ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์ วิศวกรรมคอมพิวเตอร์

สาขา **ภาควิชา**

เรื่อง เทคนิคการเรียนรู้แบบกึ่งมีผู้สอนสำหรับการจำแนกฟังก์ชันของโปรตีน

Semi-Supervised Learning for Protein Function Classification

นามผู้วิจัย นางสาวพัทริกา ณ พิกุล

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(ผู้ช่วยศาสตราจารย์ยอดเยี่ยม ทิพย์สุวรรณ์, Ph.D.)

รักษาราชการแทน หัวหน้าภาควิชา .....

(รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่..... **เดือน**..... **พ.ศ.**.....

วิทยานิพนธ์

เรื่อง

เทคนิคการเรียนรู้แบบกึ่งมีผู้สอนสำหรับการจำแนกฟังก์ชันของโปรตีน

Semi-Supervised Learning for Protein Function Classification

โดย

นางสาวพัทริกา ณ พิกุล

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2554

พัทริกา ณ พิกุล 2554: เทคนิคการเรียนรู้แบบกึ่งมีผู้สอนสำหรับการจำแนกฟังก์ชันของโปรตีน ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D. 52 หน้า

ปัญหาการทำนายฟังก์ชันการทำงานของโปรตีนเป็นปัญหาหลักทางด้านชีวสารสนเทศศาสตร์(Bioinformatics) ที่มีการค้นคว้าอย่างกว้างขวาง การสร้างตัวทำนายฟังก์ชันสำหรับข้อมูลอนุกรมโปรตีน (Protein sequence) ให้สามารถทำนายคลาสได้อย่างมีประสิทธิภาพนั้น จะต้องอาศัยข้อมูลที่เราทราบคลาส (Labeled Data) เป็นจำนวนมาก ซึ่งข้อมูลประเภทนี้มีอยู่อย่างจำกัด ในขณะที่ข้อมูลส่วนใหญ่ที่มี เป็นข้อมูลที่เราไม่ทราบคลาส (Unlabeled Data) ทำให้ได้ผลการทำนายที่ไม่ดีพอเนื่องจากข้อมูลที่เราทราบคลาสนั้นมีจำนวนน้อยเกินไป

งานวิจัยนี้นำเทคนิคการเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) มาแก้ปัญหากรณีข้อมูลที่ทราบคลาสมีจำนวนน้อย โดยนำข้อมูลที่ทราบและไม่ทราบคลาสมาทำการฝึกสอนร่วมกัน ในงานวิจัยนี้นำเสนอเกณฑ์การเลือกข้อมูลที่ใช้ค่าการเทียบเคียงลำดับอนุกรมโปรตีนและสัมประสิทธิ์แจคการ์ดมาเป็นเกณฑ์พิจารณา ในส่วนของเกณฑ์การหยุดสอนนั้น จะพิจารณาจากอัตราการเปลี่ยนแปลงของค่าสัมประสิทธิสหสัมพันธ์กำลังสอง ระหว่างข้อมูลในแต่ละคลาสและข้อมูลที่ถูกเลือกในแต่ละรอบของการฝึกสอน ผลการทดลองบนฐานข้อมูล UniProtKB/Swiss-Prot ที่มีข้อมูลที่ทราบคลาสจำนวน 17,407 ตัวอย่าง และข้อมูลที่ไม่ทราบคลาสจำนวน 47,619 ตัวอย่าง และวัดผลตัวจำแนกด้วยค่าความแม่น, ค่าความเที่ยง, รีคอล และ เอฟเมเชอร์ จากผลการทดลองพบว่าค่าความแม่นเมื่อเทียบกับเทคนิคที่ใช้เกณฑ์การเลือกข้อมูลแจคการ์ดและเกณฑ์การหยุดค่าเฉลี่ยกำลังสองต่ำสุดเพิ่มขึ้น 0.3% ค่าความเที่ยงเพิ่มขึ้น 1.00% ค่ารีคอลดีขึ้น 2.83% และค่าเอฟเมเชอร์ดีขึ้น 1.91%

/ /

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Pattarika Na Pikul 2011: Semi-Supervised Learning for Protein Function Classification. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Kitsana Waiyamai, Ph.D. 52 pages.

Protein function is one of active bioinformatics research topics. To predict protein function with high accuracy, traditional classification techniques require large amount of labeled data for training. Unfortunately, labeled data is very hard to obtain, while unlabeled data is abundant.

In this research, we develop a semi-supervised learning technique for protein function classification. Pairwise alignment and Jaccard coefficient are used for composing data selection criteria, while Square Correlation between objects in a cluster and selected data in each training round is considered as stopping criterion for a self-training algorithm. Experimental results using UniProtKB/Swiss-Prot which contains 17,407 labeled and 47,619 unlabeled protein sequences show that our technique yields good performance on both training and test sets. Our proposed method generate classifier with more accuracy, precision, recall and F-measure than which is using only Jaccard coefficient for data selection and minimum mean square error as stopping criterion 0.3%, 1.00%, 2.83%, 1.91% respectively. In addition, the overlapping between prediction results on unknown genes also demonstrates the effectiveness of the developed technique.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ขอกราบขอบพระคุณ บิดา มารดา และเพื่อนๆ ทุกท่าน ที่สนับสนุน ให้กำลังใจ เป็นที่ปรึกษา คอยแนะนำสิ่งต่างๆ ในการเล่าเรียนตลอดมา

ขอกราบขอบพระคุณ รศ.ดร.กฤษณะ ไวยมัย อาจารย์ที่ปรึกษาวิทยานิพนธ์หลักที่ให้ความช่วยเหลือ คอยชี้แนะแนวทางหลักการและช่วยตรวจแก้วิทยานิพนธ์ฉบับนี้จนเสร็จลุล่วงไปด้วยดี ขอขอบพระคุณ ผศ.ดร.ยอดเยี่ยม ทิพย์สุวรรณ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่ช่วยแนะนำและตอบคำถาม ขอบพระคุณ คุณธนภัทร ฆังคะจิตร (พี่ชาย) และ คุณกำธร พันธุมะผล (ปอ) ที่กรุณาให้คำปรึกษาแนะนำเกี่ยวกับวิธีการและเทคนิคการทำงานวิจัย และขอขอบคุณ คุณศรณัฐ เจนถนอมม้า (ม่ำ) ที่ช่วยแนะนำเทคนิคการเขียนและตรวจทานของวิทยานิพนธ์ฉบับนี้

สุดท้ายขอขอบคุณเพื่อน ๆ พี่ ๆ น้อง ๆ ที่ห้องปฏิบัติการวิจัยการวิเคราะห์ข้อมูลและสืบค้นความรู้ (DAKDL) และเพื่อนๆ พี่ๆ น้องๆ MCPE สำหรับคำแนะนำและกำลังใจที่ดีเสมอมา

พัทริกา ณ พิกุล

มิถุนายน 2554

สารบัญ

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

คำอธิบายสัญลักษณ์และคำย่อ

BLAST	=	Basic Local Alignment Search Tool
BLASTN	=	Nucleotide - BLAST
BLASTP	=	Protein - BLAST
BLASTX	=	Translated nucleotide - BLAST
BLOSUM62	=	Blocks of Amino Acid Substitution Matrix 62
C	=	Classifier
FN	=	False Negative
FP	=	False Positive
GO	=	Gene Ontology
H-bond	=	Hydrocarbon Bond
is-a	=	ความสัมพันธ์ที่มีคุณสมบัติการถ่ายทอดจากคอนเซพแม่ไปยังคอนเซพลูก
L	=	เซตของข้อมูลที่ทราบคลาส (Labeled)
KDD	=	Knowledge Discovery from very large Database
M_A	=	เซตโมทีฟของข้อมูลสายโปรตีน A
M_B	=	เซตโมทีฟของข้อมูลสายโปรตีน B
part-of	=	ความสัมพันธ์ที่หมายถึงการเป็นส่วนประกอบ
PSI-BLAST	=	Position-Specific Iterative-BLAST
PSSM	=	Position-specific scoring matrix
SVMs	=	Semi-supervised Support Vector Machine
TN	=	True Negative
TP	=	True Positive
U	=	เซตของข้อมูลที่ยังไม่ทราบคลาส (Unlabeled)
U'	=	เซตของข้อมูลที่สุ่มเลือกมาจากเซตของข้อมูลที่ยังไม่ทราบคลาส
UniProtKB	=	UniProt Knowledgebase
V	=	Validation

เทคนิคการเรียนรู้แบบกึ่งมีผู้สอนสำหรับการจำแนกฟังก์ชันของโปรตีน

Semi-Supervised Learning for Protein Function Classification

คำนำ

การสร้างตัวจำแนกคลาส สำหรับข้อมูลอนุกรมโปรตีน (Protein sequence) ให้สามารถจำแนกคลาสได้อย่างมีประสิทธิภาพจะต้องอาศัยข้อมูลที่เราทราบคลาสเป็นจำนวนมาก ซึ่งข้อมูลประเภทนี้มีอยู่อย่างจำกัด ในขณะที่ข้อมูลที่ไม่ทราบคลาสนั้นมีอยู่ทั่วไป ข้อมูลที่ได้จากฐานข้อมูล UniProtKB/Swiss-Prot (UniProt Knowledgebase, 2007) เราพบว่า มีจำนวนของสายโปรตีน 21,630 สาย แต่ข้อมูลที่เราทราบคลาส (GO: Gene Ontology) มีจำนวนเพียง 1,126 สาย คิดเป็นอัตราส่วนของจำนวนข้อมูลที่ทราบคลาสต่อจำนวนข้อมูลที่ไม่ทราบคลาสประมาณ 17 เท่า หากเราใช้เฉพาะข้อมูลที่ทราบคลาสในการสร้างตัวจำแนกคลาสจะพบว่าข้อมูลที่เราทราบคลาสนั้นมีจำนวนน้อยเกินไป และข้อมูลส่วนใหญ่ที่มีเป็นข้อมูลที่เราไม่ทราบคลาส (Unlabeled Data) ทำให้ได้ผลการจำแนกที่ไม่ดีพอ

วิธีการแก้ปัญหาข้อมูลในการฝึกสอนมีจำกัดคือ การหาวิธีที่จะเพิ่มจำนวนข้อมูลที่ทราบคลาส เราจำเป็นจะต้องใช้ผู้เชี่ยวชาญทางด้านชีววิทยามาทำการกำหนดคลาสให้กับข้อมูลซึ่งใช้เวลามาก ดังนั้นหากเราสามารถนำข้อมูลที่ไม่ทราบคลาสมาใช้ในกระบวนการสร้างตัวจำแนกข้อมูลและสามารถทำให้ได้ตัวจำแนกที่มีความแม่นยำมากขึ้น ย่อมเป็นสิ่งที่เป็นประโยชน์และน่าสนใจ

การเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) (Huang *et al*., 2001; Li *et al*., 2005) เป็นวิธีแก้ปัญหากรณีข้อมูลที่ทราบคลาสมีจำนวนน้อย เราสามารถสร้างตัวจำแนกคลาสโดยใช้ข้อมูลที่ทราบและไม่ทราบคลาสมาทำการฝึกสอนร่วมกัน หากได้รับการฝึกสอนด้วยวิธีที่เหมาะสมกับชนิดของข้อมูลแล้ว จะทำให้ผลการจำแนกคลาสมีความถูกต้องแม่นยำสูงกว่าการที่เราทำการฝึกสอนด้วยข้อมูลที่ทราบคลาสเพียงอย่างเดียว (Huang *et al*., 2001; Oliver *et al*., 2003)

ในปัจจุบันมีวิธีเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนหลายวิธี สำหรับข้อมูลอนุกรมโปรตีนแล้ว วิธีที่น่าสนใจคือวิธีฝึกสอนด้วยตนเอง (self-training) วิธีการนี้ทำการฝึกสอนโดยวิธีเพิ่มจำนวนข้อมูลที่ทราบคลาส ด้วยข้อมูลที่ไม่ทราบคลาสที่มีความคล้ายคลึงมากที่สุด อย่างไรก็ตาม

การเรียนรู้ประเภทนี้มีข้อจำกัดเกี่ยวกับการเลือกข้อมูลที่ไม่ทราบคลาสเข้ามาเพื่อสร้างตัวจำแนก ระหว่างที่ทำการฝึกสอน หากเลือกข้อมูลที่ผิดคลาส จะมีผลต่อการฝึกสอนในรอบต่อไปด้วย ทำให้ตัวจำแนกที่สร้างได้ให้ความแม่นจำในการจำแนกคลาสลดลง (Shahshahani *et al*., 1994) ข้อควรระวังอีกข้อหนึ่งคือเกณฑ์การหยุดที่ไม่ดี ซึ่งจะทำให้ข้อมูลที่นำมาสร้างตัวจำแนก อาจรวมข้อมูลคลาสที่ผิดเข้ามามากเกินไป จนส่งผลให้ตัวจำแนกที่ได้ให้ผลการจำแนกคลาสที่ไม่ดีนัก

งานวิจัยนี้จึงนำเสนอเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนพร้อมด้วยเกณฑ์การเลือกข้อมูลที่ไม่ทราบคลาสเข้ามาระหว่างที่ทำการฝึกสอน และเกณฑ์การหยุดสอนเพื่อสร้างตัวจำแนกฟังก์ชันของโปรตีน ให้ตัวจำแนกสามารถทำนายได้แม่นยำยิ่งขึ้น

วัตถุประสงค์

1. ศึกษาและพัฒนาเทคนิคการจำแนกข้อมูลแบบกึ่งมีผู้สอน เพื่อแก้ปัญหาในกรณีที่มีข้อมูลจำกัด

2. ประยุกต์เทคนิคการจำแนกข้อมูลแบบกึ่งมีผู้สอนการจำแนกการทำงานของโปรตีน เพื่อให้สามารถจำแนกข้อมูลได้แม่นยำยิ่งขึ้น

2.1 พัฒนาเกณฑ์ในการเลือกข้อมูลที่ไม่ทราบคลาสเข้ามาระหว่างที่ทำการฝึกสอน

2.2 พัฒนาเกณฑ์ในการหยุดสอนตัวจำแนก

การตรวจเอกสาร

ในหัวข้อนี้จะกล่าวถึงความรู้พื้นฐานและงานวิจัยที่เกี่ยวข้องกับการทำงานวิจัย อันได้แก่ ความรู้พื้นฐานทางชีววิทยาของโปรตีน เทคนิคการจำแนกข้อมูลโดยวิธีเรียนรู้แบบกึ่งมีผู้สอน และงานอื่นๆที่เกี่ยวข้อง

1. โปรตีน

โปรตีนจัดเป็นส่วนประกอบที่สำคัญของเซลล์ และเอนไซม์ทุกชนิด (อมรา, 2540) ดังนั้น จึงมีความสำคัญต่อการเกิดลักษณะต่างๆ โครงสร้าง ตลอดจนหน้าที่และกระบวนการต่างๆใน สิ่งมีชีวิต โปรตีนจัดเป็นสารที่มีโมเลกุลขนาดใหญ่ ประกอบด้วยสายโพลีเปปไทด์ (Polypeptide) จำนวน 1 สายหรือมากกว่านั้น โดยที่ส่วนประกอบย่อยของสายโพลีเปปไทด์ คือ กรดอะมิโน (Amino Acid) 20 ชนิดที่เรียงต่อกันเป็นสายยาว (Lesk, 2004) โปรตีนแต่ละชนิดจะมีลำดับกรดอะ-มิโนไม่เหมือนกันเป็นเอกลักษณ์ ดังนั้นโปรตีนทุกชนิดทุกตัวจะต้องมีข้อมูลลำดับกรดอะมิโน โดยข้อมูลเพิ่มเติมอื่นๆ เกี่ยวกับโปรตีนเช่น แหล่งที่มาของโปรตีน ฟังก์ชันการทำงานของโปรตีน โครงสร้างโปรตีน ก็จะมีการอ้างอิงมาที่ลำดับกรดอะมิโนนี้ ซึ่งข้อมูลเหล่านี้มีการเก็บรวบรวมอยู่ในลักษณะที่ใช้งานได้ง่ายเช่น เป็นฐานข้อมูลบนเครือข่ายอินเทอร์เน็ต ทั้งนี้การเก็บข้อมูลโปรตีนที่สมบูรณ์ที่สุดเป็นความร่วมมือระหว่างหลายสถาบันที่เก็บรวบรวมข้อมูลโปรตีนที่เรียกกลุ่มตัวเองว่า UNIPROT (Apweiler, 2004) สามารถใช้งานฐานข้อมูลผ่านอินเทอร์เน็ตได้ที่ uniprot.org

1.1 โครงสร้างโปรตีน

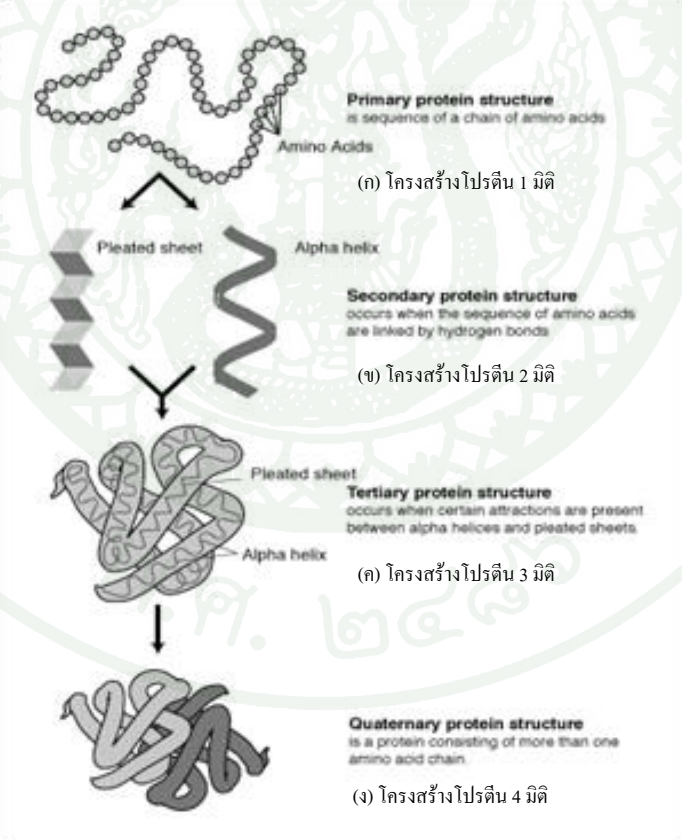
โครงสร้างของสายโปรตีนแบ่งออกเป็น 4 ระดับ คือ

