การจัดหมวดหมู่เอกสารภาษาไทยด้วยเครือข่ายฟังก์ชันฐานรัศมี

Thai Document Categorization with Radial Basis Function Network

นิเวศ จิระวิชิตชัย [1] ปริญญา สงวนสัตย์ [2] และพยุง มีสัจ[3]

[1]ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

[2]คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยรังสิต

[3]ภาควิชาครุศาสตร์ไฟฟ้า คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Emails: nivet99@hotmail.com, sanguansat@yahoo.com, pym@kmutnb.ac.th

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการจัดหมวดหมู่เอกสารภาษาไทย ด้วยเครือข่ายฟังก์ชันฐานรัศมี (Radial Basis Function Network) โดยใช้วิธีการลดคุณลักษณะร่วมกับอัลกอริทึมเครื่องจักรการเรียนรู้ จากการทดลองพบว่าเมื่อลดคุณลักษณะด้วยค่าสถิติไคสแควร์(Chisquare) และเรียนรู้ด้วยเครือข่ายฟังก์ชันฐานรัศมีให้ค่าความถูกต้อง (Accuracy) เท่ากับ 85.64 % และค่า F-Measure เท่ากับ 85.80 % ซึ่งการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ สามารถลดขนาดของข้อมูลลงถึง 93.11% โดยไม่ส่งผลต่อประสิทธิภาพในการจัดหมู่เอกสารแต่อย่างใด

คำสำคัญ: การลดมิติข้อมูล ,เครื่องจักรการเรียนรู้, การจำแนกประเภทเอกสาร

Abstract

This research presented Thai document categorization with Radial Basis Function Network. The feature reduction of the document is done before processing by machine learning. The experimental results showed that reducing the feature by Chisquare method and process with Radial Basis Function Network yielded a very high classification with the Accuracy equal to *85.64* % *and F-Measure equal* to *85*.80 % *Data reduction with Chisquare to reduce the size of the data to 93.11%.*

Keywords*: feature reduction, machine learning, text classification*

1. บทนำ

ปัจจุบันการขยายตัวด้านการใช้งานระบบคอมพิวเตอร์ และอินเทอร์เน็ตตลอดช่วงระยะเวลาที่ผ่านมา มีแนวโน้มในการใช้งานเพิ่มมากขึ้นอย่างรวดเร็ว ส่งผลให้เกิดการสร้างและเก็บข้อมูลหลากหลายชนิดในรูปแบบอิเล็กทรอนิกส์มากมาย ซึ่งหนึ่งในข้อมูลอิเล็กทรอนิกส์เหล่านี้คือ ข้อมูลประเภทเอกสาร เช่น จดหมายอิเล็กทรอนิกส์ เว็บเพจ เอกสารข่าว ไฟล์งานเอกสารต่าง ๆ ซึ่งเป็นข้อมูลที่มีปริมาณและเนื้อหาที่หลากหลายมากขึ้น ทำให้ยากต่อการค้นหาและจัดเก็บหมวดหมู่เอกสาร ดังนั้นการสืบค้นและการจัดการเอกสารจะง่าย และเป็นไปตามความต้องการ ต้องอาศัยการจัดแบ่งเอกสารเป็นกลุ่ม หรือหมวดหมู่ให้สอดคล้องและตรงกับดัชนี เพื่อให้จัดเก็บและค้นคืนเอกสารได้อย่างรวดเร็วและมีประสิทธิภาพ จึงมีความจำเป็นต้องอาศัยผู้เชี่ยวชาญในการจัดกลุ่มข้อมูล ฉะนั้นจึงเป็นการยากในการที่จะจัดกลุ่มหรือแยกประเภทเอกสาร ยิ่งหากเอกสารมีปริมาณมากขึ้นทุก ๆ วัน ซึ่งต้องพึ่งพาทรัพยากรบุคคลในการจัดหมวดหมู่เอกสารเหล่านี้มากตามไปด้วยเช่นกัน ทำให้มีการคิดค้นพัฒนากระบวนการ ในการจัดหมวดข้อมูลที่มีขนาดใหญ่เหล่านี้ให้เป็นไปแบบอัตโนมัติ เพื่อที่จะสามารถจำแนกกลุ่มข้อมูล เพื่อใช้ประโยชน์จากข้อมูลและการจัดการกับข้อมูลให้มีประสิทธิภาพ รองรับการสืบค้นจากผู้ใช้งานเอกสารอย่างถูกต้องและเหมาะสม [1,2]

