การผสานคุณลักษณะทางสถิติกับนิวรอลเน็ตเวิร์กเพื่อจำแนกผู้ใช้บนข้อความอิสระขนาดสั้น

นายวรุตม์ โรจน์รุ่งวศินกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

A COMBINATION OF STATISTICAL FEATURES AND NEURAL NETWOKRS TO CLASSIFY USERS ON SHORT FREE TEXT

Mr.Warut Roadrungwasinkul

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University

Academic Year 2010

หัวข้อวิทยานิพนธ์	การผสานคุณลักษณะทางสถิติกับนิวรอลเน็ตเวิร์กเพื่อจำแนกผู้ใช้บนข้อความอิสระขนาดสั้น
โดย	นายวรุตม์ โรจน์รุ่งวศินกุล
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ
(อาจารย์ ดร.นัทที นิภานันท์)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.ชลวิช นัทธี)

วรุตม์ โรจน์รุ่งวศินกุล : การผสานคุณลักษณะทางสถิติกับนิวรอลเน็ตเวิร์กเพื่อจำแนกผู้ใช้บนข้อความอิสระขนาดสั้น. (A COMBINATION OF STATISTICAL FEATURES AND NEURAL NETWOKRS TO CLASSIFY USERS ON SHORT FREE TEXT) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร.สุกรี สินธุภิญโญ, 77หน้า.

การจำแนกผู้ใช้จากระยะเวลาในการพิมพ์ เป็นหนึ่งในวิธีการยืนยันตัวตนที่น่าสนใจในปัจจุบัน เพราะสามารถใช้งานได้โดยไม่ต้องติดตั้งอุปกรณ์เพิ่มเติม และยังสามารถใช้ร่วมกับวิธีการจำแนกด้วยชื่อผู้ใช้และรหัสผ่านแบบเดิมได้อีกด้วย งานวิจัยที่ผ่านมาอาจแบ่งได้เป็นสองประเภท คืองานวิจัยที่ศึกษาการใช้ข้อความที่ถูกกำหนดไว้และงานวิจัยที่ศึกษาการใช้ข้อความอิสระ งานวิจัยส่วนมากนั้นจะเป็นศึกษาการใช้ข้อความที่ถูกกำหนดไว้และมีบางงานวิจัยที่รายงานผลการจำแนกได้ดีมาก มีเพียงส่วนน้อยเท่านั้นที่ศึกษาการใช้ข้อความอิสระ และงานวิจัยเหล่านั้นยังต้องการข้อความอิสระขนาดยาวเพื่อให้ได้ผลการจำแนกที่ดี

วิทยานิพนธ์นี้จึงนำเสนอวิธีการใช้ข้อความอิสระขนาดสั้นเพื่อให้ได้ผลการจำแนกที่ดีขึ้นกว่างานวิจัยที่เคยมีมา โดยใช้วิธีการแปลงข้อมูลเวลาในการพิมพ์จากตัวอย่างให้เป็นเวกเตอร์ของคุณลักษณะทางสถิติที่สามารถนำไปใช้งานกับนิวรอลเน็ตเวิร์กได้ และยังเสนอการผสานคุณลักษณะเพื่อให้ได้ผลการจำแนกที่ดีขึ้นอีกด้วย จากผลการทดลองพบว่าวิธีการที่นำเสนอให้ผลการจำแนกผู้ใช้ได้ดีกว่าวิธีอื่นเมื่อใช้ข้อความอิสระขนาดสั้น และให้ผลการจำแนกเทียบเท่ากันกับวิธีอื่นเมื่อใช้ข้อความอิสระที่มีความยาวมากขึ้น

ภาควิชา.วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต.....

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา....2553...

5270486421 : MAJOR COMPUTER ENGINEERING

KEYWORDS : KEYSTROKE DYNAMICS / FREE TEXT / USER CALSSIFICATION / NEURAL NETWORK / PROBABILITY DISTRIBUTION / COMBINATION OF FEATURES

WARUT ROADRUNGWASINKUL : A COMBINATION OF STATISTICAL FEATURES AND NEURAL NETWOKRS TO CLASSIFY USERS ON SHORT FREE TEXT. THESIS ADVISOR : ASST. PROF. SUKREE SINTHUPINYO, Ph.D., 77 pp.

Currently, user classification using keystroke latency patterns is one of the interesting authentication methods because this method does not require any additional devices and can be combined with traditional username-password authentication. Previous research can be categorized into two groups, namely fixed-text and works free-text. Most of the works concerned fixed-text and some of them reported a very good result. Only a few works concerned free-text and those works still require long free-text input sample to obtain a good classification result.

Thus, this thesis proposes a method to use short length free-text input and obtain better user classification result. This method consists of how to transform keystroke latencies from a sample into a vector of statistical features that can be used with neural network and also proposes a combination of features to obtain better user classification result. The results show that the proposed method yields better classification result when using short length free-text input and also gives comparable results when using longer free-text input

Department : Computer Engineering Student's Signature .....

Field of Study : Computer Engineering Advisor's Signature .....

Academic Year : 2010

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้มิอาจเสร็จสมบูรณ์ได้หากปราศจากความช่วยเหลือ จากอาจารย์ที่ปรึกษา ผศ.ดร.สุกรี สินธุภิญโญ ผู้ซึ่งจุดประกายความคิด ให้คำแนะนำและข้อคิดเห็นที่เป็นประโยชน์ต่องานวิจัยชิ้นนี้เสมอมา ผู้วิจัยขอกราบขอพระคุณเป็นอย่างสูงมา ณ ที่นี้

กราบขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ศ.ดร.บุญเสริม กิจสิริกุล อ.ดร.นัทที นิภานันท์ และ ผศ.ดร.ชลวิช นัทธี ที่สละเวลามาให้ข้อเสนอแนะและข้อคิดเห็นที่เป็นประโยชน์ต่อการพัฒนาวิทยานิพนธ์ฉบับนี้

ขอขอบคุณจุฬาลงกรณ์มหาวิทยาลัยที่ได้ประสิทธิ์ประสาทความรู้ในด้านวิชาการ เป็นสถานที่เก็บเกี่ยวประสบการณ์การใช้ชีวิตในมหาวิทยาลัย และทำให้ได้พบกับอาจารย์ เพื่อน พี่ และน้อง ที่คอยช่วยเหลือ แบ่งปันและเติมเต็มช่วงหนึ่งของชีวิตที่มีค่ายิ่ง

สุดท้ายนี้ ขอขอบคุณทุกคนในครอบครัว ที่คอยสนับสนุน เข้าใจ และเป็นกำลังใจตลอดมา จนสามารถทำวิทยานิพนธ์นี้ได้เสร็จสมบูรณ์

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

หน้า

สารบัญภาพ

หน้า

หน้า

บทที่ 1
บทนำ

1.1. ที่มาและความสำคัญของปัญหา

ในปัจจุบัน อินเทอร์เน็ตได้เข้ามามีบทบาทกับชีวิตของคนเรามากขึ้นกว่าแต่ก่อน ด้วยเครือข่ายที่ขยายตัวมากขึ้น มีโปรแกรมประยุกต์ที่เข้าถึงและเชื่อมต่อการใช้งานของผู้ใช้เป็นวงกว้าง แต่การขยายตัวนี้ก็มาพร้อมกับความเสี่ยงในแง่ของความมั่นคงและภาวะส่วนตัว (Security and Privacy) ของผู้ใช้ที่เพิ่มขึ้น โดยทั่วไปในโลกของอินเทอร์เน็ต ระบบมักให้ผู้ใช้ระบุตัวตนด้วยข้อมูลต่าง ๆ เช่น ชื่อผู้ใช้ รหัสผ่าน หรือข้อมูลส่วนตัวอื่น ๆ ซึ่งสามารถถูกผู้โจมตีดักจับและขโมยได้ไม่ยากนัก เมื่อผู้โจมตีล่วงรู้ข้อมูลเหล่านั้นแล้วก็จะสามารถแสดงตนเป็นผู้ใช้ได้โดยระบบไม่สามารถล่วงรู้ได้เลยว่าผู้ที่กำลังควบคุมและใช้งานอยู่นั้นไม่ใช่ผู้ใช้ที่แท้จริง

ชีวมาตร (Biometrics) เป็นอีกหนึ่งแนวทางหนึ่งที่สามารถนำมาใช้เพื่อลดความเสี่ยงดังกล่าวได้ โดยชีวมาตร คือการระบุตัวบุคคลโดยอาศัยข้อมูลที่เป็นลักษณะเฉพาะของบุคคลนั้น ๆ ซึ่งอาจจะเป็นลักษณะเชิงกายภาพ (Physiological) หรือลักษณะเชิงพฤติกรรม (Behavioral) ก็ได้ ข้อมูลเหล่านี้จะเป็นการยืนยันในอีกระดับหนึ่งว่าผู้ที่กำลังแสดงตนแก่ระบบนั้นเป็นตัวผู้ใช้จริง ๆ ตัวอย่างลักษณะที่ใช้ในการจำแนกที่เป็นที่นิยมในปัจจุบัน เช่น ลายนิ้วมือ ใบหน้า ดวงตา และเสียง เป็นต้น หนึ่งในลักษณะที่น่าสนใจที่สามารถนำมาใช้เป็นชีวมาตรได้ คือการจำแนกผู้ใช้โดยอาศัยระยะเวลาในการพิมพ์

การใช้ระยะเวลาในการพิมพ์เป็นข้อมูลในการจำแนกผู้ใช้นั้น มีข้อดีที่เหนือกว่าข้อมูลลักษณะอื่นอยู่ 2 ประการ ประการแรก คือ การเก็บข้อมูลระยะเวลาในการพิมพ์นั้นไม่ต้องอาศัยอุปกรณ์ใด ๆ เพิ่มเติมนอกเหนือไปจากคีย์บอร์ดที่ใช้ตามปกติอยู่แล้ว โดยเพิ่มเพียงโปรแกรมเพื่อเก็บข้อมูลเวลาเท่านั้น ประการที่สอง คือ การเก็บข้อมูลระยะเวลาในการพิมพ์นั้นไม่ได้เป็นการรบกวนการใช้งานปกติของผู้ใช้ เพราะผู้ใช้ต้องพิมพ์ข้อความต่าง ๆ ในช่วงเวลาที่ใช้งานระบบอยู่แล้ว อย่างไรก็ดี ข้อมูลระยะเวลาในการพิมพ์เป็นข้อมูลลักษณะเชิงพฤติกรรมที่อาจเปลี่ยนแปลงได้โดยขึ้นอยู่กับหลาย ๆ ปัจจัย เช่น ความแปรปรวนของระยะเวลาในการพิมพ์ที่มีอยู่ตามธรรมชาติ สภาพของผู้ใช้ที่อาจทำให้พิมพ์ได้ช้าลง การเปลี่ยนคีย์บอร์ดที่ทำให้ผู้ใช้ไม่คุ้นเคย หรือแม้แต่ข้อความที่แตกต่างกันที่ผู้ใช้ได้พิมพ์ลงไปก็ตาม ต่างทำให้ข้อมูลระยะเวลาในการพิมพ์นั้น

อาจมีความแปรปรวนได้มาก ซึ่งการออกแบบวิธีการจำแนกเพื่อให้ได้ความแม่นสูงนั้นจำเป็นต้องพิจารณาถึงปัจจัยเหล่านี้ด้วย

ในงานวิจัยที่ใช้ข้อมูลระยะเวลาในการพิมพ์ในการจำแนกผู้ใช้นั้น อาจแบ่งตามประเภทข้อความที่ใช้ได้เป็น 2 ประเภทใหญ่ ๆ คือ ประเภทที่ใช้ข้อความที่ถูกกำหนดไว้ (Fixed text) และประเภทที่ใช้ข้อความอิสระ (Free text) โดยประเภทแรกจะพิจารณาระยะเวลาในการพิมพ์กับเฉพาะข้อความที่กำหนดไว้แล้วเท่านั้น เช่น ชื่อผู้ใช้ รหัสผ่าน ชื่อ-นามสกุลของผู้ใช้ หรือข้อความอื่น ๆ ที่ถูกกำหนดไว้แล้วแต่แรก ส่วนในประเภทหลังนั้นจะพิจารณาข้อมูลระยะเวลาในการพิมพ์จากข้อความใด ๆ ก็ได้ที่ผู้ใช้พิมพ์เข้ามา

ข้อดีของงานวิจัยที่ใช้ข้อความอิสระ คือ สามารถนำไปใช้ได้โดยมีข้อจำกัดน้อยกว่า และยังสามารถประยุกต์ใช้ให้ระบบทนต่อการโจมตีแบบเล่นซ้ำ (Replay Attack) ซึ่งกระทำโดยผู้โจมตีดักเก็บข้อมูลระยะเวลาในการพิมพ์ไว้ และส่งข้อมูลเดิมซ้ำไปยังระบบเมื่อมีการร้องขออย่างไรก็ดี ในปัจจุบันงานวิจัยประเภทที่ใช้ข้อความที่ถูกกำหนดไว้มีจำนวนค่อนข้างมาก และให้ผลความแม่นที่ค่อนข้างดี ต่างกับงานวิจัยประเภทที่ใช้ข้อความอิสระที่ยังมีจำนวนน้อยอยู่ให้ผลความแม่นที่ยังไม่ดีเท่าแบบที่ใช้ข้อความที่ถูกกำหนดไว้ และยังต้องใช้ข้อความที่มีขนาดความยาวมากเป็นตัวอย่างทดสอบ ถึงจะให้ผลความแม่นในระดับที่น่าพึงพอใจ งานวิจัยนี้จึงเน้นไปที่การพัฒนาวิธีการจำแนกผู้ใช้โดยอาศัยระยะเวลาในการพิมพ์ กับข้อความอิสระที่มีขนาดความยาวน้อย ๆ เพื่อให้ได้ผลความแม่นในระดับที่สูงขึ้นเมื่อเทียบกับงานวิจัยอื่น ๆ ก่อนหน้านี้

การจำแนกผู้ใช้ด้วยระยะเวลาในการพิมพ์นั้น หากได้รับการพัฒนาให้มีความแม่นสูงก็ จะสามารถนำมาใช้เสริมความปลอดภัยให้กับระบบได้ในหลายกรณี เช่น การเสริมความปลอดภัยในการยืนยันตัวตนเข้าสู่ระบบโดยใช้ร่วมกับวิธีการแบบเดิม การตรวจหาและป้องกันผู้รุกรุกระบบที่รับข้อมูลโดยการพิมพ์เป็นหลัก การยืนยันตัวบุคคลในรูปแบบลายเซ็นอิเล็กทรอนิกส์สำหรับโปรแกรมประยุกต์ต่าง ๆ เช่น อีเมล การตอบข้อความบนเว็บบอร์ด การใช้งานเว็บไซต์เครือข่ายสังคมออนไลน์ (Social Network Website) หรือ การใช้งานโปรแกรมส่งข้อความแบบทันที (Instant Messaging) เป็นต้น และหากสามารถศึกษาหาวิธีที่ทำให้ใช้ข้อความอิสระที่มีขนาดความยาวน้อย ๆ โดยยังมีความแม่นสูงได้ ก็จะสามารถนำไปใช้งานได้สะดวกขึ้นในหลาย ๆ สถานการณ์

1.2. วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการจำแนกตัวบุคคลโดยอาศัยระยะเวลาในการพิมพ์ ที่สามารถทำงานได้ดีกับข้อความอิสระที่มีความยาวน้อย ๆ ซึ่งเป็นการลดข้อจำกัดของข้อความที่เป็นอินพุต เพื่อให้สามารถนำไปใช้กับโปรแกรมประยุกต์ได้หลากหลายมากขึ้น อีกทั้งยังเป็นพื้นฐานของงานวิจัยอื่น ๆ ต่อไป

1.3. ขอบเขตของการวิจัย

1) ทำการทดลองในระดับการจำแนก (classification) โดยระบบจะได้รับรู้ข้อมูลบางส่วนของผู้ใช้ทุก ๆ คนที่เกี่ยวข้องเพื่อเป็นข้อมูลฝึก

2) ทำการทดลองกับข้อมูลที่เป็นข้อความอิสระ (free text) ที่ตัดตอนมาจากการใช้งานคอมพิวเตอร์ตามปกติของผู้ใช้ 33 คน คนละ 15000 ตัวอักษร ใช้เวลาในการเก็บข้อมูลประมาณ 1-2 สัปดาห์ โดยข้อมูลจะที่เก็บจะถูกแบ่งเป็นตัวอย่างที่มีความยาวแตกต่างกัน ตั้งแต่ 100 ถึง 1000 ตัวอักษร เพื่อศึกษาผลการจำแนกกับความยาวของตัวอย่างทดสอบ

3) ผลการทดลองที่แสดง เป็นค่าเฉลี่ยจากการทดสอบแบบไขว้ข้าม 10 พับ (10-fold cross validation)

1.4. ขั้นตอนและวิธีดำเนินการวิจัย

1) ศึกษาปัญหาและงานวิจัยเกี่ยวกับการจำแนกตัวบุคคลด้วยข้อมูลการพิมพ์ที่มีผู้ได้นำเสนอมาแล้ว
2) กำหนดหัวข้อปัญหาของงานวิจัย
3) ศึกษาวิธีการเรียนรู้ของเครื่อง และวิธีอื่น ๆ ที่เกี่ยวข้อง เพื่อสร้างแนวคิดในการแก้ไขปัญหา
4) สร้างแนวคิด ปรับปรุง และพัฒนา
5) ทำการทดลองแนวคิดและสมมติฐาน
6) วิเคราะห์ผลการทดลอง
7) สรุปผลและเรียบเรียงวิทยานิพนธ์

1.5. คุณค่าทางวิชาการ

1) นำเสนอวิธีการจำแนกตัวบุคคลโดยอาศัยระยะเวลาในการพิมพ์ ที่สามารถทำงานได้ดีกับข้อความอิสระที่มีความยาวน้อย ๆ

2) สามารถนำการจำแนกตัวบุคคลโดยอาศัยระยะเวลาในการพิมพ์ไปใช้งานในโปรแกรมประยุกต์ต่าง ๆ ได้หลากหลายยิ่งขึ้น โดยมีข้อจำกัดในการใช้งานน้อยลง

3) เป็นพื้นฐานของงานวิจัยอื่น ๆ ในอนาคต

1.6. ผลงานตีพิมพ์จากวิทยานิพนธ์

1) หัวเรื่อง “User Recognition Via Keystroke Latencies Using SOM and Backpropagation Neural Network” โดย สุกรี สินธุภิญโญ วรุตม์ โรจน์รุ่งวศินกุล และ จรูญ จันแทน ในบันทึกการประชุม “ICROS-SICE International Joint Conference 2009” ซึ่งจัดขึ้น ณ เมืองฟุกุโอกะ ประเทศญี่ปุ่น ระหว่างวันที่ 18-21 สิงหาคม 2552

2) หัวเรื่อง “A Combination of Statistical Features and Neural Networks to Classify Users Based on Free Text” โดย วรุตม์ โรจน์รุ่งวศินกุล และ สุกรี สินธุภิญโญ ในบันทึกการประชุม “2010 IRAST International Congress on Computer Applications and Computational Science (CACS 2010)” ซึ่งจัดขึ้น ณ ประเทศสิงคโปร์ ระหว่างวันที่ 4-6 ธันวาคม 2553

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1. ทฤษฎีที่เกี่ยวข้อง

2.1.1. ข้อมูลทั่วไปของระยะเวลาในการพิมพ์

ในการพิจารณาข้อมูลระยะเวลาในการพิมพ์นั้น ข้อมูลเบื้องต้นที่เราจะได้รับ คือ เวลาที่ผู้ใช้กดและปล่อยปุ่มหนึ่ง ๆ บนคีย์บอร์ด ซึ่งเรียกข้อมูลเหล่านี้ว่าคีย์สโตรกไดนามิกส์ (Keystroke Dynamics)

ข้อมูลที่ง่ายที่สุดที่จะได้จากคีย์สโตรกไดนามิกส์ คือ ระยะเวลาที่ผู้ใช้กดปุ่มหนึ่ง ๆ (Keystroke Duration) ซึ่งหาได้จากการนำเวลาที่ผู้ใช้ปล่อยปุ่มลบออกด้วยเวลาที่ผู้ใช้กดปุ่ม ซึ่งก็เป็นหนึ่งคุณลักษณะที่ได้รับความนิยมในงานการจำแนกผู้ใช้ ข้อมูลอีกประเภทที่เป็นที่นิยมเช่นกัน คือการพิจารณาไดกราฟ (Digraph) ซึ่งก็คือการพิจารณาแต่ละคู่ตัวอักษรที่ผู้ใช้พิมพ์ คุณลักษณะที่จะได้จากการพิจารณาก็คือระยะเวลาระหว่างไดกราฟ (Digraph Latency) ที่หาได้จากเวลาที่ผู้ใช้กดปุ่มที่สองลบออกด้วยเวลาที่ผู้ใช้ปล่อยปุ่มแรก Hosseinzadeh, D. และ Krishnan, S. [1] ได้นำเสนอคุณลักษณะที่สามารถหาเพิ่มเติมได้จากการพิจารณาไดกราฟ โดยแยกการพิจารณาระยะเวลาระหว่างไดกราฟออกมาเป็น 4 แบบตามภาพที่ 2-1 คือ

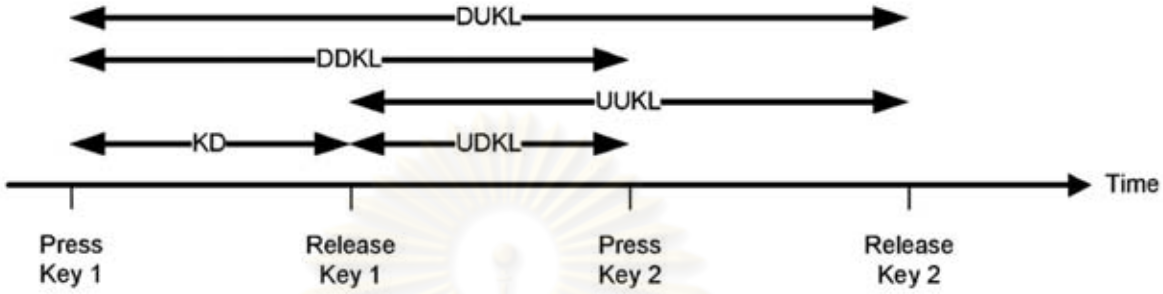
1) DDKL หรือ Press-to-Press Digraph Latency คือระยะเวลาตั้งแต่ผู้ใช้กดปุ่มแรกจนถึงผู้ใช้กดปุ่มที่สอง

2) DUKL หรือ Press-to-Release Digraph Latency คือระยะเวลาตั้งแต่ผู้ใช้กดปุ่มแรกจนถึงผู้ใช้ปล่อยปุ่มที่สอง

3) UDKL หรือ Release-to-Press Digraph Latency คือระยะเวลาตั้งแต่ผู้ใช้ปล่อยปุ่มแรกจนถึงผู้ใช้กดปุ่มที่สอง

4) UUKL หรือ Release-to-Release Digraph Latency คือระยะเวลาตั้งแต่ผู้ใช้ปล่อยปุ่มแรกจนถึงผู้ใช้ปล่อยปุ่มที่สอง

นอกจากนี้ยังมีอีกหลาย ๆ งานวิจัยที่ใช้หลาย ๆ คุณลักษณะร่วมกัน แต่เพื่อความง่ายในการเปรียบเทียบในงานวิจัยนี้จะใช้แค่ DDKL หรือ Press-to-Press Digraph Latency เพียงคุณลักษณะเดียวเท่านั้น



ภาพที่ 2-1 แผนภาพคุณลักษณะที่ได้จากการพิจารณาไดกราฟ [1]

2.1.2. การแจกแจงความน่าจะเป็นของตัวแปรสุ่ม

ตามธรรมชาติแล้ว เราเชื่อว่าเหตุการณ์ต่าง ๆ สามารถถูกจำลองได้ด้วยการแจกแจงความน่าจะเป็นของตัวแปรสุ่มแบบต่าง ๆ เช่นเดียวกันกับในงานวิทยานิพนธ์ฉบับนี้ที่เราจะจำลองความน่าจะเป็นของระยะเวลาในการพิมพ์ไดกราฟหนึ่ง ๆ ด้วยการแจกแจงความน่าจะเป็นของตัวแปรสุ่ม ตัวแปรสุ่มนั้นอาจแบ่งได้เป็นตัวแปรสุ่มแบบต่อเนื่อง และตัวแปรสุ่มแบบไม่ต่อเนื่อง ในขั้นแรกจะแนะนำสัญลักษณ์ที่สำคัญในเรื่องของการแจกแจงความน่าจะเป็นของตัวแปรสุ่ม ดังนี้

ฟังก์ชันแจกแจงแบบไม่ต่อเนื่อง *(Probability mass function หรือ p.m.f)*

ฟังก์ชันการแจกแจงแบบไม่ต่อเนื่อง เป็นฟังก์ชันที่ใช้นิยามการแจกแจงความน่าจะเป็นของตัวแปรสุ่มแบบไม่ต่อเนื่อง ใช้สัญลักษณ์ p_X โดย $p_X(k) = P(X = k)$ ฟังก์ชันแจกแจงนี้จะอธิบายความน่าจะเป็นเมื่อตัวแปรสุ่มมีค่าแต่ละตำแหน่ง เนื่องจากความน่าจะเป็นรวมของทุก ๆ เหตุการณ์มีค่าเท่ากับ 1 ดังนั้น ผลรวมของฟังก์ชันแจกแจงแบบไม่ต่อเนื่องจึงมีค่าเท่ากับ 1 เช่นกัน อธิบายได้จาก $\sum_{k \in \Omega} p_x(k) = 1$ โดย Ω คือ ปริภูมิตัวอย่าง (sample space)

ฟังก์ชันแจกแจงแบบต่อเนื่อง *(Probability density function หรือ p.d.f)*

ฟังก์ชันแจกแจงแบบต่อเนื่อง เป็นฟังก์ชันที่ใช้นิยามการแจกแจงความน่าจะเป็นของตัวแปรสุ่มแบบต่อเนื่อง มีสมบัติเป็นฟังก์ชันต่อเนื่อง f จากจำนวนจริงไปสู่จำนวนจริงบวก และ $\int_{-\infty}^{\infty} f(x)dx = 1$ ใช้สัญลักษณ์ f_X แทนฟังก์ชันแจกแจงแบบต่อเนื่องของตัวแปรสุ่ม X

สำหรับตัวแปรสุ่มต่อเนื่อง X ใด ๆ การคำนวณหาความน่าจะเป็นที่ X จะมีค่าอยู่ในช่วงใด ๆ สามารถหาได้จากพื้นที่ใต้กราฟหรือการหาปริพันธ์ โดยคำนวณจากสมการ

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

ฟังก์ชันการแจกแจงสะสม (*Cumulative distribution function* หรือ *c.d.f.)*

ฟังก์ชันแจกแจงสะสมของตัวแปรสุ่มสุ่ม X ใช้สัญลักษณ์ F_X นิยามได้โดย $F_X(x) = P(X < x) = \int_{-\infty}^{x} f(y)dy$ หรืออีกนัยหนึ่ง เราสามารถหาความน่าจะเป็นที่ X จะมีค่าอยู่ในช่วงใด ๆ ในรูปของฟังก์ชันการแจกแจงสะสมได้จากสมการ

$$P(a \leq X \leq b) = F_X(b) - F_X(a)$$

2.1.2.1. การแจกแจงความน่าจะเป็นของตัวแปรสุ่มแบบปกติ

ตัวแปรสุ่มแบบปกติ เป็นตัวแปรสุ่มแบบต่อเนื่องที่สำคัญเพราะการจำลองแบบระบบในธรรมชาติหลาย ๆ ครั้งสามารถทำได้ด้วยตัวแปรสุ่มชนิดนี้ และยังเกี่ยวข้องกับการจำลองแบบระยะเวลาที่แตกต่างกันของการพิมพ์ไดกราฟหนึ่ง ๆ ซึ่งจะใช้ในวิทยานิพนธ์ฉบับนี้ด้วย

ให้ตัวแปรสุ่ม X เป็นตัวแปรสุ่มที่มีการแจกแจงแบบปกติ สามารถนิยามฟังก์ชันการแจกแจงความน่าจะเป็นของ X ได้ดังสมการที่ 1

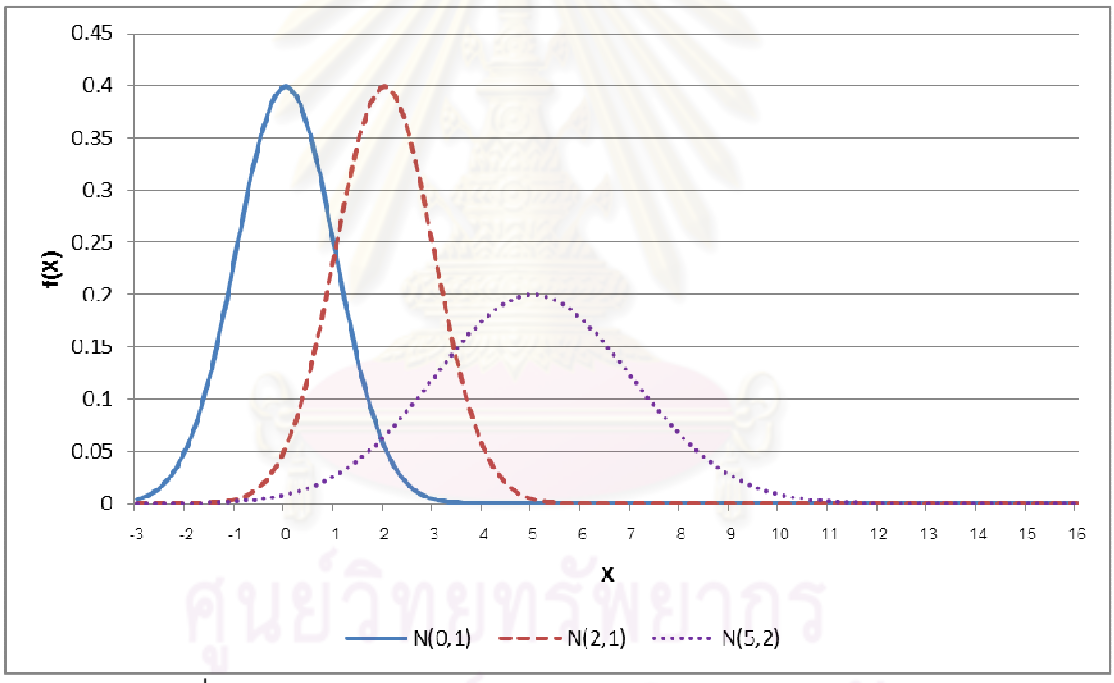
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty \qquad (1)$$

โดยมี μ และ σ^2 เป็นพารามิเตอร์ของการแจกแจง ใช้สัญลักษณ์ $N(\mu, \sigma^2)$

คุณสมบัติของฟังก์ชันการแจกแจงแบบปกติมีดังนี้

1) สมมาตรรอบค่าคาดหวัง (μ)
2) ฟังก์ชันมีค่าตั้งแต่ $-\infty$ ถึง ∞
3) ตำแหน่งของการแจกแจงขึ้นอยู่กับค่าคาดหวัง (μ)
4) ความโด่ง ความแบนของจุดยอดของฟังก์ชันการแจกแจงขึ้นอยู่กับความแปรปรวน (σ^2)

จากคุณสมบัติข้างต้น สามารถแสดงการแจกแจงความน่าจะเป็นแบบปกติที่มีค่าคาดหวังและความแปรปรวนที่แตกต่างกันได้ดังภาพที่ 2-2



ภาพที่ 2-2 รูปร่างของการแจกแจงความน่าจะเป็นของตัวแปรสุ่มแบบปกติ

ในการหาความน่าจะเป็นในช่วงใด ๆ ของตัวแปรสุ่มแบบปกติ จำเป็นจะต้องหาฟังก์ชันการแจกแจงสะสมหรือปริพันธ์ของ f_X ซึ่งไม่สามารถหาออกมาเป็นรูปปิดได้ ทำให้ไม่สะดวกในการคำนวณ เราจึงนิยามตัวแปรสุ่มแบบปกติมาตรฐาน (Standard Normal Distribution) เพื่อช่วยในการหาค่าความน่าจะเป็นของตัวแปรสุ่มแบบปกติอื่น ๆ ได้ โดยนิยามตัวแปรสุ่ม Z เป็นตัวแปรสุ่มแบบปกติมาตรฐาน เมื่อ Z มีฟังก์ชันการแจกแจงความน่าจะเป็น ดังสมการที่ 2

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty \tag{2}$$

นั่นคือ Z เป็นตัวแปรสุ่มแบบปกติมาตรฐาน เมื่อ Z เป็นตัวแปรสุ่มแบบปกติที่มีคาดคาดหวังเป็น 0 และมีความแปรปรวนเป็น 1 หรือ $Z \sim N(0,1)$

ถ้ากำหนดให้ $X \sim N(\mu, \sigma^2)$ และ $Z \sim N(0,1)$ แล้วจะได้ว่า $Z = \frac{X-\mu}{\sigma}$ และเราสามารถหาความน่าจะเป็นของ X ได้ด้วยความสัมพันธ์ $P(X \leq x) = P(Z \leq \frac{x-\mu}{\sigma})$

ถึงแม้ว่าฟังก์ชันการแจกแจงสะสมของการแจกแจงแบบปกติมาตรฐาน จะไม่สามารถหารูปปิดได้ แต่เราสามารถใช้การประมาณค่าในการคำนวณและเก็บค่าไว้ได้ ซึ่งจะช่วยให้การคำนวณหาความน่าจะเป็นทำได้สะดวกยิ่งขึ้น

การประมาณค่าพารามิเตอร์ของการแจกแจง μ และ σ^2 จากตัวอย่างข้อมูล โดยใช้ตัวประมาณค่า (estimator) $\hat{\mu}$ และ $\hat{\sigma}^2$ สามารถทำได้โดยการคำนวณ ดังนี้

ให้ตัวอย่าง $X_1, X_2, \ldots, X_n$ มาจากตัวแปรสุ่ม $X \sim N(\mu, \sigma^2)$ จะได้ว่า

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

จากสมการทั้งสอง จะสามารถประมาณค่าคาดหวังและความแปรปรวนของการแจกแจงแบบปกติได้

2.1.2.2. การแจกแจงความน่าจะเป็นของตัวแปรสุ่มแบบล็อกนอร์มัล

ในงานวิจัยนี้จะพิจารณาตัวแปรสุ่มแบบล็อกนอร์มัล เฉพาะในแง่ของการแจกแจงความน่าจะเป็น ซึ่งคล้ายคลึงกับตัวแปรสุ่มแบบปกติ ความหมายของตัวแปรสุ่มแบบล็อกนอร์มัล คือเมื่อพิจารณาการแจกแจงของค่าฟังก์ชันลอการิทึมของตัวแปรสุ่มนั้น ๆ แล้วจะพบว่ามีการแจกแจงแบบปกติ กล่าวคือสามารถนิยามตัวแปรสุ่ม X ว่ามีการแจกแจงแบบล็อกนอร์มัล ได้ เมื่อมีตัวแปรสุ่ม $Y = \log(X)$ และ $Y \sim N(\mu, \sigma^2)$

จากนิยามข้างต้น จะได้ว่าค่าความน่าจะเป็นในช่วงใด ๆ ของตัวแปรสุ่ม X สามารถหาได้จาก $P(X \leq x) = P(Y \leq \log(x))$ ซึ่งการหาค่าความน่าจะเป็น และการประมาณค่าพารามิเตอร์ของตัวแปรสุ่ม Y สามารถทำได้ดังที่กล่าวไว้ในหัวข้อการแจกแจงความน่าจะเป็นของตัวแปรสุ่มแบบปกติ