1.1.1 โปรตีนที่มีโครงสร้าง 1 มิติ (Primary structure) โครงสร้างของโปรตีนในระดับนี้จะพิจารณาการเรียงตัวกันของกรดอะมิโนในลักษณะเป็นสายยาว (string) ดังภาพที่ 1(ก) โดยที่ลำดับของสายโปรตีนถูกกำหนดจากข้อมูลทางพันธุกรรม โครงสร้างของโปรตีนในระดับนี้นำมาใช้เป็นข้อมูลตั้งต้นในงานวิจัยนี้

1.1.2. โปรตีนที่มีโครงสร้าง 2 มิติ (Secondary structure) เป็นโครงสร้างที่เกิดขึ้นจาก H-bond ระหว่างหมู่ carboxyl และหมู่ amino Secondary structure ที่พบบ่อยในธรรมชาติ ได้แก่ α-Helix และ β-Pleated sheet ดังภาพที่ 1 (ข)

1.1.3. โปรตีนที่มีโครงสร้าง 3 มิติ (Tertiary structure) เป็นรูปร่างตลอดทั้งสายของโปรตีน โดยที่การหมุนม้วนพับไปมาของสายขึ้นอยู่กับแรงในระดับโมเลกุลของกรดอะมิโนแต่ละตัวในตำแหน่งที่เหมาะสม ดังภาพที่ 1 (ค)

1.1.4. โปรตีนที่มีโครงสร้าง 4 มิติ (Quaternary Structure) เป็นโครงสร้างของโปรตีนที่ประกอบด้วย polypeptide มากกว่า 1 สาย เกิดจากโปรตีนที่มีโครงสร้าง 3 มิติ ของ polypeptide แต่ละสายมารวมกัน ดังภาพที่ 1(ง)



ภาพที่ 1 โครงสร้างโปรตีนแบบต่างๆ (ก) โครงสร้างโปรตีน 1 มิติ, (ข) โครงสร้างโปรตีน 2 มิติ, (ค) โครงสร้างโปรตีน 3 มิติ, (ง) โครงสร้างโปรตีน 4 มิติ

1.2 ฐานข้อมูลโปรตีนและฟังก์ชันโปรตีน

ปัจจุบันฐานข้อมูลโปรตีนที่สมบูรณ์ที่สุดคือฐานข้อมูลที่เกิดจากความร่วมมือจากหลายสถาบันที่เรียกว่า UNIPROT โดยการเก็บข้อมูลเกี่ยวกับฟังก์ชันโปรตีนมีพื้นฐานเริ่มต้นจากการนำผลการวิจัยของโปรตีนนั้นมาใส่ลงไปในฐานข้อมูล และมีการระบุคุณสมบัติสำคัญของโปรตีนเป็นคำสำคัญ (keywords) รวมทั้งรายละเอียดการทำงานของโปรตีนนั้น ซึ่งทำให้ผู้สนใจโดยทั่วไปสามารถคัดเลือกข้อมูลโปรตีนเฉพาะตัวที่สนใจออกมาใช้งานได้สะดวกได้ นอกจากนี้ยังมีฐานข้อมูลอื่นที่น่าสนใจ เช่น ฐานข้อมูลฟังก์ชันโปรตีนที่มีการจัดรวบรวมเป็นหมวดหมู่เป็นการเฉพาะ เช่น ฐานข้อมูลโปรตีนเอนไซม์ที่เว็บไซต์ http://www.expasy.org/enzyme/ (Appel, 1994) ฐานข้อมูล Amino Acid-Nucleotide Interaction ที่เว็บไซต์ http://aant.icmb.utexas.edu/ (Hoffman *et al.*, 2004) หรือ ระบบ ontology ของฟังก์ชันโปรตีนขึ้นมา เช่นที่ Gene Ontology ที่เว็บไซต์ http://geneontology.org/ (The Gene Ontology Consortium, 2000) ฯลฯ

1.2.1 ยีนออนโทโลยี (Gene Ontology)

โครงการภววิทยายีนหรือ ยีนออนโทโลยี (Gene Ontology) เกิดจากความร่วมมือ เพื่อกำหนดคำอธิบายผลิตภัณฑ์ยีน ในฐานข้อมูลต่างกันให้คำอธิบายผลิตภัณฑ์ยีนนั้นมีมาตราฐาน สามารถเข้าใจได้ตรงกัน โดยเป้าหมายของโครงการยีนออนโทโลยีมี 3 ประการคือ ประการแรกเรื่องการพัฒนาและการบำรุงรักษา ประการที่สองการให้คำอธิบายประกอบของผลิตภัณฑ์ยีนระหว่างฐานข้อมูลที่เกี่ยวข้องกับสมาคมที่ทำการสร้างออนโทโลยี และประการสุดท้ายคือการพัฒนาเครื่องมือที่ใช้งานได้ง่ายในการบำรุงรักษาและใช้ในการพัฒนาออนโทโลยี

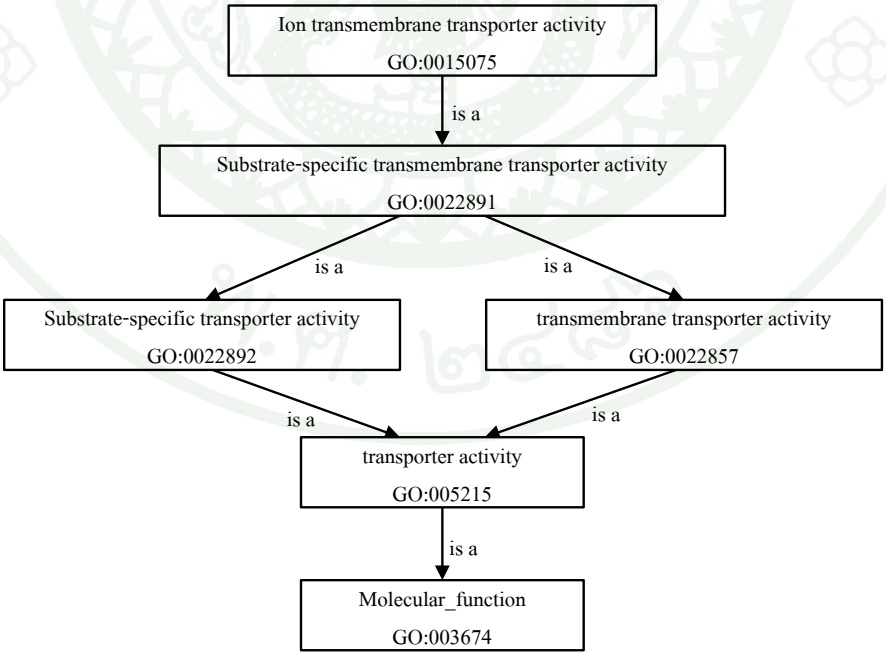
1.2.1.1 เทอมในยีนออนโทโลยี (Terms in the Gene Ontology)

ส่วนประกอบหลักของยีนออนโทโลยี คือ คำศัพท์เฉพาะที่เรียกว่าเทอม ซึ่งแต่ละเทอมในยีนออนโทโลยีจะมีคำว่า “GO:” แล้วตามด้วยตัวเลขจำนวน 7 หลักที่มีเอกลักษณ์ (ไม่ซ้ำกัน) เช่น GO:nnnnnnn และชื่อคำศัพท์เฉพาะ(term name) เช่น cell, fibroblast growth factor receptor binding หรือ signal transduction เป็นต้น ซึ่งแต่ละเทอมจะอยู่บนออนโทโลยีเพียงออนโทโลยีเดียวเท่านั้น

1.2.1.2 โครงสร้างของออนโทโลยี (Ontology structure)

เทอมหรือ ออนโทโลยีคอนเซพเชื่อมโยงกันด้วย 2 ความสัมพันธ์ คือ is-a และ part-of โดย is-a แทนความสัมพันธ์ที่มีคุณสมบัติการถ่ายทอด คุณสมบัติของคอนเซพแม่ไปยังคอนเซพลูก เมื่อ A is-a B มีความหมายว่า A เป็นคอนเซพที่สืบทอดมาจาก B เช่น nuclear chromosome is-a chromosome อธิบายได้ว่า nuclear chromosome เป็นคอนเซพที่สืบทอดมาจาก chromosome ส่วน part-of แทนความสัมพันธ์ที่หมายถึงการเป็นส่วนประกอบ เมื่อ C part-of D มีความหมายว่าคอนเซพ C ทุกตัวจะประกอบไปด้วยคอนเซพ D เสมอ ตัวอย่างเช่น nucleas part-of cell อธิบายได้ว่า nucleas ต้องเป็นส่วนประกอบของ cell แต่ทุก nucleas ไม่จำเป็นต้องประกอบด้วย cell

ความสัมพันธ์เหล่านี้จะจัดเรียงเป็นโครงสร้างกราฟแบบไม่มีวนรอบ (Direct Acyclic Graph) ดังภาพที่ 2 โดยโหนด (Node) แทน ออนโทโลยีคอนเซพ หรือเทอม และเส้นเชื่อมระหว่างโหนด (Edge) แทนความ สัมพันธ์ระหว่างออนโทโลยีคอนเซพ ซึ่งโครงสร้างกราฟแบบไม่มีวนรอบนั้นใกล้เคียงกับโครงสร้างต้นไม้ แต่แตกต่างตรงที่โหนดลูกสามารถมีโหนดพ่อได้มากกว่า 1 โหนด ตามกฎในยีนออนโทโลยี ที่กำหนดไว้ว่า เมื่อโหนดลูกสามารถอธิบายผลิตภัณฑ์ยีนนั้นได้ดังนั้น โหนดพ่อจะต้องมีความสัมพันธ์กับ ผลิตภัณฑ์ยีนนั้นเหมือนกัน



ภาพที่ 2 บางส่วนของยีนออนโทโลยี

1.3 ตัวแทนสายโปรตีนประเภทโมทีฟ

เนื่องจากฐานข้อมูลที่มีข้อมูลโปรตีนเอนไซม์ส่วนมากอยู่ในรูปของสายลำดับโปรตีน (protein sequences) ดังนั้นตัวแทนสายโปรตีนที่เหมาะสมกับข้อมูลลักษณะนี้ก็คือรูปแบบลำดับสายโปรตีน (protein sequence pattern) ที่เรียกว่าโมทีฟ (motif) (Weber *et al.*, 1982) โดยโมทีฟทำหน้าที่เป็นตัวแทนสายโปรตีนของกลุ่มโปรตีนหนึ่ง โมทีฟแสดงคุณสมบัติของกลุ่มโปรตีน องค์ประกอบพื้นฐานของโมทีฟประกอบด้วยสิ่งที่เรียกว่า กลุ่มแทนที่กรดอะมิโน บริเวณอนุรักษ์ และ gap โดยแต่ละโมทีฟก็คือการเรียงลำดับกันของกลุ่มแทนที่ บริเวณอนุรักษ์ และ gap นั่นเอง

ฐานข้อมูลที่เก็บรวบรวมเกี่ยวกับโมทีฟต่างๆเอาไว้ คือ PROSITE DATABASES โดยจัดเก็บรูปแบบ และคุณสมบัติต่างๆของโมทีฟเหล่านั้นเอาไว้ ฐานข้อมูลนี้สร้างขึ้นครั้งแรกโดย Amos Bairoch ในปี 1988 ซึ่งมีการจัดการปรับปรุงจนถึงปัจจุบันโดยสถาบัน expasy ลักษณะสำคัญของสายรูปแบบลักษณะที่สร้างขึ้นโดย PROSITE ก็คือการสร้างโดย “ผู้เชี่ยวชาญ” ไม่ใช่ระบบอัตโนมัติ ทำให้ลักษณะสายรูปแบบลักษณะของ PROSITE แฝงด้วย “ความรู้ในปฏิกิริยาชีวเคมี” ที่น่าเชื่อถือในระดับหนึ่ง ยกตัวอย่างเช่น

[ST]-x-[RK] หมายถึงเริ่มต้นด้วยกรดอะมิโนรหัส S หรือ T ก็ได้ ตามด้วยกรดอะมิโนรหัสอะไรก็ได้อีก 1 ตัว แล้วปิดท้ายด้วยกรดอะมิโนรหัส R หรือ K รวมแล้วมีกรดอะมิโนทั้งหมด 3 ตำแหน่งอยู่ในสายรูปแบบลักษณะนี้

ปัจจุบันสามารถใช้งานข้อมูล PROSITE ได้ที่เว็บไซต์ www.expasy.org

ในงานวิจัยนี้นำเอาข้อมูลจากฐานข้อมูล PROSITE มาตรวจสอบความถูกต้องของการหาโมทีฟ

1.4 ค่าคะแนนจากการเทียบเคียงคู่อนุกรม (Alignment score)

ค่าคะแนนที่เครื่องคอมพิวเตอร์คำนวณได้โดยใช้อัลกอริทึมในการเปรียบเทียบสายลำดับ ซึ่งมีค่าขึ้นกับจำนวนลำดับเบสหรือกรดอะมิโนที่ตรงกัน (matches) จำนวนเบสที่เกิดการแทนที่ (Substitutions), จำนวนเบสที่เกิดการแทรก (Insertions), และจำนวนเบสที่ถูกเอาออก (Deletions (gaps)) ค่าคะแนนของลำดับที่ตรงกันกับลำดับที่เกิดการแทนที่ได้มาจากเมทริกซ์ของ

คะแนน (Scoring matrix) เช่น ในกรณีการเปรียบเทียบสายลำดับโปรตีน จะใช้เมทริกซ์บลอสซัม BLOSUM หรือเมทริกซ์แพม PAM และเลือกใช้การหักคะแนนเมื่อมีการเปิดช่องว่างที่เหมาะสมกับเมทริกซ์ที่ถูกเลือก ค่าคะแนนของการเปรียบเทียบลำดับเบสจะอยู่ในหน่วย log odds มักอยู่ในรูป log ฐานสอง (Bit units) ค่าคะแนนสูงแสดงว่าสายลำดับที่นำมาเปรียบเทียบมีความคล้ายกันมาก

2. การทำเหมืองข้อมูล (Data Mining)

กระบวนการค้นหาสารสนเทศที่สำคัญ หรือ การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจจากฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) หรือที่เรียกกันว่า การทำเหมืองข้อมูล (Data Mining) เป็นสาขาหนึ่งในวิทยาศาสตร์คอมพิวเตอร์ที่กำลังได้รับความสนใจอย่างสูงในปัจจุบัน เทคนิคต่าง ๆ ที่ใช้ในกระบวนการ KDD จะถูกรวบรวมมาจากเทคนิคในหลายสาขาวิชา ได้แก่ Machine Learning, Statistics, Database และอื่น ๆ

ดวยเทคนิคของ KDD ขอมูลขนาดใหญจะถูกวิเคราะหและดึงความรูหรือสิ่งที่สำคัญออกมารวบรวมใหอยูในรูปฐานความรู (Knowledge Base) เพื่อใชสำหรับการสืบคนสิ่งที่ตองการซึ่งไมสามารถสืบคนไดจากวิธีการของระบบจัดการฐานขอมูล (DBMS) โดยทั่วไป เชน การวิเคราะหหาความสัมพันธของขอมูลหรือการทำนายปรากฏการณตาง ๆ ของขอมูลที่จะเกิดขึ้นในอนาคต เทคนิคตาง ๆ เหลานี้สามารถนำไปใชใหเกิดประโยชนไดในหลาย ๆ สาขา เชน ในดานชีวสารสนเทศ (Bioinformatics) เราสามารถใช้เทคนิคของ KDD วิเคราะหไดวาเอ็นไซม์ มีฟังก์ชันการทำงานแบบใด เปนตน

KDD มีหนาที่หลัก 2 ประการ ซึ่งในแตละหนาที่จะประกอบไปดวยเทคนิคยอยหลาย ๆ เทคนิค โดยสามารถอธิบายได ดังนี้

ก. หนาที่สืบคนความรูเพื่อแสดงคุณสมบัติหรือลักษณะเฉพาะของขอมูลในฐานขอมูล (Descriptive mining tasks) เทคนิคใน KDD ที่ทำหนาที่ดังกลาว เชน Association Rules Discovery Sequential Patterns Discovery และ Time-Series Analysis เปนตน

ข. หนาที่สืบคนความรูเพื่อสรุปแบบจำลองของขอมูลเพื่อใชในการทำนายขอมูลใหม (Predictive mining tasks) เทคนิคใน KDD ที่ทำหนาที่ดังกลาว เชน Data Classification, Data

Clustering และ Data Prediction เป็นต้น

2.1 Data Classification

Data Classification (Han and Kamber, 2000) เป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ หรือ ดาต้าไมน์นิง (Data Mining) เทคนิคการจำแนกประเภทข้อมูลเป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ จากกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลสอนระบบ (training data) ที่แต่ละแถวของข้อมูลประกอบด้วยลักษณะประจำแอทริบิวท์ (attribute) จำนวนมาก จุดประสงค์ของการจำแนกประเภทข้อมูลคือการสร้างโมเดลการแยกแอทริบิวท์หนึ่งโดยขึ้นกับแอทริบิวท์อื่น (Pang-Ning, 2005)

โมเดลที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาคลาสในข้อมูลที่ยังไม่ได้แบ่งกลุ่มในอนาคตได้ เทคนิคการจำแนกประเภทข้อมูลนี้ได้นำไปประยุกต์ใช้ในหลายด้าน เช่น การตรวจสอบความผิดปกติ และการวิเคราะห์ทางการแพทย์ เป็นต้น

