วิทยานิพนธ์

ออนไลน์อัลกอริทึมสำหรับแก้ปัญหาการทำนายแบบมัลติคลาสที่แต่ละเหตุการณ์มีน้ำหนักความสำคัญไม่เท่ากัน

ONLINE LEARNING ALGORITHM FOR MULTICLASS IMPORTANCE WEIGHTED PREDICTION PROBLEM

นายอดิศักดิ์ สุภีสุน

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2551

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง ออนไลน์อัลกอริทึมสำหรับแก้ปัญหาการทำนายแบบมัลติคลาสที่แต่ละเหตุการณ์มีน้ำหนักความสำคัญไม่เท่ากัน

Online Learning Algorithm for Multiclass Importance Weighted Prediction Problem

นามผู้วิจัย นายอดิศักดิ์ สุภีสุน

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(ผู้ช่วยศาสตราจารย์จิตร์ทัศน์ ฝักเจริญผล, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(รองศาสตราจารย์กฤษณะ ไวยมัย, D.U.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 29 **เดือน** พฤษภาคม **พ.ศ.** 2551

วิทยานิพนธ์

เรื่อง

ออนไลน์อัลกอริทึมสำหรับแก้ปัญหาการทำนายแบบมัลติคลาสที่แต่ละเหตุการณ์มีน้ำหนักความสำคัญไม่เท่ากัน

Online Learning Algorithm for Multiclass Importance Weighted Prediction Problem

โดย

นายอดิศักดิ์ สุภีสุน

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

พ.ศ. 2551

อดิศักดิ์ สุภีสุน 2551: ออนไลน์อัลกอริทึมสำหรับแก้ปัญหาการทำนายแบบมัลติคลาสที่แต่ละเหตุการณ์มีน้ำหนักความสำคัญไม่เท่ากัน ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์จิตร์ทัศน์ ฝักเจริญผล, Ph.D. 48 หน้า

ในวิทยานิพนธ์ฉบับนี้ เรานำเสนอการแก้ปัญหาการเรียนรู้ออนไลน์ที่ตัวอย่างมีน้ำหนักความสำคัญสองวิธีการ วิธีแรกใช้การเรียนรู้แบบลดรูป กล่าวคือ เราแสดงวิธีการลดรูปปัญหาการเรียนรู้ออนไลน์แบบมีน้ำหนักไปยังปัญหาการเรียนรู้ออนไลน์มาตรฐาน โดยการลดรูปดังกล่าวทำให้ขีดจำกัดด้านบนของความผิดพลาดของอัลกอริทึมเพิ่มขึ้นไม่เกิน c_{max} เท่า เมื่อ c_{max} คือค่าน้ำหนักที่มากที่สุด ในส่วนของวิธีการที่สอง เรานำเสนออัลกอริทึมที่เป็นการปรับปรุงอัลกอริทึมเพอร์เซพตรอนอย่างง่าย และสามารถทำงานได้ในสถานการณ์เดียวกับเพอร์เซพตรอน อัลกอริทึมที่สองนี้สามารถใช้กับปัญหาการเรียนรู้แบบหลายประเภทได้ด้วย โดยใช้วิธีลดรูปจากปัญหาการเรียนรู้หลายประเภทไปเป็นปัญหาการเรียนรู้สองประเภท ซึ่งเราได้พิสูจน์ว่า การลดรูปการเรียนรู้ดังกล่าว 2 วิธีการที่แตกต่างกันนั้นสมมูลกัน

จากวิธีการที่เรานำเสนอ 2 วิธีข้างต้น วิธีที่สองมีการรับประกันเชิงทฤษฎีที่ดีกว่าวิธีแรกในหลายๆ กรณี รวมทั้งให้ผลการทดลองที่ดีกว่า อย่างไรก็ตาม วิธีแรกเป็นวิธีที่มีความเป็นทั่วไปมากกว่า เรายังได้ทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพของอัลกอริทึมที่ได้นำเสนอกับชุดข้อมูลทดสอบจริงและชุดข้อมูลสังเคราะห์

อดิศักดิ์ สุภีสุน จิตร์ทัศน์ ฝักเจริญผล 26 / พฤษภาคม 2551

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Adisak Supeesun 2008: Online Learning Algorithm for Multiclass Importance Weighted Prediction Problem. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Assistant Professor Jittat Fakcharoenphol, Ph.D. 48 pages.

This thesis considers an online learning problem with importance weighted examples. We present two approaches. The first one uses learning reduction, i.e., we show how to reduce online importance weighted learning problems to standard online learning problems with a factor $c_{\max}$ overhead on mistake bounds, $c_{\max}$ is the maximum weight of examples. We also present another online algorithm based on perceptron learning algorithm. The second algorithm is a simple modification of perceptron algorithm and work in this same setting and by modifying the proof of gentile we can proof weighted loss bound of this algorithm. For the multiclass setting, the second algorithm can work in the setting by using multiclass-binary learning reduction. We also show two different method, the multi-vector method and its particular all-pair variant, for reducing multiclass problems to binary problems are equivalent.

While the second algorithm performs better on experiments and has a better theoretical guarantee in many cases, the first one is more general. We also perform experiments on real data sets and synthetic data sets.

Adisak Supeesun — Student's signature

Jittat Fakcharoenphol — Thesis Advisor's signature

26 / May / 2008

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลงได้ด้วยความช่วยเหลือจากผู้ช่วยศาสตราจารย์จิตร์ทัศน์ ฝักเจริญผล ประธานกรรมการที่ปรึกษา ผู้ให้คำปรึกษาและแนวทางการทำวิจัย ตลอดจนข้อเสนอแนะที่เป็นประโยชน์ต่องานวิจัยนี้ รองศาสตราจารย์กฤษณะ ไวยมัย กรรมการที่ปรึกษาร่วม ที่กรุณาให้คำปรึกษา และข้อเสนอแนะที่มีคุณค่าเพื่อให้วิทยานิพนธ์นี้สมบูรณ์ยิ่งขึ้น

ข้าพเจ้าขอขอบคุณเพื่อนๆ นิสิตปริญญาโท, สมาชิกกลุ่มวิจัยเชิงทฤษฎีทุกท่านที่ช่วยเหลือและให้คำปรึกษาแก่ข้าพเจ้าจนสามารถทำงานวิจัยชิ้นนี้สำเร็จ ขอขอบคุณเจ้าหน้าที่ธุรการ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาตร์ที่คอยอำนวยความสะดวกด้านประสานงาน และการดำเนินงานต่างๆ

อดิศักดิ์ สุภีสุน

พฤษภาคม 2551

สารบัญ

หน้า

สารบัญ	(1)
สารบัญตาราง	(2)
สารบัญภาพ	(3)
คำนำ	1
วัตถุประสงค์	4
การตรวจเอกสาร	5
วิธีการ	15
ผลการวิจัย	21
สรุป	46
เอกสารและสิ่งอ้างอิง	47

สารบัญตาราง

ตารางที่ **หน้า**

1 ผลการเปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึมแบบลดรูป, อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ และอัลกอริทึมเพอร์เซพตรอน กับชุดข้อมูลแบบสองประเภท 25

2 ผลการเปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึมแบบลดรูป, อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ และอัลกอริทึมเพอร์เซพตรอน เมื่อทำการลดรูปจากปัญหาการทำนายหลายประเภทเป็นปัญหาการทำนายหลายประเภทด้วยวิธีการหลายเวคเตอร์ กับชุดข้อมูลแบบหลายประเภท 40

3 ผลการเปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึมแบบลดรูป, อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ และอัลกอริทึมเพอร์เซพตรอน เมื่อทำการลดรูปจากปัญหาการทำนายหลายประเภทเป็นปัญหาการทำนายหลายประเภทด้วยวิธีการทุกคู่ กับชุดข้อมูลแบบหลายประเภท 40

สารบัญภาพ

ภาพที่		หน้า
1	ตัวอย่างการทำนายด้วยเพอร์เซพตรอน	8
2	อัลกอริทึมแบบลดรูป	16
3	อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ	17
4	เวคเตอร์น้ำหนักของวิธีการทุกคู่	19
5	การกระจายตัวของข้อมูลสังเคราะห์ชุดที่ 1	23
6	การกระจายตัวของข้อมูลสังเคราะห์ชุดที่ 2	24
7	การกระจายตัวของข้อมูลสังเคราะห์ชุดที่ 3	24

ออนไลน์อัลกอริทึมสำหรับแก้ปัญหาการทำนายแบบมัลติคลาสที่แต่ละเหตุการณ์มีน้ำหนักความสำคัญไม่เท่ากัน