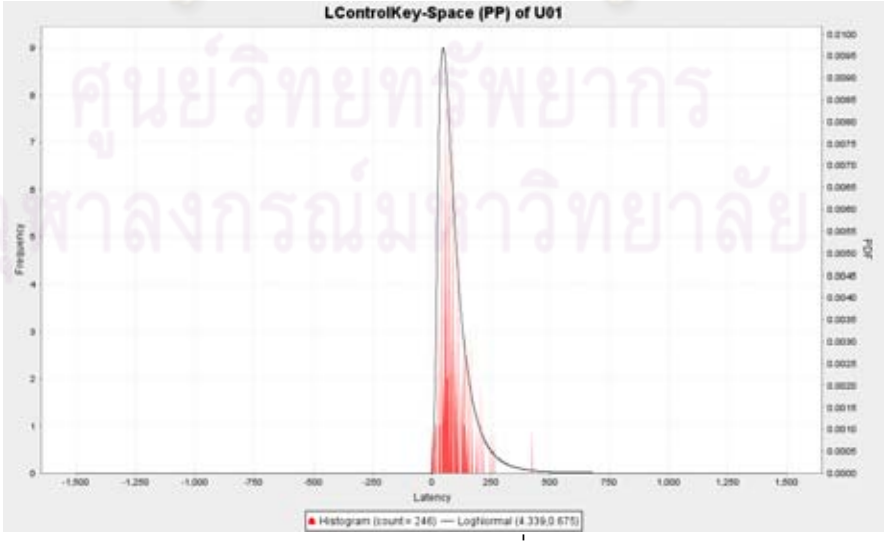
ตัวแปรสุ่มแบบล็อกนอร์มัล นั้นมีความสำคัญต่องานวิจัยนี้เนื่องจาก ในงานวิจัยของ Montalv, J. และ Freire, E. [2] ได้นำเสนอว่าการกระจายตัวของระยะเวลาระหว่างไดกราฟมีการกระจายตัวเป็นแบบล็อกนอร์มัล ซึ่งค่าความน่าจะเป็นจากการแจกแจงรูปแบบนี้น่าจะถูกนำมาใช้เป็นวิธีหนึ่งที่ช่วยในการจำแนกผู้ใช้ได้

2.1.2.3. การคำนวณความน่าจะเป็นจากการแจกแจงความถี่สัมพัทธ์ของฮิสโตแกรม

จากหัวข้อข้างต้น หากเราไม่มั่นใจว่าจะสามารถประมาณค่าพารามิเตอร์ของตัวแปรสุ่มแบบล็อกนอร์มัล ให้ตรงกับข้อมูลจริงทั้งหมดได้ ทางเลือกหนึ่งที่สามารถทำได้คือทำการสร้างฮิสโตแกรมเพื่อแจกแจงความถี่ของข้อมูลจากตัวข้อมูลทั้งหมดแทน จากนั้นจึงคำนวณความน่าจะเป็นโดยอาศัยการแจกแจงความถี่สัมพัทธ์ (Relative frequency)

การแจกแจงความถี่สัมพัทธ์ คือการสรุปข้อมูลทั้งหมดว่า ในช่วงต่าง ๆ มีความถี่ของข้อมูลที่เกิดขึ้นเป็นร้อยละเท่าใดของจำนวนข้อมูลทั้งหมด จะเห็นได้ว่าผลรวมของความถี่สัมพัทธ์ของทุก ๆ ช่วงนั้นมีค่าเท่ากับ 1 ทำให้สามารถมองการแจกแจงความถี่สัมพัทธ์เป็นการแจกแจงความน่าจะเป็นแบบไม่ต่อเนื่องรูปแบบหนึ่งได้ โดยค่าความน่าจะเป็นก็คือค่าความถี่สัมพัทธ์ในแต่ละช่วงนั่นเอง

เราจะใช้การแจกแจงความถี่สัมพัทธ์เป็นอีกหนึ่งรูปแบบของการแจกแจงความน่าจะเป็น เพื่อเปรียบเทียบกับการแจกแจงความน่าจะเป็นแบบล็อกนอร์มัล (ดังตัวอย่างในภาพที่ 2-3) ในการทดลองการจำแนกผู้ใช้ โดยกำหนดให้แต่ละช่วงมีขนาด 1 มิลลิวินาที และจุดกึ่งกลางช่วง คือจุดที่ค่าของมิลลิวินาทีเป็นจำนวนเต็มบวก



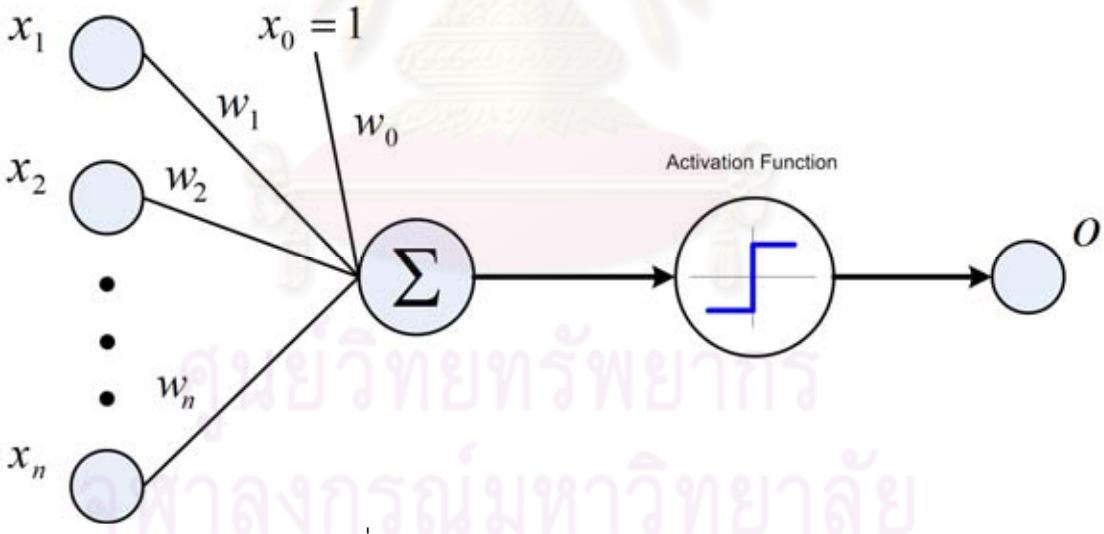
ภาพที่ 2-3 เปรียบเทียบรูปร่างของการแจกแจงความถี่ของฮิสโตแกรมและการประมาณค่าการแจกแจงความถี่แบบล็อกนอร์มัล ของระยะเวลาระหว่างไดกราฟ

2.1.3. นิวรอลเน็ตเวิร์ก

นิวรอลเน็ตเวิร์ก เป็นแบบจำลองทางคณิตศาสตร์ที่จำลองการทำงานบางส่วนมาจากสมองมนุษย์ อันประกอบไปด้วยเซลล์สมองจำนวนมากที่เชื่อมต่อกัน รับ-ส่งสัญญาณไฟฟ้าเคมีกันในเซลล์ประสาทใกล้เคียง โดยอาศัยหลักการที่เมื่อเซลล์ประสาทได้รับสัญญาณไฟฟ้าเคมีเข้ามาเกินค่าหนึ่ง เซลล์จะถูกกระตุ้นและส่งสัญญาณไปยังเซลล์อื่น ๆ ต่อไป หน่วยย่อยที่สุดที่เราจำลองขึ้น โดยมีเพียงหน่วยเดียวในการจำลองลักษณะของเซลล์ประสาท มีชื่อว่า เพอร์เซ็ปตรอน (Perceptron)

2.1.3.1. เพอร์เซ็ปตรอน

เพอร์เซ็ปตรอนนั้นเป็นแบบจำลองที่ง่ายที่สุด มีโครงสร้างดังภาพที่ 2-4 เพอร์เซ็ปตรอนรับเวกเตอร์ของค่าอินพุตที่เป็นจำนวนจริงแล้วคำนวณหาผลรวมเชิงเส้นแบบถ่วงน้ำหนักของค่าอินพุต มีการเพิ่มค่าขีดแบ่ง (θ) โดยเพิ่มข้อมูลอินพุตเทียม $x_0 = 1$ และปรับตัวถ่วงน้ำหนัก $w_0 = -\theta$ เมื่อหาผลรวมเชิงเส้นได้แล้วจะนำผลลัพธ์ไปผ่านฟังก์ชันกระตุ้น (activation function) ซึ่งจะใช้ฟังก์ชันสองขั้ว (bipolar function) ซึ่งจะให้เอาต์พุตเป็น 1 กับ -1 หรืออาจใช้ฟังก์ชันไบนารี (binary function) ที่ให้เอาต์พุตเป็น 1 กับ 0 แทนในบางงานก็ได้



ภาพที่ 2-4 โครงสร้างของเพอร์เซ็ปตรอน [3]

ในงานการจำแนกประเภท การแยกประเภทของอินพุตโดยเพอร์เซ็ปตรอนจะอยู่ในรูปเส้นตรง หรือเป็นระนาบตัดสินใจหลายมิติ (hyperplane decision surface) เมื่ออินพุตมีมากกว่าสองมิติ โดยความชันของเส้นตรงนั้นจะขึ้นอยู่กับเวกเตอร์ของน้ำหนัก ดังนั้น ปัญหาการเรียนรู้ของเพอร์เซ็ปตรอนก็คือการปรับค่าเวกเตอร์ของน้ำหนักให้เหมาะสม เพื่อให้แยกประเภทของข้อมูลสอนได้อย่างถูกต้อง กฎการเรียนรู้ของเพอร์เซ็ปตรอน (perceptron learning rule) เป็น

อัลกอริทึมที่ใช้สำหรับการสอนเพอร์เซ็ปตรอนโดยการพยายามปรับค่าเวกเตอร์น้ำหนักให้เหมาะสมกับการจำแนกได้

อัลกอริทึมเริ่มต้นจากการสุ่มค่าน้ำหนัก จากนั้นจะเปรียบเทียบตัวอย่างสอนทีละตัวว่าเพอเซ็ปตรอนสามารถจำแนกตัวอย่างสอนได้ถูกต้องหรือไม่ หากสามารถจำแนกได้ถูกต้องแสดงว่าระนาบนั้นดีแล้ว แต่หากไม่ถูกต้องก็ต้องปรับน้ำหนักดังสูตร $\Delta w_i = \alpha(t - o)x_i$ โดย t คือเอาต์พุตจริงของตัวอย่างฝึก o คือเอาต์พุตของเพอร์เซ็ปตรอน และ α คืออัตราการเรียนรู้ (learning rate) ซึ่งเป็นตัวเลขบวกจำนวนน้อย ๆ ซึ่งจะส่งผลต่อการลู่เข้าของเพอร์เซ็ปตรอน ถ้ามีค่ามากเพอร์เซ็ปตรอนจะเรียนรู้ได้เร็ว แต่อาจทำไม่สำเร็จเนื่องจากการปรับค่าหยาบเกินไป ถ้ามีค่าน้อยก็อาจเสียเวลาเรียนรู้นาน การเรียนรู้จะทำซ้ำเป็นรอบ ๆ จนกว่าเพอร์เซ็ปตรอนจะสามารถแยกตัวอย่างได้ถูกต้องทั้งหมด หรือถึงเงื่อนไขหยุดที่กำหนดไว้

เพอร์เซ็ปตรอนมีความสามารถในการเรียนรู้ที่จำกัด โดยสามารถเรียนรู้ได้ในระดับของฟังก์ชันแยกเชิงเส้นได้ (linearly separable function) เท่านั้น

การเรียนรู้โดยใช้กฎการเรียนรู้ของเพอร์เซ็ปตรอนนั้น จะลู่เข้าเมื่อเพอร์เซ็ปตรอนสามารถจำแนกตัวอย่างได้ถูกต้องทั้งหมด ซึ่งในข้อมูลจริงบางครั้งไม่สามารถทำเช่นนั้นได้ จึงมีกฎการเรียนรู้อีกแบบหนึ่งที่หาค่าของเวกเตอร์น้ำหนักที่ให้ค่าผิดพลาดน้อยที่สุดแทน นั่นคือ กฎเดลตา (delta rule) ซึ่งกฎนี้ใช้หลักการของการเคลื่อนลงตามความชัน (gradient descent) เพื่อหาคำตอบจากปริภูมิของเวกเตอร์น้ำหนักที่เป็นไปได้แทน

กฎเดลตาจะหาเวกเตอร์น้ำหนักที่ให้ค่าผิดพลาดน้อยสุดจากการหาอนุพันธ์ทางคณิตศาสตร์ ดังนั้นเราจำเป็นต้องใช้ฟังก์ชันกระตุ้นที่หาอนุพันธ์ได้ เราจึงเปลี่ยนไปใช้ฟังก์ชันกระตุ้นแบบฟังก์ชันเชิงเส้น (linear function) แทน นั่นก็คือค่าเอาต์พุตจะมีค่าเท่ากับผลรวมเชิงเส้นแบบถ่วงน้ำหนัก ซึ่งการจำแนกกลุ่มตัวอย่าง อาจใช้การจำแนกจากเครื่องหมายแทน

การหาเวกเตอร์น้ำหนักที่ให้ค่าผิดพลาดต่ำสุด เริ่มจากการนิยามฟังก์ชันค่าผิดพลาดจากการสอน (training error function) ที่ขึ้นกับเวกเตอร์น้ำหนัก ได้ดังสมการที่ 3

$$E(\vec{w}) = \frac{1}{2}\sum_{d \in D}(t_d - o_d)^2 \tag{3}$$

โดย D เป็นเซ็ตของตัวอย่างสอน t_d เป็นเอาต์พุตเป้าหมายของตัวอย่าง d และ o_d เป็นเอาต์พุตของเพอร์เซ็ปตรอนสำหรับตัวอย่าง d

ฟังก์ชันค่าผิดพลาดการสอนจะมีลักษณะเป็นฟังก์ชันพาราโบลาของ $\vec{w}$ ซึ่งจะมีจุดต่ำสุดเพียงจุดเดียว แนวคิดที่เราจะหาเวกเตอร์น้ำหนัก $\vec{w}$ ที่ให้ค่าผิดพลาดสูงสุดนั้น จะเริ่มจากการสุ่มค่าของ $\vec{w}$ ขึ้นมาก่อน จากนั้นหาความชันของเวกเตอร์สัมผัสค่าผิดพลาด ณ ตำแหน่ง

$\vec{w}$ นั้น ๆ เมื่อหาได้แล้วจึงค่อย ๆ ปรับค่า $\vec{w}$ ไปตามความชัน จนกระทั่งความชันของเวกเตอร์สัมผัสเป็น 0 นั่นคือถึงจุดต่ำสุดแล้ว ไม่จำเป็นต้องปรับค่า $\vec{w}$ อีกต่อไป

อย่างไรก็ดีถึงแม้ใช้กฎเดลตาในการเรียนรู้แล้ว เพอร์เซ็ปตรอนก็ยังไม่สามารถเรียนรู้ฟังก์ชันแยกเชิงเส้นไม่ได้ (linearly non-separable function) ได้อย่างถูกต้องสมบูรณ์ จึงมีแบบจำลองที่ซับซ้อนยิ่งขึ้นเพื่อสามารถแยกฟังก์ชันประเภทนี้ได้ นั่นคือ นิวรอลเน็ตเวิร์กแบบหลายชั้น (Multilayer Neural Network)

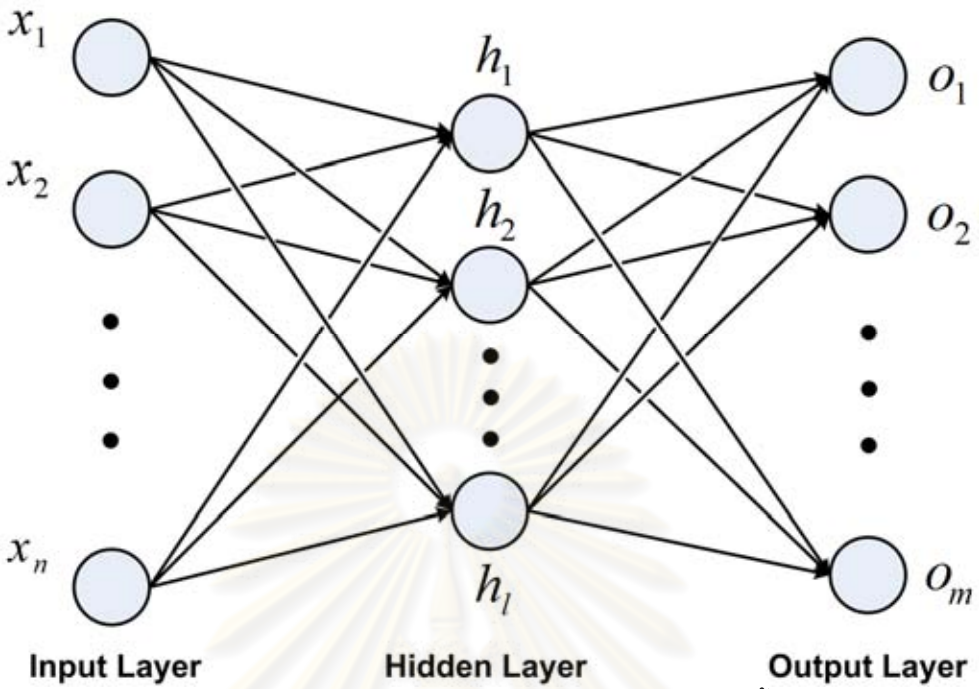
2.1.3.2. นิวรอลเน็ตเวิร์กแบบหลายชั้น

จากที่กล่าวมาข้างต้นนั้น เพอร์เซ็ปตรอนจะสามารถเรียนรู้ฟังก์ชันแยกได้เชิงเส้นเท่านั้นถึงจะมีประสิทธิภาพ เราจึงนำเพอร์เซ็ปตรอนหลาย ๆ ตัวมาต่อเชื่อมกัน เพื่อสร้างเป็นนิวรอลเน็ตเวิร์กแบบหลายชั้นที่สามารถแสดงผิวตัดสินใจไม่เชิงเส้น (non-linear decision surface) เพื่อให้สามารถเรียนรู้ฟังก์ชันที่ซับซ้อนมากขึ้นได้

นิวรอลเน็ตเวิร์กแบบหลายชั้นที่นิยมใช้คือ นิวรอลเน็ตเวิร์กป้อนไปข้างหน้าแบบหลายชั้น (multilayer feedforward network) มีโครงสร้างหลายชั้นประกอบด้วยชั้นอินพุต ชั้นแฝง และชั้นเอาต์พุต ดังภาพที่ 2-5 การเชื่อมต่อจะมีไปเป็นชั้น ๆ โดยไม่ข้ามชั้นกัน จากชั้นอินพุตไปยังชั้นแฝง ถ้ามีชั้นแฝงหลายชั้นก็จะเชื่อมต่อกันไปเรื่อย ๆ และสุดท้ายจากชั้นแฝงไปยังชั้นเอาต์พุต แต่ละเส้นเชื่อมจะมีค่าถ่วงน้ำหนักกำกับอยู่ เมื่อมีอินพุตเข้ามา ชั้นอินพุตก็จะส่งผลไปยังชั้นแฝง ชั้นแฝงจะพยายามแปลงข้อมูลที่ได้รับมาให้สามารถแยกได้ด้วยเส้นตรงเส้นเดียว (linearly separable) ก่อนที่จะส่งต่อไปยังชั้นเอาต์พุต

การปรับค่าเวกเตอร์น้ำหนักนั้น จะใช้อัลกอริทึมการแพร่กระจายย้อนกลับ (backpropagation algorithm) ซึ่งอาศัยการเคลื่อนลงตามความชันเช่นกัน ดังนั้นฟังก์ชันกระตุ้นที่จะใช้จึงต้องสามารถหาอนุพันธ์ได้ตลอด จึงนิยมใช้ฟังก์ชันเชิงเส้น หรือฟังก์ชันซิกมอยด์ (sigmoid function) ซึ่งมีลักษณะคล้ายฟังก์ชันไบนารี แต่เป็นฟังก์ชันต่อเนื่องโดยตลอด ฟังก์ชันซิกมอยด์นิยามได้ โดยสมการที่ 4

$$\sigma(y) = \frac{1}{1 + e^{-y}} \tag{4}$$



ภาพที่ 2-5 โครงสร้างของนิวรอลเน็ตเวิร์กแบบหลายชั้น [3]

คุณสมบัติที่ดีของฟังก์ชันซิกมอยด์นั้นคือ สามารถหาอนุพันธ์ได้ในรูปปิดที่คำนวณง่าย ดังสมการที่ 5

$$\frac{d\sigma(y)}{dy} = \sigma(y)(1-\sigma(y)) \tag{5}$$

อัลกอริธึมการแพร่กระจายย้อนกลับนั้นมีการทำงานคล้ายกฎเดลตา โดยเริ่มต้นจากการนิยามฟังก์ชันค่าผิดพลาดจากการสอนสำหรับเน็ตเวิร์ก $E(\vec{w})$ ดังสมการที่ 6

$$E(\vec{w}) = \frac{1}{2}\sum_{d \in D}\sum_{k \in outputs}(t_{kd} - o_{kd})^2 \tag{6}$$

โดย $outputs$ คือเซตของโหนดในชั้นเอาต์พุตในเน็ตเวิร์ก t_{kd} และ o_{kd} เป็นค่าเอาต์พุตเป้าหมายและค่าเอาต์พุตที่ได้จากเน็ตเวิร์กตามลำดับ ของโหนดเอาต์พุตที่ k ของตัวอย่างที่ d

หลังจากนิยามฟังก์ชันค่าผิดพลาดแล้วจะหาค่าเวกเตอร์น้ำหนักที่ให้ค่าผิดพลาดต่ำสุด แต่ในกรณีของนิวรอลเน็ตเวิร์กป้อนไปข้างหน้าแบบหลายชั้นนี้จะมีจุดที่ให้ค่าผิดพลาดต่ำสุดหลายที่ ดังนั้นคำตอบที่ได้จึงมักเป็นค่าต่ำสุดเฉพาะที่ (local minima)

ในงานวิจัยนี้ เราเลือกใช้นิวรอลเน็ตเวิร์กป้อนไปข้างหน้าแบบหลายชั้น มาช่วยในการจำแนกผู้ใช้จากคุณลักษณะที่ได้จากการแจกแจงความน่าจะเป็นของไดกราฟต่าง ๆ ของผู้ใช้แต่ละคน

2.2. งานวิจัยที่เกี่ยวข้อง

งานวิจัยในด้านการใช้ระยะเวลาในการพิมพ์ เพื่อการจำแนกหรือการยืนยันตัวบุคคลนั้น เริ่มมีมาตั้งแต่ปี ค.ศ. 1975 แต่เพิ่งมีจำนวนมากขึ้นและได้ผลการทดลองที่ดีขึ้นในช่วงปลายของทศวรรษ 1990 จนถึงปัจจุบัน เกณฑ์หนึ่งที่สามารถใช้แบ่งประเภทของงานได้ คือ ประเภทของข้อความที่ใช้ นั่นคือข้อความที่ถูกกำหนดไว้และข้อความอิสระ งานวิจัยส่วนมากจะศึกษาเฉพาะกับการใช้ข้อความที่ถูกกำหนดไว้แล้ว ส่วนงานที่ศึกษาข้อความอิสระนั้นยังมีจำนวนน้อยอยู่

งานวิจัยส่วนมากพิจารณาในเรื่องการยืนยันตัวตนเป็นหลัก (Authentication) กล่าวคือ ปัญหาของงานคือการรับตัวอย่างการพิมพ์ตัวอย่างหนึ่งพร้อมระบุว่าเป็นของใคร ระบบต้องตัดสินใจว่าจะยอมรับหรือปฏิเสธว่าเจ้าของตัวอย่างการพิมพ์นั้นเป็นของบุคคลที่กล่าวอ้างจริง ค่าที่จะใช้วัดประสิทธิภาพของระบบนั้นจะพิจารณาจาก False Alarm Rate (FAR) ซึ่งเป็นอัตราการปฏิเสธผู้ใช้ที่เป็นเจ้าของตัวอย่างการพิมพ์นั้นจริง ๆ โดยจะส่งผลให้ผู้ใช้เกิดความรำคาญ และ Imposter Pass Rate (IPR) ซึ่งเป็นอัตราการยอมรับผู้ใช้ที่ไม่ได้เป็นเจ้าของตัวอย่างการพิมพ์นั้นจริง โดยจะส่งผลให้ระบบไม่ปลอดภัยเพราะถูกโจมตีสำเร็จ ระบบที่ดีควรมีค่าทั้งสองนี้น้อย ๆ อย่างไรก็ดียังมีอีกค่าหนึ่งที่ใช้พิจารณาคือ Equal Error Rate (EER) ซึ่งคือค่าของ FAR และ IPR ณ ตำแหน่งที่ตั้งค่าระบบแล้วได้ค่าทั้งสองออกมาเท่ากัน

มีเพียงบางงานวิจัยที่จะพูดถึงการจำแนกตัวบุคคล (Classification) ด้วย เพราะในบางครั้งพื้นฐานของการตัดสินใจในด้านการยืนยันตัวตนนั้นก็มาจากการจำแนกตัวบุคคล โดยค่าที่ใช้วัดประสิทธิภาพของระบบนั้นพิจารณาจากความแม่น (Accuracy) ในการจำแนกตัวบุคคล

ในแง่ของคุณลักษณะที่เลือกใช้ที่นิยมมากที่สุดคือคุณลักษณะของเวลา มีบางงานวิจัยที่ใช้คุณลักษณะอื่น ๆ เช่น ระดับแรงกด (key pressure) หรือการให้ความสำคัญกับความยากง่ายของข้อความที่พิมพ์ แต่ก็ยังไม่เป็นที่นิยมเนื่องจากต้องใช้อุปกรณ์เสริมพิเศษ หรือต้องเพิ่มความยุ่งยากโดยไม่ได้ทำให้ผลความแม่นดีขึ้นอย่างมีนัยสำคัญ

ในด้านเทคนิกที่ใช้นั้นมีอยู่หลากหลายมาก เช่น การวัดความแตกต่างด้วยคุณลักษณะทางสถิติแบบค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน การนิยามการวัดระยะห่างระหว่าง

ตัวอย่างสองตัว การใช้เทคนิกการเรียนรู้ของเครื่องต่าง ๆ อาทิ นิวรอลเน็ตเวิร์ก (Neural Network) หรือ ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm) เป็นต้น

ต่อไปนี้จะยกตัวอย่างงานวิจัยโดยแบ่งออกเป็นประเภทต่าง ๆ ดังนี้

2.2.1. **งานวิจัยที่ใช้ข้อความที่ถูกกำหนดไว้** (Fixed Text)

งานวิจัยประเภทนี้จะศึกษาเฉพาะกับตัวอย่างการพิมพ์ที่เป็นข้อความที่ถูกกำหนดไว้ เช่น ชื่อผู้ใช้ รหัสผ่าน ชื่อ นามสกุลของผู้ใช้ หรือข้อความที่ถูกกำหนดไว้อื่น ๆ งานวิจัยที่มีอยู่ส่วนมากจะอยู่ในประเภทนี้ ซึ่งแต่ละงานก็จะมีวิธีที่หลากหลายแตกต่างกันไป

เราอาจแบ่งย่อยข้อความที่ถูกกำหนดไว้ได้อีก โดยแบ่งเป็นข้อความที่มีขนาดสั้นมาก เช่น รหัสผ่าน ในการใช้ข้อความประเภทนี้เราสามารถนำค่าระยะเวลามาใช้เป็นเวกเตอร์อินพุตได้ทันที ตัวอย่างงานวิจัยลักษณะนี้ คือ

Joyce, R. และ Gupta, G. [4] ได้เสนอผลงานไว้ในปี 1990 ซึ่งน่าจะถือเป็นผลงานหนึ่งที่เป็นพื้นฐานของงานวิจัยต่อมาได้ โดยผู้เขียนได้กล่าวไว้ว่าลักษณะการพิมพ์นั้นมีความคลายคลึงกับลายมือหรือลายมือชื่อ จึงน่าจะสามารถนำมาใช้ในการระบุตัวบุคคลได้ โดยรูปแบบนั้นจะยิ่งชัดเจนถ้าเป็นข้อความที่พิมพ์บ่อย ๆ ผู้เขียนทดลองในแง่มุมของการยืนยันตัวตนโดยเลือกใช้ชื่อผู้ใช้ รหัสผ่าน ชื่อและนามสกุล ของแต่ละคนเป็นข้อความที่จะนำมาทดสอบ โดยนำเวลาที่ผ่านไปของการพิมพ์แต่ละตัวอักษรในข้อความเหล่านั้นมาเรียงเป็นเวกเตอร์ การเปรียบเทียบทำโดยใช้ขนาดของเวกเตอร์ โดยกำหนดค่าขีดแบ่งว่าจะยอมรับหรือไม่ยอมรับผู้ใช้โดยดูจากส่วนเบี่ยงเบนมาตรฐาน (standard deviation) ซึ่งถือได้ว่าเป็นการพิจารณาระยะเวลาในการพิมพ์โดยภาพรวม ที่มีข้อจำกัดคือเป็นการพิมพ์ข้อความที่เหมือนกัน นอกจากนี้ยังเสนอวิธีการที่พิจารณารูปร่างของเวกเตอร์ โดยการพิจารณาความชัน (slope) ระหว่างเวลาแต่ละจุดในเวกเตอร์อีกด้วย

Lammers, A. และ Langenfeld, S. [5] ได้เสนอผลงานไว้ในปี 1991 โดยได้ศึกษาการนำนิวรอลเน็ตเวิร์กมาใช้ในการยืนยันตัวตนผู้ใช้ โดยใช้เฉพาะรหัสผ่านเท่านั้น ผลการทดลองพบว่าการใช้รหัสผ่านที่มีความยาวขนาดกลาง (ประมาณ 7 ตัวอักษร) และระยะการยอมรับที่น้อย จะให้ผลออกมาดีที่สุด

Chen, L., Weng, L. และ Chee, L. [6] ได้เสนอผลงานไว้ในปี 2007 โดยเสนอวิธีการยืนยันตัวตนโดยอาศัย ARTMAP-FD ซึ่งถือเป็นนิวรอลเน็ตเวิร์กประเภทหนึ่ง โดยมีจุดเด่นอยู่ที่วิธีการนี้ ข้อมูลฝึกไม่จำเป็นต้องมีข้อมูลเกี่ยวกับผู้โจมตี ซึ่งในการใช้งานจริงนั้นจะไม่รู้ข้อมูลเหล่านี้ล่วงหน้า การฝึก ARTMAP-FD นั้นจะใช้เพียงข้อมูลของผู้ใช้ที่เป็นเจ้าของจริง ๆ เท่านั้น โดยจะมองข้อมูลของผู้ใช้เป็นข้อมูลที่ปกติ และมองข้อมูลของผู้โจมตีว่าเป็นความผิดปกติ ดังนั้นสิ่งที่

ทำก็คือการตรวจสอบความผิดปกตินั่นเอง ข้อมูลที่ใช้ในการทดลองจะเป็นข้อมูลเวลาของการพิมพ์รหัสผ่านเดียวกัน โดยพิมพ์คนละ 10 ครั้ง โดยให้ผู้ใช้ได้ทดลองพิมพ์จนคุ้นเคยก่อนถึงจะเก็บข้อมูล นอกจากนี้ผู้เขียนยังได้เสนอการใช้ข้อมูลระดับแรงกด จากคีย์บอร์ดที่รับแรงกดได้มาร่วมในการยืนยันตัวตนด้วย โดยผลการทดลองพบว่าการใช้ข้อมูลแรงกดร่วมด้วยนั้นได้ผลที่ดีกว่าการใช้ข้อมูลเวลาเพียงอย่างเดียวอยู่เล็กน้อย

Lee, J., Choi, S และ Byung-Ro, M. [7] ได้เสนอผลงานไว้ในปี 2007 โดยศึกษาการยืนยันตัวตนด้วยระยะเวลาในการการพิมพ์รหัสผ่าน โดยใช้ระยะเวลาในการพิมพ์ตัวอักษรและระยะเวลาระหว่างไดกราฟเป็นคุณลักษณะ ส่วนวิธีการตัดสินใจนั้นใช้ปริภูมิสมมติฐานที่มีลักษณะเป็นรูปกลมรี (ellipsoidal hypothesis space) ซึ่งสร้างด้วยขั้นตอนวิธีเชิงพันธุกรรม นอกจากนั้นยังเสนอวิธีการปรับตัวเพื่อให้รองรับกับการเปลี่ยนแปลงของระยะเวลาในการพิมพ์อีกด้วย

งานวิจัยอีกประเภทหนึ่งจะพิจารณาข้อความที่ถูกกำหนดไว้ ที่เป็นข้อความที่แตกต่างกัน หรือเป็นข้อความที่มีขนาดยาวขึ้น ซึ่งการใช้ค่าระยะเวลามาเป็นอินพุตตรง ๆ นั้นอาจไม่เหมาะสม จึงเริ่มใช้คุณลักษณะทางสถิติของระยะเวลามาเป็นอินพุตแทน คุณลักษณะทางสถิติที่เป็นที่นิยมมากที่สุดในงานวิจัยต่าง ๆ คือ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน ยกตัวอย่างงานวิจัยประเภทนี้ได้ดังต่อไปนี้

Bergadano, F., Gunetti, D. และ Picardi, C. [8] ได้เสนอผลงานไว้ในปี 2002 โดยเน้นว่าข้อมูลระยะเวลาในการพิมพ์นั้นจะมีการเปลี่ยนแปลงไปตามเวลาได้มากกว่าข้อมูลชีวมาตรประเภทอื่น ๆ ดังนั้นจึงเสนอวิธีการเปรียบเทียบโดยพิจารณาการเปลี่ยนแปลงเป็นสำคัญ โดยใช้ระดับความไร้ระเบียบ (Degree of Disorder) แทนระยะห่างของตัวอย่างแต่ละตัว

ผู้เขียนจะใช้ข้อความยาว 683 ตัวอักษรเป็นตัวเพื่อให้ผู้ใช้พิมพ์ โดยจะใช้ระยะเวลาระหว่างไตรกราฟ (พิจารณาทีละ 3 ตัวอักษร โดยใช้ระยะเวลาระหว่างการกดตัวแรกถึงการกดตัวสุดท้าย) เป็นข้อมูลที่ใช้เปรียบเทียบ ในการทดลองนี้มีผู้ใช้ที่ร่วมในการทดลอง 44 คน โดยแต่ละคนพิมพ์ข้อความที่กำหนด 5 ครั้ง ผู้เขียนทำการทดลองทั้งการจำแนกตัวบุคคล และการยืนยันตัวบุคคล โดยการจำแนกตัวบุคคลนั้นจะจำแนกตามผู้ใช้ที่มีระดับความไร้ระเบียบของตัวอย่างโดยเฉลี่ยเมื่อเทียบกับตัวอย่างทดสอบแล้วมีค่าน้อยที่สุด ซึ่งได้ผลการจำแนกที่ดีมาก นั่นคือทำการจำแนกได้ถูกต้องทั้งหมดเมื่อใช้ 4 ใน 5 ตัวอย่างของแต่ละผู้ใช้เป็นข้อมูลฝึก และในส่วนของการยืนยันตัวบุคคลนั้นจะเพิ่มผู้ทดลองที่เป็นผู้โจมตีอีก 110 คน โดยการจะยอมรับหรือไม่ยอมรับผู้ใช้นั้นจะนอกจากจะอิงกับการจำแนกตัวบุคคลได้แล้ว ตัวอย่างนั้นต้องมีค่าระยะห่างไม่

มากเกินที่กำหนดไว้อีกด้วย นอกจากนั้น แนวคิดการเปรียบเทียบเฉพาะส่วนที่เหมือนกันของผู้เขียนได้เป็นพื้นฐานสำคัญของงานวิจัยที่ใช้ข้อความอิสระที่เกิดขึ้นภายหลังโดยผู้เขียนอีกด้วย

Curtin, M. et al. [9] ได้เสนอผลงานไว้ในปี 2006 โดยศึกษาระยะเวลาในการพิมพ์กับข้อความที่มีขนาดยาว โดยพิจารณาในแง่มุมของการจำแนกตัวบุคคล ผู้เขียนได้ใช้คุณลักษณะที่หลากหลาย ซึ่งโดยหลักจะเป็นค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของระยะเวลาในการพิมพ์ตัวอักษร (Keystroke Duration) และระยะเวลาระหว่างไดกราฟ เฉพาะส่วนที่พบบ่อย ๆ ในข้อความ และยังมีลักษณะอื่น ๆ เช่น จำนวนการกดปุ่มควบคุม (Control Key) หรือปุ่มลูกศร และจำนวนครั้งในการคลิกเมาส์แต่ละปุ่มอีกด้วย โดยผู้เขียนมีการคัดกรองเอาค่าที่มีความผิดพลาดสูง (Outlier) ออกไป และทำการ standardize เพื่อให้แต่ละคุณลักษณะมีความสำคัญเท่ากัน การจำแนกจะพิจารณาจากระยะทางยูคลิด (Euclidean distance) และเลือกตัวที่เป็นเพื่อนบ้านใกล้สุด (nearest neighbor) เพื่อจำแนก ผลการทดลองพบว่าจะให้ผลที่ดีที่สุดเมื่อใช้ข้อความเดียวกัน ผลจะน้อยลงเมื่อพิจารณาข้อความที่ต่างกัน และจะยิ่งน้อยลงเมื่อขนาดความยาวของข้อความนั้นน้อยลงด้วย