2.1.1 การจำแนกประเภทข้อมูลโดยอัลกอริทึม (Support Vector Machine)

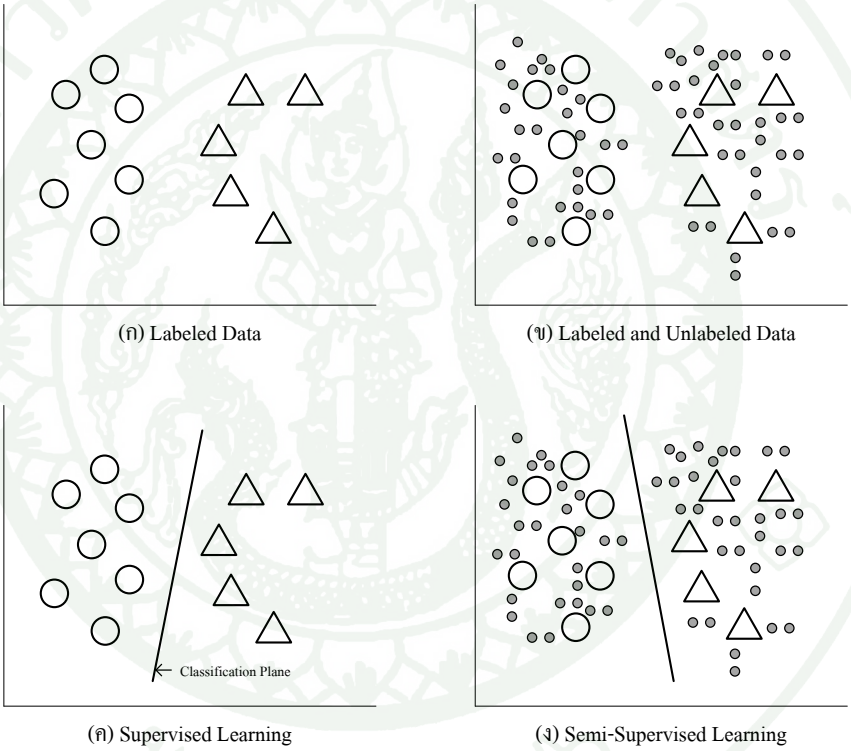
Support Vector Machine หรือ SVM คือตัวจำแนกประเภทแบบเส้นตรง (linear classifier) ที่ทำงานโดยใช้ เคอเนลฟังก์ชัน (kernel function) และสร้างโดยใช้หลักการหาขอบที่กว้างที่สุดและการแก้ปัญหาแบบมีข้อแม้ ข้อดีสำคัญของ SVM คือ สามารถจำแนกรูปแบบ (pattern) ที่มีความสลับซับซ้อน (complex) ได้อย่างมีประสิทธิภาพ ข้อมูลตัวอย่างที่จะให้อัลกอริทึมเรียนรู้ ด้วยวิธีการแบบ SVM นี้เป็นการเรียนรู้แบบมีผู้สอน

2.2 การเรียนรู้แบบกึ่งมีผู้สอน Semi-Supervised Learning

การเรียนรู้แบบเดิมใช้ข้อมูลที่ทราบคลาสในการฝึกสอนโมเดล (Xiaojin Zhu, 2005) ซึ่งข้อมูลที่เราทราบคลาสนั้นหาได้ยาก และใช้คนทำซึ่งใช้เวลามาก ในขณะที่ข้อมูลที่ไม่ทราบคลาส มีอยู่โดยทั่วไป การเรียนรู้แบบกึ่งมีผู้สอน อาศัยข้อมูลที่เราทราบคลาสร่วมกับข้อมูลที่เราไม่ทราบคลาสในการฝึกสอนเพื่อสร้างโมเดลที่มีประสิทธิภาพมากขึ้น ปัจจุบันนี้มีเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนหลายวิธี และแบ่งเป็น 5 ประเภทหลักคือ โมเดลแบบเพิ่มพูน (Generative Model)

(Ratanamahatana *et al*. 2005) วิธีการเชิงกราฟ (Graph Based Method)(Blum *et al*., 2001) วิธีการแบ่งบริเวณที่ความหนาแน่น (Density Based Approach)(Chapelle *et al*., 2006) วิธีการฝึกสอนร่วมกัน (Co-training)(Cohen *et al*., 2004) และวิธีการฝึกสอนด้วยตนเอง(Self-training)(Li *et al*., 2005)

ในงานวิจัยนี้สนใจการจำแนกข้อมูลแบบกึ่งมีผู้สอนด้วยตนเอง (Self Training) และวิธีการจำแนกแนกข้อมูลแบบกึ่งมีผู้สอน โดยวิธีการฝึกสอนร่วมกัน เนื่องจากเป็นวิธีที่เข้าใจง่าย และใช้ฟังก์ชันระยะทางในการคำนวน



ภาพที่ 3 ตัวอย่างของการเรียนรู้แบบมีผู้สอน (Supervised Learning) และการเรียนรู้แบบ กึ่งมีผู้สอน (Semi-supervised Learning) ในการจำแนกไบนารี่คสาส
(ก) แสดงข้อมูลไบนารี่คลาสที่ต้องการนำมาจำแนก
(ข) ข้อมูลที่ไม่ทราบผลเฉลยแทนด้วยรูปวงกลมเล็กๆสีเขียว
(ค) เส้นแบ่งคลาสที่ได้จากการเรียนรู้แบบมีผู้สอน ซึ่งอาจแบ่งขอบเขตข้อมูลผิดพลาด
(ง) เมื่อนำข้อมูลที่ไม่ทราบคลาสมาใช้ร่วมในการเรียนรู้แบบกึ่งมีผู้สอน เส้นแบ่งขอบเขตข้อมูลพาดผ่านบริเวณที่ความหนาแน่นต่ำ ระหว่างข้อมูลสองกลุ่ม ซึ่งมีผลทำให้จำแนกข้อมูลได้แม่นยำกว่าการเรียนรู้แบบมีผู้สอน

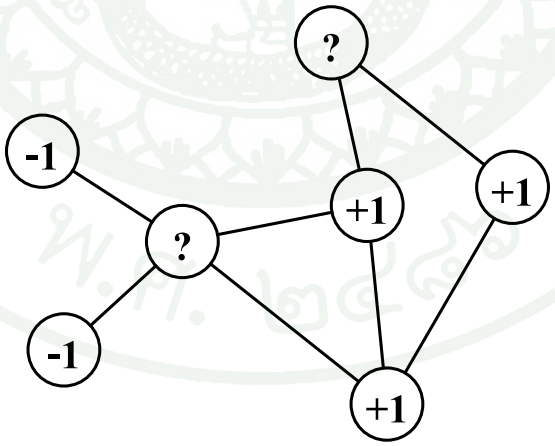
2.2.1 ประเภทของการเรียนรู้แบบกึ่งมีผู้สอน

2.2.1.1 โมเดลแบบเพิ่มพูน (Generative Model)

เป็นเทคนิคที่เกิดขึ้นในช่วงแรกๆ ของการเรียนรู้แบบกึ่งมีผู้สอน การเรียนรู้ประเภทนี้มีสมมติฐานว่า ข้อมูลเกิดจากการรวมตัวกันของการกระจายตัวหลายๆแบบ (Mixture distribution) ซึ่งการกระจายตัวของข้อมูลแบบผสมสามารถสร้างได้โดยอาศัยข้อมูลที่ไม่ทราบคลาส วิธีนี้ถูกนำไปประยุกต์ใช้กับการจำแนกประเภทตัวหนังสือ (text classification) และการจัดเรียงใบหน้าคน (face orientation discrimination) ตัวอย่างอัลกอริทึมของโมเดลแบบเพิ่มพูน คือ EM (expected maximization), Fisher kernel สำหรับการเรียนรู้แบบไม่ต่อเนื่อง (discriminative learning) ข้อดีของโมเดลแบบเพิ่มพูน เป็นโมเดลที่สามารถอธิบายการเกิดของข้อมูลได้ดี

2.2.1.2 วิธีการเชิงกราฟ (Graph-based Method)

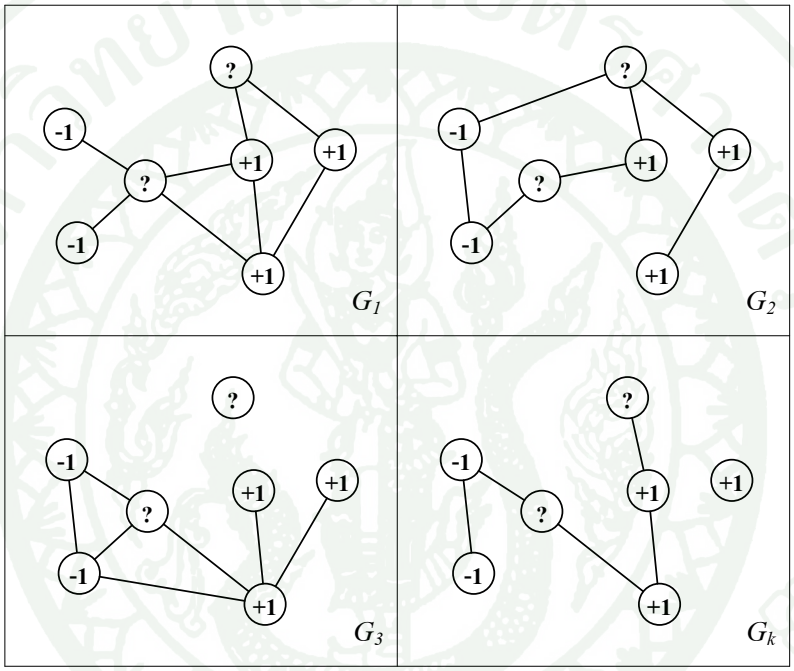
เป็นการเรียนรู้โดยอาศัยโครงสร้างข้อมูลแบบกราฟ ข้อมูลจะถูกเก็บอยู่ในรูปแบบกราฟ โหนดของกราฟแสดงค่าต่างๆของข้อมูล เส้นเชื่อมแสดงค่าความเหมือนหรือความเกี่ยวข้องกับโหนดเพื่อนบ้านดังภาพที่ 3



ภาพที่ 4 โครงสร้างข้อมูล สำหรับการเรียนรู้แบบกึ่งมีผู้สอน โดยวิธีเชิงกราฟ โดยโหนดแทนข้อมูลและเส้นเชื่อมแสดงระยะทางระหว่างข้อมูล

ที่มา: Koji Tsuda *et al*. (2005)

งานวิจัยที่ใช้เทคนิคนี้คือ การจำแนกโปรตีนโดยใช้หลายเครือข่าย (Koji Tsuda *et al*., 2005) งานวิจัยนี้ใช้ข้อมูล protein interaction network เครือข่ายโปรตีน จากแหล่งข้อมูล (Data source) หลายๆแหล่ง ซึ่งจะได้เซตของกราฟ ที่แต่ละกราฟจะได้มาจากแหล่งข้อมูลคนละแหล่ง ดังภาพที่ 4 นั่นคือกราฟแต่ละอัน จะมีข้อมูลบางส่วนที่ไม่ขึ้นต่อกัน (Independent) และ ข้อมูลบางส่วนสามารถนำมาเติมเต็มเข้าด้วยกันได้ (Complementary part) นั่นคือ เอาข้อมูลจากกราฟหลายๆตัวมาช่วยสร้างกราฟตัวใหม่ที่มีความสมบูรณ์กว่า



ภาพที่ 5 เครือข่ายโปรตีน ที่สร้างได้จาก protein interaction network จากฐานข้อมูลต่างๆ

ที่มา: Koji Tsuda *et al.* (2005)

2.2.1.3 วิธีการแบ่งบริเวณที่ความหนาแน่น (Density-based Approach)

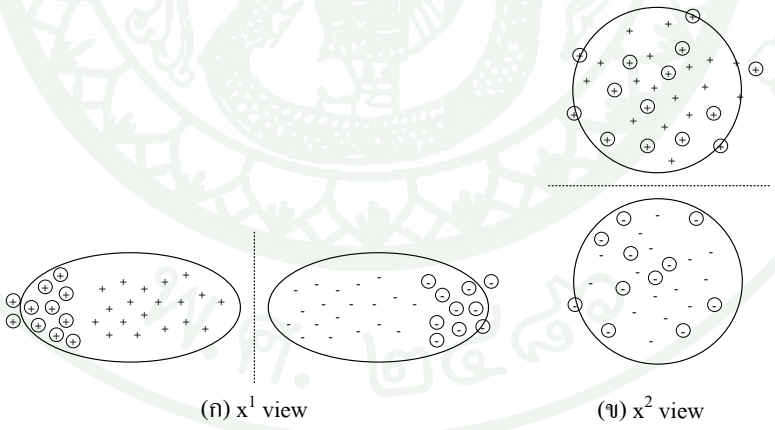
วิธีการแบ่งบริเวณที่ความหนาแน่น มีสมมติฐานว่า เส้นแบ่งคลาสน่าจะอยู่ในบริเวณที่ความหนาแน่นของข้อมูลต่ำ ตัวอย่างอัลกอริทึม เช่น Transductive SVMs (S3VM), Entropy Minimization เป็นต้น

งานวิจัยที่ใช้เทคนิคการแบ่งบริเวณที่ความหนาแน่น มักเน้นการแก้ปัญหาเรื่องการแปลงตัวแทนข้อมูล (data representation) (Weston, J. *et al*., 2004) ได้ศึกษาและทดลองสร้างระบบในการทำนายโปรตีน โดยใช้เคอร์เนล (kernel) แทนในสมการหาระยะทางระหว่างข้อมูล เพื่อนำไปใช้กับตัวจำแนกแบบ SVM ผู้วิจัยได้นำเสนอ String Kernel สำหรับปัญหาการแปลงตัวแทนข้อมูลของอนุกรมโปรตีน

2.2.1.4 วิธีการฝึกสอนร่วมกัน (Co-training)

สมมติฐานของ การเรียนรู้แบบกึ่งมีผู้สอนโดยวิธีการฝึกสอนร่วมกัน คือคุณลักษณะ (feature) ของข้อมูลไม่ขึ้นต่อกัน และไม่สามารถแบ่งแยกได้ (Blum *et al*., 1998) นั่นคือหากมีเซตของคุณลักษณะที่ไม่ขึ้นต่อกัน และคุณลักษณะแต่ละเซต สามารถนำมาใช้สร้างตัวจำแนกได้เหมือนกันก็สามารถ ใช้อัลกอริทึมการฝึกสอนร่วมกันได้

การเรียนรู้แบบนี้ จะทำการเรียนรู้โดยทำการฝึกสอนข้อมูลโดยใช้คุณลักษณะคนละประเภทกัน ในการสร้างตัวจำแนก และทำการเพิ่มข้อมูลฝึกสอนโดยใช้ผลจากการทำนายที่มีค่าความเชื่อมั่นสูงสุดของตัวจำแนกอีกตัวหนึ่งดังภาพที่ 6



ภาพที่ 6 ข้อมูลที่ถูกทำนายและมีค่าความเชื่อมั่นสูงในข้อมูลที่ใช้เซตของคุณลักษณะเซตแรก (X^1) แทนด้วยภาพวงกลมเล็ก จะถูกนำไปรวมกับข้อมูลฝึกของข้อมูลที่ใช้เซตคุณลักษณะที่สอง (X^2)

ที่มา: Chapelle *et al.* (2006)

อัลกอริทึม: การฝึกสอนร่วมกัน

1. แบ่ง feature ของข้อมูลเป็นสองเซตที่ไม่ขึ้นต่อกัน
2. ทำการฝึกสอนบน feature เซตแต่ละ เซต แยกกัน
3. ผลจากการทำนายที่ได้จากตัวจำแนกหนึ่ง จะถูกนำไปรวมกับข้อมูลฝึกสอนของอีกตัวจำแนกหนึ่ง
4. ทำซ้ำ ข้อ 2

2.1.1.5 วิธีการฝึกสอนด้วยตนเอง (Self-training)

วิธีการนี้ทำการฝึกสอนโดยวิธีเพิ่มจำนวนข้อมูลที่ทราบคลาส ด้วยข้อมูลที่ไม่ทราบคลาสที่มีความคล้ายคลึงมากที่สุด ในแต่ละรอบของการฝึกสอน โมเดลที่ได้จะถูกนำไปจำแนกข้อมูลที่ไม่ทราบคลาส จากนั้นข้อมูลที่ค่าผลเฉลยมีค่าความเชื่อมั่นสูงสุด จะถูกนำไปรวมกับข้อมูลฝึกสอน จำนวนของข้อมูลฝึกสอนจะเพิ่มขึ้นเรื่อยๆ ในทุกรอบของการฝึกสอน อัลกอริทึมจะหยุดทำงานก็ต่อเมื่อ ตัวจำแนกให้ค่าความผิดพลาดน้อยที่สุด หรือ ถึงเกณฑ์การหยุดสอน

อัลกอริทึม: การฝึกสอนด้วยตนเอง

1. สร้างตัวจำแนกคลาสจากกลุ่มข้อมูลที่ทราบคลาส
2. ใช้ตัวจำแนกคลาสที่ได้จากขั้นตอนที่ 1 จำแนกคลาส กลุ่มข้อมูลที่ไม่ทราบคลาส
3. เลือกข้อมูลที่มีความคล้ายมากที่สุดจากการจำแนกในขั้นตอนที่ 2 แล้วเพิ่มเข้ากับกลุ่มข้อมูลที่ทราบคลาส
4. สร้างตัวจำแนกคลาสโดยรวมข้อมูลจากขั้นตอนที่ 3
5. ทำขั้นตอนที่ 2 ถึง 4 ซ้ำ จนกระทั่งได้ตัวจำแนกคลาสที่เหมาะสมสำหรับการจำแนกข้อมูล

