การวิเคราะห์จดหมายข่าวขยะด้วยวิธีการจัดกลุ่ม

SPAM Mail Analysis Based Clustering Algorithm

สืบพงศ์ ฉายจันทร์

50925372

งานนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
กุมภาพันธ์ 2556

คณะกรรมการควบคุมงานนิพนธ์และคณะกรรมการสอบงานนิพนธ์ได้พิจารณางานนิพนธ์ของ สืบพงศ์ ฉายจันทร์ ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมงานนิพนธ์

ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร อาจารย์ที่ปรึกษา

คณะกรรมการสอบงานนิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.ชิดชนก เหลือสินทรัพย์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)

คณะวิทยาการสารสนเทศ อนุมัติให้รับงานนิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยบูรพา

..... คณบดีคณะวิทยาการสารสนเทศ
(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)
วันที่ 28 เดือน กุมภาพันธ์ พ.ศ. 2556

กิตติกรรมประกาศ

งานนิพนธ์เรื่องการวิเคราะห์จดหมายข่าวขยะด้วยวิธีการจัดกลุ่ม (SPAM Mail Analysis Based Clustering Algorithm) สำเร็จลุล่วงได้ด้วยดีเพราะได้รับความกรุณาจากอาจารย์ที่ปรึกษางานนิพนธ์ ผศ.ดร.กฤษณะ ชินสาร ที่ได้กรุณาเสียสละเวลาอันมีค่า อนุเคราะห์ให้คำปรึกษา คำแนะนำตลอดจนกรุณาช่วยตรวจแก้ไขข้อบกพร่องต่างๆ ให้มีความถูกต้องครบถ้วนสมบูรณ์ ทำให้ผู้เขียนงานนิพนธ์ได้รับแนวทางในการศึกษาค้นคว้าหาความรู้ และประสบการณ์อย่างกว้างขวางในการทำงานนิพนธ์ครั้งนี้

ขอกราบขอบพระคุณคณาจารย์ทุกท่าน คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ที่ได้ประสิทธิ์ประสาทความรู้ทางวิชาการ และข้อแนะนำที่เป็นประโยชน์ตลอดระยะเวลาการศึกษา

ขอขอบพระคุณมารดา และครอบครัวฉายจันทร์ ที่ให้โอกาสในการเข้ารับการศึกษา เป็นกำลังใจและเป็นแรงผลักดันจนประสบความสำเร็จในการศึกษาครั้งนี้

ขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ ชาวเทคโนโลยีสารสนเทศ ทุกท่านที่ให้ความช่วยเหลือในด้านการเรียนและเป็นกำลังใจที่ดี ตลอดระยะเวลาการศึกษาครั้งนี้

ขอขอบคุณมหาวิทยาลัยบูรพา ที่สรรสร้างบัณฑิตใหม่ทุกรุ่น ดังคำขวัญที่ว่า สุโข ปญฺญาปฏิลาโภ (ความได้ปัญญา ให้เกิดสุข)

สุดท้ายนี้ หากงานนิพนธ์นี้มีคุณงามความดี ก่อเกิดประโยชน์แก่ผู้สนใจบ้าง ขอมอบความดีอันพึงมีทั้งหมดแก่คณาจารย์ที่ได้ประสิทธิ์ประสาทวิชา และผู้มีพระคุณทุกท่าน ที่ได้ให้ความช่วยเหลือและห่วงใยมาโดยตลอด

สืบพงศ์ ฉายจันทร์

กุมภาพันธ์ 2556

50925372: สาขาวิชา: เทคโนโลยีสารสนเทศ วท.ม. (เทคโนโลยีสารสนเทศ)
คำสำคัญ: จดหมายข่าวขยะ, วิเคราะห์องค์ประกอบหลัก, การจัดกลุ่มข้อมูล

สืบพงศ์ ฉายจันทร์: การวิเคราะห์จดหมายข่าวขยะด้วยวิธีการจัดกลุ่ม (SPAM Mail Analysis Based Clustering Algorithm) อาจารย์ผู้ควบคุมงานนิพนธ์: ดร.กฤษณะ ชินสาร, 60 หน้า. ปี พ.ศ. 2556.

งานนิพนธ์นี้ได้นำเสนอผลงานศึกษา การวิเคราะห์จดหมายข่าวขยะด้วยวิธีจัดกลุ่ม โดยใช้ข้อมูลจาก Ling Spam Corpus เป็นชุดข้อมูลจดหมายที่ใช้สำหรับเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ ซึ่งในการดำเนินงานเพื่อเพิ่มประสิทธิภาพของการวิเคราะห์ จึงได้มีขั้นตอนการวิเคราะห์องค์ประกอบหลักของชุดข้อมูลทั้งหมด เพื่อทำการลดจำนวนตัวแปรหรือคำของจดหมายให้มีเฉพาะตัวแปรที่มีความสัมพันธ์ของชุดข้อมูล แล้วจึงนำเสนอขั้นตอนวิธีการวิเคราะห์จดหมายข่าวขยะประกอบด้วย ขั้นตอนวิธีการเบย์เซียน ขั้นตอนวิธีซี4.5 และขั้นตอนวิธีเคเนียเรส เนเบอร์ ผลปรากฏว่าค่าเฉลี่ยความถูกต้องสูงสุดคือวิธีซี4.5 มีค่าเฉลี่ย 97.276% รองลงมาคือขั้นตอนวิธีเคเนียเรส เนเบอร์ (ที่ k=5) มีค่าเฉลี่ย 95.7695% และสุดท้ายขั้นตอนวิธีการเบย์เซียน มีค่าเฉลี่ย 84.889% ซึ่งจากผลการดำเนินงาน ค่าเฉลี่ยความถูกต้องของวิธีซี 4.5 ที่มีค่าสูงสุด ได้นำไปพัฒนาเป็นโปรแกรมต้นแบบระบบตรวจจับจดหมายข่าวขยะ อย่างมีประสิทธิภาพตามการทดลอง

50925372: MAJOR: INFORMATON TECHNOLOGY; M.Sc. (INFORMATON TECHNOLOGY)

KEY WORDS: SPAM E-mail, Principle Component Analysis, Clustering Algorithm

SEUBPONG CHAICHAN: SPAM Mail Analysis Based Clustering Algorithm.

ADVISOR: KRISANA CHINNASARN, PH.D. 60 P. 2013

The Literary Work proposes the analysis of spam e-mail by classification using Ling Spam Corpus as e-mail of data set. They'll be used for training data set and testing data set in order to enhance efficiency of analysis. Therefore, the processes of Principal Component Analysis of all primary data sets are investigated to decrease the variables (e-mail wording tokens). It is resulted to have the only variable related to e-mail data sets. Subsequently, analysis techniques of spam e-mail are selected to 3 types; Bayesian methodology, C4.5 algorithm and K-Nearest Neighbor algorithm. Consequently, max average accuracy is to C4.5 97.276%, K-Nearest Neighbor equals to 95.7695%, and Bayesian 84.889% respectively. The highest result of this research is C4.5 algorithm. Then it will be developed to prototype of spam e-mail detection system effectively.

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูปภาพ

บทที่ 1
บทนำ

1.1 ที่มาของงานนิพนธ์

สืบเนื่องจากความนิยมในการใช้งานระบบคอมพิวเตอร์ส่วนบุคคล ระบบคอมพิวเตอร์แม่ข่าย และระบบเครือข่ายคอมพิวเตอร์ที่มีอย่างแพร่หลาย ขยายไปจนทำให้เกิดระบบเครือข่ายขนาดใหญ่ครอบคลุมทั่วโลกหรือที่เรียกว่าระบบเครือข่ายอินเตอร์เน็ต

จึงทำให้เกิดช่องทางที่ใช้ติดต่อสื่อสารผ่านระบบอินเตอร์เน็ตและได้รับความนิยมอย่างสูง นั้นก็คือการใช้อีเมลล์ (E-mail) หรือจดหมายอิเล็คโทรนิค ซึ่งเป็นการติดต่อสื่อสารที่ได้รับความนิยมอย่างแพร่หลาย แต่ทว่าการใช้จดหมายอิเล็คโทรนิคก็ได้รับผลกระทบที่ได้เกิดมาพร้อมกับระบบคอมพิวเตอร์ในยุคไอที นั่นก็คือไวรัสคอมพิวเตอร์ (Virus Computer) ที่แพร่กระจายตามการขยายตัวของระบบคอมพิวเตอร์และระบบเครือข่ายอินเตอร์เน็ต

อีเมลล์ได้กลายเป็นส่วนหนึ่งของปัญหาในการแพร่กระจายไวรัส ไม่ว่าจะเป็นอีเมลล์โฆษณา อีเมลล์ที่ไม่ได้มีเนื้อหาตามที่ผู้รับอีเมลล์ต้องการ ทั้งหมดนี้มีอีกชื่อหนึ่งคือ สแปมอีเมลล์ (SPAM E-mail) หรือจดหมายข่าวขยะ ซึ่งมักจะเกิดขึ้นตามมาพร้อมกับอีเมลล์ไวรัสคอมพิวเตอร์หรือเครื่องคอมพิวเตอร์ที่ได้ถูกไวรัสคอมพิวเตอร์เล่นงาน (Infected) และกลายเป็นผู้ส่งจดหมายข่าวขยะเอง ซึ่งปัญหาการส่งจดหมายข่าวขยะ (SPAM E-mail) นับวันยิ่งทวีคูณความรุนแรงขึ้นในยุคไอที (IT)

จึงได้มีการคิดงานวิจัยที่จะเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีในการพิจารณาจดหมายข่าวขยะด้วยวิธีการจัดกลุ่ม จากเทคนิคการวิเคราะห์ตัวประกอบหลัก ตลอดจนการเปรียบเทียบกับขั้นตอนวิธีในการพิจารณาจดหมายข่าวขยะด้วยวิธีต่างๆ ที่ได้ศึกษา

1.2 ความสำคัญของปัญหา

1.2.1 ขั้นตอนในการวิเคราะห์อีเมลล์ด้วยวิธีการต่างๆ นั้นจะใช้ตัวแปรจากคำที่ปรากฏอยู่ในจดหมายโดยตรง ทำให้ให้เกิดตัวแปรจำนวนมากตามจำนวนอีเมลล์ที่ต้องการวิเคราะห์

1.2.2 การวิเคราะห์จดหมายข่าวขยะที่มีจำนวนตัวแปรมากทำให้ผลการวิเคราะห์ หรือคำนวณต้องใช้เวลาขึ้นตามจำนวนตัวแปร

1.2.3 จากการศึกษางานนิพนธ์ต่างๆ พบว่าชุดข้อมูลที่ใช้ทดสอบ เป็นการนำตัวแปรจากจดหมายมาวิเคราะห์โดยตรง จึงมีความคิดที่พัฒนาขั้นตอนเพื่อทำให้จำนวนตัวแปรลดลง แต่ประสิทธิภาพในการวิเคราะห์จดหมายข่าวขยะยังคงเหมือนเดิมหรือดีขึ้น

1.3 วัตถุประสงค์ของการศึกษา

1.3.1 เพื่อศึกษาขั้นตอนวิธี ที่ใช้ในการจำแนกจดหมายข่าวขยะ โดยการคำนวณความถูกต้องทางสถิติได้

1.3.2 ทำการการวิเคราะห์ตัวประกอบหลัก (Principal Component Analysis) จากรูปแบบคำข้อมูลของจดหมายข่าวขยะ

1.3.3 นำขั้นตอนวิธีการวิเคราะห์จดหมายข่าวขยะเพื่อเปรียบเทียบผลในการจัดกลุ่มจดหมายข่าวขยะ

1.3.4 เป็นแนวทางการศึกษาขั้นสูง เพื่อพัฒนาระบบสารสนเทศในการกรองจดหมายข่าวขยะ

1.4 ขอบเขตของการศึกษา

1.4.1 วิเคราะห์รูปแบบของชุดตัวอย่างจดหมายข่าวขยะจาก Ling Spam Corpus ซึ่งเป็นแฟ้มข้อความ (text file) จำนวน 2,893 ฉบับ จะมีเฉพาะหัวข้อจดหมาย (Subject) และเนื้อจดหมาย (Content) และได้ตัดในส่วนของหัวจดหมายอิเล็คโทรนิคส์ (E-mail Header) ออก

1.4.2 ชุดข้อมูล Ling Spam Corpus ได้ผ่านกระบวนการ Lemm_Stop แล้ว คือผ่านกระบวนการลดรูปของคำและคำที่เป็น Stop-list แล้ว โดยมีจำนวนคำทั้งสิ้น 59,829 คำ

1.4.3 ชุดข้อมูลที่วิเคราะห์มีทั้งในส่วนที่เป็นชุดข้อมูลการเรียนรู้ (training data set) และชุดข้อมูลทดสอบ (testing data set) อยู่ด้วยกัน

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้ขั้นตอนวิธีที่เหมาะสมจากกระบวนการวิเคราะห์ตัวประกอบหลัก ในการวิเคราะห์การกระจายตัวของข้อมูล

1.5.2 ได้ผลลัพธ์จากระบบสารสนเทศที่ใช้ในการพิจารณาจดหมายข่าวขยะ

1.5.3 สามารถนำผลลัพธ์ของงานวิจัยในการวิเคราะห์จดหมายข่าวขยะ เพื่อเป็นต้นแบบในการพัฒนาระบบตรวจจับจดหมายข่าวขยะอย่างมีประสิทธิภาพ

1.6 ขั้นตอนและแผนการดำเนินงาน

ลำดับ	รายละเอียด	ตค. 2555		พย. 2555		ธค. 2555		มค. 2556		กพ. 2556		หมายเหตุ
1	ศึกษาเตรียมข้อมูลที่ใช้ใจงานวิจัย	←	—	—	→							
2	ออกแบบวางแผนการดำเนินงาน			←	—	→						
3	ดำเนินงาน			←	—	—	—	—	—	—	→	
4	สรุปผลการดำเนินงาน							←	—	—	→	

ตารางที่ 1-1 ตารางขั้นตอนและแผนการดำเนินงาน

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะนำเสนอแนวคิดทฤษฎีและขั้นตอนวิธีที่เกี่ยวข้องต่างๆ ที่ได้ใช้ในการทดลอง ซึ่งจะมีขั้นตอนวิธีตั้งแต่กระบวนการเตรียมชุดข้อมูลด้วย การวิเคราะห์ตัวประกอบหลัก และขั้นตอนวิธีในการจำแนกจดหมายข่าวขยะต่างๆ ดังนี้

- ขั้นตอนวิธีเบย์เซียน
- ทฤษฎีต้นไม้ตัดสินใจซี4.5
- ขั้นตอนวิธีเคเนียเรสต์เนเบอร์ (*k*-Nearest Neighbor: *k*-NN)

2.1 สแปมอีเมลล์หรือจดหมายข่าวขยะ

สแปมอีเมลล์หรือจดหมายข่าวขยะ คือจดหมายที่ผู้รับแต่ละคนไม่ต้องการ มีลักษณะเป็นจดหมายขยะประเภทหนึ่ง ผู้ที่ส่งจดหมายเหล่านั้นจะถูกเรียกว่าสแปมเมอร์ (SPAMMER) จดหมายดังกล่าวจะมีเนื้อหาที่ต้องการโฆษณา ชักชวน หรือแม้กระทั่งหลอกลวงไปยังผู้ที่ได้รับจำนวนมาก

ลักษณะจดหมายข่าวขยะจะถูกส่งจากสแปมเมอร์ไปยังผู้รับต่างๆ ทั่วโลก สแปมเมอร์ไม่จำเป็นต้องรู้จักหรือคุ้นเคยผู้รับปลายทาง เพียงแค่ทราบหรือมีอีเมลล์แอดเดรส และจะกระทำการโดยใช้ซอฟต์แวร์ให้ส่งจดหมายแบบหว่านแหตามเท่าที่จะสามารถทำได้ เป้าหมายของสแปมเมอร์มักจะเป็นโฆษณาเชิญชวนให้ซื้อสินค้าหรือขายของ โดยเจ้าของสินค้าอาจจะให้แฮ็คเกอร์ (Hacker) ทำการกระจายสแปมของเว็บหรือสินค้าตน และมักจะมีข้อความหลอกลวง ซึ่งบางครั้งจดหมายข่าวขยะของผู้รับคนหนึ่ง อาจไม่ใช่จดหมายข่าวขยะของอีกคนหนึ่งได้

ปัจจุบันการโจมตีจากแฮ็คเกอร์ก็มักจะใช้จดหมายข่าวขยะเป็นเครื่องมือที่ใช้ในการโจมตีด้วย เช่น การใช้ Bounce E-mail หรือจดหมายตีกลับ การฝังไวรัส โค้ด หรือ ลิ้งค์ (link) เพื่อให้ผู้เปิดจดหมายดังกล่าวติดไวรัสหรือโทรจัน และแม้กระทั่งหลอกด้วยข้อความทางจิตวิทยาได้

หัวข้อจดหมาย (Subject)	เนื้อหาจดหมาย (Content)	วันที่
FunPageLan	Anna Kournikova Topless	May 27
FunPageLand	*** A Rose For A Rose ***	May 27
Driver	Learn How to Get Paid to Drive your car	May 27
shawnda Oshinsky	Your Casino Winnings	May 27

FlowGo Fun Flash	A Little Valentine's Puss <--New Flowgo Fun	May 27
Eat to Lose	Eat your way Thinner in 10 days Guaranteed!	May 28
Marni Crofutt	Hot casino offers for you	May 28
debt-man NjExOQ	Let us help you get out of debt	May 28
ernest Arboleda	Find a longer long-term creation today	May 28
Int'l Wine Cellars	Get 3 Bottles of Wine FREE + a $29.95 Corkscr...	May 29
FlipPhone	Motorola i60 Flip Phone, With Direct Connect,...	May 29
Coleman Marv	Coleman Marv	May 29
Sean Francis	Sean Francis	May 29
Dollar Store	Own a Dollar Store for less than a dollar a d...	May 29
Coffee Lovers	Taste the Difference. Try Gevalia - Get a FRE...	May 30
BMW Offer from Digit...	Prize Entry for 2 BMW Mini Coopers #968895435	May 30
Its Natural	Increase bust size naturally… Guaranteed	May 30
cupids arrow MDExOQ	Web to door flower delivery to your valentine	May 30
FunPageLand	*** I Love You (in 8 languages) and Lots More...	May 31
SmilePop	Shed a tear for fallen heroes	May 31
Needahug@funstun.com	Warm your HEART with the best fun all week!	May 31
FlowGo Fun Flash	Prickly Lovin'<-- New Valentine's Flowgo Fun	May 31
"Royal" brokins dfps...	Hot casino offers for you	May 31

ตาราง 2-1 ตัวอย่างจดหมายข่าวขยะ

2.2 การวิเคราะห์ตัวประกอบหลัก (Principal Components Analysis)

เป็นวิธีที่ใช้ในการวิเคราะห์องค์ประกอบ (Components) จากองค์ประกอบที่แท้จริง (Real Factors) ซึ่งได้จากการคำนวณจากเมตริกซ์สหสัมพันธ์ (Correlation Matrix) จุดมุ่งหมายของการวิเคราะห์ส่วนประกอบสำคัญ คือสามารถกะประมาณ (Estimate) เมตริกซ์ สหสัมพันธ์และสามารถหาสมการลักษณะเมตริกซ์สหสัมพันธ์ (The characteristic equation of the matrix) 2 กลุ่มค่า (Sets of values) คือ

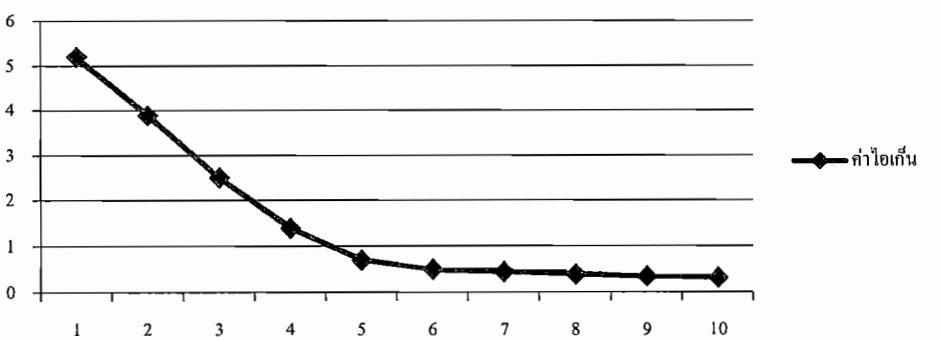
2.2.1 ไอเกนเวคเตอร์ (The characteristic vactors of the matrix : Laten vactors : Eigenvactors) สัญลักษณ์ V_a , V_b , ... ตามลำดับ ซึ่งเป็นคอลัมภ์หรือแถวของน้ำหนักของแต่ละตัวแปรในเมตริกซ์ ถ้ามี 6 ตัวแปรก็จะมีค่าน้ำหนัก 6 ค่า เช่น (V_a , V_b , ..., V_f) และมีค่าน้ำหนักองค์ประกอบที่สอดคล้อง (Corresponding) กับ องค์ประกอบต่างๆ คือ (F_a, F_b , ... , F_f) ได้มาจากเวคเตอร์คูณด้วยรากที่สองของค่าไอเกน (Eigenvalue) ขององค์ประกอบนั้น

2.2.2 ค่าไอเกน (The characteristic roots : Latent roots : Eigenvalues) ใช้สัญลักษณ์ I_a ซึ่งก็คือ ผลรวมของกำลังสองค่าน้ำหนักองค์ประกอบแต่ละองค์ประกอบ หาค่าเฉลี่ยนสัดส่วนของความแปรปรวนจากองค์ประกอบ หากผลรวมเฉลี่ยนในองค์ประกอบใดมีค่าสูง ก็อธิบายได้มากและจะถูกสกัดออกมาจะมีค่าสูงที่สุด

2.2.3 การพิจารณาจำนวนตัวประกอบที่เหมาะสม ค่าไอเกนที่หาได้จากความแปรปรวนจะถูกนำมาใช้พิจารณาจำนวนตัวประกอบหลักดังนี้

1. พิจารณาจากร้อยละความแปรปรวนสะสม ถ้าร้อยละความแปรปรวนสะสมของตัวประกอบหลัก m ตัวแรกจากจำนวนทั้งหมด p ตัว เป็นอย่างต่ำร้อยละ 80 ดังนั้น ตัวประกอบหลักควรจะมีจำนวนเท่ากับ m เช่น สมมติหากมีตัวประกอบหลักทั้งหมด 10 ตัว (C_1, C_2, ...C_{10}) p=1..10 หาก 4 ตัวแรก ที่มีค่าสะสมของร้อยละของความแปรปรวนมากกว่าร้อยละ 80 ดังนั้นจำนวนตัวประกอบที่เหมาะสมคือ m=3 (C_1, C_2, C_3, C_4)

2. ใช้กราฟในการพิจารณาจำนวนตัวประกอบหลักที่เหมาะสม โดยการพล๊อตค่าไอเกน ดังรูปที่ 2-1 หากตัวประกอบหลักตัวที่ m+1 มีค่าไอเกนต่ำมากเมื่อเปรียบเทียบกับตัวที่ m ก็ควรเลือกตัวประกอบที่เหมาะสมคือ m ตัว



รูปที่ 2-1 กราฟแสดงค่าไอเก็น

3. พิจารณาจากค่าเฉลี่ยความแปรปรวน หากค่าแปรปรวนของตัวประกอบหลักตัวใดน้อยกว่าค่าแปรปรวนเฉลี่ยให้ตัดออก

$$C_i \geq \frac{\sum_{i=1}^{p} I_i}{p} \qquad (2\text{-}1)$$

ทั้งนี้ในการพิจารณาจำนวนของตัวประกอบหลักที่เหมาะสมควรพิจารณาจากทั้ง 3 วิธี

2.3 ขั้นตอนวิธีเบย์เซียน (Bayesian Algorithm)