เนื่องจากงานวิจัยด้านการหมวดหมู่เอกสารส่วนใหญ่ ในปัจจุบันมุ่งเน้นไปในการหมวดหมู่เอกสารภาษาอังกฤษ ยังขาดงานวิจัยด้านการจัดหมวดหมู่เอกสารภาษาไทย จากความสำคัญของปัญหาดังกล่าว ผู้วิจัยจึงมีแนวคิดที่จะพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทย โดยนำเสนอแบบจำลองการจัดหมวดหมู่ของเอกสารภาษาไทย โดยงานวิจัยนี้ได้นำเสนอวิธีการลดคุณลักษณะของเอกสาร ด้วยค่าสถิติไคแสควร์ (Chisquare) [3] และทำการทดสอบประสิทธิภาพการจำแนกประเภทของเอกสารภาษาไทยด้วยวิธีวิธีเครือข่ายฟังก์ชันฐานรัศมี ในบทความนี้ประกอบด้วยส่วนต่าง ๆ ดังนี้ ส่วนที่ 2 กล่าวถึงทฤษฎีที่เกี่ยวข้อง ส่วนที่ 3 กล่าวถึงวิธีการวิจัย ส่วนที่ 4 ผลการทดลอง และส่วนที่ 5 สรุปผลและข้อเสนอแนะ

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การตัดคำ (Word Segmentation)

การประมวลผลจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิภาพนั้นมีปัญหาเบื้องต้นคือ การตัดคำในภาษาไทย ซึ่งลักษณะการเขียนในภาษาไทยนั้นจะมีการเขียนติดต่อกันเป็นสายอักขระ โดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำ ดังเช่นภาษาอังกฤษซึ่งใช้ช่องว่าง(Space) คั่นระหว่างคำ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่เพื่อแบ่งสายอักขระไทยออกเป็นคำ

จึงได้มีผู้คิดค้นพัฒนาวิธีการตัดคำแบ่งได้เป็น หลักการตัดคำโดยใช้กฎ (Rule Base Approach) หลักการตัดคำโดยใช้อัลกอริทึม (Algorithm Approach) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary Approach) และหลักการตัดคำโดยใช้คลังข้อมูล (Corpus Base Approach) แต่ละวิธีการต่างๆ ก็ให้ผลในด้านความถูกต้อง ความรวดเร็วของการทำงานและปริมาณการใช้ทรัพยากรที่แตกต่างกัน จากการศึกษาเปรียบเทียบประสิทธิภาพวิธีดังกล่าว พบว่าวิธีตัดคำที่เหมาะสมกับการจัดหมวดหมู่เอกสารคือ วิธีการตัดคำแบบยาวที่สุด (Longest Matching) ซึ่งมีวิธีการตรวจสอบสายอักขระ (String) ที่เข้ามาจากซ้ายไปขวากับพยางค์ที่เก็บไว้ในพจนานุกรม ในกรณีที่ตรวจสอบแล้วปรากฏว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม ก็ให้เลือกแบ่งพยางค์โดยเลือกพยางค์ที่ยาวที่สุด แล้วทำต่อไปเรื่อย ๆ จนจบสายอักขระ [4]

2.2 การกำจัดคำหยุด (Stop-Word List Removal)

เป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลงคำที่ไม่มีนัยสำคัญ ในที่นี้หมายถึง คำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้ว ไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ตัวอย่างเช่น คำบุพบท เป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน คำสันธานเป็นคำที่ทำหน้าที่เชื่อมคำกับคำ เป็นต้น คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสาร และปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืน หรือการจำแนกหมวดหมู่ ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการเพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์และ ลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล [5-6]

2.3 การหารากศัพท์ (Stemming)

เป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลงและ เพิ่มประสิทธิภาพในการค้นคืน หรือ การจำแนกหมวดหมู่ การหารากศัพท์ของคำภาษาไทยนั้น จะใช้วิธีการรวบรวมคำศัพท์ที่มีความหมายคล้ายกัน หรือมีรากศัพท์เดียวกันไว้เป็นรายการคำศัพท์ หรือจัดเก็บในคลังคำ เพื่อใช้ในการเปรียบเทียบหารากศัพท์[5-6]

2.4 การสกัดคุณลักษณะ (Feature Extraction)

การสกัดคุณลักษณะเอกสาร คือการดึงคุณลักษณะ (Feature) ของเอกสารออกมา ซึ่งการดึงคุณลักษณะออกมานั้น ต้องกำหนดก่อนว่าจะใช้อะไรเป็นตัวแทนคุณลักษณะของเอกสาร และใช้ค่าใดแทนคุณลักษณะเอกสารนั้น จากการสำรวจงานวิจัยที่ผ่านมาทั้งในประเทศและต่างประเทศพบว่าส่วนใหญ่จะใช้คำเป็นตัวแทนคุณลักษณะของเอกสารและใช้ค่าความถี่ของคำเป็นค่าของคุณลักษณะ นอกจากการใช้คำเดี่ยวแล้วยังสามารถใช้ วลี หรือกลุ่มของคำ ประโยค แทนคุณลักษณะของเอกสารได้เช่นกัน ตัวแทนคุณลักษณะของเอกสารที่นิยมใช้ในการจัดหมวดหมู่เอกสารข้อความคือ ถุงคำ (Bag of words) ซึ่งจะเก็บอยู่ในรูปแบบของเวกเตอร์ ในงานวิจัยนี้ใช้คุณลักษณะแบบคำเดียว (Single word) [7-8]