อุปกรณ์และวิธีการ

อุปกรณ์

1. เครื่องคอมพิวเตอร์

จำนวน 1 เครื่อง

1.1 หน่วยประมวลผล Intel Core 2 Quad ความเร็ว 2.51 กิกกะเฮิรตซ์

1.2 หน่วยความจำหลักขนาด 4 กิกกะไบต์

1.3 หน่วยความจำสำรอง(ฮาร์ดดิสก์) ขนาด 500 กิกกะไบต์

1.4 ติดตั้งระบบปฏิบัติการวินโดวส์เอ็กซ์พี เซอร์วิซแพ็ค 3 (Windows XP SP3)

1.5 ติดตั้งโปรแกรมวิชวลสตูดิโอดอทเน็ต (Visual Studio.NET) รุ่น 2005

1.6 ติดตั้งโปรแกรมระบบฐานข้อมูลไมโครซอฟท์เอสคิวแอลเซิฟเวอร์ (Microsoft SQL Server) รุ่น 2005

2. ฐานข้อมูล

จำนวน 2 ชุด

2.1 ฐานข้อมูล UniProtKB/Swiss-Prot

2.2 ฐานข้อมูล PROSITE

วิธีการ

1. ขั้นตอนการเตรียมข้อมูล

รวบรวมข้อมูลอนุกรมโปรตีนจากฐานข้อมูล UniProtKB/Swiss-Prot กับข้อมูลโมทีฟในฐานข้อมูล Prosite (Prosite Database, 2007)จากนั้นทำการค้นหาจำนวนเซตของฟังก์ชันที่ปรากฏในฐานข้อมูลเพื่อกำหนดเลขคลาสเซต และทำการแทนที่ฟังก์ชันเซตด้วยเลขอ้างอิงคลาสเซต (Class Set ID) ให้กับแถวข้อมูลที่ทราบคลาสฟังก์ชัน ตัวอย่างเช่น ข้อมูลที่มีฟังก์ชัน GO:0015075, GO:0003674 จะแทนด้วย ClassID = 1 ดังตารางที่ 1

ตารางที่ 1 เลขอ้างอิงคลาสเซตที่แทนกลุ่มของยีนออนโทโลยี เทอม

Class ID	Functions	Class ID	Functions
1	GO:0015075, GO:0003674	2	GO:0022857, GO:0015075

หลังจากแทนที่คลาสเซตด้วย Class ID แล้วทำการนับจำนวนแถวของข้อมูลจัดกลุ่มตาม Class ID แล้วเรียงลำดับจำนวนแถวของแต่ละกลุ่มจากมากไปน้อยดังตารางที่ 2 จากนั้นกรองข้อมูลที่จำนวนแถวข้อมูลในกลุ่มมีค่าน้อยกว่า 5 แถวออกไปจากข้อมูลที่ทราบคลาสฟังก์ชัน

การเตรียมแถวข้อมูลที่นำไปฝึกสอน ได้จากวนรอบอ่านเลขอ้างอิงโมทีฟจากฐานข้อมูลในเลขอ้างอิง Class ID ตัวอย่างแถวข้อมูลดังนี้

1 0:1 5:1 17:1 2122:1

1 0:1 5:1 17:1 2122:1 344:1

1 0:1 5:1 17:1 2122:1 344:1

1 0:1 5:1 17:1 2122:1 477:1

2 0:1 8:1 10:1 513:1

2 0:1 8:1 10:1 513:1

โดยข้อมูลตำแหน่งแรกคือคลาสเซ็ต ตำแหน่งถัดมาคือเลขอ้างอิงโปรตีน ตามด้วยเลขอ้างอิงโมทีฟที่พบในแต่ละสายโปรตีน

ตารางที่ 2 ตัวอย่างสัดส่วนของคลาสในข้อมูลฝึกสอน

Class ID	Count
1813	66
270	61
1	59
1989	59
1261	58
1998	56
1824	54
1467	47
1689	47
1792	46
2008	46
366	46
390	45
1075	43
1142	42
1844	42
1687	41
236	41
723	41
285	38

2. การสร้างตัวจำแนกฟังก์ชันโปรตีน

ตัวจำแนกฟังก์ชันโปรตีนสร้างจากข้อมูลที่ทราบคลาสร่วมกับข้อมูลอีกส่วนหนึ่งที่ไม่ทราบคลาส งานวิจัยนี้เลือกอัลกอริทึม Support Vector Machine (SVM) ในการเรียนรู้ และใช้เทคนิคการเรียนรู้แบบฝึกสอนด้วยตนเองที่มีการทำงานวนรอบ มีการเลือกข้อมูลที่ไม่ทราบคลาสเพื่อรวมกับข้อมูลที่ทราบคลาสในแต่ละรอบของการเรียนรู้ และหยุดการเรียนรู้เมื่อถึงเกณฑ์การหยุดที่ผู้ใช้กำหนด

2.1 อัลกอริทึม

การสร้างตัวจำแนกฟังก์ชันโปรตีนจะใช้วิธีการฝึกสอนด้วยตนเอง โดยการทำงานของอัลกอริทึมมีขั้นตอนดังนี้

2.1.1 แบ่งกลุ่มข้อมูล

กำหนดให้ L เป็นเซตของข้อมูลที่ทราบคลาส และ U เป็นเซตของข้อมูลที่ยังไม่ทราบคลาส และ V คือเซตของข้อมูลตรวจทาน (Validation Set) ซึ่งจากภาพจะเห็นว่า การฝึกสอนจะเริ่มที่การแบ่งข้อมูลออกเป็นเซต L, U และ V

2.1.2 ฝึกสอนตัวจำแนก

ใช้ข้อมูลเซต L เป็นข้อมูลฝึกสอน และ SVM เป็นอัลกอริทึมในการเรียนรู้

2.1.3 ตรวจสอบเกณฑ์การหยุด

นำข้อมูลเซต V ไปทำนายคลาสฟังก์ชันโดยโมเดลจำแนกฟังชันที่ได้จากขั้นก่อนหน้า จากนั้นตรวจสอบเกณฑ์การหยุด หากถึงเกณฑ์การหยุด ให้หยุดการฝึกสอนตัวจำแนก ถ้ายังไม่ถึงเกณฑ์การหยุดให้ทำกระบวนการถัดไป

2.1.4 สุ่มข้อมูลเซต U' จากข้อมูลเซต U

2.1.5 ทำนายข้อมูลที่ไม่ทราบคลาส

โมเดลสำหรับจำแนก C ได้จากการฝึกสอนในกระบวนการข้างต้นและจะนำโมเดล C ไปทำนายชุดข้อมูลเซต U' ที่ได้จากการสุ่ม

2.1.6 คำนวณค่าความคล้ายคลึง

ผลการทำนายคลาสฟังก์ชันของข้อมูลเซต U' จะนำมาใช้ในขั้นต่อมาเพื่อทำการคำนวนค่าความคล้าย (Similarity) ระหว่างข้อมูลในเซต L และ เซต U' ที่มีคลาสฟังก์ชัน

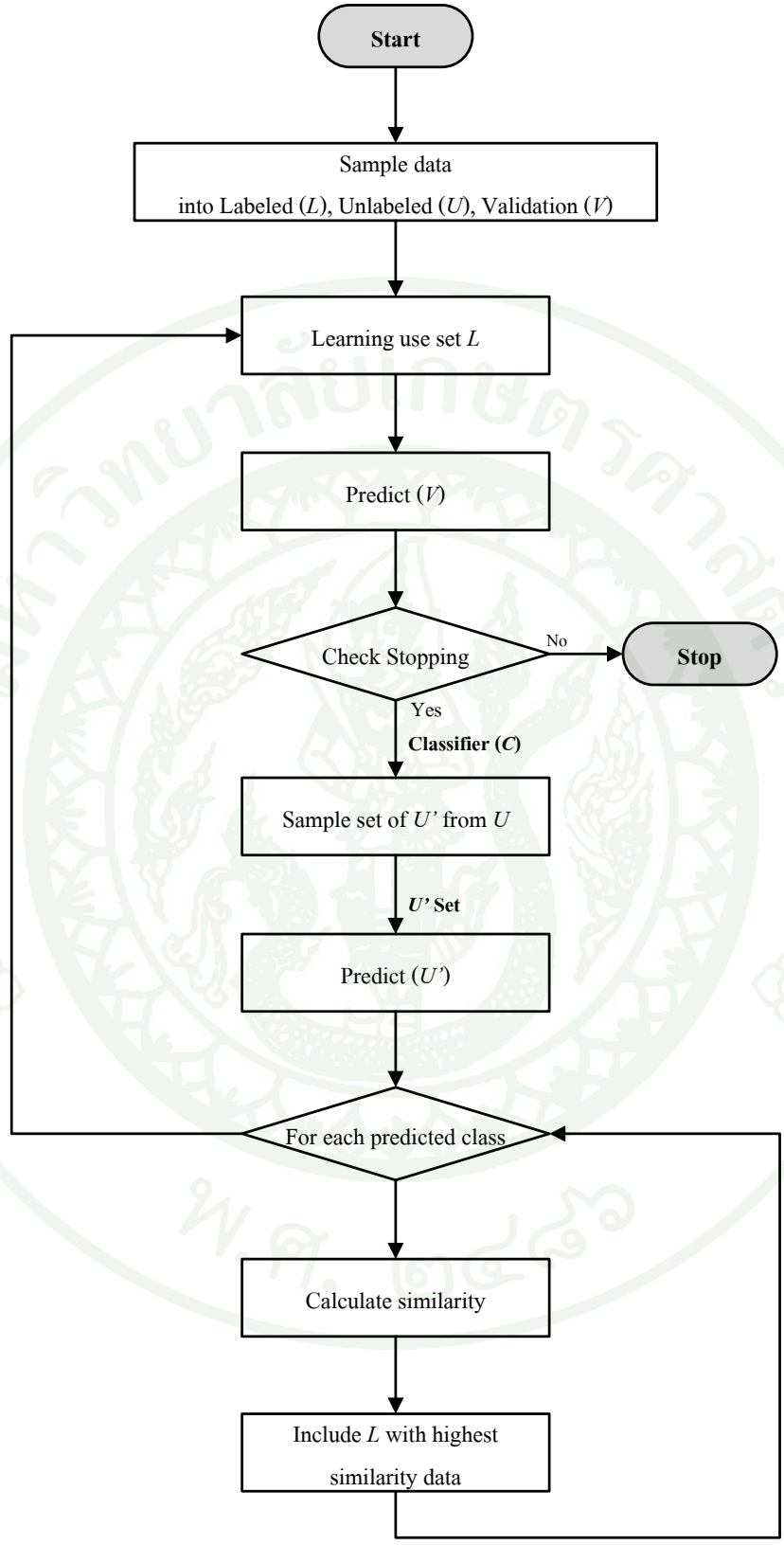
เดียวกัน เลือกแถวข้อมูลที่มีค่าความคล้ายสูงสุดในแต่ละคลาสเก็บไว้ในเซต L' แถวข้อมูลที่ถูกเลือกให้ย้ายออกจากเซต U

2.1.7 เตรียมข้อมูลฝึกสอนใหม่

นำข้อมูลเซต L' รวมเข้าในข้อมูลเซตฝึกสอน L

2.1.8 ย้อนกลับไปยังขั้นตอนที่ 2.1.2

แผนภูมิการทำงานของอัลกอริทึมแสดงดังภาพที่ 7



ภาพที่ 7 แผนภูมิการทำงานของอัลกอริทึม

3. การคำนวณค่าความคล้ายคลึงระหว่างคู่ข้อมูล

การเลือกข้อมูลที่ไม่ทราบคลาสเข้ามาเพื่อสร้างตัวจำแนกระหว่างที่ทำการฝึกสอน ทำโดยการเปรียบเทียบความคล้ายคลึงระหว่างคู่ข้อมูลในข้อมูลฝึกสอนและเซตข้อมูลที่ไม่ทราบคลาส การวัดความคล้ายคลึงระหว่างคู่สายอนุกรมโปรตีนสามารถทำได้โดยใช้สัมประสิทธิ์แจคการ์ด (Fisher *et al*, 1915) หรือความคล้ายคลึงที่ได้จากการเทียบเคียงคู่อนุกรมโปรตีน (Sean R. et al., Anuj R. *et al*., 2008)

3.1 การคำนวนค่าความคล้ายคลึงระหว่างคู่ข้อมูลโดยวิธีสัมประสิทธิ์แจคการ์ด (Jaccard Coefficient)

สูตรการคำนวณสัมประสิทธิ์แจคการ์ด

$$Jaccard\ Coefficient = \frac{|M_A \cap M_B|}{|M_A \cup M_B|} \quad (1)$$

เมื่อ M_A = เซตโมทีฟของข้อมูลสายโปรตีน A

M_B = เซตโมทีฟของข้อมูลสายโปรตีน B

โดย $|M_A \cap M_B|$ คือจำนวนข้อมูลที่ปรากฏอยู่ในเซต M_A และ M_B ส่วน $|M_A \cup M_B|$ คือจำนวนข้อมูลที่ปรากฏอยู่ในเซต M_A หรือ M_B

3.2 การคำนวนค่าความคล้ายคลึงระหว่างคู่ข้อมูลโดยวิธี PSI-BLAST

BLAST หรือ Basic Local Alignment Search Tool คือโปรแกรมที่ทำหน้าที่ในการค้นหาความเหมือนหรือความแตกต่าง (identity and similarity) ของ สายอนุกรมโปรตีนที่เราต้องการค้นหากับสายอนุกรมโปรตีน ที่มีอยู่ในฐานข้อมูลโดยมีอัลกอริทึม ในการคำนวณที่แตกต่างกันไปตามแต่วัตถุประสงค์ของการใช้งาน ซึ่งในปัจจุบันได้มีการออกแบบโปรแกรม ให้เลือกใช้งานตามแต่วัตถุประสงค์ของงานที่ต้องการศึกษา โดยโปรแกรมที่นิยมเลือกใช้งานได้แก่ BLASTN, BLASTP, BLASTX, TBLASTN, TBLASTX และ PSI-BLAST ซึ่งแต่ละชนิดมีความแตกต่างกันออกไป

PSI-BLAST ใช้สำหรับค้นหาข้อมูลของโปรตีนจากฐานข้อมูลโดยใช้ข้อมูลของโปรตีน ที่ต้องการเปรียบเทียบลงไป แต่ต่างกับ BLASTP ตรงที่ PSI-BLAST ใช้การค้นหาในลักษณะ Position-Specific Iterative BLAST ซึ่งมี อัลกอรึทึมแบบ PSSM (Position-specific scoring matrix) ถูกออกแบบมาโดยมีสมมติฐานว่าโปรตีนมีลักษณะโครงสร้าง ตำแหน่งองค์ประกอบของกรดอะมิโนที่แตกต่างกัน ในการทดลองนี้ใช้ BLOSUM62 เป็น Scoring Matrix

BLOSUM62 Substitution Matrix เป็นการพิจารณาการเรียงตัวของกรดอะมิโนในสายอนุกรมโปรตีนสองสาย คำนวนโดยใช้อัลกอริทึมในการเปรียบเทียบสายลำดับซึ่งมีค่าขึ้นกับจำนวนลำดับกรดอะมิโนที่ตรงกัน (matches) ที่เกิดการแทนที่ (substitutions) การแทรก (insertions) และอะมิโนที่ถูกเอาออก (deletions or gaps)

3.3 การคำนวนค่าความคล้ายคลึงระหว่างคู่ข้อมูลโดยวิธีสมิธ-วอเตอร์แมน (Smith-Waterman)

สมิธ-วอเตอร์แมน เป็นเกณฑ์ที่ใช้ในการวัดความคล้ายคลึงในอัลกอริทึมการเรียนรู้แบบฝึกสอนด้วยตนเองแบบหนึ่ง (Anuj R. *et al*., 2008) วิธีนี้เป็นวิธีเทียบความเหมือนของสายอนุกรมอักษระโดยใช้วิธีทางกำหนดการพลวัต (Dynamics Programming) ซึ่งค่าคะแนนที่ได้จาก Scoring Matrix เป็นค่าคะแนนมากสุดของวิธีการเรียงตัวของอักขระ

วิธีสมิท-วอเตอร์แมน ทำการแก้ปัญหาโดยการแบ่งออกเป็นปัญหาย่อย ๆ (sub-problems) จากนั้นทำการคำนวณเพื่อแก้ปัญหา แล้วนำคำตอบที่คำนวณได้บันทึกลงในตารางหรือเมทริกซ์ (Smith, 1981) วิธีทางกำหนดการพลวัตมักใช้ในการแก้ปัญหาที่มีคำตอบที่น่าจะเป็นไปได้หลายคำตอบและจำเป็นต้องได้คำตอบที่เหมาะสมที่สุด อัลกอริทึมนี้ถูกใช้เพื่อหาการเปรียบเทียบสายลำดับที่เหมาะสม โดยใช้ระบบการให้คะแนนจากการเปรียบเทียบสายลำดับหลาย ๆ แบบ

3.3.1 ขั้นตอนในการคิดค่าความคล้ายคลึงระหว่างคู่ข้อมูลโดยวิธี สมิธ-วอเตอร์แมน

ขั้นตอนในการคำนวณมีดังนี้

3.3.1.1 สร้างเมตริกซ์ M ขนาด mxn เมื่อ m คือความยาวของสายอนุกรมเส้นที่ 1 และ n คือความยาวของสายอนุกรมเส้นที่ 2

3.3.1.2 คำนวณ M[i][j] บนฐานของ Scoring Matrix ที่กำหนด

M[i][j] คำนวนจากสูตรที่ 2 และ 3 ดังนี้

$$M[i][0] = 0, M[0][j] = 0 \quad (2)$$

$$M[i][j] = \begin{cases} M[i-1][j-1] + S[s_i][t_i], \text{ถ้า } s_i\ t_j \\ M[i-1][j] - d, \text{ถ้า } s_i - \\ M[i][j-1] - d, \text{ถ้า } -t_j \end{cases} \quad (3)$$