การเรียนรู้แบบเบย์ เป็นเทคนิคที่เลือกใช้ในการเรียนรู้เพื่อสร้างรูปแบบความน่าจะเป็น โดยเบย์เซียนเป็นวิธีการเรียนรู้ในสมมติฐานของความไม่ขึ้นต่อกันอย่างมีเงื่อนไข จึงใช้เบย์เซียนในการอธิบายความไม่ขึ้นต่อกันอย่างมีเงื่อนไข (Conditional Independence) ระหว่างตัวแปรหรือคุณสมบัติ โดยสามารถใส่เหตุการณ์ให้อยู่ในรูปของโครงสร้างข่ายงานและตารางความน่าจะเป็นอย่างมีเงื่อนไขดังสมการ

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \qquad (2\text{-}2)$$

จากสมการความน่าจะเป็นแบบมีเงื่อนไขที่จะเกิดเหตุการณ์ A เมื่อมีเหตุการณ์ B ปรากฏอยู่ ดังนั้น

$P(B|A)$ คือ ความน่าจะเป็นแบบมีเงื่อนไขของเหตุการณ์ B ต่อเหตุการณ์ A

$P(A)$ คือ ความน่าจะเป็นของเหตุการณ์ A

$P(B)$ คือ ความน่าจะเป็นของเหตุการณ์ B

ในการเรียนรู้สแปมจำเป็นต้องให้สามารถจดจำ วิเคราะห์ และสร้างฐานข้อมูลที่เก็บจำนวนครั้งของคำแต่ละคำที่ปรากฏทั้งที่เป็นจดหมายที่ดี (HAM) และจดหมายข่าวขยะ (SPAM) เช่นการกรองจดหมายที่มีเหตุการณ์ ของจดหมายที่มีคำว่า **'Viagra'** จำนวนมากปรากฏอยู่ จะทำให้ระบบทราบว่าจดหมายใหม่ที่ได้รับหากมีเหตุการณ์เดียวกันนั้นเป็น จดหมายข่าวขยะ ก็จะสามารถเรียนรู้ได้เองโดยอัตโนมัติ

การคำนวณความน่าจะเป็นของเหตุการณ์จดหมายใหม่จะเป็นจดหมายข่าวขยะ เมื่อมีจดหมายใหม่เข้ามา จะต้องมีการวิเคราะห์เนื้อหาของจดหมาย โดยแบ่งออกเป็นคำและทำการค้นหาคำสำคัญที่จะสามารถบ่งชี้ได้ว่า เป็นจดหมายที่ดีหรือจดหมายข่าวขยะ จะได้ว่า

$$P(Spam|token) = \frac{P(token|Spam)*P(Spam)}{P(token)} \quad (2\text{-}3)$$

ซึ่งกำหนดให้

P(SPAM|token) คือ ความน่าจะเป็นแบบมีเงื่อนไขที่**จะเป็นจดหมายข่าวขยะ** เมื่อมีเหตุการณ์คำ (token) ปรากฏอยู่

P(token|SPAM) คือ ความน่าจะเป็นแบบมีเงื่อนไขของจดหมายข่าวขยะใดใด ที่จะมีคำ (token) ปรากฏอยู่

P(SPAM) คือ ความน่าจะเป็นเหตุการณ์ก่อนที่จดหมายใดใด จะเป็นจดหมายข่าวขยะ

P(token) คือ ความน่าจะเป็นเหตุการณ์ก่อนหน้าที่จดหมายใดใด จะพบคำ (token) ที่ระบุเอาไว้

สมมติข้อมูลตัวอย่างของจดหมายข่าวขยะ

จดหมาย	จำนวนจดหมายดี (HAM)	จำนวนจดหมายข่าวขยะ (SPAM)	ผลรวม
จดหมายทั้งหมด	400	600	1000
จดหมายที่มีคำว่า “free”	200	500	700
จดหมายที่มีคำว่า “price”	10	90	100

ตารางที่ 2-2 ตัวอย่างข้อมูลจำนวนคำของจดหมาย

ความน่าจะเป็นแบบมีเงื่อนไข เหตุการณ์ของจดหมายข่าวขยะใดใด ที่จะมีคำ (token) ปรากฏอยู่

$$P(free|SPAM) = \frac{500}{600} = 0.83$$

$$P(price|SPAM) = \frac{90}{600} = 0.15$$

ดังนั้น ความน่าจะเป็นที่จะเป็นสแปมจากเหตุการณ์ดังกล่าวสามารถทำได้ดังนี้

$$P(SPAM|free) = \frac{0.83*0.6}{0.7} = 0.71$$

$$P(SPAM|price) = \frac{0.15*0.6}{0.1} = 0.9$$

จากความน่าจะเป็นดังกล่าว สามารถนำไปคำนวณจากการเรียนรู้แบบเบย์ได้ว่า

$$P(SPAM|free,price) = \frac{P(free,price|SPAM)*P(SPAM)}{P(free,price)}$$

$$P(HAM|free,price) = \frac{P(free,price|HAM)*P(HAM)}{P(free,price)}$$

ดังนั้น

$$1 = \frac{P(free,price|SPAM)*P(SPAM)}{P(free,price)} + \frac{P(free,price|HAM)*P(HAM)}{P(free,price)}$$

ถ้าสมมติให้

X = P(free|price)*P(SPAM) = P(free|SPAM)*P(price|SPAM)*P(SPAM)

Y = P(free|price)*P(HAM) = P(free|HAM)*P(price|HAM)*P(HAM)

$$P(SPAM|free,price) = \frac{X}{X+Y}$$

$$P(HAM|free,price) = \frac{Y}{X+Y}$$

แทนค่า X = 0.83*0.15*0.6 = 0.075

Y = 0.5*0.025*0.4 = 0.005

ดังนั้น $P(SPAM|free,price) = \frac{0.075}{0.075+0.005}$

P(SPAM|free,price) = 0.9375

แต่ละโทเคน (Token) จะนำไปคำนวณค่าความน่าจะเป็น และจะถูกรวมคะแนน เพื่อจะนำไปคำนวณคะแนนทั้งหมดให้กับจดหมายฉบับนั้นๆ สำหรับการประเมินค่าของ Graham จะนำค่าความน่าจะเป็นที่มีความสำคัญมากที่สุด 15 อันดับแรก สำหรับโทเคนพบอยู่ทั้งในจดหมายที่ดี (HAM) และปรากฏอยู่บนจดหมายข่าวขยะ (SPAM) จะไม่นำมารวมค่าความน่าจะเป็นเพราะไม่มีความสำคัญต่อการจัดจำแนก

2.4 ทฤษฎีต้นไม้ตัดสินใจซี4.5 (C4.5 Decision tree)

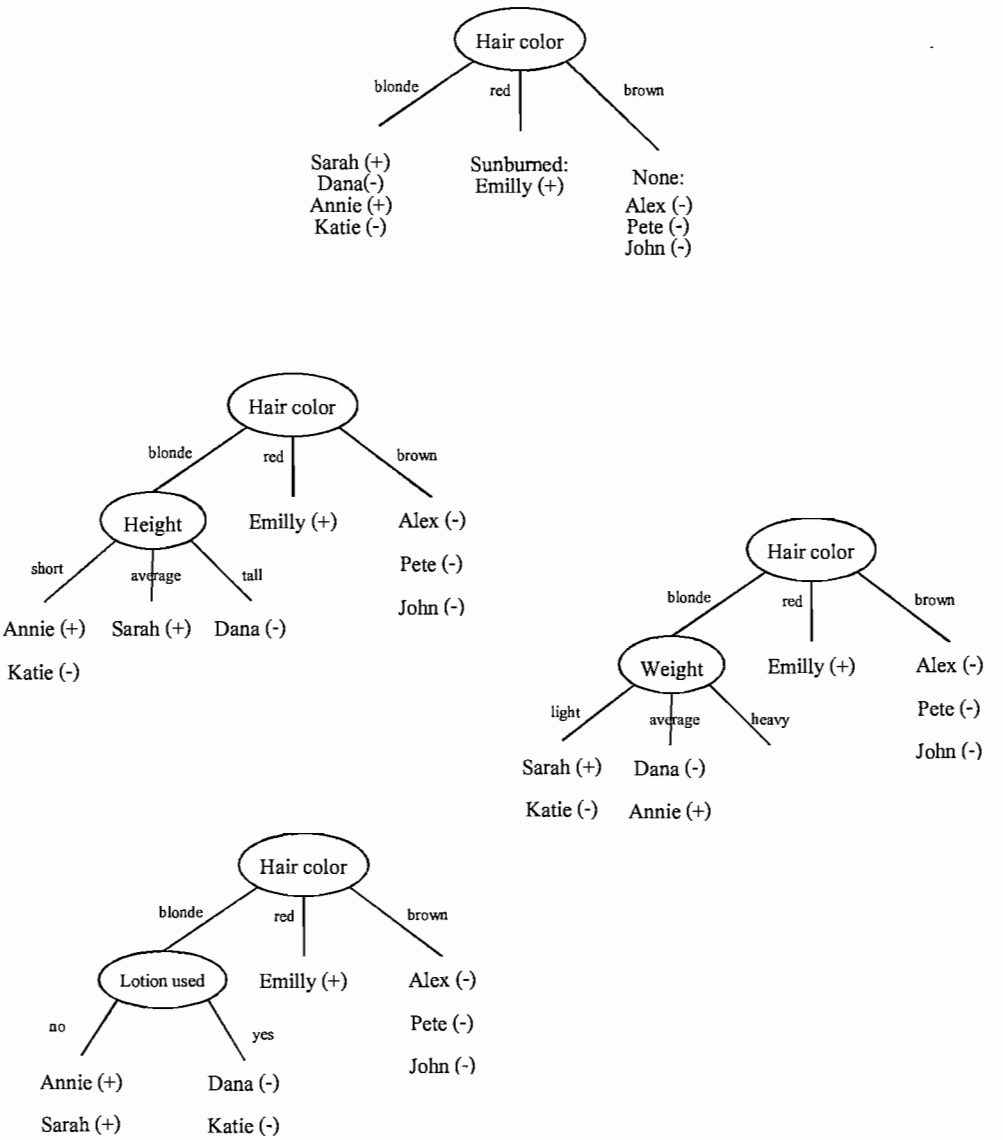
เป็นกระบวนการตัดสินใจโดยใช้การจำแนกประเภท (class) ของตัวอย่างที่มีค่าต่างๆ กัน ได้มีการนำมาใช้กับระบบผู้เชี่ยวชาญกันอย่างกว้างขวาง โดยเริ่มจากการป้อนข้อมูลตัวอย่างเข้าไปในระบบ ซึ่งข้อมูลตัวอย่างนั้นสามารถเป็นข้อมูลที่เป็นเชิงบวก หรือข้อมูลเชิงลบก็ได้ อีกทั้งข้อมูลตัวอย่างสามารถมีได้มากกว่า 2 ประเภท

ในการจำแนกประเภทข้อมูลจะประกอบไปด้วยคุณสมบัติ (attribute) ที่จะนำไปใช้สำหรับการประกอบการตัดสินใจของต้นไม้ และข้อมูลสุดท้ายคือชื่อ (name) ซึ่งจะนำไปใช้สำหรับการอ้างอิงของการจำแนกข้อมูล ต้นไม้ตัดสินใจประกอบด้วยบัพ (node) และกิ่ง (link) ที่ต่อกับบัพ บัพที่ปลายสุดเรียกว่า บัพใบ (leaf node) หรืออีกชื่อหนึ่งว่าใบ

บัพจะแสดงคุณสมบัติและกิ่งแสดงคำของคุณสมบัตินั้น ส่วนใบจะแสดงประเภท ในการสร้างต้นไม้ตัดสินใจ ทำโดยสร้างบัพทีละบัพเพื่อตรวจสอบคุณสมบัติของตัวอย่าง จากนั้นจึงแยกตัวอย่างลงตามค่าของกิ่ง ทำซ้ำจนกระทั่งตัวอย่างในใบแต่ละใบอยู่ในประเภทเดียวกันทั้งหมด

Name	Hair	Height	Weight	Lotion	Result
Sarah	blonde	average	light	no	sunburned
Dana	blonde	tall	average	yes	none
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned
Emily	red	average	heavy	no	sunburned
Pete	brown	tall	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none

ตารางที่ 2-3 ตัวอย่างข้อมูลต้นไม้ตัดสินใจ



รูปที่ 2-2 ตัวอย่างการสร้างต้นไม้ตัดสินใจ

2.4.1 ไอดี3 อัลกอริทึ่ม (ID3 Algorithm)

ฟังก์ชัน (Gain Function) ใช้วัดความสามารถในการตัดสินใจเลือกการแยกตัวอย่างของคุณสมบัติที่จะใช้เป็นรากหรือบัพในต้นไม้ตัดสินใจ คุณสมบัติที่มีค่าเกนสูงที่สุดจากการคำนวณค่าเกนจากกลุ่มตัวอย่างของข้อมูล จะนำมาเป็นรากหรือบัพ ซึ่งฟังก์ชันเกนจะประกอบไปด้วยการคำนวณค่าต่างๆ ดังนี้

ทฤษฎีสารสนเทศ (Information Theory) จะเป็นความรู้สำหรับนำไปใช้คำนวณโดยมาจากค่าความน่าจะเป็นของข้อมูล

ค่าสารสนเทศของข้อมูล = $-\log_2$ (ความน่าจะเป็นของข้อมูล) (2-4)

เอนโทรปี (Entropy) จะเป็นชุดของข้อมูลที่ใช้วัดค่าสารสนเทศโดยเฉลี่ยเพื่อระบุประเภทของข้อมูลที่คำนวณได้ เขียนแทนด้วย I(M) ซึ่งคำนวณจากสูตร

$$I(M) = \sum_i^n - \mathrm{P}(m_i)\ log_2\ \mathrm{P}(m_i) \tag{2-5}$$

ค่าเกน (Gain) เป็นคุณสมบัติบัพของเหตุการณ์ ที่จะนำมาใช้ในการแบ่งคุณสมบัติของบัพราก สามารถคำนวณได้จากการลบค่าเอนโทรปีทั้งหมดดังนี้

$$Gain(x) = I(T) - I_x\ (T) \tag{2-6}$$

ดังนั้นเราสามารถพิจารณาคุณสมบัติ hair color เป็นบัพราก โดยคำนวนจากหาค่าเกนได้ดังนี้

$$\begin{aligned} Gain(hair\ color) &= \left[-\left(\frac{3}{8}\right)\log_2\left(\frac{3}{8}\right) - \left(\frac{5}{8}\right)\log_2\left(\frac{5}{8}\right)\right] \\ &- \left[\frac{4}{8}\left(-\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)\right.\right. \\ &\left.\left.-\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)\right) + \frac{1}{8}\left(-\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right)\right) + \frac{3}{8}\left(-\left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right)\right) \quad\right] = 0.45 \end{aligned}$$

เมื่อคำนวณ คุณสมบัติอื่นๆ ก็จะได้ดังต่อไปนี้

$$Gain(height) = 0.26\ Gain(weight) = 0.01\ Gain(lotion) = 0.34$$

จากค่าที่คำนวณได้คุณสมบัติ hair color มีค่ามากที่สุดเป็นบัพแรกของต้นไม้ตัดสินใจ แต่ยังมีคุณสมบัติอื่นๆ ที่ยังแยกออกจากกันจากค่าของ blonde จึงจำเป็นต้องพิจารณาอื่นๆ ที่ตกมายังกิ่งถัดไป โดยใช้ฟังก์ชันเกนได้ค่าดังนี้

$$Gain(height) = 0.5\ Gain(weight) = 0.0\ Gain(lotion) = 1.0$$

ค่าที่ได้จะสามารถนำคุณสมบัติ lotion มีค่าเกนมากที่สุดมาแบ่งแยกข้อมูลได้เป็นต้นไม้ตัดสินใจต่อไป

2.4.2 ซี4.5 อัลกอริทึ่ม (C4.5 Algorithm)

เป็นส่วนขยายของอัลกอริทึ่ม ID3 ซึ่งพัฒนาโดย Ross Quinlan ซึ่งใช้ต้นไม้ประกอบการตัดสินใจสำหรับใช้ในการจำแนกข้อมูล ซึ่งจะนำค่าเกนและเอนโทรปี เหมือนกับ ID3 แต่มีส่วนสำคัญที่เพิ่มเติมดังนี้

ก. สามารถใช้งานได้ทั้งข้อมูลที่มีคุณลักษณะแบบต่อเนื่อง (Continuous) และแบบไม่ต่อเนื่อง (Discrete) สำหรับส่วนของข้อมูลแบบต่อเนื่อง อัลกอริทึ่ม ซี4.5 จะสร้างจุดเริ่ม (Threshold) และแยกคุณลักษณะนั้นออกเป็น 2 ส่วน คือส่วนที่มีค่ามากกว่าและ ค่าน้อยกว่าและเท่ากับ ค่าที่ใช้ในการสร้างจุดเริ่ม

ข. สามารถใช้กับข้อมูลฝึกสอน (Training Data set) ที่ไม่มีค่าของคุณสมบัติได้ โดยสามารถให้กำหนดเครื่องของคุณสมบัตินั้นเป็น "?" และไม่นำค่านั้นมาคำนวณหาค่าคาดคะเนของข้อมูลเอนโทรปี

ค. สามารถใช้งานได้กับค่าที่มีความผิดปกติหรือมีความเสียหาย

ง.สามารถทำการปรับแต่งต้นไม้ประกอบการตัดสินใจ (Pruning Trees) ในขณะสร้างได้

อัลกอริทึ่มซี 4.5 ได้มีคิดอัตราส่วนของเกน Gain Ratio ในการตัดสินใจโหนดของรากต้นไม้ดังนี้

$$GainRatio(x) = \frac{Gain(x)}{SplitInfoGain(x)} \tag{2-7}$$

$$SplitInfo(x) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} \log_i \frac{|Si|}{|S|} \tag{2-8}$$

การแบ่งช่วงของข้อมูล จะนำไปใช้ในการสร้างต้นไม้ในการตัดสินใจ หรือเรียกได้ว่าเป็นการแบ่งค่าต่อเนื่องออกเป็นช่วงย่อยๆ Discretization เช่น ข้อมูลจริงมีค่าเป็น 8 จะมีการจัดอยู่ในกลุ่มที่ 3 ซึ่งจะมีค่าของช่วงย่อย ตั้งแต่ 7 ถึง 9 โดยซี 4.5 จะถือว่าในข้อมูลชุดนั้นจะมีค่าของคุณสมบัติเป็น 3 เพื่อใช้ในการคำนวณ

ชุดข้อมูล
5
8
2
7
10
4
12
8

ข้อมูลที่ได้ผ่านการ Discretization จะถูกจัดกลุ่มออกเป็นดังนี้
1. 1-3
2. 4-6
3. 7-9
4. 10-12

ตารางที่ 2-4 ตัวอย่างการแบ่งช่วงข้อมูลของซี4.5

2.5 ขั้นตอนวิธีเคเนียเรสต์เนเบอร์ (*k*-Nearest Neighbor: *k*-NN)

คือวิธีในการจัดแบ่งคลาสซึ่งจะใช้ตัดสินใจ ว่าคลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวน ("K" ใน K-nearest neighbor) ของเงื่อนไขที่เหมือนหรือใกล้เคียงกันมากที่สุด จำนวน k ตัว จากข้อมูลชุดตัวอย่างด้วยการหาผลรวมจากการนับ (count up)

เทคนิคของ KNN เป็นการหาวิธีการวัดระยะห่างระหว่างแต่ละแอตทริบิวในข้อมูลที่จะนำมาคำนวณค่า ที่มีลักษณะคล้ายคลึงกันมากที่สุด k ตัว เพื่อจัดกลุ่มให้กับข้อมูลตัวใหม่ จากข้อมูลในตารางประกอบด้วย n คอลัมน์ และหาสมาชิกที่มีความคล้ายคลึงกัน ด้วยการวัดระยะห่าง Standard Euclidean distance จากแถว x ใดๆ ให้เขียนแทนเวกเตอร์ได้ *(x, a(x))* โดย a คือฟังก์ชั่น เมื่อแทนค่า x ในฟังก์ชั่นเขียนได้ดังนี้

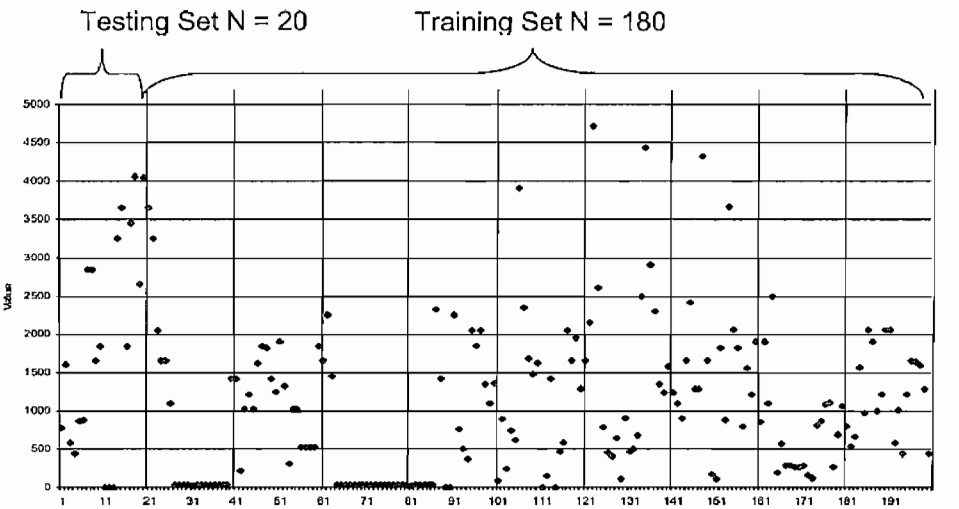
$$\left(a_1(x), a_2(x), \ldots, a_n(x)\right) \tag{2-9}$$

เมื่อ $a_1(x)$ หมายถึงค่าที่อยู่ในคอลัมน์ที่ 1 ของแถว x ระยะห่างระหว่าง 2 ตัวอย่างข้อมูล x_iกับ x_j เขียนแทนด้วย $d(x_i, x_j)$ เมื่อหาระยะห่างด้วย Standard Euclidean distance สามารถเขียนแทนด้วย

$$d(X_i, X_j) = \sqrt{\sum_{r=1}^{n} (a_r(X_i) - a_r(X_j))^2} \tag{2-10}$$

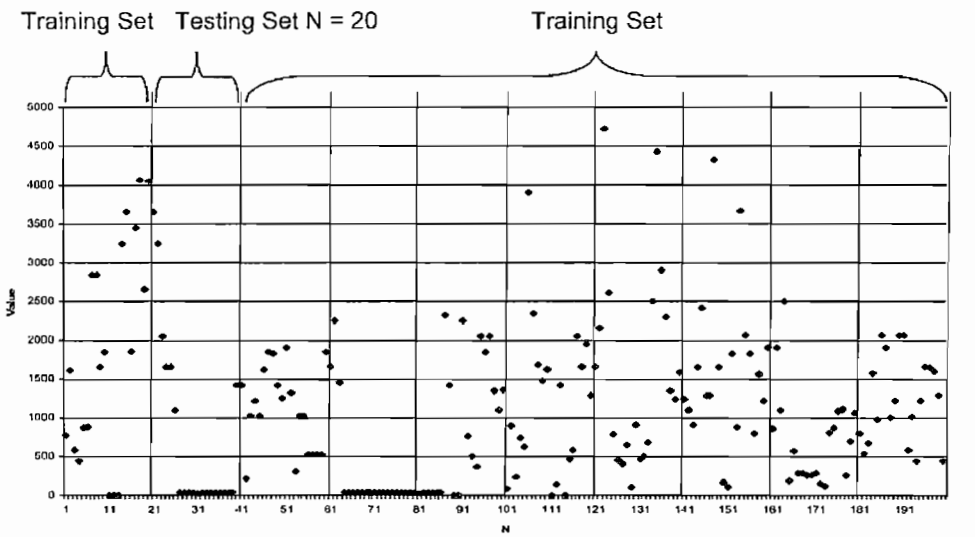
2.6 วิธีการ K-Fold Cross-Validation

เป็นวิธีการแบ่งชุดข้อมูลเรียนรู้ (Training data set) และ ชุดข้อมูลทดสอบ (Test data set) โดยการแบ่งข้อมูลออกเป็น k_i, k_i = 1,2,3,...,n ชุด แล้วให้นำชุดข้อมูล 1 ชุดมาทำเป็นชุดข้อมูลสำหรับการทดสอบ ส่วนที่เหลือ k-1 นำมาใช้สำหรับชุดข้อมูลการเรียนรู้ หลังจากนั้นให้ทำการวนรอบการทดสอบจำนวน k รอบ ซึ่งจะเปลี่ยนชุดข้อมูลทดสอบไปเรื่อยๆ จนครบ ซึ่งมักจะนิยมใช้ 10-Fold เพราะได้ผลความถูกต้อง (Accuracy) เป็นที่น่าพอใจ และมีระยะเวลาการทำงาน (Time Complexity) ไม่มาก