2.5 การสร้างดัชนี (indexing)

เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบ ที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสารเรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสารสำหรับใช้ในกระบวนการเรียนรู้ วัตถุประสงค์ของการสร้างดัชนีคือ การคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร หรืออาจจะเรียกได้ว่าการหาค่าน้ำหนัก (term weighting) การสร้างดัชนี โดยทั่วไปที่นิยมใช้กัน จะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม ซึ่งงานวิจัยนี้ใช้การสร้างดัชนีจากสูตร tf-idf weighting [5-8]

$$\log(f_{ik}+1) * \log(\frac{N}{n_i})$$

f_{ik} = เป็นความถี่ของคำ i ในเอกสาร k

N = จำนวนเอกสารทั้งหมดรวมของทุกกลุ่ม

n_i = จำนวนเอกสารทั้งหมดที่มีคำ i เกิดขึ้น

2.6 การเลือกคุณลักษณะ (Feature Selection)

จากที่กล่าวมาจะพบว่าเอกสารนั้น มีแนวโน้มที่จะเพิ่มปริมาณสูงขึ้นทุกวัน ทำให้เอกสารที่ทำการทดสอบ มีจำนวนคุณลักษณะที่สกัดได้จำนวนมาก ซึ่งจำนวนคุณลักษณะที่มากนั้น ส่งผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่โดยทั่วไป ไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ และจำนวนคุณลักษณะที่มากส่งผลถึงการใช้ทรัพยากรและหน่วยความจำของระบบจำนวนมากที่ใช้ในการเรียนรู้ ซึ่งงานวิจัยนี้ใช้ค่าสถิติไคสแควร์ เป็นตัวเลือกคุณลักษณะในการจัดหมวดหมู่เอกสาร โดยค่าสถิติไคสแควร์ ได้จากการทำนายจากกลุ่มเอกสาร โดยการดูจากการมีอยู่หรือไม่มีอยู่ของคำในเอกสารที่ทำการเรียนรู้ [3,9]

2.7 ค่าสถิติไคสแควร์ (Chisquare)

ค่าสถิติไคสแควร์ เป็นการวิเคราะห์ความความสัมพันธ์ระหว่างคำที่ปรากฏในเอกสารและกลุ่มเอกสาร โดยค่าสถิติไคสแคว์ที่ใช้ในการพิจารณาคือค่าเฉลี่ย กับค่าสูงสุด สามารถนิยามได้โดย [3]

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

A คือ จำนวนเอกสารจากกลุ่ม c ซึ่งมีคำ t

B คือ จำนวนเอกสารซึ่งมีคำ t แต่ไม่ได้อยู่ในกลุ่ม c

C คือ คือ จำนวนเอกสารจากกลุ่ม c ซึ่งไม่มีคำ t

D คือ จำนวนเอกสารที่ไม่ได้อยู่ในกลุ่ม c และไม่มีคำ t

N คือ จำนวนเอกสารทั้งหมด

$A = \#(t,c)$	$C = \#(\neg t,c)$
$B = \#(t,\neg c)$	$D = \#(\neg t, \neg c)$

ภาพที่ 1 ตารางสถิติไคสแควร์

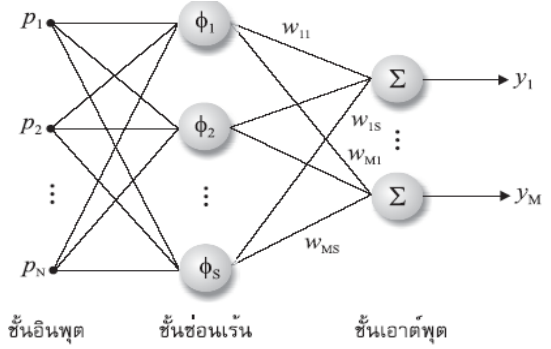
2.8 อัลกอริทึมเครือข่ายฟังก์ชันฐานรัศมี (Radial Basis Function Network)

เครือข่ายฟังก์ชันฐานรัศมีจัดเป็นเครือข่ายไปข้างหน้าประเภทหนึ่ง ที่ได้รับการยอมรับว่ามีประสิทธิภาพสูงเครือข่ายหนึ่ง โดยเครือข่าย Radial Basis Function Network (RBF) แตกต่างไปจากเครือข่ายเพอร์เซ็พตรอนแบบหลายชั้น (multi-layer perceptron) ตรงที่เครือข่าย RBF นั้นมีชั้นซ่อนเร้นเพียงชั้นเดียว เครือข่ายฟังก์ชันฐานรัศมีสามารถพิจารณาเป็นฟังก์ชันการส่ง (mapping function) ของความสัมพันธ์ระหว่างคู่รูปแบบอินพุตและเอาต์พุตได้ โดยการเรียนรู้ของเครือข่ายเป็นการปรับค่าน้ำหนักประสาทให้ได้ฟังก์ชันการส่งที่เหมาะที่สุด ดังนั้นเราสามารถกล่าวได้ว่าเครือข่าย RBF คือกระบวนการปรับเส้นโค้ง (curve fitting) ระหว่างข้อมูลอินพุตกับเอาต์พุต สถาปัตยกรรมของเครือข่าย RBF ประกอบไปด้วยชั้นของนิวรอน 3 ชั้น ดังนี้ [10]