เมื่อ s,t = อนุกรมอักขระในสายโปรตีน

m = ความยาวของสายอนุกรมเส้นที่ 1

n = ความยาวของสายอนุกรมเส้นที่ 2

'-' = ช่องว่างในสายอนุกรมโปรตีน

3.3.1.3 หาเส้นทางที่ให้ค่าคะแนนสูงสุด

3.3.2 ความแตกต่างระหว่างการเทียบเคียงคู่อนุกรมโปรตีนโดยวิธีสมิท-วอเตอร์แมน และวิธี PSI-Blast

จุดประสงค์ในการเทียบเคียงระหว่างคู่อนุกรมโปรตีนโดยวิธีสมิท-วอเตอร์แมน คือการหารูปแบบ (Pattern) ของอนุกรมโปรตีนที่ยาวที่สุดที่เป็นไปได้ระหว่างคู่อนุกรมโปรตีน หรือเรียกว่า (Global Alignment) โดยการตรวจหาลักษณะที่เหมือนกันในลำดับ ส่วนวิธี PSI-Blast นั้นเป็นวิธีเทียบเคียงแบบเปรียบเทียบบางส่วนของสายอนุกรมลำดับ (Local Alignment) ที่ทดสอบความคล้ายกันระหว่างคู่อนุกรมโปรตีนที่ได้จากแทนที่ นำออก และแทรกอะมิโนลงไปในข้อมูลอนุกรม

3.4 การคำนวนค่าความคล้ายคลึงระหว่างคู่ข้อมูลโดยวิธี MPScore

MPScore หรือ Motif Pairwise Score เป็นวิธีคำนวณความคล้ายคลึงระหว่างคู่ข้อมูลที่นำเสนอในงานวิจัยนี้ที่ปรับปรุงจากการวิธีการวัดความคล้ายคลึงโดยใช้สัมประสิทธิ์แจคการ์ด เนื่องจากวิธีสัมประสิทธิ์แจคการ์ด จะให้ค่าน้ำหนักระหว่างค่าในแอทริบิวต์เป็น 0 เมื่อมีค่าต่างกัน

ส่วนวิธี MPScore จะให้ค่าน้ำหนักระหว่างค่าในแอทริบิวต์อยู่ระหว่าง 0 และ 1 ดังสมการที่ 4

สูตรการคำนวณ MPScore

$$MPScore = \frac{\sum_{i=1}^{|M_A|} \underset{m_b \in M_B}{argmax}\, Score(m_{ai}, m_b)}{|M_A| \times |M_B|} \quad (4)$$

เมื่อ M_A = เซตโมทีฟของข้อมูลสายโปรตีน A

M_B = เซตโมทีฟของข้อมูลสายโปรตีน B

m_a = โมทีฟในเซตโมทีฟของข้อมูลสายโปรตีน A

m_b = โมทีฟในเซตโมทีฟของข้อมูลสายโปรตีน B

$Score(m_{ai}, m_b)$ = ค่าคะแนนที่ได้จากวิธีเทียบความเหมือนระหว่างคู่ข้อมูลสมิท-วอเตอร์แมน

3.5 การคำนวนค่าความคล้ายคลึงระหว่างคู่ข้อมูลโดยวิธี MP-Mix

สูตรการคำนวณ MPMix

$$MPMix = \begin{cases} JScore, & MPScore = 0 \\ MPScore, & JScore = 0 \\ JScore \times MPScore, & Otherwise \end{cases} \quad (5)$$

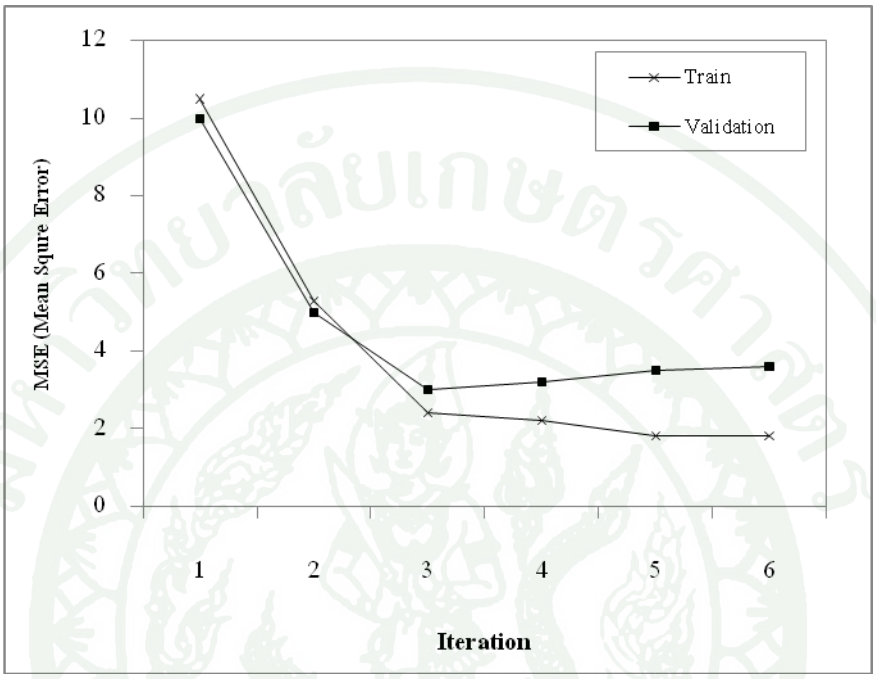
เมื่อ $MPScore$ = เซตโมทีฟของข้อมูลสายโปรตีน A

$JScore$ = เซตโมทีฟของข้อมูลสายโปรตีน B

4. เกณฑ์ที่ใช้พิจารณาการหยุดสอน

เกณฑ์การหยุดที่ถูกใช้โดยทั่วไป จะพิจารณาโดยดูจากค่าความผิดพลาดกำลังสองเฉลี่ย (mean square error, MSE) ที่ได้บนข้อมูลฝึกสอน (Training Set) และข้อมูลตรวจทาน (Validation Set) ค่าความผิดพลาดกำลังสองเฉลี่ยจะถูกคำนวนในแต่ละรอบของการฝึกสอน ค่าความผิดพลาดกำลังสองเฉลี่ยจะลดลงเมื่อจำนวนรอบเพิ่มขึ้นเมื่อทดสอบทั้งในข้อมูลฝึกสอนและข้อมูล

ตรวจทาน รอบที่ควรหยุดสอนคือรอบที่ค่าความผิดพลาดกำลังสองเฉลี่ยในข้อมูลตรวจทานเริ่มเพิ่มขึ้น นั่นคือตัวจำแนกที่ให้ค่าความผิดพลาดกำลังสองเฉลี่ยต่ำสุดบนข้อมูลตรวจทานเป็นตัวจำแนกที่ดีที่สุด ค่าความผิดพลาดกำลังสองในแต่ละรอบแสดงดังภาพที่ 8

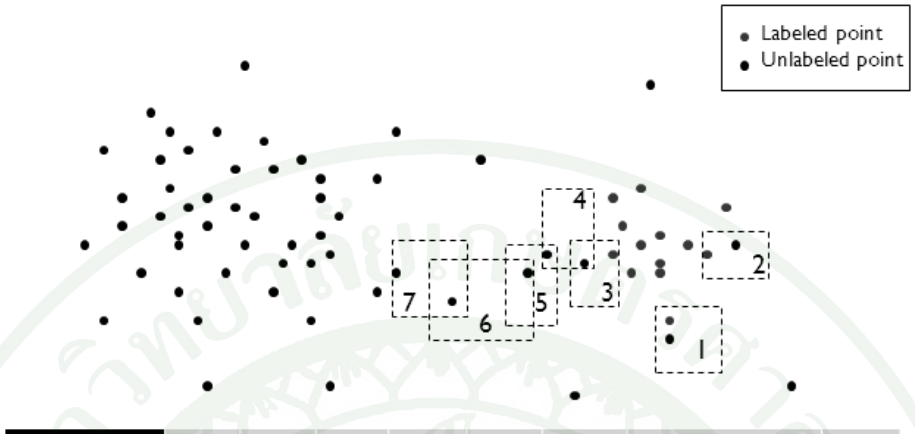


ภาพที่ 8 เปรียบเทียบค่าความผิดพลาดกำลังสองเฉลี่ยบนข้อมูลฝึกสอนและข้อมูลตรวจทานในแต่ละรอบของการฝึกสอนตัวจำแนกโปรตีนฟังก์ชัน รอบที่ 3 เป็นรอบถูกเลือกให้เป็นรอบที่ควรหยุดสอนตัวจำแนก

เกณฑ์การหยุดที่ใช้ในงานวิจัยนี้ มีแนวคิดที่ว่าข้อมูลที่อยู่ในคลาสเดียวกันควรมีความสัมพันธ์กันมาก และข้อมูลที่อยู่ต่างคลาสกันควรมีความสัมพันธ์กันน้อย (พัทริกา, 2552) ดังนั้นเมื่อพิจารณาว่าในแต่ละรอบข้อมูลจากถูกย้ายจากเซต U' ไปยังเซต L ดังภาพที่ 8 ค่าความสัมพันธ์ของรอบปัจจุบันควรใกล้เคียงในรอบก่อนหน้าหรือไม่ควรลดลงมากนัก เมื่อใดก็ตามที่ค่าความสัมพันธ์ลดลงเกินเกณฑ์ที่ผู้ใช้กำหนด แสดงว่าข้อมูลที่ไม่มีความใกล้เคียงกันกับเซต L ถูกย้ายเข้ามามากเกินไป

ในแต่ละรอบ i ของการฝึกสอนตัวจำแนกอัลกอริทึม จะมีการบันทึกค่าความสัมพันธ์ Square Correlation ไว้ในทุกๆรอบดังภาพที่ 9 เพื่อใช้คำนวนอัตราการเปลี่ยนแปลงของ

ความสัมพันธ์ ChangeRate ดังสมการที่ 6 อัลกอริทึมจะหยุดการฝึกสอนเมื่อค่าที่คำนวณได้เกินค่าที่กำหนด



iteration	1	2	3	4	5	6	...	i
correlation	0.87	0.87	0.88	0.86	0.86	0.82		

ภาพที่ 9 สัมประสิทธิ์สหสัมพันธ์ในแต่ละรอบของการรันอัลกอริทึม

สูตรการคำนวณ ChangeRage

$$ChangeRate = \frac{|Corr_i - Corr_{i-1}|}{Corr_{max} - Corr_{min}} \quad (6)$$

โดย $Corr_i$ = ค่าสัมประสิทธิสหสัมพันธ์ในรอบที่ i

$Corr_{i-1}$ = ค่าสัมประสิทธิสหสัมพันธ์ในรอบที่ i-1

$Corr_{max}$ = ค่าสัมประสิทธิสหสัมพันธ์มากที่สุด

$Corr_{min}$ = ค่าสัมประสิทธิสหสัมพันธ์น้อยที่สุด

5. วิธีการวัดผล

งานวิจัยนี้ทำการทดลองบนชุดข้อมูลจาก UniProtKB/Swiss-Prot ที่มีจำนวนข้อมูลฝึกสอน 9,905 ข้อมูลทดสอบ 3,751 ข้อมูลตรวจทาน 3,751 และข้อมูลที่ไม่ทราบคลาส 47,619 แถว จำนวนกลุ่มฟังก์ชัน 763 คลาส ซึ่งข้อมูลที่ใช้มีสัดส่วนในแต่ละคลาสไม่สมดุล (Unbalance Data) และคลาสที่ต้องการจำแนกมีจำนวนมาก (Multiple Class)

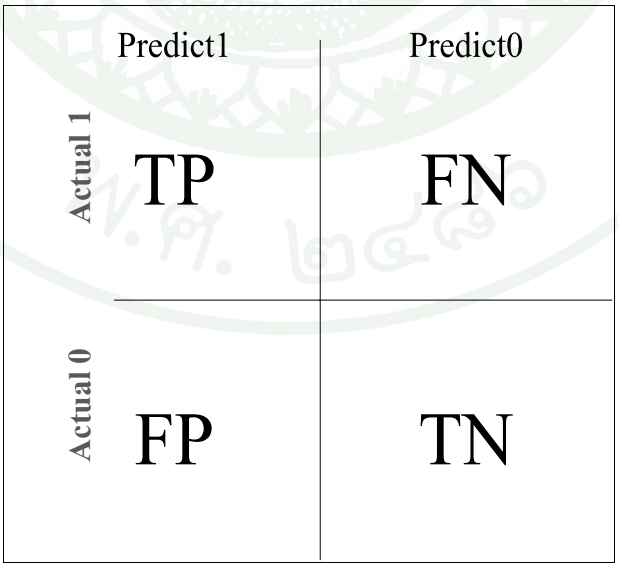
การวัดประสิทธิภาพของตัวทำนายในงานวิจัยนี้คำนวณโดยใช้ค่าความเที่ยง (Precision), รี-คอล (Recall), เอฟเมเชอร์ (F-Measure) และค่าความแม่น (Accuracy) ดังสมการที่ 7, 8, 9 และ 10 ตามลำดับ

ค่าความเที่ยง เป็นการวัดประสิทธิภาพตัวจำแนก ถ้าให้คลาสที่เราสนใจคือคลาสบวกและคลาสอื่นๆเป็นคลาสลบความเที่ยง คือความถูกต้องของคลาสบวกที่ตัวจำแนกทำนายได้ คำนวนจาก จำนวนข้อมูลที่จำแนกถูกในคลาสบวก หรือ TP (True Positive) ข้อมูลที่จำแนกผิดในคลาสบวกหรือ FP (False Positive) แสดงใน Confusion Matrix ดังภาพที่ 10

รีคอลเป็นมาตรวัดประสิทธิภาพที่แสดงความถูกต้องของตัวจำแนกว่าในจำนวนผลเฉลยที่เป็นคลาสบวกทั้งหมด ตัวจำแนกสามารถทำนายได้แม่นเป็นสัดส่วนเท่าใด

เอฟเมเชอร์เป็นการวัดประสิทธิภาพของตัวจำแนกโดยภาพรวม และเนื่องจากเป็นการทดลองของข้อมูลที่มีคลาสมากกว่าสองคลาส ค่าเอฟเมเชอร์ คำนวนโดยวิธี Macro Averaging F-measure % (Yimming Yang *et al.*, 1999) ซึ่งได้ค่าเฉลี่ยของเอฟเมเชอร์ ในแต่ละคลาส

ค่าความแม่นเป็นอีกหนึ่งมาตรวัด ที่บอกสัดส่วนของแถวข้อมูลที่ตัวจำแนกทำนายถูกเทียบกับจำนวนข้อมูลทดสอบทั้งหมด



ภาพที่ 10 Confusion Matrix

สูตรการวัดประสิทธิภาพของตัวจำแนกฟังก์ชันโปรตีน

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F - Measure = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

เมื่อ $TP(True\ Positive)$ = จำนวนตัวอย่างที่ทายถูกในคลาสบวก

$FP(False\ Positive)$ = จำนวนตัวอย่างที่ทายผิดเป็นคลาสบวก

$TN(True\ Negative)$ = จำนวนตัวอย่างที่ทายถูกในคลาสลบ

$FN(True\ Negative)$ = จำนวนตัวอย่างที่ทายผิดเป็นคลาสลบ

ผลและวิจารณ์

ผล

1. การทดลอง

เพื่อทดสอบประสิทธิภาพของตัวจำแนกฟังก์ชันโปรตีนเมื่อมีการปรับปรุงเกณฑ์การเลือกข้อมูลและเกณฑ์การหยุดสอนเปรียบเทียบกับวิธีการอื่นๆ งานวิจัยนี้ใช้ข้อมูลจากฐานข้อมูล UniportKB/Swissport เป็นข้อมูลทดสอบและออกแบบการทดลองเพื่อวัดประสิทธิภาพของตัวจำแนกเมื่อใช้เกณฑ์การเลือกข้อมูลที่แตกต่างกัน นอกจากนั้นยังทำการทดลองเพื่อวัดประสิทธิภาพของตัวจำแนกเมื่อใช้เกณฑ์การหยุด MSE เทียบกับวิธี ChangeRate การทดลองแบ่งเป็น 11 การทดลอง รายละเอียดดังตารางที่ 3

ในการทดลอง จะฝึกสอนตัวจำแนกด้วย อัลกอริทึม SVM กำหนดพารามิเตอร์ Kernel = Linear ใช้ไลบารี LibSVM (Chih-Chung Chang, 2001)

ตารางที่ 3 รายละเอียดการทดลอง

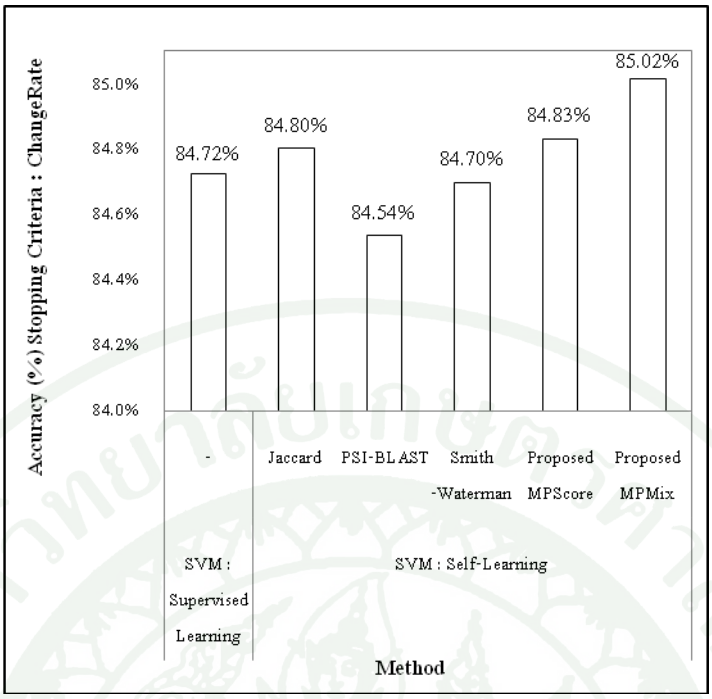
การทดลองที่	เกณฑ์การเลือกข้อมูล	เกณฑ์การหยุดสอน
1	Jaccard	MSE
2	PSI-Blast	MSE
3	Smith-Waterman	MSE
4	MPScore	MSE
5	MPMix	MSE
6	Jaccard	ChangeRate
7	PSI-Blast	ChangeRate
8	Smith-Waterman	ChangeRate
9	MPScore	ChangeRate
10	MPMix	ChangeRate
11	SVM	-