Online Learning Algorithm for Multiclass Importance Weighted Prediction Problem

คำนำ

ในปัจจุบันมีการนำเอาวิธีการทางด้านการเรียนรู้ด้วยเครื่องจักร (machine learning) ไปประยุกต์ใช้กับการแก้ปัญหาของงานหลาย ๆประเภท กระบวนการทั่วไปจะเริ่มจากการหาตัวอย่างที่สุ่มจากการกระจาย (distribution) เดียวกับสถานการณ์ที่ใช้จริง แล้วนำตัวอย่างไปใช้เพื่อสร้างตัวทำนายที่มีคุณภาพ จากนั้นนำตัวทำนายที่ได้ไปใช้จริง ลักษณะของปัญหาดังกล่าวจัดเป็น*ปัญหาการเรียนรู้แบบออฟไลน์* กล่าวคือ ในการฝึกสอนตัวทำนายนั้น อัลกอริทึมจะได้ข้อมูลสำหรับการฝึกสอนทั้งหมดมาก่อน งานวิจัยนี้สนใจ*ปัญหาการเรียนรู้แบบออนไลน์* กล่าวโดยย่อคือ อัลกอริทึมจะสร้างตัวทำนายพร้อมๆ กับที่ได้รับข้อมูลทดสอบ ปัญหาลักษณะนี้เกิดขึ้นในกรณีที่ผู้สร้างตัวทำนายไม่สามารถหาตัวอย่างการเรียนรู้ได้เนื่องจากไม่มีแหล่งข้อมูลหรือไม่ทราบการกระจายที่แท้จริงหรือใกล้เคียงกับที่ต้องการ หรือในกรณีที่การกระจายของตัวอย่างและคำตอบเปลี่ยนแปลงไม่แน่นอน

ในงานวิจัยนี้สนใจปัญหาการเรียนรู้แบบออนไลน์ในกรณีที่*ตัวอย่างมีน้ำหนักความสำคัญ* (Importance weighted online learning) ในปัญหานี้อัลกอริทึมจะทำงานเป็นรอบๆ ในแต่ละรอบอัลกอริทึมจะได้รับตัวอย่าง x และจะต้องทำนายประเภทของ x ก่อนที่จะได้รับคำเฉลย y และน้ำหนักความสำคัญ c ของ x ถ้าอัลกอริทึมทำนายผิด อัลกอริทึมจะถูกปรับด้วยค่า c ประสิทธิภาพของอัลกอริทึมจะวัดจากผลรวมของการถูกปรับ เป้าหมายก็คือต้องการให้ผลรวมดังกล่าวน้อยที่สุด ส่วนปัญหาการเรียนรู้ที่ตัวอย่างมีน้ำหนักความสำคัญนั้นมีปรากฏขึ้นในการนำการเรียนรู้ด้วยเครื่องจักรไปใช้ในสถานการณ์จริง โดยมากมักพบในการนำไปประยุกต์ใช้ในทางธุรกิจที่ค่าใช้จ่ายกับผลตอบแทนของลูกค้าแต่ละรายไม่เท่ากัน สำหรับสถานการณ์แบบออนไลน์มักพบในการเสนอขายสินค้าให้กับลูกค้า เป็นต้น

ปัญหาการเรียนรู้ทั้งแบบที่ตัวอย่างไม่มีน้ำหนักความสำคัญและแบบที่ตัวอย่างมีน้ำหนักความสำคัญ ที่กล่าวมายังสามารถแบ่งได้อีก 2 กลุ่ม ได้แก่ ปัญหาการเรียนรู้สองประเภท โดยปัญหานี้จะมีคำตอบอยู่สองทางเลือก และปัญหาการเรียนรู้หลายประเภท ซึ่งมีคำตอบมีมากกว่าสองทางเลือก สำหรับปัญหาหลักของงานวิจัยนี้ คือ ปัญหาการเรียนรู้ออนไลน์แบบหลายประเภทที่ตัวอย่างมีน้ำหนักความสำคัญ โดยงานวิจัยนี้เสนอวิธีการสร้างอัลกอริทึมสำหรับปัญหาดังกล่าว 2 วิธีการ

วิธีแรก สร้างอัลกอริทึมสำหรับปัญหาการเรียนรู้แบบออนไลน์ที่มีน้ำหนักความสำคัญจากอัลกอริทึมแบบออนไลน์มาตรฐาน (นั่นคืออัลกอริทึมที่ใช้ในงานตัวอย่างไม่มีน้ำหนักความสำคัญ) วิธีนี้จะสามารถสร้างอัลกอริทึมสำหรับเซตของตัวอย่างและเซตของประเภทใดๆ ก็ได้ (ทำนายแบบสองประเภทหรือหลายประเภทก็ได้) ขึ้นกับอัลกอริทึมออนไลน์ที่ใช้เป็นฐาน แต่เซตของน้ำหนักของตัวอย่างจะต้องเป็นจำนวนเต็ม และได้พิสูจน์ประสิทธิภาพของอัลกอริทึมที่สร้างได้เทียบกับอัลกอริทึมฐาน เทคนิคที่ใช้ในการพิสูจน์คือการลดรูปการเรียนรู้ (learning reduction) กล่าวคือ เราได้พิสูจน์ว่าอัลกอริทึมที่ได้จะสามารถรับประกันความผิดพลาดได้ไม่แย่กว่า $c_{\max}$ เท่าของความผิดพลาดของอัลกอริทึมฐาน เมื่อ $c_{\max}$ คือค่าน้ำหนักที่มากที่สุดของข้อมูล

วิธีที่สอง เป็นการปรับปรุงอัลกอริทึมเพอร์เซพตรอน (Perceptron) เพื่อให้ทำงานกับข้อมูลที่มีน้ำหนักความสำคัญ เช่นเดียวกับอัลกอริทึมเพอร์เซพตรอน อัลกอริทึมนี้ทำงานได้กับตัวอย่างบน $\Re^n$ และสามารถทำนายแบบสองประเภท อย่างไรก็ตามในกรณีนี้ เราสามารถพิสูจน์ประสิทธิภาพการทำงานได้ดีกว่าการใช้วิธีแรก กล่าวคือ ถ้า $c_{\max}$ และ $c_{\min}$ คือค่าน้ำหนักมากที่สุดและน้อยที่สุดของตัวอย่าง (ไม่จำเป็นต้องเป็นจำนวนเต็มก็ได้) ให้ θ เป็นอัตราส่วน $c_{\max}/c_{\min}$ เราสามารถแสดงได้ว่าความผิดพลาดที่ได้จะไม่เกิน $\theta L + c_{\max}\theta^2 C + \theta^2\sqrt{c_{\min} LC}$ เมื่อ L เป็นค่าความผิดพลาดแบบมีน้ำหนักความสำคัญของตัวทำนายเชิงเส้นใดๆ และ C เป็นพารามิเตอร์ที่ขึ้นกับความซับซ้อนของตัวทำนายเชิงเส้นนั้น ค่าความผิดพลาดนี้สามารถเทียบได้กับอัลกอริทึมเพอร์เซพตรอนในกรณีที่ตัวอย่างไม่มีน้ำหนักความสำคัญ (หรือกรณีที่ $c_{\max} = c_{\min} = 1$) ที่มีค่า $L + C + \sqrt{LC}$ เนื่องจากค่าความผิดพลาดข้างต้นขึ้นกับค่า θ เรายังได้ทำการปรับปรุงวิธีการในการพิสูจน์ค่าดังกล่าวใหม่ ด้วยวิธีนี้เราสามารถพิสูจน์ประสิทธิภาพการทำงานได้ดีขึ้น นั่นคือ ค่าความผิดพลาดที่ได้จะไม่เกิน $L + c_{\max} C + \sqrt{c_{\max} LC}$

อย่างไรก็ตามวิธีการที่สองที่นำเสนอใช้ได้กับกรณีที่ตัวอย่างมีสองประเภท ในกรณีที่ตัวอย่างมีหลายประเภท งานวิจัยนี้สนใจการแก้ปัญหาด้วยวิธีการลดรูป กล่าวคือทำการลดรูปปัญหาการเรียนรู้แบบหลายประเภทไปเป็นปัญหาการเรียนรู้แบบสองประเภท โดยได้ทำการศึกษาวิธีการลดรูปสองวิธีการ ได้แก่ วิธีการหนึ่งต่อทั้งหมด (One-against-All) และวิธีการหนึ่งต่อหนึ่ง (One-against-One) และได้ทำการพิสูจน์ว่า ในกรณีที่ใช้อัลกอริทึมเพอร์เซพตรอนเป็นฐานการลดรูปด้วยวิธีการหนึ่งต่อทั้งหมดและวิธีการหนึ่งต่อหนึ่งนั้นสมมูลกัน

วัตถุประสงค์

ศึกษาและปรับปรุงอัลกอริทึมการเรียนรู้แบบออนไลน์ สำหรับปัญหาการเรียนรู้แบบมีน้ำหนักความสำคัญ

การตรวจเอกสาร

นิยามพื้นฐาน

นิยาม *ภารกิจการเรียนรู้แบบมีการแนะนำ* คือ คู่ลำดับ (K,Y,ℓ) โดยที่ K เป็นเซตของประเภท ที่ได้รับการแนะนำระหว่างขั้นตอนการฝึกสอน, Y เป็นเซตของประเภทที่ได้จากการทำนาย และ $\ell : K \times Y \to [0,\infty)$ เป็นฟังก์ชันความสูญเสีย

นิยาม *ปัญหาการเรียนรู้แบบมีการแนะนำ* คือ คู่ลำดับ (D,X,T) โดยที่ $T = (K,Y,\ell)$ เป็นภารกิจการเรียนรู้แบบมีการแนะนำ, X เป็นเซตของลักษณะ (feature) และ D เป็นการกระจายตัวบน $X \times K$

เป้าหมายของการแก้ปัญหาการเรียนรู้แบบมีการแนะนำคือ การหาสมมติฐาน $h : X \to Y$ ที่ทำให้ค่าเฉลี่ยของความสูญเสีย $E_{(x,k)\sim D}\ell(k,h(x))$ น้อยที่สุด

นิยาม *อัลกอริทึมการเรียนรู้แบบมีการแนะนำ* สำหรับภารกิจ (K,Y,ℓ) คือ กระบวนการสร้างแผนที่สำหรับเซตจำกัดของกลุ่มตัวอย่าง $(X \times K)^*$ ใดๆ ไปยังสมมติฐาน $h : X \to Y$

เนื่องจากในงานวิจัยนี้สนใจเฉพาะการเรียนรู้แบบมีการแนะนำ ดังนั้นต่อไปจะเรียกการเรียนรู้แบบมีการแนะนำ แบบย่อ ว่า “การเรียนรู้” และสมมติให้เซตของประเภทที่ได้รับการแนะนำระหว่างขั้นตอนการฝึกสอน K เท่ากับ เซตของประเภทที่ได้จากการทำนาย Y

การเรียนรู้แบบออฟไลน์

ให้ X เป็นเซตของตัวอย่างทั้งหมดที่เป็นไปได้, Y เป็นเซตจำกัดของประเภทของตัวอย่างทั้งหมดที่เป็นไปได้, S เป็นเซตของกลุ่มตัวอย่างที่สุ่มโดยอิสระต่อกันจากการกระจายตัว D ซึ่งเป็นการกระจายตัวบน $X \times Y$ สมมติฐานใดๆ คือฟังก์ชัน $h : X \to Y$ ในการเรียนรู้หนึ่งๆ เราจะจำกัดการพิจารณาสมมติฐานให้อยู่ในเซตหนึ่งๆ เท่านั้น กล่าวคือ เราจะให้ H เป็นเซตของสมมติฐานทั้งหมดที่สนใจ *อัลกอริทึมการเรียนรู้แบบออฟไลน์* คืออัลกอริทึมที่รับกลุ่มตัวอย่าง S มาใช้สร้าง

สมมติฐาน $h \in H$ ประสิทธิภาพของสมมติฐานดังกล่าวจะวัดโดยความผิดพลาดซึ่งมีนิยามเป็น $\Pr_{(x,y)\sim D}[h(x) \neq y]$

จากนิยามข้างต้นจะเห็นว่าอัลกอริทึมการเรียนรู้แบบออฟไลน์จะมีเอาท์พุทเป็นสมมติฐาน หรือ *ตัวทำนาย* และตัวทำนายดังกล่าวสามารถนำไปใช้ในการทำนายประเภทตัวอย่างอื่นๆ ต่อไป

การเรียนรู้แบบออนไลน์

ให้ X เป็นเซตของตัวอย่างที่เป็นไปได้, Y เป็นเซตจำกัดของประเภทของตัวอย่างทั้งหมดที่เป็นไปได้ ในสถานการณ์การเรียนรู้แบบออนไลน์ อัลกอริทึมจะไม่ได้รับกลุ่มตัวอย่าง S ที่สมบูรณ์ก่อนเริ่มทำงาน หากแต่จะได้รับตัวอย่างไปพร้อมๆ กับการใช้งานจริง โดยมีการทำงานเป็นรอบ ๆ ดังนี้ เมื่อเริ่มต้นแต่ละรอบ อัลกอริทึมจะได้รับตัวอย่าง $x \in X$ จากนั้นจะต้องทำนายประเภทของตัวอย่างนั้นทันที เมื่ออัลกอริทึมทำนายเสร็จ จะได้รับคำเฉลย $y \in Y$ เพื่อนำไปใช้ปรับปรุงความถูกต้องของการจำแนกในรอบถัดๆ ไป การวัดประสิทธิภาพของอัลกอริทึมจะวัดจากจำนวนครั้งทั้งหมดที่ตอบผิด

โดยเนื้อแท้แล้ว *อัลกอริทึมการเรียนรู้แบบออนไลน์* สามารถระบุโดยใช้ฟังก์ชัน $f : H \times X \times Y \to H$ สำหรับปรับปรุงสมมติฐาน กล่าวคือในการทำงานแต่ละรอบอัลกอริทึมจะมีสมมติฐาน h_t สำหรับจำแนกประเภทของตัวอย่าง $x_t \in X$ อัลกอริทึมจะตอบ $\hat{y}_t = h_t(x_t)$ ถ้าอัลกอริทึมตอบผิด นั่นคือ $\hat{y}_t \neq y_t$ อัลกอริทึมจะปรับปรุงสมมติฐานโดยใช้ฟังก์ชัน f กล่าวคือ $h_{t+1} = f(h_t, x_t, y_t)$ ไม่เช่นนั้นจะให้ $h_{t+1} = h_t$

จากนิยามจะเห็นว่าในการเรียนรู้แบบออนไลน์จะมีสมมติฐานที่เปลี่ยนแปลงได้ตลอดกระบวนการการทำงาน ซึ่งต่างจากอัลกอริทึมการเรียนรู้แบบออฟไลน์ที่ใช้เพียงสมมติฐานเดียวในการตอบคำถาม

ในการพิจารณาการทำงานของอัลกอริทึมการเรียนรู้แบบออนไลน์ เราจะพิจารณาลำดับของข้อมูลป้อนเข้าและผลลัพธ์ของอัลกอริทึมบนลำดับป้อนเข้านั้นๆ กล่าวคือ จะเรียกลำดับ (x_1,y_1), (x_2,y_2) …(x_m,y_m) ว่า *ลำดับข้อมูลป้อนเข้า* ประสิทธิภาพการทำงานของอัลกอริทึมออนไลน์ f ในลำดับดังกล่าวจะวัดโดยใช้ฟังก์ชันความสูญเสีย $Loss : Y \times Y \to \Re$ ซึ่งจะมีค่าเท่ากับ

$$\sum_{t=1}^{m} Loss(y_t, \hat{y}_t)$$

เมื่อ $\hat{y}_i$ คือผลลัพธ์ของอัลกอริทึมในการทำงานรอบที่ t โดยทั่วไปฟังก์ชันความสูญเสียที่นิยมใช้จำนวนครั้งที่ตอบผิดพลาด (mistake) คือ $M(y, \hat{y}) = I[y \neq \hat{y}]$ เมื่อ I เป็นฟังก์ชันชี้วัด (indicator function)

การเรียนรู้แบบออนไลน์แบบมีน้ำหนักความสำคัญ

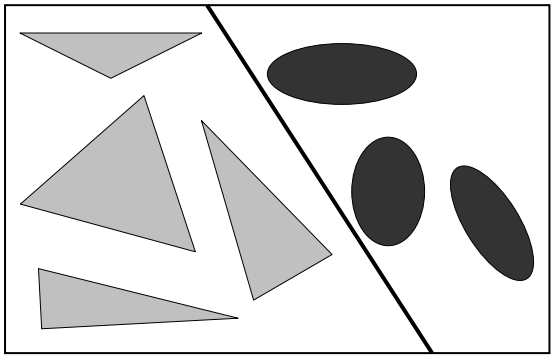
ในกรณีที่ตัวอย่างมีน้ำหนักความสำคัญ นั่นคือตัวอย่างจะระบุน้ำหนัก c จากเซต $C \subseteq [0, \infty)$ อัลกอริทึมจะทราบน้ำหนักของตัวอย่างเมื่อได้ทำนายประเภทของตัวอย่างนั้นแล้ว *อัลกอริทึมเรียนรู้ออนไลน์แบบมีน้ำหนักความสำคัญ*สามารถนิยามได้ในลักษณะเดียวกันกับอัลกอริทึมแบบออนไลน์ทั่วไป เพียงแต่ฟังก์ชัน $f : H \times X \times Y \times C \to H$ จะรับน้ำหนักของตัวอย่างไปใช้ในการปรับสมมติฐานด้วย เรียกลำดับ $(x_1,y_1,c_1), (x_2,y_2,c_2), \ldots (x_m,y_m,c_m)$ ว่า *ลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก* และจะวัดประสิทธิภาพของอัลกอริทึมออนไลน์ f ในลำดับดังกล่าวเป็น

$$\sum_{t=1}^{m} c_t \cdot Loss(y_t, \hat{y}_t)$$

เมื่อ $\hat{y}_t$ คือผลลัพธ์ของอัลกอริทึมในการทำงานรอบที่ t และ $Loss$ เป็นฟังก์ชันความสูญเสีย

อัลกอริทึมเพอร์เซพตรอน (Perceptron Algorithm)

อัลกอริทึมเพอร์เซพตรอน Rosenblatt (1958) เป็นอัลกอริทึมการเรียนรู้ออนไลน์แบบหนึ่งที่ใช้ในการสร้างตัวทำนายเชิงเส้น สำหรับการทำนายประเภทของกลุ่มตัวอย่างแบบสองประเภท แนวคิดหลักของอัลกอริทึมคือพยายามหาเส้นตรงที่แบ่งแยกกลุ่มตัวอย่างทั้งสองประเภทออกจากกัน ดังแสดงในภาพที่ 1



ภาพที่ 1 ตัวอย่างการทำนายด้วยเพอร์เซพตรอน

อัลกอริทึมเพอร์เซพตรอนจะทำงานกับตัวทำนายเชิงเส้น ซึ่งนิยามโดยเวคเตอร์น้ำหนัก $\mathbf{w} \in \Re^n$ โดยอัลกอริทึมจะรับตัวอย่าง $\mathbf{x}_t \in \Re^n$ แล้วทำนาย $\hat{y}_t = \text{sgn}\langle \mathbf{x}_t, \mathbf{w}_t \rangle$ เมื่อ sgn คือฟังก์ชันเครื่องหมาย จากนั้นจะได้รับเฉลย y_t ถ้าทำนายผิด $\hat{y}_t \neq y_t$ อัลกอริทึมจะทำการปรับปรุงเวคเตอร์น้ำหนัก

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_t \mathbf{x}_t$$

สำหรับการวิเคราะห์ประสิทธิภาพของอัลกอริทึมเพอร์เซพตรอนมีอยู่ 2 แนวทางได้แก่ การวิเคราะห์ในกรณีมีตัวทำนายเชิงเส้นใดๆ ที่สามารถแบ่งกลุ่มตัวอย่างทั้งสองกลุ่มออกจากกันได้ และการวิเคราะห์ในกรณีไม่มีตัวทำนายเชิงเส้นใดๆ ที่สามารถแบ่งกลุ่มตัวอย่างทั้งสองกลุ่มออกจากกันได้

ทฤษฎีบทที่ 1 Novikoff (1962) *สำหรับลำดับข้อมูลป้อนเข้า* $S = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)) \in (\Re^n \times \{-1,+1\})^m$ *ใดๆ ถ้า* $R = \max_t \|\mathbf{x}_t\|_2$ *และมีเวคเตอร์ขนาดหนึ่งหน่วย* $\mathbf{u}$ $\left(\|\mathbf{u}\|_2 = 1\right)$ *ที่ทำให้สำหรับทุกๆ ตัวอย่างใน* S *มี* $y_t \cdot \langle \mathbf{u}, \mathbf{x}_t \rangle \geq \gamma$ *แล้วจำนวนครั้งที่อัลกอริทึมเพอร์เซพตรอนทำนายผิดมีค่าไม่เกิน* $(R/\gamma)^2$

ทฤษฎีบทข้างต้นเป็นการวิเคราะห์เฉพาะกรณีมีตัวทำนายเชิงเส้นใดๆ ที่สามารถแบ่งกลุ่มตัวอย่างทั้งสองกลุ่มออกจากกันได้ ส่วนในกรณีไม่มีตัวทำนายเชิงเส้นใดๆ ที่สามารถแบ่งกลุ่มตัวอย่างทั้งสองกลุ่มออกจากกันได้ จะกล่าวถึงในส่วนถัดไป

ประสิทธิภาพของตัวทำนายเชิงเส้น

ในการพิสูจน์ประสิทธิภาพของอัลกอริทึมการเรียนรู้แบบออนไลน์ นิยมพิสูจน์เทียบประสิทธิภาพกับสมมติฐาน h^* จากเซตของสมมติฐาน H' บางเซต ซึ่งไม่จำเป็นต้องเท่ากับ H เซตของสมมติฐานที่นิยมใช้ ในกรณีที่เซตของตัวอย่าง $X = \Re^n$ คือเซตของตัวจำแนกเชิงเส้น h^* ทั้งหมดที่นิยามด้วยเวคเตอร์ $\mathbf{u} \in \Re^n$ กล่าวคือ $h^*(\mathbf{x}) = \text{sgn}\langle \mathbf{x}, \mathbf{u} \rangle$ เมื่อ sgn คือฟังก์ชันเครื่องหมาย

ประสิทธิภาพของตัวจำแนกเชิงเส้นที่นิยามด้วยเวคเตอร์ $\mathbf{u}$ สามารถวัดได้หลายแบบ เช่น วัดโดยใช้จำนวนครั้งที่ตอบผิดพลาด ในที่นี้จะวัดโดยใช้ฟังก์ชันความสูญเสียแบบฮิงก์ (Hinge loss) ที่มีนิยามเป็น

$$\ell(\mathbf{x}, y, \mathbf{u}) = \max\{0, 1 - y \cdot \langle \mathbf{u}, \mathbf{x} \rangle\} \tag{1}$$

ดังนั้นความสูญเสียแบบฮิงก์ $L(h^*)$ ทั้งหมดบนลำดับข้อมูลป้อนเข้าแบบไม่มีน้ำหนัก $(\mathbf{x}_1, y_1)$, …,$(\mathbf{x}_m, y_m)$ จะมีค่าเป็น $\sum_{t=1}^{m} \ell(\mathbf{x}_t, y_t, \mathbf{u})$

ในกรณีที่ข้อมูลป้อนเข้ามีน้ำหนัก จะนิยามความสูญเสียทั้งหมดแบบฮิงก์ที่มีน้ำหนัก $L_w(h^*)$ ของตัวจำแนกเชิงเส้นที่นิยามด้วยเวคเตอร์ $\mathbf{u}$ บนลำดับ $(\mathbf{x}_1, y_1, c_1)$, $(\mathbf{x}_2, y_2, c_2)$, …$(\mathbf{x}_m, y_m, c_m)$ เป็น $L_w(h^*) = \sum_{t=1}^{m} c_t \cdot \ell(\mathbf{x}_t, y_t, \mathbf{u})$

สำหรับการวิเคราะห์ประสิทธิภาพของอัลกอริทึมเพอร์เซพตรอนในกรณีไม่มีตัวทำนายเชิงเส้นใดๆ Gentile (2003) ได้ทำการวิเคราะห์ไว้ดังทฤษฎีบทต่อไปนี้

ทฤษฎีบทที่ 2 Gentile (2003) *สำหรับลำดับข้อมูลป้อนเข้า* $S = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)) \in (\Re^n \times \{-1, +1\})^m$ *ใดๆ ถ้า* $R = \max_t \|\mathbf{x}_t\|_2$ *และมีตัวทำนายเชิงเส้น* h^* *ที่นิยามด้วยเวคเตอร์* $\mathbf{u}$ *โดยที่* $L(h^*) \le L$ *และให้* $C = R^2 \|\mathbf{u}\|_2^2$ *แล้วจำนวนครั้งที่อัลกอริทึมเพอร์เซพตรอนทำนายผิดบน* S *มีค่าไม่เกิน* $L + C + \sqrt{LC}$

อันที่จริงแล้วทฤษฎีบทที่ 2 นี้เป็นรูปแบบเฉพาะของทฤษฎีบทที่ว่าด้วยขีดจำกัดบนของอัลกอริทึมพีนอร์ม (p-Norm Algorithm) ทำนายผิดใน Gentile (2003) ซึ่งในบทพิสูจน์ดังกล่าวมีการวัดระยะห่างระหว่าง 2 เวคเตอร์ โดยใช้ฟังก์ชันดังนี้

$$d_{\mathbf{f}}(\mathbf{u},\mathbf{w}) = \frac{1}{2}\|\mathbf{u}\|_p^2 + \frac{1}{2}\|\mathbf{w}\|_q^2 - \langle \mathbf{u}, \mathbf{f}(\mathbf{w}) \rangle \tag{2}$$

เมื่อ $\frac{1}{p}+\frac{1}{q}=1$ และ ฟังก์ชัน $\mathbf{f}:\Re^n \to \Re^n$ มีนิยามเป็น $\mathbf{f}=(f_1,\ldots,f_n)$ โดยที่

$$f_i(\mathbf{w}) = \frac{\operatorname{sgn}(w_i)|w_i|^{q-1}}{\|\mathbf{w}\|_q^{q-2}} \ , \ \mathbf{w}=(w_1,\ldots,w_n) \in \Re^n$$

ฟังก์ชันระยะห่างตามสมการ (2) เรียกว่า *การลู่ออกของเบรกแมน* (*Bregman divergence*) สำหรับฟังก์ชัน $\mathbf{f}$ มีอินเวอร์สฟังก์ชันเป็นฟังก์ชัน $\mathbf{f}^{-1}:\Re^n \to \Re^n$ ซึ่งมีนิยามเป็น $\mathbf{f}^{-1}=(f_1^{-1},\ldots,f_n^{-1})$ โดยที่

$$f_i^{-1}(\boldsymbol{\theta}) = \frac{\operatorname{sgn}(\theta_i)|\theta_i|^{p-1}}{\|\theta\|_p^{p-2}} \ , \ \boldsymbol{\theta}=(\theta_1,\ldots,\theta_n) \in \Re^n$$

สำหรับการพิสูจน์ทฤษฎีบทข้างต้นจะต้องใช้ทฤษฎีบทย่อย 2 ทฤษฎีดังนี้

ทฤษฎีบทย่อยที่ 1 *ให้* $\mathbf{u},\mathbf{w},\mathbf{x} \in \Re^n$, $a \in \Re$ *และ* $\mathbf{w}' = \mathbf{f}^{-1}(\mathbf{f}(\mathbf{w}) + a\mathbf{x}), \boldsymbol{\theta} = \mathbf{f}(\mathbf{w})$ *แล้ว จะได้ว่า*

$$a\big(\langle \mathbf{u},\mathbf{x} \rangle - \langle \mathbf{w},\mathbf{x} \rangle\big) = d_{\mathbf{f}}(\mathbf{u},\mathbf{w}) - d_{\mathbf{f}}(\mathbf{u},\mathbf{w}') + d_{\mathbf{f}}(\mathbf{w},\mathbf{w}')$$

ทฤษฎีบทย่อยที่ 2 *ให้* $\mathbf{u},\mathbf{w},\mathbf{x} \in \Re^n$, $a \in \Re$, $p \geq 2$ *และ* $\mathbf{w}' = \mathbf{f}^{-1}(\mathbf{f}(\mathbf{w}) + a\mathbf{x}), \boldsymbol{\theta} = \mathbf{f}(\mathbf{w})$ *แล้ว จะได้ว่า*

$$d_{\mathbf{f}}(\mathbf{w},\mathbf{w}') \leq \frac{a^2}{2}(p-1)\|\mathbf{x}\|_2^2$$

สำหรับนิยามฟังก์ชันการลู่ออกของเบรกแมน, ทฤษฎีบทย่อยที่ 1 และทฤษฎีบทย่อยที่ 2 นั้นมีอยู่ใน Gentile (2003) ซึ่งในงานวิจัยนี้จะทำการปรับปรุงนิยามและทฤษฎีบทย่อยดังกล่าว เพื่อใช้สำหรับการพิสูจน์ทฤษฎีบทที่ว่าด้วยค่าขีดจำกัดบนของความสูญเสียของอัลกอริทึมที่ได้นำเสนอ ซึ่งเป็นอัลกอริทึมที่ปรับปรุงมาจากเพอร์เซพตรอน

การลดรูปการเรียนรู้

ในการแก้ปัญหาการเรียนรู้มีแนวทางการแก้ปัญหาอยู่ 2 แนวทาง แนวทางแรกคือ การแก้ปัญหาการเรียนรู้ต่างๆ โดยอิสระจากวิธีการอื่น หรือกล่าวคือพยายามแก้ปัญหาการเรียนรู้นั้นๆ โดยตรง ส่วนแนวทางที่สองคือ แก้ปัญหาโดยวิธีการแบบลดรูป นั่นคือการลดรูปจากปัญหาที่ต้องการจะแก้ไปเป็นปัญหาที่มีการแก้และวิเคราะห์ไปแล้ว ซึ่งมีข้อดีคือสามารถใช้อัลกอริทึมและทฤษฎีของปัญหาดังกล่าวกับปัญหาที่ต้องการจะแก้ได้ หรือกล่าวอีกนัยหนึ่ง การลดรูปการเรียนรู้ R จากภารกิจ $T=(K,Y,\ell)$ ไปยังภารกิจ $T'=(K',Y',\ell')$ คือกระบวนการที่ใช้อัลกอริทึมเรียนรู้สำหรับ T' เป็นกล่องดำในการแก้ภารกิจ T โดยรับประกันว่าถ้าแก้ปัญหาย่อยที่สร้างขึ้นโดย R ได้ดี แล้วจะแก้ปัญหาตั้งต้นได้ดีด้วย

นิยาม *การลดรูปการเรียนรู้* $R(S,A)$ จากภารกิจ (K,Y,ℓ) ไปยังภารกิจ (K',Y',ℓ') คือ กระบวนการที่รับข้อมูลที่ประกอบด้วย เซตจำกัดของข้อมูลป้อนเข้า $S\in(X\times K)^*$ และ อัลกอริทึมเรียนรู้ A สำหรับภารกิจ (K',Y',ℓ') และให้ผลลัพธ์เป็นสมมติฐาน $h:X\to Y$ ซึ่งเป็นกลุ่มของสมมติฐานย่อยที่ได้จาก A

วิธีการหนึ่งต่อทั้งหมด (One-against-All Method)

วิธีการแบบหนึ่งต่อทั้งหมดเป็นวิธีการในการลดรูปการเรียนรู้จากปัญหาการเรียนรู้แบบหลายประเภทไปเป็นปัญหาการเรียนรู้แบบสองประเภทวิธีหนึ่ง ซึ่งจะทำการลดรูปการเรียนรู้ k ประเภทไปเป็นการเรียนรู้สองประเภท สำหรับการเรียนรู้แบบหลายประเภท จะใช้เซตของตัวทำนายแบบสองประเภท $B=\{b^r:r\in\{1,2,\ldots k\}\}$ ซึ่งตัวทำนายแต่ละตัวถูกฝึกสอนโดยแผนที่ $(x,y)\to(x,I(y=r))$ จากตัวอย่างการเรียนรู้แบบหลายประเภทไปยังตัวอย่างการเรียนรู้แบบสองประเภท โดยวิธีการหนึ่งต่อทั้งหมดจะทำนาย r เมื่อ $b^r(x)=1$

เมื่อพิจารณาวิธีการหนึ่งต่อทั้งหมดที่กล่าวมาในข้างต้น จะเห็นว่าเวลาที่ใช้ในการทำนายแต่ละตัวอย่างมีค่าเป็น $O(k)$ และวิธีการดังกล่าวสามารถสร้างตัวทำนายแบบหลายประเภทจากตัวทำนายแบบสองประเภทใดๆ ก็ได้ สำหรับอัลกอริทึมที่ใช้สร้างตัวทำนายแบบสองประเภทที่สนใจในงานวิจัยนี้ คือ อัลกอริทึมเพอร์เซพตรอน ในการสร้างตัวทำนายแบบหลายประเภทโดยใช้เพอร์เซพตรอนเป็นฐานมีวิธีการหลักๆ ได้แก่ วิธีการหลายเวคเตอร์ และ วิธีการเวคเตอร์เดียว

วิธีการหลายเวคเตอร์ (Multi-vector Method)

ให้ Y เป็นเซตจำกัดของประเภทของตัวอย่างทั้งหมดที่เป็นไปได้ วิธีการหลายเวคเตอร์จะทำนายโดยใช้เซตของเวคเตอร์น้ำหนัก $W = \{w^r : r \in Y\}$ ซึ่งประกอบด้วย $|Y|$ เวคเตอร์ สำหรับตัวอย่าง $\mathbf{x}_t$ จะทำนาย

$$\hat{y}_t = \arg\max_{r \in Y} \langle \mathbf{w}_t^r, \mathbf{x}_t \rangle$$

จากนั้นเมื่อได้รับเฉลย y_t ถ้าทำนายผิด $\hat{y}_t \neq y_t$ จะทำการปรับปรุงเวคเตอร์น้ำหนัก

$$\mathbf{w}_{t+1}^{y_t} \leftarrow \mathbf{w}_t^{y_t} + \mathbf{x}_t \ , \ \mathbf{w}_{t+1}^{\hat{y}_t} \leftarrow \mathbf{w}_t^{\hat{y}_t} - \mathbf{x}_t$$

และ $\mathbf{w}_{t+1}^r \leftarrow \mathbf{w}_t^r$ สำหรับทุกๆ $r \in Y \setminus \{y_t, \hat{y}_t\}$

วิธีการเวคเตอร์เดียว (Single-vector Method)

วิธีการเวคเตอร์เดียวจะทำนายโดยใช้เวคเตอร์น้ำหนัก $\mathbf{w}_t$ เพียงตัวเดียวในการทำนายตัวอย่าง $\mathbf{x}_t$ อย่างไรก็ตาม การทำนายด้วยวิธีการนี้จะต้องทำการสร้างแผนที่ตัวอย่าง ด้วยฟังก์ชันแผนที่ $\phi : \Re^n \times Y \rightarrow \Re^d$ นั่นคือ สำหรับตัวอย่าง $\mathbf{x}_t$ จะทำนาย

$$\hat{y}_t = \arg\max_{r \in Y} \langle \mathbf{w}_t, \phi(\mathbf{x}_t, r) \rangle$$

จากนั้นเมื่อได้รับเฉลย y_t ถ้าทำนายผิด $\hat{y}_t \neq y_t$ จะทำการปรับปรุงเวคเตอร์น้ำหนัก

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \phi(\mathbf{x}_t, y_t) - \phi(\mathbf{x}_t, \hat{y}_t)$$

สำหรับฟังก์ชันแผนที่ สามารถสร้างได้หลายวิธี ยกตัวอย่างวิธีง่ายๆ คือ ใช้ตัวอย่างแรกที่เจอในแต่ละประเภท เพื่อทำการสร้างแผนที่สำหรับตัวอย่าง $\mathbf{x}_t$ และประเภทนั้นๆ หรือกล่าวคือ สำหรับแต่ละประเภท $r \in Y$ ให้ $\mathbf{p}^r \in \Re^n$ เป็นตัวอย่างแรกของประเภท r ที่เจอในลำดับข้อมูลป้อนเข้า, สามารถนิยามแผนที่ $\phi(\mathbf{x}_t, r)$ ซึ่งเป็นเวคเตอร์บน $\Re^n$ ได้เป็น

$$\phi_i(\mathbf{x}_t, r) = x_{t,i} p_t^r$$

เมื่อ $\phi_i(\mathbf{x}_t, r), \mathbf{x}_{t,i}$ และ p_i^r เป็นค่าในมิติที่ i ของเวคเตอร์ $\phi(\mathbf{x}_t, r), \mathbf{x}_t$ และ $\mathbf{p}^r$ ตามลำดับ

วิธีการหนึ่งต่อหนึ่ง (One-against-One Method)

วิธีการแบบหนึ่งต่อหนึ่งเป็นวิธีการในการลดรูปการเรียนรู้การเรียนรู้แบบหลายประเภทไปเป็นภารกิจการเรียนรู้แบบสองประเภทอีกวิธีการหนึ่ง, ให้ X เป็นเซตของตัวอย่างที่เป็นไปได้ และ $Y = \{1,2,\ldots k\}$ วิธีการแบบหนึ่งต่อหนึ่งจะทำนายโดยใช้เซตของตัวทำนายแบบสองประเภท $B = \{b^{ij} : i \in Y, j \in Y, i \neq j\}$ โดยที่ตัวทำนาย b^{ij} เป็นฟังก์ชันจาก X ไปยัง $\{i,j\}$ และ $b^{ij}(x) = b^{ji}(x)$ วิธีการหนึ่งต่อหนึ่งจะทำนาย

$$\hat{y} = \arg\max_{r \in Y} \sum_{j \in Y : j \neq r} I[r = b^{rj}(x)]$$

เมื่อ I เป็นฟังก์ชันชี้วัด, พิจารณาวิธีการหนึ่งต่อหนึ่งที่กล่าวมาในข้างต้น จะเห็นว่าวิธีการนี้ใช้ตัวทำนายแบบสองประเภทเป็นจำนวน $\binom{k}{2}$ ตัวทำนาย และใช้เวลาในการทำนายแต่ละตัวอย่างเป็น $O(k^2)$ ซึ่งมากกว่าวิธีการแบบหนึ่งต่อทั้งหมด

อย่างไรก็ตามในกรณีการเรียนรู้แบบออฟไลน์การเปรียบเทียบประสิทธิภาพการทำนายของทั้งสองวิธีการยังไม่เป็นที่แน่ชัดว่าวิธีใดดีกว่ากัน และยังเป็นปัญหาเปิดที่สำคัญปัญหาหนึ่ง (Rifkin and Klautau (2004))

งานวิจัยในอดีต

งานวิจัยชิ้นนี้เกี่ยวข้องกับงานวิจัยทางการเรียนรู้ด้วยเครื่องจักร 2 กลุ่มใหญ่ ๆ กล่าวคือ งานที่เกี่ยวข้องกับการเรียนรู้ที่ไวต่อค่าใช้จ่าย (Cost-sensitive learning) และงานที่เกี่ยวข้องกับการเรียนรู้แบบออนไลน์

การเรียนรู้ที่ไวต่อค่าใช้จ่าย. งานวิจัยด้านนี้มีมากมาย โดยเฉพาะในในหลาย ๆ ปีที่ผ่านมา ได้มีความสนใจในปัญหาดังกล่าว เช่น Zadrozny *et al*. (2003); Elkan (2001); Abe *et al*. (2004); Domingos (1999); Liu (2006) โดยในการแก้ปัญหามีหลายวิธี เช่น การพยายามปรับอัลกอริทึมหนึ่งๆ ให้ทำงานได้กับข้อมูลที่มีน้ำหนัก Liu (2006) หรืออาจจะเป็นการพยายามสร้างวิธีทั่วไปในการสร้างอัลกอริทึมสำหรับปัญหานี้ โดยอาจมีการประมาณการกระจายตัวของข้อมูล Zadrozny and Elkan (2001) หรืออาจจะใช้วิธีการลดรูปการเรียนรู้ไปยังปัญหาการทำนายแบบไม่มีน้ำหนัก Zadrozny *et al*. (2003) ซึ่งอัลกอริทึมแรกที่นำเสนอในงานวิจัยนี้ก็ใช้แนวคิดของการลดรูปเช่นเดียวกัน

การเรียนรู้แบบออนไลน์. การเรียนรู้แบบออนไลน์เป็นสาขาที่มีการทำวิจัยอย่างกว้างขวาง สำหรับเนื้อหาหลัก ดูหนังสือ Cesa-Bianchi and Lugosi (2006) อัลกอริทึมพื้นฐานของการเรียนรู้แบบออนไลน์มีหลายแบบ ที่ใช้ในงานวิจัยนี้คือเพอร์เซพตรอน Rosenblatt (1958) ซึ่ง Novikoff (1962) ได้ทำการวิเคราะห์ประสิทธิภาพของอัลกอริทึมนี้ในกรณีที่มีเส้นแบ่งระหว่างข้อมูลสองกลุ่มได้ ส่วนในงานวิจัยนี้เราใช้การวิเคราะห์ประสิทธิภาพในแบบ Gentile (2003) ที่สามารถพิจารณากรณีที่ไม่มีเส้นแบ่งระหว่างข้อมูลสองกลุ่มได้ อัลกอริทึมแบบออนไลน์สำหรับปัญหาการทำนายหลายประเภทที่ใช้เพอร์เซพตรอนมีมากมาย เช่น งานของ Fink *et al*. (2006) ที่สามารถใช้เป็นกล่องดำในการลดรูปในอัลกอริทึมแรกที่งานวิจัยนี้นำเสนอได้ด้วย

งานวิจัยของ Cammer *et al*. (2006) ได้นำเสนอและวิเคราะห์อัลกอริทึมที่เรียกว่า Passive-Aggressive สำหรับปัญหาการเรียนรู้แบบออนไลน์ อัลกอริทึมดังกล่าวสามารถจัดการกับปัญหาการเรียนรู้แบบหลายประเภทที่ไวต่อค่าใช้จ่ายได้ด้วย อย่างไรก็ตาม น้ำหนักความสำคัญของข้อมูลอัลกอริทึมนี้รองรับได้จะขึ้นกับประเภทของข้อมูลเท่านั้น ในขณะที่ในงานเช่นของ Zadrozny *et al*. (2003) และงานวิจัยนี้น้ำหนักความสำคัญจะขึ้นกับข้อมูลแต่ละตัว

วิธีการ

งานวิจัยนำเสนอวิธีการในการแก้ปัญหาการเรียนรู้ที่ตัวอย่างมีน้ำหนักความสำคัญทั้งแบบสองประเภทและแบบหลายประเภทโดยใช้อัลกอริทึมการเรียนรู้แบบออนไลน์ สำหรับแนวคิดหลักของการแก้ปัญหาคือให้อัลกอริทึมการเรียนรู้แบบออนไลน์ทำการเรียนรู้ตามค่าน้ำหนักความสำคัญของแต่ละตัวอย่าง ในส่วนของการแก้ปัญหาการเรียนรู้สองประเภทแบบมีน้ำหนักความสำคัญได้นำเสนออัลกอริทึมการเรียนรู้แบบออนไลน์ 2 อัลกอริทึม โดยอัลกอริทึมแรกจะทำการทำซ้ำข้อมูลตามค่าน้ำหนักความสำคัญของตัวอย่าง ส่วนอีกอัลกอริทึมหนึ่งปรับปรุงมาจากอัลกอริทึมเพอร์เซพตรอนอัลกอริทึม กล่าวคือถ้าอัลกอริทึมทำนายผิดจะทำการปรับปรุงเวคเตอร์น้ำหนักโดยการถ่วงน้ำหนักของตัวอย่างด้วยค่าน้ำหนักความสำคัญของตัวอย่าง สำหรับการแก้ปัญหาการเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญจะใช้วิธีการการลดรูปจากปัญหาการเรียนรู้หลายประเภทไปเป็นปัญหาการเรียนรู้สองประเภทโดยใช้อัลกอริทึมที่นำเสนอในส่วนของการแก้ปัญหาการเรียนรู้สองประเภทแบบมีน้ำหนักความสำคัญเป็นฐาน

วิธีการแก้ปัญหาการเรียนรู้สองประเภทแบบมีน้ำหนักความสำคัญ

สำหรับวิธีการแก้ปัญหาการเรียนรู้สองประเภทแบบมีน้ำหนักความสำคัญเรานำเสนออัลกอริทึมการเรียนรู้แบบออนไลน์ 2 อัลกอริทึมได้แก่อัลกอริทึมแบบลดรูปและอัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ

อัลกอริทึมแบบลดรูป

ในหัวข้อนี้จะนำเสนอการลดรูปการเรียนรู้ในแบบออนไลน์ กล่าวคือเราจะทำการลดรูปปัญหาการเรียนรู้แบบมีน้ำหนักความสำคัญไปเป็นปัญหาการเรียนรู้แบบไม่มีน้ำหนักความสำคัญด้วยกระบวนการแบบออนไลน์ ซึ่งแนวคิดหลักของวิธีการการลดรูปการเรียนรู้แบบออนไลน์นี้คือจะทำการแปลงลำดับข้อมูลป้อนเข้าแบบมีน้ำหนักไปเป็นลำดับข้อมูลป้อนเข้าแบบไม่มีน้ำหนักและทำการทำนายตัวอย่างบนลำดับที่ได้จากการแปลงโดยใช้อัลกอริทึมการเรียนรู้แบบออนไลน์ใดๆ ที่ไม่สนใจค่าน้ำหนักความสำคัญของตัวอย่าง การลดรูปดังกล่าวทำให้ได้อัลกอริทึมลดรูปการเรียนรู้แบบออนไลน์ ซึ่งเราจะเรียกว่า *อัลกอริทึมแบบลดรูป* (*Online-Reduction Algorithm*) ซึ่งในที่นี้จะเขียนแทนด้วย A^w สังเกตุว่า อัลกอริทึมแบบลดรูป A^w สามารถทำนายตัวอย่างแบบมีน้ำหนัก

ความสำคัญได้โดยเรียกใช้อัลกอริทึมการเรียนรู้แบบออนไลน์ A นั่นคือ A เป็นเหมือนกล่องดำของ A^w

ภาพที่ 2 แสดงการทำงานของอัลกอริทึมแบบลดรูป A^w ที่สร้างจากอัลกอริทึมการเรียนรู้ออนไลน์ A ที่มีฟังก์ชัน f ในการปรับตัวทำนาย โดย h_t เป็นสมมติฐานที่ A ใช้ในการทำนายรอบที่ t จากอัลกอริทึมแบบลดรูป จะเห็นว่า A^w จะเรียกใช้ A ในการทำนายประเภทของ x_t จากนั้น A^w จะได้รับเฉลย y_t และ c_t โดยสมมติให้ c_t มีค่าเป็นจำนวนเต็มบวก (สามารถพิจารณาได้ว่า x_t, y_t และ c_t เป็นตัวอย่างหนึ่งบนลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก Q) จากนั้น A^w จะทำการทำซ้ำ x_t ออกมาอีก c_t ตัว แล้วเรียกใช้ A เรียนรู้จากข้อมูลดังกล่าว ซึ่งเท่ากับว่า A^w ทำการแปลงลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก Q ไปเป็นลำดับข้อมูลป้อนเข้าแบบไม่มีน้ำหนัก

Algorithm *Online-Reduction*

1. For $t = 1, 2, 3, \ldots$
2. Receive an instance x_t
3. Predict $\hat{y}_t \leftarrow h_t(x_t)$
4. Receive correct label y_t and cost c_t
5. $h_{t+1} \leftarrow h_t$
6. For $i = 1$ to c_i
7. $x_{ti} = x_t$
8. If $h_{t+1}(x_{ti}) \neq y_t$

ภาพที่ 2 อัลกอริทึมแบบลดรูป

สังเกตุว่า อัลกอริทึมแบบลดรูป A^w สามารถใช้อัลกอริทึมการเรียนรู้แบบออนไลน์ A ใดๆ เป็นกล่องดำก็ได้ ดังนั้น A^w สามารถทำนายได้ในลักษณะเดียวกับอัลกอริทึมกล่องดำ A ที่นำมาใช้ ยกตัวอย่างเช่น ถ้า A เป็นอัลกอริทึมการเรียนรู้แบบหลายประเภท A^w ก็จะสามารถทำนายแบบหลายประเภทด้วย

อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ

ในหัวข้อนี้จะนำเสนอการปรับปรุงอัลกอริทึมเพอร์เซพตรอนเพื่อให้สามารถทำนายตัวอย่างที่มีน้ำหนักความสำคัญได้อย่างมีประสิทธิภาพ จากการทำงานของอัลกอริทึมเพอร์เซพตรอนที่กล่าวถึงในส่วนของการตรวจเอกสาร จะเห็นว่าเพอร์เซพตรอนทำงานกับตัวอย่างซึ่งเวคเตอร์ ดังนั้นเราสามารถใช้แนวคิดหลักของงานวิจัยนี้ที่กล่าวไว้ว่า ให้อัลกอริทึมการเรียนรู้แบบออนไลน์ทำการเรียนรู้ตามค่าน้ำหนักความสำคัญของแต่ละตัวอย่าง โดยเราเสนอให้ทำการปรับปรุงอัลกอริทึมเพอร์เซพตรอนโดยใช้วิธีการปรับค่าเวคเตอร์น้ำหนักของเพอร์เซพตรอน ในรอบที่ทำนายผิดด้วยถ่วงน้ำหนักของตัวอย่างด้วยค่าน้ำหนักความสำคัญของตัวอย่าง โดยเราจะเรียกอัลกอริทึมที่ได้จากการปรับปรุงนี้ว่า *อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ* หรือเรียกโดยย่อว่า *IWP*

Algorithm *IWP*

1. INITIALIZE: $\mathbf{w}_1 = 0$
2. For $t = 1, 2, 3, \ldots$
3. Receive an instance $\mathbf{x}_t$
4. Predict $\hat{y}_t \leftarrow \mathrm{sgn}\langle \mathbf{w}_t, \mathbf{x}_t \rangle$
5. Receive correct label y_t and cost c_t
6. If $\hat{y}_t \neq y_t$
7. $\mathbf{w}_{t+1} = \mathbf{w}_t + y_t c_t \mathbf{x}_t$

ภาพที่ 3 อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ

ภาพที่ 3 แสดงการวิธีการทำงานของเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ สังเกตุว่าอัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญใช้ได้เฉพาะการทำนายแบบสองประเภทเช่นเดียวกับอัลกอริทึมเพอร์เซพตรอน ในหัวข้อถัดไปเราจะนำเสนอวิธีการที่ทำให้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญสามารถทำนายแบบหลายประเภทได้

วิธีการแก้ปัญหาเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญ

ในการแก้ปัญหาการเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญ เรานำเสนอให้ใช้วิธีการลดรูปจากปัญหาการเรียนรู้หลายประเภทไปเป็นปัญหาการเรียนรู้สองประเภท โดยใช้อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญที่ได้นำเสนอในส่วนที่ผ่านมาเป็นฐาน สำหรับวิธีการลดรูปจากปัญหาการทำนายเรียนรู้หลายประเภทไปเป็นปัญหาการเรียนรู้สองประเภทที่เรานำมาใช้ ได้แก่ วิธีการหลายเวคเตอร์ และวิธีการทุกคู่

วิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐาน

วิธีการหลายเวคเตอร์เป็นวิธีการลดรูปจากปัญหาการเรียนรู้หลายประเภทไปเป็นปัญหาการทำนายสองประเภทซึ่งเป็นรูปแบบหนึ่งของวิธีการลดรูปแบบหนึ่งต่อทั้งหมด สำหรับการแก้ปัญหาการเรียนรู้หลายประเภทแบบไม่มีน้ำหนักความสำคัญโดยใช้วิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนเป็นฐาน เราได้กล่าวถึงแล้วในหัวข้อการตรวจเอกสาร ในส่วนนี้เราจะนำเสนอการปรับปรุงวิธีการดังกล่าวให้สามารถทำนายข้อมูลที่ตัวอย่างมีน้ำหนักความสำคัญได้ดี โดยเราจะเรียกวิธีการดังกล่าวว่า *วิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐาน* (Multi-vector based on *IWP*) ซึ่งมีวิธีการดังต่อไปนี้

สำหรับปัญหาการเรียนรู้ k ประเภท สมมติให้ $Y = \{1, 2, \ldots, k\}$ เป็นเซตประเภทของตัวอย่าง วิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐานจะทำนายโดยใช้เซตของเวคเตอร์น้ำหนัก $W = \{\mathbf{w}^r : r \in Y\}$ ซึ่งประกอบด้วย k เวคเตอร์ สำหรับการทำนายรอบที่ t ใดๆ จะใช้สมมติฐานการทำนาย $h_t : \Re \to \Re$ ในการคำนวณคะแนนของแต่ละประเภท $r \in Y$ ซึ่งจะมีค่าคะแนนเป็น

$$h_t(\mathbf{x}_t, r) = \left\langle \mathbf{w}_t^r, \mathbf{x}_t \right\rangle \tag{3}$$

วิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐานจะทำนาย

$$\hat{y}_t = \arg\max_{r \in Y} h_t(\mathbf{x}_t, r)$$

จากนั้นเมื่อได้รับเฉลย y_t ถ้าทำนายผิด $\hat{y}_t \neq y_t$ จะทำการปรับปรุงเวคเตอร์น้ำหนัก

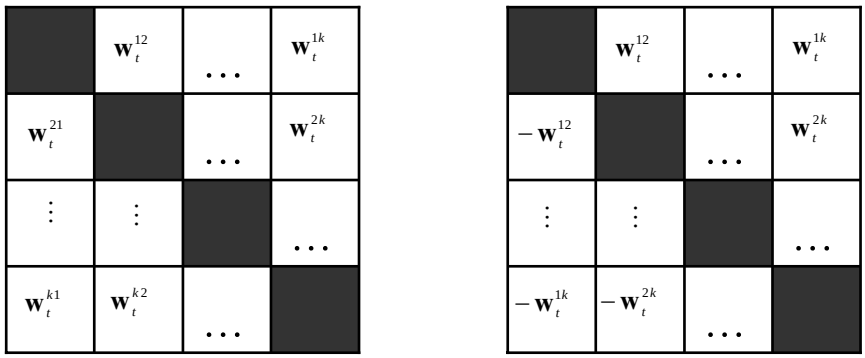
$$\mathbf{w}_{t+1}^{y_t} \leftarrow \mathbf{w}_t^{y_t} + c_t \mathbf{x}_t \text{ , } \mathbf{w}_{t+1}^{\hat{y}_t} \leftarrow \mathbf{w}_t^{\hat{y}_t} - c_t \mathbf{x}_t$$

และ $\mathbf{w}_{t+1}^{r} \leftarrow \mathbf{w}_t^{r}$ สำหรับทุกๆ $r \in Y \setminus \{y_t, \hat{y}_t\}$

วิธีการทุกคู่ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐาน

จากวิธีการหลายเวคเตอร์ซึ่งเป็นรูปแบบหนึ่งของวิธีการลดรูปจากปัญหาการเรียนรู้หลายประเภทไปเป็นปัญหาการเรียนรู้สองประเภทแบบหนึ่งต่อทั้งหมด เราได้ทำการนิยามวิธีการลดรูปแบบหนึ่งต่อหนึ่ง โดยใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐาน เพื่อการแก้ปัญหาการเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญ โดยเราจะเรียกวิธีการดังกล่าวว่า *วิธีการทุกคู่ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐาน* (All-pair based on *IWP*) ซึ่งมีวิธีการดังต่อไปนี้