Hocquet, S., Ramel, J. และ Cardot, H. [10] ได้เสนอผลงานไว้ในปี 2007 โดยผู้เขียนเสนอว่าการตั้งค่าขีดแบ่งและค่าพารามิเตอร์ของระบบสำหรับการยืนยันตัวตนของผู้ใช้ทุก ๆ คนไว้เท่ากันนั้นอาจทำให้ได้ผลที่ไม่ดีเท่าที่ควร ผู้เขียนจึงเสนอวิธีโดยให้จัดกลุ่ม (cluster) ของผู้ใช้ก่อนแล้วจึงกำหนดค่าขีดแบ่งและค่าพารามิเตอร์แยกกันสำหรับแต่ละกลุ่ม ผู้เขียนได้ใช้คุณลักษณะที่แตกต่างกันทั้งหมด 31 แบบในการอธิบายตัวอย่างแต่ละตัว ซึ่งจะทำให้การจัดกลุ่มเป็นไปอย่างยากลำบาก จึงทำการลดมิติด้วยวิธีการวิเคราะห์องค์ประกอบหลัก (principal component analysis : PCA) แล้วจึงนำไปจัดกลุ่มด้วยวิธี K-mean ผลการทดลองพบว่าการตั้งต่าขีดแบ่งและค่าพารามิเตอร์แยกกันในแต่ละกลุ่มนั้นให้ค่าผิดพลาดที่ต่ำกว่าการกำหนดเป็นค่าค่าเดียวสำหรับทุก ๆ คนจริง

2.2.2. **งานวิจัยที่ใช้ข้อความอิสระ** (Free Text)

งานวิจัยประเภทนี้จะศึกษากับตัวอย่างข้อมูลการพิมพ์ที่ไม่จำกัดรูปแบบ โดยจะต้องสามารถเปรียบเทียบข้อมูลการพิมพ์ของข้อความที่แตกต่างกันได้ งานวิจัยประเภทนี้ยังมีอยู่ค่อนข้างน้อย และยังให้ผลการทดลองไม่ดีเทียบเท่ากับงานวิจัยที่ใช้ข้อความที่ถูกกำหนดไว้ ซึ่งได้ยกตัวอย่างมาดังต่อไปนี้

Monrose, F. และ Rubin, A. [11] ได้เสนอผลงานไว้ในปี 1997 โดยเป็นผลงานชิ้นแรก ๆ ที่ใช้ข้อความอิสระในการทดลอง ผู้เขียนใช้คุณลักษณะเป็นระยะเวลาในการพิมพ์

ตัวอักษรและระยะเวลาระหว่างไดกราฟเช่นกัน โดยจะใช้ค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของระยะเวลาเหล่านั้น โดยผู้เขียนเสนอวิธีการในการเปรียบเทียบมา 3 วิธี คือ

1) การแทนแต่ละข้อความด้วยเวกเตอร์ของค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานดังกล่าว จากนั้นจะใช้ระยะทางยูคลิด (Euclidean Distance) เป็นวิธีการเปรียบเทียบ โดยหากในข้อความที่สนใจมีจำนวนตัวอักษรหรือไดกราฟบางประเภทปรากฏอยู่น้อยเกินไป จะถูกปรับค่าเฉลี่ยของตัวอักษรหรือไดกราฟนั้นให้เป็น 0 ซึ่งจะทำให้เกิดความแตกต่างกับข้อความอื่นค่อนข้างมาก
2) ใช้การเปรียบเทียบจากคะแนนที่ได้จากการแจกแจงความน่าจะเป็นโดยเฉลี่ยของแต่ละตัวอักษรหรือไดกราฟ นำมารวมเข้าด้วยกันสำหรับทุก ๆ ตัวอักษรหรือไดกราฟ โดยจะใช้การแจกแจงความน่าจะเป็นแบบปกติ ที่มีพารามิเตอร์เป็นค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานที่ได้จากข้อมูลฝึกนั่นเอง
3) ใช้การเปรียบเทียบคะแนนที่ได้จากความน่าจะเป็นเช่นกัน แต่จะมีการถ่วงน้ำหนักตามความบ่อยครั้งที่แต่ละตัวปรากฏในข้อมูลฝึก

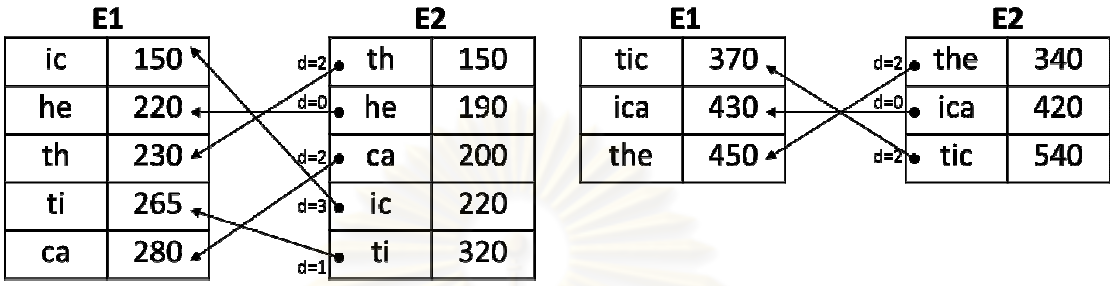
ผลการทดลองปรากฏว่าวิธีที่ใช้คะแนนความน่าจะเป็นแบบถ่วงน้ำหนักให้ผลออกมาดีที่สุด แต่ก็ได้ผลที่ต่ำมาก อย่างไรก็ดีงานวิจัยนี้ก็เป็นพื้นฐานให้กับงานวิจัยที่ใช้ข้อความอิสระในเวลาต่อมา

Gunetti, D. และ Picardi, C. [12] ได้เสนอผลงานไว้ในปี 2005 ซึ่งถือว่าเป็นงานวิจัยที่ศึกษาบนข้อความอิสระที่รายงานผลการทดลองออกมาดีที่สุดงานหนึ่ง

งานวิจัยของ D. Gunetti และ C. Picardi ใช้วิธีการเปรียบเทียบระยะห่าง (distance) ระหว่างตัวอย่างสองตัว ในการจำแนกตัวบุคคลจะจำแนกโดยเปรียบเทียบระยะห่างเฉลี่ยระหว่างตัวอย่างทดสอบกับตัวอย่างทั้งหมดของผู้ใช้แต่ละคน แล้วตัดสินใจจำแนกตัวอย่างทดสอบให้เป็นของผู้ใช้ที่มีระยะห่างเฉลี่ยน้อยที่สุด โดยผู้เขียนได้กำหนดวิธีวัดระยะหว่างไว้สองแบบ คือ ระยะห่างสัมพัทธ์ (R Distance) และระยะห่างสัมบูรณ์ (A Distance) ซึ่งในการเปรียบเทียบระยะห่างทั้งสองแบบนั้น จะพิจารณาเฉพาะไดกราฟที่มีอยู่ซ้ำกันในตัวอย่างทั้งสองเท่านั้น

ระยะห่างสัมพัทธ์ เกิดขึ้นจากแนวคิดที่ว่าระยะเวลาในการพิมพ์สามารถเปลี่ยนแปลงได้จากปัจจัยต่าง ๆ แต่ก็ควรจะเปลี่ยนแปลงไปในทิศทางเดียวกัน ดังนั้นระยะห่างสัมพัทธ์จะไม่ได้ใช้ค่าของระยะเวลาระหว่างไดกราฟตรง ๆ แต่จะใช้ระดับความไร้ระเบียบ (Degree of Disorder) แทน ซึ่งระดับความไร้ระเบียบ คือการเปรียบเทียบตำแหน่งการเรียงลำดับ

ของตัวอย่างทั้งสอง ว่ามีการเรียงลำดับระยะเวลาระหว่างไดกราฟที่ต่างกันไปมากเท่าไร ค่านี้จะถูกหารด้วยขนาดของลำดับที่เปรียบเทียบยกกำลังสองแล้วหารด้วยสอง ซึ่งเป็นระดับความไร้ระเบียบสูงสุด เพื่อเป็นการปรับให้ระยะห่างสัมพัทธ์มีค่าไม่เกิน 1 (ตัวอย่างการคำนวณระดับความไร้ระเบียบแสดงอยู่ในภาพที่ 2-6)



ภาพที่ 2-6 ภาพแสดงการคำนวณระดับความไร้ระเบียบของตัวอย่าง E1 และ E2 เปรียบเทียบในส่วนที่เป็นไดกราฟและไตรกราฟ [12]

ระยะห่างสัมบูรณ์ เกิดขึ้นเพื่อลบช่องโหว่ของการใช้ระยะห่างสัมพัทธ์แต่เพียงอย่างเดียว โดยมีแนวคิดว่าตัวอย่างทั้งสองถ้ามาจากผู้ใช้คนเดียวกัน ระยะเวลาระหว่างไดกราฟก็ไม่ควรห่างกันมากเกินไป ระยะห่างสัมบูรณ์เกิดจากการนับว่ามีไดกราฟที่มีระยะเวลาห่างกันไม่เกินค่าขีดแบ่งที่กำหนดไว้ (t) เป็นจำนวนเท่าไรเมื่อเทียบกับจำนวนไดกราฟทั้งหมด โดยการเปรียบเทียบแต่ละไดกราฟนั้นใช้การนำระยะเวลาของทั้งคู่มาหารกันโดยให้ตัวเศษเป็นตัวที่มีค่ามากกว่าเสมอ ระยะห่างสัมบูรณ์คิดจากนำ 1 ลบออกด้วยอัตราส่วนของไดกราฟที่มีระยะห่างไม่เกินค่าขีดแบ่งกับจำนวนไดกราฟทั้งหมด ซึ่งระยะห่างสัมบูรณ์ก็จะมีค่าไม่เกิน 1 เช่นกัน (ตัวอย่างการนับจำนวนไดกราฟที่มีความคล้ายคลึงกันแสดงอยู่ในภาพที่ 2-7)

E1		E2		
280	ca	220	(280/200=1.400)	
220	he	190	(220/190=1.157)	(similar pair)
150	ic	220	(220/150=1.466)	
230	th	150	(230/150=1.533)	
265	ti	320	(320/265=1.207)	(similar pair)

ภาพที่ 2-7 ภาพแสดงการนับจำนวนไดกราฟที่มีระยะเวลาต่างกันไม่เกินค่าขีดแบ่งเมื่อกำหนดค่าขีดแบ่ง t = 1.25 [12]

ระยะห่างที่ใช้พิจารณาในการจำแนกตัวบุคคลนั้น เกิดจากการนำค่าระยะห่างสัมพัทธ์และระยะห่างสัมบูรณ์มารวมกัน อย่างไรก็ดีการคำนวณระยะห่างทั้งแบบสัมพัทธ์และสัมบูรณ์นั้น อาจนิยามเพิ่มเติมให้นอกจากจะพิจารณาไดกราฟแล้ว ยังสามารถพิจารณาระยะเวลาระหว่างไตรกราฟ (trigraph) โฟร์กราฟ (4-graph) หรือมากกว่านั้นได้ โดยคำนวณระยะห่างเหมือนกับกรณีที่พิจารณาไดกราฟ เมื่อได้ระยะห่างมาแล้วจะต้องถ่วงน้ำหนักให้มีค่าลดลงตามอัตราส่วนของจำนวนไตรกราฟกับจำนวนไดกราฟ เนื่องจากโดยปกติเมื่อพิจารณาข้อความเดียวกันจะพบว่าจำนวนของไตรกราฟ จะมีน้อยกว่าจำนวนไดกราฟอยู่แล้ว

จากการทดลองการจำแนกตัวบุคคลกับผู้ใช้ 40 คน แต่ละคนพิมพ์ 15 ข้อความที่ยาวประมาณ 750-900 ตัวอักษร ด้วยวิธีการทดสอบแบบ leave-one-out พบว่าผลการจำแนกตัวบุคคลนั้นได้ความแม่นสูงสุดโดยมีข้อผิดพลาดเพียง 0.16% เมื่อใช้ระยะห่างสัมพัทธ์ของ 2 3 และ 4 ตัวอักษร รวมกับระยะห่างสัมบูรณ์ของ 2 และ 3 ตัวอักษร

ถึงแม้วิธีของ D. Gunetti และ C. Picardi จะให้ผลการทดลองที่ดีมาก แต่วิธีนี้ยังมีจุดอ่อนในกรณีที่ตัวอย่างทดสอบมีความยาวน้อย ๆ ซึ่งจะให้ผลความแม่นที่ไม่ค่อยดีนัก

Sim, T. และ Janakiraman, R. [13] ได้กล่าวไว้ว่าการพิจารณาแต่เพียงระยะเวลาระหว่างไดกราฟโดยไม่สนใจประเภทของไดกราฟ (generic digraph) นั้นอาจไม่ใช่คุณลักษณะที่เหมาะสมกับงานวิจัยที่ใช้ข้อความอิสระในการจำแนกผู้ใช้หรือการยืนยันตัวผู้ใช้ ผู้เขียนทำการทดลองโดยสร้างฮิสโตแกรมจากระยะเวลาการพิมพ์ (ซึ่งสามารถมองเป็นการแจกแจงความน่าจะเป็นได้) และใช้ Bhattacharyya distance [14] เพื่อพิจารณาความคล้ายคลึงของฮิสโตแกรมแต่ละอัน ถ้าฮิสโตแกรมซ้อนทับกันน้อยแสดงว่าสามารถใช้แบ่งแยกได้ดี ผลการทดลองพบว่าการพิจารณาแต่เพียงระยะเวลาระหว่างไดกราฟนั้นไม่สามารถใช้จำแนกผู้ใช้ได้ดี ผู้เขียนได้เสนอการใช้ระยะเวลาระหว่างไดกราฟโดยพิจารณาถึงข้อความที่เป็นบริบท นอกจากนั้นผู้เขียนได้เสนอวิธีการจำแนกผู้ใช้โดยอาศัยการพิจารณาระยะเวลาระหว่างไดกราฟที่ปรากฏอยู่ในคำที่พบบ่อยในภาษาอังกฤษ 10 อันดับแรกอีกด้วย

Tappert, C., Villani, M. และ Cha, S. [15] ได้เสนอผลงานที่ศึกษาต่อมาจากงานเดิมของ Curtin, M. et al. [9] โดยขยายมาทำการศึกษาระหว่างข้อความที่กำหนดไว้กับข้อความอิสระ รวมทั้งการพิมพ์บนคีย์บอร์ดของเครื่องคอมพิวเตอร์ตั้งโต๊ะและเครื่องคอมพิวเตอร์วางตักด้วย

ในด้านคุณลักษณะที่ใช้นั้น มีความหลากหลายและมีจำนวนมากกว่าเดิม โดยขยายเป็นค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาในการพิมพ์ตัวอักษร ระยะเวลาระหว่างไดกราฟแบบปล่อยถึงกด ระยะเวลาระหว่างไดกราฟแบบกดถึงกด นอกจากนั้นยังคำนึงถึงร้อยละการกดปุ่มที่ไม่ใช่ตัวอักษรและปุ่มเมาส์ ระยเวลาที่ใช้ทั้งหมด และจำนวนการกดปุ่มทั้งหมดอีกด้วย

นอกจากนี้ ยังมีการศึกษาลักษณะความสัมพันธ์ในแง่มุมต่าง ๆ ของการพิมพ์ตัวอักษรและไดกราฟ และสร้างแผนภูมิต้นไม้เพื่อแสดงความสัมพันธ์ โดยแผนภูมิต้นไม้นี้จะได้ใช้งานเมื่อมีการพิมพ์ตัวอักษรหรือไดกราฟนั้น ๆ น้อยกว่าค่าที่กำหนด (fallback threshold) ซึ่งจะถือว่าค่าที่คำนวณได้นั้นจะไม่แม่นยำ และจะทำการ fallback โดยจะคำนวณค่าของคุณลักษณะในระดับที่สูงกว่าแทน

ผลการทดลองพบว่าจะให้ผลการยืนยันตัวตนได้ดีกว่า เมื่อใช้ข้อความประเภทเดียวกัน (เป็นข้อความที่กำหนดไว้ทั้งหมด หรือเป็นข้อความอิสระทั้งหมด) และเมื่อเป็นการพิมพ์บนคีย์บอร์ดประเภทเดียวกัน งานวิจัยนี้มีความน่าสนใจที่การเลือกใช้คุณลักษณะที่แตกต่างกันจำนวนมาก และมีการจัดการเมื่อมีบางตัวอักษรหรือไดกราฟมีจำนวนน้อยเกินไป แต่การใช้ค่าเหล่านั้นก็ใช้เพียงแค่ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานเท่านั้น

2.2.3. **งานวิจัยที่ใช้การผสานหลายคุณลักษณะ**

โดยทั่วไปแล้วการใช้หลาย ๆ คุณลักษณะหรือวิธีการมาผสานกันนั้น มักจะให้ผลลัพธ์ที่ดีกว่าการใช้เพียงคุณลักษณะหรือวิธีการเดียว Teh, P., Teoh, A., Tee, C. และ Ong, T. [16] ได้เสนอผลงานไว้ในปี 2009 โดยเสนอวิธีการใช้หลายคุณลักษณะ และหลายวิธีการมาประกอบกันอย่างเป็นระบบ เพื่อเพิ่มความถูกต้องในการยืนยันตัวตน โดยผู้เขียนได้ใช้คุณลักษณะของระยะเวลาในการพิมพ์ตัวอักษร และระยะเวลาระหว่างไดกราฟแบบต่าง ๆ รวมกัน และยังได้เสนอวิธีการเปรียบเทียบอีกสองแบบ คือ เปรียบเทียบด้วยการแจกแจงความน่าจะเป็นแบบปกติ (Gaussian probability density function) และการเปรียบเทียบจากการวัดความคล้ายคลึงของทิศทาง (Direction similarity measure) หรือการวัดความชันนั่นเอง นอกจากนี้ผู้เขียนยังได้นำเสนอวิธีการรวมผลจากหลายองค์ประกอบเข้าด้วยกัน โดยแบ่งเป็นระดับการให้คะแนน (score level) คือ ผลรวม ผลรวมแบบถ่วงน้ำหนัก ผลคูณ และการเลือกค่าที่มากที่สุดหรือน้อยที่สุด และในระดับการตัดสินใจ (decision level) คือ การให้ผ่านเมื่อมีส่วนใดส่วนหนึ่งให้ผ่านและการให้ผ่านเมื่อทุก ๆ ส่วนให้ผ่าน ซึ่งวิธีการเหล่านี้สามารถนำมาใช้เป็นโครงสร้างพื้นฐานสำหรับงานวิจัยอื่น ที่ใช้หลายคุณลักษณะหรือหลายวิธีการได้

ตัวอย่างงานวิจัยที่ใช้การผสานหลายคุณลักษณะ มีดังต่อไปนี้

Hosseinzadeh, D. และ Krishnan, S. [1] ได้เสนอผลงานไว้ในปี 2007 โดยนอกจากจะเสนอการพิจารณาระยะเวลาระหว่างไดกราฟทั้ง 4 แบบ และเสนอวิธีการเปรียบเทียบคุณลักษณะจากข้อความรหัสผ่าน โดยใช้การจำลองการผสานความน่าจะเป็นแบบเกาส์ (Gaussian mixture modeling) จากนั้นทำการทดลองเพื่อค้นหาว่าการผสานไดกราฟแบบใดที่ให้ผลการทดลองดีที่สุด ผู้เขียนยังเสนอโพรโตคอล ในการทำการวิจัยเกี่ยวกับระยะเวลาในการพิมพ์อีกด้วย โดยระบุในแง่ของ การเลือกใช้คุณลักษณะ การเลือกข้อความ วิธีการเก็บข้อมูล และจำนวนตัวอย่างที่ควรใช้เพื่อให้ผลการทดลองมีความน่าเชื่อถือ เป็นต้น

นอกจากงานวิจัยของ Chen, L., Weng, L. และ Chee, L. [6] แล้ว ยังมีอีกงานวิจัยหนึ่ง ที่ได้พิจารณาการใช้ระดับแรงกดเป็นอีกคุณลักษณะในการทดลอง คือ งานวิจัยของ Lv, H. และ Wang, W. [17] ซึ่งได้เสนอไว้ในปี 2006 โดยการทดลองนี้จะใช้คีย์บอร์ดที่สามารถรับรู้แรงกดได้ โดยคุณลักษณะของแรงกดที่ใช้มีอยู่ 5 ลักษณะคือ ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐาน พิสัย Positive Energy Center (PEC) และ Negative Energy Center (NEC) โดยทำการวัดระยะของแต่ละโพรไฟล์โดยใช้ระยะทางยูคลิด และวิธีไดนามิกไทม์วอร์ปปิง (dynamic time warping) เปรียบเทียบกับการใช้คุณลักษณะของคีย์สโตรกไดนามิกส์ ผลการทดลองพบว่าการใช้ลักษณะของแรงกดนั้นให้ค่า EER ต่ำกว่าการใช้ลักษณะของคีย์สโตรกไดนามิกส์อยู่เล็กน้อย และการใช้ทั้งสองลักษณะร่วมกันก็ทำให้ค่าความผิดพลาดลดลงไปอีก งานวิจัยนี้ทำให้เห็นว่าเราสามารถใช้ลักษณะของแรงกดในการยืนยันตัวบุคคลได้ แต่ค่าผิดพลาดอาจไม่ได้ลดลงจนถึงระดับที่คุ้มค่าที่จะต้องใช้คีย์บอร์ดพิเศษเหล่านี้

สำหรับงานวิจัยของ Gunetti, D. และ Picardi, C. [12] ก็อาจถือว่าเป็นงานวิจัยที่ใช้การผสานหลายคุณลักษณะได้เช่นกัน เพราะวิธีการนี้ใช้ทั้งการวัดแบบ R measure และ A Measure ร่วมกันเพื่อตัดสินใจจำแนกผู้ใช้ ดังรายละเอียดที่กล่าวไปข้างต้น

บทที่ 3
การออกแบบคุณลักษณะและวิธีการจำแนกผู้ใช้

งานวิจัยด้านการจำแนกตัวบุคคลโดยอาศัยข้อมูลระยะเวลาการพิมพ์ที่ใช้งานกับข้อความอิสระนั้นยังมีจำนวนน้อย งานวิจัยของ Gunetti, D. และ Picardi, C. [12] ถือว่าเป็นงานที่มีผลความแม่นในการจำแนกสูงสุดเท่าที่พบมา แต่ว่างานวิจัยนั้นให้ผลที่ไม่ค่อยดีกับข้อความอิสระที่มีขนาดสั้น ๆ วิทยานิพนธ์ฉบับนี้จึงออกแบกการเลือกคุณลักษณะและวิธีการโดยอาศัยการวิเคราะห์จุดอ่อนของวิธีการดังกล่าว และนำมาปรับปรุงหรือเปลี่ยนแปลงเพื่อให้ได้ผลความแม่นที่ดีขึ้นเมื่อใช้กับข้อความอิสระขนาดสั้น

จากการทดลองพบว่า การคำนวณระยะห่างทั้งสองวิธีของ Gunetti, D. และ Picardi, C. [12] นั้นให้ผลไม่ดีเมื่อคำนวณจากข้อความที่มีขนาดสั้น ซึ่งคาดว่าเมื่อข้อความมีขนาดสั้น ในการเปรียบเทียบกันแล้วจะพบว่ามีไดกราฟที่ซ้ำกันน้อย ทำให้ค่าระยะทางที่ได้จากการวัดไม่สามารถอธิบายความแตกต่างได้ดี เมื่อนำค่าระยะทางเหล่านี้ไปใช้ในการจำแนกผู้ใช้จึงทำให้ได้ความแม่นน้อย

จากเหตุผลดังกล่าว จึงออกแบบวิธีการจำแนกด้วยวิธีอื่นที่ไม่ใช่การวัดระยะทางระหว่างสองตัวแทน โดยได้เลือกใช้นิวรอลเน็ตเวิร์กมาช่วยจำแนก เพราะการใช้นิวรอลเน็ตเวิร์กนั้นเป็นการสร้างโพรไฟล์จากชุดข้อมูลฝึกทั้งหมดก่อน จากนั้นจึงนำข้อมูลทดสอบมาเปรียบเทียบกับโพรไฟล์ที่ถูกสร้างขึ้น

ในการออกแบบคุณลักษณะนั้น มีข้อสันนิษฐานบางประการเกี่ยวกับระยะเวลาในการพิมพ์ของผู้ใช้ ประการแรก คือ เมื่อพิจารณาระยะเวลาระหว่างไดกราฟของผู้ใช้แต่ละคน จะมีบางไดกราฟที่มีการกระจายตัวอย่างหนาแน่นอยู่ในบางช่วงเวลา ซึ่งช่วงเวลานั้นจะไม่ซ้ำกับการกระจายตัวของไดกราฟประเภทนั้นของผู้ใช้คนอื่นเลยหรือคล้ายคลึงกับเพียงบางผู้ใช้เท่านั้น การใช้ไดกราฟประเภทที่มีลักษณะเช่นนี้จะเป็นส่วนสำคัญที่ใช้ในการจำแนกผู้ใช้แต่ละคนออกจากกัน ประการที่สอง คือ ระยะเวลาระหว่างไดกราฟของผู้ใช้นั้นอาจเปลี่ยนแปลงไปตามเวลาหรือด้วยเหตุผลอื่น ๆ แต่การเปลี่ยนแปลงนั้นจะเป็นไปในแนวทางเดียวกับสำหรับไดกราฟทุก ๆ ประเภท

จากข้อสันนิษฐานข้อแรก ทำให้นิวรอลเน็ตเวิร์กเป็นตัวเลือกที่เหมาะสมที่จะนำมาใช้งาน เพราะนิวรอลเน็ตเวิร์กสามารถให้ค่าน้ำหนักที่แตกต่างกันสำหรับแต่ละคุณลักษณะเพื่อบ่งบอกระดับความสำคัญของคุณลักษณะที่มีผลต่อการจำแนกได้เป็นอย่างดีอีกด้วย

ข้อจำกัดอย่างหนึ่งของการใช้นิวรอลเน็ตเวิร์ก คือ จำนวนอินพุตของเน็ตเวิร์กจะต้องมีจำนวนคงที่ การนำค่าระยะเวลาระหว่างคู่ไดกราฟมาใช้เป็นอินพุตโดยตรงจึงเป็นไปไม่ได้

ในกรณีที่ข้อความที่ใช้เป็นข้อความอิสระ จึงมีความจำเป็นที่จะต้องแปลงค่าระยะเวลาระหว่างไดกราฟมาเป็นคุณลักษณะในรูปแบบอื่น ที่มีจำนวนอินพุตคงที่ไม่ว่าข้อความจะมีลักษณะเป็นอย่างไรก่อน จึงจะนำมาใช้งานกับนิวรอลเน็ตเวิร์กได้

หากพิจารณาวิธีการคำนวณระยะทางของ Gunetti, D. และ Picardi, C. [12] จะพบว่า การใช้ R Measure นั้นจะยังสามารถทำนายผลการทดสอบได้คงเดิมแม้ระยะเวลาระหว่างคู่ไดกราฟจะเปลี่ยนไปแต่เปลี่ยนไปในทางเดียวกันทั้งหมด และการใช้ A Measure จะจำกัดความแตกต่างของระยะเวลาระหว่างไดกราฟให้อยู่ในช่วงที่ยอมรับได้ จากข้อสันนิษฐานข้อที่สอง จึงออกแบบการเลือกใช้คุณลักษณะที่รองรับการเปลี่ยนแปลงของระยะเวลาระหว่างไดกราฟ เพื่อให้สอดคล้องกับความแปรปรวนที่มีอยู่ในธรรมชาติของระยะเวลาในการพิมพ์ และยังได้ออกแบบคุณลักษณะที่เป็นการเปรียบเทียบความคล้ายคลึงกันของระยะเวลาในการพิมพ์ของผู้ใช้อีกอย่างหนึ่ง นอกจากนี้ยังเสนอวิธีการใช้คุณลักษณะทั้งสองร่วมกันเพื่อให้ได้ผลการจำแนกที่ดีขึ้นอีกด้วย

3.1. การใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟเป็นคุณลักษณะสำหรับเวกเตอร์อินพุต

การใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟเพื่อเป็นคุณลักษณะนั้น เป็นที่นิยมในงานวิจัยที่สนใจข้อความที่ถูกกำหนดไว้ แต่ก็สามารถนำมาปรับใช้กับงานที่ใช้ข้อความอิสระได้ไม่ยาก ข้อสังเกตอันดับแรกคือ เวกเตอร์อินพุตจะต้องมีรูปแบบที่คงที่ไม่ว่าข้อความที่จะนำมาทดสอบนั้นจะเป็นเช่นไร เพื่อให้สามารถนำมาใช้กับนิวรอลเน็ตเวิร์กได้

ขั้นตอนแรกของการสร้างอินพุตเวกเตอร์ คือการกำหนดรูปแบบของเวกเตอร์อินพุต โดยกำหนดให้เป็นค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของไดกราฟแต่ละประเภทที่ได้พบขณะแปลงข้อความอิสระมาเป็นอินพุตเวกเตอร์ การกำหนดรูปแบบนั้นทำได้โดย พิจารณาข้อความทุก ๆ ข้อความในชุดข้อมูลฝึกและนำประเภทไดกราฟที่พบทั้งหมดในชุดข้อมูลฝึกมาเรียงต่อกัน โดยจะเรียงในลำดับใดก็ได้ และกำหนดรายการของประเภทไดกราฟนั้นว่าเป็น*รูปแบบของเวกเตอร์อินพุต*

ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานนั้นจะเป็นตัวแทนที่ดีของข้อมูลเมื่อข้อมูลมีปริมาณมากเพียงพอ ดังนั้น เพื่อลดโอกาสที่จะพบข้อมูลที่ผิดปกติหรือข้อมูลของประเภทไดกราฟที่ปรากฏอยู่น้อยในชุดข้อมูลฝึก ในการกำหนดรูปแบบอินพุต จะทำการตัดประเภทไดกราฟที่พบในชุดข้อมูลฝึกน้อยกว่า 30 ครั้งออกไปด้วย

หลังจากได้รูปแบบของเวกเตอร์อินพุตที่สมบูรณ์แล้ว ขั้นตอนต่อไปคือการแปลงข้อความในชุดข้อมูลฝึกเพื่อเป็นเวกเตอร์อินพุตสำหรับใช้ฝึกนิวรอลเน็ตเวิร์ก การแปลงข้อความ

เป็นเวกเตอร์อินพุต คือการพิจารณาระยะเวลาระหว่างไดกราฟที่ปรากฏอยู่ในข้อความ นำมาคำนวณหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานแยกตามประเภทของไดกราฟ แล้วนำมาจัดเรียงให้มีรูปแบบตรงกันกับรูปแบบที่กำหนดไว้ ในการคำนวณนั้นมีหลักการพื้นฐาน ดังนี้

1) หากในข้อความมีข้อมูลระยะเวลาการพิมพ์ของไดกราฟประเภทนั้น ๆ อยู่มากกว่า 2 ครั้ง ให้คำนวณหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานตามปกติ
2) หากในข้อความมีข้อมูลระยะเวลาการพิมพ์ของไดกราฟประเภทนั้น ๆ อยู่เพียงครั้งเดียว ให้ค่าเฉลี่ยมีค่าเป็นค่าระยะเวลานั้น และกำหนดให้ส่วนเบี่ยงเบนมาตรฐานมีค่าเป็น 0
3) หากในข้อความไม่ปรากฏข้อมูลระยะเวลาการพิมพ์ของไดกราฟประเภทนั้น ๆ เลย ให้กำหนดทั้งค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานมีค่าเป็น 0
4) หากในข้อความมีข้อมูลระยะเวลาระหว่างไดกราฟที่ไม่ได้ปรากฏอยู่ในรูปแบบ ให้ละทิ้งข้อมูลระยะเวลาเหล่านั้น

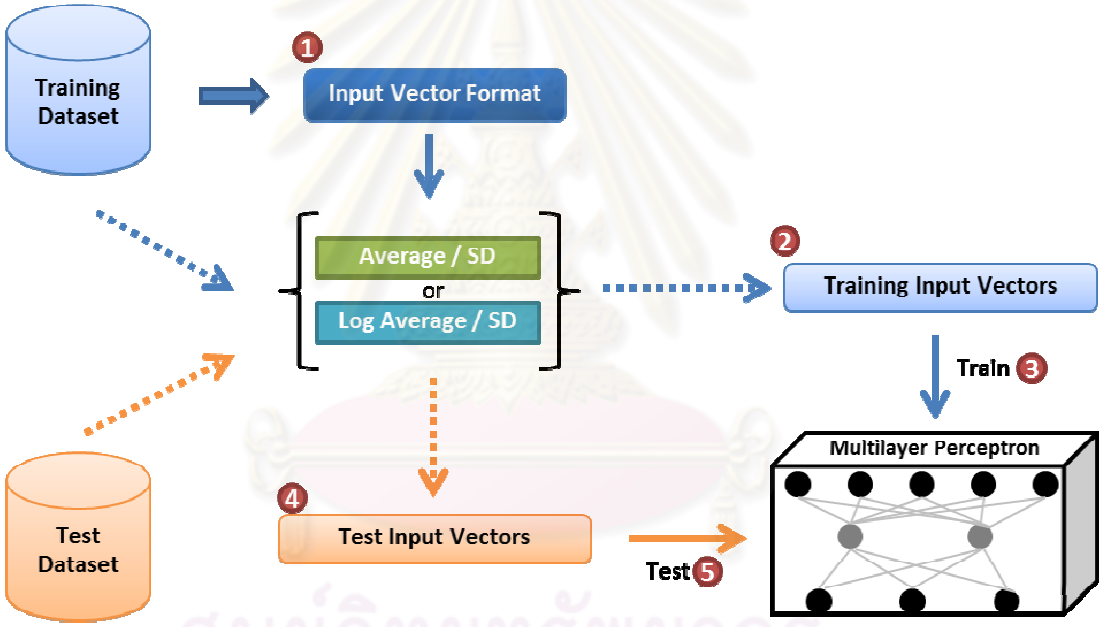
เมื่อแปลงชุดข้อมูลฝึกให้เป็นเวกเตอร์อินพุตสำหรับฝึกได้แล้ว ก็นำเวกเตอร์อินพุตเหล่านั้นไปให้นิวรอลเน็ตเวิร์กเรียนรู้ เพื่อสร้างโมเดลในการจำแนกออกมา เมื่อต้องการทำการทดสอบกับข้อมูลทดสอบ ก็นำข้อความที่จะทดสอบมาแปลงเป็นเวกเตอร์อินพุตด้วยวิธีเดียวกัน ก็จะสามารถทำการจำแนกโดยใช้นิวรอลเน็ตเวิร์กที่เรียนรู้แล้วได้

อีกประการหนึ่ง ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานนั้นจะเป็นตัวแทนที่ดีถ้าข้อมูลมีการกระจายตัวเป็นแบบปกติ อย่างไรก็ดี Montalv, J. และ Freire, E. [2] ได้เสนอไว้ในงานวิจัยว่า การกระจายตัวของระยะเวลาระหว่างไดกราฟนั้นมีลักษณะเป็นแบบล็อกนอร์มัล ดังนั้น หากต้องการให้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานยังคงเป็นตัวแทนข้อมูลที่ดีอยู่ เราจะพิจารณา*ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมของระยะเวลาระหว่างไดกราฟ* เป็นคุณลักษณะอีกแบบหนึ่งที่จะนำมาทดสอบด้วย โดยการสร้างนั้นทำเหมือนกับการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานแบบปกติทุกประการ ยกเว้นแต่การคำนวณหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานตอนแปลงข้อความเป็นเวกเตอร์อินพุตเท่านั้น ที่จะทำการหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมของระยะเวลาระหว่างไดกราฟแทน

การใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานนั้นสามารถรองรับการเปลี่ยนแปลงของระยะเวลาระหว่างไดกราฟดังข้อสันนิษฐานข้อที่สองได้ รายละเอียดของเรื่องนี้จะแสดงในบทถัดไป (หัวข้อ 4.4.3)

โดยสรุป การใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟเป็นคุณลักษณะมีขั้นตอน ดังนี้ (ขั้นตอนต่างๆสามารถแสดงได้ดังภาพที่ 3-1)

1) กำหนดรูปแบบของเวกเตอร์อินพุต โดยการพิจารณาชุดข้อมูลฝึก
2) แปลงข้อมูลฝึกเป็นชุดของเวกเตอร์อินพุต โดยการคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟ หรือ ค่าลอการิทึมของระยะเวลาระหว่างไดกราฟ
3) สร้างโมเดลการจำแนก โดยให้นิวรอลเน็ตเวิร์กเรียนรู้ชุดข้อมูลฝึก
4) แปลงข้อมูลทดสอบเป็นชุดของเวกเตอร์อินพุตด้วยวิธีเดียวกัน
5) ทำการทดสอบด้วยชุดข้อมูลทดสอบ