2. ผลการทดลอง

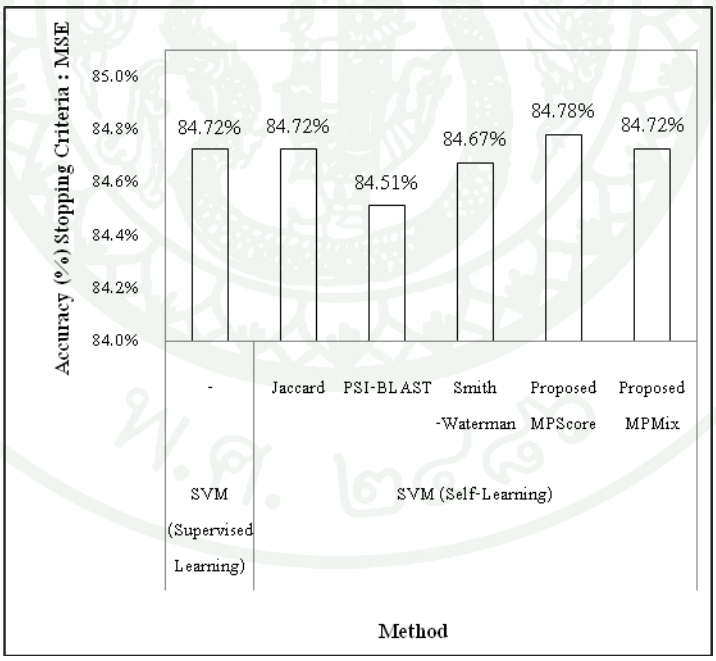
2.1 ผลการทดลองด้านความแม่น (Accuracy)

เมื่อให้เกณฑ์การหยุดคือ ChangeRate และปรับเกณฑ์การเลือกข้อมูลเป็นแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix ตามลำดับ พบว่าตัวจำแนกที่ใช้เกณฑ์การเลือกข้อมูลแต่ละเกณฑ์ให้ค่าความแม่นที่แตกต่างกันไม่มากนักแสดงดังภาพที่ 11 ค่า ความแม่นที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูลแจคการ์ด เท่ากับ 84.70%, MPScore 84.83% และ MPMix 85.02% สูงกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน (Supervised Learning) ที่ให้ค่าความแม่น 84.72% ในทางกลับกันค่าที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูล PSI-Blast 84.54% และสมิธ-วอเตอร์แมน 84.70% ให้ค่าความแม่น ที่ต่ำกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เทียบเกณฑ์การเลือกข้อมูลในทุกเกณฑ์เมื่อใช้ ChangeRate เป็นเกณฑ์การหยุดพบว่า เกณฑ์การเลือกข้อมูล MPMix ให้ความแม่น สูงสุด เกณฑ์การเลือกข้อมูล PSI-Blast ให้ค่าต่ำสุด

เมื่อให้เกณฑ์การหยุดคือ MSE และปรับเกณฑ์การเลือกข้อมูลเป็นแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix ตามลำดับ พบว่าตัวจำแนกที่ใช้เกณฑ์การเลือกข้อมูลแต่ละเกณฑ์ให้ค่าความแม่น ที่แตกต่างกันไม่มากนักแสดงดังภาพที่ 12 ค่าความแม่น ที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูลแจคการ์ด เท่ากับ 84.72% และ MPMix 84.72% ไม่ต่างกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน ค่าที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูล MPScore 84.78% ให้ค่าความแม่น สูงกว่า ส่วนสมิธ-วอเตอร์แมน 84.75% และ PSI-Blast 84.51% มีค่าต่ำกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เทียบเกณฑ์การเลือกข้อมูลในทุกเกณฑ์เมื่อใช้เกณฑ์การหยุด MSE พบว่าเกณฑ์การเลือกข้อมูล MPScore ให้ความแม่น สูงสุด เกณฑ์การเลือกข้อมูล PSI-Blast ให้ค่าต่ำสุด

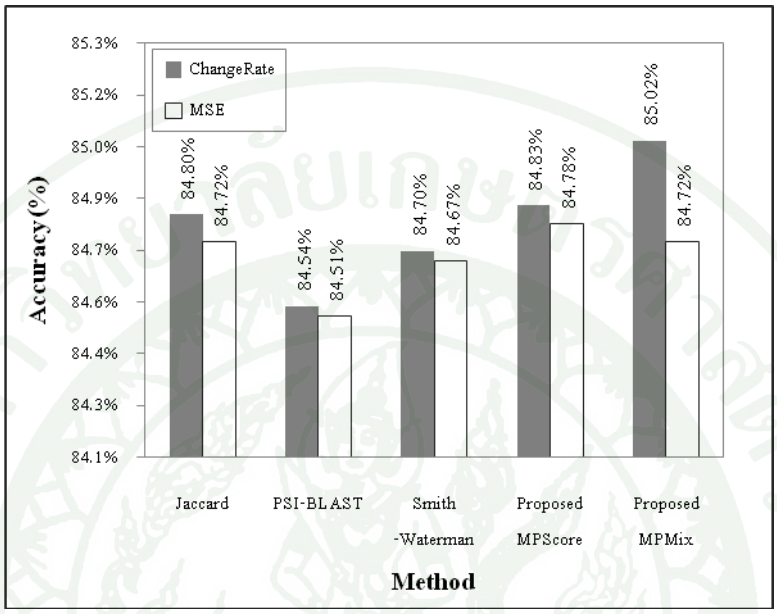


ภาพที่ 11 ผลการทดลองด้านค่าความแม่น โดยมีเกณฑ์การหยุด Change Rate



ภาพที่ 12 ผลการทดลองด้านความแม่น โดยมีเกณฑ์การหยุด MSE

ในมุมมองสำหรับเกณฑ์การหยุด MSE และ ChangeRate พบว่าเกณฑ์การหยุดทั้งสองวิธีให้ผลความแม่น ที่แตกต่างกันแสดงดังภาพที่ 13 แนวโน้มของผลที่ได้จากการใช้เกณฑ์การหยุด ChangeRate ให้ค่าความแม่น สูงกว่าในเกณฑ์การหยุดแบบ MSE



ภาพที่ 13 เปรียบเทียบผลการทดลองด้านความแม่น ทั้ง 2 เกณฑ์การหยุดสอน

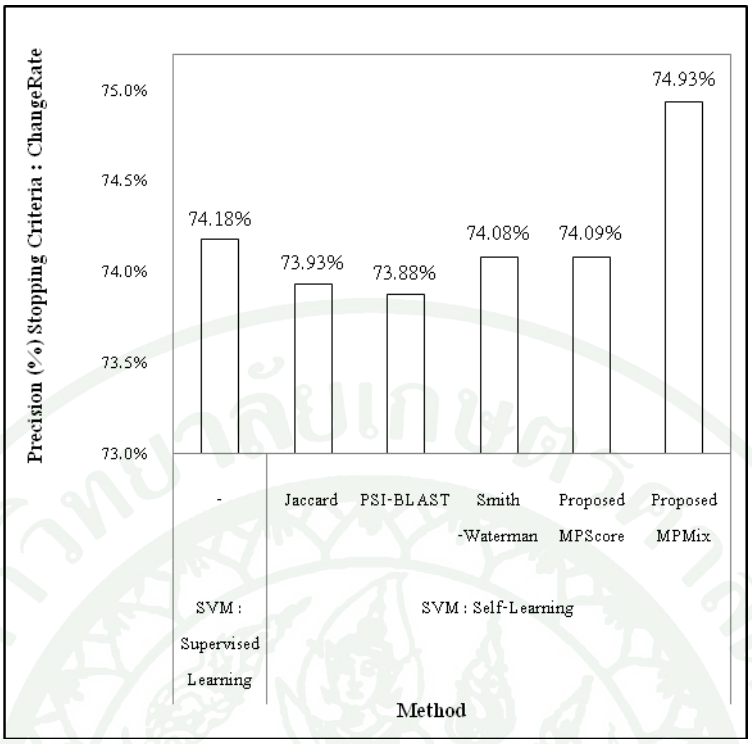
2.1.1 สรุปผลการทดลอง ด้านความแม่น

เทคนิคการเรียนรู้แบบเรียนรู้ด้วยตนเองที่เป็นการเรียนรู้แบบกึ่งมีผู้สอนให้ประสิทธิภาพของตัวแจกแนกด้านความแม่นสูงขึ้นเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เมื่อปรับวิธีในการเลือกข้อมูลวิธีที่แตกต่างกันพบว่าเกณฑ์การเลือกข้อมูลมีผลต่อความแม่นยำของตัวจำแนก เกณฑ์การเลือกข้อมูลวิธี MPMix ให้ค่าความแม่นสูงสุดและมีค่ามากกว่าการฝึกสอนข้อมูลโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน นอกจากนั้นเกณฑ์การหยุดสอนก็มีผลต่อความแม่นของตัวจำแนกเช่นเดียวกัน เกณฑ์การหยุดสอน MPMix ที่งานวิจัยนี้นำเสนอให้ค่าความแม่นสูงกว่าวิธี MSE

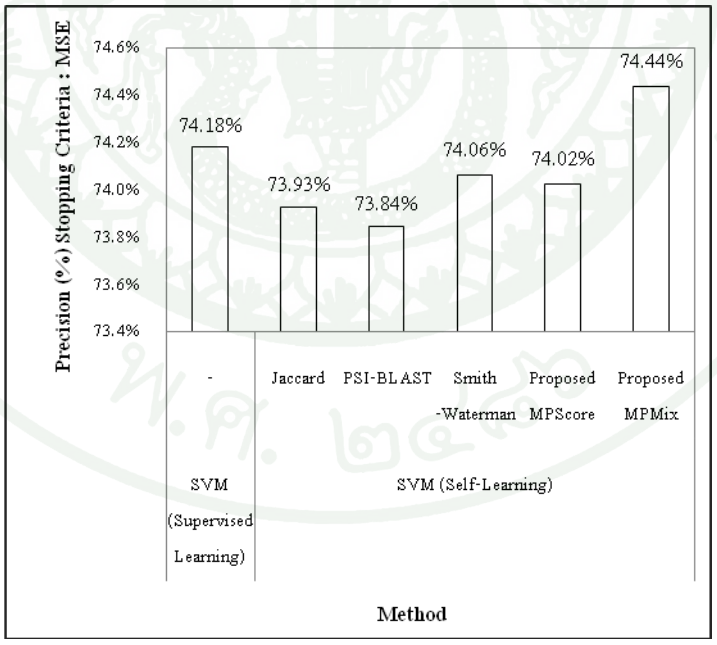
2.2 ผลการทดลองด้านความเที่ยง

เมื่อให้เกณฑ์การหยุดคือ ChangeRate และปรับเกณฑ์การเลือกข้อมูลเป็นแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix ตามลำดับ พบว่าตัวจำแนกที่ใช้เกณฑ์การเลือกข้อมูลแต่ละเกณฑ์ให้ค่าความเที่ยง ที่แตกต่างกันไม่มากนักแสดงดังภาพที่ 14 ค่าความเที่ยง ที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูลแจคการ์ด เท่ากับ 73.93%, PSI-Blast 73.88%, สมิธ-วอเตอร์แมน 73.08% และ MPScore 74.093% ต่ำกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน (Supervised Learning)ที่ให้ค่าความเที่ยง 74.18% ในทางกลับกันค่าที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูล MPMix คือ 74.93% ให้ค่าความเที่ยง ที่สูงกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เทียบเกณฑ์การเลือกข้อมูลในทุกเกณฑ์เมื่อใช้ ChangeRate เป็นเกณฑ์การหยุดพบว่า เกณฑ์การเลือกข้อมูล MPMix ให้ความเที่ยง สูงสุด เกณฑ์การเลือกข้อมูล PSI-Blast ให้ค่าต่ำสุด

เมื่อให้เกณฑ์การหยุดคือ MSE และปรับเกณฑ์การเลือกข้อมูลเป็นแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix ตามลำดับ พบว่าตัวจำแนกที่ใช้เกณฑ์การเลือกข้อมูลแต่ละเกณฑ์ให้ค่าความเที่ยง ที่แตกต่างกันดังภาพที่ 15 ค่าความเที่ยง ที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูลแจคการ์ด เท่ากับ 73.93%, PSI-Blast 73.84%, สมิธ-วอเตอร์แมน 74.06% และ MPScore 72.02% ต่ำกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน (Supervised Learning)ที่ให้ค่าความเที่ยง 74.18% พิจารณาค่าที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูล MPMix คือ 74.44% ให้ค่าความเที่ยง ที่สูงกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เทียบเกณฑ์การเลือกข้อมูลในทุกเกณฑ์เมื่อใช้ MSE เป็นเกณฑ์การหยุดพบว่า เกณฑ์การเลือกข้อมูล MPMix ให้ความเที่ยง สูงสุด เกณฑ์การเลือกข้อมูล PSI-Blast ให้ค่าต่ำสุดเช่นเดียวกับการทดลองที่กำหนดให้เกณฑ์การหยุดคือ ChangeRate

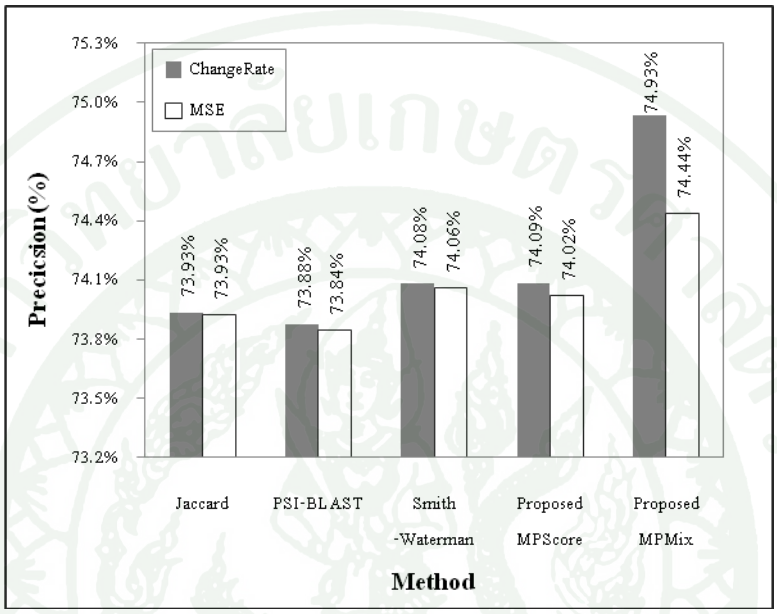


ภาพที่ 14 ผลการทดลองด้านค่าความเที่ยง โดยมีเกณฑ์การหยุด Change Rate



ภาพที่ 15 ผลการทดลองด้านค่าความเที่ยง โดยมีเกณฑ์การหยุด MSE

ในมุมมองสำหรับเกณฑ์การหยุด MSE และ ChangeRate เทียบกับเกณฑ์การเลือกข้อมูลแจคการ์ด, PSI-Blast,สมิธ-วอเตอร์แมน, MPScore และ MPMix พบว่าเกณฑ์การหยุดทั้งสองวิธีให้ผลความเที่ยง ที่แตกต่างกัน แนวโน้มของผลที่ได้จากการใช้เกณฑ์การหยุด ChangeRate ให้ค่าความเที่ยง สูงกว่าโดยเฉลี่ยเมื่อเทียบกับตัวจำแนกที่ใช้เกณฑ์การหยุดแบบ MSE ดังภาพที่ 16



ภาพที่ 16 เปรียบเทียบผลการทดลองด้านความเที่ยง ทั้ง 2 เกณฑ์การหยุด

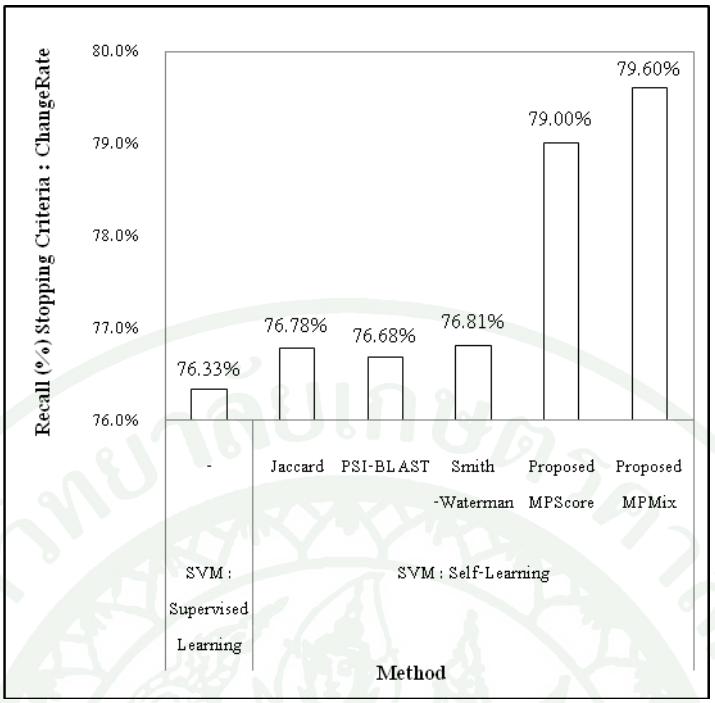
2.2.1 สรุปผลการทดลอง ด้านความเที่ยง

เทคนิคการเรียนรู้แบบเรียนรู้ด้วยตนเองที่เป็นการเรียนรู้แบบกึ่งมีผู้สอนให้ประสิทธิภาพของตัวแจกแนกด้านความเที่ยง ต่ำกว่าโดยเฉลี่ยเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เมื่อปรับวิธีในการเลือกข้อมูลวิธีที่แตกต่างกันพบว่าเกณฑ์การเลือกข้อมูลมีผลต่อความแม่นยำของตัวจำแนกในแง่ค่าความเที่ยง เกณฑ์การเลือกข้อมูลวิธี MPMix ให้ค่าความเที่ยง สูงสุดและเป็นเกณฑ์การเลือกวิธีเดียวที่ให้ค่ามากกว่าการฝึกสอนข้อมูลโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน นอกจากนั้นเกณฑ์การหยุดสอนก็มีผลต่อความแม่นของตัวจำแนกเช่นเดียวกัน เมื่อศึกษาถึงผลของเกณฑ์การหยุดสอนที่มีต่อค่าความเที่ยง พบว่า เกณฑ์การหยุดสอน ChangeRate ที่งานวิจัยนี้นำเสนอให้ค่าความเที่ยง สูงกว่าวิธี MSE เมื่อทดสอบกับทุกเกณฑ์การเลือกข้อมูล