ชั้นอินพุต แต่ละอินพุตจะแทนคุณลักษณะของเวกเตอร์อินพุต เหมือนกับในเครือข่ายเพอร์เซ็พตรอนแบบหลายชั้นทั่วๆไป ในที่นี้เวกเตอร์อินพุตมีขนาดเท่ากับ N

ชั้นซ่อนเร้น แต่ละนิวรอนในชั้นซ่อนเร้นจะมีฟังก์ชันถ่ายโอนซึ่งมีลักษณะพิเศษ ที่ซึ่งให้ผลตอบสนองของฟังก์ชันที่ขึ้นอยู่กับระยะห่างระหว่างอินพุตกับจุดศูนย์กลางของฟังก์ชัน กล่าวคือถ้าเวกเตอร์อินพุตอยู่ใกล้จุดศูนย์กลางมาก เอาต์พุตที่ได้จะมาก ถ้าเวกเตอร์อินพุตอยู่ห่างออกจากจุดศูนย์กลาง เอาต์พุตที่ได้จะลดลงตามลำดับ ในที่นี้จำนวนนิวรอนในชั้นซ่อนเร้นมีขนาดเท่ากับ S

ชั้นเอาต์พุต มีหน้าที่รวมเอาต์พุตที่ได้จากแต่ละนิวรอนในชั้นซ่อนเร้น เครือข่ายให้เอาต์พุตในรูปของเวกเตอร์ขนาดเท่ากับ M



ภาพที่ 2 เครือข่าย Radial Basis Function Network

ดังนั้นเราสามารถพิจารณาเครือข่าย RBF เป็นการฟังก์ชันการส่งระหว่างปริภูมิของอินพุต $p \in \mathbf{R}^{N\times 1}$ ไปยังปริภูมิของเอาต์พุต $y \in \mathbf{R}^{M\times 1}$ ได้ จากเครือข่าย RBF ในรูปข้างต้น จะได้ว่าเอาต์พุตของเครือข่ายมีค่าเท่ากับ

$$y_i = \sum_{k=1}^{S} w_{ik}\phi_k(p, c_k)$$

$$= \sum_{k=1}^{S} w_{ik}\phi_k(\| p, c_k \| 2)$$

โดยที่ $\phi_k(.)$ = ฟังก์ชันส่งค่าจาก $\mathbf{R}^{+}$ ไปยัง $\mathbf{R}$

$\| . \| 2$ = ฟังก์ชันระยะทางแบบยุคลิด

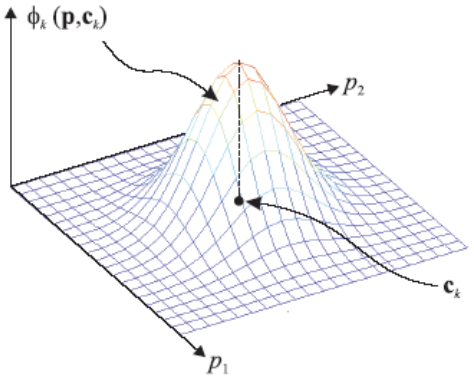
w_{ik} = ค่าน้ำหนักประสาทในชั้นซ่อนเร้น

S = จำนวนนิวรอนในชั้นซ่อนเร้น

$c_k \in \mathbf{R}^{N\times 1}$ = เวกเตอร์จุดศูนย์กลางของ RBF ในปริภูมิของเวกเตอร์อินพุต

สำหรับแต่ละนิวรอนในชั้นซ่อนเร้น ค่าระยะทางยุคลิดระหว่างเวกเตอร์จุดศูนย์กลาง c_k กับเวกเตอร์อินพุต p จะถูกคำนวณ เอาต์พุตของนิวรอนในชั้นซ่อนเร้นนี้จะได้จากฟังก์ชัน $\phi_k(.)$ ซึ่งเป็นฟังก์ชันแบบไม่เป็นเชิงเส้น สุดท้ายแล้วเอาต์พุตของเครือข่ายจะได้จากผลรวมของค่าน้ำหนักประสาทกับเอาต์พุตของนิวรอนจากชั้นซ่อนเร้น โดยฟังก์ชัน $\phi_k(.)$ ที่ใช้ในเครือข่าย RBFในงานวิจัยนี้คือ ฟังก์ชันเกาส์เซียน