ภาพที่ 3-1 ภาพแสดงขั้นตอนการดำเนินงานเมื่อใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟเป็นคุณลักษณะสำหรับเวกเตอร์อินพุต

3.2. การใช้ค่าเฉลี่ยความน่าจะเป็นเพื่อเป็นคุณลักษณะสำหรับเวกเตอร์อินพุต

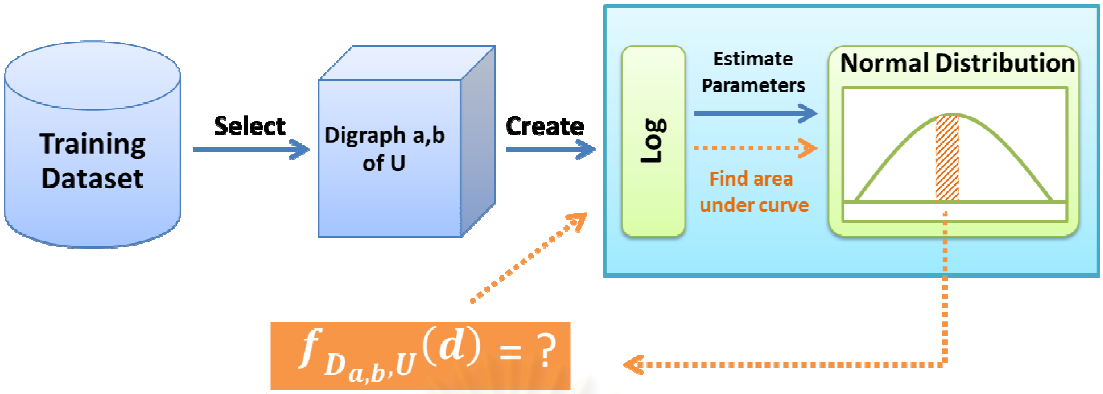
คุณลักษณะอีกแบบหนึ่งที่น่าจะใช้ในการจำแนกผู้ใช้ได้ คือการเปรียบเทียบค่าระยะเวลาระหว่างไดกราฟที่อยู่ในข้อความที่ต้องการทดสอบ กับการแจกแจงความน่าจะเป็นของระยะเวลาระหว่างไดกราฟประเภทนั้น ๆ ของผู้ใช้แต่ละคนที่ได้มาจากชุดข้อมูลฝึก ว่าไดกราฟที่มีระยะเวลาดังในข้อความทดสอบนั้น มีความคล้ายคลึงหรือน่าจะถูกสร้างขึ้นจากผู้ใช้แต่ละคนมากน้อยเท่าใด

หากนำระดับความคล้ายคลึงของไดกราฟที่นำทดสอบเมื่อเปรียบเทียบกับข้อมูลที่มีอยู่ของแล้วของผู้ใช้แต่ละคนมาเป็นคุณลักษณะ การใช้นิวรอลเน็ตเวิร์กในการเรียนรู้ก็จะเป็นการให้ค่าน้ำหนักหรือให้ความสำคัญของไดกราฟแต่ละประเภทเพื่อให้สามารถจำแนกได้อย่างถูกต้อง ซึ่งสอดคล้องกับข้อสันนิษฐานข้อแรก การใช้คุณลักษณะนี้กับนิวรอลเน็ตเวิร์กจึงน่าจะให้ผลการจำแนกที่ดีได้

ในขั้นตอนแรก จะต้องทำการสร้างโพรไฟล์ความน่าจะเป็นจากชุดข้อมูลฝึกก่อน เพื่อที่จะทำการแปลงจากข้อความที่ใช้ฝึกให้เป็นเวกเตอร์อินพุตได้ในภายหลัง การสร้างโพรไฟล์ความน่าจะเป็นทำได้โดย พิจารณาข้อมูลระยะเวลาระหว่างไดกราฟในชุดข้อมูลฝึก จากนั้นสร้างฟังก์ชันการแจกแจงความน่าจะเป็นของระยะเวลาระหว่างไดกราฟ โดยสร้างแยกกันในไดกราฟแต่ละประเภทจากผู้ใช้แต่ละคน เช่นเดียวกับการกำหนดรูปแบบของเวกเตอร์อินพุตในหัวข้อที่ผ่านมา ถ้าจะให้ฟังก์ชันการแจกแจงความน่าจะเป็นจะอธิบายธรรมชาติของการกระจายตัวของระยะเวลาได้ ก็จะต้องมีข้อมูลมากเพียงพอที่จะสร้างฟังก์ชันดังกล่าว ดังนั้น ในการพิจารณาข้อมูลเพื่อสร้างฟังก์ชันการแจกแจงความน่าจะเป็นนั้น จะไม่พิจารณาประเภทของไดกราฟของผู้ใช้ที่มีข้อมูลของระยะเวลาอยู่ไม่ถึง 30 ข้อมูล

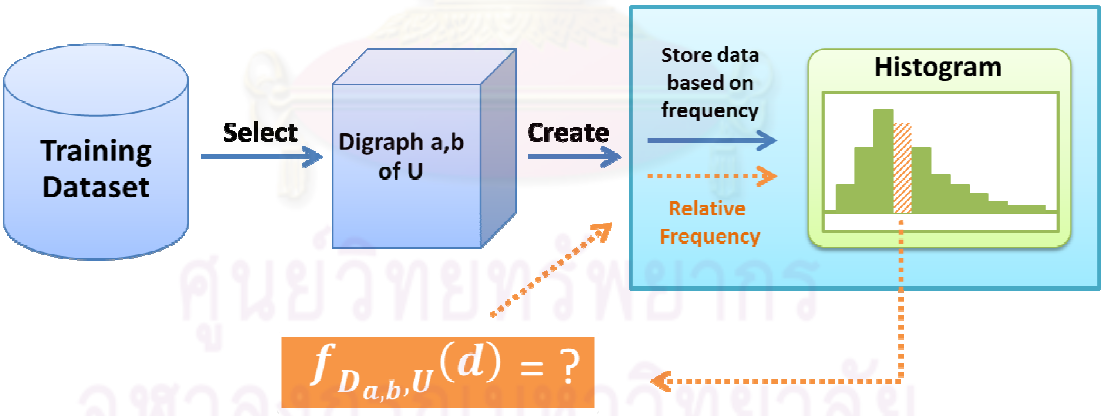
กำหนดให้ $f_{D_{a,b},U}(x)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นของระยะเวลาระหว่างไดกราฟ a,b ของผู้ใช้ U ฟังก์ชันนี้จะถูกสร้างขึ้นจากการพิจารณาข้อมูลระยะเวลาระหว่างไดกราฟ a,b ของผู้ใช้ U ทุก ๆ ข้อมูลในชุดข้อมูลฝึก โดยเราเสนอวิธีการสร้างฟังก์ชันดังกล่าว 2 วิธี คือ *สร้างจากการประมาณค่าพารามิเตอร์ของการแจกแจงความน่าจะเป็นแบบล็อกนอร์มัล* และ *สร้างจากการคำนวณความถี่สัมพัทธ์ของฮิสโตแกรม*

เนื่องจากเราทราบว่า ระยะเวลาระหว่างไดกราฟนั้นมีการแจกแจงความน่าจะเป็นแบบล็อกนอร์มัล การสร้างฟังก์ชันการแจกแจงความน่าจะเป็นจากการประมาณค่าพารามิเตอร์ของการแจกแจงความน่าจะเป็นแบบล็อกนอร์มัลนั้นจึงเป็นเรื่องที่สามารถทำได้ เพื่อความง่ายในการคำนวณ จึงประมาณค่าพารามิเตอร์โดยทำการหาค่าลอการิทึมของระยะเวลาระหว่างไดกราฟ จากนั้นจึงนำมาประมาณค่าพารามิเตอร์ตามแบบการแจกแจงความน่าจะเป็นแบบปกติ สำหรับการหาค่าความน่าจะเป็น จะคำนวณจากพื้นที่ใต้กราฟของฟังก์ชันการแจกแจงแบบต่อเนื่อง โดยมีความกว้างเป็น 1 มิลลิวินาที กล่าวคือ หากต้องการหาความน่าจะเป็นที่จะเกิดไดกราฟที่มีระยะเวลา x มิลลิวินาที จะหาจากพื้นที่ใต้กราฟของฟังก์ชันการแจกแจงแบบต่อเนื่อง ตั้งแต่ระยะเวลา x-0.5 มิลลิวินาที ถึง x+0.5 มิลลิวินาที ซึ่งขั้นตอนการสร้างฟังก์ชันและการคำนวณความน่าจะเป็นสามารถแสดงได้ดังภาพที่ 3-2



ภาพที่ 3-2 ภาพแสดงการสร้างฟังก์ชันการแจกแจงความน่าจะเป็นจากการประมาณค่าพารามิเตอร์ของการแจกแจงแบบล็อกนอร์มัล และการหาค่าความน่าจะเป็น

ในอีกแง่มุมหนึ่ง หากเราคิดว่าในบางครั้งการประมาณการแจกแจงความน่าจะเป็นของระยะเวลาระหว่างไดกราฟให้เป็นแบบล็อกนอร์มัล อาจไม่ได้มีลักษณะเหมือนการกระจายตัวที่แท้จริง เราอาจใช้การแจกแจงความถี่หรือฮิสโตแกรมในการแจกแจงความน่าจะเป็นแทนได้ สำหรับการหาค่าความน่าจะเป็น สามารถหาได้โดยพิจารณาค่าความถี่สัมพัทธ์เป็นค่าความน่าจะเป็นจากฟังก์ชันการแจกแจงแบบไม่ต่อเนื่องได้ทันที โดยขั้นตอนการสร้างฟังก์ชันแจกแจงและการคำนวณความน่าจะเป็นสามารถแสดงได้ดังภาพที่ 3-3



ภาพที่ 3-3 ภาพแสดงการสร้างฟังก์ชันการแจกแจงความน่าจะเป็นจากฮิสโตแกรม และการหาค่าความน่าจะเป็น

เมื่อสร้างโพรไฟล์ความน่าจะเป็นจากข้อมูลฝึกเสร็จสิ้นแล้ว ขั้นตอนต่อไปจะเป็นการแปลงข้อความในชุดข้อมูลฝึกเพื่อเป็นเวกเตอร์อินพุตที่นิวรอลเน็ตเวิร์กจะใช้ในการเรียนรู้ เพื่อสร้างโมเดลในการจำแนกผู้ใช้ต่อไป โดยการแปลงนั้นจะเป็นการพิจารณาว่า ระยะเวลาระหว่างไดกราฟที่ปรากฏอยู่ในข้อความนั้นมีความคล้ายคลึงหรือน่าจะถูกสร้างมาจากผู้ใช้แต่ละคนมาก

น้อยเพียงใด หรืออาจกล่าวได้ว่าเมื่อพิจารณาเทียบกับฟังก์ชันการแจกแจงความน่าจะเป็นแต่ละฟังก์ชันในโพรไฟล์ความน่าจะเป็นแล้วได้ค่าความน่าจะเป็นออกมาเป็นเท่าใด เนื่องจากในข้อความหนึ่ง ๆ อาจมีข้อมูลระยะเวลาระหว่างไดกราฟประเภทเดียวกันซ้ำกันหลาย ๆ ครั้ง การคำนวณจึงใช้ค่าเฉลี่ยความน่าจะเป็นแทน ซึ่งการแปลงมีรายละเอียดแต่ละขั้นตอนดังนี้

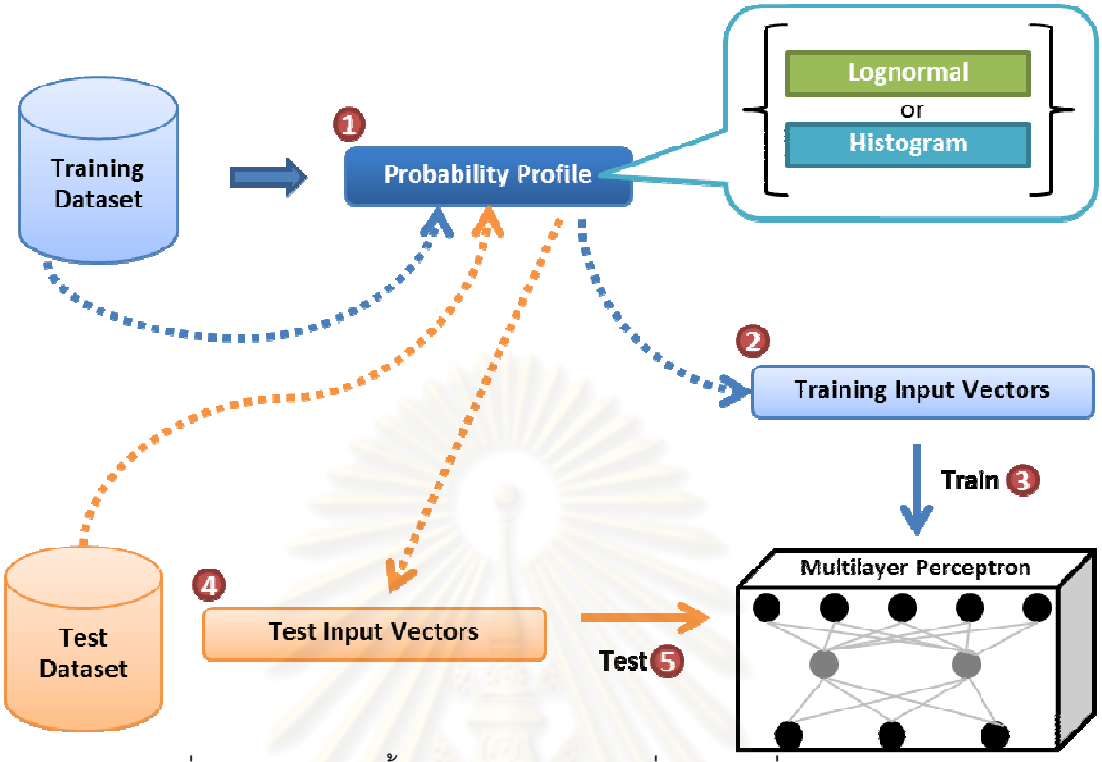
1) พิจารณาฟังก์ชันการแจกแจงความน่าจะเป็นของระยะเวลาระหว่างไดกราฟ $f_{D_{a,b},U}(x)$ ในโพรไฟล์ความน่าจะเป็นทีละฟังก์ชัน
2) พิจารณาข้อมูลระยะเวลาระหว่างไดกราฟ a, b ในข้อความที่จะนำมาแปลงทุก ๆ ข้อมูล คำนวณหาค่าเฉลี่ยความน่าจะเป็นที่ข้อมูลเหล่านั้นจะถูกสร้างมาจากฟังก์ชัน $f_{D_{a,b},U}(x)$ ดังสมการที่ 1

$$AvgProb_{f_{D_{a,b},U}}(S) = \frac{\sum_{d \in D_{a,b} \text{ in } S} \{f_{D_{a,b},U}(d)\}}{|D_{a,b} \text{ in } S|} \quad (1)$$

3) หากในข้อความไม่ปรากฏข้อมูลระยะเวลาการพิมพ์ของไดกราฟประเภทนั้น ๆ เลย ให้กำหนดค่า ณ ตำแหน่งนั้น ๆ ของเวกเตอร์อินพุตเป็น 0 เนื่องจากถือว่าค่าความน่าจะเป็นที่ข้อความนี้จะเป็นของผู้ใช้คนนั้น เมื่อพิจารณาด้วยไดกราฟประเภทดังกล่าวมีค่าเป็น 0
4) พิจารณาฟังก์ชันการแจกแจงความน่าจะเป็นของระยะเวลาระหว่างไดกราฟฟังก์ชันอื่น ๆ จนครบ นำค่าความน่าจะเป็นเฉลี่ยที่คำนวณได้จากทุก ๆ ฟังก์ชันมาเป็นเวกเตอร์อินพุตต่อไป

ลำดับขั้นตอนข้างต้นสามารถแสดงได้ดังภาพที่ 3-4

การใช้ค่าเฉลี่ยความน่าจะเป็นมาเป็นคุณลักษณะสำหรับเวกเตอร์อินพุตนั้น น่าจะเป็นตัวแทนที่ดีกว่าการใช้เพียงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน เพราะมีการเปรียบเทียบกับการแจกแจงความน่าจะเป็นของระยะเวลาระหว่างไดกราฟโดยตรง อย่างไรก็ดี การใช้คุณลักษณะแบบนี้อาจไม่รองรับการเปลี่ยนแปลงของระยะเวลาระหว่างไดกราฟตามข้อสันนิษฐานข้อที่ 2 ดังที่จะกล่าวในบทถัดไป (หัวข้อ 4.4.3)



ภาพที่ 3-4 ภาพแสดงขั้นตอนการดำเนินงานเมื่อใช้ค่าเฉลี่ยความน่าจะเป็นเพื่อเป็นคุณลักษณะสำหรับเวกเตอร์อินพุต

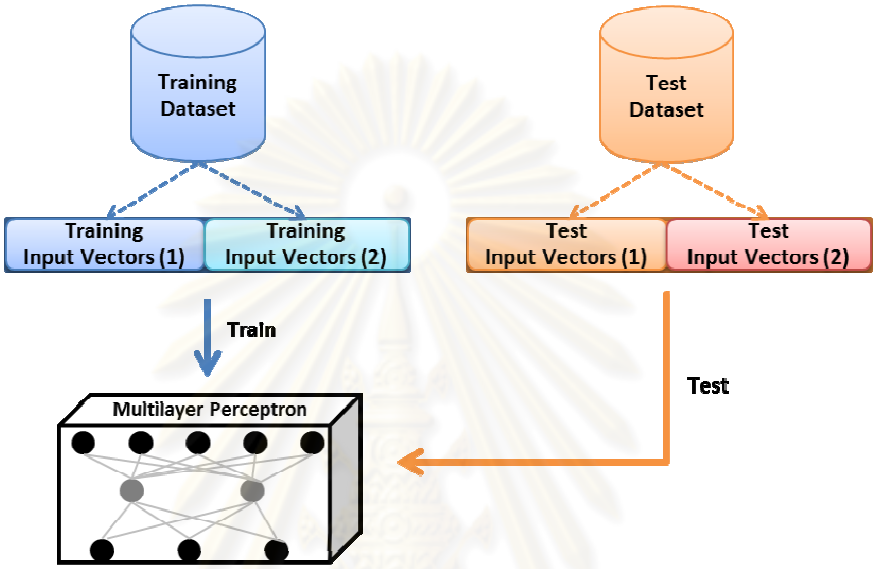
3.3. การผสานหลายคุณลักษณะในการจำแนกผู้ใช้

โดยทั่วไปแล้ว การผสานคุณลักษณะหรือตัวจำแนกเพื่อทำนายผลลัพทธ์สุดท้ายมักจะให้ผลดีกว่าการใช้แต่ละคุณลักษณะหรือตัวจำแนกเดี่ยว เพราะจะมีข้อมูลที่มากขึ้นซึ่งจะช่วยให้การจำแนกได้ผลความแม่นสูงขึ้น

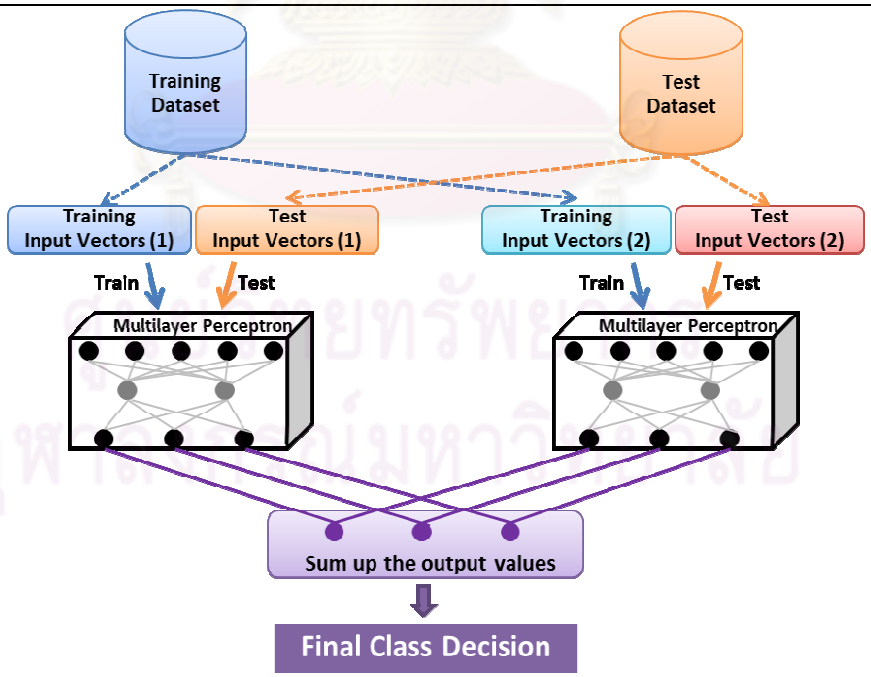
เราจะเลือกคุณลักษณะที่ให้ผลความแม่นสูงกว่าในการทดลองการจำแนกโดยใช้คุณลักษณะเดี่ยว ๆ เพื่อนำมาทดลองการจำแนกโดยการผสานคุณลักษณะ กล่าวคือ จะเลือกคุณลักษณะที่ให้ผลความแม่นสูงกว่า ระหว่างการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟและค่าลอการิทึมของระยะเวลาระหว่างไดกราฟ มาผสานกับคุณลักษณะที่ให้ผลความแม่นสูงกว่า ระหว่างการใช้ค่าเฉลี่ยความน่าจะเป็นจากฟังก์ชันซึ่งประมาณด้วยการแจกแจงความน่าจะเป็นแบบล็อกนอร์มัล และการคำนวณความน่าจะเป็นจากฮิสโตแกรม

การผสานนั้นอาจทำได้ในสองรูปแบบ รูปแบบแรกคือการนำเวกเตอร์อินพุตของคุณลักษณะทั้งสองมาต่อกัน จากนั้นจึงนำไปให้นิวรอลเน็ตเวิร์กเรียนรู้ทีเดียว เรียกการผสานลักษณะนี้ว่า *การผสานอินพุต (Input Combination)* วิธีที่สองนั้นคือ การนำเวกเตอร์อินพุตของคุณลักษณะทั้งสองไปให้นิวรอลเน็ตเวิร์กเรียนรู้แยกจากกัน การทดสอบกับตัวอย่างทดสอบก็ทำ

แยกกันเช่นกัน แต่แทนที่จะนำผลที่ได้จากการจำแนกโดยนิวรอลเน็ตเวิร์กแต่ละเน็ตเวิร์กมาใช้เลย ก็นำผลลัพธ์มาผสานกัน โดยนำค่าที่ได้จากโหนดในชั้นเอาต์พุตของนิวรอลเน็ตเวิร์กทั้งคู่มาบวกกันเป็นทีละคู่ จากนั้นพิจารณาว่าผลรวมของโหนดคู่ใดมีค่ามากที่สุด ก็ให้จำแนกผลสุดท้ายเป็นผู้ใช้นั้น ๆ เรียกการผสานลักษณะนี้ว่า *การผสานผลลัพธ์ (Result Combination)* ซึ่งขั้นตอนของวิธีการทั้งสองสามารถแสดงได้ดังภาพที่ 3-5



ก) Input Combination



ข) Result Combination

ภาพที่ 3-5 ภาพแสดงวิธีการผสานคุณลักษณะ

ก) การผสานอินพุต ข) การผสานผลลัพธ์

บทที่ 4
ผลการทดลองและวิเคราะห์ผล

ในบทนี้จะทำการวิเคราะห์ความสามารถในการจำแนกผู้ใช้เมื่อใช้วิธีการที่ได้นำเสนอ โดยเริ่มจากการกำหนดตัววัดผล การออกแบบการทดลองและเครื่องมือที่ใช้ในการทดลอง ผลการทดลองที่ได้จากการใช้วิธีที่ได้นำเสนอมาจำแนกผู้ใช้บนข้อความอิสระที่มีความยาวต่าง ๆ กัน โดยเปรียบเทียบผลการทดลองกับวิธีอื่น และการวิเคราะห์ผลการทดลอง เพื่อแสดงให้เห็นความแตกต่างระหว่างวิธีที่นำเสนอ กับวิธีอื่น ๆ ที่มีมาก่อนหน้านี้

4.1. ตัววัดผล

ในวิทยานิพนธ์ฉบับนี้ ได้กำหนดตัววัดผลการทดลองออกเป็น 3 อย่าง คือ

1) ความแม่น (Accuracy) : ความแม่นเป็นตัววัดหลัก ที่จะแสดงให้เห็นถึงความสามารถในการจำแนกผู้ใช้ของแต่ละวิธีการ ความแม่นนั้นเป็นร้อยละซึ่งคำนวณจากอัตราของตัวอย่างที่ระบบสามารถจำแนกผู้ใช้ได้อย่างถูกต้อง ต่อจำนวนอย่างทั้งหมดในการทดลองแต่ละครั้ง
2) False Positive ของผู้ใช้แต่ละคน : คำนวณจากร้อยละของจำนวนครั้งที่ระบบจำแนกตัวอย่างเป็นของผู้ใช้คนนั้น แต่จริง ๆ แล้วตัวอย่างนั้นไม่ได้เป็นของผู้ใช้คนนั้น ต่อจำนวนตัวอย่างทั้งหมดในการทดลองแต่ละครั้ง False Positive จะแสดงให้เห็นโอกาสที่ระบบจะถูกบุกรุกสำเร็จ เมื่อผู้บุกรุกปลอมแปลงเป็นผู้ใช้คนหนึ่ง ๆ ในระบบ
3) False Negative ของผู้ใช้แต่ละคน : คำนวณจากร้อยละของจำนวนครั้งที่ระบบจำแนกตัวอย่างของผู้ใช้คนนั้นผิดไปเป็นตัวอย่างของผู้ใช้คนอื่น ต่อจำนวนตัวอย่างทั้งหมดในการทดลองแต่ละครั้ง False Negative จะแสดงให้เห็นถึงโอกาสที่ผู้ใช้จะเกิดความไม่สะดวกในการใช้ เนื่องจากระบบไม่ยอมให้ผู้ใช้ที่แท้จริงเข้าใช้งานระบบ

4.2. การออกแบบการทดลอง

วิทยานิพนธ์นี้จะพิจารณาผลการทดลองการจำแนกผู้ใช้ด้วยข้อความอิสระขนาดต่าง ๆ กัน ดังนั้น ในการทดสอบแต่ละครั้ง จะทำการแบ่งข้อมูลทดสอบทั้งหมดออกเป็นข้อความสั้น ๆ หลาย ๆ ข้อความที่มีความยาวแตกต่างกันไปในการทดลองแต่ละครั้ง ตั้งแต่ความยาว 100 ตัวอักษร ถึง 1000 ตัวอักษร เพื่อเน้นผลการทดลองจำแนกผู้ใช้ในกรณีที่ใช้ข้อความอิสระขนาดสั้น

ในส่วนของชุดข้อมูลฝึก ก็จะถูกแบ่งออกเป็นข้อความที่มีความยาว 100 500 และ 1000 ตัวอักษร เช่นกัน เพื่อศึกษาผลของความยาวของข้อความฝึกที่ส่งผลกระทบต่อผลการจำแนกผู้ใช้

ในการทดสอบ จะทำการทดสอบแบบไขว้ข้าม 10 พับ (10-fold cross validation) โดยผลการทดสอบที่แสดงจะเป็นผลที่เฉลี่ยมาจากการทดสอบทั้ง 10 ครั้ง

ในการจำแนก จะใช้นิวรอลเน็ตเวิร์กแบบหลายชั้นเป็นตัวจำแนก โดยในการทดลองนี้ได้ใช้โปรแกรม WEKA [18] ในการสร้างนิวรอลเน็ตเวิร์ก

4.3. ผลการทดลอง

4.3.1. การจำแนกด้วยวิธีของ D. Gunetti และ C. Picardi

วิธีการของ D. Gunetti และ C. Picardi นั้นถือว่าเป็นงานวิจัยที่รายงานผลลัพธ์การจำแนกผู้ใช้ด้วยข้อความอิสระที่ให้ผลดีที่สุดในบรรดางานวิจัยที่มีมา วิทยานิพนธ์ฉบับนี้จึงนำวิธีการนี้มาทดลองซ้ำเพื่อเปรียบเทียบผลลัพธ์กับวิธีที่นำเสนอ

ในงานวิจัยของ D. Gunetti และ C. Picardi ได้นำเสนอวิธีการวัดระยะห่างระหว่างตัวอย่างสองตัวไว้เป็นจำนวนมาก เช่น ระยะห่างสัมพัทธ์ของไดกราฟ (R_2), ระยะห่างสัมพัทธ์ของไดกราฟและไตรกราฟ ($R_{2,3}$), ระยะห่างสัมพัทธ์ของไดกราฟ ไตรกราฟ และ 4-กราฟ ($R_{2,3,4}$), ระยะห่างสัมบูรณ์ของไดกราฟ (A_2) และ ระยะห่างสัมบูรณ์ของไดกราฟและไตรกราฟ ($A_{2,3}$) โดยได้เสนอวิธีการวัดระยะทางที่ให้ผลดีที่สุดในการทดลองการจำแนกผู้ใช้ไว้สามแบบ นั่นคือ การใช้ระยะห่างสัมพัทธ์ของไดกราฟควบคู่กับระยะห่างสัมบูรณ์ของไดกราฟ (R_2+A_2), การใช้ระยะห่างสัมพัทธ์ของไดกราฟและไตรกราฟควบคู่กับระยะห่างสัมบูรณ์ของไดกราฟ ($R_{2,3}$+A_2) และการใช้ระยะห่างสัมพัทธ์ของไดกราฟ ไตรกราฟและ 4-กราฟ ควบคู่กับระยะห่างสัมบูรณ์ของไดกราฟและไตรกราฟ ($R_{2,3,4}$+$A_{2,3}$)

ในวิทยานิพนธ์ฉบับนี้จึงทำการทดลองโดยใช้วิธีการวัดระยะห่างทั้งสามวิธีข้างต้น จากผลการทดลองพบว่า การใช้ระยะห่างแบบ R_2+A_2 นั้นให้ผลการทดลองที่ดีที่สุด จึงทำการทดลองเพิ่มเติมโดยใช้การแบ่งชุดข้อมูลฝึกให้เป็นแต่ละตัวอย่างมีความยาวที่ 100 500 และ 1000 ตัวอักษรตามลำดับ ดังแสดงผลในตารางที่ 4-1

ตารางที่ 4-1 ตารางแสดงความแม่น (ค่าเฉลี่ยร้อยละ ± ส่วนเบี่ยงเบนมาตรฐาน)
ในการจำแนกผู้ใช้ เมื่อใช้วิธีการของ D. Gunetti และ C. Picardi

Length of Test Sample (Characters)	Length of Training Sample (Characters)				
	1000			500	100
	$R_{2,3,4}+A_{2,3}$	$R_{2,3}+A_2$	R_2+A_2	R_2+A_2	R_2+A_2
100	13.115 ± 1.01	15.749 ± 1.09	33.375 ± 2.28	25.617 ± 1.54	10.128 ± 1.72
200	23.952 ± 2.17	33.608 ± 3.47	54.442 ± 2.09	45.853 ± 2.22	21.081 ± 1.69
300	33.963 ± 3.23	48.15 ± 2.64	66.988 ± 3.04	60.499 ± 3.15	32.098 ± 2.37
400	44.44 ± 2.52	57.175 ± 5.46	73.813 ± 3.98	67.579 ± 4.05	38.885 ± 3.41
500	52.527 ± 4.13	64.953 ± 3.96	80.087 ± 3.96	75.457 ± 3.27	46.261 ± 4.74
600	60.759 ± 5.91	72.121 ± 5.16	85 ± 3.15	81.666 ± 4.65	58.637 ± 4.79
700	62.425 ± 3.9	72.122 ± 6.11	86.817 ± 5.35	82.424 ± 5.85	59.245 ± 4.37
800	63.335 ± 4.27	72.88 ± 5.41	85.738 ± 5.35	82.727 ± 5.06	61.818 ± 4.99
900	73.942 ± 5.38	82.729 ± 5.35	91.213 ± 5.43	89.698 ± 5.57	73.336 ± 4.91
1000	76.972 ± 4.33	84.244 ± 6.67	90.607 ± 4.62	89.698 ± 3.83	77.275 ± 6.1

จากผลการทดลองจะเห็นได้ว่าวิธีการนี้ถึงแม้จะให้ผลการจำแนกที่ดีในกรณีที่ข้อความทดสอบมีขนาดยาว แต่กับข้อความทดสอบที่มีขนาดสั้นนั้นจะให้ผลการจำแนกที่ไม่ดี นอกจากนั้นยังพบว่า ความยาวของข้อความฝึกนั้นก็มีผลต่อผลการจำแนก โดยผลการจำแนกจะดีขึ้นเมื่อข้อความฝึกนั้นมีความยาวมากขึ้น

4.3.2. การจำแนกโดยการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟ

ในการทดลองนี้จะทำการทดสอบโดยใช้ทั้งค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟและของค่าลอการิทึมของระยะเวลาระหว่างไดกราฟ ทำการทดลองโดยแบ่งชุดข้อมูลฝึกออกเป็นข้อความที่มีความยาว 100 500 และ 1000 ตัวอักษร ได้ผลความแม่นในการจำแนกผู้ใช้ ดังตารางที่ 4-2

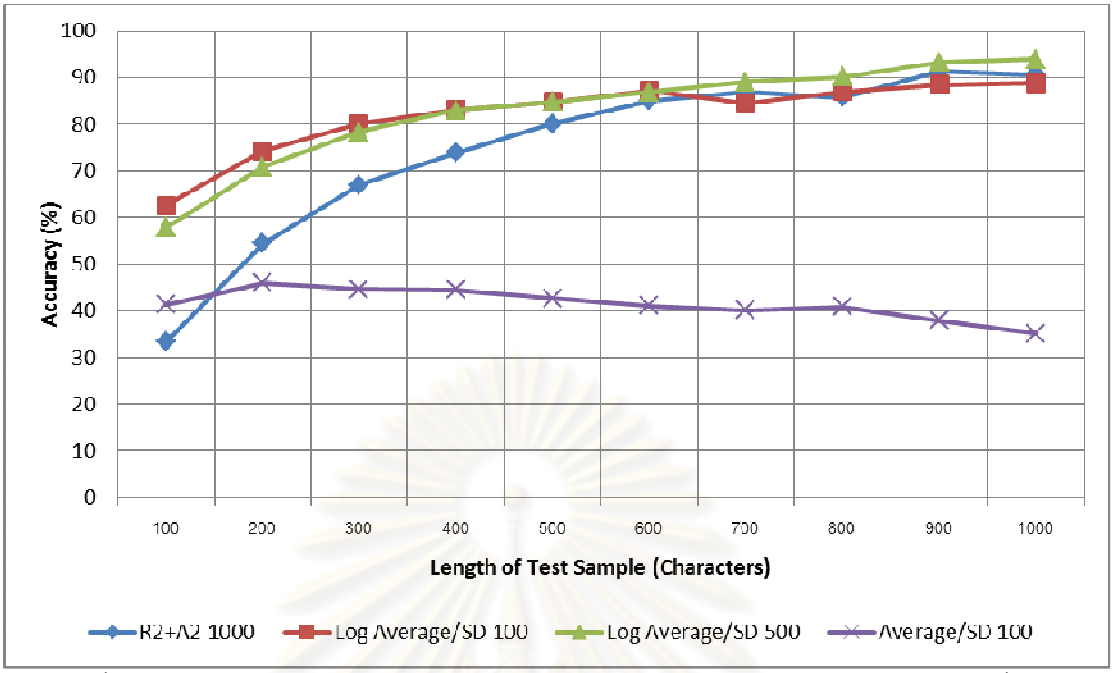
ตารางที่ 4-2 ตารางแสดงความแม่น (ค่าเฉลี่ยร้อยละ ± ส่วนเบี่ยงเบนมาตรฐาน) ในการจำแนกผู้ใช้ เมื่อใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟเป็นคุณลักษณะ