รูปภาพที่ 2-3 ตัวอย่างขั้นตอนของการแบ่งชุดข้อมูลเพื่อทดสอบในรอบที่ 1

จะเห็นว่าการแบ่งชุดข้อมูลแบบ k-Fold ที่ k = 10 เมื่อจำนวนชุดข้อมูลทั้งหมด N = 200 ในรอบแรกจะใช้ชุดข้อมูลจำนวน 1-20 ในการทดสอบ และให้ชุดข้อมูลอีก 180 เป็นชุดข้อมูลการเรียนรู้



รูปภาพที่ 2-4 ตัวอย่างขั้นตอนของการแบ่งชุดข้อมูลเพื่อทดสอบในรอบที่ 2

ภาพที่ 2-4 เป็นการทำงานในรอบที่ 2 โดยใช้ชุดข้อมูล 21-30 เป็นในการทดสอบ และให้ชุดข้อมูล 31-200 และ 1-20 เป็นชุดข้อมูลการเรียนรู้

2.7 การคำนวณความแม่นยำ

การวัดความแม่นยำในการจำแนก จะใช้วิธีการหาอัตราส่วนระหว่างจำนวนผลลัพธ์ที่จำแนกกับจำนวนข้อมูลทั้งหมด โดยสามารถแสดงได้ดังนี้

$$Accuracy\ Rate(\%) = \frac{Number\ of\ Correcive\ Result}{Amount\ of\ Total\ data\ set} \times 100 \qquad (2\text{-}11)$$

เพื่อประเมินถึงประสิทธิภาพของขั้นตอนวิธีในการจำแนกจดหมายข่าวขยะ จากกลุ่มข้อมูลจะแบ่งประเภทผลลัพธ์การทดลองพื้นฐานให้อยู่ในรูปของตาราง Confusion Matrix

Prediction class	Actual expectation	
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True Negative)

ตารางที่ 2-5 ประเภทผลลัพธ์การทดลองพื้นฐาน

จากตารางที่ 2-5 อธิบายผลที่ได้จากจำแนกกลุ่มผลลัพธ์พื้นฐานได้ดังนี้

- TP (True Positive) คือจำนวนผลการจำแนกในคลาสบวก (Positive) และได้ผลเป็นบวกถูกต้อง (True)
- FP (False Positive) คือจำนวนผลการจำแนกในคลาสบวก (Positive) และได้ผลเป็นลบผิดพลาด (False)
- TN (True Negative) คือจำนวนผลการจำแนกในคลาสลบ (Negative) และได้ผลเป็นลบถูกต้อง (True)
- FN (False Negative) คือจำนวนผลการจำแนกในคลาสลบ (Negative) และได้ผลเป็นบวกผิดพลาด (False)

เพื่อวัดประสิทธิภาพความแม่นยำของผลลัพธ์ สามารถนำไปหาความแม่นยำได้ต่อไปนี้

Precision คือค่าความแม่นยำ ที่จะแสดงให้เห็นว่าระบบที่ทดสอบมีความแม่นยำเพียงใด

Recall คือค่าระลึกหรือค่าความครบถ้วน ที่แสดงให้เห็นเมื่อระบบได้ทำการดึงผลลัพธ์แล้วมีความถูกต้องเพียงใด

F-Measure คืออัตราการรู้จำ หรือค่าเฉลี่ยที่ให้ความแม่นยำและความครบถ้วนเท่าๆกัน

$$Precision = \frac{TP}{TP+FP} \tag{2-12}$$

$$Recall = \frac{TP}{TP+FN} \tag{2-13}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2\text{-}14)$$

2.8 งานวิจัยที่เกี่ยวข้อง

เสนีย์ ทรัพย์บุญเลิศมา และ ศาสตรา วงศ์ธนวสุ นำเสนองานวิจัยเรื่อง การเปรียบเทียบขั้นตอนวิธีตัวกรองสแปมอีมลล์ โดยได้นำเสนอแนวคิดการเปรียบเทียบวิธีการกรองจดหมายข่าวขยะ ด้วยขั้นตอนวิธีต่างๆ ได้แก่ วิธีข่ายงานเบย์ วิธีเบย์แบบง่าย และวิธีต้นไม้ตัดสินใจ เพื่อเปรียบเทียบประสิทธิ์ภาพความแม่นยำ ซึ่งทดลองกับชุดข้อมูลตัวอย่างจดหมาย Ling Spam Corpus ทราบว่าวิธีข่ายงานเบย์ให้ประสิทธิภาพดีที่สุด โดยความแม่นยำร้อยละ 96.44 สำหรับวิธีเบย์แบบง่ายให้ค่าความแม่นยำร้อยละ 95.23 และวิธีต้นไม้ตัดสินใจให้ค่าความแม่นยำร้อยละ 94.12

ชัดชัย แก้วตา และ อัจฉรา มหาวีรวัฒน์ นำเสนอการวินิจฉัยคดีด้วยเทคนิคต้นไม้ตัดสินใจ
โดยประยุกต์ใช้เทคนิคต้นไม้ตัดสินใจ (Decision tree techniques) เพื่อจำแนกความผิดซึ่งถูกอธิบายด้วยชุดของคุณสมบัติ (Attributes) ออกมาเป็นมาตราต่างๆ ที่เหมาะสมกับคดีความนั้น และเปรียบเทียบความถูกต้องของการจำแนกข้อมูลด้วยอัลกอริทึ่มซี4.5

สิทธิโชค มุกดาสกุลภิบาล นำเสนองานวิจัย เรื่อง การวัดประสิทธิภาพของขั้นตอนวิธี ตัวจำแนก C4.5, ADTree และ Naïve Bayes ในการจำแนกข้อมูลการซุกซ่อนสิ่งเสพติดสำหรับไปรษณีย์ระหว่างประเทศ ได้กล่าวถึงการวัดผลด้านความแม่นยำ และประสิทธิภาพด้านของตัวจำแนก 3 ชนิด คือ C4.5, ADTree และเบย์ส เพื่อนำไปใช้กับข้อมูลการซุกซ่อนส่งเสพย์ติดสำหรับไปรษณีย์ระหว่างประเทศ

เฉลิมพล ณ สงขลา และ เกริก ภิรมย์โสภา นำเสนองานวิจัยเรื่อง การเพิ่มประสิทธิภาพการจัดจำแนกอีเมลล์สแปมภาษาไทยด้วยโปรแกรมตัดคำไทยคูวส์ โดยนำเสนอแนวคิดการจำแนกประเภทของสแปมอีเมลล์ภาษาไทยด้วยการเรียนรู้แบบเบย์ ซึ่งระบบได้ใช้โปรแกรมตัดคำไทยคูวส์ร่วมกับการเรียนรู้แบบเบย์ของ Spamassassin Version 3.2.5

Buddhika N. Kottahachchi และ Arjun R. Narayanswamy นำเสนองานวิจัยเรื่อง SpamWallah A Rule-Based E-mail Spam Detection System โดยนำเสนอแนวคิดวิธีการคัดกรองสแปมอีเมลล์ โดยวิธีแยกแยะจากแนวคิดพื้นฐานของมนุษย์ว่าจดหมาย ดังกล่าวมีความข้องเกี่ยวกันหรือไม่ จากฐานกฎความรู้ทั้งหมด 103 กฎ โดยจะมีการให้คะแนนในแต่ละกฎ แล้วจึงคำนวณเป็นความเกี่ยวข้อง หรือไม่เกี่ยวข้องของจดหมายฉบับนั้นๆ ออกมา

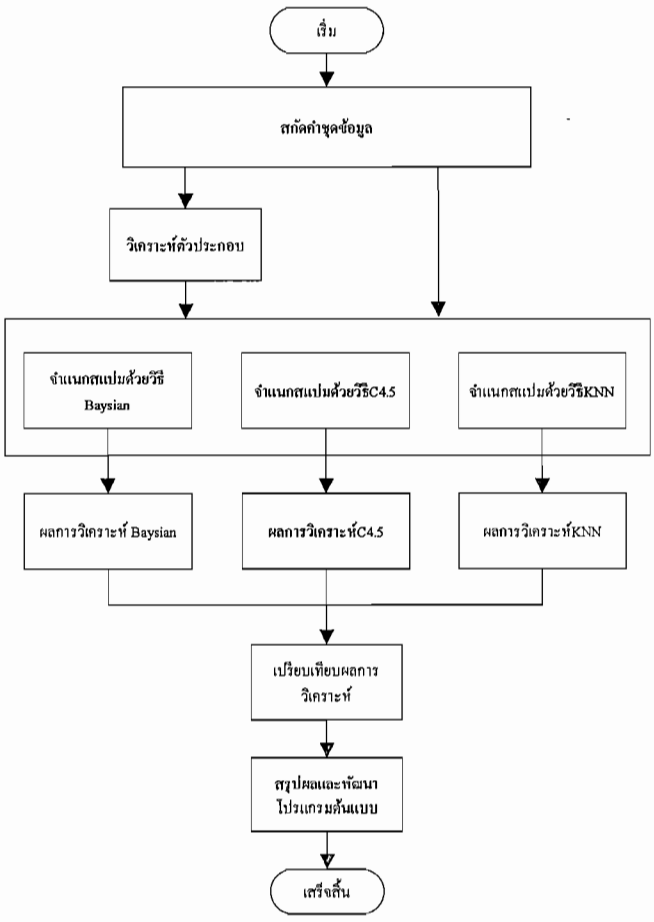
หทัยชนก กรชี นำสเนองานวิจัยเรื่อง การคำนวณน้ำหนักของคุณลักษณะข้อมูลสำหรับตัวจำแนกนาอีฟเบยส์ โดยวิธีการจัดกลุ่มข้อมูลและต้นไม้ตัดสินใจ ซึ่งได้แบ่งกลุ่มด้วยอัลกอริทึ่ม

เคมีน (K-Means) และอัลกอริทึ่มสองขั้นตอนคือ คลัสเตอร์ฟีเจอร์ทรี (Cluster Feature Tree) กับวิธีการลำดับชั้นแบบ Agglomerative พร้อมทั้งเปรียบเทียบประสิทธิภาพความแม่นยำของตัวจำแนกนาอีเบยส์ (Naïve Bayes)

บทที่ 3
วิธีการดำเนินงาน

3.1 ภาพรวมของการดำเนินงาน

ในบทนี้เป็นการนำเสนอวิธีการดำเนินงาน ในการวิเคราะห์สแปมอีเมลล์ จะสามารถแสดงเป็นแผนภาพได้ดังนี้



ภาพที่ 3-1 ภาพรวมการดำเนินงาน

3.2 การสกัดคำชุดข้อมูล

ในขั้นตอนแรกของวิธีดำเนินการทดลอง จะทำการสกัดคำของชุดข้อมูล ก่อนที่จะนำไปทำการทดสอบซึ่งมี 2 ขั้นตอนดังนี้

3.2.1 ขั้นตอนการเตรียมชุดข้อมูล โดยนำข้อมูลสแปมอีเมลล์ที่ใช้มาจาก Ling Spam Corpus จะมีอีเมลล์ทั้งหมดเป็นจำนวน 2,893 ฉบับ ประกอบด้วย อีเมลล์ปกติ ในที่นี้ขอเรียกว่าแฮมอีเมลล์ (HAM) 2,412 ฉบับ และสแปมอีเมลล์อีก 481 ฉบับ คิดเป็นสแปมอีเมลล์จากอีเมลล์ปกติร้อยละ 16.63 เนื้อหาของอีเมลล์จะผ่านกระบวนการ stemming และ stopping มาเรียบร้อยแล้ว สำหรับ tag HTML และสิ่งที่แนบมาด้วย จะไม่นำมาร่วมทดสอบ

การแบ่งชุดข้อมูลทดสอบได้แบ่งชุดสำหรับการเรียนรู้ Training set และ การทดสอบ Testing set ด้วยหลักของ K-Fold cross-validation จะทำการแบ่งข้อมูลออกเป็น 10 ส่วน ในการทดลองทั้งหมด 10 ครั้ง โดย

ครั้งแรก แบ่งชุดข้อมูลส่วนแรกเป็นชุดทดสอบ และแบ่งชุดข้อมูลส่วนที่ 2, 3, 4, ..., 10 เป็นชุดเรียนรู้

ครั้งที่สอง แบ่งชุดข้อมูลส่วนที่สองเป็นชุดทสอบ และแบ่งชุดข้อมูลส่วนที่ 1, 3, 4, ..., 10 เป็นชุดเรียนรู้

ครั้งที่สอง แบ่งชุดข้อมูลส่วนที่สามเป็นชุดทสอบ และแบ่งชุดข้อมูลส่วนที่ 1, 2, 4, ..., 10 เป็นชุดเรียนรู้

ทำต่อไปเรื่อยๆ จนกระทั้งครั้งที่ 10 แบ่งชุดข้อมูลส่วนที่สิบเป็นชุดทสอบ และแบ่งชุดข้อมูลส่วนที่ 1, 2, 3, ..., 9 เป็นชุดเรียนรู้

นำผลของค่าการหาจำนวนของสแปมอีเมลล์มาหาค่าเฉลี่ย ขอความถูกต้อง

3.2.2 ขั้นตอนนับจำนวนคำที่จะนำไปใช้คำนวณ ในการนับจำนวนของคำของอีเมลล์ โดยจะจัดอันดับของคำที่มีจำนวนมากที่สุด 1000 อันดับแรกที่จะนำมาใช้งาน โดยยกเว้นเครื่องหมายสัญลักษณ์ได้แก่ ‘ ’ (เว้นวรรค), ‘#’ (เครื่องหมายแฮช หรือ เครื่องหมายชาร์พ), ‘@’ (เครื่องหมายแอซซาย), ‘;’(เครื่องหมายเซมิโคล่อน), ‘/’ (เครื่องหมายสแลช), ‘\’ (เครื่องหมายแบ็คสแลช), ‘*’ (เครื่องหมายดอกจัน), ‘~’ (เครื่องหมายเกรพ), ‘?’ (เครื่องหมายคำถาม)

จดหมายลำดับที่ (m=2,893)	จำนวนความถี่ของคำที่พบจน ถึงคำที่ n (n=1000)							เหตุการณ์ (SPAM or HAM)
	free	price	cash	language	spent	...	คำที่ n	
1	0	2	0	0	0	...	X_{1n}	HAM
2	0	3	1	3	0	...	X_{2n}	HAM
3	0	3	0	5	1	...	X_{3n}	HAM
4	3	0	6	0	0	...	X_{4n}	SPAM
5	4	1	1	0	2	...	X_{5n}	SPAM
...	...	...	...	...	...	...	...	...
m	5	2	0	0	6	...	X_{mn}	SPAM

ตารางที่ 3-1 ข้อมูลของจำนวนคำจากชุดจดหมาย

3.3 ขั้นตอนการลดจำนวนมิติข้อมูลโดยวิเคราะห์ตัวประกอบหลัก (Principal Component Analysis : PCA)

จากตารางข้อมูลจำนวนคำจากอีเมลล์ทั้งหมด 1,000 คำแรกจากค่าสูงสุด จากจำนวนอีเมลล์ทั้งสิ้น 2,983 ฉบับ มีเมทริกซ์ขนาด 2,983 x 1,000 จำเป็นต้องทำการวิเคราะห์ตัวประกอบหลักเพื่อลดจำนวนตัวแปร

3.3.1 ขั้นตอนการหาเมทริกซ์ค่าแปรปรวนร่วม โดยนำเมทริกซ์มาหาค่าความแปรปรวนร่วม จากสมการ

$$C = \frac{1}{M}\sum_{1}^{M}\sigma_i\sigma_i^r \qquad (3\text{-}1)$$

$$= AA^T$$

โดยที่ A คือ $[\sigma_1, \sigma_2, \ldots, \sigma_M]$ เป็นชุดข้อมูลค่าเบี่ยงเบนมาตรฐาน

C คือ เมทริกซ์ความแปรปรวนร่วม

A^T คือ กลุ่มของเมทริกซ์ความแปรปรวนร่วม

3.32 ขั้นตอนการคำนวนหา ไอเกนเวคเตอร์ (v) และค่าไอเกน (μ) ของเมทริกซ์ความแปรปรวนร่วม

$$A^T A v_i = \mu_i v_i \qquad (3\text{-}2)$$

โดยที่ A คือ กลุ่มของเมทริกซ์ความแปรปรวนร่วม

μ_i คือ ค่าไอเกนหรือ Eigen value

v_i คือ ไอเกนเวคเตอร์หรือ Eigen vector

3.3.3 การพิจารณาจำนวนตัวประกอบหลัก ได้นำแนวทางการพิจารณา 2 แนวทาง ดังนี้คือ

- พิจารณาจากร้อยละความแปรปรวนสะสม ของตัวประกอบหลัก m ตัวแรก เป็นอย่างต่ำร้อยละ 80

- พิจารณาจากกราฟที่พล็อตจากค่าไอเก็น โดยตัวประกอบหลักตัวที่ m+1 มีค่าไอเก็นต่ำมากเมื่อเทียบกับตัวที่ m

- พิจารณาจากค่าเฉลี่ยความแปรปรวน โดยเลือกตัวประกอบหลักตัวที่ มีค่ามากกว่าค่าเฉลี่ยแปรปรวน

3.4 การจำแนกจดหมายข่าวขยะด้วยขั้นตอนวิธีเบย์เซียน

การจัดอันดับของกลุ่มคำที่เป็นดังกล่าว ด้วยการทำวิเคราะห์ตัวประกอบ สามารถลดจำนวนของตัวแปร (คำ) ให้เป็นตัวแปรที่มีความสัมพันธ์ จากนั้นจึงนำค่าของตัวแปรใหม่ที่ได้ไปคำนวณ จำนวนของคำที่เหมาะสมจะนำมาคำนวณ SPAM อีเมลล์ การคำนวณดังกล่าวแบ่งตามรอบ K-Fold

จากพื้นฐานของคำที่นำไปใช้

$$P(Spam|token) = \frac{P(token|Spam)*P(Spam)}{P(token)} \qquad (3\text{-}3)$$

เปลี่ยนเป็น

$$P(Spam|PC_m) = \frac{P(PC_m|Spam)*P(Spam)}{P(PC_m)} \qquad (3\text{-}4)$$

จากวิธีการคำนวณแบบเบส์โดยใช้ ปัจจัยที่ได้จากการทำ Factor Analysis โดยจะรวบรวมผลลัพธ์ที่ได้ไปเปรียบเทียบความถูกต้องแม่นยำ (Accuracy) ที่คำนวนอีกครั้ง

3.5 การจำแนกจดหมายข่าวขยะด้วยวิธีต้นไม้ตัดสินใจซี4.5 (C4.5)

การเรียนรู้ต้นไม้ตัดสินใจซี4.5 ได้นำขั้นตอนวิธีของ อัลกอริทึ่มไอดี3 (ID3) มาใช้ในการเรียนรู้ต้นไม้ตัดสินใจจากข้อมูลที่ได้จากการวิเคราะห์ตัวประกอบหลัก โดยทำการสร้างแบบจำลองต้นไม้ด้วยวิธีการ ID3 ซึ่งแอททริบิวต์ เป็นค่าจากตัวแปรของการวิเคราะห์องค์ประกอบหลักดังนี้

Function ID3 (Examples, Target_Attribute, Attributes)

- Create a root node for the tree;
- If all Examples are positive, Return the single-node tree Root, with label = + ;
- If all Examples are negative, Return the single-node tree Root, with label = - ;
- If number of predicting Attributes is empty, then Return the single node tree Root, with label = most common value of the target Attribute in the Examples;
- Otherwise Begin
    - A = the Attribute that classifies examples;
    - Decision Tree attribute for Root = A;
    - For each possible value, v_i, of A;
        - Add a new tree branch below Root, corresponding to the test A = v_i ;
        - Let Examples(v_i), be the subset of examples that have the value v_i for A;
        - If Examples(v_i) is empty;
            - Then below this new branch add a leaf node with label = most common target value in the examples;
        - Else below this new branch add the subtree ID3 (Examples(v_i), Target_attribute, Attributes – {A});
- End;
- Return Root;

3.6 การจำแนกจดหมายข่าวขยะด้วยวิธีเคเนียเรสต์เนเบอร์ *k*-Nearest Neighbor (*k*-NN)

จะใช้วิธีหาจุดที่ใกล้เคียงที่ได้เท่ากับจำนวน k (*k*-NN) ที่ต้องการเช่น k=5 ก็จะขยายขอบเขตของข้อมูลไปจนกว่าจะพบ จากนั้นจะดูว่าข้อมูลที่พิจารณามีความน่าจะเป็นว่าอยู่ใกล้กับ

ข้อมูลจุดไหนมากสุด จากนั้นก็ทำการจำแนกกลุ่มของข้อมูล ซึ่งในที่นี้เราจะนำมาใช้ในการพิจารณาจดหมายว่าเป็นสแปมอีเมลล์หรือไม่ โดยมีขั้นตอนดังนี้

3.6.1 วัดระยะห่างของข้อมูลโดยวิธีการวัดระยะห่างยูคลิเดียน (Euclidean Distance) ดังสมการที่ 2.8 ในบทที่ผ่านมา

3.6.2 เมื่อทำการวัดระยะทางระหว่างข้อมูลที่ทำการทดสอบกับข้อมูลที่เป็นชุดตัวอย่างเป็นที่ เรียบร้อยแล้ว จากนั้นจึงเรียงลำดับข้อมูลของระยะห่างจากน้อยไปหามาก เพื่อใช้เป็นข้อมูลในการพิจารณาของสแปมอีเมลล์

3.6.3 ขยายขอบเขตของข้อมูลตามจำนวน k-NN มาพิจารณา (จากข้อมูลที่เรียงลำดับน้อยไปมาก) เพื่อหา Distance ที่น้อยที่สุด จากจำนวน k ที่กำหนด เพื่อให้ Class Label จากการ Vote ของ Majority Vote

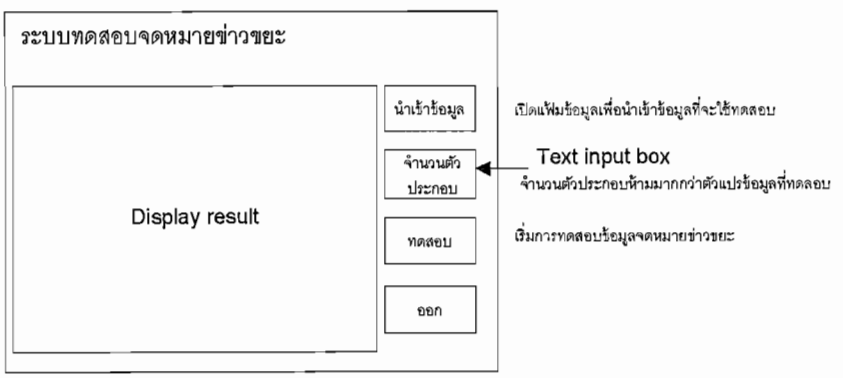
3.7 ผลลัพธ์ความถูกต้อง

เมื่อได้ผลลัพธ์จากการพิจารณาจดหมายข่าวขยะเรียบร้อยแล้ว สามารถเปรียบเทียบความถูกต้องของการพิจารณาเป็นร้อยละดังนี้

$$\text{Corrective Accuracy } (c) = \frac{\text{Corrective Testing Data set}}{\text{Amount Data set}} \times 100 \qquad (3\text{-}5)$$

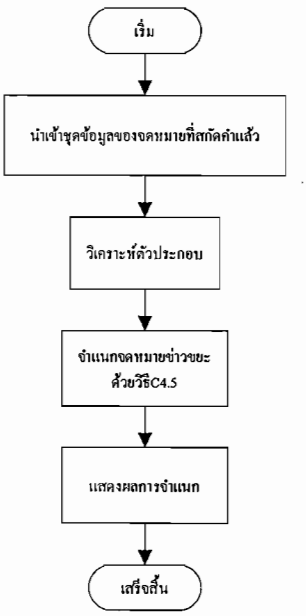
3.8 ออกแบบต้นแบบโปรแกรมวิเคราะห์จดหมายข่าวขยะ

ผลลัพธ์จากการพิจารณาจดหมายข่าวขยะ จะนำไปทดลองสร้างโปรแกรมต้นแบบสำหรับนำไปจำแนกจดหมายข่าวขยะ โดยออกแบบหน้าจอการทำงานดังนี้



ภาพที่ 3-2 หน้าจอต้นแบบโปรแกรม

ขั้นตอนการทำงานของต้นแบบโปรแกรมวิเคราะห์จดหมายข่าวขยะนั้น มีดังนี้



ภาพที่ 3-3 ออกแบบขั้นตอนการทำงานโปรแกรม

การทำงานของต้นแบบโปรแกรมวิเคราะห์จดหมายข่าวขยะมีดังนี้

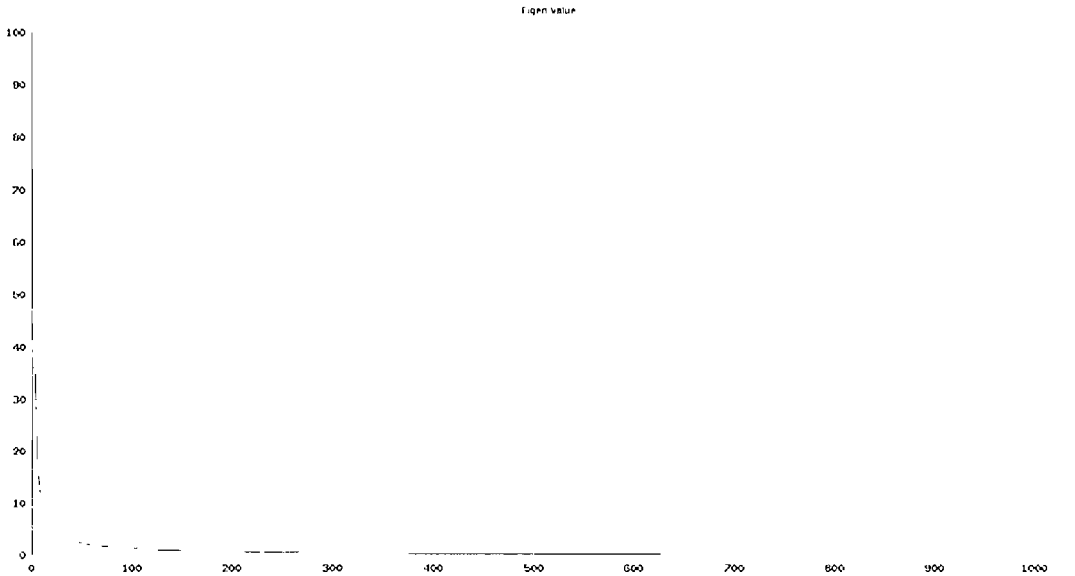
- นำเข้าไฟล์ข้อมูลที่มีการสกัดคำจำนวน 1000 คำเรียบร้อยแล้ว
- คำนวนค่าไอเกนและเวคเตอร์ไอเกน เพื่อวิเคราะห์ตัวประกอบหลัก
- แบ่งชุดข้อมูลสำหรับนำไปทดสอบ
- ใส่ค่าจำนวนตัวประกอบหลักที่จะนำไปวิเคราะห์
- วิเคราะห์จดหมายข่าวขยะด้วยวิธีซี 4.5

บทที่ 4
ผลการดำเนินงาน

จากการวิเคราะห์และออกแบบขั้นตอนการดำเนินงานดังที่ได้กล่าวในบทที่ 3 ในบทนี้จะกล่าวถึงผลการดำเนินงาน จากชุดข้อมูลจดหมายข่าวขยะ (สแปมอีเมลล์) Ling Spam Corpus มีจำนวนคำทั้งสิ้น 59,829 คำ โดยได้ทำการตัดคำที่นำมาใช้ทดลองจำนวน 1,000 คำ ด้วยการนับความถี่ของคำ (Term frequency) สูงสุดตามลำดับ ที่ปรากฏในอีเมลล์แต่ละฉบับ จากนั้นจึงได้ทำการลดตัวแปรจากขั้นตอนวิเคราะห์ตัวประกอบหลัก และใช้วิธีการจำแนกสแปมด้วยวิธีการคำนวณ โดยได้ผลการทดลองดังนี้

4.1 การพิจารณาลดตัวแปรด้วยวิธีวิเคราะห์ตัวประกอบหลัก

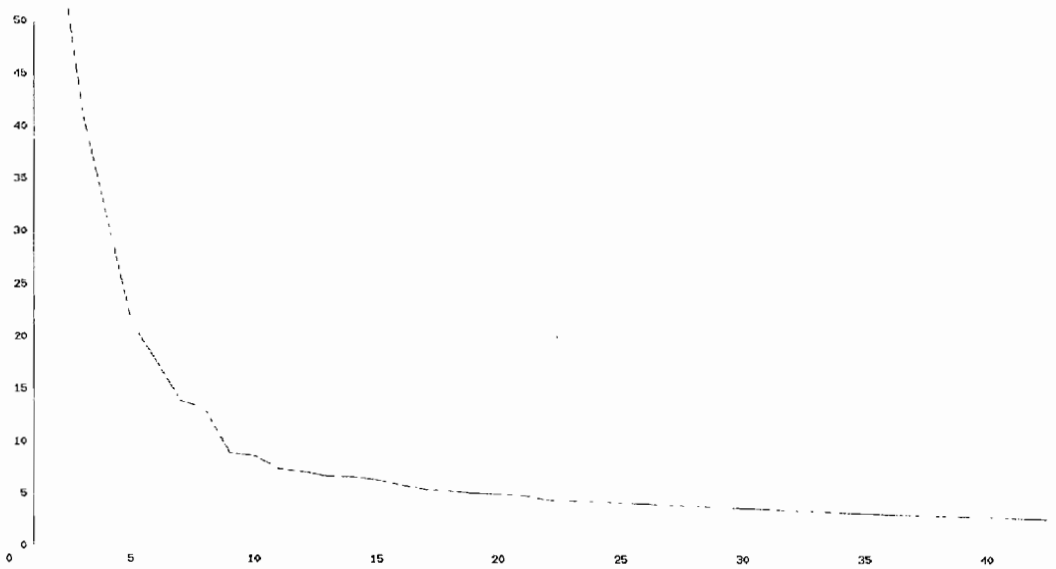
จากขั้นตอนการวิเคราะห์ตัวประกอบหลักสามารถแสดงค่าไอเกนได้ดังภาพ



รูปที่ 4-1 ค่าไอเกนจากชุดข้อมูลจำนวน 1000 ตัวประกอบหลัก

4.1.1 พิจารณาจำนวนตัวประกอบหลักสำหรับที่จะนำไปวิเคราะห์ต่อไป โดยใช้กราฟพบว่า ชุดข้อมูลของตัวประกอบหลักที่ควรนำมาพิจารณาคือจำนวน 100 ตัวประกอบหลักแรก และ

ไม่ควรเกิน 200 ตัวประกอบ ดังนั้นจึงทำการขยายภาพจากการพล๊อตกราฟที่ 50 ตัวประกอบหลักแรก ได้ดังภาพ



รูปที่ 4-2 ค่าไอเกนจากการขยาย

พบว่าตัวประกอบหลักที่ 9 และ 10 มีค่า แตกต่างกับตัวประกอบที่ 8 ดังนั้นตัวประกอบหลักที่ 1 - 8 ไป จึงควรนำไปทำการวิเคราะห์ต่อ

ลำดับที่	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
ค่าไอเกน	90.39579	58.44924	40.9132	30.97092	21.20991	17.63984	13.80334	12.9932

ตารางที่ 4-1 ค่าไอเกนจำนวน 8 ตัว

4.1.2 พิจารณาจำนวนตัวประกอบหลักสำหรับที่จะนำไปวิเคราะห์ต่อไป โดยค่าร้อยละความแปรปรวนสะสมที่มีค่ามากกว่าร้อยละ 80 พบว่ามีจำนวนตัวประกอบหลัก ทั้งหมดจำนวน 140 ตัวแรกที่มีค่ามากกว่า ร้อยละ 80 ที่จะนำไปทำการวิเคราะห์ต่อ

4.1.3 พิจารณาจำนวนตัวประกอบหลักสำหรับที่จะนำไปวิเคราะห์ต่อไป โดยตัวประกอบที่มีความแปรปรวน มากกว่าค่าเฉลี่ยของค่าความแปรปรวนทั้งหมด ซึ่งค่าเฉลี่ยของค่าความ

แปรปรวนจากจำนวนตัวประกอบหลัก 1000 ตัว ทั้งหมดมีค่าเท่ากับ 0.719431 ดังนั้นจำนวนตัวประกอบที่มีค่าความแปรปรวนมากกว่าค่าเฉลี่ยจะมีทั้งหมด 158 ตัวแรก ที่จะนำไปวิเคราะห์ต่อ

4.2 ผลการวิเคราะห์จดหมายข่าวขยะจากการวิเคราะห์ตัวประกอบหลัก

จากผลการดำเนินงานวิเคราะห์ตัวประกอบหลักที่ผ่านมา ได้ทำการทดลองวิเคราะห์จดหมายข่าวขยะ ด้วยโปรแกรม WEKA มีค่าพารามิเตอร์ต่างๆ เช่น วิธีซี4.5 มีค่าความเชื่อมั่น (Confidence Factor) เท่ากับ 0.25 ซึ่งเป็นค่าโดยปริยาย วิธีเค-เอ็นเอ็น มีการกำหนดค่าคือ 1, 3, 5, 7, 9 และ 11 ซึ่งวิธีเบย์เซียนไม่มีการกำหนดค่าพารามิเตอร์

ในครั้งแรกใช้จำนวนตัวประกอบหลักที่ 1-8 จาก ซึ่งพิจารณาจำนวนตัวประกอบหลักโดยใช้กราฟจากรูปที่ 4-2 ได้ผลดังนี้

	Bayes	C4.5	KNN-1	KNN-3	KNN-5	KNN-7	KNN-9	KNN-11
Accuracy Rate (%)	87.3833	97.7186	97.7532	97.9884	98.168	97.8223	97.6288	97.5907

ตารางที่ 4-2 ค่าความถูกต้องของการพิจารณจดหมายข่าวขยะจากตัวประกอบหลักที่ 1-8

ผลทดลองพิจารณาวิเคราะห์จดหมายข่าวขยะ ครั้งที่สอง ด้วยจำนวนตัวประกอบหลัก 140 ตัว ซึ่งพิจารณจำนวนตัวประกอบหลัก จากร้อยละความแปรปรวนสะสมมากกว่า ร้อยละ 80 ได้ผลดังนี้

	Bayes	C4.5	KNN-1	KNN-3	KNN-5	KNN-7	KNN-9	KNN-11
Accuracy Rate (%)	88.4895	97.7878	94.0892	94.6079	94.2966	94.1583	94.1725	94.0618

ตารางที่ 4-3 ค่าความถูกต้องของการพิจารณสแปมจากตัวประกอบหลักที่ 1-140

ผลทดลองพิจารณาวิเคราะห์จดหมายข่าวขยะ ครั้งที่สาม ด้วยจำนวนตัวประกอบหลัก 158 ตัว ซึ่งพิจารณจำนวนตัวประกอบหลัก จากค่าเฉลี่ยความแปรปรวนมากกว่า 0.7194 ได้ผลดังนี้

	Bayes	C4.5	KNN-1	KNN-3	KNN-5	KNN-7	KNN-9	KNN-11
Accuracy Rate (%)	83.8340	97.8259	94.6252	94.2968	94.4596	94.2313	94.2105	93.9304

ตารางที่ 4-4 ค่าความถูกต้องของการพิจารณสแปมจากตัวประกอบหลักที่ 1-158

จะเห็นได้ว่า จากการดำเนินงานในครั้งแรกค่าความถูกต้องของการพิจารณาสแปมด้วยวิธี KNN-5 มีค่าสูงที่สุด รองลงมาคือ C4.5 และ Bayes ต่ำที่สุด ในครั้งที่สอง ค่าความถูกต้องของการพิจารณาสแปมด้วยวิธี C4.5 มีค่าสูงที่สุด รองลงมาคือ KNN-3 และ Bayes ต่ำที่สุด และในครั้งที่สาม ค่าความถูกต้องของการพิจารณาสแปมด้วยวิธี C4.5 มีค่าสูงที่สุด รองลงมาคือ KNN-1 และ Bayes ต่ำที่สุด

ดังนั้นจึงได้กำหนดขอบเขตของการทดลองใหม่จากช่วงของค่าไอเกนดังนี้

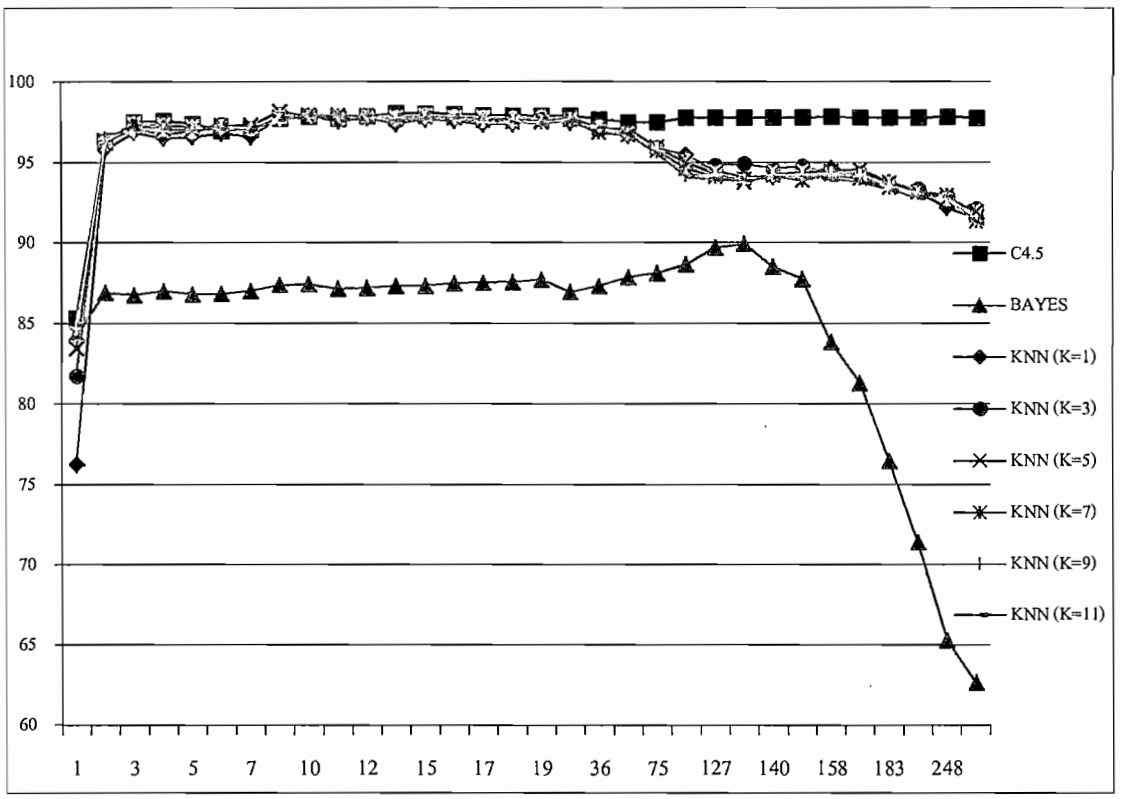
ลำดับที่	ค่าไอเกน	จำนวนมิติ (ตัวประกอบหลัก)
1	>=90.0	1
2	>=50.0	2
3	>=40.0	3
4	>=30.0	4
5	>=21.0	5
6	>=17.0	6
7	>=13.0	7
8	>=12.0	8
9	>=8.0	10
10	>=7.3	11
11	>=7.0	12
12	>=6.5	14
13	>=6.0	15
14	>=5.7	16
15	>=5.3	17
16	>=5.2	18

ลำดับที่	ค่าไอเกน	จำนวนมิติ (ตัวประกอบหลัก)
17	>=5.0	19
18	>=4.0	26
19	>=3.0	36
20	>=2.0	56
21	>=1.5	75
22	>=1.0	114
23	>=0.9	127
24	>=0.8456	135
25	>=0.81	140
26	>=0.8	142
27	>=0.7194	158
28	>=0.7	161
29	>=0.6	183
30	>=0.5	210
31	>=0.4	248
32	>=0.3	302

ตารางที่ 4-5 ค่าไอเกนจำนวน 32 ช่วง

จากตารางที่ 4-5 ได้มีการกำหนดขอบเขตโดยใช้ค่าไอเกนเป็นเกณฑ์อยู่ที่ 302 ตัวแปร ซึ่งเป็น 3 ใน 10 ของจำนวนตัวแปรทั้งหมด เพื่อแสดงให้เห็นประสิทธิภาพของการลดจำนวนตัวแปรอย่างชัดเจน

จากการพิจารณาค่าไอเกนและจำนวนตัวแปรในแต่ลำดับ จึงได้พิจารณาเลือกแนวโน้มที่สัมพันธ์กันระหว่างค่าไอเกนและตัวแปรได้ 32 ช่วงลำดับ

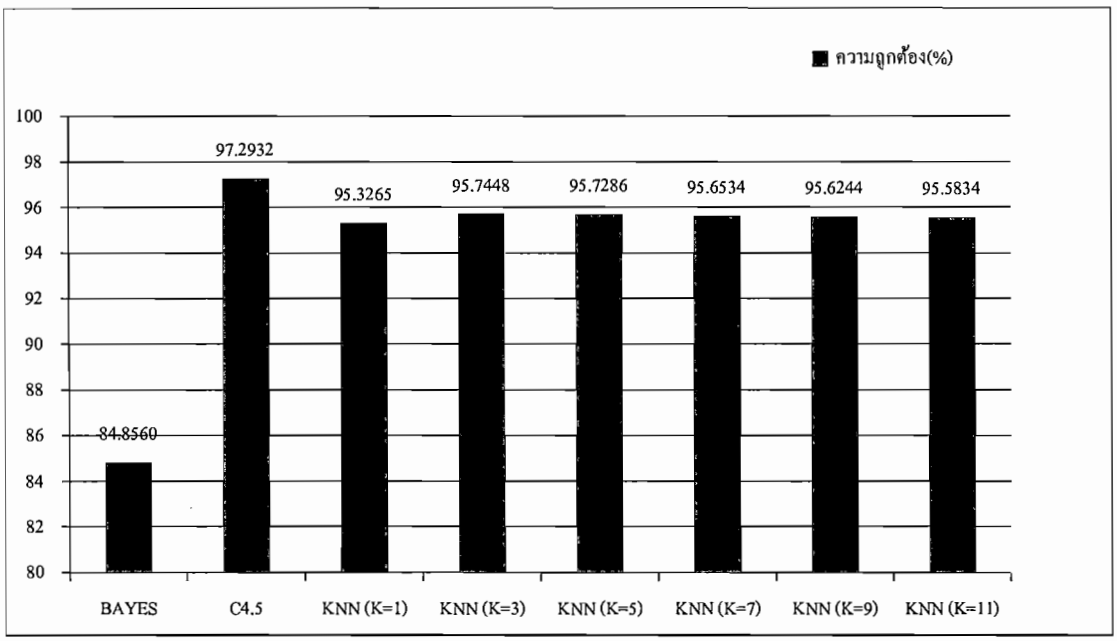


รูปที่ 4-3 ค่าความถูกต้อง (%) ต่อจำนวนตัวประกอบหลัก

จากผลการทดลองจะเห็นแนวโน้มของความถูกต้อง ของการดำเนินแต่ละวิธีค่อยๆ เพิ่มขึ้นจนกระทั่งถึงจุดหนึ่ง จึงมีแนวโน้มค่อยๆ ลดลงตามจำนวนตัวประกอบหลักที่เพิ่มขึ้น โดยการคำนวณเบย์จะมีค่าความถูกต้องสูงสุด 89.9412% ที่จำนวนตัวประกอบหลัก 135 ตัว หรือมีค่าไอเกนมากกว่าหรือเท่ากับ 0.8456 แล้วจากนั้นค่าความถูกต้องจะลดลงอย่างเห็นได้ชัด ซึ่งจะใกล้เคียงกับขั้นตอนวิธีเค-เอ็นเอ็นที่ k=5 จะมีค่าความถูกต้องสูงสุด 98.168% ที่จำนวนตัวประกอบหลัก 8 ตัว หรือมีค่าไอเกนมากกว่าหรือเท่ากับ 12.0

สำหรับขั้นตอนการดำเนินเงานซี 4.5 แนวโน้มความถูกต้องจะค่อยๆเพิ่มขึ้น และมีแนวโน้มและคงที่เมื่อจำนวนตัวประกอบเพิ่มขึ้น โดยจะมีค่าความถูกต้องสูงสุด 98.0643% ที่จำนวนตัวประกอบหลัก 14 ตัวหรือมีค่าไอเกนมากกว่าหรือเท่ากับ 6.5 ซึ่งจะแตกต่างจาก ขั้นตอนวิธีการคำนวณเบย์ และขั้นตอนวิธีเค-เอ็นเอ็น ที่แนวโน้มความถูกต้องลดลงเมื่อมีจำนวนตัวประกอบหลักมากขึ้น

ดังนั้นมาหาค่าเฉลี่ยความถูกต้องของตัวประกอบหลักทั้งหมด ในการพิจารณาจดหมายข่าวขยะในแต่ละวิธี สามารถแสดงได้ดังนี้



รูปที่ 4-4 ค่าเฉลี่ยความถูกต้อง (%) ในแต่ละวิธี

จากรูปที่ 4-4 ค่าเฉลี่ยความถูกต้อง จากการทดลองทั้งหมดพบว่าขั้นตอนวิธีซี 4.5 มีค่าสูงที่สุดตามมาด้วยเค-เอ็นเอ็น ซึ่งมีค่าเฉลี่ยความถูกต้องใกล้เคียงกันแต่เค สุดท้ายคือขั้นตอนวิธีเบย์เซียนมีค่าเฉลี่ยความถูกต้องน้อยที่สุด