$$\phi_k(p) = e^{-p^2/\sigma^2}$$



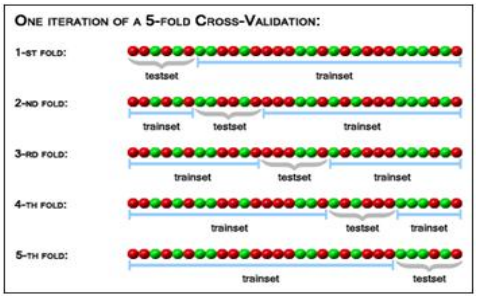
ภาพที่ 3 ฟังก์ชันฐานรัศมีแบบเกาส์เซียน

2.9 การวัดประสิทธิภาพ

การทดสอบเพื่อเปรียบเทียบประสิทธิภาพในการจำแนกหมวดหมู่เอกสารภาษาไทยนั้น สามารถวัดที่ประสิทธิภาพของการจำแนกข้อมูล ตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) และค่า F-Measure ซึ่งคำนวณได้ดังตาราง Confusion matrix ดังภาพ [8]

		actual value		
		p	*n*	total
prediction outcome	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

ภาพที่ 4 Confusion matrix



ภาพที่ 5 K - fold cross-validation

TP = จำนวนตัวอย่างที่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าอยู่กลุ่ม C_j

FP = จำนวนตัวอย่างที่ไม่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าอยู่กลุ่ม C_j

FN = จำนวนตัวอย่างที่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่กลุ่ม C_j

TN = จำนวนตัวอย่างที่ไม่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่กลุ่ม C_j

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

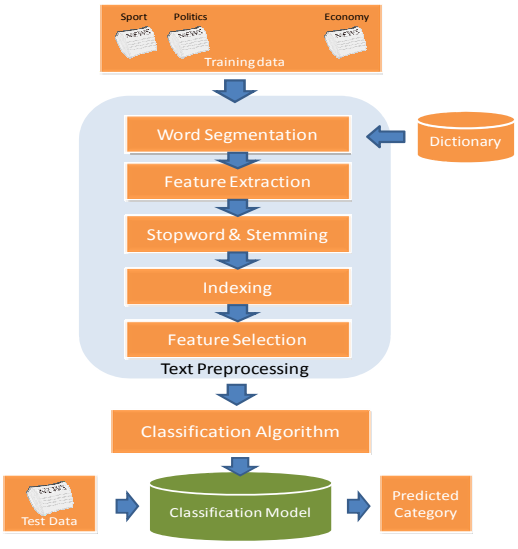
2.10 ครอสวาลิเดชั่น (Cross Validation)

เป็นวิธีการในการคาดการณ์ค่าความผิดพลาดของโมเดลที่นำเสนอ โดยพื้นฐานของวิธีการครอสวาลิเดชั่นคือ การสุ่มตัวอย่าง (Resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆและ นำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำครอสวาลิเดชั่นมักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (Network architecture), โมเดลในการคัดแยกประเภท(Classification model) เช่นในการจำแนกข้อมูลโดยใช้เทคนิคของ Data mining เช่น Neural Network หรือ Decision Tree นั้น จะต้องมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ แต่ในบางครั้งอาจเกิดปัญหาจากการเลือกข้อมูลที่ดีและ ง่ายมาเป็นข้อมูลชุดทดสอบ ทำให้ผลการจำแนกประเภทนั้นดีเกินจริง ดังนั้นจะมีการคิดวิธี k-fold cross validation ขึ้นมาแก้ปัญหา

โดยการแบ่งข้อมูลแบบ K - fold cross-validation คือการแบ่งข้อมูลออกเป็น K ชุดเท่าๆกัน และทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก K - 1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ดังตัวอย่างต่อไปนี้ K - fold Cross Validation (K = 5) ชุดข้อมูลหลังจากทำการแบ่งออกเป็น 5 ชุดข้อมูลย่อยเท่าๆกัน โดยแต่ละกล่องคือชุดข้อมูลย่อย 1 ชุด ดังภาพ