Length of Test Sample (Characters)	Average / SD			Log Average / SD		
	100	500	1000	100	500	1000
100	41.2929 ± 3.01	19.2121 ± 3.29	12.2424 ± 2.81	62.5859 ± 3.46	57.8788 ± 3.88	44.1414 ± 4.3
200	45.9307 ± 3.04	31.6450 ± 2.87	18.7446 ± 3.55	74.2425 ± 4.36	70.8225 ± 4.48	57.6191 ± 4.84
300	44.6667 ± 4.33	39.6364 ± 3.37	25.2727 ± 4.12	80.1212 ± 4.36	78.3636 ± 3.28	65.5758 ± 4.91
400	44.5454 ± 5.76	41.9192 ± 4.32	28.1818 ± 4.75	83.1313 ± 3.37	82.9293 ± 4.3	68.8889 ± 6.46
500	42.7272 ± 6.56	46.6667 ± 6.09	30.303 ± 6.55	84.8485 ± 3.56	84.8485 ± 3.87	72.3233 ± 4.86
600	41.0606 ± 5.69	48.1818 ± 6.77	33.9394 ± 7.73	87.2727 ± 4.05	86.8182 ± 3.92	75.6061 ± 5.07
700	40.1515 ± 9.04	48.3333 ± 5.91	34.9999 ± 7.37	84.5454 ± 4.78	89.0909 ± 3.33	78.1818 ± 5.35
800	40.9091 ± 9.61	51.3636 ± 7.05	35.9091 ± 6.89	86.9697 ± 4.85	90.1515 ± 3.8	80.1515 ± 4.54
900	37.8786 ± 5.93	51.8182 ± 10.54	37.8788 ± 11.54	88.4848 ± 4.91	93.3333 ± 4.47	83.9394 ± 4.96
1000	35.1514 ± 6.88	51.2121 ± 7.88	38.1818 ± 9.92	88.7879 ± 4.75	93.9394 ± 3.5	82.1212 ± 5.24

หากพิจารณาผลความแม่นจะพบว่าการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมของระยะเวลาระหว่างไดกราฟนั้น จะให้ผลดีกว่าการใช้แต่ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานธรรมดา ซึ่งน่าจะเป็นเพราะว่าการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมนั้นสามารถเป็นตัวแทนของข้อมูลทั้งหมดได้ดีกว่าการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานธรรมดาดังที่สันนิษฐานไว้

ในส่วนของความยาวของข้อมูลฝึกนั้นพบว่า การทดลองที่ใช้ข้อมูลฝึกที่มีความยาว 100 ตัวอักษรนั้นให้ผลความแม่นที่ดีที่สุด โดยเฉพาะกับการทดสอบข้อความที่มีความยาวน้อย ๆ ซึ่งน่าจะเป็นเพราะ การหาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของข้อมูลนั้นเป็นการตัดทอนตัวข้อมูลลง จากข้อมูลระยะเวลาทั้งหมดเหลือเพียงค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานเท่านั้น หากแบ่งข้อมูลฝึกออกมามีความยาวมากเกินไป ข้อมูลของระยะเวลาก็อาจถูกตัดทอนออกไปมาก ทำให้ได้ผลการทดสอบที่ไม่ดีนัก

หากเปรียบผลความแม่นของวิธีนี้กับวิธีของ D. Gunetti และ C. Picardi จะพบว่าในการทดลองที่ให้ผลความแม่นดีที่สุดของวิธีนี้ (Log Average/SD 100) นั้นให้ผลสูงกว่าวิธีของ D. Gunetti และ C. Picardi มากในกรณีที่เปรียบเทียบข้อความทดสอบที่มีขนาดสั้น และให้ผลที่ใกล้เคียงกันเมื่อเปรียบเทียบข้อความทดสอบที่มีขนาดยาวขึ้น ดังที่แสดงในภาพที่ 4-1



ภาพที่ 4-1 ภาพแสดงการเปรียบเทียบผลความแม่น (ร้อยละ) ระหว่างวิธีการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน กับ วิธีของ D. Gunetti และ C. Picardi

4.3.3. การจำแนกโดยการใช้ค่าเฉลี่ยความน่าจะเป็น

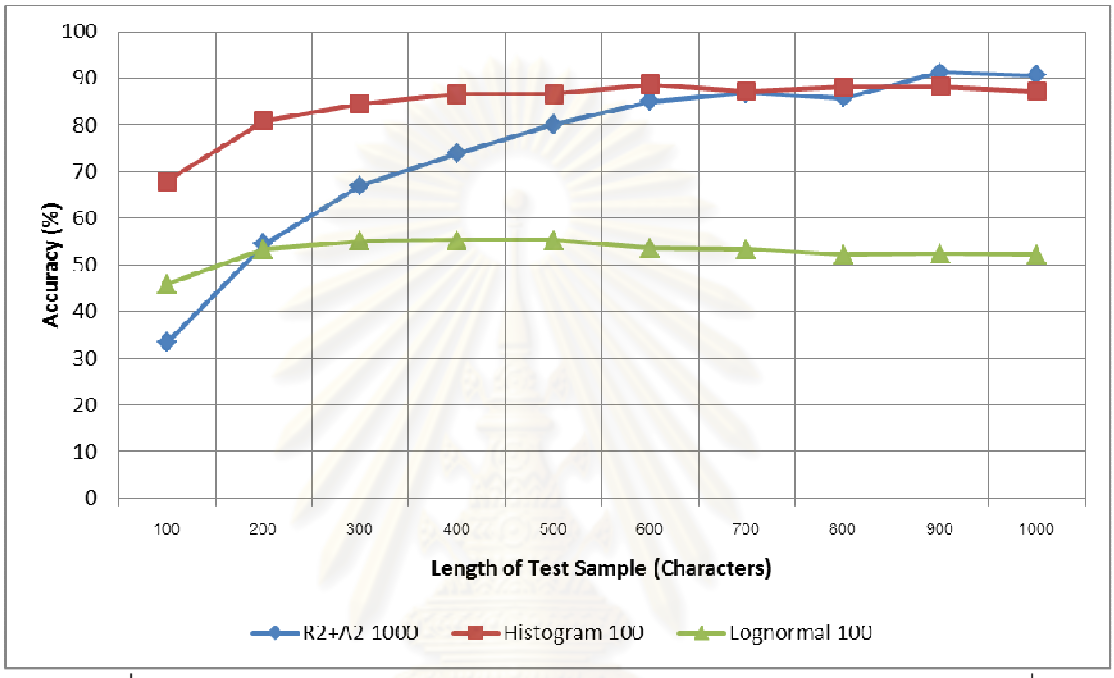
ในการทดลองนี้จะทำการทดสอบโดยใช้ค่าเฉลี่ยความน่าจะเป็นซึ่งคำนวณจากการสร้างโพรไฟล์ความน่าจะเป็นทั้งสองวิธี คือ การประมาณค่าพารามิเตอร์ของการแจกแจงแบบล็อกนอร์มัล และการสร้างฮิสโตแกรม ทำการทดลองโดยแบ่งชุดข้อมูลฝึกออกเป็นข้อความที่มีความยาว 100 500 และ 1000 ตัวอักษร ได้ผลความแม่นในการจำแนกผู้ใช้ ดังตารางที่ 4-3

ตารางที่ 4-3 ตารางแสดงความแม่น (ค่าเฉลี่ยร้อยละ ± ส่วนเบี่ยงเบนมาตรฐาน) ในการจำแนกผู้ใช้ เมื่อใช้ค่าเฉลี่ยความน่าจะเป็นเป็นคุณลักษณะ

Length of Test Sample (Characters)	Lognormal			Histogram		
	100	500	1000	100	500	1000
100	45.7778 ± 3.67	12.9697 ± 2.67	3.1515 ± 0.14	67.899 ± 2.4	34.202 ± 3.95	18.0404 ± 3.44
200	53.3334 ± 3.77	21.8182 ± 4.58	3.2468 ± 0.31	80.8658 ± 2.45	61.4719 ± 5.2	40.2597 ± 5.11
300	55.1515 ± 4.44	26.9697 ± 6.01	3.3333 ± 0.43	84.5455 ± 2.77	73.3939 ± 5.05	53.2121 ± 4.56
400	55.3536 ± 4.15	29.798 ± 6.5	3.5354 ± 0.71	86.5657 ± 4.72	78.2829 ± 6.5	63.4344 ± 6.83
500	55.3536 ± 4.12	32.0202 ± 5.87	3.5354 ± 0.71	86.6667 ± 3.08	84.8485 ± 3.5	70.101 ± 6.5
600	53.6364 ± 3.59	32.7273 ± 6.03	3.6364 ± 1.06	88.6364 ± 3.21	86.3636 ± 2.86	71.9697 ± 7.6
700	53.3333 ± 4.94	34.0909 ± 6.24	3.6364 ± 1.06	87.2727 ± 5.01	88.9394 ± 3.78	76.3636 ± 7.11
800	52.1212 ± 3.73	33.0303 ± 6.69	3.4849 ± 1.02	88.0303 ± 3.53	90.1515 ± 3.8	79.697 ± 6.48
900	52.4243 ± 4.05	31.5151 ± 8.59	3.3333 ± 0.96	88.1818 ± 5.04	92.4243 ± 5.93	84.8485 ± 7.42
1000	52.1212 ± 4.69	33.6363 ± 9.2	3.6364 ± 1.28	87.2727 ± 4.24	93.9394 ± 5.53	84.5455 ± 7.06

หากพิจารณาผลความแม่นจะพบว่าการใช้ค่าเฉลี่ยความน่าจะเป็นที่โพรไฟล์ความน่าจะเป็นสร้างมาจากฮิสโตแกรมนั้น จะให้ผลดีกว่าการใช้แต่ค่าเฉลี่ยความน่าจะเป็นที่โพรไฟล์ความน่าจะเป็นสร้างมากจากการประมาณค่าพารามิเตอร์ของการแจกแจงแบบล็อกนอร์มัล ซึ่งน่าจะเกิดจากความซับซ้อนของข้อมูลประกอบกับเมื่อแบ่งแยกข้อมูลระยะเวลาระหว่างไดกราฟตามประเภทของผู้ใช้แต่ละคนแล้ว อาจจะเหลือจำนวนข้อมูลน้อยเกินไปจนไม่อาจนำมาประมาณค่าพารามิเตอร์ของการแจกแจงความน่าจะเป็นเพื่อใช้เป็นตัวแทนของการกระจายตัวของข้อมูลที่แท้จริงได้อย่างมีประสิทธิภาพ การใช้ฮิสโตแกรมในการประมาณจึงมีความรัดกุมมากกว่า กล่าวคือ การประมาณค่าพารามิเตอร์ของการแจกแจงแบบล็อกนอร์มัลด้วยข้อมูลจำนวนน้อย ๆ เมื่อประมาณแล้วจะได้รูปร่างของการแจกแจงที่ค่อนข้างกระจาย ทำให้เมื่อนำข้อมูลระยะเวลามาทดสอบ ก็จะได้ค่าความน่าจะเป็นสูงหรือปานกลางถึงแม้ว่าข้อมูลที่นำมาทดสอบนั้นจะไม่ตรงกับข้อมูลจริงในชุดข้อมูลฝึกเลยก็ตาม แต่การใช้ฮิสโตแกรมนั้น เราจะได้รูปร่างของการแจกแจงที่มีลักษณะเป็นยอดแคบ ๆ หลาย ๆ ยอด เมื่อนำข้อมูลมาทดสอบ หากข้อมูลที่เข้ามานั้นไม่ตรงกับข้อมูลจริงในชุดข้อมูลฝึก ก็จะได้ค่าความน่าจะเป็นที่ต่ำมากหรือเป็น 0 แต่ถ้าข้อมูลที่เข้ามานั้นตรงกับข้อมูลบางข้อมูลในชุดข้อมูลฝึก ก็จะได้ค่าความน่าจะเป็นที่สูง ซึ่งน่าจะเป็นสาเหตุที่ทำให้การสร้างโพรไฟล์ความน่าจะเป็นจากฮิสโตแกรมนั้นให้ผลดีกว่าการสร้างโพรไฟล์ความน่าจะเป็นจากการประมาณค่าพารามิเตอร์ของการแจกแจงแบบล็อกนอร์มัล

หากเปรียบผลความแม่นของวิธีนี้กับวิธีของ D. Gunetti และ C. Picardi จะพบว่าในการทดลองที่ให้ผลความแม่นดีที่สุดของวิธีนี้ (Histogram 100) นั้นให้ผลสูงกว่าวิธีของ D. Gunetti และ C. Picardi มากในกรณีที่เปรียบเทียบข้อความทดสอบที่มีขนาดสั้น และให้ผลที่ใกล้เคียงกันเมื่อเปรียบเทียบข้อความทดสอบที่มีขนาดยาวขึ้น ดังที่แสดงในภาพที่ 4-2



ภาพที่ 4-2 ภาพแสดงการเปรียบเทียบผลความแม่น (ร้อยละ) ระหว่างวิธีการใช้ค่าเฉลี่ยความน่าจะเป็น กับ วิธีของ D. Gunetti และ C. Picardi

ในแง่ของค่า False Positive และ False Negative หากเปรียบเทียบค่าเฉลี่ยของค่าทั้งสองจากผู้ใช้ทุกๆคนแล้วจะพบว่า ค่าเฉลี่ยของค่า False Positive และ False Negative จากผู้ใช้แต่ละคนของการจำแนกด้วยวิธีการที่นำเสนอนั้น จะมีค่าต่ำกว่าการจำแนกด้วยวิธีการของ D. Gunetti และ C. Picardi โดยการจำแนกโดยใช้ค่าเฉลี่ยความน่าจะเป็นแบบ Histogram จะให้ค่า False Positive และ False Negative เฉลี่ยที่ต่ำที่สุด ดังสรุปได้ในตารางที่ 4-4 (สำหรับตารางแสดงผลค่า False Positive และ False Negative แยกสำหรับแต่ละผู้ใช้นั้นจะแสดงในภาคผนวก)

ตารางที่ 4-4 ตารางเปรียบเทียบค่า False Positive และ False Negative เฉลี่ยทุกผู้ใช้ (ร้อยละ ± ส่วนเบี่ยงเบนมาตรฐาน) ในการจำแนกผู้ใช้ด้วยวิธีต่างๆ

Length of Test Sample (Characters)	False Positive			False Negative		
	R2 + A2	Log Average/SD	Histogram	R2 + A2	Log Average/SD	Histogram
100	2.0439 ± 2.77	1.1338 ± 0.46	0.9728 ± 0.46	2.0439 ± 0.54	1.1338 ± 0.48	0.9728 ± 0.39
200	1.3734 ± 1.75	0.7805 ± 0.44	0.5798 ± 0.32	1.3734 ± 0.59	0.7805 ± 0.43	0.5798 ± 0.36
300	1.0016 ± 1.35	0.6024 ± 0.42	0.4683 ± 0.35	1.0016 ± 0.61	0.6024 ± 0.38	0.4683 ± 0.35
400	0.7775 ± 1.06	0.5112 ± 0.37	0.4071 ± 0.43	0.7775 ± 0.63	0.5112 ± 0.36	0.4071 ± 0.34
500	0.6004 ± 0.86	0.4591 ± 0.43	0.404 ± 0.41	0.6004 ± 0.53	0.4591 ± 0.35	0.404 ± 0.35
600	0.4454 ± 0.73	0.3857 ± 0.38	0.3443 ± 0.43	0.4454 ± 0.59	0.3857 ± 0.32	0.3443 ± 0.34
700	0.404 ± 0.65	0.4683 ± 0.53	0.3857 ± 0.5	0.404 ± 0.58	0.4683 ± 0.39	0.3857 ± 0.36
800	0.4408 ± 0.66	0.3949 ± 0.33	0.3627 ± 0.43	0.4408 ± 0.55	0.3949 ± 0.33	0.3627 ± 0.37
900	0.2847 ± 0.52	0.3489 ± 0.44	0.3581 ± 0.49	0.2847 ± 0.57	0.3489 ± 0.37	0.3581 ± 0.4
1000	0.2755 ± 0.52	0.3398 ± 0.45	0.3857 ± 0.5	0.2755 ± 0.55	0.3398 ± 0.36	0.3857 ± 0.43

4.3.4. การผสานคุณลักษณะ

ในการทดลองนี้จะทำการทดสอบการผสานคุณลักษณะทั้งสองแบบ คือ การผสานอินพุตและการผสานผลลัพธ์ โดยจะนำคุณลักษณะในการทดลองที่ให้ผลความแม่นในการจำแนกสูงสุดของแต่ละแบบมาเป็นคุณลักษณะในการผสาน ซึ่งก็คือ “Log Average/SD 100” และ “Histogram 100” โดยผลการทดลองการผสานคุณลักษณะ แสดงอยู่ในตารางที่ 4-5

ตารางที่ 4-5 ตารางแสดงความแม่น (ค่าเฉลี่ยร้อยละ ± ส่วนเบี่ยงเบนมาตรฐาน) ในการทดลองการผสานคุณลักษณะ

Length of Test Sample (Characters)	Input Combination	Result Combination
100	62.2828 ± 3.51	73.8384 ± 3.32
200	74.9351 ± 3.93	85.4546 ± 1.84
300	80.9091 ± 3.85	89.4545 ± 2.12
400	83.1313 ± 4.04	91.7172 ± 2.47
500	85.5556 ± 3.37	92.4243 ± 2.49
600	88.0303 ± 5.32	93.4848 ± 4.17
700	86.9697 ± 4.75	93.3333 ± 2.49
800	88.0303 ± 3.94	94.2424 ± 2.12
900	89.697 ± 4.33	92.7273 ± 4.78
1000	89.394 ± 3.57	94.2424 ± 4.15

จะเห็นได้ว่า จากผลการทดลอง การผสานผลลัพธ์ให้ผลที่ดีกว่าการผสานอินพุต และหากเปรียบเทียบกับกับวิธีการอื่น ๆ แล้ว การผสานผลลัพธ์จะให้ผลความแม่นที่ดีที่สุด และหากเทียบกับวิธีของ D. Gunetti และ C. Picardi จะพบว่า การผสานผลลัพธ์นั้นให้ผลสูงกว่าวิธีของ D. Gunetti และ C. Picardi มากในกรณีที่เปรียบเทียบข้อความทดสอบที่มีขนาดสั้น และให้ผลที่ใกล้เคียงกันเมื่อเปรียบเทียบข้อความทดสอบที่มีขนาดยาวขึ้น ดังที่แสดงในภาพที่ 4-3

ภาพที่ 4-3 ภาพแสดงการเปรียบเทียบผลความแม่น (ร้อยละ) ของทุกวิธี

4.4. วิเคราะห์ผลการทดลอง

4.4.1. ข้อมูลที่ใช้ในการทดลอง

ในหัวข้อนี้จะกล่าวถึงลักษณะข้อมูลที่ใช้ในการทดลอง วิธีการเก็บข้อมูล การกระจายตัวของระยะเวลาระหว่างไดกราฟของข้อมูลทั้งหมด และการกระจายตัวของจำนวนไดกราฟของผู้ใช้แต่ละคน

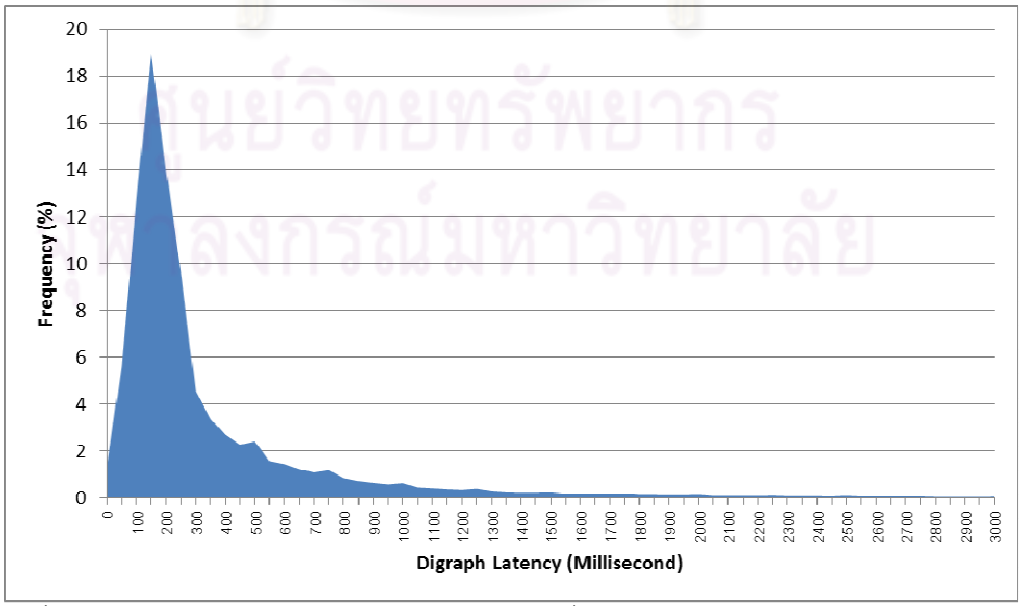
ในการทดลอง เราต้องการใช้ข้อมูลระยะเวลาในการพิมพ์ที่ได้จากการใช้งานปกติของผู้ใช้ จึงได้พัฒนาโปรแกรมเพื่อเก็บข้อมูลเหล่านั้น เมื่ออาสาสมัครรันโปรแกรมนั้น โปรแกรมจะทำงานโดยเก็บข้อมูลการพิมพ์ของผู้ใช้ แยกเป็นข้อมูลของปุ่มที่ถูกกด เวลาที่ถูกกด ลักษณะ (กด-

ปล่อย) และภาษาที่ใช้ (ไทย-อังกฤษ) อาสาสมัครจะได้รับคำแนะนำให้รันโปรแกรมไว้เป็นระยะเวลาหนึ่ง (ประมาณ 1-2 อาทิตย์)

เมื่อได้ข้อมูลของอาสาสมัครแล้วจะนำข้อมูลเป็นสร้างเป็นชุดข้อมูล โดยทำการเลือกจากข้อมูลที่เก็บได้จากผู้ใช้แต่ละคนเป็นจำนวนตัวอักษรที่เท่า ๆ กัน ซึ่งในการทดลองนี้ใช้ 15,000 ตัวอักษร โดยตัดจาก 7,500 ตัวอักษรแรกและอีก 7,500 ตัวอักษรจากข้างท้าย เพื่อให้ครอบคลุมข้อมูลที่ได้จากระยะเวลาที่ห่างกันพอสมควร ข้อมูลที่ถูกตัดแล้วจะถูกนำมาแบ่งเป็นความยาวต่าง ๆ อีกครั้ง เพื่อเป็นตัวอย่างฝึกและตัวอย่างทดสอบในการทดลองต่อไป

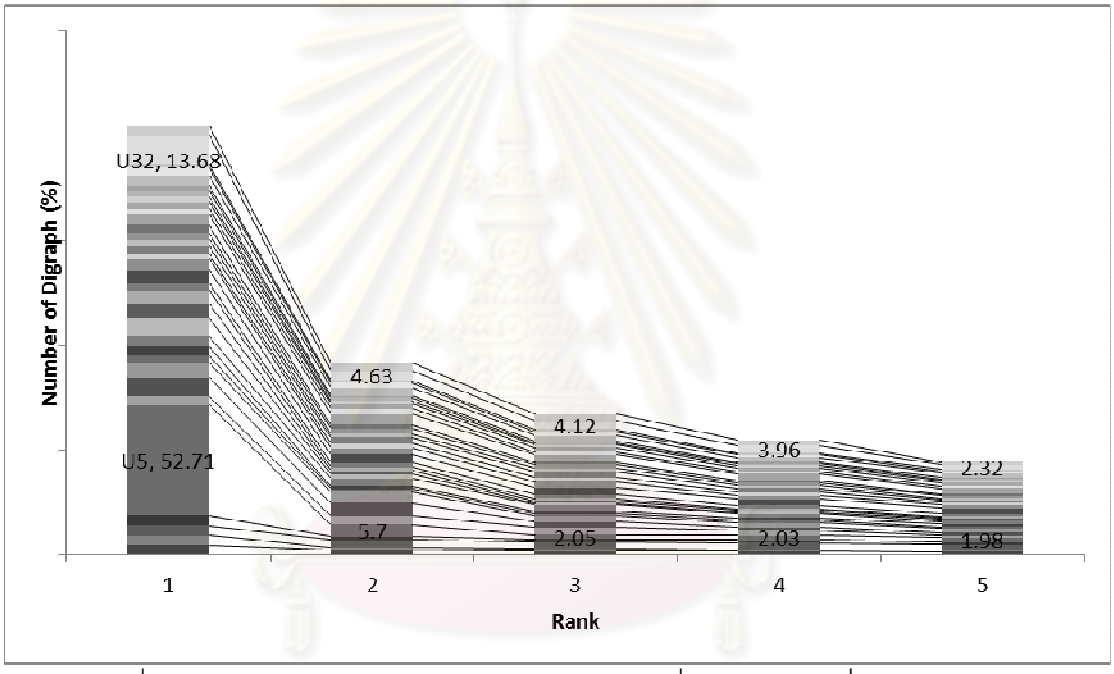
เพื่อลดข้อมูลที่มีค่าผิดไปจากปกติมาก ๆ เช่น ผู้ใช้หยุดพิมพ์เป็นเวลานานเกินไปกว่าจะพิมพ์ตัวอักษรต่อไป หรือ เป็นข้อมูลการใช้งานในช่วงเวลาที่แตกต่างกัน เป็นต้น เนื่องจากข้อมูลเหล่านั้นไม่ใช่ข้อมูลที่ควรนำมาพิจารณา เพราะอาจทำให้เกิดผลที่ผิดพลาดได้ ดังนั้นระบบจะคำนึงเฉพาะค่าระยะเวลาที่ไม่เกิน 1500 มิลลิวินาทีเท่านั้น

หากพิจารณาการกระจายตัวของระยะเวลาระหว่างไดกราฟของผู้ใช้ทุก ๆ คน จะเป็นดังภาพที่ 4-4 ซึ่งจะพบว่า การกระจายตัวของระยะเวลาระหว่างไดกราฟนั้น จะกระจุกตัวอยู่ในช่วง 100-400 มิลลิวินาที และจะมีไดกราฟที่มีระยะเวลามากกว่านั้นเป็นจำนวนน้อยลงเรื่อย ๆ ดังนั้น เมื่อพิจารณาความเป็นจริงในการพิมพ์ของคนทั่วไป กับการกระจายตัวของไดกราฟที่มีระยะเวลาระหว่างไดกราฟเป็นค่าต่าง ๆ การเลือกพิจารณาเฉพาะค่าระยะเวลาระหว่างไดกราฟที่ไม่เกิน 1500 มิลลิวินาทีนั้นเป็นค่าที่เหมาะสมแล้ว และการเลือกพิจารณาเช่นนี้จะทำให้ข้อมูลหายไปประมาณร้อยละ 9.382



ภาพที่ 4-4 ภาพการกระจายตัวของจำนวนไดกราฟที่มีระยะเวลาระหว่างไดกราฟเป็นค่าต่าง ๆ

หากจะคาดเดาพฤติกรรมการใช้งานของผู้ใช้ อาจทำได้โดยดูการกระจายตัวของจำนวนไดกราฟของประเภทที่มีจำนวนอยู่มากเป็นอันดับต้นของผู้ใช้แต่ละคน จำนวนไดกราฟของประเภทที่มีจำนวนสูงสุด 5 อันดับแรกของผู้ใช้แต่ละคนแสดงอยู่ในภาพที่ 4-5 จะเห็นได้ว่ามีผู้ใช้บางคน (U5, U32) ที่มีลักษณะของการกระจายตัวที่แปลกกว่าของผู้ใช้คนอื่น ซึ่งอาจเกิดจากลักษณะการใช้งานที่แตกต่างจากผู้ใช้คนอื่น เพราะมีการพิมพ์คู่ไดกราฟบางคู่มากกว่าปกติ เนื่องจากการทดลองนี้พิจารณาการใช้งานปกติของผู้ใช้แต่ละคน และข้อมูลที่มีลักษณะเช่นนั้นมีจำนวนไม่มาก จึงตัดสินใจใช้ข้อมูลทั้งหมดของผู้ใช้ทุก ๆ คนในการทดลอง



ภาพที่ 4-5 ภาพแสดงสัดส่วนของไดกราฟ (ร้อยละ) ที่มีจำนวนมากที่สุด 5 อันดับแรกของผู้ใช้แต่ละคน

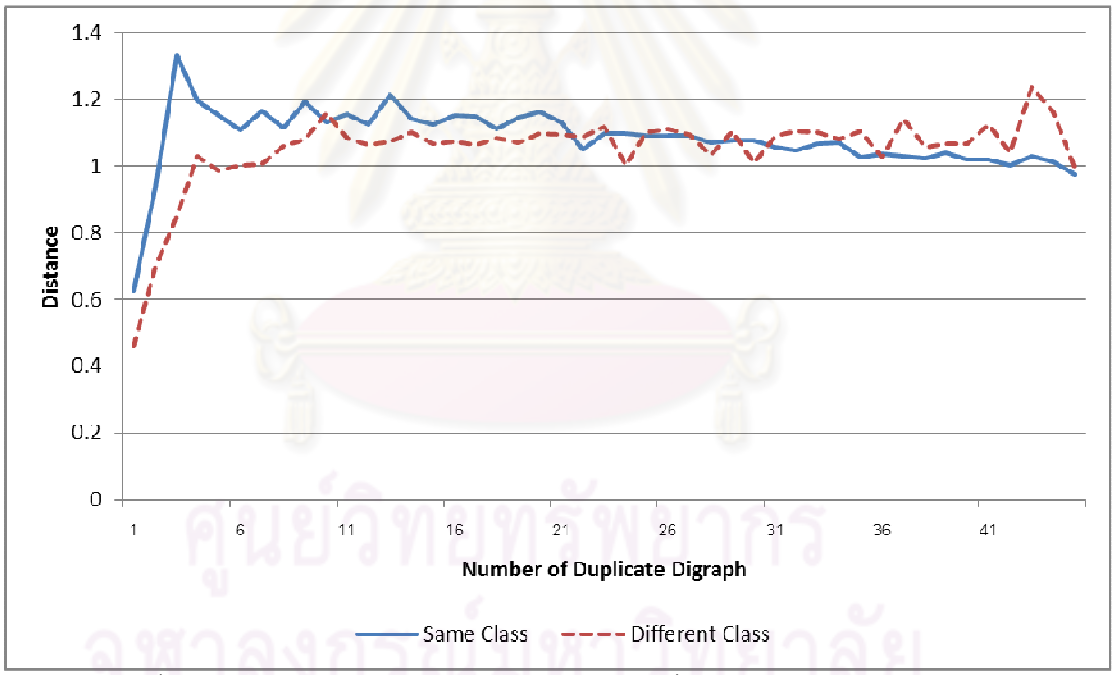
4.4.2. วิเคราะห์ผลของจำนวนไดกราฟที่ซ้ำกันที่มีผลต่อความแม่นเมื่อทดสอบด้วยวิธีการของ D. Gunetti และ C. Picardi

ดังที่กล่าวไว้ข้างต้น วิธีการของ D. Gunetti และ C. Picardi นั้นถือว่าเป็นงานวิจัยที่รายงานผลลัพธ์การจำแนกผู้ใช้ด้วยข้อความอิสระที่ให้ผลดีที่สุดในบรรดางานวิจัยที่มีมา แต่ก็มีจุดอ่อนอยู่ที่ไม่สามารถจำแนกได้ดีในกรณีที่ข้อความทดสอบมีขนาดสั้น ในหัวข้อนี้จะทำการศึกษาเพื่อพิจารณาการคำนวณระยะห่างของข้อความโดยละเอียด โดยพิจารณาในแง่มุมของจำนวน

ไดกราฟที่ซ้ำกันเวลาเปรียบเทียบข้อความทั้งสอง และความถี่ของจำนวนครั้งที่เปรียบเทียบที่มีจำนวนไดกราฟที่ซ้ำกันเป็นค่าต่าง ๆ เพื่อดูผลกระทบต่อผลความแม่นในการจำแนกผู้ใช้

ในกรณีนี้ จะพิจารณาการทดสอบเพียง 1 Fold สำหรับการจำแนกผู้ใช้ด้วยวิธีการของ D. Gunetti และ C. Picardi ซึ่งใช้วิธีการวัดระยะแบบ R_2+A_2 โดยข้อความฝึกมีความยาว 1000 ตัวอักษร และข้อความทดสอบมีความยาว 100 ตัวอักษร โดยพิจารณาเฉพาะกับข้อความทดสอบที่ระบบจำแนกผิด และพิจารณาการคำนวณระยะห่างจากข้อความทดสอบไปยังแต่ละข้อความในชุดข้อมูลฝึกของผู้ใช้ที่เป็นเจ้าของที่แท้จริง (มาจากผู้ใช้คนเดียวกันกับข้อความทดสอบ) และของผู้ใช้ที่ระบบทำนายไปผิด (ผู้ใช้คนละคนกับข้อความทดสอบ) เท่านั้น

อันดับแรกพิจารณาค่าเฉลี่ยระยะ (R_2+A_2) โดยเทียบกับจำนวนไดกราฟที่ซ้ำกันระหว่างการเปรียบกับกับข้อมูลฝึกที่เป็นผู้ใช้คนเดียวกัน (Same class) กับข้อมูลทดสอบ และกับข้อมูลฝึกที่เป็นผู้ใช้คนละคนกัน (Different class) ซึ่งได้ผลการทดสอบดังภาพที่ 4-6

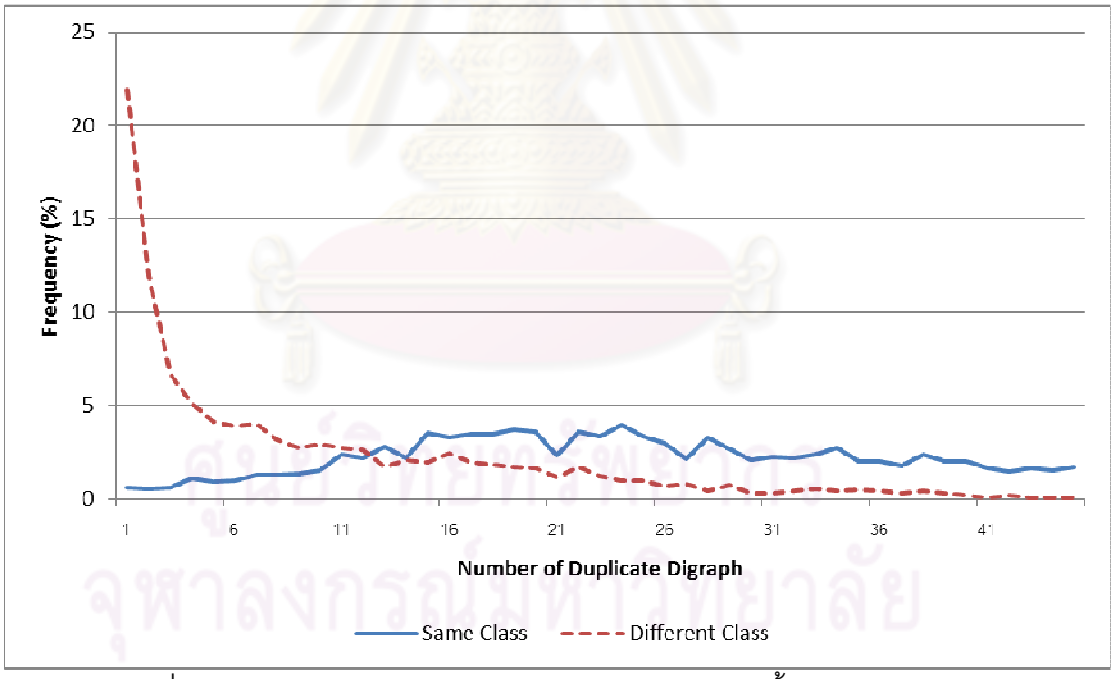


ภาพที่ 4-6 ภาพแสดงการเปรียบเทียบระยะห่างเฉลี่ยของการเปรียบเทียบตัวอย่างเมื่อจำนวนไดกราฟที่ซ้ำกันมีค่าแตกต่างกัน

ในการทดสอบนั้น หากค่าระยะห่างมีค่ามาก หมายความว่าตัวอย่างทั้งสองมีความแตกต่างกันมาก ซึ่งหากเทียบข้อความของผู้ใช้คนละคนกัน ก็ควรจะมีระยะห่างที่มากกว่าการเปรียบเทียบข้อความของผู้ใช้คนเดียวกัน แต่จะเห็นว่าที่ระดับจำนวนไดกราฟที่ซ้ำกันน้อย ๆ นั้นผลระยะห่างเฉลี่ยของการเปรียบเทียบกับตัวอย่างของผู้ใช้คนละคนกันนั้นกลับมีค่าน้อยกว่า