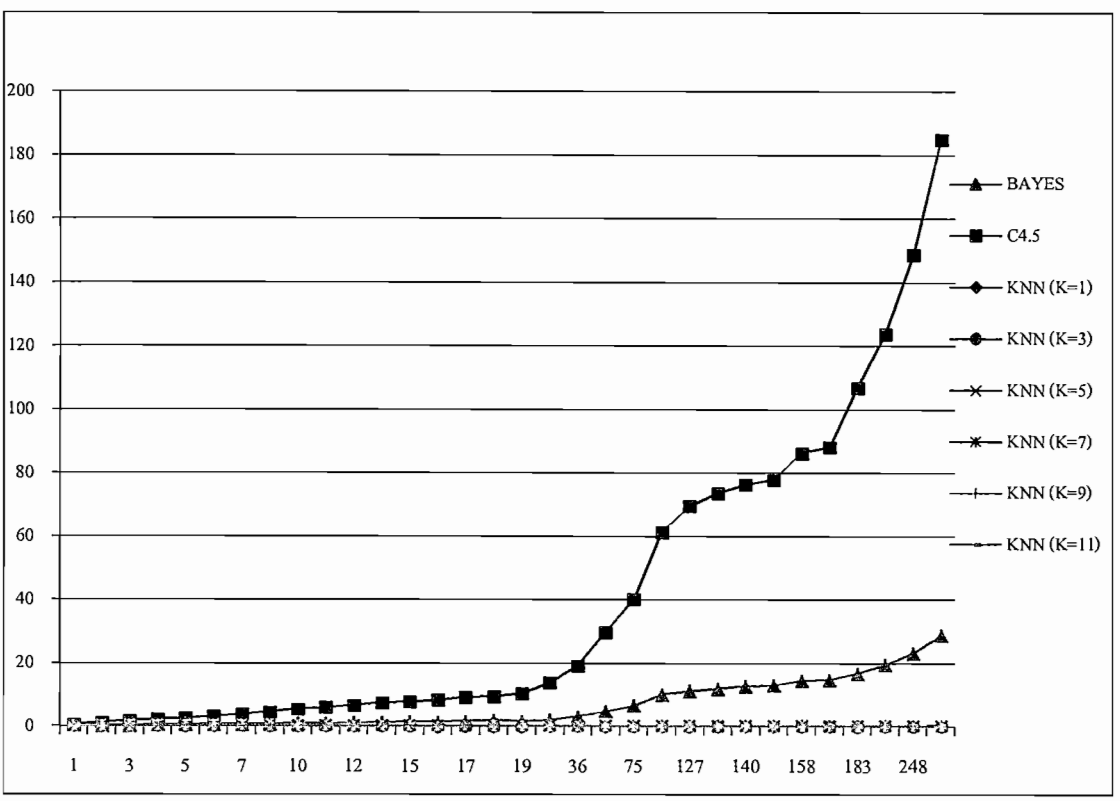
ในการวัดค่าความแม่นยำสามารถ วิเคราะห์ได้จากผลของค่าเฉลี่ยของ Precision และค่า Recall และ F-Measure ได้ดังนี้

	Bayes	C4.5	KNN-1	KNN-3	KNN-5	KNN-7	KNN-9	KNN-11
Precision	0.8918	0.9808	0.9770	0.9768	0.9752	0.9736	0.9720	0.9707
Recall	0.9382	0.9877	0.9669	0.9725	0.9743	0.9752	0.9763	0.9773
F-Measure	0.9069	0.9840	0.9719	0.9745	0.9746	0.9742	0.9740	0.9738

ตารางที่ 4-6 ค่าเฉลี่ยการวัดความแม่นยำ

จากค่าเฉลี่ยในการวัดความแม่นยำในแต่ละวิธีเห็นว่า ค่า Precision จะอยู่ในระดับสูง ดังนั้นภาพรวมของความแม่นยำอยู่ในระดับสูง ขั้นตอนวิธีซี 4.5 มีค่าสูงสุด 0.9808 ตามมาด้วยขั้นตอนวิธีเค-เอ็นเอ็น 0.9770 (k=1) และขั้นตอนวิธีของเบย์ มีค่าต่ำสุด ซึ่งตีความหมายโดยรวมได้ว่า ความแม่นยำของวิธีซี 4.5 มีความน่าเชื่อในการจำแนกจดหมายข่าวสูงสุด

ในส่วนของการดำเนินงาน สามารถแสดงเวลาที่ใช้ในการดำเนินงานวิเคราะห์จดหมายข่าวขยะในแต่ละวิธี โดยแบ่งเป็น เวลาที่ใช้ในการเรียนรู้จดหมายข่าวขยะและเวลาที่ใช้ในการทดสอบจดหมายข่าวขยะของแต่ละวิธี ดังนี้

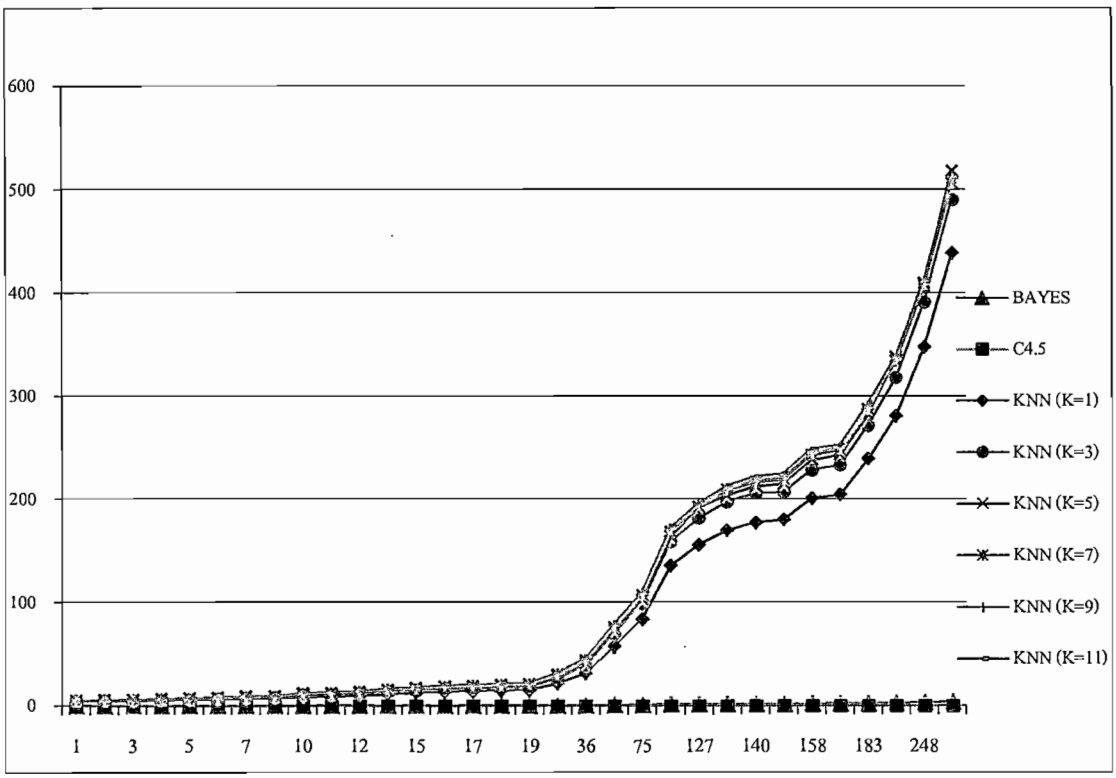


รูปที่ 4-5 เวลาที่ใช้ในการเรียนรู้จดหมายข่าวขยะในแต่ละวิธี (วินาที)

จากรูปที่ 4-5 ผลการใช้เวลาในการเรียนรู้จดหมายข่าวขยะ จะเห็นได้ว่าเวลาที่ใช้ในการเรียนรู้มากที่สุดคือขั้นตอนวิธีซี 4.5 โดยเวลาที่ใช้ในการเรียนรู้จะขึ้นอยู่กับจำนวนของตัวประกอบหลักที่นำมาเรียนรู้ หากจำนวนตัวประกอบหลักที่นำมาเรียนรู้มาก ก็จะใช้เวลาในการเรียนรู้นานขึ้น ซึ่งแนวโน้มการใช้เวลาจะเพิ่มขึ้นอย่าเห็นได้ชัดคือการเรียนรู้ตั้งแต่ตัวประกอบหลักจำนวน 26 ตัว

ขึ้นไป รองลงมาคือขั้นตอนวิธีเบย์เซียน ซึ่งแนวโน้มการใช้เวลาจะเพิ่มขึ้นตั้งแต่ตัวประกอบหลักจำนวน 36 ตัวขึ้น และสุดท้ายคือขั้นตอนวิธีเค-เอ็นเอ็น ที่ใช้เวลาการเรียนรู้น้อยมาก

จากผลการดำเนินงาน สามารถแสดงเวลาที่ใช้ในการพิจารณาจดหมายข่าวขยะของแต่วิธีแสดงได้ดังนี้

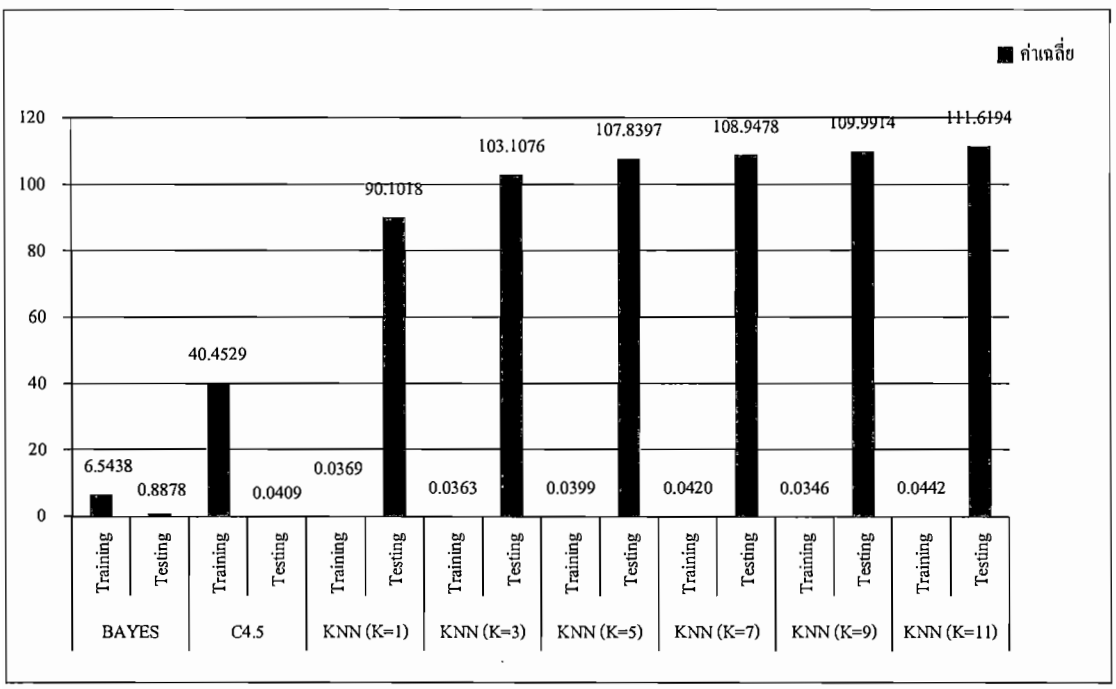


รูปที่ 4-6 เวลาที่ใช้ในการพิจารณาจดหมายข่าวขยะในแต่ละวิธี (วินาที)

จากรูปที่ 4-6 ผลการใช้เวลาในการพิจารณาจดหมายข่าวขยะ จะเห็นได้ว่าเวลาที่ใช้ในการพิจารณามากที่สุด คือขั้นตอนวิธีเค-เอ็นเอ็นโดยเวลาที่ใช้ในการพิจารณาจะขึ้นอยู่จำนวนของตัวประกอบหลักที่ใช้ในการวิเคราะห์ หากจำนวนตัวประกอบหลักที่ใช้ในการวิเคราะห์มาก โดยแนวโน้มที่ใช้เวลานานอย่างเห็นได้ชัดคือตั้งแต่ตัวประกอบหลักที่ 36 ขึ้นไป

รองลงมาคือขั้นตอนวิธีเบย์เซียน ซึ่งใช้เวลาในการพิจารณาน้อยแต่เริ่มมีแนวโน้มเพิ่มขึ้นเล็กน้อยตามจำนวนของตัวประกอบหลักที่เพิ่ม และสุดท้ายคือขั้นตอนวิธีซี 4.5 คือใช้เวลาในการพิจารณาน้อยที่สุด และไม่มีแนวโน้มเพิ่มขึ้นตามจำนวนตัวประกอบหลักที่เพิ่ม

ดังนั้นสามารถแสดงค่าเฉลี่ยของเวลาที่ใช้ในการพิจารณาจดหมายข่าวขยะของแต่ละวิธีการแสดงได้ดังนี้



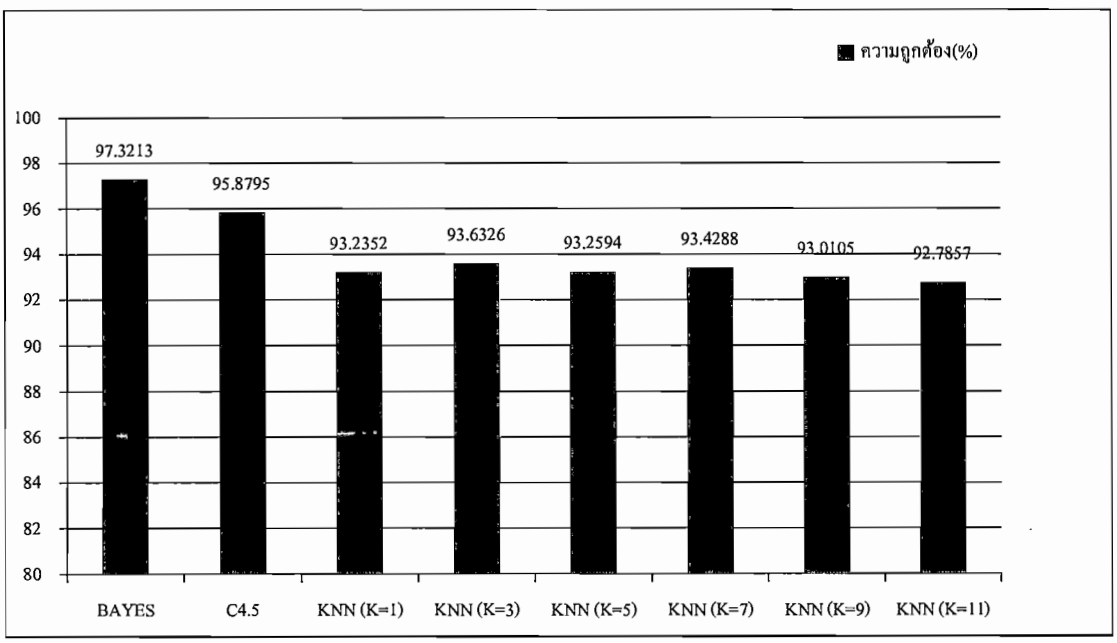
รูปที่ 4-7 เวลาเฉลี่ยที่ใช้ในแต่ละวิธี (วินาที)

จากรูปที่ 4-7 ค่าเฉลี่ยเวลาที่ใช้ในการดำเนินการทั้งหมดในแต่วิธี จากการทดลองทั้งหมดพบว่าขั้นตอนวิธีเบย์เซียนค่าที่ใช้เวลารวมทั้งในส่วนการเรียนรู้และการพิจารณาน้อยที่สุด ตามมาด้วยวิธีซี 4.5 ค่าเฉลี่ยที่ใช้เวลารวมทั้งในส่วนการเรียนรู้และการพิจารณารองลงมา ซึ่งซี 4.5 จะใช้เวลาการเรียนรู้สูงกว่าขั้นตอนวิธีเบย์เซียน แต่เวลาในการพิจารณาน้อยกว่า

และสุดท้ายขั้นตอนวิธีเค-เอ็นเอ็น มีค่าเฉลี่ยที่ใช้เวลารวมทั้งในส่วนการเรียนรู้และการพิจารณามากที่สุด ซึ่งขั้นตอนวิธีเค-เอ็นเอ็นใช้เวลาในการพิจารณาจดหมายข่าวขยะนานมากที่สุด

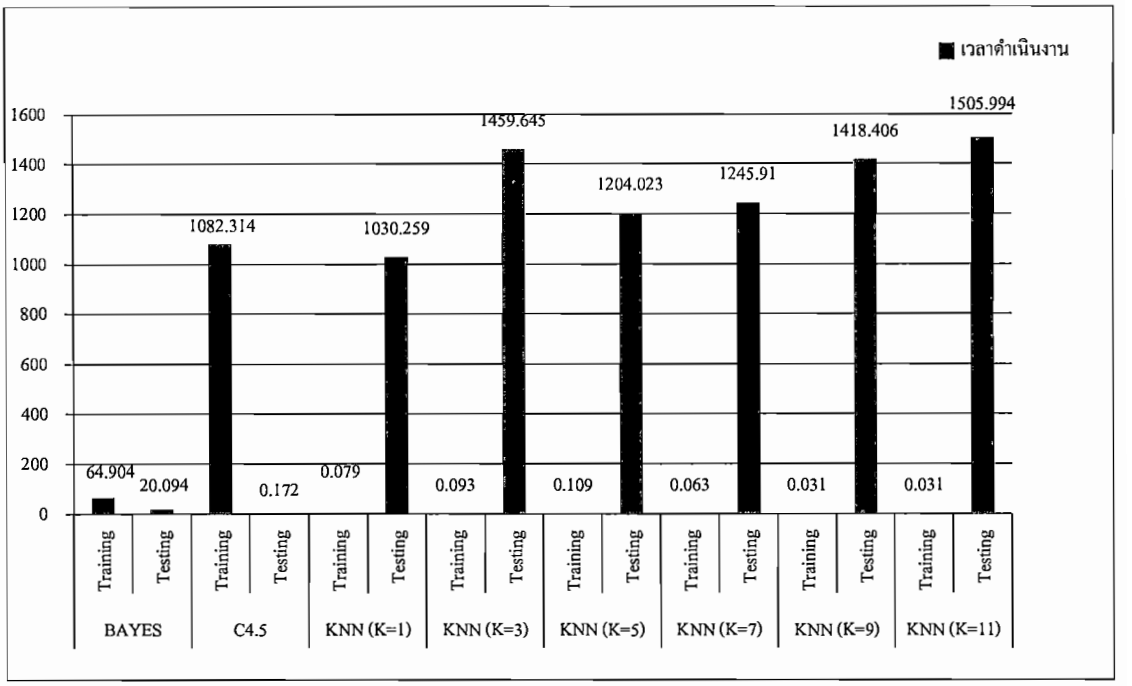
4.3 ผลการวิเคราะห์จดหมายข่าวขยะจากข้อมูลดั้งเดิมและผลการทดลองจากงานวิจัยต้นแบบ

จากผลการดำเนินงานในหัวข้อที่ผ่านมา จึงได้ทำการทดลองวิเคราะห์จดหมายข่าวขยะเพิ่มเติมโดยใช้ข้อมูลดั้งเดิมจากการตัดคำของชุดข้อมูล Ling Spam Corpus จำนวน 1000 คำ (ตัวแปร) จากจำนวนคำสูงสุด ซึ่งจะใช้ขั้นตอนวิธีการพิจารณาจดหมายข่าวขยะ เหมือนหัวข้อที่ผ่านมา แตกต่างที่ชุดข้อมูลไม่ได้ผ่านกระบวนการวิเคราะห์ตัวประกอบหลัก ด้วยโปรแกรม WEKA สามารถแสดงผลการทดลองได้ดังนี้



รูปที่ 4-8 ค่าเฉลี่ยความถูกต้อง (%) ในการพิจารณาจดหมายข่าวขยะด้วยข้อมูลดั้งเดิม

จากรูปที่ 4-8 ค่าเฉลี่ยความถูกต้อง จากการทดลองพบว่าขั้นตอนวิธีวิธีเบย์เซียนมีค่าเฉลี่ยความถูกต้องที่สูงสุดคือ 97.3213% รองลงมาคือวิธีซี 4.5 มีค่าเฉลี่ยความถูกต้อง 95.8795% และสุดท้ายคือเค-เอ็นเอ็นมีค่าเฉลี่ยความถูกต้องน้อยสุดที่ (k=3) ทำได้ 93.6326%



รูปที่ 4-9 เวลาที่ใช้ในการพิจารณาจดหมายข่าวขยะด้วยข้อมูลดั้งเดิม (วินาที)

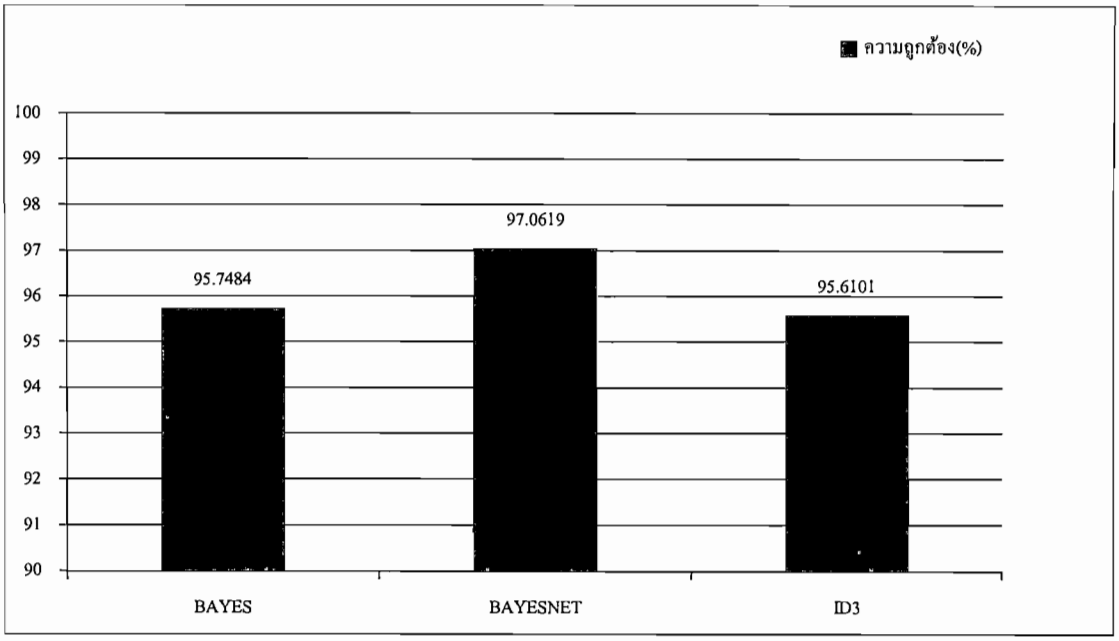
จากรูปที่ 4-9 แสดงเวลาที่ใช้ในการดำเนินการทั้งหมดในแต่ละวิธี จากการทดลองทั้งหมดพบว่าขั้นตอนวิธีเบย์เซียนค่าที่ใช้เวลารวมทั้งในส่วนการเรียนรู้และการพิจารณาน้อยที่สุด ตามมาด้วยวิธีซี 4.5 ค่าเฉลี่ยที่ใช้เวลารวมทั้งในส่วนการเรียนรู้และการพิจารณารองลงมา ซึ่งซี 4.5 จะใช้เวลาการเรียนรู้สูงกว่าขั้นตอนวิธีเบย์เซียน แต่เวลาในการพิจารณาน้อยกว่า

และสุดท้ายขั้นตอนวิธีเค-เอ็นเอ็น มีค่าเฉลี่ยที่ใช้เวลารวมทั้งในส่วนการเรียนรู้และการพิจารณามากที่สุด ซึ่งขั้นตอนวิธีเค-เอ็นเอ็นใช้เวลาในการพิจารณาจดหมายข่าวขยะนานมากที่สุด

จึงได้ทดลองตามขั้นตอนจากงานวิจัยของ เสนีย์ ทรัพย์บุญเลิศมา และศาตรา วงศ์ธนวสุ เรื่อง การเปรียบเทียบขั้นตอนวิธีตัวกรองสแปมอีเมลล์ โดยได้ดำเนินวิธีทดลองทำตามได้ผลดังนี้

ขั้นตอนวิธี	ค่าความถูกต้อง (%)
เบย์เซียน	95.7484
ข่ายงานเบย์เซียน	97.0619
ต้นไม้ตัดสินใจไอดี3	95.6101

ตารางที่ 4-7 ผลการทดลองตามงานวิจัยต้นแบบ



รูปที่ 4-10 ค่าความถูกต้อง (%) ตามงานวิจัยของเสนีย์ ทรัพย์บุญเลิศมาและศาตรา วงศ์ธนวสุ

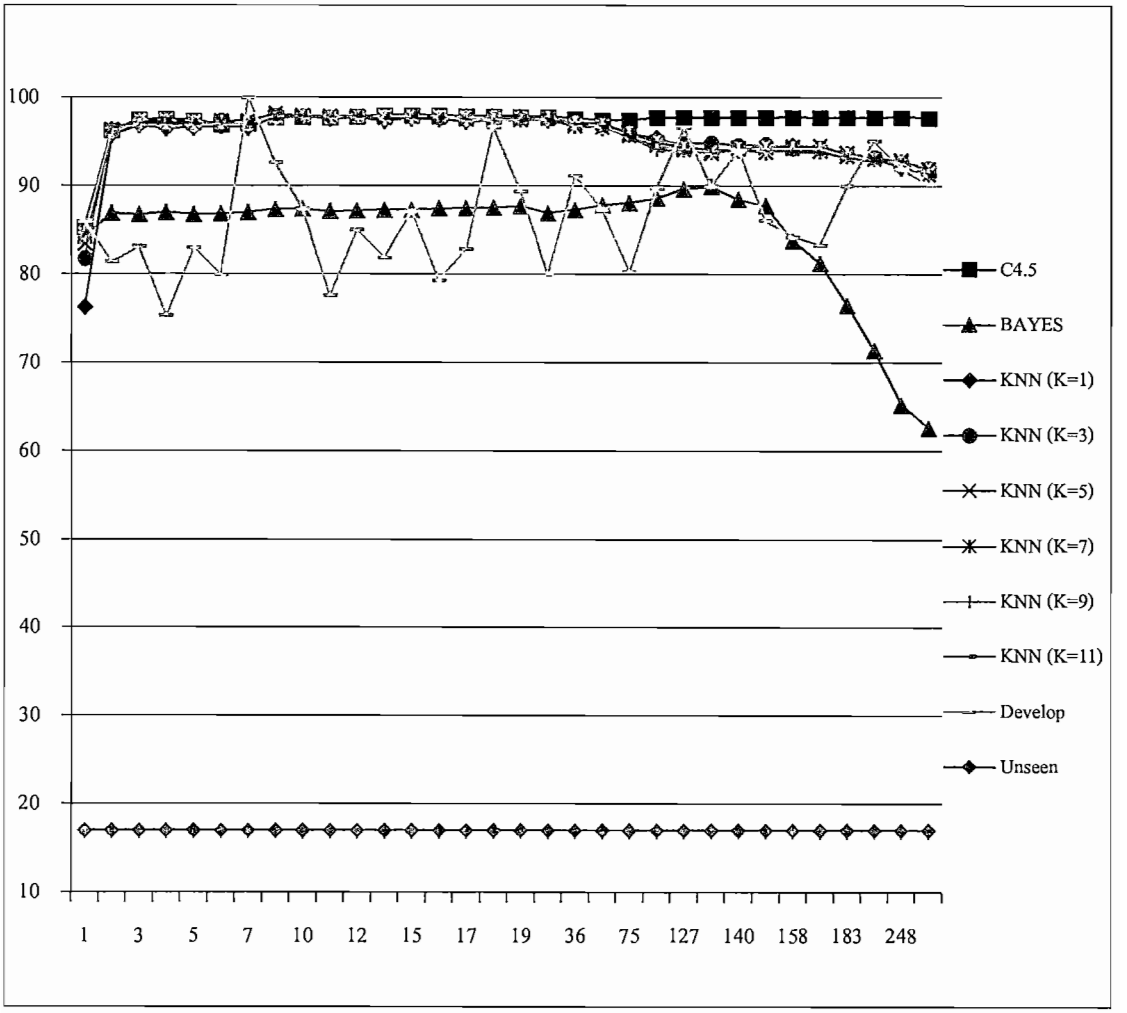
จากผลการทดลองค่าความถูกต้อง มีผลลัพธ์ใกล้เคียงกับผลการทดลองต้นฉบับซึ่งมีดังนี้ ขั้นตอนวิธีเบย์เซียน 95.23% ขั้นตอนวิธีข่ายงานเบย์เซียน 96.44% และขั้นตอนวิธีต้นไม้ตัดสินไอดี 3 มีค่าความถูกต้อง 94.12% ซึ่งพบว่าค่าความถูกต้องที่ได้มีความใกล้เคียงกันและดีขึ้นเล็กน้อย เนื่องจากวิธีการเลือกสุ่มตัวแปร เพื่อมาทดสอบจดหมายข่าวขยะ ดังนั้นผลการทดลองจึงแตกต่างกันได้

4.4 ผลการวิเคราะห์ด้วยต้นแบบโปรแกรมวิเคราะห์จดหมายข่าวขยะและผลทดลองกับชุดข้อมูลจริง

เมื่อได้ผลจากการทดลองต่างๆ แล้วนั้น เห็นว่าขั้นตอนวิธี ซี 4.5 มีค่าความถูกต้องโดยรวมอยู่ในเกณฑ์ดีที่สุดเมื่อเทียบกับขั้นตอนวิธีต่างๆ ในงานนิพนธ์นี้ และยังใช้เวลาในการทำงานอย่างรวดเร็ว ดังนั้นจึงเลือกขั้นตอนวิธีซี 4.5 ไปพัฒนาต้นแบบโปรแกรมวิเคราะห์จดหมายข่าวขยะ ซึ่งทำการทดลองโดยใช้ชุดข้อมูลกับวิธีทดลองเดียวกัน กับโปรแกรมเวก้า

จากนั้นได้ทำการทดลองกับชุดข้อมูลจริง ซึ่งได้รวบรวมจดหมายข่าวจากบริษัท อาร์ เอ็ม เอ กรุ๊ป จำกัด จำนวน 200 ฉบับ จดหมายข่าวขยะ166 ฉบับ และจดหมายปกติ 34 ฉบับ จดหมายที่ได้ทำการรวบรวม มีการนำหัวจดหมาย (E-mail Header) กับแท็กเอชทีเอ็มแอล (HTML TAG) ออก

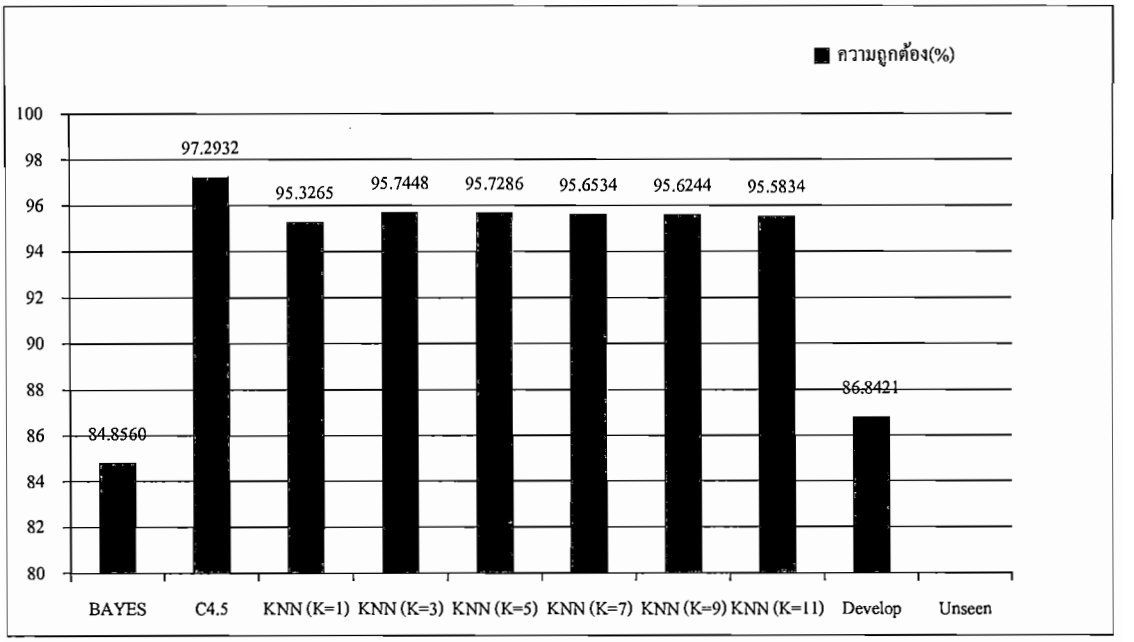
จากนั้นนำชุดสกัดคำจำนวน 1000 คำจากหัวข้อ 3.2 มาเป็นต้นแบบที่ใช้สกัดคำในชุดข้อมูลจริง ซึ่งการวิเคราะห์ตัวประกอบและการจำแนกจดหมายข่าวขยะด้วยขั้นตอนวิธีซี 4.5 ได้ดังนี้



รูปที่ 4-11 ผลการวิเคราะห์ด้วยต้นแบบโปรแกรมวิเคราะห์จดหมายข่าวขยะ

การทดลองพบว่าต้นแบบโปแกรมวิเคราะห์จดหมายข่าวขยะ สามารถจำแนกจดหมายข่าวขยะได้ค่าความถูกต้องเฉลี่ย 86.8421% แต่ผลลัพธ์ค่าความถูกต้องของแต่ละจำนวนตัวประกอบมีความสวิงอย่างมากโดย ค่าความถูกต้องสูงสุดทำได้ 100% ที่จำนวนตัวประกอบหลัก 13 ตัวและค่าความถูกต้องต่ำสุดได้ 75.2844% ที่จำนวนตัวหลัก 14 ตัว

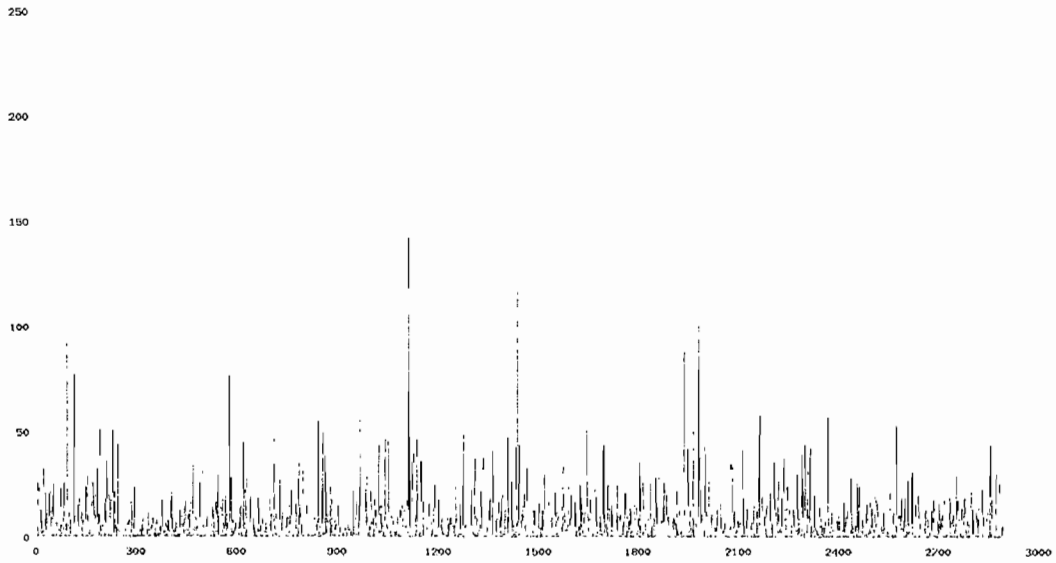
แต่เมื่อทดลองกับชุดข้อมูลจริงที่ได้ทำการรวบรวม พบว่าค่าความถูกต้องทำได้เพียง 17% เท่านั้นไม่ว่าจะใช้จำนวนตัวประกอบหลักกี่จำนวนก็ตาม



รูปที่ 4-12 ค่าเฉลี่ยความถูกต้อง (%) ด้วยต้นแบบโปรแกรมวิเคราะห์จดหมายข่าวขยะ

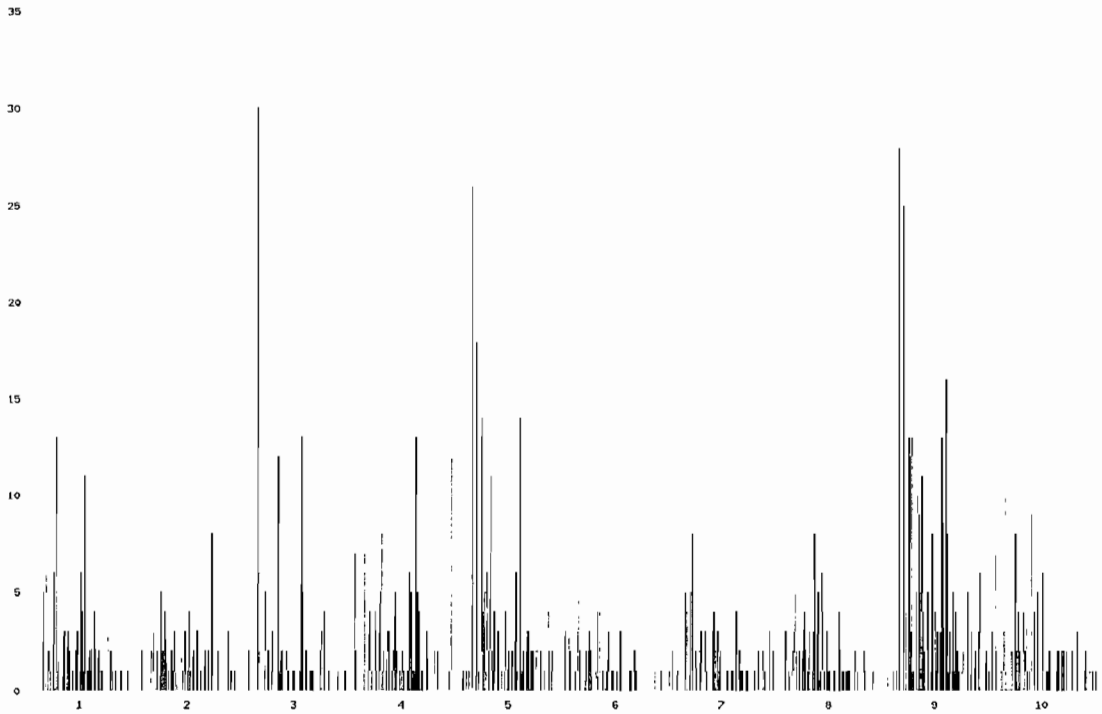
4.5 ผลการวิเคราะห์รูปแบบชุดข้อมูลจดหมาย

วิเคราะห์ในส่วนของรูปแบบชุดข้อมูล พบว่าชุดข้อมูลที่ผ่านการสกัดคำมานั้น จะเป็นความถี่ของคำ ซึ่งแต่ละค่าความถี่จะกระจายตัวแปรสุ่มแบบไม่ต่อเนื่อง มีค่าเป็นจำนวนเต็มบวกเท่านั้น ดังภาพ



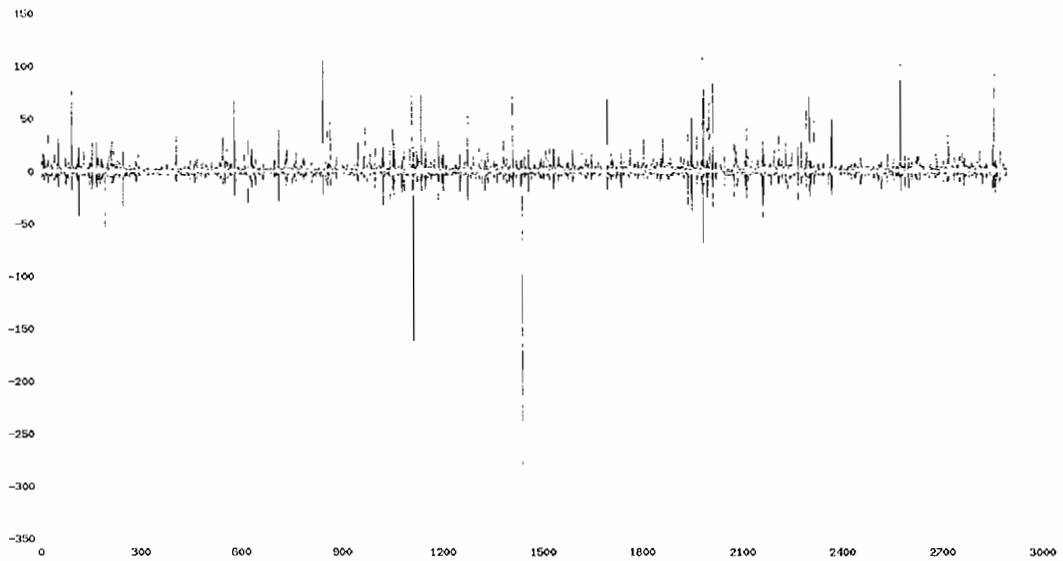
รูปที่ 4-13 กราฟการกระจายของความถี่ที่ได้จากสกัดคำจากจดหมาย 2,893 ฉบับ

เมื่อลองลดจำนวนจดหมายเหลือเพียง 10 ฉบับ เพื่อดูรายละเอียดความถี่ แสดงได้ดังภาพ



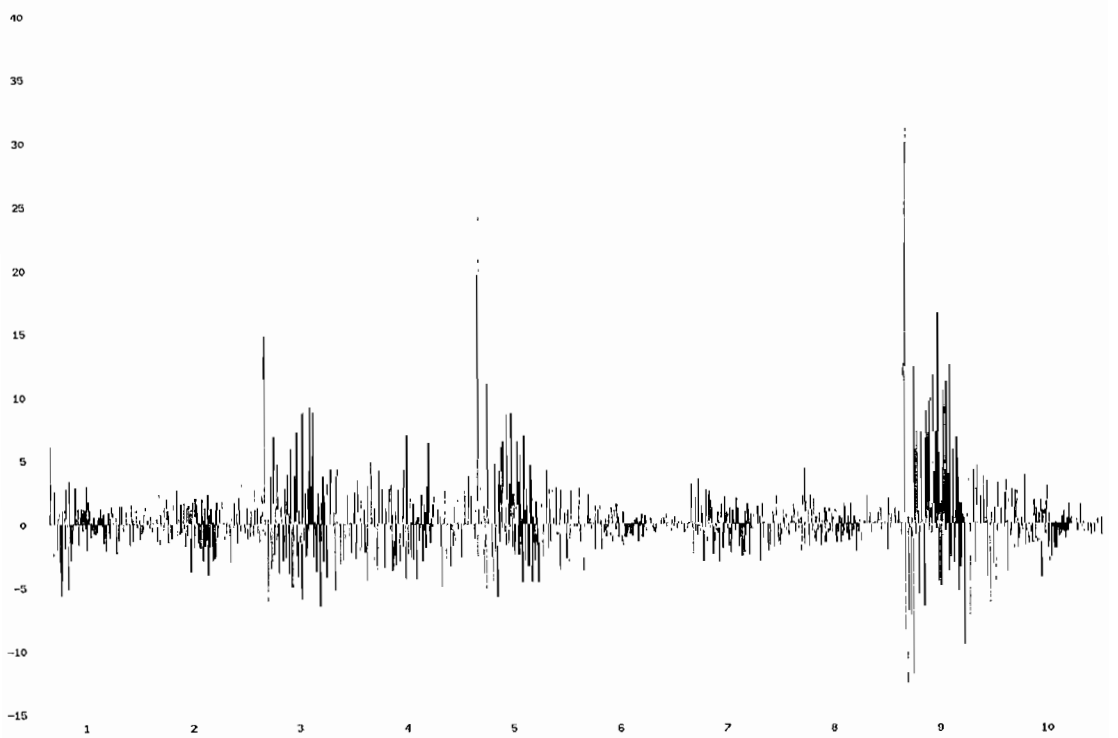
รูปที่ 4-14 กราฟการกระจายของความถี่ที่ได้จากสกัดคำจากจดหมาย 10 ฉบับ

แต่เมื่อนำชุดข้อมูลไปผ่านขั้นตอนวิเคราะห์ตัวประกอบหลัก จะทำให้ค่าของผลลัพธ์ที่ได้เปลี่ยนไปเป็นจำนวนจริงซึ่งมีทั้งค่าบวกและค่าลบ ขึ้นอยู่กับความสัมพันธ์ค่าความแปรปรวนของชุดข้อมูล ดังภาพ



รูปที่ 4-15 กราฟการกระจายของความถี่จากวิเคราะห์ตัวประกอบหลัก

เมื่อแสดงรายละข้อมูลตัวประกอบหลักจำนวน 10 ตัว พบว่าค่าของตัวแปรที่ผ่านการวิเคราะห์ตัวประกอบที่มีค่าบวกและลบนั้น มีลักษณะคล้ายการแจกแจงปกติ (Standard Normal Distribution) ที่เป็นระฆังคว่ำ ดังภาพ



รูปที่ 4-16 กราฟการกระจายของความถี่จากวิเคราะห์ตัวประกอบหลักที่ 10 ชุดข้อมูล

บทที่ 5
สรุปผลการดำเนินงาน

งานวิจัยนี้ได้นำเสนอวิธีการพิจารณาจดหมายข่าวขยะ ด้วยการลดจำนวนตัวแปรหรือความถี่ของคำในจดหมาย โดยการวิเคราะห์ตัวประกอบหลัก จากนั้นจึงจดหมายข่าวขยะด้วยวิธีต่างๆ ได้แก่การคำนวณเบย์เซียน การใช้ขั้นตอนวิธีซี4.5 และการขั้นตอนวิธี เค-เอ็นเอ็น โดยใช้ชุดข้อมูลอีเมลล์จาก Ling Spam Corpus

ผลการดำเนินงานสามารถสรุปได้ว่า การลดจำนวนตัวแปร ด้วยการวิเคราะห์ตัวประกอบหลักนั้น ยังสามารถให้ผลการพิจารณาจดหมายข่าวขยะได้เป็นที่น่าพอใจ โดยค่าเฉลี่ยความถูกต้องของการจำแนกจดหมายข่าวขยะ ด้วยวิธีซี4.5 ให้ค่าเฉลี่ยความถูกต้องมากที่สุดคือ 97.276% รองลงมาคือวิธีเค-เอ็นเอ็น (k=5) ให้ค่าเฉลี่ยความถูกต้อง 95.765% และเบยส์ ให้ค่าเฉลี่ยความถูกต้อง 84.889%

ความถูกต้องสูงสุดที่ทำได้ ด้วยวิธีเค-เอ็นเอ็น (K=5) ด้วยจำนวนตัวประกอบ 8 ตัวคือ 98.168% รองลงมาคือวิธี ซี4.5 ด้วยจำนวนตัวประกอบ 14 ตัว สูงสุดทำได้คือ 98.0643% และสุดท้าย เบยส์ ด้วยจำนวนตัวประกอบ 135 ตัว สูงสุด 89.9412%

หากเมื่อเมื่อทำการวิเคราะห์จดหมายข่าวขยะด้วยวิธีต่างๆ โดยไม่ผ่านขั้นตอนลดตัวแปรด้วยการวิเคราะห์ตัวประกอบหลักแล้ว ผลปรากฏว่า ค่าความถูกต้องมากที่สุดคือวิธีเบย์เซียนได้ 97.3213% รองลงมาคือวิธีซี 4.5 ได้ 95.3213% และสุดท้ายวิธีเค-เอ็นเอ็น (K=3) ได้ 93.2352%

ผลความถูกต้องของขั้นตอนวิธีการจำแนกจดหมายข่าวขยะ สามารถให้ผลลัพธ์ค่าความถูกต้องเป็นที่หน้าพอใจเมื่อเปรียบเทียบกับผลลัพธ์ของการจำแนกจดหมายข่าวขยะ ที่ไม่ผ่านขั้นตอนการลดตัวแปร จากตัวแปรทั้งหมด 1000 คำ ซึ่งสามารถแสดงผลได้ดังนี้

วิธีการ	ชุดข้อมูลที่ผ่านการวิเคราะห์ตัวประกอบหลัก		ชุดข้อมูลดิบ
	ค่าความถูกต้องเฉลี่ย	ค่าความถูกต้องสูงสุด	ค่าความถูกต้อง
วิธีเบย์เซียน	84.8560 %	89.9412%	97.3213%
วิธี ซี4.5	97.2932%	98.0643%	95.8795%
วิธีเค-เอ็นเอ็น (K=1)	95.3492%	97.8915%	93.2352%
วิธีเค-เอ็นเอ็น (K=3)	95.7448%	97.9883%	93.6326%
วิธีเค-เอ็นเอ็น (K=5)	95.7286%	98.168%	93.2594%
วิธีเค-เอ็นเอ็น (K=7)	95.6534 %	97.9606%	93.4288%
วิธีเค-เอ็นเอ็น (K=9)	95.6244 %	97.9433%	93.0105%
วิธีเค-เอ็นเอ็น (K=11)	95.5834%	97.9191%	92.7857%

ตารางที่ 5-1 สรุปผลค่าความถูกต้องการทดสอบจดหมายข่าวขยะ

หากวิเคราะห์ผลลัพธ์ของค่าความถูกต้องในตารางที่ 5-1 นั้นพบว่า ค่าความถูกต้องสูงสุดด้วยวิธีซี4.5 ที่ผ่านขั้นตอนวิเคราะห์ตัวประกอบหลัก สามารถให้ผลลัพธ์สูงกว่าค่าความถูกต้องด้วยวิธีเบย์เซียนที่ไม่ผ่านขั้นตอนวิเคราะห์ตัวประกอบหลัก และค่าความถูกต้องเฉลี่ยของวิธีซี4.5 ที่ผ่านขั้นตอนวิเคราะห์ตัวประกอบหลัก ใกล้เคียงกับค่าความถูกต้องของวิธีเบย์เซียนที่ไม่ผ่านขั้นตอนวิเคราะห์ตัวประกอบหลักอีกด้วย

จากการทดสอบพบว่าขั้นตอนวิธีเบย์เซียน มีประสิทธิภาพในการจำแนกจดหมายข่าวขยะลดลงหลังจากการวิเคราะห์ตัวประกอบเนื่องจาก

- ในการคำนวณความน่าจะเป็น ข้อมูลที่ใช้คำนวณจะต้องเป็นข้อมูลที่เป็นจำนวนจริงที่มีค่าเป็นบวก

- ชุดข้อมูลจดหมายที่เหมาะสมกับวิธีเบย์เซียน ควรเป็นข้อมูลที่มีลักษณะเป็นตัวแปร ที่มีการกระจายตัวแปรสุ่มแบบไม่ต่อเนื่อง และควรเป็นชุดข้อมูลที่ยังไม่ได้ผ่านการวิเคราะห์ตัวประกอบหลัก

สำหรับขั้นตอนวิธีซี 4.5 มีประสิทธิภาพในการจำแนกจดหมายข่าวขยะ ที่มีแนวโน้มดีขึ้นกว่าขั้นตอนวิธีอื่นๆ หลังจากที่ข้อมูลผ่านการวิเคราะห์ตัวประกอบ สันนิษฐานได้ว่า

- ชุดข้อมูลที่ผ่าน PCA มีกระจายตัวที่มีลักษณะคล้ายการแจกแจงปกติ ซึ่งสอดคล้องกับวิธีคำนวณค่า เกนเรโช (GainRatio) เพื่อหารากของต้นไม้ ทำให้ได้โมเดลของต้นไม้ตัดสินใจที่ดี

เมื่อพิจารณาความเร็วเฉลี่ยของการพิจารณาสแปมอีเมลล์ ขั้นตอนวิธีเบยส์มีการใช้เวลาทำงานโดยรวมน้อยที่สุดตามมาด้วย ซี4.5 และเค-เอ็นเอ็น ซึ่งอัตราความเร็ว ซี4.5 และเค-เอ็นเอ็นจะลดลงตามจำนวนตัวประกอบหลักที่เพิ่มขึ้น

วิธีการ	ชุดข้อมูลที่ผ่านการวิเคราะห์ตัวประกอบหลัก		ชุดข้อมูลดิบ	
	ค่าเฉลี่ยเวลาที่ใช้ในการเรียนรู้ (วินาที)	ค่าเฉลี่ยเวลาที่ใช้ในการทดสอบ (วินาที)	เวลาที่ใช้ในการเรียนรู้ (วินาที)	เวลาที่ใช้ในการทดสอบ (วินาที)
วิธีเบย์เซียน	6.5438	0.8878	64.904	20.094
วิธี ซี4.5	40.4529	0.0409	1082.314	0.172
วิธีเค-เอ็นเอ็น (K=1)	0.0369	90. 1018	0.079	1030.259
วิธีเค-เอ็นเอ็น (K=3)	0.0363	103. 1076	0.093	1459.645
วิธีเค-เอ็นเอ็น (K=5)	0.0399	107.8397	0.109	1204.023
วิธีเค-เอ็นเอ็น (K=7)	0.0420	108. 9478	0.063	1245.91
วิธีเค-เอ็นเอ็น (K=9)	0.0346	109.9914	0.031	1418.406
วิธีเค-เอ็นเอ็น (K=11)	0.0442	111.6194	0.031	1505.994

ตารางที่ 5-2 สรุปผลเวลาที่ใช้ในทดสอบจดหมายข่าวขยะ

จะเห็นได้ว่าขั้นตอนวิธีต่างๆ ที่ใช้ในการพิจารณา สแปมอีเมลล์นั้นมีการใช้เวลาในการทำงานแตกต่างกันตามแต่ละวิธี ซึ่งจะเห็นได้ว่าค่าเฉลี่ยที่ใช้ในการเรียนรู้และค่าเฉลี่ยที่ใช้ในการพิจารณาแตกต่างกันอย่างเห็นได้ชัด เช่นซี4.5 จะใช้ระยะเวลาในการเรียนรู้มาก แต่ใช้เวลาในการพิจารณาน้อยที่สุด ซึ่งจากขั้นตอนวิธีเค-เอ็นเอ็นจะใช้เวลาการเรียนรู้น้อยที่สุด แต่กลับใช้เวลาในการพิจารณามากที่สุด สำหรับขั้นตอนของเบยส์นั้นจะเห็นได้การทำงานโดยรวมรวดเร็วที่สุด

หากเปรียบเทียบในส่วนของประสิทธิภาพในการดำเนินงาน พบว่าการจำแนกจดหมายข่าวขยะด้วยวิธีซี4.5 นั้น หากนำตัวแปรที่มีการวิเคราะห์ตัวประกอบหลัก จะมีความเร็วสูงกว่าข้อมูลดิบอย่างเห็นได้ชัด อีกทั้งค่าความถูกต้องเฉลี่ยก็มีค่าสูงกว่าความถูกต้องของข้อมูลดิบอีกด้วย จึงเห็นว่าควรมีการพัฒนาต้นแบบของระบบสารสนเทศในการจำแนกจดหมายข่าวขยะด้วยวิธีซี 4.5

ข้อเสนอแนะ

เห็นได้ว่าผลที่ได้จากการดำเนินงาน ในแต่ละขั้นตอนวิธี เหมาะสมกับชุดข้อมูลที่ผ่านการวิเคราะห์ตัวประกอบในรูปแบบต่างๆ เช่นว่า ขั้นตอนวิธีเบยส์อาจจะเหมาะกับการพิจารณาอีเมลล์ที่มีปริมาณมาก เพราะเนื่องจากมีความรวดเร็วในการดำเนินงาน ทั้งในส่วนของการเรียนรู้และการวิเคราะห์ สำหรับซี4.5 เหมาะกับการพิจารณาอีเมลล์ที่มีความคล้ายคลึงกัน ส่วนขั้นตอนวิธีเค-เอ็นเอ็น เหมาะกับการพิจารณาอีเมลล์ที่มีชุดข้อมูลตัวอย่างน้อย

หากเปรียบเทียบกับงานวิจัยเรื่อง การเปรียบเทียบขั้นตอนวิธีตัวกรองสแปมอีเมลล์ ของ เสนีย์ ทรัพย์บุญเลิศมา และ ศาสตรา วงศ์ธนวสุ นั้นพบว่าค่าความถูกต้องในส่วนของซี 4.5 เพิ่มขึ้นเล็กน้อย ซึ่งอาจจะเป็นที่ขั้นตอนวิธีซี 4.5 ได้มีการพัฒนาจากไอดี3 (ID3) ของงานวิจัยดังกล่าว แต่ขั้นตอนวิธีของเบย์เซียนกลับลดลงอย่างเห็นได้ชัด ตามที่ได้นำเสนอมาแล้ว

สำหรับการดำเนินงานต่อไปเห็นว่า ควรมีการศึกษาเพิ่มเติมในส่วนของ เปรียบเทียบขั้นตอนวิธีในการลดจำนวนมิติตัวแปรในวิธีต่างๆ เช่นการวิเคราะห์องค์ประกอบหลักเพื่อเพิ่มประสิทธิภาพในการจัดการชุดข้อมูล และควรมีการนำชุดข้อมูลอีเมลล์นอกเหนือจากชุดข้อมูล Ling Spam Corpus และมีลักษณะต่างๆ กัน เพื่อให้ได้โมเดลของผลการทดลองที่ครอบคลุมมากขึ้น

การนำต้นแบบโปรแกรมวิเคราะห์จดหมายข่าวขยะ ไปพัฒนาต่อเพื่อประยุกต์ใช้งาน สามารถนำไปพัฒนาเป็นแอพพลิเคชั่นเพื่อไปใช้กับระบบเว็บเมลล์เซิร์ฟเวอร์ เพื่อทำการวิเคราะห์แยกจดหมายที่ดี กับจดหมายข่าวขยะก่อนเข้าสู่กล่องจดหมายส่วนตัว บนเว็บเมลล์เซิร์ฟเวอร์ได้ หรือสามารถนำไปพัฒนาเป็นแอพพลิเคชั่นเพิ่มเติม (Add-on) สำหรับโปรแกรมอีเมลล์ไคลเอนต์ เช่น ไมโครซอฟต์เอาท์ลุค (Microsoft Outlook) หรือธันเดอร์เบิร์ดได้ โดยจุดเด่นของโปรแกรมจะเปรียบเสมือนระบบวิเคราะห์จดหมายข่าวขยะส่วนตัว ซึ่งตัวโปรแกรมจะทำการคำนวณตัวประกอบหลักของแต่ผู้ใช้โดยอิสระ

บรรณานุกรม

Thomas B. Fomby. (2008). *K-Nearest Neighbours Algorithm: Prediction and Classification.* Department of Economics Southern Methodist University.

เฉลิมพล ณ สงขลา และ เกริก ภิรมย์โสภา. (2009). *การเพิ่มประสิทธิภาพการจัดจำแนกอีเมลสแปมภาษาไทยด้วยโปรแกรมตัดคำไทยคูวส์.* ภาควิชาวิศวกรรมศาสตร์. จุฬาลงกรณ์มหาวิทยาลัย.

เสนีย์ ทรัพย์บุญเลิศมา และ ศาสตรา วงศ์ธนวสุ. *การเปรียบเทียบขั้นตอนวิธีตัวกรองสแปมอีเมลล์.* ภาควิชาวิทยาการคอมพิวเตอร์. คณะวิทยาศาสตร์. มหาวิทยาลัยขอนแก่น.

นรุตม์ บุตรพลอย. (2010). *การจำแนกข้อมูลสูญหายด้วย Soft k-Nearest Neighbor.* วิทยานิพนธ์ปริญญา. วิศวกรรมศาสตรมหาบัณฑิต. สาขาวิชาวิศวกรรมคอมพิวเตอร์. บัณฑิตวิทยาลัย มหาวิทยาลัยขอนแก่น

บุญเสริม กิจศิริกุล. (2006). *ปัญญาประดิษฐ์.* ภาควิชาวิศวกรรมศาสตร์. คณะวิศวกรรมศาสตร์. จุฬาลงกรณ์มหาวิทยาลัย.

หทัยชนก กรชี. (2008). *การคำนวณน้ำหนักของคุณลักษณะข้อมูลสำหรับตัวจำแนกไนอีฟเบย์ โดยวิธีจัดกลุ่มข้อมูลและต้นไม้ตัดสินใจ.* ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์). สาขาวิทยาการคอมพิวเตอร์. มหาวิทยาลัยเกษตรศาสตร์.

สิทธิโชค มุกดาสกุลภิบาล. (2008). *การวัดประสิทธิภาพของขั้นตอนวิธี ตัวจำแนก C4.5, ADTree และ Naïve Bayes ในการจำแนกข้อมูลการซุกซ่อนสิ่งเสพติดสำหรับไปรษณีย์ระหว่างประเทศ.* ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์). สาขาวิทยาการคอมพิวเตอร์. มหาวิทยาลัยเกษตรศาสตร์.

Simon Haykin. (1998). *Neural Networks: A Comprehensive Foundation (2nd edition)*, Prentice Hall.

กัลยา วานิชย์บัญชา. (2007). *การวิเคราะห์ข้อมูลหลายตัวแปร.* กรุงเทพมหานคร: บริษัท ธรรมสาร จำกัด.

Thales Sehn Korting. *C4.5 algorithm and Multivariate Decision Trees.* Image Processing Division, National Institute for Space Research – INPE. Sao Jose dos Campos – SP. Brazil.

ปริญญา สงวนสัตย์. (2010). *คู่มือ MATLAB ฉบับสมบูรณ์.* นนทบุรี : บริษัท ไอดีซี พรีเมียร์ จำกัด.

ภาคผนวก

ภาคผนวก ก

ผลการทดลองเพิ่มเติม

ตัวอย่างตารางค่าไอเกนจำนวน 300 ตัว

ลำดับที่	ค่าไอเกน	ลำดับที่	ค่าไอเกน	ลำดับที่	ค่าไอเกน	ลำดับที่	ค่าไอเกน	ลำดับที่	ค่าไอเกน	ลำดับที่	ค่าไอเกน
1	90.39579	51	2.159202	101	1.145675	151	0.75538	201	0.531561	251	0.395953
2	58.44924	52	2.123504	102	1.134662	152	0.74741	202	0.525932	252	0.392523
3	40.9132	53	2.091054	103	1.122189	153	0.74183	203	0.524891	253	0.390824
4	30.97092	54	2.070663	104	1.119188	154	0.736075	204	0.521984	254	0.38924
5	21.20991	55	2.024066	105	1.11399	155	0.730954	205	0.517483	255	0.386603
6	17.63984	56	2.006304	106	1.096538	156	0.724577	206	0.516238	256	0.385095
7	13.80334	57	1.965922	107	1.084321	157	0.723884	207	0.511146	257	0.380558
8	12.9932	58	1.900817	108	1.06852	158	0.720745	208	0.510157	258	0.377296
9	8.831807	59	1.889594	109	1.063437	159	0.708461	209	0.504084	259	0.376537
10	8.59332	60	1.866052	110	1.049307	160	0.704013	210	0.500687	260	0.37447
11	7.355486	61	1.854245	111	1.041033	161	0.700644	211	0.497846	261	0.371869
12	7.03529	62	1.806141	112	1.025418	162	0.694021	212	0.493815	262	0.370614
13	6.644568	63	1.776947	113	1.022443	163	0.68646	213	0.491352	263	0.368563
14	6.539314	64	1.745129	114	1.008026	164	0.682963	214	0.487413	264	0.367052
15	6.236129	65	1.72096	115	0.999553	165	0.677831	215	0.485679	265	0.365286
16	5.748768	66	1.714532	116	0.994311	166	0.671619	216	0.484181	266	0.361863
17	5.38363	67	1.698873	117	0.984258	167	0.668185	217	0.481143	267	0.360326
18	5.205616	68	1.651964	118	0.972781	168	0.663269	218	0.478698	268	0.358031
19	5.009546	69	1.632436	119	0.972401	169	0.657932	219	0.474086	269	0.355631
20	4.897356	70	1.619907	120	0.956389	170	0.650236	220	0.47059	270	0.354556
21	4.765341	71	1.611443	121	0.945701	171	0.649157	221	0.469114	271	0.353353
22	4.377338	72	1.572051	122	0.943953	172	0.647994	222	0.465964	272	0.351585
23	4.311117	73	1.550345	123	0.93006	173	0.640638	223	0.465106	273	0.347944
24	4.231858	74	1.542408	124	0.928911	174	0.636858	224	0.459771	274	0.345691
25	4.080846	75	1.511733	125	0.921395	175	0.63083	225	0.456374	275	0.344484
26	4.014385	76	1.491765	126	0.913565	176	0.627	226	0.452749	276	0.343024
27	3.838272	77	1.474189	127	0.903422	177	0.621152	227	0.452122	277	0.341641
28	3.791324	78	1.458402	128	0.898571	178	0.619088	228	0.448111	278	0.34046
29	3.672768	79	1.448397	129	0.889044	179	0.616027	229	0.446611	279	0.33843
30	3.558834	80	1.433038	130	0.877732	180	0.61072	230	0.445106	280	0.337217
31	3.511697	81	1.418796	131	0.872723	181	0.607282	231	0.442673	281	0.335687
32	3.394533	82	1.40115	132	0.870676	182	0.604532	232	0.439955	282	0.334328
33	3.300149	83	1.394444	133	0.859423	183	0.60315	233	0.438065	283	0.333163
34	3.158887	84	1.37245	134	0.85289	184	0.593917	234	0.433518	284	0.330541
35	3.07739	85	1.353754	135	0.845622	185	0.589407	235	0.429238	285	0.328381
36	3.016574	86	1.34839	136	0.841212	186	0.586272	236	0.427635	286	0.327523
37	2.933937	87	1.33392	137	0.837933	187	0.585475	237	0.426978	287	0.326836
38	2.852721	88	1.332106	138	0.829558	188	0.582436	238	0.426588	288	0.323975
39	2.815092	89	1.309544	139	0.823827	189	0.576709	239	0.422857	289	0.320206
40	2.750967	90	1.289838	140	0.819667	190	0.569012	240	0.419565	290	0.31839
41	2.649918	91	1.279803	141	0.80882	191	0.566279	241	0.41895	291	0.316002
42	2.618197	92	1.267808	142	0.803078	192	0.565134	242	0.417021	292	0.315402
43	2.552176	93	1.230917	143	0.797315	193	0.559571	243	0.413536	293	0.313181
44	2.471846	94	1.222021	144	0.794141	194	0.554767	244	0.408918	294	0.31107
45	2.419429	95	1.210272	145	0.789634	195	0.551214	245	0.407702	295	0.309982
46	2.401622	96	1.199371	146	0.777064	196	0.546331	246	0.405907	296	0.30723
47	2.37103	97	1.187358	147	0.774583	197	0.544773	247	0.40381	297	0.306627
48	2.316634	98	1.178678	148	0.765109	198	0.543662	248	0.401153	298	0.304851
49	2.272858	99	1.173088	149	0.763249	199	0.536443	249	0.399619	299	0.304144
50	2.265157	100	1.164816	150	0.758871	200	0.534912	250	0.397252	300	0.302742

ตารางตัวอย่างผลลัพธ์ที่ผ่านการวิเคราะห์ตัวประกอบหลัก

ตัวประกอบที่ / จดหมายลำดับที่	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	....	999	1000	ชนิดจดหมาย
1	0.66632	6.0701	-0.09157	4.5088	-2.1419	-8.9005	2.5804	2.5401	-3.7499	3.8707	-2.0081	0.42466	-6.4229	1.641	-2.9727	-3.7397	-5.5648	1.013	....	-0.31729	-0.91944	HAM
2	0.41694	-1.2475	0.11027	3.934	-1.8045	-1.5491	0.16698	0.59694	1.4111	-3.0313	1.348	-0.08888	0.25807	1.944	0.96824	1.0278	-1.4786	0.097816	....	0.26294	0.41112	HAM
3	1.8709	14.723	-0.85355	19.032	9.2639	10.189	-6.1207	-2.2019	-6.5371	-2.8587	3.6403	-3.5036	4.2252	-2.8708	6.8068	-2.3704	-0.14762	4.9846	....	-0.07171	-0.89842	HAM
4	4.8065	3.7552	-0.01585	2.9745	-1.421	-3.5747	-0.11686	-0.29481	5.8446	8.159	-1.0318	4.2088	-5.4353	-0.16682	0.052261	2.5762	2.408	-1.5417	....	-0.14756	0.059308	HAM
5	5.3094	19.659	-2.7009	28.31	5.242	-0.13793	-2.8494	0.008546	-3.81	8.2268	-3.8054	-1.8362	3.6607	-5.172	10.988	6.4262	3.0665	3.1891	....	0.018944	-0.74286	HAM
6	-3.6242	-0.86092	-0.53365	1.4622	0.57561	-0.7209	1.6152	0.40658	-0.48787	-0.32203	0.11825	0.28615	-0.31845	-0.52	0.26814	0.007716	-1.7404	0.61467	....	-0.34462	-0.56206	HAM
7	2.3757	3.2089	0.20793	0.89908	-2.4436	0.57775	1.5735	2.06	5.9697	2.9484	-0.88951	3.5228	-2.9279	-0.70835	1.4004	0.13426	0.80848	-0.72183	....	0.35818	-0.39838	HAM
8	0.9796	-0.16758	0.19715	2.7206	-0.93101	-2.4483	2.447	2.1701	3.6581	-2.581	4.3437	1.7504	0.76859	0.33009	1.0909	-0.60905	-1.7686	2.2291	....	0.28879	0.76441	HAM
9	12.605	30.008	-8.412	39.324	0.30379	-12.722	-3.3687	-6.8588	-5.4542	7.715	-7.2468	-1.1513	8.4313	-11.815	12.321	5.5128	5.9176	6.1499	....	0.61263	-0.38533	HAM
10	-0.95035	2.6791	-0.19467	5.048	1.0137	-2.4427	0.17529	1.6247	-0.75636	-2.5398	1.3762	-1.2105	0.03931	0.91588	1.4119	-0.24656	0.55148	-0.48048	....	0.52256	-0.21554	HAM
11	11.829	18.141	-1.7496	9.8286	-6.7075	-13.071	14.397	3.7127	8.9736	24.644	-9.1264	2.9891	-15.432	5.9787	-6.8116	-1.5382	2.1643	-1.8882	....	0.48196	0.29089	HAM
12	-4.068	-2.3348	0.26561	0.52278	1.3641	0.66756	1.8155	0.51972	0.7726	-0.39928	-0.32376	-0.33591	-0.23594	-0.1373	0.90976	0.46804	-0.72024	0.075627	....	0.029531	0.25276	HAM
13	-4.3819	-4.8926	-0.07179	0.14696	0.28521	-0.4814	0.93619	-0.43095	0.18749	-0.15523	-0.30376	-0.24903	-0.76195	-0.05668	0.23975	0.009163	-0.01291	-0.18886	....	-0.08069	0.055559	HAM
14	-3.0968	-5.2168	0.10738	0.30604	-1.4223	0.61612	0.69508	-0.4757	0.19489	-0.72064	-1.1742	-0.87156	-0.47793	0.27059	-0.22526	0.061369	-0.18275	0.050071	....	0.19784	0.015296	SPAM
15	-4.4321	-5.3624	0.15634	-0.33015	-0.16634	0.45342	0.69042	-0.08408	-0.39927	-0.1843	-0.74351	-0.05755	0.12689	0.30448	0.052615	0.26258	-0.25394	-0.14847	....	0.022603	0.012246	SPAM
16	9.24	-3.9242	-1.0664	1.333	-5.1501	-7.7833	-2.2681	-34.788	1.6798	-1.1724	6.5685	-5.2678	-7.8029	0.852	3.4503	5.057	-7.2328	-1.6419	....	0.48861	-0.35841	SPAM
17	0.51395	-5.5048	-0.7232	-0.80513	-1.5082	-0.75791	-1.1334	-5.7957	-0.12238	2.8166	0.75622	1.7482	3.6758	-1.7442	-0.08966	-2.2241	0.59151	-3.0077	....	-0.46676	0.12396	SPAM
18	18.001	-8.2911	-2.9919	0.3928	-6.7982	3.6447	0.77978	-4.8586	-3.9488	3.8448	-7.8491	2.5747	5.8941	3.2789	0.94112	-1.8652	-1.5878	-0.42181	....	-0.42293	-1.0743	SPAM
19	-3.1465	-5.3334	0.32805	-0.54976	0.32833	-0.12048	0.31126	-0.23946	-0.70082	0.92029	-0.10126	0.44263	1.1446	0.151	-0.09652	0.059059	-0.40439	-1.211	....	0.10827	-0.30301	SPAM
20	-4.1622	-5.4087	0.1312	-0.34918	0.36828	0.21543	0.76249	-0.17652	0.12723	-0.19026	-0.04344	-0.06504	-0.62318	0.1209	0.181	-0.25886	-0.24695	0.15356	....	-0.14948	0.005057	SPAM
21	-3.573	-4.9913	-0.16297	0.1028	0.12254	0.6822	0.32292	-0.98566	-0.4881	0.80656	-0.5957	0.61317	0.91963	-0.12579	0.061882	-0.5816	0.17404	-0.84009	....	0.060014	0.51619	SPAM
22	-4.8395	-5.1583	0.23009	-0.51662	0.50102	-0.11609	0.61506	-0.48235	-0.26549	-0.28532	-0.38796	-0.22911	-0.39243	-0.17063	-0.05358	0.26761	0.099253	-0.27676	....	0.19134	-0.12928	SPAM
23	3.6231	-4.8272	-1.8441	-1.5214	-2.6024	-1.4969	-2.615	-8.0501	0.78658	3.5748	-0.42306	0.88065	5.428	-0.69637	0.92063	-2.5396	-0.17891	-4.2476	....	-0.41925	-0.84769	SPAM
24	-4.1383	-4.9878	0.093682	-0.08022	-0.16932	-0.22887	0.33354	-0.38939	-0.75504	0.52881	-0.87796	0.40333	0.25087	-0.20789	-0.15377	-0.33318	-0.16578	-0.25046	....	0.2759	0.11567	SPAM
25	-2.4329	-5.7318	0.40929	-0.54172	0.058832	0.76917	0.69632	0.38939	-0.0265	0.48043	-0.36493	0.2776	0.15887	0.21809	0.19744	-0.49303	0.43764	-0.34335	....	0.030099	-0.02753	SPAM
26	-1.4376	-4.9513	-0.77893	-0.14381	-0.08607	-1.4306	0.017725	-3.7837	-0.61233	0.63638	-0.919	-0.12598	-0.54764	-1.0049	-0.38313	0.27585	1.8473	-1.3286	....	-0.4022	-0.33479	SPAM
27	-3.6128	-5.2949	-0.13477	-0.22275	-0.53963	-0.26378	0.24555	-2.5214	-0.40917	0.47241	-0.60804	0.1439	0.43255	-0.31248	-0.19502	-0.69832	0.21787	-0.44712	....	0.13138	0.009883	SPAM
28	-4.3108	-5.0202	0.081884	-0.08396	-0.14335	-0.27251	0.483	-0.82443	-0.47002	0.27959	-0.62503	0.22429	0.1006	-0.28543	-0.19018	-0.27047	0.12921	-0.48387	....	0.28189	0.018439	SPAM
29	1.5834	-5.2623	-1.4693	-1.058	-1.432	-1.2798	-1.4686	-6.6508	0.46694	1.7674	-0.82008	0.38771	2.6731	-0.57926	0.57229	-1.9351	0.13664	-2.6891	....	1.1241	0.07158	SPAM
30	-3.201	-5.3349	0.26691	-0.51071	0.24445	-0.09418	0.31314	-0.2508	-0.7364	0.9326	-0.06678	0.44033	1.1331	0.12407	-0.09074	0.025757	-0.37106	-1.2191	....	0.15244	-0.21455	SPAM
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	....	...	...	...
2891	-4.6485	-5.2662	0.2292	-0.394	0.39162	0.061057	0.65182	-0.29565	-0.35942	-0.07444	-0.499	-0.1187	-0.37528	-0.17668	0.028861	0.096827	0.068307	-0.17783	0.62615	0.064353	-0.10587	SPAM
2892	-2.041	-5.102	0.15648	0.008924	-0.85599	0.11312	0.37638	-0.47619	-0.68499	1.6218	0.26025	0.18524	1.033	0.29699	-0.06113	-0.63888	0.058338	-0.30624	0.21981	0.61351	-0.08681	SPAM
2893	1.6879	-5.1034	-0.98106	-1.2811	-1.6815	-1.2755	-1.8988	-5.1283	0.80625	1.5771	0.68587	1.0817	2.8378	-1.4144	0.67841	-1.7361	0.1523	-2.6756	1.8194	-0.64846	-0.30943	SPAM

ตารางผลการวิเคราะห์จดหมายข่าวขยะจากตัวแปรที่ผ่านการวิเคราะห์องค์ประกอบหลัก

Test No	EigenValue	Dimention	BAYES	C4.5	KNN (K=1)	KNN (K=3)	KNN (K=5)	KNN (K=7)	KNN (K=9)	KNN (K=11)	Develop
1	>=90.0	1	84.3069	85.3094	76.253	81.7525	83.4082	84.3761	84.6699	84.9119	85.9661
2	>=50.0	2	86.8994	96.2669	95.7829	96.2737	96.1632	96.3705	96.5156	96.6607	81.4015
3	>=40.0	3	86.7611	97.5112	96.8545	97.1342	97.3038	97.373	96.9269	97.0685	83.1355
4	>=30.0	4	87.0031	97.5804	96.5088	96.7022	97.0619	97.3038	97.0063	97.0719	75.2844
5	>=21.0	5	86.7957	97.4075	96.5779	96.9894	96.9927	97.2001	96.9651	96.9857	82.9943
6	>=17.0	6	86.8303	96.9582	96.7853	97.3660	97.2693	97.1656	97.2071	97.1310	79.915
7	>=13.0	7	87.0031	97.0964	96.6125	97.1725	97.3038	97.0964	97.0549	96.9857	100
8	>=12.0	8	87.3833	97.7186	97.7532	97.9884	98.168	97.8223	97.6288	97.5907	92.6429
9	>=8.0	10	87.4179	97.8223	97.8915	97.8673	97.8569	97.9606	97.9434	97.9192	87.4216
10	>=7.3	11	87.1414	97.7186	97.5804	97.6738	97.8915	97.8569	97.9088	97.8985	77.5639
11	>=7.0	12	87.2105	97.8569	97.7878	97.8879	97.8915	97.8915	97.8948	97.9087	85.0654
12	>=6.5	14	87.3142	98.0643	97.4075	97.7602	97.6495	97.7532	97.6979	97.5666	81.8506
13	>=6.0	15	87.3142	98.0297	97.6841	97.8950	97.7878	97.8569	97.7635	97.7185	87.1811
14	>=5.7	16	87.487	97.9952	97.5804	97.7739	97.7186	97.7532	97.8155	97.7497	79.3353
15	>=5.3	17	87.5216	97.926	97.3384	97.7324	97.5458	97.7532	97.5320	97.6114	82.8551
16	>=5.2	18	87.5562	97.926	97.4075	97.7082	97.6841	97.4767	97.4661	97.4489	96.6817
17	>=5.0	19	87.729	97.8915	97.6495	97.7637	97.5458	97.4767	97.4628	97.4317	89.3889
18	>=4.0	26	86.934	97.8915	97.4767	97.6391	97.7878	97.6149	97.6460	97.5527	79.9878
19	>=3.0	36	87.3142	97.6841	96.9927	97.1863	97.0964	96.8199	97.1344	97.1897	91.1903
20	>=2.0	56	87.8673	97.5112	96.6816	96.8789	97.0964	96.6816	96.6921	96.7508	87.3848
21	>=1.5	75	88.1092	97.5112	95.8521	95.9318	95.9558	95.7138	95.6930	95.5547	80.2964
22	>=1.0	114	88.6623	97.7878	95.4718	95.1089	94.9879	94.6077	94.2172	94.1168	89.7713
23	>=0.9	127	89.6993	97.7878	94.4694	94.8152	94.2966	94.0892	94.1654	93.9060	96.614
24	>=0.8456	135	89.9412	97.7878	94.0892	94.9152	93.709	93.9509	93.8440	93.8542	90.0135
25	>=0.81	140	88.4895	97.7878	94.0892	94.6079	94.2966	94.1583	94.1725	94.0618	94.0552
26	>=0.8	142	87.7636	97.7878	94.3311	94.7566	94.3311	93.8472	94.1931	94.0928	86.1013
27	>=0.7194	158	83.8340	97.8259	94.6252	94.2968	94.4596	94.2313	94.2105	93.9304	84.2738
28	>=0.7	161	81.2997	97.7878	94.3311	94.3730	94.5385	93.9509	94.1277	93.7717	83.306
29	>=0.6	183	76.495	97.7878	93.7781	93.6781	93.7435	93.3979	93.3911	93.2874	90.0121
30	>=0.5	210	71.4138	97.7878	93.1559	93.3221	93.1213	93.1213	93.0247	92.7966	95.1614
31	>=0.4	248	65.261	97.8223	92.1535	92.7691	92.7065	92.9485	92.7312	92.9314	91.9466
32	>=0.3	302	62.6339	97.7532	91.4967	92.1125	91.9461	91.2893	91.2790	91.2132	90.1484
		Correct(%)	84.8560	97.2932	95.3265	95.7448	95.7286	95.6534	95.6244	95.5834	86.8421
		max	89.9412	98.0643	97.8915	97.9884	98.168	97.9606	97.9434	97.9192	100.0000

ตารางเวลาดำเนินงานวิเคราะห์จดหมายข่าวขยะ

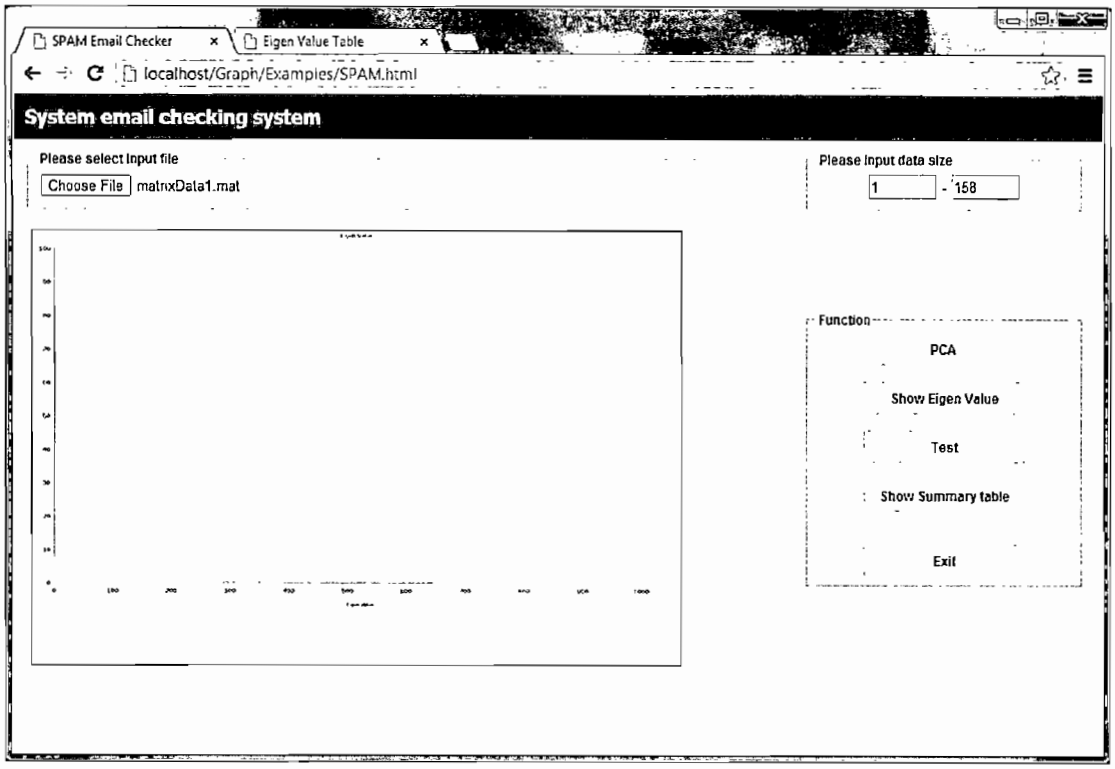
Test No	EigenValue	Dimention	BAYES		C4.5		KNN (K=1)		KNN (K=3)		KNN (K=5)		KNN (K=7)		KNN (K=9)		KNN (K=11)	
			Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
1	>=90.0	1	0.235	0.185	0.594	0.079	0.078	5.025	0.016	4.884	0.016	4.821	0.016	4.881	0.031	4.898	0.03	4.914
2	>=50.0	2	0.202	0.061	1.154	0.015	0.046	5.085	0.062	5.179	0.109	5.227	0.015	5.449	0.032	5.412	0.031	5.724
3	>=40.0	3	0.185	0.063	1.733	0.015	0.063	5.127	0.03	5.351	0.016	5.35	0.032	5.725	0.047	5.771	0.031	5.865
4	>=30.0	4	0.486	0.125	2.259	0.064	0.032	5.601	0.062	5.978	0.093	6.165	0.063	6.443	0.063	6.474	0.016	6.795
5	>=21.0	5	0.467	0.078	2.71	0	0.032	5.863	0.046	6.597	0.048	6.647	0.047	6.851	0.047	6.95	0.06	7.023
6	>=17.0	6	0.558	0.076	3.268	0.048	0.078	6.282	0.03	6.928	0.062	7.12	0.048	7.717	0.016	7.785	0.016	8.079
7	>=13.0	7	0.591	0.094	3.834	0	0.015	7.042	0.015	7.91	0.032	8.145	0.031	8.561	0	8.643	0.047	9.126
8	>=12.0	8	0.669	0.109	4.532	0.048	0.045	7.272	0.032	8.096	0.016	8.32	0.045	9.05	0	9.135	0.045	9.217
9	>=8.0	10	0.857	0.155	5.6	0.046	0.045	8.796	0.047	10.519	0	11.089	0.031	11.759	0.032	12.163	0.016	12.425
10	>=7.3	11	0.886	0.171	6.079	0.076	0.015	9.343	0.031	10.683	0.016	11.412	0.03	12.092	0.046	12.736	0.015	13.142
11	>=7.0	12	1.013	0.202	6.794	0.016	0.047	9.911	0.031	11.998	0	12.832	0.048	13.291	0.047	14.115	0.015	14.364
12	>=6.5	14	1.139	0.154	7.582	0.047	0.047	11.215	0.031	13.524	0.047	14.71	0.047	15.163	0.015	16.049	0.047	16.549
13	>=6.0	15	1.355	0.171	7.9	0.03	0.046	12.416	0.077	14.599	0.047	15.673	0.015	16.629	0.063	16.959	0.046	17.646
14	>=5.7	16	1.323	0.168	8.51	0.016	0.062	13.257	0	16.38	0.032	17.476	0.032	18.077	0.032	18.967	0.031	19.686
15	>=5.3	17	1.542	0.199	9.373	0.031	0.077	13.997	0.061	16.805	0.031	17.983	0.032	18.967	0.015	19.766	0.047	20.522
16	>=5.2	18	1.967	0.223	9.607	0.016	0	14.429	0.094	17.396	0.031	18.713	0.094	19.831	0.047	20.669	0.046	21.092
17	>=5.0	19	1.598	0.297	10.52	0.032	0.015	15.109	0	18.424	0.015	19.719	0.015	20.79	0.016	21.464	0.048	22.29
18	>=4.0	26	2.113	0.372	13.972	0	0.016	21.633	0.032	26.427	0.016	28.424	0.031	30.19	0.062	31.24	0.03	32.144
19	>=3.0	36	3.155	0.453	19.285	0.031	0.015	31.718	0.016	38.76	0.032	41.279	0.031	43.604	0	44.991	0.061	46.087
20	>=2.0	56	4.84	0.714	29.943	0.031	0	57.548	0.062	69.497	0.063	72.644	0.079	75.725	0.046	77.606	0.047	79.268
21	>=1.5	75	6.66	0.912	40.224	0.091	0.016	84.126	0.015	99.47	0.046	102.73	0.031	106.642	0.078	108.607	0.047	110.34
22	>=1.0	114	10.227	1.463	61.398	0.092	0.062	135.707	0.015	158.637	0.031	163.791	0.061	168.093	0.047	170.976	0.063	172.977

ตารางเวลาดำเนินงานวิเคราะห์จดหมายข่าวขยะ

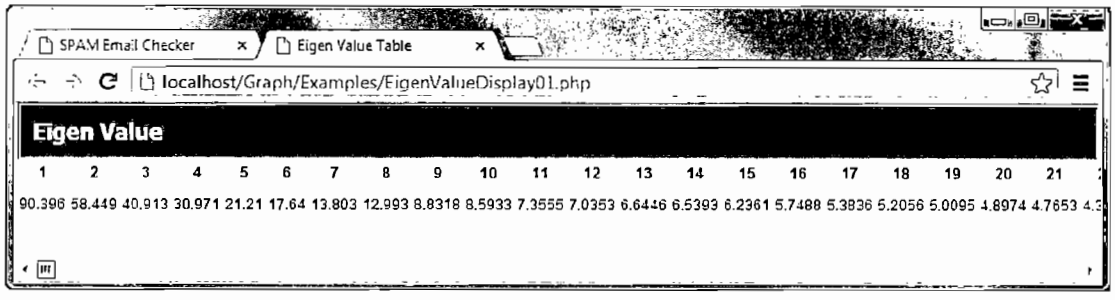
Test No	EigenValue	Dimention	BAYES		C4.5		KNN (K=1)		KNN (K=3)		KNN (K=5)		KNN (K=7)		KNN (K=9)		KNN (K=11)	
			Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
23	>=0.9	127	11.284	1.572	69.646	0.031	0.016	155.851	0.016	181.541	0.077	191.955	0.016	192.525	0.03	194.518	0.079	196.564
24	>=0.8456	135	12.063	1.621	73.667	0.015	0.016	169.875	0.048	197.477	0	203.678	0.032	208.795	0.016	210.411	0.048	213.26
25	>=0.81	140	12.845	1.645	76.485	0.048	0.015	177.762	0.062	205.871	0.047	212.275	0.016	216.722	0.03	217.834	0.091	222.196
26	>=0.8	142	13.159	1.672	77.98	0.094	0.048	180.497	0.015	206.962	0.032	214.869	0.092	218.901	0.047	220.381	0.031	224.959
27	>=0.7194	158	14.608	1.883	86.429	0.032	0.016	200.869	0.046	228.485	0.063	237.904	0.031	242.588	0.015	244.987	0.095	248.983
28	>=0.7	161	14.789	2.089	88.274	0.078	0.046	204.488	0.032	233.09	0	243.113	0.047	247.398	0.015	250.191	0.047	252.757
29	>=0.6	183	16.879	2.261	107.026	0.031	0.032	239.35	0	271.828	0.047	282.085	0.047	287.42	0.031	289.103	0.031	293.341
30	>=0.5	210	19.561	2.445	124.034	0.016	0.064	280.894	0.046	317.962	0.045	338.129	0.047	334.158	0.062	335.4	0.064	339.514
31	>=0.4	248	23.306	3.087	148.975	0.030	0.03	347.919	0.046	391.441	0.06	409.02	0.062	406.59	0.032	408.647	0.063	413.509
32	>=0.3	302	28.841	3.689	185.107	0.109	0.047	439.251	0.046	490.745	0.108	517.572	0.079	505.701	0.047	506.876	0.03	511.464
		Average	6.5438	0.8878	40.4529	0.0409	0.0369	90.1018	0.0363	103.1076	0.0399	107.8397	0.0420	108.9478	0.0346	109.9914	0.0442	111.6194

ภาคผนวก ข

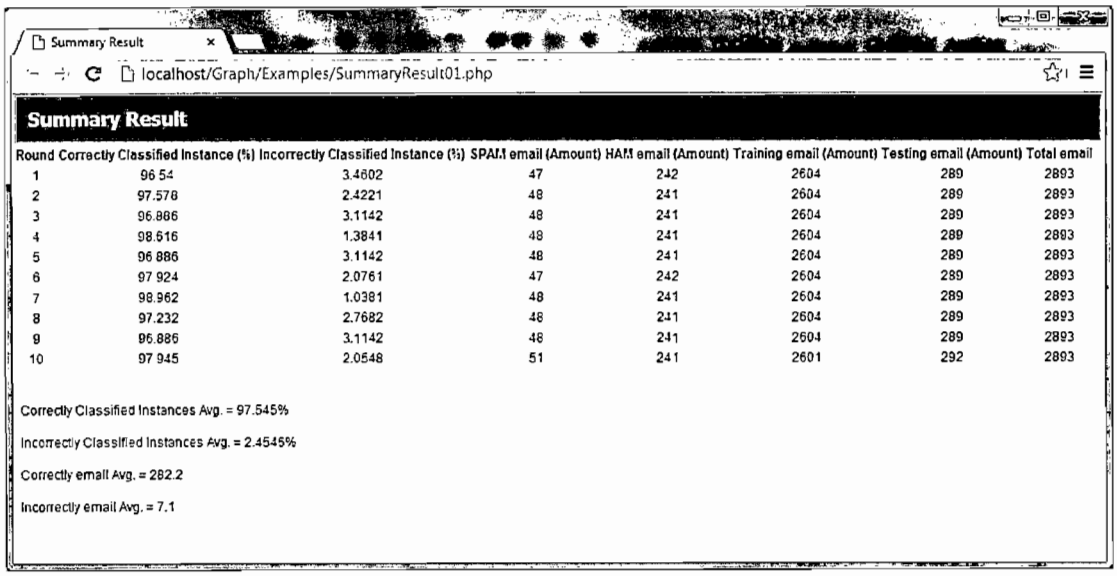
ต้นแบบโปรแกรมจำแนกจดหมายข่าวขยะ



รูปภาพที่ ข-1 หน้าจอแสดงโปรแกรม



รูปภาพที่ ข-2 หน้าจอแสดงค่าไอเกน



รูปภาพที่ ข-3 หน้าจอแสดงผลการทดสอบจดหมายข่าวขยะ