3.วิธีการดำเนินการวิจัย

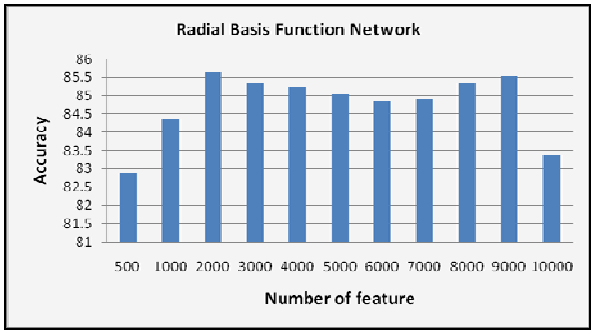
งานวิจัยนี้ทำการทดลองการลดมิติของคุณลักษณะร่วมกับ อัลกอริทึมในการจำแนกประเภทเรียนรู้แบบมีผลเฉลย (Supervised Learning) โดยมุ่งเน้นจำแนกหมวดหมู่เอกสารภาษาไทย โดยทดสอบกับเอกสารประเภทข่าวอิเล็กทรอนิกส์จากหนังสือพิมพ์ไทยรัฐ จำนวน 10 กลุ่ม ได้แก่ กลุ่มการศึกษา กลุ่มบันเทิง กลุ่มสังคม กลุ่มการเมือง กลุ่มเทคโนโลยี กลุ่มกีฬา กลุ่มข่าวต่างประเทศ กลุ่มเกษตร กลุ่มเศรษฐกิจ กลุ่มกรุงเทพ โดยมีจำนวนกลุ่มตัวอย่างทั้งหมด 12000 เอกสาร เป็นกลุ่มตัวอย่างเรียนรู้ และทำการทดสอบด้วยวิธี 10-fold cross validation โดยแบบจำลองที่นำเสนอในงานวิจัยนี้ทำการลดขนาดคุณลักษณะของข้อมูลด้วย ค่าสถิติไคแสควร์ (Chisquare) โดยพิจารณาจากค่าสูงสุด ซึ่งผลที่ได้จากการลดขนาดของมิติดังกล่าว จะถูกส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลยด้วยอัลกอริทึมเครือข่ายฟังก์ชันฐานรัศมี (RBF) ทำการเรียนรู้ ทำการทดสอบเปรียบเทียบประสิทธิภาพด้านความถูกต้อง ความแม่นยำ โดยวัดที่ประสิทธิผลของการจำแนกตามแนวคิดทางด้านการค้นคืนสารสนเทศ



ภาพที่ 6 ขั้นตอนวิธีจำแนกประเภทของเอกสารภาษาไทย

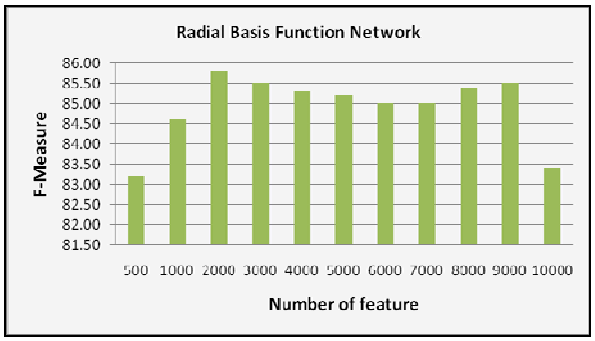
4.ผลการทดลอง

การทดลองลดคุณลักษณะด้วยวิธี Chi-square และเรียนรู้ด้วยอัลกอริทึมเครือข่ายฟังก์ชันฐานรัศมี (RBF) ฟังก์ชันเกาส์เซียน บนแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทย การทดสอบอัลกอริทึมในการสังเคราะห์โมเดลใช้ซอฟต์แวร์ Weka version 3.6 พารามิเตอร์ดั้งเดิม (Default parameter) ทำการเปรียบเทียบประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) และค่า F-Measure ซึ่งได้ผลการทดลองดังนี้



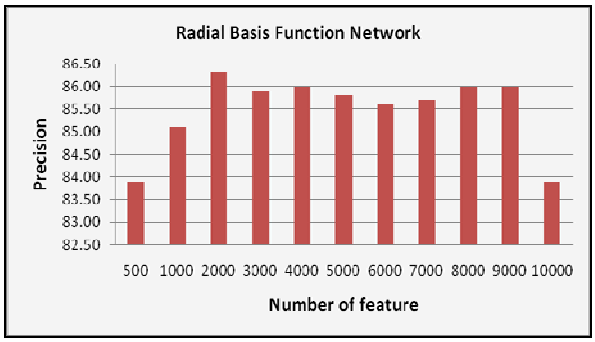
ภาพที่ 7 ผลเปรียบเทียบประสิทธิภาพด้าน Accuracy

ผลทดลองเมื่อสกัดคุณลักษณะคำมาจากกลุ่มเอกสาร ที่ใช้ในการเรียนรู้ได้มาทั้งหมด 29047 คุณลักษณะ เมื่อลดคุณลักษณะด้วย Chi-square ที่ระดับแตกต่างกันและเรียนรู้ด้วยเครือข่ายฟังก์ชันฐานรัศมีพบว่า เมื่อใช้คุณลักษณะของเอกสารจำนวน 10000 คุณลักษณะมาทดสอบให้ค่าความถูกต้อง(Accuracy) เท่ากับ 83.39 % แต่เมื่อลดมิติคุณลักษณะลงเหลือ 2000 คุณลักษณะ ให้ค่าความถูกต้องสูงขึ้นเท่ากับ 85.64 % มากกว่าข้อมูลต้นฉบับที่ยังไม่ลดมิติ โดยสามารถลดมิติของข้อมูลได้ถึง 93.11% (29047-2000 / 29047) โดยไม่กระทบต่อค่าความถูกต้องแต่อย่างใด แต่เมื่อลดมิติลงมากกว่าระดับดังกล่าวพบว่า ค่าความถูกต้อง เริ่มถดถอยลงตามจำนวนของมิติข้อมูลที่ลดลงตามลำดับ