โดยค่าเฉลี่ยระยะห่างของการเปรียบเทียบข้อความของผู้ใช้คนละคนกันนั้น กลับมีค่ามากกว่าค่าเฉลี่ยระยะห่างของการเปรียบเทียบข้อความของผู้ใช้คนเดียวกัน เมื่อจำนวนไดกราฟที่ซ้ำกันมีค่ามากขึ้น (ประมาณ 30 ขึ้นไป) อย่างไรก็ดี ถึงแม้ว่าในช่วงที่จำนวนไดกราฟที่ซ้ำกันมีค่าน้อยนั้นค่าความแตกต่างของระยะห่างเฉลี่ยของทั้งสองแบบก็มีค่าต่างกันไม่มาก ดังนั้นจึงพิจารณาจำนวนครั้งการเปรียบเทียบ โดยมีจำนวนไดกราฟที่ซ้ำกันเป็นค่าต่าง ๆ ร่วมด้วย เพราะในการจำแนกผู้ใช้ด้วยวิธีนี้นั้น จะใช้ค่าเฉลี่ยของระยะห่างระหว่างข้อความทดสอบเทียบกับทุก ๆ ข้อความของผู้ใช้แต่ละคนในชุดข้อมูลฝึก จากนั้นจึงพิจารณาจำแนกข้อความทดสอบว่ามาจากผู้ใช้ที่มีค่าเฉลี่ยระยะห่างระหว่างข้อความน้อยที่สุด ดังนั้นจำนวนครั้งของการทดสอบจึงเป็นเหมือนค่าถ่วงน้ำหนักให้กับค่าเฉลี่ยที่ได้ปรากฏดังภาพที่ 4-6 ในการพิจารณาจำแนกข้อความว่ามาจากผู้ใช้คนใด

ทำการหาจำนวนครั้งการเปรียบเทียบซึ่งมีจำนวนไดกราฟที่ซ้ำกันเป็นจำนวนที่แตกต่างกัน โดยพิจารณาเป็นร้อยละของจำนวนการเปรียบเทียบทั้งหมด ได้ผลดังภาพที่ 4-7



ภาพที่ 4-7 ภาพแสดงการเปรียบเทียบร้อยละของจำนวนครั้งในการเปรียบเทียบเมื่อจำนวนไดกราฟที่ซ้ำกันมีค่าแตกต่างกัน

จะพบว่าการเปรียบเทียบข้อความที่มาจากผู้ใช้คนเดียวกันนั้น การกระจายตัวของจำนวนครั้งการเปรียบเทียบจะมีลักษณะคล้ายการแจกแจงแบบปกติ แต่การเปรียบเทียบข้อความที่มาจากผู้ใช้คนละคนกันนั้น จะมีจำนวนครั้งสูงมากในกรณีที่มีจำนวนไดกราฟที่ซ้ำกัน

น้อย ๆ เนื่องจากเมื่อเป็นข้อความที่มาจากผู้ใช้คนละคนกัน โอกาสในการพิมพ์ไดกราฟที่ซ้ำกันจึงน้อยกว่าข้อความที่มาจากผู้ใช้คนเดียวกัน และจากผลการทดลองในภาพที่ 4-6 จะพบว่าในช่วงที่มีจำนวนไดกราฟที่ซ้ำกันน้อย ๆ นั้น ค่าระยะห่างเฉลี่ยของการเปรียบเทียบระหว่างผู้ใช้คนละคนกันมีค่าน้อยกว่าการเปรียบเทียบระหว่างผู้ใช้คนเดียวกัน เมื่อค่าเฉลี่ยเหล่านั้นถูกถ่วงน้ำหนักด้วยร้อยละของจำนวนครั้งของการเปรียบเทียบแล้ว ก็จะทำให้ค่าเฉลี่ยของระยะห่างระหว่างข้อความทดสอบกับผู้ใช้ที่เป็นคนละคนกันนั้นมีค่าน้อยกว่า จึงเป็นสาเหตุให้ระบบจำแนกผู้ใช้ออกมาผิดพลาดและมีความแม่นต่ำนั่นเอง

4.4.3. วิเคราะห์ผลการทดลองเมื่อข้อความทดสอบมีระยะเวลาระหว่างไดกราฟที่เปลี่ยนไป

จากข้อสันนิษฐานที่ว่า ระยะเวลาระหว่างไดกราฟของผู้ใช้นั้นอาจเปลี่ยนแปลงไปตามเวลาหรือด้วยเหตุผลอื่น ๆ แต่การเปลี่ยนแปลงนั้นจะเป็นไปในแนวทางเดียวกับสำหรับไดกราฟทุก ๆ ประเภท ในหัวข้อนี้จะทำการทดสอบเพื่อเปรียบเทียบผลความแม่นในการจำแนกผู้ใช้ในการทดลองด้วยวิธีต่าง ๆ เมื่อระยะเวลาระหว่างไดกราฟมีค่าเปลี่ยนไปในทางเดียวกัน

นิยามข้อมูลระยะเวลาระหว่างไดกราฟ $X_i = X_1, X_2, \ldots, X_n$ คือข้อมูลระยะเวลาระหว่างไดกราฟจริงจากชุดข้อมูลทดสอบ ทำการสร้างชุดข้อมูลทดสอบชุดใหม่ $X'_i = cX_1, cX_2, \ldots, cX_n$ เพื่อเป็นชุดข้อมูลที่ระยะเวลาระหว่างไดกราฟเปลี่ยนแปลงไปในทางเดียวกัน โดยค่าคงที่ $c > 0$ เป็นตัวบ่งบอกระดับความเปลี่ยนแปลง การทดสอบในหัวข้อนี้จะทดสอบเพื่อดูความแม่นในการจำแนกเมื่อ $c = 0.5, 1.5$ และ 2

หากพิจารณาการจำแนกด้วยนิวรอลเน็ตเวิร์กจะพบว่า นิวรอลเน็ตเวิร์กจะทำการจำแนกตัวอย่างหนึ่ง ๆ ได้ผลลัพธ์เหมือนเดิมถ้าข้อมูลอินพุตทุก ๆ ข้อมูลของตัวอย่างนั้นเปลี่ยนแปลงไปในทางเดียวกันและเปลี่ยนแปลงไปไม่มากนัก ในหัวข้อนี้จึงพิจารณาการเปลี่ยนแปลงของค่าอินพุตเมื่อใช้คุณลักษณะแบบต่าง ๆ ร่วมกับผลการทดสอบ

4.4.3.1. ด้วยการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟเป็นคุณลักษณะ

ในการทดลองนี้ สิ่งที่ใช้เป็นคุณลักษณะคือ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของระยะเวลาระหว่างไดกราฟ ซึ่งการเปลี่ยนแปลงของระยะเวลาระหว่างไดกราฟจะส่งผลโดยตรงกับค่าของอินพุต โดยในหัวข้อนี้จะพิจารณาทั้งการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน กับการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึม

ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของข้อมูลสามารถหาได้ ดังสมการที่ 1 และสมการที่ 2 ตามลำดับ

$$\bar{x} = \frac{\sum x}{n} \tag{1}$$

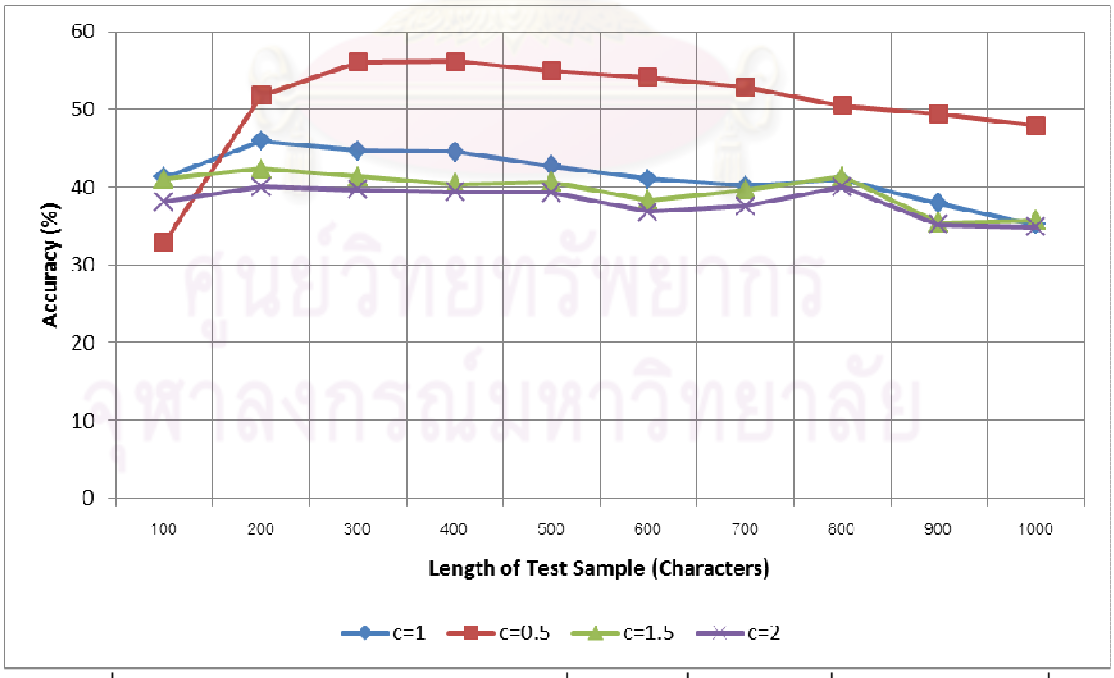
$$SD_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \tag{2}$$

หากพิจารณาชุดข้อมูลใหม่ จะคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานใหม่ได้ดังสมการที่ 3 และสามการที่ 4 ตามลำดับ

$$\bar{x}' = \frac{\sum cx}{n} = c\bar{x} \tag{3}$$

$$SD_x' = \sqrt{\frac{\sum (x' - \bar{x}')^2}{n}} = \sqrt{\frac{\sum (cx - c\bar{x})^2}{n}} = c \times \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = c\, SD_x \tag{4}$$

หรืออาจกล่าวได้ว่าทั้งค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานจะเพิ่มขึ้น c เท่า ตามการเปลี่ยนแปลงของระยะเวลาระหว่างไดกราฟ ซึ่งจากการทดลองจะได้ผลดังภาพที่ 4-8



ภาพที่ 4-8 ภาพผลความแม่นในการจำแนกเมื่อใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานเมื่อระยะเวลาระหว่างไดกราฟมีค่าเปลี่ยนไป

ในการพิจารณาค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมของระยะเวลาระหว่างไดกราฟ จะได้ดัง สมการที่ 5 และ สมการที่ 6 ตามลำดับ

$$\bar{y} = \frac{\Sigma \log(x)}{n} \tag{5}$$

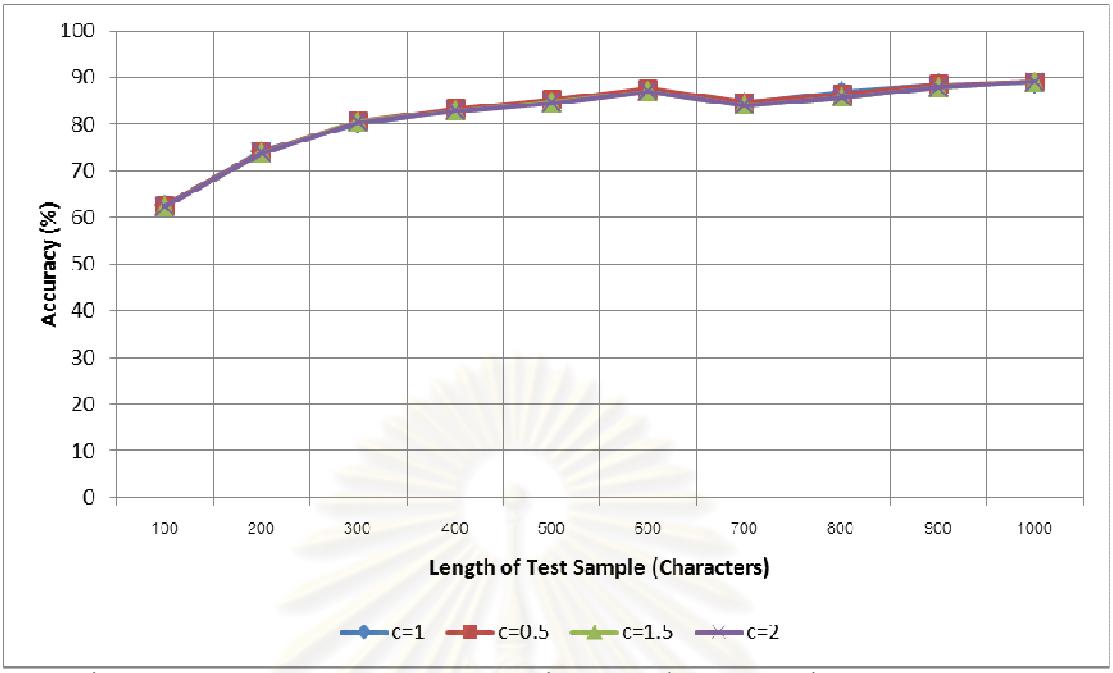
$$SD_y = \sqrt{\frac{\Sigma(\log(x) - \bar{y})^2}{n}} \tag{6}$$

หากพิจารณาชุดข้อมูลใหม่ จะคำนวณค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานใหม่ได้ดังสมการที่ 7 และสมการที่ 8 ตามลำดับ

$$\bar{y}' = \frac{\Sigma \log(cx)}{n} = \frac{\Sigma \log(c) + \log(x)}{n} = \frac{n\log(c) + \Sigma \log(x)}{n} = \log(c) + \bar{y} \tag{7}$$

$$\begin{aligned} SD'_y &= \sqrt{\frac{\Sigma(\log(cx) - \bar{y}')^2}{n}} \\ &= \sqrt{\frac{\Sigma(\log(c) + \log(x) - \log(c) - \bar{y})^2}{n}} \\ &= \sqrt{\frac{\Sigma(\log(x) - \bar{y})^2}{n}} \\ &= SD_y \end{aligned} \tag{8}$$

หรืออาจกล่าวได้ว่าทั้งค่าเฉลี่ยของระยะเวลาระหว่างไดกราฟ จะเพิ่มขึ้นเท่ากับค่าลอการิทึมของค่า c แต่ส่วนเบี่ยงเบนมาตรฐานจะมีค่าเท่าเดิม ซึ่งจากการทดลองจะได้ผลดังภาพที่ 4-9 ซึ่งจะเห็นได้ว่าผลการทดลองแทบจะไม่ต่างกันเลย เพราะค่าอินพุตเปลี่ยนแปลงไปเพียงเล็กน้อยเท่านั้น

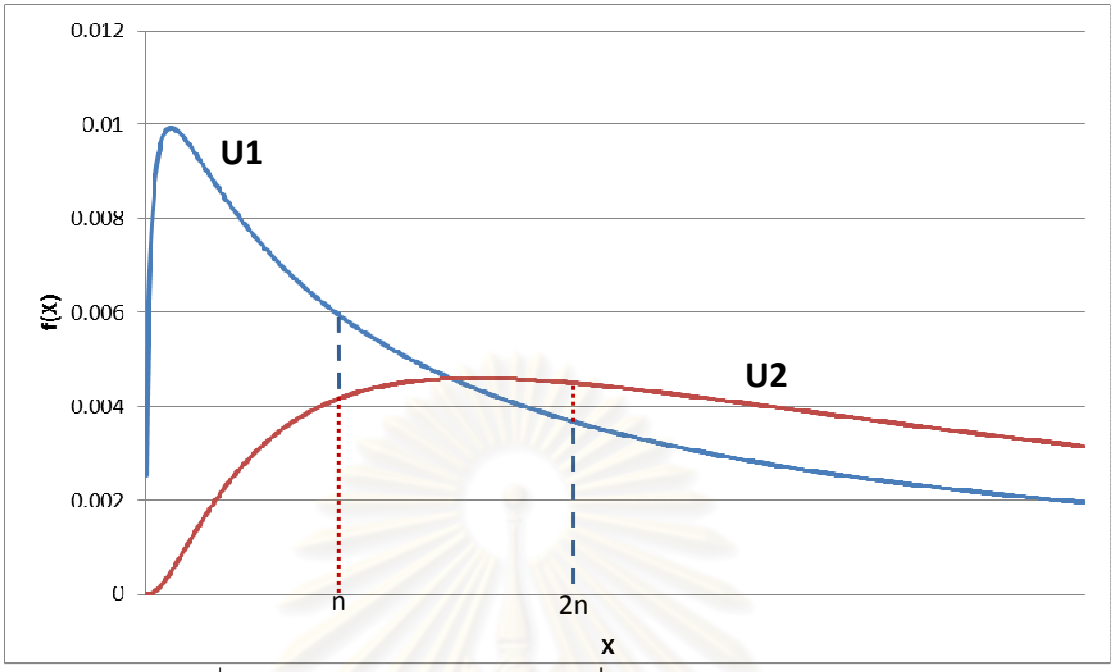


ภาพที่ 4-9 ภาพผลความแม่นในการจำแนกเมื่อใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมเมื่อระยะเวลาระหว่างไดกราฟมีค่าเปลี่ยนไป

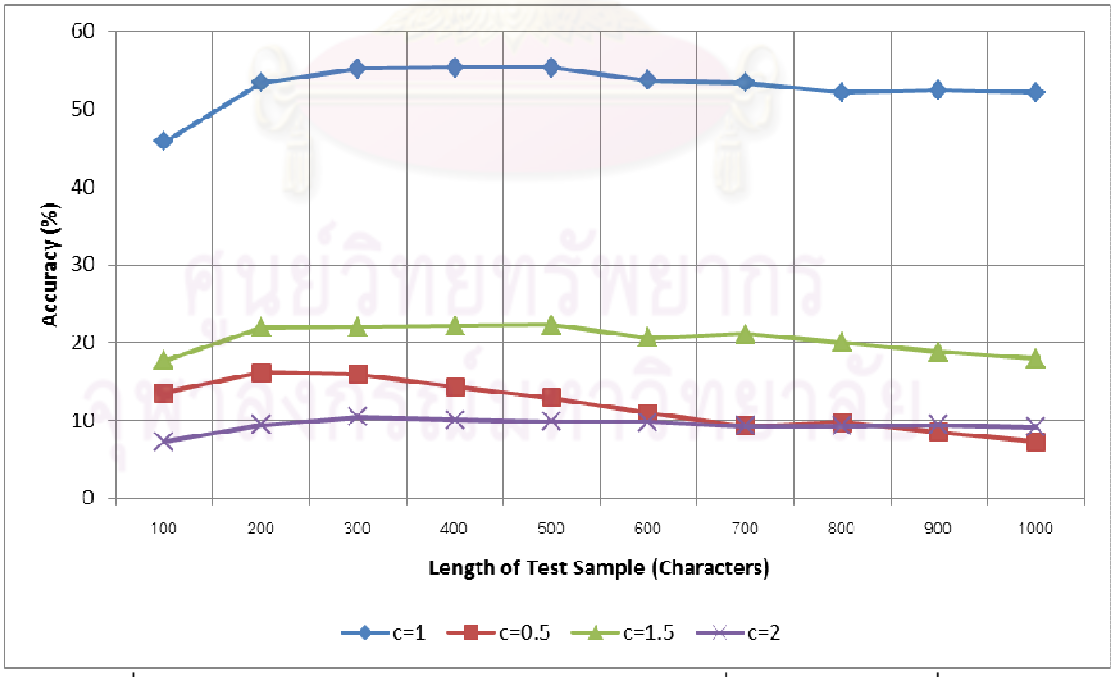
4.4.3.2. การทดสอบด้วยการใช้ค่าเฉลี่ยความน่าจะเป็นเป็นคุณลักษณะ

ในการทดลองนี้ สิ่งที่ใช้เป็นคุณลักษณะคือ ค่าเฉลี่ยความน่าจะเป็นที่ไดกราฟในตัวอย่างทดสอบจะเป็นของผู้ใช้แต่ละคน ซึ่งคุณลักษณะแบบนี้อาจไม่รองรับการเปลี่ยนแปลงของระยะเวลาระหว่างไดกราฟ เพราะค่าความน่าจะเป็นอาจไม่เปลี่ยนแปลงไปในทางเดียวกันสำหรับทุก ๆ อินพุตเมื่อระยะเวลาระหว่างไดกราฟเปลี่ยนแปลงไป ดังเช่นในภาพที่ 4-10 กล่าวคือ เมื่อค่าระยะเวลาระหว่างไดกราฟเป็น n ไดกราฟนี้มีโอกาสจะเป็นของ U1 มากกว่า U2 แต่ถ้าระยะเวลาระหว่างไดกราฟเปลี่ยนไปเป็น 2n กลับทำให้ไดกราฟนี้มีโอกาสจะเป็นของ U2 มากกว่า U1 ซึ่งเหตุการณ์เช่นนี้มีโอกาสที่จะทำให้นิวรอลเน็ตเวิร์กทำนายผลออกมาผิดพลาดได้

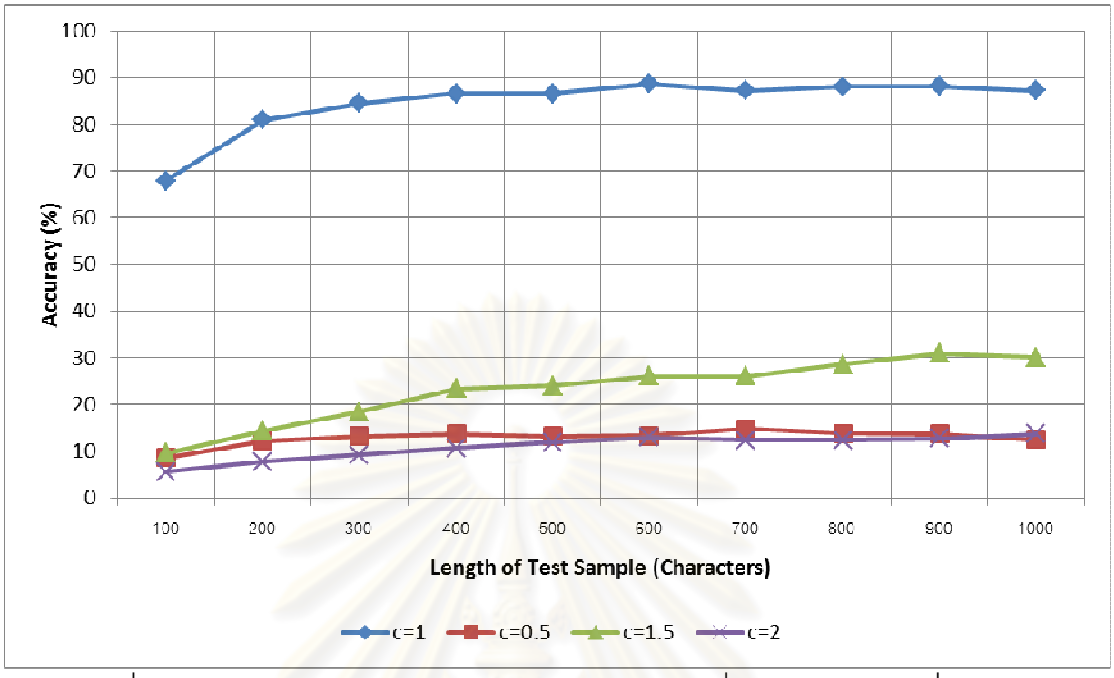
ผลการทดลองการจำแนกผู้ใช้ด้วยคุณลักษณะค่าเฉลี่ยความน่าจะเป็น ทั้งแบบที่ใช้การประมาณค่าพารามิเตอร์ของการแจกแจงแบบล็อกนอร์มัล และการใช้ความถี่สัมพัทธ์ของฮิสโตแกรม ได้ผลดังภาพที่ 4-11 และ ภาพที่ 4-12 ตามลำดับ ซึ่งจะเห็นว่าเมื่อระยะเวลาระหว่างไดกราฟเปลี่ยนไป ผลการทดลองจะมีความแม่นลดลงอย่างมาก



ภาพที่ 4-10 ภาพแสดงตัวอย่างการเปลี่ยนแปลงของค่าความน่าจะเป็น
เมื่อค่าระยะเวลาระหว่างไดกราฟเพิ่มขึ้น 2 เท่า



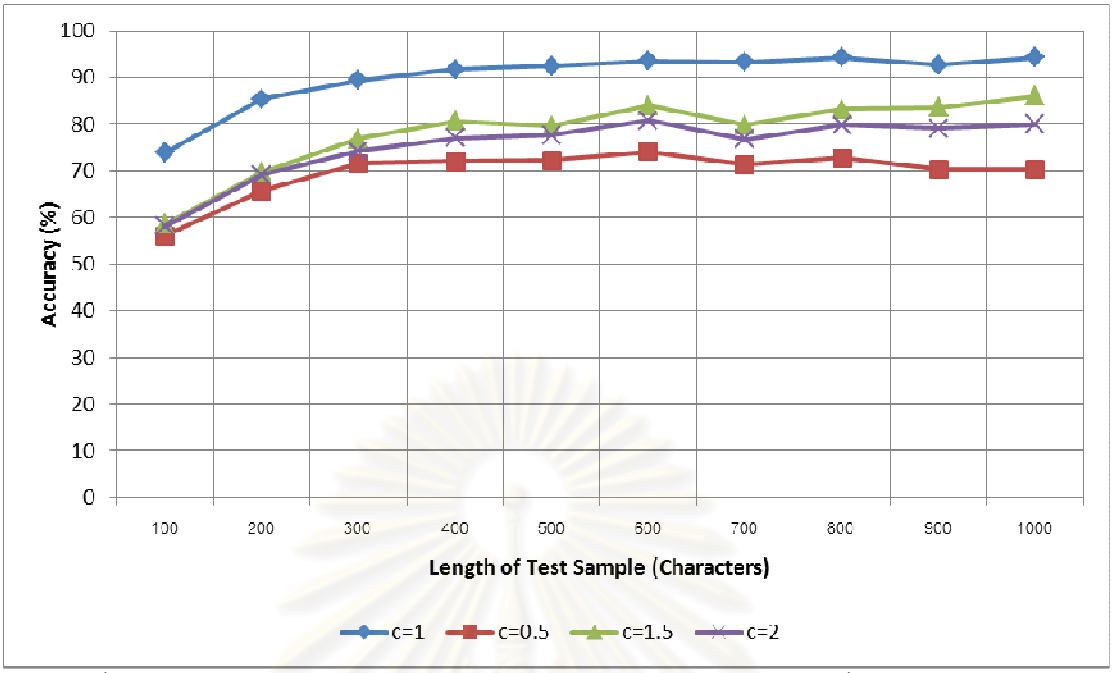
ภาพที่ 4-11 ภาพผลความแม่นในการจำแนกโดยใช้ค่าเฉลี่ยความน่าจะเป็นที่สร้างจาก
การแจกแจงแบบล็อกนอร์มัลเมื่อระยะเวลาระหว่างไดกราฟมีค่าเปลี่ยนไป



ภาพที่ 4-12 ภาพผลความแม่นในการจำแนกโดยใช้ค่าเฉลี่ยความน่าจะเป็นที่สร้างจากฮิสโตแกรมเมื่อระยะเวลาระหว่างไดกราฟมีค่าเปลี่ยนไป

4.4.3.3. การทดสอบการผสานคุณลักษณะ

ในหัวข้อนี้จะทดลองการผสานผลลัพธ์ เมื่อค่าระยะเวลาระหว่างไดกราฟของชุดข้อมูลทดสอบมีค่าเปลี่ยนไป ซึ่งผลจะแสดงดังภาพที่ 4-13 ซึ่งจะเห็นว่า เมื่อผสานผลลัพธ์ที่ได้จากการจำแนกด้วยวิธีทั้งสอง วิธีการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมของระยะเวลาระหว่างไดกราฟ ซึ่งรองรับการเปลี่ยนแปลงของค่าระยะเวลาระหว่างไดกราฟจะช่วยให้ผลการจำแนกไม่ลดลงมากนัก ในขณะที่วิธีใช้ค่าเฉลี่ยความน่าจะเป็นซึ่งไม่รองรับการเปลี่ยนแปลงของระยะเวลาระหว่างไดกราฟ จะทำให้ผลการจำแนกในกรณีปกติ หรือเมื่อระยะเวลาระหว่างไดกราฟเปลี่ยนแปลงไปไม่มาก มีผลดีกว่าการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานเพียงอย่างเดียว

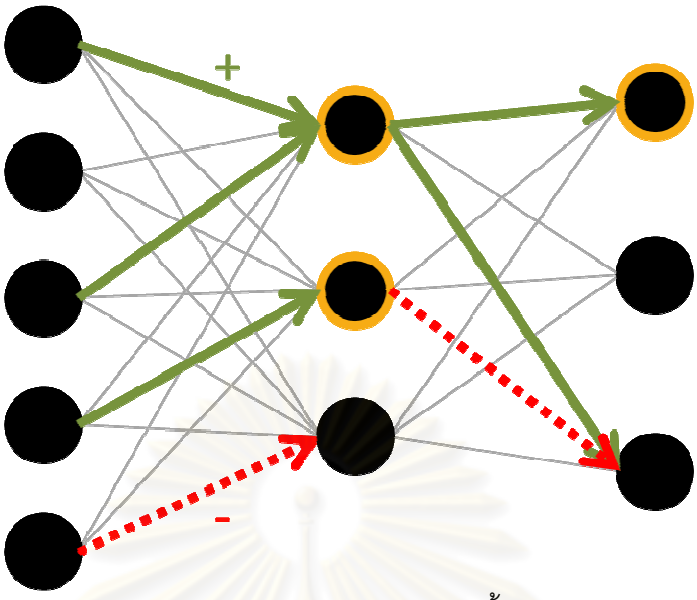


ภาพที่ 4-13 ภาพผลความแม่นในการจำแนกโดยใช้การผสานผลลัพธ์เมื่อระยะเวลาระหว่างไดกราฟมีค่าเปลี่ยนไป

4.4.4. วิเคราะห์ผลค่าน้ำหนักในนิวรอลเน็ตเวิร์ก

ข้อสันนิษฐานอย่างหนึ่งที่น่าจะทำให้การใช้นิวรอลเน็ตเวิร์กและคุณลักษณะแบบต่าง ๆ ให้ผลการจำแนกผู้ใช้ที่ดีกว่าวิธีของ D. Gunetti และ C. Picardi คือ การใช้นิวรอลเน็ตเวิร์กในการจำแนกนั้น นิวรอลเน็ตเวิร์กจะทำการให้ค่าน้ำหนักกับอินพุตต่าง ๆ เพื่อบ่งบอกความสำคัญของอินพุตนั้น ๆ ที่ส่งผลต่อการจำแนกออกเป็นผู้ใช้แต่ละคน ซึ่งวิธีของ D. Gunetti และ C. Picardi ไม่มีการทำงานในลักษณะนี้

จากโครงสร้างของนิวรอลเน็ตเวิร์กที่ประกอบไปด้วยชั้นอินพุต ชั้นแฝง และชั้นเอาต์พุต ในการจำแนกตัวอย่างนั้น โหนดในชั้นอินพุตจะรับค่าของอินพุตทุก ๆ ค่า และส่งต่อไปยังชั้นแฝง โหนดแต่ละโหนดในชั้นแฝงจะมีค่าน้ำหนักที่กำกับอินพุตแต่ละค่า ซึ่งมีทั้งค่าน้ำหนักที่เป็นบวกและเป็นลบ ค่าน้ำหนักที่เป็นบวกจะบ่งบอกว่าอินพุตนั้น ๆ ส่งผลให้โหนดในชั้นแฝงนี้ถูกกระตุ้นมากหรือน้อยเพียงใด เช่นเดียวกัน ค่าน้ำหนักที่เป็นลบก็จะบ่งบอกว่าอินพุตนั้น ๆ ยับยั้งการถูกกระตุ้นของโหนดในชั้นแฝงนี้มากเพียงใด จากนั้นโหนดแต่ละโหนดในชั้นแฝงที่ถูกกระตุ้นก็จะส่งผ่านค่าไปยังโหนดในชั้นเอาต์พุตต่อไป ค่าผลลัพธ์ของโหนดในชั้นเอาต์พุตจะบ่งบอกว่าตัวอย่างตัวนี้ควรจะเป็นของผู้ใช้คนใดโดยพิจารณาโหนดที่มีค่าผลลัพธ์มากที่สุด ผลลัพธ์นี้จะได้มาจากค่าที่โหนดในชั้นแฝงส่งผ่านมา ประกอบกับค่าน้ำหนักที่โหนดในชั้นเอาต์พุตแต่ละโหนดกำกับไว้กับโหนดในชั้นแฝงแต่ละโหนด ในลักษณะเดียวกันกับที่โหนดในชั้นแฝงมีค่าน้ำหนักกำกับอินพุต



ภาพที่ 4-14 ภาพแสดงตัวอย่างการถูกกระตุ้นหรือยับยั้งของโหนดในนิวรอลเน็ตเวิร์ก

จากภาพที่ 4-14 แสดงให้เห็นถึงการถูกกระตุ้นของโหนดในชั้นต่าง ๆ โดยโหนดในชั้นแฝงจะถูกกระตุ้นจากอินพุต โดยอินพุตที่มีค่าน้ำหนักกำกับไว้มากในทางบวกก็จำเพิ่มโอกาสที่โหนดในชั้นแฝงจะถูกกระตุ้นได้ โหนดในชั้นแฝงที่ถูกกระตุ้นจะส่งผลถึงโหนดในชั้นเอาต์พุต โดยโหนดในชั้นแฝงที่ถูกกระตุ้น อาจส่งผลกระตุ้นหรือยับยั้งการจำแนกให้เป็นผู้ใช้คนหนึ่ง ๆ ได้ ขึ้นอยู่กับค่าน้ำหนักที่กำกับโหนดนั้น

ดังนั้น หากจะพิจารณาว่าอินพุตใดส่งผลกับการจำแนกเป็นผู้ใช้แต่ละคน อาจหาได้จากการพิจารณาที่โหนดในชั้นเอาต์พุตแต่ละโหนดว่า โหนดในชั้นแฝงโหนดใดที่มีค่าน้ำหนักกำกับไว้สูง ทั้งในทางบวกและทางลบ จากนั้นจึงพิจารณาโหนดในชั้นแฝงเหล่านั้นว่า โหนดในชั้นอินพุตโหนดใดที่มีค่าน้ำหนักกำกับไว้สูงทั้งในทางบวกและทางลบเช่นกัน

ในหัวข้อนี้จะทำการยกตัวอย่าง เพื่อแสดงอินพุตที่มีผลต่อการจำแนกเป็นผู้ใช้แต่ละคน โดยจะแสดงโหนดในชั้นแฝงที่มีค่าน้ำหนักสูงสุด 10 อันดับ ทั้งทางบวกและทางลบ และแสดงอินพุตที่มีค่าน้ำหนักสูงที่สุด 10 อันดับ ทั้งในทางบวกและลบ ของแต่ละโหนดในชั้นแฝงเช่นกัน ทั้งนี้จะทำการแยกเป็นส่วนที่ส่งผลให้จำแนกเป็นผู้ใช้คนนั้น ๆ คือ

1) โหนดอินพุตที่มีค่าน้ำหนักเป็นบวกมากที่สุด ของโหนดในชั้นแฝงที่มีค่าน้ำหนักมากเป็นบวกที่สุด ซึ่งจะส่งผลให้จำแนกเป็นผู้ใช้คนนี้มากยิ่งขึ้น
2) โหนดอินพุตที่มีค่าน้ำหนักเป็นลบมากที่สุด ของโหนดในชั้นแฝงที่มีค่าน้ำหนักเป็นลบมากที่สุด ซึ่งอินพุตในลักษณะนี้จะยับยั้งไม่ให้โหนดในชั้นแฝงไปยับยั้งการจำแนก ซึ่งจะส่งผลให้จำแนกเป็นผู้ใช้คนนี้ได้มากยิ่งขึ้นเช่นกัน

3) โหนดอินพุตซึ่งมีค่าน้ำหนักเป็นบวกมากที่สุด ของโหนดในชั้นฮินเดนที่มีค่าน้ำหนักเป็นลบมากที่สุด ซึ่งอินพุตในลักษณะนี้จะไปกระตุ้นให้โหนดในชั้นแฝงไปยับยั้งการจำแนก ซึ่งจะส่งผลให้จำแนกเป็นเป็นผู้ใช้คนนี้ได้น้อยลง

4) โหนดอินพุตซึ่งมีน้ำหนักเป็นลบมากที่สุด ของโหนดในชั้นแฝงที่มีค่าน้ำหนักเป็นบวกมากที่สุด ซึ่งอินพุตในลักษณะนี้จะยับยั้งไม่ให้โหนดในชั้นแฝง ไปส่งผลในการจำแนก ทำให้จำแนกเป็นผู้ใช้คนนี้ได้น้อยลง