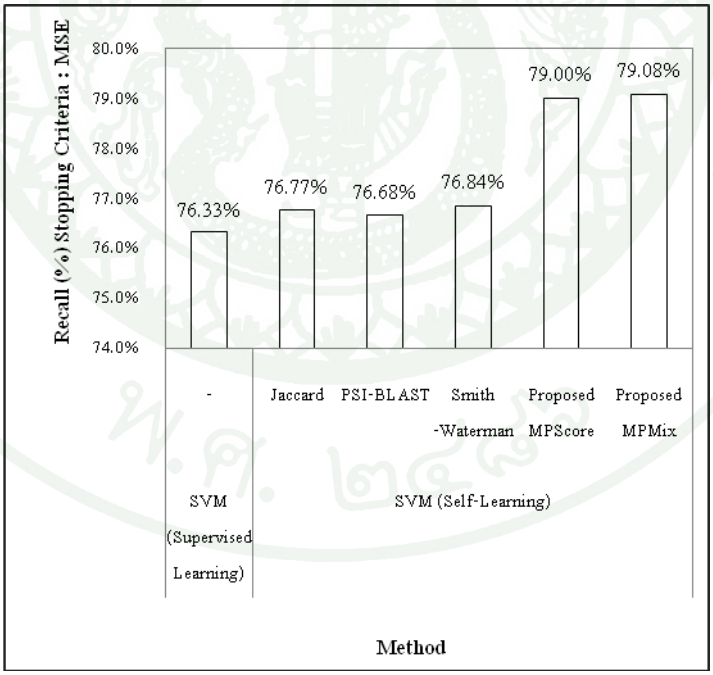
2.3 ผลการทดลองด้านรีคอล

เมื่อให้เกณฑ์การหยุดคือ ChangeRate และปรับเกณฑ์การเลือกข้อมูลเป็นแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix ตามลำดับ พบว่าตัวจำแนกที่ใช้เกณฑ์การเลือกข้อมูลแต่ละเกณฑ์ให้ค่ารีคอล ที่แตกต่างกันอย่างชัดเจนดังภาพที่ 17 ค่ารีคอล ที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูลแจคการ์ด เท่ากับ 76.78%, PSI-Blast 76.68% , สมิธ-วอเตอร์แมน 76.81%, MPScore 79.00% และ MPMix 79.60% จะเห็นว่าเกณฑ์การเลือกข้อมูลทุกเกณฑ์ ให้ค่าสูงกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอนที่ให้ค่ารีคอล 76.33% เกณฑ์การเลือกข้อมูล MPMix ให้รีคอล สูงสุด เกณฑ์การเลือกข้อมูล PSI-Blast ให้ค่าต่ำสุดในบรรดาเกณฑ์การเลือกข้อมูลทุกเกณฑ์ที่ทดลอง

เมื่อให้เกณฑ์การหยุดคือ MSE และปรับเกณฑ์การเลือกข้อมูลเป็นแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix ตามลำดับ พบว่าตัวจำแนกที่ใช้เกณฑ์การเลือกข้อมูลแต่ละเกณฑ์ให้ค่ารีคอล ที่แตกต่างกันอย่างเห็นได้ชัดแสดงดังภาพที่ 18 ค่ารีคอล ที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูลแจคการ์ด เท่ากับ 76.77%, PSI-Blast 76.68% , สมิธ-วอเตอร์แมน 76.84%, MPScore 79.00% และ MPMix 79. 08% จะเห็นว่าเกณฑ์การเลือกข้อมูลทุกเกณฑ์ ให้ค่าสูงกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอนที่ให้ค่ารีคอล 76.33% เทียบเกณฑ์การเลือกข้อมูลในทุกเกณฑ์เมื่อใช้เกณฑ์การหยุด MSE พบว่าเกณฑ์การเลือกข้อมูล MPMix ให้รีคอล สูงสุด เกณฑ์การเลือกข้อมูล PSI-Blast ให้ค่าต่ำสุด

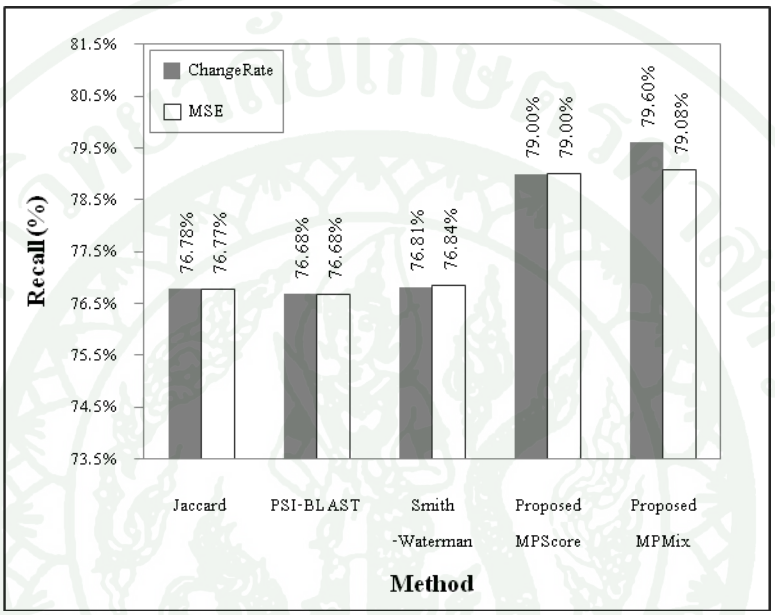


ภาพที่ 17 ผลการทดลองด้านค่ารีคอล โดยมีเกณฑ์การหยุด Change Rate



ภาพที่ 18 ผลการทดลองด้านค่ารีคอล โดยมีเกณฑ์การหยุด MSE

ในมุมมองสำหรับเกณฑ์การหยุด MSE และ ChangeRate เทียบกับเกณฑ์การเลือกข้อมูลแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix พบว่าเกณฑ์การหยุดทั้งสองวิธีให้ผลรีคอล ที่ไม่แตกต่างกันมากนัก แนวโน้มของผลที่ได้จากการใช้เกณฑ์การหยุด ChangeRate ให้ค่ารีคอล สูงกว่าเล็กน้อยโดยเฉลี่ยเมื่อเทียบกับตัวจำแนกที่ใช้เกณฑ์การหยุดแบบ MSE แสดงดังภาพที่ 19



ภาพที่ 19 เปรียบเทียบผลการทดลองด้านรีคอล ทั้ง 2เกณฑ์การหยุด

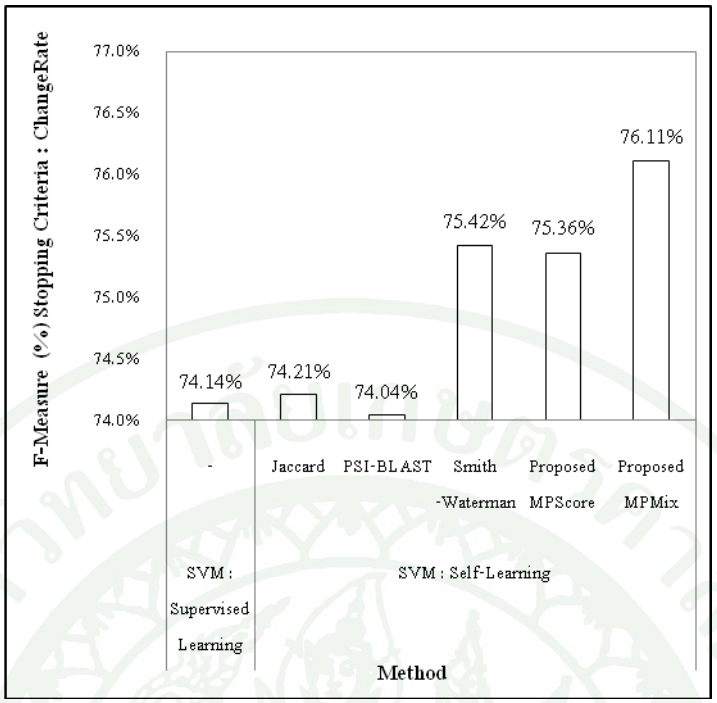
2.3.1 สรุปผลการทดลอง ด้านรีคอล

เทคนิคการเรียนรู้แบบเรียนรู้ด้วยตนเองที่เป็นการเรียนรู้แบบกึ่งมีผู้สอนให้ประสิทธิภาพของตัวแจกแนกด้านรีคอล สูงขึ้นเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เมื่อปรับวิธีในการเลือกข้อมูลวิธีที่แตกต่างกันพบว่าเกณฑ์การเลือกข้อมูลมีผลต่อความแม่นยำของตัวจำแนก เกณฑ์การเลือกข้อมูลวิธี MPMix ให้ค่ารีคอล สูงสุดและมีค่ามากกว่าการฝึกสอนข้อมูลโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน นอกจากนั้นเกณฑ์การหยุดสอนก็มีผลต่อความแม่นของตัวจำแนกเช่นเดียวกัน เกณฑ์การหยุดสอน ChangeRate ที่งานวิจัยนี้นำเสนอให้ค่ารีคอล สูงกว่าวิธี MSE เล็กน้อย

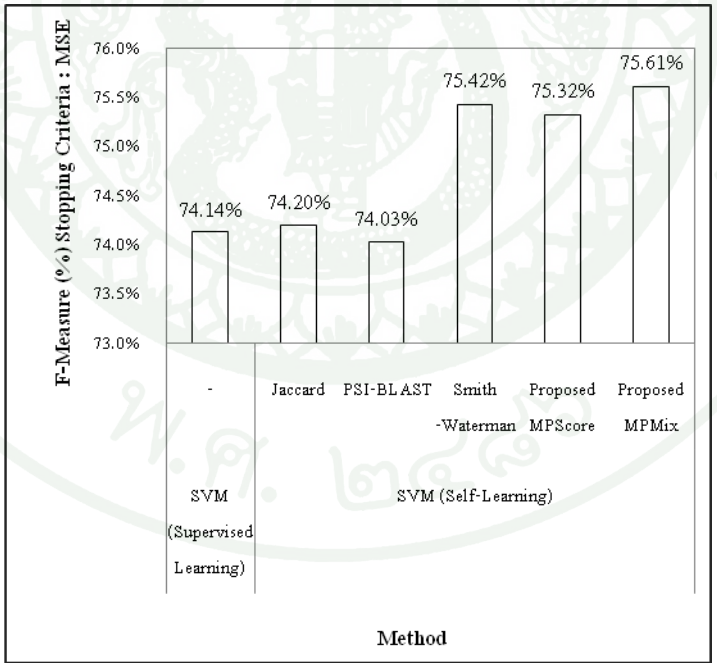
2.4 ผลการทดลองด้านเอฟเมเชอร์

เมื่อให้เกณฑ์การหยุดคือ ChangeRate และปรับเกณฑ์การเลือกข้อมูลเป็นแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix ตามลำดับ พบว่าตัวจำแนกที่ใช้เกณฑ์การเลือกข้อมูลแต่ละเกณฑ์ให้ค่าเอฟเมเชอร์ ที่แตกต่างกันมากแสดงดังภาพที่ 20 ค่าเอฟเมเชอร์ ที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูลแจคการ์ด เท่ากับ 74.21%, สมิธ-วอเตอร์แมน 75.42%, MPScore 75.36% และ MPMix 76.11% สูงกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน ที่ให้ค่าเอฟเมเชอร์ 74.14% ในทางกลับกันค่าที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูล PSI-Blast 74.21% ให้ค่าเอฟเมเชอร์ ที่ต่ำกว่าเมื่อเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เทียบเกณฑ์การเลือกข้อมูลในทุกเกณฑ์เมื่อใช้ ChangeRate เป็นเกณฑ์การหยุด พบว่า เกณฑ์การเลือกข้อมูล MPMix ให้เอฟเมเชอร์ สูงสุด เกณฑ์การเลือกข้อมูล PSI-Blast ให้ค่าต่ำสุด

เมื่อให้เกณฑ์การหยุดคือ MSE และปรับเกณฑ์การเลือกข้อมูลเป็นแจคการ์ด, PSI-Blast, สมิธ-วอเตอร์แมน, MPScore และ MPMix ตามลำดับ พบว่าตัวจำแนกที่ใช้เกณฑ์การเลือกข้อมูลแต่ละเกณฑ์ให้ค่าเอฟเมเชอร์ ที่แตกต่างกันแสดงดังภาพที่ 21 ค่าเอฟเมเชอร์ ที่ได้จากตัวจำแนกที่เรียนรู้ด้วยตนเองเมื่อใช้เกณฑ์การเลือกข้อมูลแจคการ์ด เท่ากับ 74.20%, สมิธ-วอเตอร์แมน 75.42% , MPScore 75.32%, และ MPMix 75.61% เพิ่มขึ้นเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน โดยเกณฑ์การเลือกข้อมูล MPMix ให้ค่าเอฟเมเชอร์ สูงสุด เกณฑ์ PSI-Blast ให้ค่าต่ำสุด

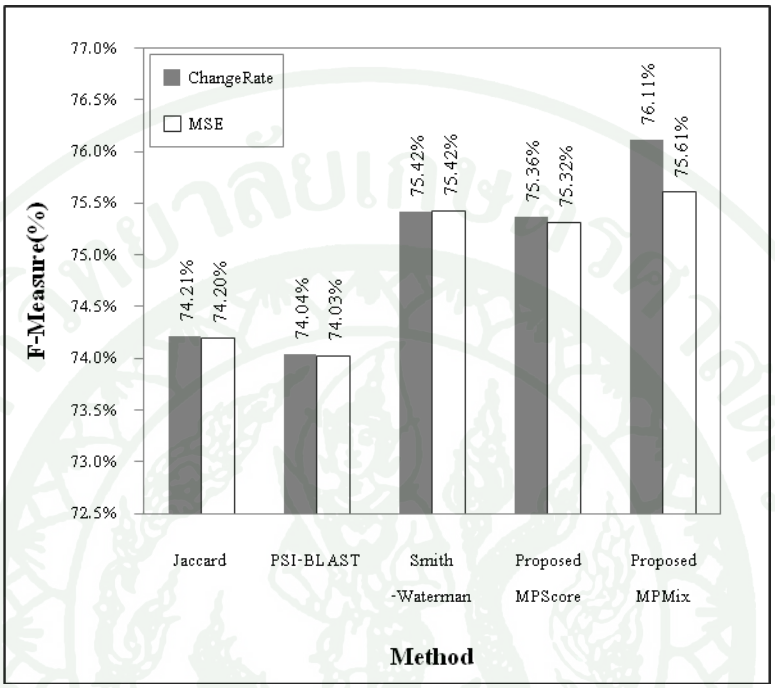


ภาพที่ 20 ผลการทดลองด้านค่าเอฟเมเชอร์ โดยมีเกณฑ์การหยุด Change Rate



ภาพที่ 21 ผลการทดลองด้านค่าเอฟเมเชอร์ โดยมีเกณฑ์การหยุด MSE

จากภาพที่ 22 ทำการเปรียบเทียบเอฟเมเชอร์ ระหว่างตัวจำแนกที่ใช้เกณฑ์การหยุด MSE และเกณฑ์การหยุด ChangeRate พบว่าตัวจำแนกที่ใช้เกณฑ์การหยุด ChangeRate ให้ค่าเอฟเมเชอร์ สูงกว่าในเกณฑ์การหยุด MSE



ภาพที่ 22 เปรียบเทียบผลการทดลองด้านเอฟเมเชอร์ ทั้ง 2เกณฑ์การหยุด

2.3.1 สรุปผลการทดลอง ด้านเอฟเมเชอร์

เทคนิคการเรียนรู้แบบเรียนรู้ด้วยตนเองที่เป็นการเรียนรู้แบบกึ่งมีผู้สอนให้ประสิทธิภาพของตัวแจกแนกด้านเอฟเมเชอร์ สูงขึ้นเทียบกับตัวจำแนกที่เรียนรู้แบบมีผู้สอน เมื่อปรับวิธีในการเลือกข้อมูลวิธีที่แตกต่างกันพบว่าเกณฑ์การเลือกข้อมูลมีผลต่อความแม่นยำของตัวจำแนก เกณฑ์การเลือกข้อมูลวิธี MPMix ให้ค่าเอฟเมเชอร์ สูงสุดและมีค่ามากกว่าการฝึกสอนข้อมูลโดยใช้เทคนิคการเรียนรู้แบบมีผู้สอน นอกจากนั้นเกณฑ์การหยุดสอนก็มีผลต่อความแม่นของตัวจำแนกเช่นเดียวกัน เกณฑ์การหยุดสอน MPMix ที่งานวิจัยนี้นำเสนอให้ค่าเอฟเมเชอร์ สูงกว่าวิธี MSE ดังนั้นเกณฑ์การหยุด ChangeRate และเกณฑ์การเลือกข้อมูล MPMix นำเสนอในงานวิจัยนี้ให้ค่าประสิทธิภาพเอฟเมเชอร์ สูงกว่าทุกวิธีที่นำมาทดลอง