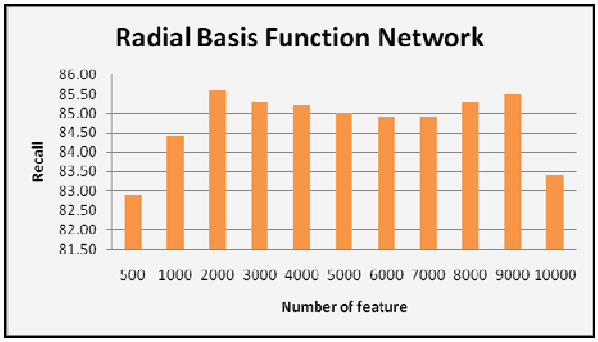


ภาพที่ 8 ผลเปรียบเทียบประสิทธิภาพด้าน F-measure

เมื่อเปรียบเทียบประสิทธิภาพด้าน F-Measure ผลทดลองพบว่า เมื่อใช้คุณลักษณะของเอกสารจำนวน 10000 คุณลักษณะมาทดสอบให้ค่า F-Measure เท่ากับ 83.40 % แต่เมื่อลดมิติคุณลักษณะลงเหลือ 2000 คุณลักษณะ ให้ค่าความถูกต้องสูงขึ้นเท่ากับ 85.80 % มากกว่าข้อมูลต้นฉบับที่ยังไม่ลดมิติ แต่เมื่อลดมิติลงมากกว่าระดับดังกล่าวพบว่า ค่า F-Measure เริ่มถดถอยลงตามจำนวนของมิติข้อมูลที่ลดลงตามลำดับ

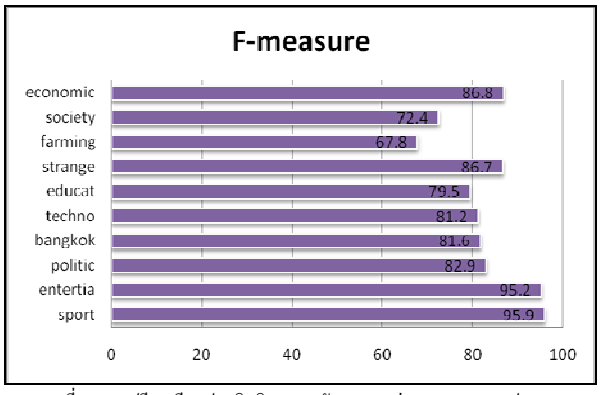


ภาพที่ 9 ผลเปรียบเทียบประสิทธิภาพด้าน Precision



ภาพที่ 10 ผลเปรียบเทียบประสิทธิภาพด้าน Recall

เมื่อเปรียบเทียบประสิทธิภาพด้าน ค่าความแม่นยำ (Precision) และค่าความระลึก(Recall) ผลทดลองพบว่าเป็นไปในทิศทางเดียวกันคือเมื่อใช้คุณลักษณะของเอกสารจำนวน 10000 คุณลักษณะมาทดสอบให้ค่าความแม่นยำเท่ากับ 83.90 % ค่าความระลึกเท่ากับ 83.40 % แต่เมื่อลดมิติคุณลักษณะลงเหลือ 2000 คุณลักษณะ ให้ค่าความแม่นยำสูงขึ้นเท่ากับ 86.30 % ค่าความระลึก 85.60 % มากกว่าข้อมูลต้นฉบับที่ยังไม่ลดมิติ แต่เมื่อลดมิติลงมากกว่าระดับดังกล่าวพบว่า ค่าทั้งสองเริ่มถดถอยลงตามจำนวนของมิติข้อมูลที่ลดลงตามลำดับ



ภาพที่ 11 ผลเปรียบเทียบประสิทธิภาพการจัดหมวดหมู่เอกสารแยกตามประเภท

เมื่อพิจารณาประสิทธิภาพการจัดหมวดหมู่เอกสารแยกตามประเภท ด้วยอัลกอริทึม RBF ที่พารามิเตอร์ 2000 คุณลักษณะ ซึ่งให้ประสิทธิภาพในการจำแนกสูงสุดพบว่า กลุ่มข่าวที่จำแนกถูกต้องมากที่สุดคือ กลุ่มข่าวกีฬา (sport) 95.9% รองลงมาคือกลุ่มข่าวบันเทิง(entertia) 95.2% เศรษฐกิจ (Economic) 86.8% ต่างประเทศ (strange) 86.7% การเมือง(politic) 82.9% กรุงเทพ (Bangkok) 81.6% เทคโนโลยี (techno) 81.2% การศึกษา (educat) 79.5% สังคม (society) 72.4% และการเกษตร (farming)67.8%ตามลำดับ