ตัวอย่างการจำแนกเมื่อใช้ Log Average/SD 100 เป็นคุณลักษณะ จากการทดลองใน Fold ที่ 1 โดยข้อความฝึกมีความยาว 100 ตัวอักษร และข้อความทดสอบมีความยาว 100 ตัวอักษร อินพุตที่ส่งผลดีต่อการจำแนกเป็นผู้ใช้ U1 จะแสดงในตารางที่ 4-6 และ 4-7 และอินพุตที่ส่งผลไม่ให้จำแนกเป็นผู้ใช้ U1 จะแสดงในตารางที่ 4-8 และ 4-9 ช่องที่แรเงา หมายถึงอินพุตที่ซ้ำกันที่ส่งผลให้จำแนกหรือไม่จำแนกเป็นผู้ใช้ U1 ในตารางนั้น ๆ

สำหรับตัวอย่างการจำแนกเมื่อใช้ Histogram 100 เป็นคุณลักษณะ จากการทดลองใน Fold ที่ 1 โดยข้อความฝึกมีความยาว 100 ตัวอักษร และข้อความทดสอบมีความยาว 100 ตัวอักษร อินพุตที่ส่งผลดีต่อการจำแนกเป็นผู้ใช้ U1 จะแสดงในตารางที่ 4-10 และ 4-11 และอินพุตที่ส่งผลไม่ให้จำแนกเป็นผู้ใช้ U1 จะแสดงในตารางที่ 4-12 และ 4-13

ในตารางที่ 4-10 และ 4-11 ช่องที่แรเงา คืออินพุตที่เป็นการกระจายตัวของความน่าจะเป็นในการพิมพ์ไดกราฟต่าง ๆ ของผู้ใช้ U1 ซึ่งจะเห็นได้ว่าอินพุตลักษณะนี้ จะปรากฏอยู่ในลำดับต้น ๆ หรือหมายความว่ามีความสำคัญสูงที่จะส่งผลให้จำแนกผู้ใช้ออกเป็น U1 และหากพิจารณาตารางที่ 4-12 และ 4-13 ร่วมด้วยก็จะพบว่าไม่มีอินพุตที่เป็นการกระจายตัวของความน่าจะเป็นในการพิมพ์ไดกราฟต่าง ๆ ของผู้ใช้ U1 เลย

ในตารางที่ 4-12 และ 4-13 ช่องที่แรเงา คืออินพุตที่เป็นการกระจายตัวของความน่าจะเป็นในการพิมพ์ไดกราฟประเภทเดียวกันกับที่แรเงาในตารางที่ 4-10 และ 4-11 แต่เป็นของผู้ใช้คนอื่น อินพุตลักษณะนี้เป็นส่วนหนึ่งของอินพุตที่สามารถจำแนกผู้ใช้ออกจากกันได้ดี กล่าวคือ หากอินพุตของตัวอย่างที่เข้ามา หากค่าอินพุตนั้น ๆ มีค่าโน้มเอียงไปทางไดกราฟของผู้ใช้ U1 ก็จะมีโอกาสมากที่จะจำแนกออกมาเป็น U1 ในทางกลับกัน หากอินพุตนั้น ๆ มีค่าโน้มเอียงไปทางไดกราฟของผู้ใช้คนอื่น ๆ ก็มีโอกาสที่โหนดในชั้นแฝงจะยับยั้งไม่ให้จำแนกออกมาเป็น U1 ได้ ซึ่งจะเห็นว่า มีอินพุตประเภทนี้อยู่ในตารางที่ 4-12 และ 4-13 พอสมควร

ตารางที่ 4-6 ตารางแสดงอินพุตที่มีค่าน้ำหนักเป็นบวกสูง และโหนดในชั้นแฝงมีค่าน้ำหนักเป็นบวกสูง เมื่อใช้คุณลักษณะ Log Average/SD

Hidden Weight Rank (+)	Node number	Input Weight Rank (+)									
		1	2	3	4	5	6	7	8	9	10
1	66	mean_Down-Down	mean_LControlKey-z	mean_ค-สระอะ	mean_LShiftKey-Home	sd_Down-Down	mean_ด-RShiftKey	mean_สระเอ-ก	mean_z-LControlKey	mean_LControlKey-w	mean_Back-NumPad1
2	41	mean_Up-Down	mean_a-m	mean_RShiftKey-ไม้โท	mean_LControlKey-w	mean_ป-RShiftKey	mean_v-e	mean_Down-Up	mean_Escape-Return	mean_RShiftKey-สระอี	mean_g-r
3	35	mean_Up-Down	mean_สระอือ-ฝ	mean_Comma-Space	mean_t-c	mean_ค-สระอะ	mean_LShiftKey-d	mean_LControlKey-w	mean_Up-Up	mean_ฐ-ถ	mean_ฝ-สระอะ
4	47	mean_LControlKey-w	mean_t-c	mean_LControlKey-LShiftKey	mean_F5-F5	mean_Return-l	mean_Return-Up	mean_LShiftKey-ย	mean_F5-Down	mean_Left-Delete	mean_x-x
5	80	mean_LShiftKey-สระอุ	mean_ก-LShiftKey	mean_LShiftKey-ไม้เอก	mean_ก-สระอิ	mean_LShiftKey-ไม้โท	mean_End-Delete	mean_สระอี-LShiftKey	mean_Down-Return	mean_NumPad1-NumPad6	mean_e-Back
6	60	sd_Down-Down	mean_Space-Space	sd_Down-Up	mean_Up-Down	mean_ง-Return	mean_RControlKey-f	mean_e-t	sd_Up-Up	mean_Delete-Down	mean_สระเอ-ค
7	44	mean_ว-สระอะ	mean_LShiftKey-ง	mean_สระอะ-ค	mean_x-x	mean_Space-Return	mean_สระอิ-Space	mean_ง-ว	mean_Return-Escape	mean_สระอา-Space	mean_z-z
8	68	mean_LControlKey-w	mean_สระเอ-ห	mean_Next-Next	mean_Return-Down	mean_สระอะ-Space	mean_NumPad5-Add	mean_e-Space	mean_ป-ไม้โท	mean_ป-Space	mean_LShiftKey-Home
9	45	mean_LControlKey-w	mean_F6-F6	mean_D5-Return	mean_Back-LControlKey	mean_x-x	mean_PageUp-PageUp	mean_Next-PageUp	mean_Down-Down	mean_Slash-Slash	mean_Up-Right
10	71	mean_Down-Down	mean_Up-Up	mean_LShiftKey-LShiftKey	mean_Return-LControlKey	mean_ย-ย	mean_อ-สระอา	mean_Space-d	mean_m-Space	mean_Left-Right	mean_c-u

ตารางที่ 4-7 ตารางแสดงอินพุตที่มีค่าน้ำหนักเป็นลบสูง และโหนดในชั้นแฝงมีค่าน้ำหนักเป็นลบสูง เมื่อใช้คุณลักษณะ Log Average/SD

Hidden Weight Rank (-)	Node number	Input Weight Rank (-)									
		1	2	3	4	5	6	7	8	9	10
1	62	mean_LControl Key-w	mean_LControl Key-Up	mean_Return-Space	mean_Left-LControlKey	mean_u-n	mean_ด-Space	mean_ตัวอะ-ตัวแอ	mean_v-Space	mean_Back-Down	mean_ด-ย
2	70	mean_Down-Down	mean_x-x	mean_LControl Key-w	mean_n-Return	mean_Down-z	mean_ก-ร	mean_LShiftKey -ว	mean_Down-Up	mean_Tab-p	mean_LShiftKey -Left
3	73	mean_n-Return	mean_Space-Return	mean_ท-ไม่ห้น อากาศ	mean_u-t	mean_a-m	mean_ก-ไม่ใช	mean_LControl Key-w	mean_ตัวอะ-Return	mean_ค-ตัวอะ	mean_ไม่ห้น อากาศ-ย
4	38	mean_Up-Down	sd_Down-Down	sd_Up-Up	sd_Up-Down	mean_LControl Key-w	mean_RControl Key-RControlKey	mean_F8-F8	mean_LControl Key-f	mean_ม-ไม่ห้น อากาศ	sd_Down-Up
5	56	mean_Down-Down	sd_Down-Down	mean_LControl Key-z	mean_ตัวอา-ไ	mean_ไ-ง	mean_ไม่เอก-ไ	mean_Oemtilde-Back	mean_ก-ว	mean_ไม่ห้น อากาศ-ม	mean_e-a
6	79	mean_ค-ไม่เอก	mean_LShiftKey -End	mean_ตัวะแอ-Back	mean_ตัวะอะ-Return	mean_Down-Up	mean_Return-Escape	mean_ๆ-Return	mean_Space-Return	mean_ด-Back	mean_ข-ตัวะอา
7	57	sd_Down-Down	mean_Up-Down	mean_Down-Down	sd_Down-Up	mean_Down-LControlKey	sd_Up-Up	mean_Up-LControlKey	mean_ด-Back	mean_Oemtilde-Back	mean_RControl Key-RControlKey
8	33	mean_Up-Up	mean_g-Back	mean_Down-End	mean_Down-LControlKey	mean_LControl Key-s	mean_Space-ตัวะแอ	mean_LShiftKey -ร	mean_ร-ด	mean_ตัวะเอ-ม	mean_ย-ม
9	48	mean_Down-Up	sd_Up-Down	mean_ตัวะอะ-Return	mean_ย-Return	mean_a-m	mean_Up-Down	mean_RControl Key-RControlKey	mean_Return-s	mean_LControl Key-w	mean_F6-F6
10	64	mean_Up-Down	sd_Up-Up	sd_Down-Down	sd_Up-Down	mean_Left-LShiftKey	mean_F10-F10	mean_ย-ด	mean_Back-ไ	mean_LShiftKey -Right	mean_LControl Key-w

ตารางที่ 4-8 ตารางแสดงอินพุตที่มีค่าน้ำหนักเป็นบวกสูง และโหนดในชั้นแฝงมีค่าน้ำหนักเป็นลบสูง เมื่อใช้คุณลักษณะ Log Average/SD

Hidden Weight Rank (-)	Node number	Input Weight Rank (Input Weight Type : +)									
		1	2	3	4	5	6	7	8	9	10
1	62	sd_Down-Down	sd_Up-Up	mean_ไม้เอก-สระอะ	mean_v-Down	sd_Down-Up	mean_LControl Key-v	mean_Space-ไ	mean_Right-Down	mean_LControl Key-f	mean_Up-Down
2	70	mean_Up-Down	sd_Up-Down	sd_Up-Up	mean_LControl Key-v	sd_Down-Down	mean_LControl Key-z	mean_NumPad5 -Return	mean_ย-ๆ	mean_Space-LControlKey	mean_v-Down
3	73	mean_LShiftKey -ง	mean_ป-สระอา	mean_ป-Return	mean_Return-Escape	mean_ค-ไม้หันอากาศ	mean_Up-Up	mean_Tab-Tab	mean_u-s	mean_NumPad4 -NumPad0	mean_ค-สระอุ
4	38	mean_Down-Up	mean_Down-Down	mean_Up-Up	mean_LShiftKey -ง	mean_Add-Return	mean_พ-สระอี	mean_Right-Return	mean_ม-Space	mean_ง-ง	mean_น-น
5	56	mean_ค-สระอะ	mean_ป-ป	mean_น-ไม้โท	mean_LShiftKey -D6	mean_ย-ย	mean_g-o	mean_x-Right	mean_Back-Oemtilde	sd_Up-Down	mean_ก-ม
6	79	mean_RControl Key-RControlKey	mean_Delete-Down	mean_Next-Next	mean_Space-ท	mean_สระอา-Return	mean_c-Down	mean_พ-LShiftKey	mean_z-LControlKey	mean_ป-Return	mean_ๆ-ๆ
7	57	mean_Down-Up	mean_Up-Up	mean_Space-Space	mean_LControl Key-c	mean_Down-Right	mean_ๆ-ๆ	mean_อ-ไม้เอก	mean_ท-สระอา	mean_สระอา-ว	mean_o-n
8	33	mean_ค-ไม้โท	mean_NULL-NULL	mean_LShiftKey -ค	mean_n-Return	mean_สระอิ-Back	mean_a-p	mean_Right-Back	mean_a-t	mean_F2-F2	mean_LControl Key-v
9	48	mean_Down-Down	sd_Down-Down	sd_Down-Up	mean_RShiftKey -สระอุ	mean_Space-ๆ	mean_NumPad5 -Add	mean_m-o	mean_Up-Return	mean_Space-t	mean_Add-Return
10	64	mean_Down-Down	mean_n-Return	mean_Subtract-Return	mean_ก-Return	mean_Delete-Down	mean_LShiftKey -ค	mean_ว-สระอา	mean_Delete-Up	mean_น-ไม้หันอากาศ	mean_จ-สระอา

ตารางที่ 4-9 ตารางแสดงอินพุตที่มีค่าน้ำหนักเป็นลบสูง และโหนดในชั้นแฝงมีค่าน้ำหนักเป็นบวกสูง เมื่อใช้คุณลักษณะ Log Average/SD

Hidden Weight Rank (+)	Node number	Input Weight Rank (Input Weight Type : -)									
		1	2	3	4	5	6	7	8	9	10
1	66	sd_Up-Down	mean_Next-Next	mean_ไม้เอก-สระอะ	mean_Delete-Down	mean_LShiftKey-ง	mean_x-x	mean_RControlKey-RControlKey	mean_c-c	mean_Up-Down	mean_ก-สระอา
2	41	mean_Down-Down	mean_LShiftKey-สระอุ	mean_z-LControlKey	mean_Next-Next	mean_LShiftKey-ไม้โท	mean_Right-Right	mean_n-LShiftKey	mean_Left-Up	mean_Oemtilde-LShiftKey	mean_ม-Back
3	35	mean_อ-ไม้เอก	mean_p-r	mean_ค-ไม้เอก	mean_ข-ไม้ไหม้อากาศ	mean_D0-Point	mean_ไม้เอก-Return	mean_LControlKey-x	mean_Oemtilde-p	mean_ง-n	mean_RControlKey-Home
4	47	mean_Down-Down	mean_พ-สระอี	mean_ไม้เอก-Return	mean_ว-สระอะ	mean_Back-Space	mean_NumPad5-NumPad5	mean_ไม้โท-Space	mean_l-o	mean_p-i	mean_Up-Down
5	80	sd_Down-Down	mean_ก-ไม้โท	mean_s-m	mean_RShiftKey-ค	mean_Escape-Return	mean_Return-Escape	mean_RShiftKey-ไม้โท	mean_LControlKey-x	mean_"]-"]	mean_n-RShiftKey
6	60	mean_ค-สระอะ	mean_Down-Up	mean_ค-Back	mean_v-LControlKey	mean_LControlKey-LShiftKey	mean_ไม้เอก-ล	mean_Return-s	mean_สระเอ-อ	mean_Space-Left	mean_Back-ค
7	44	mean_Back-ไม้โท	mean_ม-Return	mean_Oemtilde-Space	mean_ไม้เอก-Return	mean_RShiftKey-สระอุ	mean_PageUp-PageUp	mean_ก-สระอา	mean_ย-Return	mean_Oemtilde-RShiftKey	mean_Right-Back
8	68	mean_Up-Down	mean_RShiftKey-ไม้โท	mean_ค-Back	mean_x-x	mean_ไม้เอก-Back	mean_NumPad1-NumPad1	mean_RShiftKey-สระอือ	mean_อ-สระอะ	mean_ว-Return	mean_z-z
9	45	mean_Down-Up	mean_Back-Oemtilde	mean_F7-F7	mean_RControlKey-RControlKey	mean_F8-F8	mean_i-Back	mean_Right-Delete	mean_ง-ค	mean_c-c	mean_D2-Return
10	71	mean_Up-Down	sd_Up-Down	sd_Up-Up	mean_s-t	mean_g-Space	mean_สระเอ-อ	mean_i-l	mean_Space-Space	mean_ก-ไม้โท	mean_น-ก

ตารางที่ 4-10 ตารางแสดงอินพุตที่มีค่าน้ำหนักเป็นบวกสูง และโหนดในชั้นแฝงมีค่าน้ำหนักเป็นบวกสูง เมื่อใช้คุณลักษณะ Histogram

Hidden Weight Rank (+)	Node number	Input Weight Rank (+)									
		1	2	3	4	5	6	7	8	9	10
1	81	U01_LControlKey-w	U18_สระอะ-Return	U16_Up-Down	U01_Space-Oemtilde	U01_ไม้เอก-สระอา	U01_Next-Next	U30_RControlKey-Home	U30_r-e	U01_ม-ไม้เอก	U13_d-a
2	42	U09_Right-Right	U09_LControlKey-x	U09_a-n	U09_LControlKey-v	U09_Left-Left	U09_Back-RShiftKey	U01_LControlKey-w	U03_RShiftKey-สระอือ	U09_Back-Oemtilde	U01_Back-Back
3	57	U01_LControlKey-w	U01_Space-Oemtilde	U27_LControlKey-LShiftKey	U01_ไม้เอก-สระอา	U01_Back-Back	U01_Next-Next	U11_x-x	U03_Right-Right	U01_ม-ไม้เอก	U01_น-ม
4	78	U07_Down-Up	U07_Up-Down	U21_Down-Return	U15_LShiftKey-Home	U16_LControlKey-s	U24_ไม้หัน-อากาศ-ไม้โท	U21_Back-Back	U27_LControlKey-LShiftKey	U17_LControlKey-v	U07_Space-Oemtilde
5	45	U08_Down-Up	U08_Back-Back	U08_Down-Down	U08_a-r	U08_Space-Oemtilde	U08_Up-Up	U32_RControlKey-RControlKey	U08_r-Space	U08_LControlKey-s	U08_Slash-Slash
6	73	U05_x-x	U04_Space-Oemtilde	U13_Back-CapsLock	U13_D0-D1	U04_LControlKey-v	U05_z-x	U05_z-z	U11_x-x	U13_RShiftKey-D9	U04_LControlKey-c
7	79	U08_Down-Up	U25_NumPad0-Return	U25_a-m	U25_RControlKey-F3	U25_NumPad1-NumPad0	U22_Down-Left	U16_Right-Down	U08_Down-Down	U16_Up-Down	U08_a-r
8	43	U21_s-t	U33_NumPad0-NumPad0	U26_a-t	U18_F7-F7	U32_Up-Up	U21_Down-Right	U18_NumPad1-NumPad9	U18_สระอือ-ไม้เอก	U18_F8-Down	U11_จ-สระอา
9	39	U04_LControlKey-v	U04_Space-Oemtilde	U04_LControlKey-a	U04_LControlKey-c	U04_Back-Back	U04_a-Delete	U04_ไม้หัน-อากาศ-ไม	U04_Decimal-NumPad6	U04_Delete-Delete	U04_Return-Return
10	38	U23_Return-LControlKey	U25_NumPad0-Return	U21_Return-LControlKey	U31_NumPad0-NumPad0	U30_d-e	U21_NumPad0-NumPad0	U31_LControlKey-สระแอ	U25_Down-Left	U25_RShiftKey-F11	U21_Right-Left

ตารางที่ 4-11 ตารางแสดงอินพุตที่มีค่าน้ำหนักเป็นลบสูง และโหนดในชั้นแฝงมีค่าน้ำหนักเป็นลบสูง เมื่อใช้คุณลักษณะ Histogram

Hidden Weight Rank (-)	Node number	Input Weight Rank (-)									
		1	2	3	4	5	6	7	8	9	10
1	48	U01_LControlKey-w	U01_ไม้เอก-สระอา	U01_Next-Next	U01_Back-Back	U01_Back-Oemtilde	U01_ไ-ม	U01_Space-Oemtilde	U01_Down-LControlKey	U01_ม-ไม้เอก	U01_LControlKey-s
2	51	U01_LControlKey-w	U16_Up-Down	U01_ไม้เอก-สระอา	U01_ไ-ม	U03_Back-Back	U01_Next-Next	U01_Back-Oemtilde	U01_Back-Back	U16_LControlKey-s	U03_LShiftKey-Down
3	37	U10_p-s	U11_Delete-Down	U11_LControlKey-v	U10_t-Space	U32_RControlKey-RControlKey	U24_v-RControlKey	U25_a-m	U01_t-a	U01_LControlKey-w	U24_RControlKey-RControlKey
4	56	U02_LControlKey-v	U01_LControlKey-w	U28_NumPad1-NumPad1	U19_LShiftKey-OemMinus	U01_Space-Oemtilde	U23_Return-LControlKey	U01_ไม้เอก-สระอา	U01_ม-ไม้เอก	U12_LControlKey-c	U01_Next-Next
5	58	U07_Down-Up	U07_Up-Down	U24_v-RControlKey	U01_LControlKey-w	U01_Back-Back	U07_Space-Oemtilde	U25_a-m	U01_ไ-ม	U03_Right-Right	U07_j-k
6	76	U04_LControlKey-v	U30_d-e	U07_Down-Up	U07_Up-Down	U04_LControlKey-a	U30_RControlKey-Home	U07_Back-Back	U04_a-Delete	U04_LControlKey-c	U31_Space-RShiftKey
7	74	U16_LControlKey-s	U31_Down-Down	U22_Right-Back	U27_LControlKey-ย	U29_LShiftKey-Left	U26_Oemtilde-LShiftKey	U29_Left-Right	U18_สระอะ-Return	U29_Return-Up	U10_ไม้เอก-น
8	44	U10_p-s	U15_LShiftKey-Home	U28_Right-Right	U19_t-Space	U21_s-t	U16_Up-Down	U25_a-m	U22_Down-Left	U30_a-t	U22_Back-Quote
9	77	U33_Down-Left	U21_Down-Right	U21_LControlKey-v	U10_p-s	U12_ย-น	U02_น-Return	U28_ง-Return	U33_Space-LShiftKey	U14_Space-Oemtilde	U24_สระอา-Space
10	53	U16_LControlKey-s	U11_x-x	U16_Up-Down	U32_LShiftKey-Right	U15_LShiftKey-Home	U32_j-RControlKey	U14_ว-ย	U16_Right-Down	U16_Down-Up	U32_RControlKey-RControlKey

ตารางที่ 4-12 ตารางแสดงอินพุตที่มีค่าน้ำหนักเป็นบวกสูง และโหนดในชั้นแฝงมีค่าน้ำหนักเป็นลบสูง เมื่อใช้คุณลักษณะ Histogram

Hidden Weight Rank (-)	Node number	Input Weight Rank (Input Weight Type : +)									
		1	2	3	4	5	6	7	8	9	10
1	48	U04_LControlKey-v	U04_Space-Oemtilde	U04_LControlKey-c	U04_ไม้ทับอากาศ-ม	U04_LControlKey-a	U05_x-x	U04_a-Delete	U04_Decimal-NumPad6	U07_Down-Up	U07_Up-Down
2	51	U11_Delete-Down	U11_Down-Up	U11_Down-Delete	U11_D0-D0	U11_Right-Left	U30_c-c	U04_LControlKey-v	U11_Delete-Right	U11_สระแอ-ล	U28_NumPad1-NumPad1
3	37	U02_LControlKey-v	U23_Return-LControlKey	U23_LControlKey-v	U15_LShiftKey-Home	U23_LControlKey-x	U23_Down-NumPad1	U23_Down-Tab	U23_NumPad1-NumPad2	U19_e-r	U14_D9-D9
4	56	U08_Down-Up	U33_NumPad0-NumPad0	U33_Right-Right	U17_สระแอ-ห	U33_t-e	U03_Right-Right	U11_Delete-Down	U08_Down-Down	U08_o-r	U09_Up-Up
5	58	U19_n-n	U21_Down-Return	U09_LControlKey-v	U21_Down-Right	U09_LControlKey-x	U19_LShiftKey-SemiColon	U29_Down-Up	U19_Up-Down	U13_l-e	U32_LShiftKey-Right
6	76	U05_x-x	U11_x-x	U05_z-x	U10_p-s	U17_Space-Space	U05_x-z	U05_z-z	U10_Space-LShiftKey	U10_t-Space	U17_t-Space
7	74	U12_LControlKey-c	U11_D0-D0	U28_NumPad1-NumPad1	U27_t-e	U13_u-t	U19_e-r	U02_LControlKey-v	U14_D0-D0	U19_LShiftKey-OemMinus	U13_D1-D0
8	44	U05_x-x	U11_x-x	U27_t-e	U30_RControlKey-Home	U05_z-x	U33_s-t	U27_LControlKey-สระแอ	U03_Right-Right	U31_LControlKey-ย	U05_z-z
9	77	U27_LControlKey-LShiftKey	U30_NumPad2-NumPad0	U19_i-t	U22_Down-Left	U27_p-r	U27_c-Space	U30_r-o	U30_o-u	U03_LShiftKey-Left	U22_ไม้เอก-น
10	53	U17_LControlKey-c	U28_NumPad1-NumPad1	U07_Down-Up	U28_NumPad2-Return	U07_Space-Oemtilde	U07_Up-Down	U07_o-n	U04_LControlKey-v	U07_f-f	U28_Delete-Delete

ตารางที่ 4-13 ตารางแสดงอินพุตที่มีค่าน้ำหนักเป็นลบสูง และโหนดในชั้นแฝงมีค่าน้ำหนักเป็นบวกสูง เมื่อใช้คุณลักษณะ Histogram

Hidden Weight Rank (+)	Node number	Input Weight Rank (Input Weight Type : -)									
		1	2	3	4	5	6	7	8	9	10
1	81	U08_Down-Up	U21_Down-Right	U21_Return-LControlKey	U27_ธ-ดีระ๊ะ	U23_Return-LControlKey	U08_e-r	U29_LShiftKey-Left	U21_Down-Return	U08_o-r	U22_Space-Space
2	42	U07_Down-Up	U07_Up-Down	U07_Space-Oemtilde	U32_RControlKey-RControlKey	U07_RShiftKey-n	U24_v-RControlKey	U07_Back-Oemtilde	U24_RControlKey-RControlKey	U22_Right-Back	U11_D0-D0
3	57	U08_Down-Up	U30_c-o	U04_LControlKey-v	U10_p-s	U25_a-m	U32_RControlKey-RControlKey	U32_c-a	U24_RControlKey-RControlKey	U28_Down-Down	U20_Left-Left
4	78	U14_D9-D9	U14_D0-D0	U05_x-x	U22_Down-Left	U33_NumPad0-NumPad0	U05_z-x	U33_s-t	U19_o-r	U10_p-s	U33_Right-Right
5	45	U05_x-x	U11_x-x	U05_z-x	U25_RShiftKey-F11	U05_x-z	U13_Back-Back	U05_z-z	U04_Down-Down	U21_Return-LControlKey	U11_Delete-Down
6	73	U08_Down-Up	U11_Delete-Down	U08_Back-Back	U30_d-e	U30_c-o	U30_NumPad1-NumPad0	U30_LControlKey-z	U31_NumPad0-NumPad0	U32_RControlKey-RControlKey	U06_F1-F1
7	79	U07_Down-Up	U07_Up-Down	U24_v-RControlKey	U03_ธ-ดีระ๊ะดิด	U07_o-n	U21_Down-Right	U21_LControlKey-v	U03_Right-Right	U07_Space-Oemtilde	U18_Return-Up
8	43	U24_v-RControlKey	U17_LControlKey-c	U16_Up-Down	U16_Right-Left	U31_LControlKey-c	U27_LControlKey-LShiftKey	U30_i-n	U30_c-t	U19_o-r	U30_Down-Up
9	39	U11_x-x	U05_x-x	U33_NumPad0-NumPad0	U30_Return-Return	U20_Return-Return	U26_Return-Return	U06_Left-Left	U19_Return-Return	U19_i-t	U12_Back-Back
10	38	U27_LControlKey-LShiftKey	U32_Up-Left	U33_Down-Up	U19_Space-Back	U33_t-e	U16_Down-Up	U19_Up-Up	U19_LShiftKey-SemiColon	U27_t-e	U27_u-p

ต่อไปนี้จะยกตัวอย่างสถานการณ์การจำแนกข้อความทดสอบข้อความหนึ่ง เพื่อแสดงให้เห็นถึงความสามารถในการจำแนกข้อความขนาดสั้นของวิธีที่นำเสนอ ข้อความทดสอบนี้เป็นข้อความของ U1 ในการทดลองจำแนกข้อความทดสอบยาว 100 ตัวอักษร ของการทดลอง fold ที่ 1 ซึ่งจะแสดงการรายละเอียดการจำแนกด้วยวิธีการจำแนกแต่ละวิธี

ในการจำแนกด้วยวิธีการของ D. Gunetti และ C. Picardi ระบบจะเปรียบเทียบข้อความทดสอบ กับข้อความฝึกของผู้ใช้แต่ละคน และพิจารณาระยะห่างเฉลี่ย สำหรับข้อความทดสอบชุดนี้ เมื่อเทียบกับข้อความฝึกของ U1 จะพบว่ามีข้อความที่มีไดกราฟซ้ำกันอยู่บ้างจำนวน 9 จาก 13 ข้อความ คำนวณระยะเฉลี่ยได้ 1.1123 แต่เมื่อเทียบกับข้อความฝึกของ U28 จะพบว่ามีข้อความที่มีไดกราฟซ้ำกันอยู่บ้างจำนวน 8 จาก 13 ข้อความ คำนวณระยะเฉลี่ยได้ 0.9158 เท่านั้น จึงทำให้ระบบจำแนกข้อความนี้ว่าเป็นของ U28 ซึ่งเป็นคำตอบที่ผิด โดยสาเหตุน่าจะเกิดจาก ขณะที่เทียบกับข้อความของ U28 นั้น ในการเปรียบเทียบแต่ละครั้งมีจำนวนไดกราฟที่ซ้ำกันน้อยมาก จึงทำให้การคำนวณค่าระยะออกมาไม่เหมาะสมเท่าที่ควร ดังรายละเอียดในตารางที่ 4-14

ตารางที่ 4-14 ตารางแสดงรายละเอียดการคำนวณระยะห่างของตัวอย่าง โดยวิธีการจำแนกของ D. Gunetti และ C. Picardi

Compared With		1	2	3	4	5	6	7	8	9	Average
U28	Distance	0	1.4	1.42	1	1	0.72	1.12	0.67	-	0.9158
	Length	1	10	10	1	1	6	17	3	-	6.125
U1	Distance	1.08	1.05	1.15	1.24	1.13	1.37	1.04	0.91	1.04	1.1123
	Length	24	30	25	20	28	30	24	27	37	27.2222

สำหรับการเปรียบเทียบโดยการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมของระยะเวลาระหว่างไดกราฟนั้น สามารถจำแนกได้อย่างถูกต้อง โดยเมื่อทำการจำแนกเสร็จแล้ว โหนดในชั้นเอาต์พุตของ U1 มีค่าสูงถึง 0.97 ในขณะที่ โหนดที่มีค่าสูงเป็นอันสองคือโหนดของ U19 มีค่าเพียง 0.028 เท่านั้น หากนำรายการโหนดในชั้นแฝงที่ถูกกระตุ้นทั้งหมดเมื่อทำการจำแนกข้อความนี้ มาพิจารณาร่วมกับค่าน้ำหนักที่กำกับโหนดในชั้นแฝงที่มีอันดับสูงสุด 10 อันดับทั้งทางบวกและทางลบ ของ U1 และ U19 จะเห็นว่าสำหรับข้อความนี้ นิวรอลเน็ตเวิร์กได้กำหนดค่าน้ำหนักที่เหมาะสมเพื่อให้สามารถจำแนกข้อความนี้ได้ จากรายละเอียดในตารางที่ 4-15 จะเห็นว่า เมื่อพิจารณา U1 โหนดในชั้นแฝงที่มีค่าน้ำหนักสูงสุดในทางบวก 10 อันดับแรกถูกกระตุ้นทั้งหมด และมีโหนดที่มีค่าน้ำหนักสูงสุดทางลบเพียงโหนดเดียวที่ถูกกระตุ้น ในขณะที่เมื่อ

พิจารณา U19 จะเห็นว่า มีเพียง 7 ใน 10 โหนดที่มีค่าน้ำหนักสูงสุดทางบวกเท่านั้นที่ถูกกระตุ้น แต่มีโหนดที่มีค่าน้ำหนักสูงสุดทางลบอยู่ถึง 3 ใน 10 โหนดที่ถูกกระตุ้น จึงทำให้ผลลัพธ์ที่โหนดในชั้นเอาต์พุตมีค่าแตกต่างกันอย่างเห็นได้ชัด และทำให้จำแนกข้อความนี้ได้อย่างถูกต้อง

ตารางที่ 4-15 ตารางแสดงโหนดในชั้นแฝงที่มีค่าน้ำหนักสูงสุดและถูกกระตุ้น ในการจำแนกข้อความของ U1 ด้วยวิธี Log Average / SD

Activated Hidden Nodes		*35*	*36*	*37*	*39*	*40*	*41*	*42*	*44*	*45*	*47*
		49	*51*	*52*	*54*	*59*	*60*	*61*	*62*	*65*	*66*
		67	*68*	*71*	*74*	*76*	*78*	*80*	*81*		
Rank		1	2	3	4	5	6	7	8	9	10
U1	+	66	41	35	47	80	60	44	68	45	71
	-	62	70	73	38	56	79	57	33	48	64
U19	+	50	51	74	48	59	42	40	61	47	70
	-	82	43	78	38	81	53	71	63	56	34

ในทำนองเดียวกัน หากพิจารณาวิธีการใช้ค่าเฉลี่ยความน่าจะเป็น จะพบว่าเมื่อจำแนกเสร็จแล้วจะได้โหนดในชั้นเอาต์พุตของ U1 มีค่า 0.794 และโหนดในชั้นเอาต์พุตที่มีค่าสูงเป็นอันดับสองคือโหนดของ U5 ซึ่งมีค่า 0.103 ซึ่งทำให้จำแนกได้อย่างถูกต้อง และหากพิจารณาโหนดในชั้นแฝงที่ถูกกระตุ้นควบคู่กับค่าน้ำหนักเช่นเดียวกับวิธีการที่ผ่านมา ดังรายละเอียดในตารางที่ 4-16 จะเห็นว่า สำหรับ U1 มีโหนดที่มีค่าน้ำหนักสูงสุดทางบวกถึง 8 ใน 10 โหนดที่ถูกกระตุ้น ส่วน U5 มีเพียง 6 ใน 10 โหนดเท่านั้นที่ถูกกระตุ้น ในขณะที่ในทางลบ ผู้ใช้ทั้งสองคนมีโหนดที่ถูกกระตุ้นเพียง 1 โหนดเท่ากัน จึงทำให้ค่าของโหนดเอาต์พุตมีความแตกต่าง แม้ไม่เท่าในตัวอย่างที่ผ่านมา แต่ก็ยังสามารถจำแนกได้อย่างถูกต้องเช่นกัน

ตารางที่ 4-16 ตารางแสดงโหนดในชั้นแฝงที่มีค่าน้ำหนักสูงสุดและถูกกระตุ้น ในการจำแนกข้อความของ U1 ด้วยวิธี Histogram

Activated Hidden Nodes		*34*	*38*	*42*	*45*	*52*	*56*	*57*	*60*	*65*	*68*
		70	*73*	*75*	*78*	*79*	*80*	*81*			
Rank		1	2	3	4	5	6	7	8	9	10
U1	+	81	42	57	78	45	73	79	43	39	38
	-	48	51	37	56	58	76	74	44	77	53
U5	+	34	73	68	60	52	48	44	46	76	38
	-	40	54	50	72	41	77	43	45	49	67

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1. สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอวิธีการจำแนกผู้ใช้โดยอาศัยข้อมูลระยะเวลาในการพิมพ์ข้อความอิสระที่ได้ผลการจำแนกดีเมื่อใช้ข้อความอิสระขนาดสั้น โดยได้เสนอการใช้คุณลักษณะสองแบบ คือ การใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของข้อมูลระยะเวลาระหว่างไดกราฟ และการใช้ค่าเฉลี่ยของความน่าจะเป็น ร่วมกับนิวรอลเน็ตเวิร์ก นอกจากนี้ยังเสนอวิธีการผสานการใช้คุณลักษณะทั้งสองเพื่อให้ได้ผลความแม่นที่ดียิ่งขึ้น

ในงานวิจัยนี้ได้ศึกษาวิธีการจำแนกผู้ใช้โดยอาศัยข้อมูลเวลาในการพิมพ์ข้อความอิสระที่เคยมี และพบว่างานวิจัยนั้นยังมีจุดอ่อนอยู่ที่ไม่สามารถทำงานได้ดีเมื่อใช้ข้อความอิสระที่มีขนาดสั้น เพราะวิธีนั้นใช้การเปรียบเทียบเพื่อคำนวณระยะห่างของตัวอย่างสองตัวโดยใช้เฉพาะประเภทของไดกราฟที่มีซ้ำกันทั้งสองตัวอย่างในการเปรียบเทียบเท่านั้น เมื่อข้อความมีขนาดสั้นจำนวนไดกราฟที่ซ้ำกันจะน้อยลง ทำให้ระยะห่างที่คำนวณได้ไม่เหมาะสม เมื่อนำไปใช้จำแนกจึงไม่สามารถจำแนกได้ดี

ในวิทยานิพนธ์ฉบับนี้ จึงเสนอวิธีที่จะจำแนกผู้ใช้ โดยการเปรียบเทียบข้อความกับโมเดลที่สร้างขึ้นมาแทน ซึ่งได้ใช้นิวรอลเน็ตเวิร์กเป็นตัวเรียนรู้เพื่อสร้างโมเดล นอกจากนี้นิวรอลเน็ตเวิร์กจะสามารถเรียนรู้ค่าน้ำหนักที่กำกับแต่ละอินพุต ซึ่งค่าน้ำหนักเหล่านี้จะเป็นตัวบ่งบอกความสำคัญของอินพุตแต่ละตัวว่ามีผลต่อการจำแนกเป็นผู้ใช้แต่ละคนมากน้อยเพียงใด การใช้นิวรอลเน็ตเวิร์กเป็นตัวเรียนรู้ ร่วมกับการใช้คุณลักษณะที่เหมาะสม ทำให้การจำแนกผู้ใช้ด้วยข้อความอิสระที่มีขนาดสั้น ๆ ทำได้ดีขึ้น

โดยสรุป เมื่อเปรียบเทียบกันแล้ว วิธีที่นำเสนอในวิทยานิพนธ์ฉบับนี้ สามารถจำแนกผู้ใช้ได้ดีกว่ามากในการทดลองที่ใช้ข้อความอิสระขนาดสั้น ๆ ประมาณ 100-500 ตัวอักษร และให้ผลความแม่นในการจำแนกผู้ใช้ที่ใกล้เคียงกันในการทดลองที่ใช้ข้อความอิสระที่ยาวขึ้น

5.2. ข้อจำกัด

แม้ว่าวิทยานิพนธ์ฉบับนี้จะแสดงให้เห็นถึงความสามารถในการจำแนกผู้ใช้ด้วยระยะเวลาในการพิมพ์ข้อความอิสระขนาดสั้น ๆ ได้ แต่ผลงานในวิทยานิพนธ์ฉบับนี้ยังมีข้อจำกัดในการนำไปใช้งานจริงอยู่ในบางแง่มุม ดังต่อไปนี้

ประการแรกคือ ความสามารถในการรองรับจำนวนผู้ใช้ที่มากขึ้น หากในระบบที่ใช้งานจริงมีจำนวนผู้ใช้มากกว่าในการทดลองนี้มาก ๆ ระบบอาจไม่สามารถใช้งานได้หรืออาจจำแนกผู้ใช้ได้ถูกต้องน้อยลง เนื่องจาก เมื่อมีจำนวนผู้ใช้มาก ก็จะมีข้อมูลมาก ซึ่งในการใช้คุณลักษณะบางประเภทจะทำให้จำนวนอินพุตมีมากขึ้นกว่าเดิม นอกจากนี้แล้ว การมีข้อมูลจำนวนมากจะทำให้ใช้เวลาในการเรียนรู้นิวรอลเน็ตเวิร์กนานขึ้นอีกด้วย

อีกประการหนึ่ง คือ ความสามารถในการจำแนกเมื่อมีข้อมูลฝึกจำนวนน้อย ๆ ในบางสถานการณ์ในการใช้งานจริง เราอาจไม่สามารถเก็บข้อมูลจำนวนมาก ๆ จากผู้ใช้ เพื่อใช้เป็นข้อมูลฝึกได้ ดังนั้น เมื่อข้อมูลฝึกมีจำนวนน้อยลง ก็อาจทำให้ความสามารถในการจำแนกผู้ใช้ของระบบลดลงได้

5.3. ข้อเสนอแนะ

ความสามารถในการสร้างระบบเพื่อให้จำแนกผู้ใช้ด้วยการพิมพ์ข้อความอิสระสามารถนำไปเป็นพื้นฐานของระบบประยุกต์ต่าง ๆ ได้ เช่น การใช้ในการยืนยันตัวตนบุคคลต่อระบบ สามารถนำวิธีการจำแนกผู้ใช้นี้ไปเป็นส่วนหนึ่งของระบบยืนยันตัวตนได้โดยการกำหนดกฎในการยืนยันตัวตนที่อิงจากผลการจำแนกผู้ใช้ หรือนำไปใช้ในการระบุตัวตนเพื่อป้องกันการปลอมแปลงการโพสต์ข้อความบนเว็บบอร์ด หรือเว็บไซต์เครือข่ายสังคมออนไลน์ โดยพิจารณาความคล้ายคลึงของระยะเวลาในการพิมพ์ หรืออาจนำไปใช้เป็นลายเซ็นอิเล็กทรอนิกส์สำหรับอีเมลหรือเอกสารต่าง ๆ ได้ นอกจากนี้ การสร้างระบบให้ทำงานได้ดีแม้ใช้ข้อความอิสระที่มีขนาดสั้น ดังในวิทยานิพนธ์ฉบับนี้ จะช่วยลดข้อจำกัดในการนำไปใช้งานได้อีกด้วย

รายการอ้างอิง

[1] Hosseinzadeh, D., and Krishnan, S. Gaussian mixture modeling of keystroke patterns for biometric applications. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 38, 6 (2008): 816-826.

[2] Montalv, J., and Freire, E. On the equalization of keystroke timing histograms. Pattern Recogn. Lett. 27, 13 (2006): 1440–1446.

[3] บุญเสริม กิจศิริกุล. ปัญญาประดิษฐ์, 169-185. 2548.

[4] Joyce, R., and Gupta, G. Identity authentication based on keystroke latencies. Commun. ACM 33, 2 (1990): 168–176.

[5] Lammers, A., and Langenfeld, S. Identity authentication based on keystroke latencies using neural networks. J. Comput. Small Coll. 6, 5 (1991): 48-51.

[6] Chen, L., Weng, L., and Chee, L. Keystroke patterns classification using the artmap-fd neural network. Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on 1 (2007): 61-64.

[7] Lee, J., Choi, S., and Byung-Ro, M. An evolutionary keystroke authentication based on ellipsoidal hypothesis space. GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation (2007): 2090–2097.

[8] Bergadano, F., Gunetti, D., and Picardi, C. User authentication through keystroke dynamics. ACM Trans. Inf. Syst. Secur. 5, 4 (2002): 367–397.

[9] Curtin, M. et al. Keystroke biometric recognition on long-text input: A feasibility study. Proc. Int. Workshop Sci Comp/Comp Stat (IWSCCS 2006) (2006).

[10] Hocquet, S., Ramel, J., and Cardot, H. User classification for keystroke dynamics authentication. Lecture Notes in Computer Science 4642 (2007): 531-539.

[11] Monrose, F., and Rubin, A. Authentication via keystroke dynamics. CCS '97: Proceedings of the 4th ACM conference on Computer and communications security (1997): 48–56.

[12] Gunetti, D., and Picardi, C. Keystroke analysis of free text. ACM Trans. Inf. Syst. Secur. 8, 3 (2005): 312-347.

[13] Sim, T., and Janakiraman, R. Are digraphs good for free-text keystroke dynamics?. Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on (June 2007): 1-6.

[14] Duda, R., Hart, P., and Stork, D. Pattern Classification. 2nd Edition. Wiley-Interscience, 2000.

[15] Tappert, C., Villani, M., and Cha, S. Keystroke Biometric Identification and Authentication on Long-Text Input. 2009.

[16] Teh, P., Teoh, A., Tee, C., and Ong, T. A multiple layer fusion approach on keystroke dynamics. Pattern Analysis & Applications (2009).

[17] Lv, H., and Wang, W. Biologic verification based on pressure sensor keyboards and classifier fusion techniques. Consumer Electronics, IEEE Transactions on 52, 3 (2006): 1057–1063.

[18] Hall, M. et al. The WEKA data mining software: an update. SIGKDD Explorations 11, 1 (2009): 10-18.

ภาคผนวก

ผล False Positive และ False Negative ของผู้ใช้แต่ละคน จากการทดลองการจำแนกด้วยวิธีที่แตกต่างกัน

ในหัวข้อนี้จะแสดงผลค่า False Positive และ False Negative แยกสำหรับผู้ใช้แต่ละคน จากการทดลองจำแนกด้วยวิธีที่แตกต่างกัน คือ การจำแนกด้วยวิธีของ D. Gunetti และ C. Picardi จำแนกด้วยการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึมของระยะเวลาระหว่างไดกราฟเป็นคุณลักษณะ และการจำแนกด้วยการใช้ค่าเฉลี่ยความน่าจะเป็นจากฟังก์ชันแจกแจงที่สร้างด้วยฮิสโตแกรมเป็นคุณลักษณะ

เมื่อนำผลการทดลอง ที่ใช้ระยะห่างที่ให้ผลความแม่นดีที่สุด (R_2+A_2 ที่ใช้ข้อความฝึกที่มีความยาว 1000 ตัวอักษร) จากการทดลองจำแนกด้วยวิธีการของ D. Gunetti และ C. Picardi มาแจกแจงค่า False Positive และ False Negative ของผู้ใช้แต่ละคนจะได้ดังตารางที่ ก-1 และตารางที่ ก-2 ซึ่งจะเห็นได้ว่า ในกรณีที่ข้อความทดสอบมีขนาดสั้นนั้น ยังมีตัวอย่างของผู้ใช้หลาย ๆ คนที่มีค่า False Positive สูงมาก เช่น U3, U5, U6 และ U7 โดยเฉพาะ U5 ที่มีค่าเฉลี่ยถึงร้อยละ 12.2565 ในขณะที่ค่า False Negative จะมีค่าพอ ๆ กันระหว่างผู้ใช้แต่ละคน

หากเปรียบเทียบผลการทดลองกับวิธีอื่นๆที่ให้ผลความแม่นดี คือ การใช้ Log Average/SD ที่ใช้ข้อความฝึกที่มีความยาว 100 ตัวอักษร และ การใช้ Histogram ที่ใช้ข้อความฝึกที่มีความยาว 100 ตัวอักษรเช่นกัน เมื่อนำมาแจกแจงค่า มาแจกแจงค่า False Positive และ False Negative ของผู้ใช้แต่ละคน จะเห็นได้ว่าทั้งสองวิธีนั้นมีค่า False Positive และ False Negative นั้นลดลงอย่างมากในโดยเฉพาะในกรณีที่ข้อความทดสอบมีขนาดสั้นนั้น เมื่อเทียบกับการใช้วิธีการจำแนกของ D. Gunetti และ C. Picardi ดังในตารางที่ ก-3, ก-4, ก-5 และ ก-6

ตารางที่ ก-1 ตารางแสดงค่า False Positive (ค่าเฉลี่ยร้อยละ) ในการจำแนกผู้ใช้โดยใช้การวัดระยะทางแบบ R_2+A_2 โดยใช้ข้อความฝึกยาว 1000 ตัวอักษร

User	Length of Test Sample (Characters)									
	100	200	300	400	500	600	700	800	900	1000
U1	0.1422	0.1302	0.061	0	0	0	0	0	0	0
U2	2.4952	2.0828	1.7591	1.6162	0.8081	0.7576	1.0606	0.303	0.6061	0.303
U3	6.3728	3.2955	2.7901	1.2121	1.3131	0.303	0.1515	1.0606	0.303	0
U4	1.522	1.3883	0.6064	1.0101	0.5051	0.303	0.1515	0.1515	0	0
U5	12.2565	8.375	6.976	5.6566	4.4517	3.6364	3.1818	3.3333	1.8182	2.1212
U6	4.424	2.7734	1.8189	1.1111	1.0111	0.7576	0.6061	0.303	0.303	0.6061
U7	10.7135	5.8124	3.4571	2.1212	1.7182	0.9091	0.1515	1.0606	0	0
U8	2.5162	1.5172	0.9697	0.5051	0.6071	0.303	0.1515	0.4545	0	0
U9	0.6293	0.2601	0.303	0.202	0.101	0.303	0	0.1515	0	0
U10	3.2271	2.5129	1.5155	1.3131	1.0111	0.303	0	0.9091	0	0
U11	0.5479	0.4773	0.3636	0.303	0.202	0	0.1515	0	0	0
U12	0.8523	0.52	0.3636	0.303	0.101	0.1515	0.1515	0.1515	0	0
U13	0.6895	0.3906	0.2424	0.202	0.303	0.1515	0.1515	0.1515	0	0
U14	4.1787	3.3394	2.9719	2.5253	2.0212	2.5758	2.2727	1.8182	1.8182	1.5152
U15	0.8523	0.4777	0.425	0.202	0.101	0	0	0	0	0
U16	1.2375	1.3446	0.5455	0.8081	0.9091	0.4545	0.4545	0.6061	0.303	0.6061
U17	0.6697	0.3469	0.4848	0.202	0.202	0.303	0.4545	0.7576	0.6061	0.303
U18	0.3449	0.217	0.1818	0	0.101	0	0	0	0	0.303
U19	1.9695	1.2575	0.6064	0.8081	0.4071	0.1515	0.1515	0	0	0.303
U20	0.9131	0.78	0.6064	0.6061	0.101	0.303	0.6061	0.4545	0.303	0.303
U21	0.2841	0.0866	0.3034	0.101	0	0	0.303	0	0	0
U22	0.588	0.2168	0.1212	0.101	0.202	0.1515	0.1515	0.1515	0.303	0
U23	0.8723	0.6497	0.6064	0.303	0	0	0.1515	0	0.303	0.303
U24	0.5677	0.2599	0.1212	0.202	0	0	0	0.1515	0	0
U25	0.548	0.6936	0.4246	0.202	0.202	0.303	0.1515	0.303	0.303	0
U26	1.2779	1.0839	0.8489	0.9091	0.7071	0.7576	0.303	0.4545	0	0.303
U27	0.6088	0.2607	0.4246	0.303	0	0	0.1515	0	0	0
U28	1.9279	1.3874	1.0913	0.5051	1.0101	0.4545	0.4545	0.303	0	0
U29	0.3245	0.2168	0.1216	0	0	0	0	0	0	0
U30	0.3455	0.1733	0.1212	0.404	0.101	0.1515	0.303	0	0	0
U31	1.6637	0.9533	0.8485	0.6061	0.303	0.4545	0.6061	0.4545	0.6061	0.303
U32	1.6434	1.9966	0.9701	1.3131	1.3142	0.6061	0.9091	1.0606	1.8182	1.8182
U33	0.2433	0.0433	0	0	0	0.1515	0	0	0	0
Average	2.0439	1.3734	1.0016	0.7775	0.6004	0.4454	0.404	0.4408	0.2847	0.2755
SD	2.7679	1.7457	1.3539	1.0572	0.8643	0.7328	0.6543	0.6624	0.5219	0.5241

ตารางที่ ก-2 ตารางแสดงค่า False Negative (ค่าเฉลี่ยร้อยละ) ในการจำแนกผู้ใช้โดยใช้การวัดระยะทางแบบ R_2+A_2 โดยใช้ข้อความฝึกยาว 1000 ตัวอักษร

User	Length of Test Sample (Characters)									
	100	200	300	400	500	600	700	800	900	1000
U1	2.1105	1.2579	0.9708	0.8081	0.8091	0.1515	0	0.4545	0	0
U2	1.2781	0.7805	0.4242	0.6061	0.303	0	0.1515	0.6061	0	0
U3	0.7714	0.5206	0.2428	0.202	0	0	0	0	0	0
U4	2.5165	2.1675	1.7583	1.2121	1.1111	1.2121	1.0606	0.9091	0.6061	0.303
U5	2.7185	2.8615	2.9106	3.0303	2.7283	3.0303	3.0303	3.0303	3.0303	2.7273
U6	2.4348	1.7792	1.6378	1.7172	1.1132	1.0606	0.9091	0.7576	0.303	0.6061
U7	2.3335	1.4308	1.1526	0.8081	0.202	0.1515	0.1515	0.6061	0.303	0.303
U8	2.5572	1.9524	1.5769	1.5152	1.1121	0.9091	0.6061	1.0606	0.6061	0
U9	2.0495	1.127	0.5462	0.5051	0.202	0	0	0	0	0
U10	1.1972	0.5206	0.2424	0	0.2031	0	0	0.1515	0	0
U11	1.5834	0.7372	0.5458	0.101	0.101	0	0	0	0	0
U12	0.9332	0.2601	0.2428	0	0	0	0	0	0	0
U13	2.0494	1.4305	1.0909	1.4141	0.9091	1.0606	1.0606	0.7576	0.6061	0.303
U14	2.1104	1.2151	0.85	0.7071	0.9101	0.4545	0.6061	0.4545	0.9091	0.6061
U15	2.2525	1.5615	1.031	0.6061	0.4051	0.303	0.1515	0.1515	0.303	0.6061
U16	2.1913	1.3446	1.3943	0.6061	0.404	0.6061	0.1515	0.1515	0	0
U17	1.076	0.3902	0.1818	0.303	0.101	0	0	0	0	0
U18	2.577	1.7775	1.7579	1.3131	1.3152	1.0606	1.0606	1.0606	0.9091	1.5152
U19	2.333	1.5611	1.2129	0.6061	0.6061	0.303	0.1515	0.303	0	0
U20	2.2722	1.6484	1.2125	0.7071	0.6061	0.1515	0.6061	0.6061	0.6061	0.303
U21	2.638	1.9083	1.4553	1.3131	0.8081	0.9091	0.4545	0.4545	0	0.303
U22	2.5567	1.9953	1.1523	1.1111	0.8081	0.4545	0.6061	0.4545	0	0
U23	2.7801	2.298	1.8795	1.6162	0.9091	0.4545	0.6061	0.6061	0.303	0.303
U24	1.8871	0.9548	0.364	0.101	0.202	0	0	0	0	0
U25	2.2926	1.6911	1.0913	0.6061	0.404	0.1515	0	0.303	0	0
U26	1.3185	0.7365	0.1818	0.101	0.303	0.1515	0.1515	0	0	0
U27	1.8672	1.1286	0.6681	0.5051	0.2041	0.303	0.1515	0.1515	0	0
U28	1.4199	0.693	0.6061	0.303	0.404	0.1515	0.1515	0.1515	0	0
U29	2.2322	1.4323	0.9712	0.6061	0.6071	0.303	0.1515	0.303	0	0
U30	2.7189	1.9957	1.6981	1.3131	1.0101	0.6061	0.9091	0.4545	0.9091	0.9091
U31	2.1102	1.5616	0.7882	0.9091	0.8112	0.7576	0.4545	0.303	0	0.303
U32	2.1512	1.2575	0.667	0.202	0.202	0	0	0.303	0	0
U33	2.1308	1.3437	0.5458	0.202	0	0	0	0	0	0
Average	2.0439	1.3734	1.0016	0.7775	0.6004	0.4454	0.404	0.4408	0.2847	0.2755
SD	0.5417	0.5865	0.6075	0.6313	0.5313	0.5921	0.5846	0.5525	0.5727	0.545

ตารางที่ ก-3 ตารางแสดงค่า False Positive (ค่าเฉลี่ยร้อยละ) ในการจำแนกผู้ใช้โดยใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึม โดยใช้ข้อความฝึกยาว 100 ตัวอักษร

User	Length of Test Sample (Characters)									
	100	200	300	400	500	600	700	800	900	1000
U1	0.5859	0.3463	0.5455	0.101	0.303	0	0.1515	0.1515	0	0
U2	1.4545	1.6017	1.3333	0.9091	1.6162	1.3636	1.9697	0.9091	1.5152	1.5152
U3	0.3434	0.1732	0.0606	0	0	0	0	0	0	0
U4	0.9293	0.3463	0.1212	0.202	0.202	0.1515	0.303	0.1515	0	0
U5	0.101	0.0433	0	0	0	0	0	0	0	0
U6	0.2424	0.2597	0	0.101	0	0.1515	0.1515	0.1515	0	0
U7	1.5556	1.4719	0.8485	0.6061	0.5051	0	0.1515	0.4545	0.6061	0.9091
U8	1.5152	0.5628	0.303	0.101	0.101	0	0	0	0	0
U9	0.7475	0.4329	0.4848	0.6061	0.303	0.303	0.7576	0.7576	0.9091	0.9091
U10	0.7677	0.5195	0.1818	0.101	0.202	0.1515	0	0.1515	0.6061	0.6061
U11	1.3333	0.9091	0.7879	0.6061	0.202	0.1515	0.303	0.1515	0	0
U12	1.4343	1.1688	0.9091	0.7071	0.202	0.303	0.4545	0.4545	1.2121	0.9091
U13	0.7273	0.303	0.4242	0.202	0.101	0	0	0	0	0
U14	0.5859	0.4329	0.4848	0.202	0	0.1515	0.1515	0.1515	0	0
U15	1.3131	0.9091	0.303	0.8081	0.404	0.6061	0.303	0.4545	0	0
U16	1.8788	1.1255	0.6667	1.0101	0.8081	0.6061	0.4545	0.303	0.303	0
U17	1.0101	1.2121	1.2121	1.0101	0.8081	0.4545	1.6667	0.9091	0.6061	0.9091
U18	1.4747	0.8225	0.6667	1.1111	1.5152	1.0606	0.9091	0.7576	0.303	0.6061
U19	0.9697	0.3463	0.3636	0.101	0.202	0.1515	0	0	0	0
U20	1.4545	1.039	0.6667	0.404	0.5051	0.6061	0.303	0.7576	0.9091	0.6061
U21	1.4747	1.2987	0.6667	0.6061	0.5051	0.4545	0.6061	0.303	0	0
U22	1.6566	1.2554	1.0909	0.9091	0.9091	0.7576	1.5152	0.7576	1.2121	0.6061
U23	0.7879	0.303	0.0606	0.101	0	0	0	0	0	0
U24	1.3131	0.5195	0.4242	0.7071	0.303	0.7576	0.9091	0.7576	0.6061	0.303
U25	1.3535	1.2987	1.1515	1.0101	1.2121	0.9091	1.2121	1.0606	0.9091	0.6061
U26	2.0606	1.4719	1.6364	1.1111	0.9091	1.0606	1.0606	0.7576	0	1.5152
U27	1.1313	0.8225	0.3636	0.5051	0.202	0.303	0.1515	0.1515	0	0
U28	1.8384	1.5584	1.2727	0.7071	1.0101	0.9091	0.9091	0.7576	0.6061	0.6061
U29	1.1919	0.8225	1.2121	0.6061	0.7071	0.6061	0.303	0.6061	0.6061	0.303
U30	1.0101	0.6061	0.1818	0	0.202	0	0.1515	0.303	0	0
U31	0.7879	0.6061	0.6667	0.7071	0.5051	0.6061	0.6061	0.6061	0.303	0.303
U32	1.1313	0.3463	0.303	0.101	0	0	0	0	0	0
U33	1.2525	0.8225	0.4848	0.9091	0.7071	0.1515	0	0.303	0.303	0
Average	1.1338	0.7805	0.6024	0.5112	0.4591	0.3857	0.4683	0.3949	0.3489	0.3398
SD	0.4632	0.4418	0.4241	0.3688	0.4322	0.3769	0.5254	0.3272	0.4358	0.4461

ตารางที่ ก-4 ตารางแสดงค่า False Negative (ค่าเฉลี่ยร้อยละ) ในการจำแนกผู้ใช้โดยใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของค่าลอการิทึม โดยใช้ข้อความฝึกยาว 100 ตัวอักษร

User	Length of Test Sample (Characters)									
	100	200	300	400	500	600	700	800	900	1000
U1	0.8485	0.7359	0.4848	0.6061	0.5051	0.303	0.6061	0.1515	0	0.303
U2	1.6162	0.9957	0.6667	0.7071	0.404	0.303	0.1515	0.303	0.303	0
U3	0.2424	0.0866	0.0606	0.202	0	0.1515	0.1515	0	0	0
U4	1.1919	0.8225	0.7273	0.8081	0.5051	0.6061	0.9091	0.9091	0.6061	0.6061
U5	0.0404	0	0	0	0	0	0	0	0	0
U6	0.3636	0.1732	0.0606	0	0	0	0	0	0	0
U7	1.0707	0.6926	0.7273	0.6061	0.5051	0.4545	0.6061	0.6061	0.6061	0.6061
U8	1.3737	1.2554	1.3333	1.4141	1.2121	0.9091	1.0606	0.6061	1.2121	1.2121
U9	0.5859	0.303	0.1818	0.101	0.101	0	0	0.1515	0	0
U10	0.8081	0.5628	0.4242	0.404	0	0.1515	0.1515	0.1515	0	0.303
U11	1.5152	1.039	0.7273	0.8081	0.6061	0.1515	0.4545	0.303	0.9091	0.6061
U12	1.5354	1.5584	1.0303	0.7071	1.0101	0.4545	1.0606	0.7576	0.9091	0.9091
U13	0.7475	0.4329	0.3636	0.101	0.101	0.1515	0.303	0	0	0
U14	0.5253	0.1732	0.1212	0	0	0	0	0	0	0
U15	1.2525	0.6494	0.6061	0.303	0.5051	0	0	0.1515	0	0.303
U16	1.5354	1.2987	1.3939	0.9091	0.7071	0.9091	0.6061	0.4545	1.2121	0.9091
U17	1.0909	0.3896	0.2424	0	0	0	0	0	0	0
U18	1.6364	1.0823	0.6061	0.404	0.5051	0.4545	0.4545	0.4545	0.303	0.6061
U19	1.1717	0.7359	0.7879	0.5051	0.6061	0.1515	0.303	0.303	0	0
U20	1.697	1.2554	0.7879	0.5051	0.404	0.6061	0.9091	0.9091	0.303	0.303
U21	1.4949	1.2554	0.7273	0.7071	0.404	0.6061	0.6061	0.6061	0.303	0
U22	2.101	1.6883	1.4545	1.0101	0.9091	1.0606	0.9091	0.9091	0.6061	0.9091
U23	0.6869	0.4762	0.303	0.303	0.5051	0.4545	0.303	0.303	0.303	0.303
U24	1.2525	0.8658	0.7879	0.7071	1.0101	0.6061	0.7576	0.6061	0.6061	0.6061
U25	1.3535	1.0823	0.6061	0.6061	0.6061	0.7576	1.0606	0.7576	0.6061	0.303
U26	1.8788	1.2554	1.2121	1.1111	1.2121	0.6061	1.2121	1.0606	0.6061	0.9091
U27	0.9495	0.6061	0.4242	0.303	0.303	0	0	0	0.6061	0.303
U28	1.4343	0.9091	0.8485	0.7071	0.5051	0.6061	0.6061	0.6061	0.303	0
U29	1.3131	0.6061	0.4242	0.202	0.303	0.4545	0.1515	0.1515	0	0
U30	1.3333	0.9957	0.4242	0.5051	0.404	0.4545	0.4545	0.4545	0	0
U31	0.5051	0.1299	0.0606	0	0	0	0	0	0	0
U32	0.7273	0.5195	0.4242	0.5051	0.404	0.303	0.6061	0.4545	0.303	0.303
U33	1.5354	1.1255	0.8485	1.1111	0.9091	1.0606	1.0606	0.9091	0.9091	0.9091
Average	1.1338	0.7805	0.6024	0.5112	0.4591	0.3857	0.4683	0.3949	0.3489	0.3398
SD	0.4844	0.4312	0.3827	0.362	0.3518	0.3211	0.3883	0.3293	0.3739	0.3636

ตารางที่ ก-5 ตารางแสดงค่า False Positive (ค่าเฉลี่ยร้อยละ) ในการจำแนกผู้ใช้โดยใช้ค่าเฉลี่ยความน่าจะเป็นจากฮิสโตแกรม โดยใช้ข้อความฝึกยาว 100 ตัวอักษร

User	Length of Test Sample (Characters)									
	100	200	300	400	500	600	700	800	900	1000
U1	0.1616	0.1299	0.1818	0.101	0.202	0.1515	0	0	0	0
U2	0.6263	0.6926	0.8485	1.1111	1.6162	2.1212	2.4242	1.6667	2.1212	1.8182
U3	0.303	0.2597	0.2424	0.202	0.303	0.303	0.4545	0.303	0.6061	0.303
U4	0.404	0.4329	0.2424	0.404	0.303	0.1515	0.1515	0.1515	0.303	0
U5	2.4444	0.5628	0.303	0.101	0.101	0.1515	0.1515	0	0	0
U6	0.3636	0.2165	0.2424	0.303	0.101	0.1515	0.1515	0	0	0
U7	1.3333	0.5628	0.6667	0.303	0.303	0.1515	0.1515	0.303	0	0.6061
U8	1.2525	0.6926	0.4848	0.5051	0.303	0.1515	0.1515	0.303	0	0
U9	0.6869	0.303	0.1818	0.101	0.202	0.1515	0.1515	0	0	0
U10	0.6061	0.1732	0.0606	0.101	0	0.1515	0	0	0	0
U11	1.1919	0.5195	0.4848	0.101	0.101	0	0	0	0	0
U12	1.1111	0.5628	0.3636	0.101	0.303	0	0	0.1515	0.303	0.303
U13	0.6667	0.1732	0.1212	0.101	0	0	0	0	0	0
U14	0.5657	0.1732	0.2424	0.101	0.202	0.303	0.303	0.4545	0.6061	0.6061
U15	0.8485	0.5195	0.3636	0.303	0.404	0.1515	0.4545	0.1515	0	0
U16	1.1313	0.5195	0.303	0.202	0.101	0.4545	0.303	0.1515	0.303	0.303
U17	1.0707	1.3853	1.2121	2.0202	1.7172	0.9091	1.2121	1.3636	1.8182	1.5152
U18	1.0909	1.2121	0.9091	1.3131	0.6061	0.9091	1.2121	0.9091	0.6061	0.6061
U19	0.6869	0.3463	0.1818	0.303	0.202	0	0	0.303	0.303	0.6061
U20	1.4747	0.4329	0.5455	0.101	0.303	0.1515	0.1515	0.4545	0.303	0.303
U21	1.1717	0.7792	1.0303	0.404	0.9091	0.9091	0.6061	0.4545	0.303	0.303
U22	1.899	0.8658	0.7879	0.5051	0.5051	0.1515	0.4545	0.303	0.303	0.303
U23	1.0909	0.4329	0.1818	0.202	0.202	0.1515	0	0.1515	0	0
U24	1.0101	1.2121	1.4545	1.1111	0.7071	0.7576	0.7576	0.9091	0.6061	0.6061
U25	1.3535	1.0823	0.7879	0.8081	0.8081	0.9091	0.6061	0.303	0.6061	0.6061
U26	0.8081	0.8225	0.8485	0.7071	1.0101	0.4545	0.9091	1.2121	0.6061	1.8182
U27	0.8687	0.4329	0.303	0.202	0.404	0	0.1515	0.1515	0	0
U28	1.0101	0.6494	0.4242	0.5051	0.202	0.4545	0.6061	0.303	0.6061	0.6061
U29	0.9697	0.8658	0.4242	0.303	0.5051	0.4545	0.7576	0.7576	0.9091	0.9091
U30	1.2727	0.9091	0.8485	0.5051	0.5051	0.4545	0.4545	0.7576	0.6061	0.6061
U31	0.5253	0.2597	0.0606	0.101	0	0	0	0	0	0
U32	0.7273	0.5195	0.0606	0.101	0.101	0	0	0	0	0
U33	1.3737	0.4329	0.0606	0.101	0.101	0.1515	0	0	0	0
Average	0.9728	0.5798	0.4683	0.4071	0.404	0.3443	0.3857	0.3627	0.3581	0.3857
SD	0.4554	0.3206	0.3544	0.4282	0.4079	0.4257	0.4978	0.4252	0.4918	0.502

ตารางที่ ก-6 ตารางแสดงค่า False Negative (ค่าเฉลี่ยร้อยละ) ในการจำแนกผู้ใช้โดยใช้ค่าเฉลี่ยความน่าจะเป็นจากฮิสโตแกรม โดยใช้ข้อความฝึกยาว 100 ตัวอักษร

User	Length of Test Sample (Characters)									
	100	200	300	400	500	600	700	800	900	1000
U1	1.3737	0.6061	0.4242	0.5051	0.303	0.303	0.1515	0.1515	0.303	0.303
U2	0.7677	0.2165	0.0606	0.101	0	0	0	0	0	0
U3	0.1818	0	0	0	0	0	0	0	0	0
U4	0.3232	0.1299	0.1212	0	0.101	0	0.1515	0.1515	0	0
U5	0.1414	0.2597	0.303	0.404	0.6061	0.7576	0.7576	0.9091	0.9091	1.5152
U6	0.6465	0.303	0.1818	0.101	0.202	0	0	0.1515	0	0
U7	1.2929	1.1688	1.0303	0.8081	1.0101	1.0606	1.0606	0.9091	0.9091	0.6061
U8	1.0909	0.7792	0.7273	0.6061	0.5051	0.4545	0.303	0.1515	0.303	0.6061
U9	1.9596	1.4286	1.4545	1.2121	1.0101	0.7576	1.2121	1.3636	1.5152	1.2121
U10	0.6869	0.303	0.2424	0.303	0	0	0	0	0	0
U11	1.0303	0.6926	0.4242	0.101	0.202	0	0	0	0	0.303
U12	0.6667	0.3896	0.1818	0.303	0.101	0.1515	0.1515	0	0	0
U13	0.7677	0.6061	0.4242	0.7071	0.5051	0.4545	0.6061	0.6061	0.6061	0.303
U14	0.6465	0.1732	0.1818	0	0	0	0	0	0	0
U15	1.1717	0.5195	0.4242	0.101	0.202	0	0.1515	0.1515	0	0
U16	0.7879	0.5195	0.5455	0.303	0.202	0.303	0.303	0.303	0.6061	0.303
U17	0.7273	0.1732	0.1212	0	0.101	0	0	0	0	0
U18	1.3131	0.5195	0.3636	0.303	0.101	0.1515	0.303	0.1515	0.303	0.303
U19	1.2525	0.9957	0.7273	0.6061	0.9091	0.6061	0.7576	0.7576	0.6061	0.6061
U20	1.4747	1.0823	0.9091	0.8081	0.8081	0.6061	0.9091	0.9091	0.6061	0.6061
U21	1.5758	1.1688	0.9697	0.8081	1.1111	0.7576	0.7576	0.6061	0.9091	1.2121
U22	1.4545	0.9091	0.9091	0.7071	0.8081	0.6061	0.6061	0.4545	0.303	0.303
U23	1.0101	0.8658	0.8485	0.9091	0.9091	0.7576	0.9091	0.7576	0.6061	0.9091
U24	0.8687	0.2165	0.1212	0.101	0	0	0.1515	0	0	0
U25	1.2525	0.8225	0.6667	0.5051	0.6061	0.4545	0.4545	0.4545	0.303	0.303
U26	1.1313	0.6494	0.2424	0.404	0.404	0.303	0.6061	0.303	0.303	0.303
U27	0.9697	0.3463	0.303	0	0.202	0	0	0	0	0
U28	0.8283	0.3896	0.2424	0.101	0	0	0.1515	0.1515	0	0
U29	1.0101	0.4762	0.1818	0.404	0.404	0.4545	0.6061	0.6061	0.303	0.6061
U30	1.1515	0.7359	0.4848	0.404	0.6061	0.4545	0.1515	0.303	0.303	0.303
U31	0.4646	0	0	0	0	0	0	0	0	0
U32	1.0303	0.7359	0.7273	0.8081	0.7071	1.0606	0.6061	0.7576	0.9091	0.9091
U33	1.0505	0.9524	0.9091	1.0101	0.7071	0.9091	0.9091	0.9091	1.2121	1.2121
Average	0.9728	0.5798	0.4683	0.4071	0.404	0.3443	0.3857	0.3627	0.3581	0.3857
SD	0.3923	0.3571	0.3497	0.3387	0.3534	0.3444	0.3637	0.3691	0.4049	0.4303

ประวัติผู้เขียนวิทยานิพนธ์

นายวรุตม์ โรจน์รุ่งวศินกุล เกิดเมื่อวันที่ 1 ธันวาคม พ.ศ. 2529 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาปริญญาวิศวกรรมศาสตร์บัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2551 (เกียรตินิยมอันดับ 1) และเข้าศึกษาในหลักสูตรวิศวกรรมศาสตร์มหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2552