2.5 รายการโปรตีนฟังก์ชันที่ผลการทดลองเพิ่มขึ้นเมื่อใช้เกณฑ์การเลือกข้อมูล MPScore และ MPMix

จากการทดลอง เมื่อดูรายละเอียดตามฟังก์ชันที่ใช้เทคนิคการเลือกข้อมูล MPScore และ MPMix พบว่าเทคนิคดังกล่าวสามารถทำนายบางฟังก์ชันเซตได้ถูกต้องเพิ่มขึ้น จากเดิมที่ทำนายไม่ถูกเลย หรือ ทำนายได้ถูกน้อย รายละเอียดของฟังก์ชันเซตที่สามารถทำนายได้ถูกต้องเพิ่มขึ้นแสดงดังตารางที่ 4

ตารางที่ 4 รายการโปรตีนฟังก์ชันที่ผลการทดลองเพิ่มขึ้นเทียบกับการเรียนรู้แบบมีผู้สอน

Function IDS	% Improve	
	MPScore	MPMix
GO:0003252, GO:0003695, GO:0005250, GO:0005282, GO:0009218		100.00%
GO:0000162, GO:0003707, GO:0003809, GO:0003901, GO:0009890, GO:0010183, GO:0012966		12.50%
GO:0003466, GO:0005747, GO:0009246, GO:0016859		100.00%
GO:0003484, GO:0003488, GO:0003753, GO:0005335		100.00%
GO:0003707	10.00%	
GO:0003707, GO:0003901, GO:0004856, GO:0006187, GO:0009188, GO:0009190, GO:0009595, GO:0020716, GO:0020741		9.09%
GO:0003707, GO:0004856, GO:0009188, GO:0009595, GO:0009920, GO:0020716, GO:00020741		100.00%
GO:0003809, GO:0013303, GO:0019265	100.00%	50.00%
GO:0000775, GO:0005250, GO:0009219		2.63%
GO:0000775, GO:0005250, GO:0009219, GO:0009616	7.69%	
GO:0002137, GO:0005919		13.16%
GO:0002207, GO:0013794		33.33%
GO:0002627, GO:0003755, GO:0004618, GO:0005889, GO:0005919		33.33%
GO:0000162, GO:0002136, GO:0005919		100.00%
GO:0003012, GO:0003014, GO:0003048, GO:0003707, GO:0004581		37.50%
GO:0002627, GO:0004618, GO:0005889, GO:0005919	12.50%	
GO:0005303	100.00%	100.00%
GO:0003516, GO:0004906, GO:0009219		14.29%

วิจารณ์

จากผลการทดลองเปรียบเทียบค่าความแม่น,ความเที่ยง,รีคอล และเอฟเมเชอร์ ระหว่างเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนแบบเรียนรู้ด้วยตนเองและเทคนิคการเรียนรู้แบบมีผู้สอนในทุกการทดลอง พบว่าเทคนิคการเรียนรู้แบบกึ่งมีผู้สอนแบบเรียนรู้ด้วยตนเองเป็นเทคนิคที่ช่วยให้ประสิทธิภาพของตัวจำแนกโปรตีนฟังก์ชันเพิ่มขึ้นโดยภาพรวม แต่อย่างไรก็ตามประสิทธิภาพของเทคนิคการเรียนรู้แบบเรียนรู้ด้วยตนเองขึ้นอยู่กับเกณฑ์การเลือกข้อมูลและเกณฑ์การหยุดที่เหมาะสม เมื่อพิจาณาการทดลองที่ฝึกสอนตัวจำแนกข้อมูลโดยเทคนิคการเรียนรู้แบบเรียนรู้ด้วยตนเองที่ใช้เกณฑ์การเลือกข้อมูลแจคการ์ด และ PSI-Blast พบว่าวิธีแจคการ์ด ให้ประสิทธิภาพดีกว่า PSI-Blast และ สมิธ-วอเตอร์แมน เมื่อพิจารณากลับไปดูวิธีการคำนวณค่าความคล้ายคลึงแล้ว เกณฑ์การเลือกข้อมูลแบบแจคการ์ด เป็นการเทียบความเหมือนของกลุ่มโมทีฟ ซึ่งการศึกษานี้ใช้โมทีฟเป็นคุณลักษณะ (feature) ในการฝึกสอนตัวจำแนกฟังก์ชันของโปรตีน แต่ PSI-Blast และ สมิธ-วอเตอร์แมน เป็นการเทียบลำดับอนุกรมของอักขระในสายโปรตีนระหว่างคู่ข้อมูล ดังนั้นเกณฑ์การเลือกข้อมูลโดยใช้แจคการ์ด เป็นเกณฑ์ที่เหมาะสมกว่า PSI-Blast และ สมิธ-วอเตอร์แมน

เมื่อเปรียบเทียบเกณฑ์การเลือกข้อมูลวิธี PSI-Blast และ สมิธ-วอเตอร์แมน พบว่าวิธีสมิธ-วอเตอร์แมน ให้ผลการทำนายที่ดีกว่า แสดงว่าวิธีเทียบเคียงลำดับอนุกรมโปรตีน สมิธ-วอเตอร์แมน ที่เป็นการเทียงเคียงแบบ Global Alignment นำมาใช้ในเทคนิคแบบกึ่งมีผู้สอนโดยวิธีฝึกสอนด้วยตนเอง ที่มี โมทีฟเป็นคุณลักษณะ ได้ดีกว่าวิธีเทียบเคียงลำดับอนุกรมโปรตีนแบบ PSI-Blast ที่เป็นการเทียบเคียงแบบ Local Alignment

พิจารณาผลการทดลองกรณีที่ใช้เกณฑ์การเลือกข้อมูลแจคการ์ด และเกณฑ์การเลือกข้อมูล MPScore และ MPMix ที่นำเสนอในงานวิจัยนี้พบว่าตัวแนกที่ใช้เกณฑ์แจคการ์ด และ MPScore ให้ประสิทธิภาพใกล้เคียงกัน ส่วนเกณฑ์การเลือกข้อมูล MPMix ให้ประสิทธิภาพของตัวจำแนกเพิ่มขึ้นเมื่อเทียบกับวิธีแจคการ์ด เมื่อพิจารณาจากวิธีการคำนวณค่าความเหมือนของวิธี MPMixพบว่า ใช้เกณฑ์จาก MPScore และ สัมประสิทธิ์แจคการ์ด ในการพิจารณาเลือกข้อมูล แสดงว่าค่าความคล้ายคลึงทั้ง MPScore และ วิธีสัมประสิทธิ์แจคการ์ด ช่วยให้ข้อมูลที่มีผลต่อการทำนายฟังก์ชันของโปรตีนถูกเลือกเพิ่มเข้ามารวมกับข้อมูลฝึกสอนเพิ่มขึ้น

พิจารณาผลการทดลองกรณีที่ใช้เกณฑ์การหยุด MSE และเกณฑ์การหยุด ChangeRate ที่นำเสนอในงานวิจัยนี้พบว่าตัวแนกที่ใช้เกณฑ์ ChangeRate ให้ประสิทธิภาพของตัวจำแนกใกล้เคียงกับวิธี MSE แต่เกณฑ์การหยุดที่พิจารณาจากค่าความผิดพลาดกำลังสองเฉลี่ยที่ต่ำสุดจำเป็นต้องแบ่งข้อมูลส่วนหนึ่งมาใช้เป็นข้อมูลตรวจทาน (Validation Set) ดังนั้นเกณฑ์การหยุด ChangeRate มีความเหมาะสมกับการสร้างตัวจำแนกโปรตีนฟังก์ชันที่ใช้โมทีฟเป็นคุณลักษณะและมีจำนวนข้อมูลฝึกสอนน้อยมากกว่าเกณฑ์การหยุด MSE เนื่องมาจากเกณฑ์ ChangeRate ไม่จำเป็นต้องแบ่งข้อมูลออกมาจากข้อมูลที่ทราบคลาสเพื่อใช้เป็นข้อมูลตรวจทาน

สรุปและข้อเสนอแนะ

สรุป

ปัญหาการทำนายฟังก์ชันการทำงานของโปรตีนเป็นปัญหาหลักทางด้านชีวสารสนเทศศาสตร์(Bioinformatics) ที่มีการค้นคว้าอย่างกว้างขวาง การสร้างตัวทำนายฟังก์ชันสำหรับข้อมูลอนุกรมโปรตีน (Protein sequence) ให้สามารถทำนายคลาสได้อย่างมีประสิทธิภาพนั้น จะต้องอาศัยข้อมูลที่เราทราบคลาส (Labeled Data) เป็นจำนวนมาก ซึ่งข้อมูลประเภทนี้มีอยู่อย่างจำกัด ในขณะที่ข้อมูลส่วนใหญ่ที่มี เป็นข้อมูลที่เราไม่ทราบคลาส (Unlabeled Data) ทำให้ได้ผลการทำนายที่ไม่ดีพอเนื่องจากข้อมูลที่เราทราบคลาสนั้นมีจำนวนน้อยเกินไป

งานวิจัยนี้นำเทคนิคการเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) มาแก้ปัญหากรณีข้อมูลที่ทราบคลาสมีจำนวนน้อย โดยนำข้อมูลที่ทราบและไม่ทราบคลาสมาทำการฝึกสอนร่วมกัน ในงานวิจัยนี้นำเสนอเกณฑ์การเลือกข้อมูลที่ใช้ลักษณะความคล้ายคลึงบนคุณลักษณะโม-ทีฟ และค่าการเทียบเคียงลำดับอนุกรมโปรตีนระหว่างคุณลักษณะของโมทีฟมาเป็นเกณฑ์พิจารณา ในส่วนของเกณฑ์การหยุดสอนนั้น จะพิจารณาจากอัตราการเปลี่ยนแปลงของค่าสัมประสิทธิสหสัมพันธ์กำลังสอง ระหว่างข้อมูลในแต่ละคลาสและข้อมูลที่ถูกเลือกในแต่ละรอบของการฝึกสอน ผลการทดลองบนฐานข้อมูล UniProtKB/Swiss-Prot พบว่าประสิทธิภาพของตัวทำนายคลาสเพิ่มขึ้นอย่างมีนัยสำคัญ

ข้อเสนอแนะ

1. เกณฑ์การหยุดสอน ChangeRate อาจเป็นเกณฑ์การหยุดสอนที่ยังไม่ดีพอเนื่องจากยังเป็นเกณฑ์ที่ให้ค่าความเที่ยง ต่ำกว่าเทคนิค SVM แบบที่เป็นการเรียนรู้แบบมีผู้สอนอยู่เล็กน้อย ควรทำการทดลองเพิ่มเติมเทียบกับเกณฑ์การหยุดอื่นๆ เช่น อัตราการเปลี่ยนแปลงของค่าความคล้ายคลึง เป็นต้น

2. เกณฑ์การเลือกข้อมูลทำการเปรียบเทียบข้อมูลในข้อมูลฝึกสอน กับข้อมูลที่ไม่ทราบคลาส ซึ่งข้อมูลฝึกสอนบางตัวได้จากการรวมข้อมูลไม่ทราบคลาสบางตัวจากรอบก่อนหน้าเข้ากับข้อมูลฝึกสอนเดิม ดังนั้นอาจเกิดกรณีที่ขอบเขตของข้อมูลฝึกสอนถูกขยายออกไปในทิศทางที่ผิดเนื่องจากถูกเหนี่ยวนำจากข้อมูลใหม่ที่ถูกรวมเข้ามาในข้อมูลฝึกสอนในรอบก่อนหน้า

3. การทดลองในงานนี้ใช้วิธีทดลองซ้ำบนข้อมูลชุดเดิมแล้วหาค่าประสิทธิภาพเฉลี่ยบนทุกการทดลอง เนื่องจากข้อมูลฝึกสอนมีจำนวนน้อย อาจใช้วิธีการทดสอบความถูกต้องแบบข้าม (Cross Validation) เพื่อให้ผลการทดลองได้ค่าที่แม่นยำมากขึ้น

4. เกณฑ์การเลือกข้อมูลและเกณฑ์การหยุดที่นำเสนอในงานวิจัยนี้สามารถนำไปประยุกต์ใช้กับขั้นตอนในการเรียนรู้แบบการฝึกสอนร่วมกัน (co-training) ได้

เอกสารและสิ่งอ้างอิง

อมรา คัมภิรานนท์. 2540. **พันธุศาสตร์ของเซลล์.** มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ.

พัทริกา ณ พิกุล และ กฤษณะ ไวยมัย. 2552. การทำนายฟังก์ชันของโปรตีนโดยอาศัยเทคนิคการเรียนรู้แบบกึ่งมีผู้สอน, น. 540-546. *ใน* **งานสัมมนาวิชาการด้านคอมพิวเตอร์และวิศวกรรมแห่งชาติ ครั้งที่ 13**. กรุงเทพฯ.

Anuj R. Shah, Christopher S. Oehmen and Bobbie-Jo Webb-Robertson. 2008. SVM-HUSTLE – an iterative semi-supervised machine learning approach for pairwise protein remote homology detection, pp. 783-790. *In* **Bioinformatics**.

Appel, R.D., A. Bairoch and D.F. Hochstrasser. 1994. A New Generation of Information Retrieval Tools for Biologists : The Example of The EXPASY WWW Server. **Trends in biochemical sciences** 19 (6): 258-260.

Apweiler, R., A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi และ L.L. Yeh. 2004. UniProt: the Universal Protein knowledgebase. **Nucleic Acids Research** 32: Database issue D115-D119.

Blum, A. and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training, pp. 92-100. *In* **Proc. of 11th annual Conf. on Computational learning theory Madison**, Wisconsin, United States

Chapelle, O., Schölkopf, B. and Zien, A. 2006. Semi-Supervised Learning. *In* **MIT Press**. Cambridge, MA.

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM : a library for support vector machines, available source: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 26 May 2011.

Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C. and Huang, T.S. 2004. Semisupervised learning of classifiers: theory, algorithms and their application to human-computer interaction,
pp. 1553-1567. *In* **IEEE Trans. on Pattern Analysis and Machine Intelligence**.

Fisher, R.A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. **Biometrika** 4: 91.

Han, J. and M. Kamber. 2000. Data Mining Concepts and Techniques. **Morgan Kaufmann**. USA.

Huang, T.-M., Kecman, V. and Kopriva, I. 2006. Kernel Based Algorithms for Mining Huge Data Sets. **Springer-Verlag**, Berlin, Heidelberg.

Koji Tsuda, Hyunjung Shin and Bernhard Schölkopf. 2005. Fast protein classification with multiple networks. **Bioinformatics** 21(2): 59-65.

Lesk, Arthur M. 2004. Introduction to Protein science. **OXFORD University Press Inc.**

Li, M., Zhou and Z.-H. 2005. SETRED: self-training with editing, pp. 611-621. *In* **Proc. of 9th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'05)**.

Oliver Chapelle, J. Weston, and B. Schölkopf. 2003. Cluster kernels for semi-supervised learning. **NIPS** 15.

Pang-Ning Tan, Michael Steinbach and Vipin Kumar. 2005. **Introduction to Data Mining**. Pearson Addison-Wesley.

Prosite Database. 2007. Available Source: http://www.expasy.org, 26 May 2011.

Ratanamahatana, C.A. and Keogh, E. 2005. Using Relevance Feedback to Learn Both the Distance Measure and the Query in Multimedia Databases, pp. 16-23. *In* **9th Int. Conf. on Knowledge-Based & Intelligent Information & Engineering Systems**.

Sean R. Eddy. Where did the BLOSUM62 alignment score matrix come from. 2004. **Nature Biotechnology 22**: 1035.

Shahshahani, B.M. and Landgrebe, D.A. 1994. The Effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon, pp. 1087-1095. *In* **IEEE Trans on Geoscience and Remote Sensing.**

Smith Temple F., Waterman Michael S. 1981. Identification of Common Molecular Subsequences, pp. 195–197. *In* **Journal of Molecular Biology**.

UniProt Knowledgebase, **UniProt Knowledgebase Release 12.4. Oct 23, 2007.** Available Source: http://www.expasy.ch, 24 Oct 2007.

Weber, I.T., D.B. McKay and T.A. Steitz. 1982. Two helix DNA binding motif of CAP found in lac repressor and gal repressor. **Nucleic Acids Research** 10: 5085–5102.

Weston, J., C. Leslie, D. Zhou, A. Elisseeff and W. S. Noble. 2004. Semi-Supervised Protein Classification using Cluster Kernels, pp. 595-602. *In* **Advances in Neural Information Processing Systems**.

Xiaojin Zhu. Semi-Supervised Learning Literature Survey. 2005. **Computer Sciences**. University of Wisconsin-Madison.

Xinghua Lu, Chengxiang Zhai, Vanathi Gopalakrishnan and Bruce G Buchanan. 2004. Automatic annotation of protein motif function with Gene Ontology terms. **BMC Bioinformatics.** 5(1): 122.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods, pp. 42–49. *In* **SIGIR-99, 2nd ACM International Conference on Research and Development in Information Retrieval**.

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นางสาวพัทริกา ณ พิกุล
วัน เดือน ปี ที่เกิด	วันที่ 13 มกราคม 2524
สถานที่เกิด	จังหวัดเชียงราย
ประวัติการศึกษา	วศ.บ. (วิศวกรรมไฟฟ้า) คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ (บางเขน) (พ.ศ. 2542)
ตำแหน่งหน้าที่การงานปัจจุบัน	ผู้เชี่ยวชาญการศึกษาอาวุโส
สถานที่ทำงานปัจจุบัน	บริษัท แซส ซอฟท์แวร์ (ประเทศไทย) จำกัด
ผลงานดีเด่นและรางวัลทางวิชาการ	
ทุนการศึกษาที่ได้รับ	