5.สรุปผลการวิจัย

งานวิจัยนี้ได้เสนอวิธีการลดมิติของข้อมูลและทำการทดสอบประสิทธิภาพการจำแนกหมวดหมู่เอกสาร โดยใช้วิธีการลดคุณลักษณะด้วยค่าสถิติไคแสควร์ (Chisquare)ร่วมกับอัลกอริทึมเครื่องจักรการเรียนรู้ด้วยเครือข่ายฟังก์ชันฐานรัศมี (Radial Basis Function Network) เพื่อศึกษาวิธีการที่มีประสิทธิภาพในการจำแนกการจัดหมวดหมู่เอกสาร จากการทดลองพบว่าการลดคุณลักษณะของข้อมูลด้วยวิธีไคแสควร์ แล้วส่งเข้าเครื่องจักรการเรียนรู้และวัดประสิทธิภาพจากจำนวนมิติของข้อมูลที่ดีที่สุด พบว่าเครือข่ายฟังก์ชันฐานรัศมีที่จำนวน 2000 คุณลักษณะ ให้ค่าความถูกต้อง (Accuracy) เท่ากับ 85.64 % ค่าความแม่นยำ (Precision) เท่ากับ 86.30 % ค่าความระลึก(Recall) เท่ากับ 85.60 % และค่า F-Measure เท่ากับ 85.80 %

ผลการทดลองจากงานวิจัยนี้ พบว่าเราสามารถลดมิติของข้อมูลด้วยค่าสถิติไคแสควร์ (Chisquare) ลงได้มากถึง 93% ขึ้นไป โดยไม่กระทบต่อประสิทธิภาพในการจำแนกหมวดหมู่เอกสารแต่อย่างใด แต่สามารถลดทรัพยากรของระบบและลดระยะเวลาในการประมวลผลได้เป็นอย่างมาก ผลจากการวิจัยนี้สามารถนำขั้นตอนวิธีที่นำเสนอไปประยุกต์ใช้ในการลดมิติของข้อมูลกับฐานข้อมูลอื่นๆ และการสร้างระบบจัดหมวดหมู่เอกสารอัตโนมัติ เช่น การคัดกรองเอกสาร (Document Filtering) การจัดทำดัชนีอัตโนมัติเพื่อใช้ในการค้นคืนเอกสาร (Automatic Indexing for IR System) การจัดหมวดหมู่ของเว็บเพจ (Web Page Classification) เป็นต้น

เอกสารอ้างอิง

[1] Sebastiani, Fabrizio, “Machine Learning in Automated Text Categorization”. *ACM Computing Surveys (CSUR)*, 2002.

[2] Marquez, Llus, “Machine learning and natural language processing” *Technical Report Departament de Llenguatges Sistemes Informatics (LSI)*, Barcelona, Spain, 2000.

[3] นิเวศ จิระวิชิตชัย ปริญญา สงวนสัตย์ และพยุง มีสัจ. “การศึกษาทดลองเทคนิคการลดคุณลักษณะและอัลกอริทึมการจัดหมวดหมู่ของเอกสารภาษาไทย ”. วารสารวิทยาศาสตร์ลาดกระบัง ปี 2553.

[4] Charoenpornsawat ,P “Feature-based Thai Word Segmentation”. *Master's Thesis. Computer Engineering*. Chulalongkorn University, Bangkok, Thailand, 1999.

[5] Jaruskulchai, C. “An Automatic Indexing for Thai Text Retrieval”. *PhD Thesis*, George Washington University. USA, 1998.

[6] ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ.“ข้อมูลคำศัพท์ที่พบบ่อยจากฐานข้อมูลประโยคที่มาจากหนังสือพิมพ์” http://thailang.nectec.or.th/thaichar/word_thai.php

[7] วัลลภ อินทร์ฉ่ำ. ระบบการจัดหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้ SVM ร่วมกับการประมวลผลภาษา,วิทยานิพนธ์วิศวกรรมศาสตร์มหาบัณฑิต มหาวิทยาลัยเกษตรศาสตร์. 2548

[8] นิเวศ จิระวิชิตชัย ปริญญา สงวนสัตย์ และพยุง มีสัจ. “การเปรียบเทียบการคำนวณน้ำหนักดัชนี สำหรับอัลกอริทึมการจัดหมวดหมู่เอกสารภาษาไทย”. วารสารวิทยาศาสตร์ลาดกระบัง ปี 2553.

[9] Yang, Perderson, “A Comparative Study on Feature Selection in Text Categorization”. *Proceedings of ICML-97, 14th International Conference on Machine Learning*, 1997.

[10] อาทิตย์ ศรีแก้ว "ปัญญาเชิงคำนวณ" สาขาวิชาวิศวกรรมไฟฟ้า สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี