่ใช้วัดความเที่ยงตรงตามเกณฑ์ เช่น ค่าสัมประสิทธิ์สหสัมพันธ์

(Correlation Coefficient) ทั้งของเพียร์สัน (Pearson) และสเพียร์เมน (Spearman) และการทดสอบค่าที (t-test) เป็นต้น

3) ความตรงตามโครงสร้าง (Construct Validity) หมายถึงความสามารถของแบบสอบถามที่สามารถวัดได้ตรงตามโครงสร้างหรือทฤษฎี ซึ่งมักจะมีในแบบวัดทางจิตวิทยาและแบบวัดสติปัญญา สถิติที่ใช้วัดความเที่ยงตรงตามโครงสร้างมีหลายวิธี เช่น การวิเคราะห์องค์ประกอบ (Factor Analysis) การตรวจสอบในเชิงเหตุผล เป็นต้น

2. ความเที่ยง (Reliability) หมายถึง เครื่องมือที่มีความคงเส้นคงวา นั่นคือ เครื่องมือที่สร้างขึ้นให้ผลการวัดที่แน่นอนคงที่จะวัดกี่ครั้งผลจะได้เหมือนเดิม สถิติที่ใช้ในการหาค่าความเที่ยงมีหลายวิธีแต่นิยมใช้กันคือ ค่าสัมประสิทธิ์แอลฟ่าของคอนบาร์ช (Conbach's Alpha Coefficient: α coefficient) ซึ่งจะใช้สำหรับข้อมูลที่มีการแบ่งระดับการวัดแบบประมาณค่า (Rating Scale)

ขั้นที่ 6 ปรับปรุงแบบสอบถามให้สมบูรณ์

ผู้วิจัยจะต้องทำการแก้ไขข้อบกพร่องที่ได้จากผลการวิเคราะห์คุณภาพของแบบสอบถามและตรวจสอบความถูกต้องของถ้อยคำหรือสำนวน เพื่อให้แบบสอบถามมีความสมบูรณ์และมีคุณภาพผู้ตอบอ่านเข้าใจได้ตรงประเด็นที่ผู้วิจัยต้องการ ซึ่งจะทำให้ผลงานวิจัยเป็นที่น่าเชื่อถือยิ่งขึ้น

ขั้นที่ 7 จัดพิมพ์แบบสอบถาม

จัดพิมพ์แบบสอบถามที่ได้ปรับปรุงเรียบร้อยแล้วเพื่อนำไปใช้จริงในการเก็บรวบรวมข้อมูลกับกลุ่มเป้าหมาย โดยจำนวนที่จัดพิมพ์ควรไม่น้อยกว่าจำนวนเป้าหมายที่ต้องการเก็บรวบรวมข้อมูล และควรมีการพิมพ์สำรองไว้ในกรณีที่แบบสอบถามเสียหรือสูญหายหรือผู้ตอบไม่ตอบกลับ แนวทางในการจัดพิมพ์แบบสอบถามมีดังนี้

1) การพิมพ์แบ่งหน้าให้สะดวกต่อการเปิดอ่านและตอบ
2) เว้นที่ว่างสำหรับคำถามปลายเปิดไว้เพียงพอ
3) พิมพ์อักษรขนาดใหญ่ชัดเจน
4) ใช้สีและลักษณะกระดาษที่เอื้อต่อการอ่าน

2.1.3 หลักการสร้างแบบสอบถาม

1) สอดคล้องกับวัตถุประสงค์การวิจัย
2) ใช้ภาษาที่เข้าใจง่าย เหมาะสมกับผู้ตอบ
3) ใช้ข้อความที่สั้น กระทัดรัด ได้ใจความ
4) แต่ละคำถามควรมีนัย เพียงประเด็นเดียว
5) หลีกเลี่ยงการใช้ประโยคปฏิเสธซ้อน

6) ไม่ควรใช้คำย่อ

7) หลีกเลี่ยงการใช้คำที่เป็นนามธรรมมาก

8) ไม่ชี้นำการตอบให้เป็นไปแนวทางใดแนวทางหนึ่ง

9) หลีดเลี่ยงคำถามที่ทำให้ผู้ตอบเกิดความลำบากใจในการตอบ

10) คำตอบที่มีให้เลือกต้องชัดเจนและครอบคลุมคำตอบที่เป็นไปได้

11) หลีกเลี่ยงคำที่สื่อความหมายหลายอย่าง

12) ไม่ควรเป็นแบบสอบถามที่มีจำนวนมากเกินไป ไม่ควรให้ผู้ตอบใช้เวลาในการตอบแบบสอบถามนานเกินไป

13) ข้อคำถามควรถามประเด็นที่เฉพาะเจาะจงตามเป้าหมายของการวิจัย

14) คำถามต้องน่าสนใจสามารถกระตุ้นให้เกิดความอยากตอบ

2.1.4 เทคนิคการใช้แบบสอบถาม (เกียรติสุดา ศรีสุข, 2552; จินตนา ธนวิบูลย์ชัย, 2545; อุทุมพร จามรมาน, 2544; Neil J. Salkind, 2006)

วิธีใช้แบบสอบถามมี 2 วิธี คือ การส่งทางไปรษณีย์กับการเก็บข้อมูลด้วยตนเอง ซึ่งไม่ว่ากรณีใดต้องมีจดหมายระบุวัตถุประสงค์ของการเก็บข้อมูล ตลอดจนความสำคัญของข้อมูลและผลที่คาดว่าจะได้รับ เพื่อให้ผู้ตอบตระหนักถึงความสำคัญและสละเวลาในการตอบแบบสอบถาม

การทำให้อัตราตอบแบบสอบถามสูงเป็นเป้าหมายสำคัญของผู้วิจัย ข้อมูลจากแบบสอบถามจะเป็นตัวแทนของประชากรได้เมื่อมีจำนวนแบบสอบถามคืนมากว่าร้อยละ 90 ของจำนวนแบบสอบถามที่ส่งไป แนวทางที่จะทำให้ได้รับแบบสอบถามกลับคืนในอัตราที่สูง มีวิธีการดังนี้

1. มีการติดตามแบบสอบถามเมื่อให้เวลาผู้ตอบไประยะหนึ่ง ระยะเวลาที่เหมาะสมในการติดตามคือ 2 สัปดาห์ หลังครบกำหนดส่ง อาจจะติดตามมากกว่าหนึ่งครั้ง

2. วิธีการติดตามแบบสอบถาม อาจใช้จดหมาย ไปรษณีย์ โทรศัพท์ เป็นต้น

3. ในกรณีที่ขอคำถามอาจจะถามในเรื่องของส่วนตัว ผู้วิจัยต้องให้ความมั่นใจว่าข้อมูลที่ได้จะเป็นความลับ

2.1.5 ข้อเด่นและข้อด้อยของการเก็บข้อมูลโดยใช้แบบสอบถาม

การใช้แบบสอบถามในการเก็บรวบรวมข้อมูลมีข้อเด่นและข้อด้อยที่ต้องพิจารณาประกอบในการเลือกใช้แบบสอบถามในการเก็บรวบรวมข้อมูล ดังนี้

ข้อเด่นของการเก็บข้อมูลโดยใช้แบบสอบถามมีดังนี้ คือ

1. ถ้ากลุ่มตัวอย่างมีขนาดใหญ่ วิธีการเก็บข้อมูลโดยใช้แบบสอบถาม จะเป็นวิธีการที่สะดวกและประหยัดกว่าวิธีอื่น

2. ผู้ตอบมีเวลาตอบมากกว่าวิธีการอื่น

3. ไม่จำเป็นต้องฝึกอบรมพนักงานเก็บข้อมูลมากเหมือนกับวิธีการสัมภาษณ์หรือวิธีการสังเกต

4. ไม่เกิดความลำเอียงอันเนื่องมาจากการสัมภาษณ์หรือการสังเกต เพราะผู้ตอบเป็นผู้ตอบข้อมูลเอง

5. สามารถส่งแบบสอบถามให้ผู้ตอบทางไปรษณีย์ได้

6. ประหยัดค่าใช้จ่ายในการเก็บข้อมูล

ข้อด้อยของการเก็บข้อมูลโดยใช้แบบสอบถาม มีดังนี้คือ

1. ในกรณีที่ส่งแบบสอบถามให้ผู้ตอบทางไปรษณีย์ มักจะได้แบบสอบถามกลับคืนมาน้อย และต้องเสียเวลาในการติดตาม อาจทำให้ระยะเวลาการเก็บข้อมูลล่าช้ากว่าที่กำหนดไว้

2. การเก็บข้อมูลโดยวิธีการใช้แบบสอบถามจะใช้ได้เฉพาะกับกลุ่มประชากรเป้าหมายที่อ่านและเขียนหนังสือได้เท่านั้น

3. จะได้ข้อมูลจำกัดเฉพาะที่จำเป็นจริงๆ เท่านั้น เพราะการเก็บข้อมูลโดยวิธีการใช้แบบสอบถามจะต้องมีคำถามจำนวนน้อยข้อที่สุดเท่าที่จะเป็นไปได้

4. การส่งแบบสอบถามไปทางไปรษณีย์ หน่วยตัวอย่างอาจไม่ได้เป็นผู้ตอบแบบสอบถามเองก็ได้ ทำให้คำตอบที่ได้มีความคลาดเคลื่อนไม่ตรงกับความจริง

5. ถ้าผู้ตอบไม่เข้าใจคำถามหรือเข้าใจคำถามผิด หรือไม่ตอบคำถามบางข้อ หรือไม่ไตร่ตรองให้รอบคอบก่อนที่จะตอบคำถาม ก็จะทำให้ข้อมูลมีความคลาดเคลื่อนได้ โดยที่ผู้วิจัยไม่สามารถย้อนกลับไปสอบถามหน่วยตัวอย่างนั้นได้อีก

6. ผู้ที่ตอบแบบสอบถามกลับคืนมาทางไปรษณีย์ อาจเป็นกลุ่มที่มีลักษณะแตกต่างจากกลุ่มผู้ที่ไม่ตอบแบบสอบถามกลับคืนมา ดังนั้นข้อมูลที่นำมาวิเคราะห์จะมีความลำเอียงอันเนื่องมาจากกลุ่มตัวอย่างได้

2.2 กระบวนการเตรียมข้อมูล (Preprocessing)

กระบวนการเตรียมข้อมูล เป็นการเตรียมข้อมูลก่อนทำการเรียนรู้และจำแนกหมวดหมู่ประกอบด้วย

2.2.1 การตัดคำ (Word Segmentation)

การประมวลผลจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิภาพนั้น มีปัญหาเบื้องต้นคือ การตัดคำในภาษาไทย ซึ่งลักษณะการเขียนภาษาไทยจะมีการเขียนติดต่อกันเป็นสายอักขระโดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำ ดังเช่นภาษาอังกฤษซึ่งใช้ช่องว่าง (Space) คั่นระหว่างคำ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่เพื่อแบ่งสายอักขระไทยออกเป็นคำ ๆ จึงได้มีผู้คิดค้น

พัฒนาวิธีการตัดคำแบ่งได้เป็น หลักการตัดคำโดยใช้กฎ (Rule Base Approach) หลักการตัดคำโดยใช้อัลกอริทึม (Algorithm Approach) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary Approach) และหลักการตัดคำโดยใช้คลังข้อมูล (Corpus Base Approach)แต่ละวิธีการต่างๆ ก็ให้ผลในด้านความถูกต้อง ความรวดเร็วของการทำงานและปริมาณการใช้ทรัพยากรต่าง ๆ ที่แตกต่างกัน จากการศึกษาเรื่องตัดคำสำหรับการจัดหมวดหมู่เอกสารภาษาไทยพบปัญหาด้านการหาขอบเขตของคำ เนื่องจากไม่มีการเขียนแบ่งพยางค์ คำ หรือประโยค ไม่มีหลักเกณฑ์ตายตัวในการใช้ช่องว่างในภาษาเขียน การสะกดคำมีรูปแบบซับซ้อน มีคำยืม คำทับศัพท์ คำเฉพาะจำนวนมาก และคำมีความกำกวมสูง จากการศึกษาเปรียบเทียบประสิทธิภาพวิธีดังกล่าว พบว่าวิธีตัดคำที่เหมาะสมกับการจัดหมวดหมู่เอกสารคือวิธีการตัดคำแบบยาวที่สุด (Longest Matching) ซึ่งมีวิธีการตรวจสอบสายอักขระ (String) ที่เข้ามาจากซ้ายไปขวากับพยางค์ที่เก็บไว้ในพจนานกรม ในกรณีที่ตรวจสอบแล้วปรากฏว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม ก็ให้เลือกแบ่งพยางค์โดยเลือกพยางค์ที่ยาวที่สุด แล้วทำต่อไปเรื่อย ๆ จนจบสายอักขระ แต่ถ้ากรณีที่เลือกพยางค์ที่ยาวที่สุดแล้วทำให้เกิดพยางค์ที่ไม่ปรากฏในพจนานุกรมก็ยอมให้มีการย้อนรอย (Back Tracking) กลับไปเลือกพยางค์ที่ยาวรองมาแทนทำต่อไปเรื่อยๆจนสิ้นสุดสายอักขระ (Charoenpornsawat, 1999)

2.2.2 การกำจัดคำหยุด (Stop-Word List Removal)

การกำจัดคำหยุด เป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญ ในที่นี้หมายถึง คำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ประเภทของคำที่จัดว่าเป็นคำหยุดในภาษาไทย (Jaruskulchai, 1998) มีดังต่อไปนี้

1) คำบุพบท

เป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน มักใช้นำหน้าคำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อแสดงความสัมพันธ์ของคำหรือกลุ่มคำที่อยู่หลังคำบุพบทว่ามีความสัมพันธ์กับคำหรือกลุ่มคำที่อยู่หน้าคำบุพบทอย่างไร ตัวอย่างของคำบุพบท ได้แก่ ของ ใน แก่ แด่ ต่อ ตั้งแต่ โดย เมื่อ กว่า กับ เป็น ดูก่อน ช้าแต่ ทาง สู่ แค่ ฯลฯ

2) คำสันธาน

เป็นคำที่ทำหน้าที่เชื่อมคำกับคำ ประโยคกับประโยค ข้อความกับข้อความ เพื่อทำให้ประโยคนั้นกระชับ และสละสลวย โดยมักจะใช้เชื่อมใจความที่คล้อยตามกัน ใจความที่ขัดแย้งกันใจความที่เป็นเหตุเป็นผลกัน หรือใจความที่ให้เลือกอย่างใดอย่างหนึ่ง ตัวอย่างของคำสันธานได้แก่ เพราะ เพราะว่า และ หรือ จึง ดังนั้น มิฉะนั้น ทั้ง แต่ แต่ว่า ครั้น หรือไม่ก็ ฯลฯ

3) คำสรรพนาม

เป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เพื่อไม่ต้องกล่าวคำนามนั้นซ้ำอีก ตัวอย่างของคำสรรพนาม ได้แก่ ฉัน เรา เขา ดิฉัน กระผม กู คุณ ท่าน เธอ ใต้เท้า มัน สิ่ง ใคร บ้าง ต่างก็ ตัวนั้น อันโน้น อะไร ไหน ใด อย่างนี้ ฯลฯ

4) คำวิเศษณ์

เป็นคำที่ใช้ขยายคำอื่น เช่น คำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อให้มีความหมายชัดเจนและมีรายละเอียดมากยิ่งขึ้น ตัวอย่างของคำวิเศษณ์ ได้แก่ มาก น้อย ใหญ่ เล็กมหึมา มโหฬาร อ้วน โต สูง ไพเราะ เยอะ หลาย สวย หอม นุ่ม เผ็ด เช้า ค่ำ นี่ นั่น เอง ทั้งนี้ ค่ะ ครับ ขอรับ จ้า จ๊ะ วะ ไม่ หามิได้ บ่ ทำไม อย่างไร หมด อดีต ปัจจุบัน โบราณ หน้า ฯลฯ

5) คำอุทาน

เป็นคำที่แสดงอารมณ์ อาการ หรือความรู้สึกของผู้พูด รวมทั้งเป็นคำที่ใช้เสริมคำพูดตัวอย่างของคำอุทานได้แก่ เอ๊ะ อ๊ะ อ๋อ เฮ้อ ว้าย โธ่ โถ อนิจจา สิ หนอย ซะ นะ น่า แหม ตายล่ะ คุณพระช่วย ชิชะ อุวะ ไม่น่าเลย โอ้โห ฯลฯ คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสาร และปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนก หมวดหมู่ ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์ และลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล

2.2.3 การหารากศัพท์

รากศัพท์ (Stem) คือ รูปเดิมของคำที่ยังไม่ได้เติมคำอุปสรรค (Prefixes) คำปัจจัย (Suffixes) หรือยังไม่ได้เปลี่ยนรูปไป ตัวอย่างรากศัพท์ของคำแสดงในภาพที่ 2-1 และภาพที่ 2-2

คำศัพท์	รากศัพท์
organization	organize
organisational	organize
organizational	organize
organized	organize
organize	organize
organizer	organize
organizing	organize

ภาพที่ 2-1 ตัวอย่างรากศัพท์คำภาษาอังกฤษ

คำศัพท์	รากศัพท์
พิจารณา	พิจาร
พิจารณ์	พิจาร
การพิจารณา	พิจาร
พิจาร	พิจาร

ภาพที่ 2-2 ตัวอย่างรากศัพท์คำภาษาไทย

การหารากศัพท์ จึงเป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกันเพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลง และเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่

การหารากศัพท์ของคำภาษาอังกฤษมีขั้นตอนวิธีที่แน่นอนซึ่งสามารถเขียนเป็นอัลกอริทึมในการสกัดคำอุปสรรคและคำปัจจัยได้โดยไม่ต้องเทียบกับรายการคำศัพท์หรือคลังคำ เนื่องมาจากไวยากรณ์ของภาษาอังกฤษมีกฎเกณฑ์ที่แน่นอน ไม่ค่อยมีข้อยกเว้นมากนัก จึงมีความซับซ้อนน้อย ดังนั้นจึงมีผู้สร้างอัลกอริทึมสำหรับการหารากศัพท์ไว้หลายแบบ เช่น อัลกอริทึมปีเตอร์ (Porter Algorithm), อัลกอริทึมแลนแคสเตอร์ (Lancaster Algorithm) เป็นต้น

การหารากศัพท์ของคำภาษาไทยนั้น เท่าที่ศึกษามายังไม่พบว่ามีผู้คิดค้นสร้างอัลกอริทึมสำหรับการหารากศัพท์ไว้ เนื่องมาจากไวยากรณ์ภาษาไทยมีความซับซ้อนมาก และมีข้อยกเว้นหลายประการ การหารากศัพท์ของคำภาษาไทยจึงอาจใช้วิธีการรวบรวมคำศัพท์ที่มีความหมายคล้ายกันหรือมีรากศัพท์เดียวกันไว้เป็นรายการคำศัพท์ หรือจัดเก็บในคลังคำ เพื่อใช้ในการเปรียบเทียบหารากศัพท์ ซึ่งวิธีการนี้ต้องอาศัยมนุษย์เป็นผู้กำหนดไว้ก่อนว่า คำแต่ละคำมีรากศัพท์เป็นคำใด วิธีการนี้ต้องอาศัยผู้เชี่ยวชาญทางภาษาและใช้เวลาในการเก็บรวบรวมและจัดทำรายการคำศัพท์

2.2.4 การทำดัชนี

เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบ ที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสาร เรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ วัตถุประสงค์ของการสร้างดัชนีคือการคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร หรืออาจจะเรียกได้ว่าการหาค่าน้ำหนัก (Term Weighting) การสร้างดัชนี โดยทั่วไปที่นิยมใช้กัน จะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม (Sebastiani, 2002; Marquez. 2000; Aas & Eikvil, 1999)

2.3 การเรียนรู้และจำแนกหมวดหมู่

2.3.1 การเรียนรู้ด้วยคอมพิวเตอร์ เป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligent) ที่เกี่ยวข้องกับการออกแบบและพัฒนาอัลกอริทึมและวิธีการที่จะทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้ การเรียนรู้ด้วยคอมพิวเตอร์เชิงอุปนัย เป็นการค้นหากฎ ลักษณะแบบแผน หรือข้อสรุปต่าง ๆจากการสังเกตกลุ่มข้อมูลขนาดใหญ่ ส่วนการเรียนรู้เชิงอนุมาน เป็นการหาข้อสรุปจากหลักฐานหรือข้อเท็จจริงที่มีอยู่ หลักสำคัญของการวิจัยทางด้านการเรียนรู้ด้วยคอมพิวเตอร์ คือ การสกัดเอาความรู้หรือสารสนเทศจากข้อมูลโดยอัตโนมัติด้วยวิธีการคำนวณหรือวิธีการทางสถิติ ดังนั้นการเรียนรู้ด้วยคอมพิวเตอร์นั้นจึงมีความสัมพันธ์อย่างใกล้ชิดกับการทำเหมืองข้อมูล (Data Mining) และสถิติอัลกอริทึมสำหรับการเรียนรู้ด้วยคอมพิวเตอร์ถูกจัดให้อยู่ในกลุ่มของวิทยาศาสตร์หรือวิธีการที่เกี่ยวข้องกับการแบ่งแยกประเภท (Taxonomy) ซึ่งมีที่มาจากผลลัพธ์ที่ได้จากอัลกอริทึมประเภทของอัลกอริทึมอาจแบ่งได้ ดังต่อไปนี้

1) การเรียนรู้โดยอาศัยตัวอย่าง (Supervised Learning)

เป็นการเรียนรู้โดยใช้อัลกอริทึมเพื่อสร้างฟังก์ชั่นที่จะทำข้อมูลเข้าให้เป็นผลลัพธ์ที่ต้องการ การจำแนกประเภทเป็นรูปแบบหนึ่งของการเรียนรู้โดยอาศัยตัวอย่าง ตัวเรียนรู้จะต้องศึกษาพฤติกรรมของฟังก์ชั่นที่จะทำการกำหนดเวกเตอร์ให้อยู่ $[X_1,X_2,\ldots,X_N]$ ภายใต้ประเภทใดประเภทหนึ่งจากหลายประเภทโดยสังเกตจากตัวอย่างข้อมูลเข้าและตัวอย่างผลลัพธ์ของฟังก์ชั่น

2) การเรียนรู้โดยไม่อาศัยตัวอย่าง (Unsupervised Learning)

เป็นการเรียนรู้โดยการจำลองแบบของชุดข้อมูลเข้าจากลักษณะเฉพาะของข้อมูล โดยไม่ต้องใช้ตัวอย่างผลลัพธ์ที่ถูกจัดประเภทไว้แล้ว

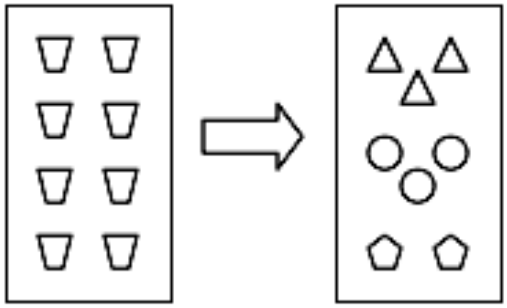
3) การเรียนรู้กึ่งอาศัยตัวอย่าง (Semi-supervised Learning)

เป็นการเรียนรู้ที่อาศัยทั้งตัวอย่างที่ยังไม่ได้จัดประเภทและตัวอย่างที่ถูกจัดประเภทไว้แล้ว เพื่อสร้างฟังก์ชั่นหรือตัวจำแนกประเภทที่เหมาะสมการวิเคราะห์การคำนวณและประสิทธิภาพของอัลกอริทึมที่ใช้สำหรับการเรียนรู้ด้วยคอมพิวเตอร์เป็นแขนงหนึ่งของทฤษฎีทางวิทยาการคอมพิวเตอร์ที่รู้จักกันในชื่อ ทฤษฎีการเรียนรู้เกี่ยวกับการคำนวณ (Computational Learning Theory)

2.3.2 การจำแนกหมวดหมู่

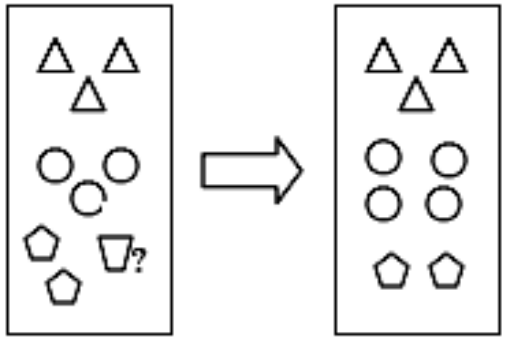
การจำแนกหมวดหมู่สามารถแบ่งได้ 2 ลักษณะ คือ การจัดกลุ่ม (Clustering) และการจำแนกหมวดหมู่ (Classification หรือ Categorization)

การจัดกลุ่มของเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยไม่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน ซึ่งจะเป็นการแบ่งกลุ่มตามลักษณะของเอกสาร โดยเอกสารที่มีลักษณะเหมือนกันจะอยู่ด้วยกัน ดังแสดงในภาพที่ 2-3



ภาพที่ 2-3 การจัดกลุ่ม

การจำแนกหมวดหมู่ของเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยที่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน โดยจะเปรียบเทียบเอกสารกับต้นแบบในแต่ละหมวดหมู่ เอกสารจะถูกจัดอยู่ในหมวดหมู่ที่ต้นแบบมีลักษณะคล้ายกับตัวมันเองมากที่สุด ดังแสดงภาพที่ 2-4



ภาพที่ 2-4 การจำแนกหมวดหมู่

การจำแนกหมวดหมู่เอกสารในภาษาต่างประเทศไม่สามารถนำมาใช้ได้โดยตรงกับการจัดหมวดหมู่ในภาษาไทย เนื่องจากมีปัญหาในการประมวลผลทางด้านภาษาไทยโดยเฉพาะอย่างยิ่งปัญหาในระดับคำ ได้แก่

1. ปัญหาการตัดแบ่งคำ ซึ่งเกิดจากภาษาไทยไม่มีช่องว่างระหว่างคำ ทำให้ไม่สามารถระบุขอบเขตของคำได้อย่างถูกต้อง

2. ปัญหาคำไม่รู้จัก ตัวอย่างเช่น คำศัพท์ด้านเทคนิค คำทับศัพท์ ซึ่งไม่มีการใช้อักษรพิเศษ เพื่อบอกความแต่งแตกระหว่างคำยืมกับคำไทย เช่นเดียวกับภาษาญี่ปุ่นหรือภาษาเกาหลี

3. ไม่มีการใช้เครื่องมือทางภาษา เช่น การใช้อักษรตัวใหญ่ เพื่อบ่งบอกชื่อเฉพาะ เช่นเดียวกับภาษาอังกฤษ

ในการจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติ หากไมมีการแก้ปัญหาในระดับคำจะไม่สามารถกำหนดดัชนีคำหรือวลี เพื่อสร้างกลุ่มต้นแบบในระบบต้นแบบสำหรับจำแนกหมวดหมู่เอกสารอัตโนมัติได้

การจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติประกอบด้วย 2 ขั้นตอน ได้แก่ การฝึกฝนระบบสำหรับจัดกลุ่มเอกสารต้นแบบ และการคัดแยกเอกสารด้วยการคำนวณความคล้ายระหว่างเอกสารอินพุตกับกลุ่มเอกสารต้นแบบ สำหรับการฝึกฝนระบบมี 2 วิธี ได้แก่ การเรียนรู้โดยอาศัยตัวอย่าง และการเรียนรู้โดยไม่อาศัยตัวอย่าง

เทคนิคการเรียนรู้โดยอาศัยตัวอย่างมีหลายวิธี เช่น นิวรอนเน็ตเวิร์ก (Neural Network), ริปเปอร์ (Ripper), ร็อกชิโอ (Rocchio) โดยมีการประยุกต์เทคนิคการประมวลผลภาษาธรรมชาติ และปรับโครงสร้างการแทนกลุ่มเอกสาร เทคนิคการประมวลผลภาษาธรรมชาติใช้สำหรับประมวลผลคำและวลี เพื่อใช้เป็นตัวแทนเอกสารที่ใช้ดัชนีคำหรือวลีเพียงอย่างเดียว ส่วนการปรับโครงสร้างการแทนกลุ่มเอกสาร ได้ใช้โครงสร้างแบบมีลำดับชั้นแทนการใช้ระดับเดียว เพราะการค้นหาเอกสารในโครงสร้างแบบมีลำดับชั้นจะตัดกลุ่มเอกสารที่ไม่เกี่ยวข้องออกไป ทำให้สามารถเพิ่มประสิทธิภาพในการค้นหาเอกสารได้เร็วขึ้น

อัลกอริทึมการจัดหมวดหมู่ (Classifier Algorithm) อัลกอริทึมในการจัดหมวดหมู่แบบการเรียนรู้โดยอาศัยตัวอย่าง สามารถแบ่งขั้นตอนวิธีการจัดหมวดหมู่เอกสารแบ่งได้เป็น 2 ขั้นตอนคือ การเรียนรู้เพื่อสร้างกลุ่มเอกสารต้นแบบและแยกหมวดหมู่ของเอกสารที่สนใจ โดยการตรวจสอบหาความคล้ายกับกลุ่มเอกสารต้นแบบ

2.4 ซัพพอร์ตเวคเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีนหรือเอสวีเอ็ม (Support Vector Machine : SVM) จัดเป็นการเรียนรู้ของเครื่องประเภทแบบการเรียนรู้โดยอาศัยตัวอย่างประเภทหนึ่ง ซึ่งมีความสามารถในการจัดหมวดหมู่ และการทำนาย (Regression) โดยเอสวีเอ็มมีพื้นฐานจะมีการคำนวณแบบเชิงเส้น (Linear) ซึ่งจัดอยู่ในประเภทมุ่งหาผลลัพธ์ที่ดีที่สุดของการเรียนรู้ (Discriminative Training) บนการเรียนรู้จากสถิติของข้อมูล ซึ่งทำงานโดยการหาค่าระยะขอบที่มากที่สุด (Maximum Margin) ของระนาบตัดสินใจ (Decision Hyperplane) ในการแบ่งแยกกลุ่มข้อมูลที่ใช้ฝึกฝนออกจากกัน

วิธีการนี้มีชื่อเรียกอีกชื่อหนึ่งว่า จัดหมวดหมู่โดยค่าระยะขอบที่มากที่สุด (Maximum Margin Classifier)

2.4.1 หลักการทำงานของเอสวีเอ็ม ข้อมูลจะถูกเขียนในรูปสมาชิกคู่อันดับดังนี้

$$\{(x_1,c_1),(x_2,c_2),(x_3,c_3),\ldots,(x_n,c_n)\} \quad (2\text{-}1)$$

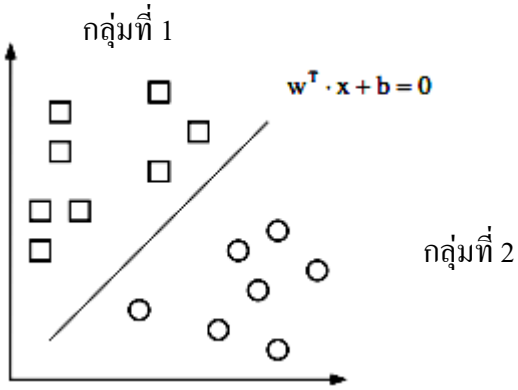
เมื่อ c_i มีค่าเป็น 1 หรือ -1 ซึ่งกำหนดให้เป็นข้อมูลแบ่งกลุ่มของข้อมูล x_i ที่มีค่า c_i เป็น 1 และ x_i ที่มีค่า c_i เป็น -1 โดยที่แต่ละ x_i เป็นค่าข้อมูลมิติของเวกเตอร์จริง เมื่อกำหนดให้ข้อมูลนี้เป็นข้อมูล สำหรับฝึกฝน ซึ่งหมายความว่าการแบ่งกลุ่มของข้อมูลมีความถูกต้อง ดังนั้นเส้นแบ่งกลุ่มข้อมูลที่ ถูกสร้างขึ้นซึ่งเป็นสมการเส้นตรงทั่วไปคือ $y = mx+b$ โดยในที่นี้จะแทน m ด้วย w^T เพื่อกำหนด เป็นสมการ (2-2)

$$w^T.x+b = 0 \quad (2\text{-}2)$$

เมื่อ w^T คือ เวกเตอร์ตั้งฉากของค่าความชัน m ของเส้นแบ่ง

b คือ ค่าคงที่ที่ได้จากค่าของแกน y ของแต่ละข้อมูล x

การแบ่งข้อมูลดังภาพที่ 2-5



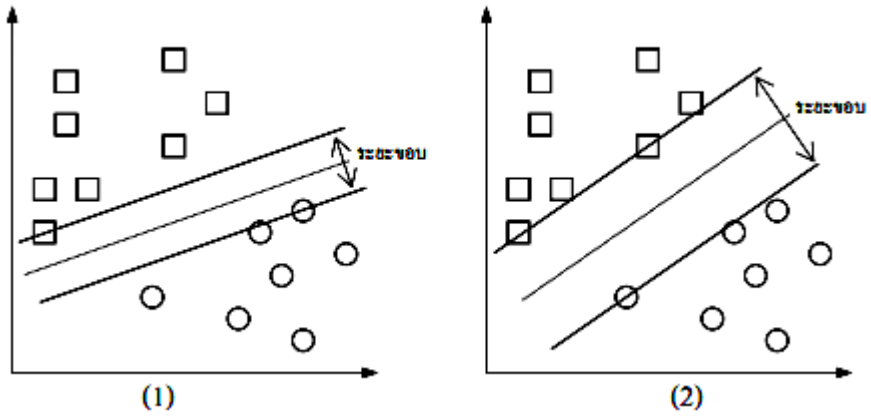
ภาพที่ 2-5 ข้อมูลสองกลุ่มที่ถูกแบ่งด้วยเส้นตรง

ในทางอุดมคติเส้นแบ่งกลุ่มที่ดีที่สุดคือเส้นแบ่งกลุ่มที่ทำให้มีระยะขอบ (Margin) มากที่สุดของเส้นคู่ขนานที่ขยายออกจากเส้นแบ่งไปสัมผัสจุดข้อมูลของทั้งสองกลุ่มที่ใกล้ที่สุด หรือกล่าวได้

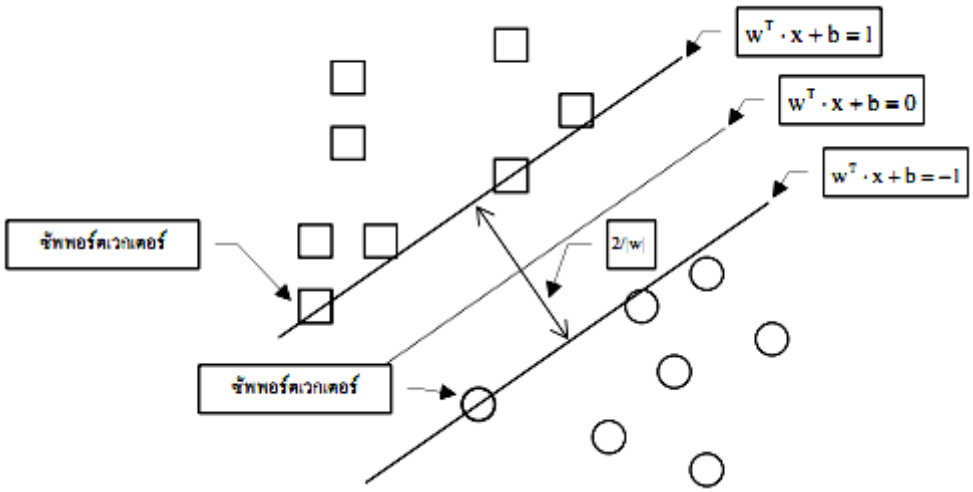
ว่าเป็นเส้นแบ่งที่มีระยะขอบกว้างที่สุด จากภาพที่ 2-6 และภาพที่ 2-7 เรียกจุดข้อมูลอย่างน้อยหนึ่งจุดจากทั้งสองกลุ่มที่สัมผัสกับเส้นขนานที่ขยายออกได้มากที่สุดว่า ซัพพอร์ตเวกเตอร์ โดยค่าของเส้นขอบทั้งสองที่ใช้แบ่งกลุ่มข้อมูลเป็นสองกลุ่มคำนวณได้จากสมการ (2-3) และสมการ (2-4)

$$w^T.x+b = 1 \quad (2\text{-}3)$$

$$w^T.x+b = -1 \quad (2\text{-}4)$$



ภาพที่ 2-6 การปรับความชันของเส้นแบ่งแล้วทำให้ได้ระยะขอบที่มากที่สุด



ภาพที่ 2-7 ระยะขอบที่กว้างที่สุดเมื่อสัมผัสกับซัพพอร์ตเวกเตอร์

เมื่อข้อมูลที่ใช้ฝึกฝนเป็นข้อมูลที่สามารถแบ่งกลุ่มได้ด้วยเส้นตรงใด ๆ ระยะที่ที่กว้างที่สุดของไฮเปอร์แพลน (Hyper plane) ทั้งสองที่ขยายออกไปจนกว่าจะพบจุดข้อมูลของทั้งสองกลุ่มคือ $2/|w|$ โดย ค่า $|w|$ มีค่าน้อยที่สุด ค่าข้อมูลแต่ละ x_i จุดจำแนกอยู่ในกลุ่มใดสามารถพิจารณาได้จากเงื่อนไขดังนี้ ถ้า $w^T.x_i+b \geq 1$ แสดงว่า x_i เป็นกลุ่มที่ 1 และถ้า $w^T.x_i+b \leq -1$ แสดงว่า x_i เป็นกลุ่มที่ 2

ในการคัดแยกจุดข้อมูลใด ๆ ที่นำมาฝึกฝนเพื่อกรองว่าเป็นกลุ่ม 1 สามารถตรวจสอบ ได้จากสมการที่ (2-5)

$$c_i(w \cdot x_i-b) \geq 1 \quad \text{เมื่อ} \quad 1 \leq i \leq n \tag{2-5}$$

ดังนั้นสามารถเขียนเป็นรูปสั้นเพื่อหาระยะขอบที่น้อยที่สุด โดยที่สามารถคำนวณการแบ่งกลุ่มได้ดังสมการที่ (2-6)

$$\text{Minimize}_{w,b} |w| \text{ โดยตรวจสอบ } c_i(w \cdot x_i-b) \geq 1 \quad \text{เมื่อ} \quad 1 \leq i \leq n \tag{2-6}$$

2.4.2 ปัญหาในการเรียนรู้ของเอสวีเอ็ม (Vapnik, 1995) ได้นำเสนอวิธีการปรับปรุงการเรียนรู้แบบ Maximum Margin โดยมีวิธีการดังนี้

2.4.2.1 การแก้ไขปัญหาความผิดพลาดในการจัดกลุ่ม (Misclassification) โดยให้สามารถผ่อนปรนหรืออนุโลมให้มีการละจุดข้อมูลที่ไม่สามารถสร้างเส้นระยะขอบที่มากที่สุดที่ดีที่สุดเพื่อการแบ่งกลุ่ม เรียกว่า ซอฟต์มาร์จิน (Soft Margin) ดังตัวอย่างในภาพที่ 2-8 เรียกจุดข้อมูลนี้ว่า เวกเตอร์อนุโลม โดยการกำหนดให้มีตัวแปรค่าความหยอน (Slack Variable) ξ_i ซึ่งใช้กำหนดค่าการอนุโลมของการจัดกลุ่มไม่ได้จากกลุ่มข้อมูล x_i และตัวแปร C โดยปรับปรุงสูตรการคำนวณใหม่ดังสมการที่ (2-7)

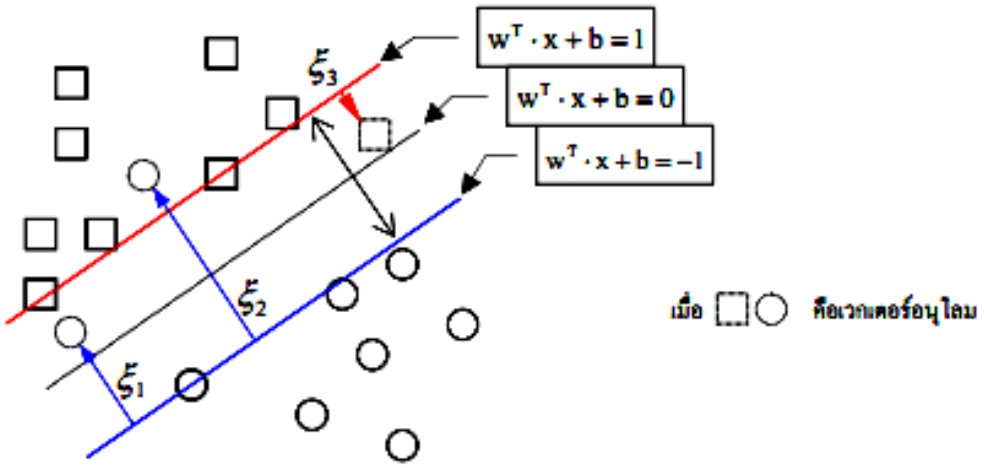
$$\text{minimize}_{w,b} |w| + C\sum_{i=1}^{n} \xi_i \tag{2-7}$$

โดยตรวจสอบ $c_i (w \cdot x_i -b) \geq 1 - \xi_i$

$\xi_i \geq 0$

$C > 0$

เมื่อ $1 \leq i \leq n$

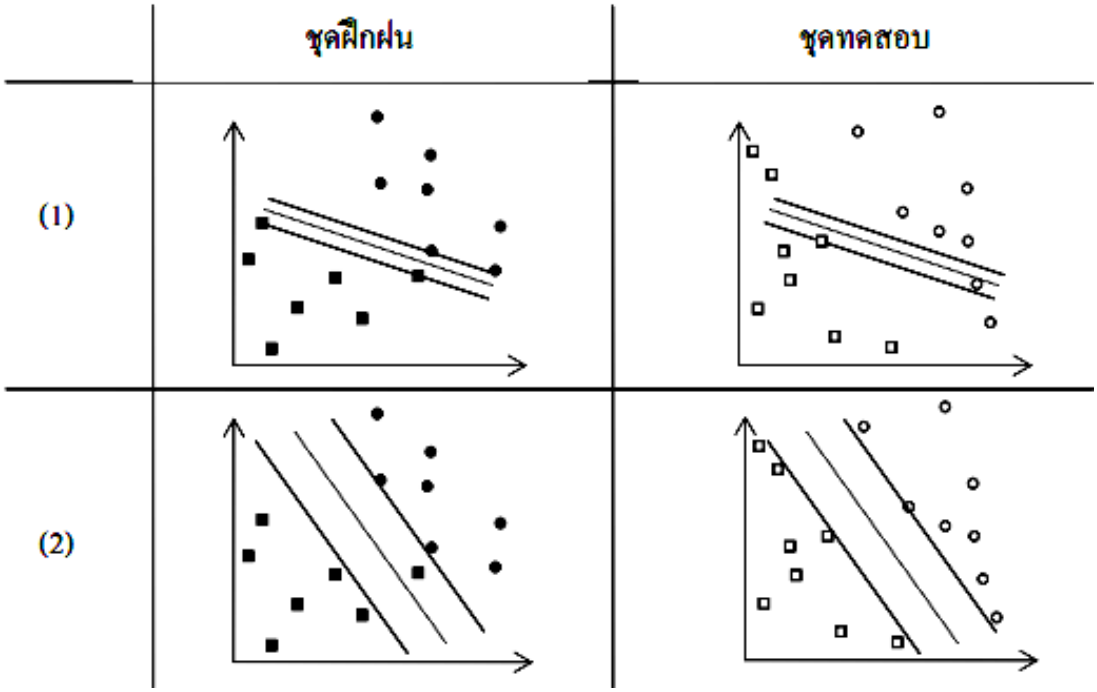


ภาพที่ 2-8 การเพิ่มตัวแปรค่าความหย่อนสำหรับเวกเตอร์อนุโลม

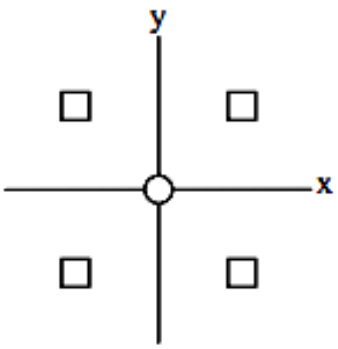
2.4.2.2 การแก้ไขปัญหาเกินพอดี (Over fitting) ซึ่งเกิดจากการปรับให้เอสวีเอ็มเรียนรู้จนได้ค่าที่ดีที่สุด แต่เมื่อนำไปทดสอบกับข้อมูลจริงปรากฏว่ามีความผิดพลาดสูง ดังภาพที่ 2-9 (1) เป็นการเรียนรู้ให้ได้เส้นแบ่งในการแบ่งกลุ่มที่ดีที่สุด แต่เมื่อนำไปทดสอบกับข้อมูลชุดทดสอบปรากฏว่าผิดพลาดมาก เมื่อเปรียบเทียบกับภาพที่ 2-9 (2) ที่มีการผ่อนปรนใหม่ค่าผิดพลาดในการเรียนรู้บ้าง เมื่อนำไปทดสอบกับข้อมูลชุดทดสอบปรากฏว่ามีความผิดพลาดน้อยลงโดยที่ $C\sum^{n}_{i=1}\xi_i$ ทำหน้าที่ในการควบคุมจำนวนจุดข้อมูลที่จัดกลุ่มไม่ได้ ซึ่งค่าของ C จะถูกกำหนดโดยผู้ใช้ ทั้งนี้ C ที่มีค่ามากหมายถึงการยอมให้มีความผิดพลาดได้มาก เมื่อนำไปประยุกต์ใช้ในการคัดแยกข้อมูลจริงมักจะพบกับข้อมูลที่ไม่สมดลซึ่งอาจเกิดจากจำนวนตัวอย่างในแต่ละกลุ่มมีจำนวนไม่เท่ากัน หรือความหนาแน่นของแต่ละคุณลักษณะ (Feature) ที่แตกต่างกันมากในแต่ละกลุ่ม โดยปกติแล้วควรมีการปรับปรุงข้อมูลตัวอย่างให้สมดุลก่อนนำเข้าสู่กระบวนการเรียนรู้จึงทำให้ระบบ สามารถเรียนรู้ได้ดีที่สุด

2.4.3 เอสวีเอ็มกับข้อมูลแบบไม่เชิงเส้น พื้นฐานของเอสวีเอ็มเริ่มใช้ในการคัดแยกกลุ่มข้อมูลในลักษณะเชิงเส้น (Linear Classifier) แต่ในบางครั้งข้อมูลจริงอาจจะมีคุณลักษณะที่ไม่เป็นเชิงเส้นดังตัวอย่างในภาพที่ 2-10 เพื่อแก้ปัญหาดังกล่าวจึงมีวิธีจำลองข้อมูลโดยเพิ่มมิติของข้อมูล (Higher Dimensional) มากขึ้น เช่น ข้อมูลเดิมมีขนาด 2 มิติ คือ แกน x และแกน y จะจำลองมิติที่ 3 ขึ้นมาดังภาพที่ 2-11 ซึ่งเป็นการจำลองมิติที่ 3 ด้วยการสร้างโครงผิวจากสมการ $f(x,y)=x^2$ และภาพที่ 2-12 เป็นการจำลองมิติที่ 3 ด้วยการสร้างโครงผิวจากสมการ $f(x,y)=x^2$ และภาพที่ 2-12 เป็นการจำลองมิติที่ 3 ด้วยการสร้างโครงผิวจากสมการ $f(x,y)=x^2+y^2$ เพื่อเป็นการยกข้อมูลให้ไปอยู่บนโครงผิวที่สร้างขึ้น หลังจากนั้นจึงคำนวณหาเส้นระนาบใดๆ ที่สามารถใช้แบ่งข้อมูลจากโครงผิวเป็นสองกลุ่ม

หลังจากนั้นจึงทำการลดมิติกลับไปเป็น 2 มิติ สิ่งที่ได้คือขอบเขตของการแบ่งข้อมูลออกเป็นสองกลุ่ม หลังจากนั้นจึงทำการลดมิติกลับไปเป็น 2 มิติ สิ่งที่ได้คือขอบเขตของการแบ่งข้อมูลออกเป็นสองกลุ่ม ฟังก์ชันที่ใช้ในการคำนวณเพื่อเพิ่มมิติของข้อมูลเรียกว่า ฟังก์ชันแกน (Kernel Function)

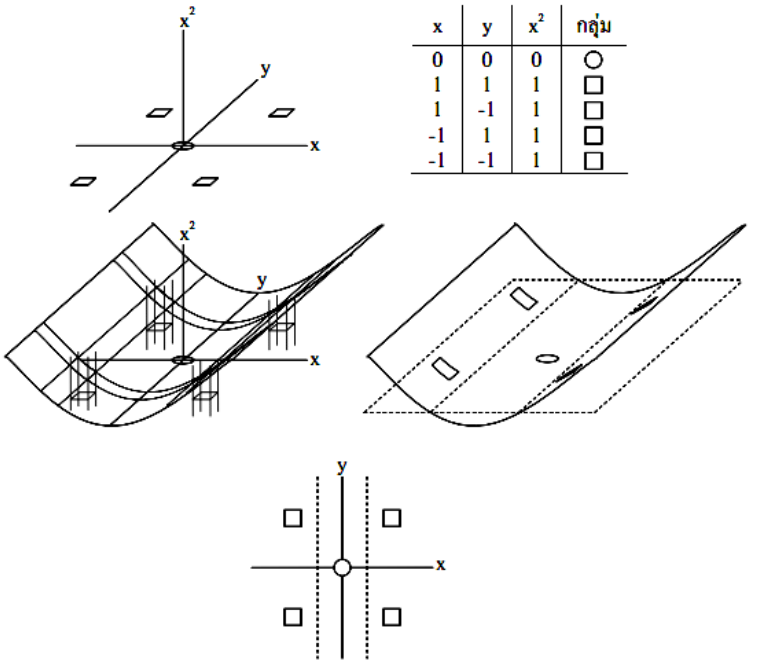


ภาพที่ 2-9 การเปรียบเทียบตัวอย่างการเรียนรู้ของเอสวีเอ็มแบบเกินพอดี



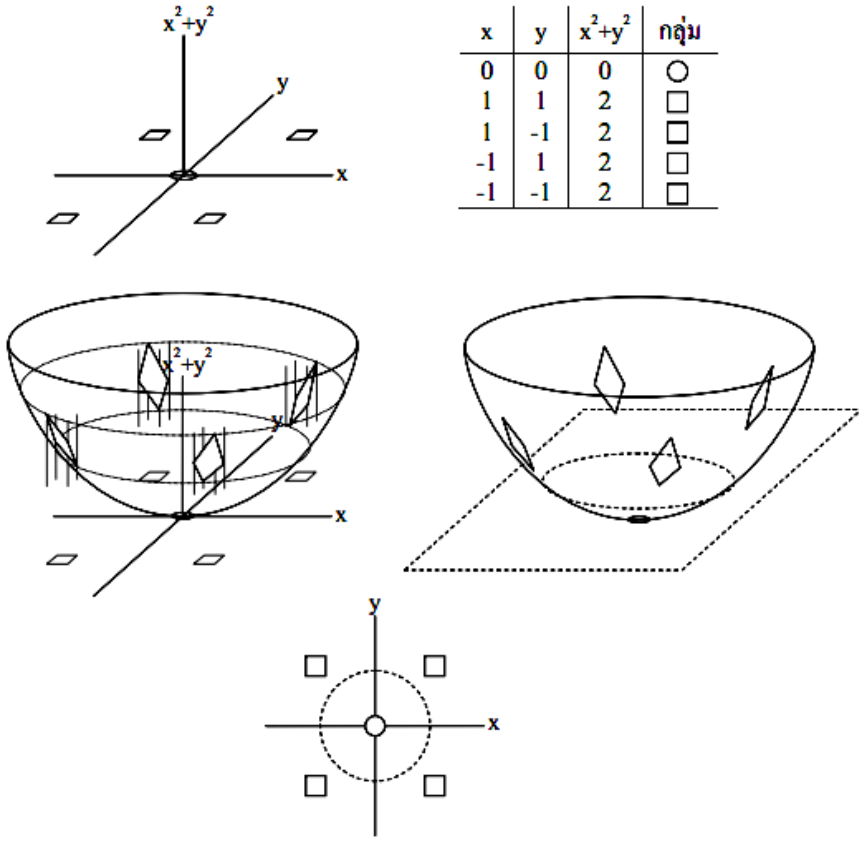
x	y	กลุ่ม
0	0	○
1	1	□
1	-1	□
-1	1	□
-1	-1	□

ภาพที่ 2-10 ตัวอย่างข้อมูลสองมิติที่ไม่สามารถใช้การคัดแยกแบบเชิงเส้นได้



x	y	x^2	กลุ่ม
0	0	0	○
1	1	1	□
1	-1	1	□
-1	1	1	□
-1	-1	1	□

ภาพที่ 2-11 การเพิ่มมิติของข้อมูลด้วย $f(x,y)=x^2$ เพื่อให้สามารถแบ่งกลุ่มข้อมูล



x	y	x^2+y^2	กลุ่ม
0	0	0	○
1	1	2	□
1	-1	2	□
-1	1	2	□
-1	-1	2	□

ภาพที่ 2-12 การเพิ่มมิติข้อมูลด้วย $f(x,y)=x^2+y^2$ เพื่อสามารถแบ่งกลุ่มข้อมูลด้วยระนาบเชิงเส้น

เมื่อข้อมูลแต่ละ x_i ถูกคำนวณใหม่มิติสูงขึ้นด้วยฟังก์ชันแกน Φ จากสมการที่ (2-6) สามารถเขียนใหม่เป็นสมการที่ (2-8)

$$\text{Minimize}_{w,b} |w| \text{ โดยตรวจสอบ } c_i(w \cdot \Phi x - b) \geq 1 \quad \text{เมื่อ } 1 \leq i \leq n \tag{2-8}$$

กำหนดผลของฟังก์ชันแกน $k(x,x')$ ให้เป็นบวกเพื่อใช้ยกข้อมูลให้อยู่ในมิติที่สูงกว่าซึ่งคำนวณได้จากอินเนอร์โปรดักท์ของพื้นที่คุณลักษณะ (Feature Space) ดังสมการ (2-9)

$$k(x, x') = \Phi x \cdot \Phi x' \tag{2-9}$$

สมการที่ (2-9) สามารถเขียนใหม่ในรูปของการรวมฟังก์ชันแกน ได้ดังสมการที่ (2-10)

$$\text{Minimize}_{w,b} |w| \quad \text{โดยตรวจสอบ } c_i(w \cdot k(x, x') - b) \quad \geq 1 \quad \text{เมื่อ } 1 \leq i \leq n \tag{2-10}$$

โดยที่ฟังก์ชันแกนพื้นฐาน ที่ใช้ในการเพิ่มมิติข้อมูลสำหรับประมวลผลข้อมูลแสดงไว้ในตารางที่ 2-1

ตารางที่ 2-1 ฟังก์ชันแกนพื้นฐานสำหรับเอสวีเอ็ม

ชนิดของฟังก์ชัน	รูปสมการ
Linear	$k(x,x') = (x^T x')$
Polynomial	$k(x,x') = (\gamma x^T x' + r)^d, \gamma > 0$
Radius Basic Function	$k(x,x') = \exp(-\gamma \|\| x - x' \|\|^2), \gamma > 0$
Sigmoid	$k(x,x') = \tanh(\gamma x^T x' + r)$
โดยที่ γ , r และ d เป็นพารามิเตอร์ของฟังก์ชันแกน	

ดังนั้นสมการพื้นฐานของเอสวีเอ็มเมื่อรวมฟังก์ชันการคำนวณเพิ่มมิติและการกำหนดเวกเตอร์อนุโลมเข้าด้วยกันสามารถเขียนใหม่ได้ดังสมการ (2-11)

$$\text{minimize}_{w,b} |w| + C\sum_{i=1}^{n} \xi_i \tag{2-11}$$

โดยตรวจสอบ $c_i (w \cdot k(x,x')-b) \geq 1$

$\xi_i \geq 0$

$C > 0$

เมื่อ $1 \leq i \leq n$

2.4.4 การใช้เอสวีเอ็มคัดแยกข้อมูลมากกว่าสองกลุ่ม (Multiclass Classification) เอสวีเอ็มมีพื้นฐานการคัดแยกแบบสองกลุ่ม (Binary Classification) ด้วยการสร้างระนาบตัดสินใจ สำหรับการประยุกต์ใช้เอสวีเอ็มในการคัดแยกกลุ่มข้อมูลมากกว่าสองกลุ่มทำได้โดยการเปรียบเทียบคัดแยกทีละคู่ ๆ หลังจากนั้นจึงหาข้อสรุปเพื่อคัดแยกว่าแต่ละข้อมูลอยู่กลุ่มใด วิธีที่นิยมประยุกต์ใช้มี 2 วิธี (Burges, 1998) คือ การคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมด (One-against-The Rest) และ การคัดแยกทีละหนึ่งต่อหนึ่ง (One-against-One) โดยรายละเอียดแต่ละวิธีมีดังนี้

2.4.4.1 วิธีการคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมด หรือบางครั้งเรียกว่า One-against-All เป็นการเลือกกลุ่มใด ๆ มาเปรียบเทียบกับกลุ่มอื่น ๆ ทีละคู่จนครบทุกกลุ่ม เมื่อมีจำนวนกลุ่มเท่ากับ k กลุ่ม วิธีการนี้จะทำการเรียนรู้การคัดแยกแบบสองกลุ่มจำนวน k รอบ โดยการเปรียบเทียบวนกลุ่มที่เลือกกับกลุ่มที่เหลือ แสดงตัวอย่างดังตารางที่ 2-2

ตารางที่ 2-2 การเปรียบเทียบแบ่งกลุ่มแบบการคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมด

กลุ่มที่เลือก		กลุ่มที่เหลือ
1	เปรียบเทียบกับ	2 จนถึงกลุ่มที่ k
2	เปรียบเทียบกับ	1, 3 จนถึงกลุ่มที่ k
3	เปรียบเทียบกับ	1, 2, 4 จนถึงกลุ่มที่ k
⋮	⋮	⋮
k	เปรียบเทียบกับ	1 จนถึงกลุ่มที่ (k-1)

การเปรียบเทียบวนทีละกลุ่มแบบนี้ทำให้ได้ฟังก์ชันตัดสินใจในการแบ่งกลุ่มจำนวน k ฟังก์ชัน คือ

$$(W^1)^T x + b_1$$
$$(W^2)^T x + b_2$$
$$\vdots$$
$$(W^k)^T x + b_k$$

ในการตัดสินใจแบ่งข้อมูลออกเป็นกลุ่มใด ใช้วิธีการคัดกรองกลุ่มที่เหลือออก ยกตัวอย่างเช่นในการตัดสินใจแบ่งข้อมูลเป็นกลุ่มที่ 1 จากจำนวน k กลุ่ม มีการคัดกรองกลุ่มที่เหลือออกคือกลุ่มที่ 2 จนถึงกลุ่มที่ k ด้วยฟังก์ชันตัดสินใจดังนี้

$$(W^{1})^{T} x + b_{1} \geq +1$$
$$(W^{2})^{T} x + b_{2} \leq -1$$
$$\vdots$$
$$(W^{k})^{T} x + b_{k} \leq -1$$

ผลที่ได้คือการคัดแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ตามลักษณะฟังก์ชันตัดสินใจที่สร้างขึ้นในแต่ละรอบ

2.4.4.2 วิธีการคัดแยกทีละหนึ่งต่อหนึ่ง เป็นการเปรียบเทียบกลุ่มข้อมูลจำนวน k กลุ่ม ด้วยการ เปรียบเทียบเพื่อคัดแยกแบบสองกลุ่มทีละคู่แบบไม่ซ้ำกันจนครบทุกกลุ่ม ตัวอย่างเช่น

(1,2),(1,3),(1,4),…,(1,k)

(2,3),(2,4),…,(2,k)

(3,4),…,(2,k)

(k-1,k)

โดยจำนวนครั้งในการเปรียบเทียบของการเรียนรู้เท่ากับ k(k-1)/2 ครั้ง ดังนั้น ฟังก์ชันตัดสินใจมีจำนวน k(k-1)/2 ฟังก์ชัน เช่นกัน ตัวอย่างเช่น หากมีจำนวนกลุ่ม 4 กลุ่ม จำนวนฟังก์ชันตัดสินใจเท่ากับ 4(4-1)/2 = 6 ฟังก์ชัน แสดงตัวอย่างดังตารางที่ 2-3

ตารางที่ 2-3 ตัวอย่างฟังก์ชันตัดสินใจของวิธีการคัดแยกทีละหนึ่งต่อหนึ่ง

$y_i = 1$	$y_i = -1$	ฟังก์ชันตัดสินใจ
กลุ่ม 1	กลุ่ม 2	$f^{12}(x) = (w^{12})^{T} x + b^{12}$
กลุ่ม 1	กลุ่ม 3	$f^{13}(x) = (w^{13})^{T} x + b^{13}$
กลุ่ม 1	กลุ่ม 4	$F^{14}(x) = (w^{14})^{T} x + b^{14}$
กลุ่ม 2	กลุ่ม 3	$F^{23}(x) = (w^{23})^{T} x + b^{23}$
กลุ่ม 2	กลุ่ม 4	$F^{24}(x) = (w^{24})^{T} x + b^{24}$
กลุ่ม 3	กลุ่ม 4	$F^{34}(x) = (w^{34})^{T} x + b^{34}$

ในการตัดสินว่าข้อมูลที่คัดแยกอยู่กลุ่มใด ใช้วิธีพิจารณาผลการโหวตสูงสุดจากการเปรียบเทียบทีละคู่ โดยแต่ละกลุ่มมีโอกาสเปรียบเทียบ จำนวน k-1 ครั้งเท่า ๆ กัน ยกตัวอย่างเช่น ตารางที่ 2-4

ตารางที่ 2-4 ตัวอย่างผลการเปรียบเทียบทีละคู่แบบไม่ซ้ำกัน

คู่เปรียบเทียบ		ผลการเปรียบเทียบ
กลุ่ม 1	กลุ่ม 2	กลุ่ม 1
กลุ่ม 1	กลุ่ม 3	กลุ่ม 1
กลุ่ม 1	กลุ่ม 4	กลุ่ม 1
กลุ่ม 2	กลุ่ม 3	กลุ่ม 2
กลุ่ม 2	กลุ่ม 4	กลุ่ม 4
กลุ่ม 3	กลุ่ม 4	กลุ่ม 3

รวมผลเปรียบเทียบแล้วนำมาพิจารณาตัดสินใจเลือกกลุ่ม ตัวอย่างดังตารางที่ 2-5 ดังนั้นการแบ่งกลุ่ม ครั้งนี้จึงถูกจัดให้เป็นกลุ่มที่ 1 ด้วยผลรวมการเปรียบเทียบทีละคู่ สูงสุดคือ 3 ครั้ง

ตารางที่ 2-5 ผลการตัดสินใจเลือกกลุ่มของการคัดแยก

	กลุ่ม 1	กลุ่ม 2	กลุ่ม 3	กลุ่ม 4
ผลรวมการเปรียบเทียบ	3	1	1	1

จากการเปรียบเทียบการทำงานของทั้งสองวิธีในการคัดแยกข้อมูลมากกว่าสองกลุ่มโดยใช้เอสวีเอ็ม พบว่าผลที่ได้มีความถูกต้องใกล้เคียงกัน แต่วิธีการคัดแยกทีละหนึ่งต่อหนึ่งใช้เวลาในการประมวลผลน้อยกว่า เนื่องจากเมื่อพิจารณามิติของข้อมูลจำนวน n ข้อมูล ที่ k กลุ่ม วิธีการคัดแยกทีละหนึ่งต่อหนึ่งใช้การเปรียบเทียบคัดแยกจำนวน k(k-1)/2 ครั้งบนข้อมูลจำนวน 2n/k ข้อมูล เมื่อพิจารณาค่าความซับซ้อนตามขนาดข้อมูลเมื่อประมวลผลด้วยตัวอย่างฟังก์ชันแกนแบบโพลิโนเมียล (Polynomial) ที่มีดีกรีเท่ากับ d คือ $\frac{k(k-1)}{2}O\left(\left(\frac{2n}{k}\right)^d\right)$

ในขณะที่วิธีการคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมดใช้การเปรียบเทียบคัดแยกจำนวน k ครั้งบนจำนวนข้อมูล n ข้อมูล ซึ่งมีค่าความซับซ้อนตามขนาดข้อมูลเมื่อประมวลผลด้วยฟังก์ชันแกนแบบโพลิโนเมียลที่มีดีกรีเท่ากับ d คือ $kO(n^d)$

ดังนั้นจะเห็นได้ว่าเมื่อเปรียบเทียบวิธีการประมวลผลด้วยฟังก์ชันแกนเดียวกันบนมิติของจำนวนข้อมูลที่เท่ากันและจำนวนกลุ่มเท่ากัน วิธีการคัดแยกทีละหนึ่งต่อหนึ่งมีค่าความซับซ้อนตามขนาด ข้อมูลน้อยกว่า เพราะฉะนั้นวิธีการคัดแยกทีละหนึ่งต่อหนึ่งจะใช้เวลาในการประมวลผลน้อยกว่าวิธีการคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมดดังนั้นในงานวิจัยนี้จึงเลือกเอสวีเอ็มที่มีความสามารถคัดแยกข้อมูลมากกว่าสองกลุ่ม โดยใช้วิธีการเปรียบเทียบแบบวิธีการคัดแยกทีละหนึ่งต่อหนึ่ง ซึ่งมีการประมวลผลที่เร็วกว่าในขณะที่ ประสิทธิภาพของทั้งสองวิธีไม่แตกต่างกัน

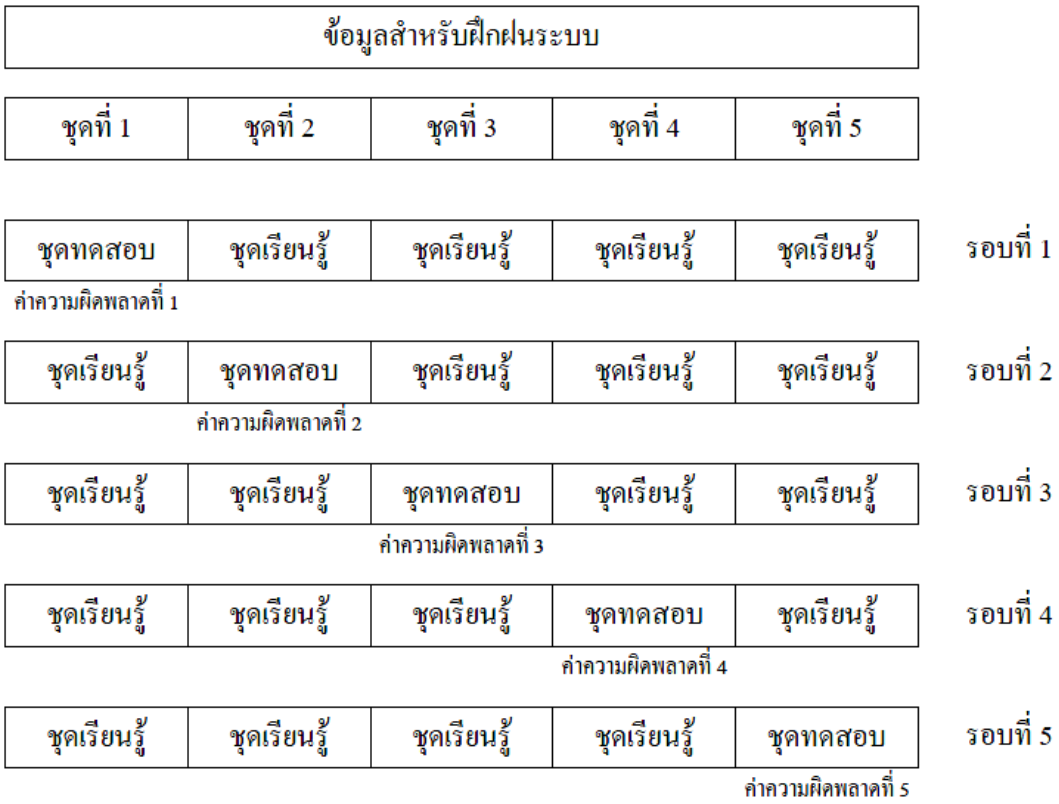
2.4.5 การตรวจสอบความถูกต้องของการเรียนรู้ สามารถสังเกตได้จากค่าความแม่นยำหรือค่าความผิดพลาดที่ได้จากการทำครอสวาลิเดชั่น (Cross Validation) ระหว่างการแบ่งชุดทดสอบในการเรียนรู้ ในที่นี้ผู้วิจัยเลือกพิจารณาความแม่นยำ ซึ่งเป็นสัดส่วนร้อยละของการคัดแยกได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด โดยใช้การครอสวาลิเดชั่นแบบเคโฟล์ด (k-fold) ซึ่งพื้นฐานของวิธีการครอสวาลิเดชั่นนี้คือการสุ่มตัวอย่าง (Sampling) ข้อมูลไปเป็นชุดฝึกฝนและชุดทดสอบแบบสลับกัน โดยเริ่ม จากแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาเป็นชุดทดสอบเพื่อจำลอง ผลลัพธ์จากการ ทำครอสวาลิเดชั่น การแบ่งชุดทดสอบแบบนี้มักถูกใช้เป็นตัวเลือกในการกำหนด โมเดล โดยสังเกตจากผลของการทดสอบในแต่ละครั้งของการปรับค่าพารามิเตอร์แต่ละโมเดล การทำครอสวาลิเดชั่นมีรูปแบบการทดลองเพื่อประเมินหลายวิธีดังนี้

2.4.5.1 วิธีแฮนด์เอาท์วาลิเดชั่น (Handout Validation) เป็นการแบ่งข้อมูลที่จะใช้ตรวจสอบโมเดลโดยการ เลือกสุ่มข้อมูลบางส่วนนำมาเป็นชุดทดสอบจากข้อมูลทั้งหมดที่ใช้เรียนรู้ด้วยการพิจารณาเลือกด้วยมือ โดยในการสุ่มจะต้องมีความสมดุลของการเป็นตัวอย่างที่ดี ส่วนข้อมูลที่เหลือจะใช้เป็นข้อมูล สำหรับฝึกฝนระบบ เมื่อฝึกฝนระบบจนได้โมเดลแล้วจึงนำโมเดลมาทดสอบหาความถูกต้องหรือ ค่าความผิดพลาดด้วยการทดสอบกับชุดข้อมูลทดสอบที่ได้สุ่มเอาไว ซึ่งวิธีการนี้เหมาะสำหรับ ข้อมูลที่นำมาทดสอบมีจำนวนไม่มาก

2.4.5.2 วิธีเคโฟล์ดครอสวาลิเดชั่น เป็นการแบ่งข้อมูลออกเป็น k ชุด เท่า ๆ กัน และทำการคำนวณค่าความผิดพลาด k รอบ โดยแต่ละรอบการคำนวณ ข้อมูลชุดหนึ่งจากข้อมูล k ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก k-1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ แล้วนำผลความถูกต้องหรือค่าความผิดพลาดของแต่ละรอบมารวมกันและหาค่าเฉลยเพื่อเป็นค่าสะท้อนประสิทธิภาพของการฝึกฝน ดังตัวอย่างในภาพที่ 2-13 เป็นการกำหนดค่า k=5 หมายถึงการนำข้อมูล ทั้งหมดมาแบ่งเป็น 5 ชุด และจะดำเนินการเรียนรู้และทดสอบจำนวน 5 รอบ โดยในแต่ละรอบจะเลือกข้อมูลมา 1 ชุดไม่ซ้ำกันมาเป็นชุดทดสอบ ด้วยการเรียนรู้ของชุดเรียนรู้ที่เหลือ เมื่อทำครบ 5 รอบ หาผลรวมของค่าความผิดพลาดหรือความแม่นยำจากชุดทดสอบในแต่ละรอบแล้วหารด้วยจำนวนรอบ ผลที่ได้เป็นค่าเฉลยของความผิดพลาดในการเรียนรู้ของโมเดล ข้อดีของวิธีการนี้

คือ ข้อมูลในแต่ละชุดที่ทำการแบ่งจะถูกทดสอบอย่างน้อย 1 ครั้ง และถูกเรียนรู้ทั้งหมด k-1 ครั้ง โดยในขั้นตอนเหล่านี้สามารถกำหนดได้ว่าต้องการขนาดข้อมูลขนาดใด และต้องการทำการคำนวณเป็นจำนวนรอบเท่าใด ซึ่งเหมาะสำหรับการประมวลผลทดสอบกับข้อมูลที่มีมิติขนาดจำนวนมาก

2.4.5.3 วิธีลีฟวันเอาท์ครอสวาลิเดชั่น (Leave-one-out cross-validation) เป็นการทำครอสวาลิเดชั่นแบบเคโฟล์ดแบบหนึ่ง เมื่อกำหนดให้ k มีค่าเท่ากับจำนวนข้อมูลทั้งหมด ดังนั้นในกรณีที่มีข้อมูล 10 ชุด จะต้องทำการวัดผลความผิดพลาดโดยวิธีการ 10-fold เป็นต้น ข้อดีของวิธีนี้คือ ความเที่ยงตรงสูงใกล้เคียงกับการฝึกฝนในจริง เนื่องจากมีความแตกต่างของข้อมูลที่ใช้ฝึกฝนจริงเพียง 1 ชุดเท่านั้น แต่วิธีนี้ต้องใช้เวลาในการประมวลผลมากกว่าเนื่องจากต้องทำการประมวลผลเป็นจำนวน k รอบ ดังนั้น เมื่อนำไปใช้กับข้อมูลจำนวนมาก ยอมต้องใช้เวลาในการประมวลผลมากขึ้นด้วย โดยมีอัตราการเติบโตของเวลาการประมวลผลเป็น $O = n^2$ ดังนั้นวิธีการนี้จึงเหมาะกับข้อมูลตัวอย่างที่มีน้อย



ภาพที่ 2-13 ตัวอย่างการแบ่งชุดข้อมูลของเคโฟล์ดครอสวาลิเดชั่นเมื่อ k = 5

2.4.6 การคัดเลือกโมเดลจากการเรียนรู้ พิจารณาจากค่าความแม่นยำของแต่ละโมเดลเมื่อนำไปเรียนรู้กับกลุ่มตัวอย่างที่ใช้ฝึกฝน ทั้งนี้ความความแม่นยำขึ้นอยู่กับการปรับพารามิเตอร์ของฟังก์ชันแกนที่เลือกใช้ โดยทั่วไปของการปรับพารามิเตอร์ระหว่างการฝึกฝนมี 2 วิธี คือวิธีการค้นหาแบบตาราง (Grid Search) และวิธีการค้นหาแบบมีรูปแบบ (Pattern Search) โดยรายละเอียดแต่ละวิธีมีดังนี้

2.4.6.1 การค้นหาแบบตารางเป็นการทดสอบค่าพารามิเตอร์ทั้งหมดในช่วงที่กำหนดทีละค่าเพื่อหาค่าพารามิเตอร์ที่ทำให้โมเดลมีค่าความแม่นยำสูงที่สุด ตัวอย่างเช่น ในกรณีที่มีพารามิเตอร์ที่ต้องปรับ 2 พารามิเตอร์ และกำหนดช่วงให้แต่ละพารามิเตอร์มีค่าช่วงที่จะต้องทดสอบตั้งแต่ 0 ถึง 10 ซึ่งมีค่าทดสอบ 11 ค่า ดังนั้นการค้นหาแบบตารางจะมีจำนวนรอบการทดสอบเท่ากับ 11*11 หรือ 121 รอบการทดสอบ เป็นต้น

2.4.6.2 วิธีการค้นหาแบบมีรูปแบบมีหลักการทำงานคือ เริ่มต้นประมวลผลด้วยค่ากึ่งกลางของช่วงค่าพารามิเตอร์ที่กำหนดและทำการกำหนดขั้นของการสุ่มค่าขึ้นและลงของแต่ละค่าพารามิเตอร์ หากค่าใดทำให้ความแม่นยำสูงขึ้น จุดกึ่งกลางในการทดสอบจะย้ายไปยังค่านั้นและทำซ้ำเช่นเดิมอีก หากพบว่าไม่มีค่าใดทำให้โมเดลมีความแม่นยำดีขึ้น ในรอบถัดไปจะมีการกำหนดค่าขั้นของการสุ่มให้ละเอียดขึ้น และทำซ้ำเช่นเดิมไปเรื่อย ๆ จนกว่าค่าความละเอียดของขั้นการสุ่มจะลดลงถึงค่าที่กำหนดให้หยุด ซึ่งวิธีนี้มีจุดดีตรงที่มีจำนวนรอบการทำงานเพื่อทดสอบน้อยกว่าการค้นหาแบบตาราง

2.4.7 ข้อจำกัดของเอสวีเอ็ม Burges ได้กล่าวถึงข้อจำกัดของการเรียนรู้ของเครื่องแบบเอสวีเอ็มไว้ดังนี้

2.4.7.1 การเลือกฟังก์ชันแกนที่เหมาะสม ซึ่งงานวิจัยส่วนใหญ่มักกำหนดฟังก์ชันแกนไว้ก่อน แล้วจึงปรับพารามิเตอร์ของฟังก์ชันแกนเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ดังนั้นหากต้องการเลือกฟังก์ชันแกนที่เหมาะสมกับข้อมูลนั้นก็ยังต้องมีการวิจัยเพิ่มเติม

2.4.7.2 ความเร็วและขนาดของข้อมูลในการเรียนรู้และทดสอบ โดยเฉพาะเมื่อใช้กับข้อมูลที่มีจำนวนซัพพอร์ตเวกเตอร์จำนวนมาก เช่น มากกว่าล้านซัพพอร์ตเวกเตอร์ เป็นต้น จะทำให้ระบบไม่สามารถทำงานได้

2.4.7.3 ข้อมูลส่วนใหญ่ไม่ได้เป็นข้อมูลแบบสองกลุ่มเสมอไป ดังนั้นการออกแบบวิธีการในการคัดแยกแบบหลายกลุ่ม เป็นประเด็นที่ต้องทำวิจัยเพื่อพัฒนาให้ตอบสนองกับข้อมูลหลากหลายแบบมากขึ้น ในงานวิจัยนี้ผู้วิจัยได้เลือกเอสวีเอ็มเป็นระบบเรียนรู้ในระบบการจัดหมวดหมู่เอกสารแบบอัตโนมัติ โดยทำการศึกษาข้อมูลที่นำมาใช้กับฟังก์ชันแกนแบบต่าง ๆ และทำการปรับพารามิเตอร์ โดยใช้วิธีการแบบการค้นหาแบบตาราง แม้ว่าจะต้องใช้จำนวนรอบใน

การทำงานมากและประมวลผลช้า เนื่องจากผู้วิจัยต้องการศึกษาคุณลักษณะของพารามิเตอร์ที่ส่งผลต่อความถูกต้องของระบบ ส่วนการตรวจสอบความถูกต้องของการปรับพารามิเตอร์ใช้การครอสวาลิเดชั่นแบบเคโฟล์ดบนการคัดแยกข้อมูลแบบหลายกลุ่มด้วยวิธีการคัดแยกทีละหนึ่งต่อหนึ่ง ซึ่งเป็นวิธีที่มีความรวดเร็วในการประมวลผลและมีประสิทธิภาพดี

2.5 เนอีฟเบย์

วิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดี เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกันมีการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnosis) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่นทำให้ผู้วิจัยเลือกวิธีการนี้มาใช้ในงานวิจัย เนื่องจากเป็นวิธีการจำแนกข้อมูลได้อย่างมีประสิทธิภาพและมีอัลกอริทึมในการทำงานที่ไม่ซับซ้อนเหมือนวิธีการอื่นๆ

กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม v_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X=\{ a_1, a_2, ..., a_n\}$ หรือใช้สัญลักษณ์ว่า $P(a_1, a_2, ..., a_n \mid v_j)$ ดังสมการที่ 2-12

$$P(a_1,a_2,...a_n \mid v_j) = \prod_{i=1}^{n} P(a_i \mid v_j) \tag{2-12}$$

โดยที่ $\prod$ หมายถึง ผลคูณของค่า $P(a_i \mid v_j)$ ทั้งหมด

$i = 1,2,3,\ldots,n$ และ $j = 1,2,3,\ldots,n$

การนำวิธีการเรียนรู้เบย์อย่างง่ายไปใช้ มีวิธีการดังต่อไปนี้ คือ

1. หาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า $P(a_1, a_2, ..., a_n \mid v_j)$ จากสมการที่ 1 มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้นๆ คือ $P(v_j)$ ได้เท่ากับ v_{NB}

2. นำค่าที่ได้ มาเปรียบเทียบกัน กลุ่มที่มีค่าความน่าจะเป็นสูงสุด คือ คำตอบ ดังนั้นจะได้ว่าวิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการที่ 2-13

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \times \prod_{i=1}^{n} P(a_i \mid v_j) \tag{2-13}$$

จากสมการที่ 2-13 สามารถเขียนเป็นอัลกอริทึมการเรียนรู้เบย์อย่างง่าย ได้ดังต่อไปนี้

- **Naive_Bayes_Learn(examples)**
  FOR EACH target value v **DO**
  $\overline{P}(v_j) \leftarrow$ estimate $P(v_j)$
  FOR EACH attribute value a of each attribute **DO**
  $\overline{P}(a_i \mid v_j) \leftarrow$ estimate $P(a_i \mid v_j)$
- **Classify_New_Example(x)**

$$v_{NB} = \arg\max_{v_j \in V} \overline{P}(v_j) \times \prod_{i=1}^{n} \overline{P}(a_i \mid v_j)$$

ภาพที่ 2-14 อัลกอริทึมการเรียนรู้เบย์อย่างง่าย

การจัดกลุ่มเอกสารเป็นเทคนิคสำคัญของการจำแนกประเภทข้อความ จุดประสงค์เพื่อจำแนกเอกสารโดยอัตโนมัติ มีอัลกอริทึมมากมายที่ถูกพัฒนาขึ้นสำหรับการจัดกลุ่มเอกสารแบบอัตโนมัติอีกหนึ่งวิธีที่นิยมใช้ทั่วกัน ก็คือ เบย์อย่างง่าย ซึ่งมีงานวิจัยที่ศึกษาอัลกอริทึมแบบเบย์อย่างง่ายมากมาย ในการนำเอกสารมาเรียนรู้เพื่อให้ได้ผลลัพธ์ที่มีความถูกต้อง

ในการเรียนรู้เพื่อจำแนกประเภทเอกสารโดยใช้เบย์อย่างง่าย สมมติว่ามีข้อความที่สนใจกับไม่สนใจ เมื่อทำการเรียนรู้แล้วต้องการทำนายว่าเอกสารหนึ่งๆ จะเป็นเอกสารที่สนใจหรือไม่ สามารถนำประยุกต์ใช้งาน เช่น การกรองข่าวสารเลือกเฉพาะข่าวที่สนใจ เป็นต้น

สมมติให้เอกสารหนึ่ง ๆ คือ ตัวอย่างหนึ่งตัว และแทนเอกสารแต่ละฉบับด้วยเวกเตอร์ของคำโดยใช้คำที่ปรากฏในเอกสารเป็นคุณสมบัติของเอกสาร กล่าวคือ คำที่หนึ่งในเอกสารเป็นคุณสมบัติตัวที่หนึ่ง คำที่สองในเอกสารเป็นคุณสมบัติตัวที่สอง ตามลำดับ ดังนั้นจะได้ว่า a_1 คือคุณสมบัติของคำที่หนึ่ง a_2 คือคุณสมบัติของคำที่สองตามลำดับ จากนั้นสร้างการเรียนรู้ข้อมูลโดยใช้ตัวอย่างสอนเพื่อประมาณค่าความน่าจะเป็นจากสมมติฐานเรื่องความไม่ขึ้นต่อกันของคุณสมบัติของเบย์อย่างง่ายทำให้ได้ สมการหาค่าความน่าจะเป็นของเอกสารตามสมการที่ (2-14)

$$P(doc \mid v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k \mid v_j) \qquad (2\text{-}14)$$

เมื่อ a_i คือ คุณสมบัติตัวที่ i

w_k คือ คำที่ k ในรายการของคำที่เรามีอยู่

$P(a_i = w_k \mid v_j)$ คือ ความน่าจะเป็นที่คำในตำแหน่งที่ i ของคำที่ k ในรายการของคำที่พบในเอกสาร v_j

$P(doc \mid v_j)$ คือ ความน่าจะเป็นของแต่ละเอกสาร v_j

อัลกอริทึมสำหรับการจำแนกประเภทข้อความโดยใช้การเรียนรู้เบย์อย่างง่ายเป็นดังต่อไปนี้

Algorithm: Learn_naive_Bayes_text(*Examples*,V)

1. Collect all words and other tokens that occur in *Examples*.
   - *Vocabulary* ← all distinct words and other tokens in *Examples*.
2. Calculate the required $P(v_j)$ and $P(w_k|v_j)$:
   - **FOR EACH** target value v_j in V **DO**
     - $docs_j$ ← subset of *Examples* for which the target value is v_j
     - $P(v_j) = \frac{|docs_j|}{Examples}$
     - $Text_j$ ← a single document created by concatenating all members of $docs_j$
     - n ← total number of words in $Text_j$ (counting duplicate words multiple times)
     - FOR EACH word w_k in *Vocabulary* **DO**
       - n ← number of times word w_k occurs in $Text_j$
       - $P(w_k|v_j) = \frac{n_k + 1}{n + |Vocabulary|}$

Algorithm: Classify_naive_Bayes_text(*Doc*)

- positions ← all word positions in *Doc* that contain tokens found in *Vocabulary*
- Return v_{NB}

$$v_{NB} = \arg\max_{v_i \in V} P(v_j) \times \prod_{i \in positions} P(a_i | v_j)$$

ภาพที่ 2-15 อัลกอริทึมการเรียนรู้เบย์อย่างง่ายสำหรับจำแนกประเภทเอกสาร

จากอัลกอริทึมด้านบน สามารถอธิบายรายละเอียดได้ดังนี้

1. นำคำที่ผ่านการตัดคำจากเอกสารมาสร้างเป็นข้อมูลตัวอย่าง

2. คำนวณหาค่าความน่าจะเป็นของกลุ่มเอกสาร $P(v_j)$ และคำนวณหาค่าความน่าจะเป็นของคำที่พบทั้งหมดในเอกสาร $P(a_i | v_j)$ โดยคำนวณได้จาก

2.1 คำนวณหาค่า $P(v_j)$ ได้จากสมการที่ (2-15)

$$P(v_j) = \frac{\text{จำนวนเอกสารในกลุ่มเอกสารนั้น ๆ}}{\text{จำนวนเอกสารทั้งหมด}} \quad (2\text{-}15)$$

2.2 คำนวณหาค่า $P(a_i | v_j)$ ได้จาก

$$P(a_i | v_j) = \frac{\text{ผลรวมความถี่ของคำที่พบในกลุ่มเอกสารนั้น ๆ}}{\text{จำนวนเอกสารในกลุ่มเอกสารนั้น ๆ}} \quad (2\text{-}16)$$

3. คำนวณหาค่า $P(v_j) \times \prod_{i=1}^{n} P(a_i \mid v_j)$ หมายถึง ค่าความน่าจะเป็นของเอกสารในแต่ละกลุ่มเอกสาร คือ ผลคูณระหว่างค่าความน่าจะเป็นของกลุ่มเอกสารกับค่าความน่าจะเป็นของคำที่พบทั้งหมดในเอกสาร

4. เลือกกลุ่มเอกสารที่มีค่าความน่าจะเป็นมากที่สุด เป็นคำตอบ

(Kruengkrai, Canasai and Jaruskulchai, Chuleerat, 2001; บุญเสริม กิจศิริกุล, 2548)

2.6 งานวิจัยที่เกี่ยวข้อง

นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ (2554) ได้นำเสนอแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทย โดยลดคุณลักษณะของเอกสารก่อนประมวลผลด้วยเครื่องจักรการเรียนรู้ เพื่อประโยชน์ในการลดมิติของข้อมูล ลดระยะเวลาประมวลผล ประหยัดทรัพยากรของระบบ และเพิ่มประสิทธิภาพในการจัดหมวดหมู่เอกสารภาษาไทย จากการทดลองพบว่าการลดคุณลักษณะด้วยวิธีอินฟอร์เมชันเกรน (Information Gain) และประมวลผลด้วยเครื่องจักรการเรียนรู้แบบต่างๆ บนแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทยโดยวัดประสิทธิภาพการจัดหมวดหมู่เอกสารที่ระดับคุณลักษณะที่ดีสุด ได้ทดสอบกับเอกสารประเภทข่าวอิเล็กทรอนิกส์จากหนังสือพิมพ์ไทยรัฐ จำนวน 10 กลุ่ม ได้แก่ กลุ่มการศึกษา กลุ่มบันเทิง กลุ่มสังคม กลุ่มการเมือง กลุ่มเทคโนโลยี กลุ่มกีฬา กลุ่มข่าวต่างประเทศ กลุ่มเกษตร กลุ่มเศรษฐกิจ กลุ่มวัฒนธรรม โดยมีจำนวนกลุ่มตัวอย่างทั้งหมด 12,000 เอกสาร โดยเป็นกลุ่มตัวอย่างเรียนรู้ (Training) จำนวน 10,000 เอกสาร และกลุ่มทดลอง (Testing) จำนวน 2,000 เอกสาร พบว่าอัลกอริทึมเอสวีเอ็มให้ประสิทธิภาพสูงสุด คือ 94.3% รองลงมาเป็นอัลกอริทึมเบย์ 86.2% อัลกอริทึมอาร์บีเอฟ (RBF) 86.1% อัลกอริทึม J48 79.7% อัลกอริทึมริปเปอร์ 78.9% อัลกอริทึมเคเอ็นเอ็น (KNN) 70.3% ตามลำดับ เมื่อพิจารณาด้านการลดขนาดคุณลักษณะจากกลุ่มตัวอย่างของอัลกอริทึมเอสวีเอ็ม พบว่าสามารถลดลงได้มากถึง 90% โดยการลดลงของคุณลักษณะดังกล่าวไม่ส่งผลให้ประสิทธิภาพในการจัดหมวดหมู่เอกสารลดลงแต่อย่างใด

พิลาพรรณ โพธิ์นรินทร์, สุพจน์ นิตย์สุวัฒน์ และชูชาติ หฤไชยะศักดิ์ (2554) ได้ทำวิจัยโดยนำข้อมูลเชิงบรรณานุกรมซึ่งเป็นข้อมูลรายงานวิจัยด้านการเกษตร จากศูนย์สารสนเทศทางการเกษตรแห่งชาติ สำนักหอสมุด มหาวิทยาลัยเกษตรศาสตร์ จำนวน 2,850 ระเบียน ในการจำแนกหมู่หมวดข้อมูล แบ่งออกเป็น 5 หมวดหมู่ คือ การปรับปรุงพันธุ์ การใช้ปุ๋ยและการปรับปรุงดิน ศัตรูพืช โรคพืช และสภาพแวดล้อม และเลือกใช้คำสำคัญมาทำการสกัดคุณลักษณะ (Feature Extraction) เพื่อดึงคุณลักษณะของคำมา มาใช้ในการทดลองจำแนกหมวดหมู่ข้อมูลแบบมีการเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยเลือกใช้อัลกอริธึมต้นไม้การตัดสินใจ นาอีฟเบย์

และซัพพอร์ทเวกเตอร์แมชชีน จากผลการทดลองอัลกอริธึมที่ดีที่สุด คือ ซัพพอร์ทเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบเรเดียลเบสิกฟังก์ชัน (Radial Basis Function) มีค่าความถ่วงดุลเท่ากับ 87.4%

กมลาสน์ อุดมล้ำเลิศ และอานนท์ รุ่งสว่าง (2553) กล่าวว่า การนำเสนอเอกสารเว็บเพจที่ถูกจัดแบ่งเป็นหมวดหมู่ตามเนื้อหาที่ปรากฏหรือที่เรียกว่า “เว็บไดเร็กทอรี่” เป็นการให้บริการในอีกรูปแบบหนึ่งที่ได้รับนิยมอย่างแพร่หลายของระบบสืบค้นข้อมูลเว็บยุคใหม่ แต่ด้วยปริมาณเว็บเพจที่มีจำนวนมากขึ้นเรื่อย ๆ จึงได้ทำการงานวิจัยเกี่ยวการจำแนกหมวดหมู่อัตโนมัติโดยใช้เครื่องจักรเรียนรู้ โดยศึกษาตัวจำแนกเว็บเพจภาษาไทยโดยเปรียบเทียบการใช้ตัวคัดแยกที่แตกต่างกัน การลดจำนวนคุณลักษณะโดยการคัดเลือกคุณลักษณะ และการให้น้ำหนัก จากผลการทดลองกับข้อมูลทดสอบ 2 ชุด ซึ่งคัดเลือกมาจากเว็บไดเร็กทอรี่ภาษาไทย 5 หมวดหมู่หลัก ทำให้เห็นว่าการลดคุณลักษณะโดยวิธีการคัดเลือกคุณลักษณะที่เหมาะสมจะไม่ส่งผลต่อความถูกต้องในการจำแนกหมวดหมู่ และการลดคุณลักษณะประกอบกับการใช้ตัวคัดแยกแบบนาอีฟเบย์ร่วมกับเทคนิคการให้น้ำหนักคุณลักษณะจะให้ค่าประสิทธิผลดีที่สุด โดยมีค่าความถูกต้อง (F-measure) 83.4%

จันทิมา พลพินิจ และอุมาภรณ์ สายแสงจันทร์ (2548) กล่าวว่า เนื่องจากระบบเวิร์ลไวด์เว็บได้รับความนิยมในการใช้งานเป็นอย่างมาก เพราะความสามารถในการให้บริการเกี่ยวกับสารสนเทศที่มีความหลากหลาย ทำให้เกิดปัญหาที่ต้องได้รับการแก้ไขเพราะปริมาณของสารสนเทศที่มีจำนวนมหาศาล ซึ่งปัญหาดังกล่าวก็คือ การค้นหาและการเลือกใช้สารสนเทศจะไม่สามารถกระทำได้ในเวลาอันรวดเร็ว การจัดกลุ่มเอกสารข้อความอัตโนมัติจึงกลายเป็นอีกทางเลือกที่สามารถช่วยให้การเข้าถึงสารสนเทศได้เร็วขึ้น จึงได้นำเสนอการจัดกลุ่มเอกสารข้อความภาษาไทยแบบอัตโนมัติบนพื้นฐานของพจนานุกรม และอัลกอริทึมซัพพอร์ตเวคเตอร์แมชชีน ชุดข้อมูลทดสอบคือเอกสารข่าวที่เก็บข้อมูลมาจากหนังสือพิมพ์ไทยรัฐและเดลินิวส์ระหว่างวันที่ 1 มกราคม-31 มีนาคม 2546 โดยจะใช้ข้อมูลจำนวน 1,000 เอกสารสำหรับการเรียนรู้และ 500 เอกสารสำหรับการทดสอบ ภายหลังจากที่ระบบพัฒนาเสร็จสิ้นได้มีการทดสอบประสิทธิภาพของระบบด้วยการวัดค่าเอฟแล้วพบว่าระบบให้ผลของความถูกต้องในการจัดกลุ่มเอกสารอยู่ระหว่างร้อยละ 77.46%-84.70%