สำหรับปัญหาการเรียนรู้ k ประเภท สมมติให้ $Y = \{1, 2, \ldots, k\}$ เป็นเซตประเภทของตัวอย่าง วิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักกความสำคัญเป็นฐานจะทำนายโดยใช้เซตของเวคเตอร์น้ำหนัก $W = \{\mathbf{w}^{ij} : i \in Y, j \in Y, i \neq j\}$ ซึ่งประกอบด้วย $k(k\text{-}1)$ นั่นคือ ในการทำนายรอบที่ t สำหรับแต่ละคู่ $i, j \in Y$ จะมีเวคเตอร์น้ำหนัก $\mathbf{w}_t^{ij}$ และ $\mathbf{w}_t^{ji}$ เวคเตอร์ ดังที่แสดงในภาพที่ 4ก โดยที่ $\mathbf{w}_t^{ij} = -\mathbf{w}_t^{ji}$ ดังที่แสดงในภาพที่ 4ข



(ก) (ข)

ภาพที่ 4 เวคเตอร์น้ำหนักของวิธีการทุกคู่

ในการทำนายรอบที่ t ใดๆ จะใช้สมมติฐานการทำนาย $h_t : \Re \to \Re$ ในการคำนวณคะแนนของแต่ละประเภท $r \in Y$ ซึ่งจะมีค่าคะแนนเป็น

$$h_t(\mathbf{x}_t, r) = \sum_{j \in Y \setminus \{r\}} \left\langle \mathbf{w}_t^{rj}, \mathbf{x}_t \right\rangle$$

นั่นคือ คะแนนของประเภท r จะเท่ากับผลรวมของคะแนนของประเภท r ที่ได้จากตัวทำนายแบบสองประเภทแต่ละตัว และจะทำนาย

$$\hat{y}_t = \arg\max_{r \in Y} h_t(\mathbf{x}_t, r)$$

จากนั้นเมื่อได้รับเฉลย y_t ถ้าทำนายผิด $\hat{y}_t \neq y_t$ จะทำการปรับปรุงเวคเตอร์น้ำหนัก โดยสำหรับแต่ละ $r \neq y_t$

$$\mathbf{w}_{t+1}^{y_t r} \leftarrow \mathbf{w}_t^{y_t r} + c_t \mathbf{x}_t$$

และ สำหรับแต่ละ $r \neq \hat{y}_t$

$$\mathbf{w}_{t+1}^{\hat{y}_t r} \leftarrow \mathbf{w}_t^{\hat{y}_t r} - c_t \mathbf{x}_t$$

จากการทำงานของทั้งสองวิธีการที่กล่าวมา จะเห็นว่าวิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐานใช้เวลาในการทำงานเป็น $O(k)$ น้อยกว่าวิธีการทุกคู่ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐานซึ่งใช้เวลาในการทำงานเป็น $O(k^2)$

ผลการวิจัย

สำหรับผลการวิจัยเราจะแบ่งออกเป็น 2 ส่วน ได้แก่ ผลการวิจัยการแก้ปัญหาการเรียนรู้สองประเภทแบบมีน้ำหนักความสำคัญ และผลการวิจัยการแก้ปัญหาการเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญ

ผลการวิจัยการแก้ปัญหาการเรียนรู้สองประเภทแบบมีน้ำหนักความสำคัญ

สำหรับผลการวิจัยการแก้ปัญหาการเรียนรู้สองประเภทแบบมีน้ำหนักความสำคัญแบ่งออกเป็น 2 ส่วน ได้แก่ ผลการวิจัยภาคทดลองและผลการวิจัยภาคทฤษฎี

ผลการวิจัยภาคทดลอง

เราได้ทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึมที่ได้นำเสนอไปในส่วนวิธีการ ได้แก่อัลกอริทึมแบบลดรูปที่ใช้เพอร์เซพตรอนเป็นกล่องดำ และอัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ (*IWP*) รวมทั้งอัลกอริทึมเพอร์เซพตรอน โดยทดลองกับชุดข้อมูลแบบสองประเภท 5 ชุดข้อมูลดังนี้

ชุดข้อมูล KDD Cup 1998 Data

ชุดข้อมูล KDD Cup 1998 Data นำมาจาก Asuncion and Newman (2007) ชุดข้อมูลนี้จะประกอบไปด้วยข้อมูล 95412 รายการ ซึ่งเป็นรายละเอียดของบุคคลที่บริจาคและไม่บริจาคเงินเพื่อการกุศลรวมถึงจำนวนเงินที่ได้บริจาค จุดประสงค์ของการใช้ข้อมูลชุดนี้ คือ เพื่อใช้ในการทดสอบอัลกอริทึมที่ใช้ในการตัดสินใจว่าจะต้องส่ง หรือ ไม่ส่งจดหมายขอรับบริจาคไปให้กับบุคคลใดบ้าง โดยเราจะให้กลุ่มบุคคลที่บริจาคจัดอยู่ในประเภทบวก และ กลุ่มบุคคลที่ไม่บริจาคจัดอยู่ในประเภทลบ สำหรับประเภทบวกค่าใช้จ่ายเมื่ออัลกอริทึมตอบผิดหรือค่าน้ำหนักความสำคัญ จะคิดจากยอดเงินที่บริจาคหักค่าแสตมป์ 0.68 ส่วนประเภทลบถ้าตอบผิดจะต้องเสียค่าแสตมป์ 0.68 ในการทดลองทำการสุ่มข้อมูลร้อยละ 10 ของรายการทั้งหมดมาใช้ และเลือกคุณลักษณะ (Features) จากคุณลักษณะที่มีค่าความสัมพันธ์ (*Correlation*) สูงสุด 10 อันดับแรกมาใช้ โดยกำหนดให้ค่าความสัมพันธ์ของคุณลักษณะที่ i มีค่าดังนี้

$$Correlation^{(i)} = \sum_{t=1}^{m} x_t^i y_t$$

โดย x_t^i คือค่าของคุณลักษณะที่ i ของตัวอย่าง $\mathbf{x}_t$

ชุดข้อมูล Auto-MPG

ชุดข้อมูล Auto-MPG นำมาจาก Data Asuncion and Newman (2007) ชุดข้อมูลประกอบไปด้วยข้อมูล 398 รายการ แต่ละรายการมี 8 คุณลักษณะ ซึ่งเป็นข้อมูลเกี่ยวกับการใช้เชื้อเพลิงของรถยนต์แต่ละคันซึ่งมีหน่วยเป็นไมล์ต่อแกลลอน (MPG) โดยปกติชุดข้อมูลนี้ใช้กับงาน Regression นั่นคือการทำนายการใช้เชื้อเพลิงของรถยนต์แต่ละคัน แต่ในการทดลองนี้นำมาใช้กับงานจำแนกประเภท โดยใช้ข้อมูล 392 รายการ (เนื่องจากอีก 6 รายการมีคุณลักษณะไม่ครบ) สำหรับตัวอย่าง $\mathbf{x}_t$ กำหนดให้ประเภท y_t และ น้ำหนักความสำคัญ c_t มีค่าดังนี้

$$y_t = \text{sgn}\left(mpg^{(t)} - \overline{mpg}\right) \text{ และ } c_t = \left|mpg^{(t)} - \overline{mpg}\right|$$

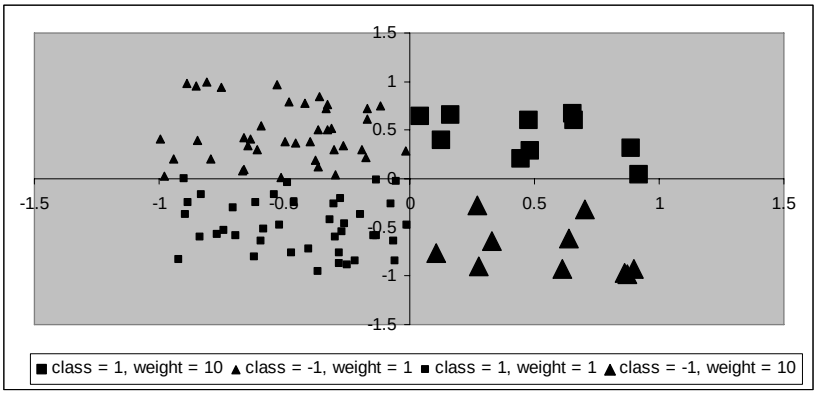
โดยที่ $mpg^{(t)}$ คือค่าคุณลักษณะ mpg ของตัวอย่าง $\mathbf{x}_t$ และ $\overline{mpg}$ คือ ค่าเฉลี่ยของคุณลักษณะ mpg

ชุดข้อมูลสังเคราะห์

เพื่อให้เปรียบเทียบประสิทธิภาพของวิธีการต่างๆ เมื่อทำนายตัวอย่างที่มีน้ำหนักความสำคัญได้อย่างชัดเจน เราทำการสร้างชุดข้อมูลสังเคราะห์ที่มีการกระจายของน้ำหนักความสำคัญในชุดข้อมูลขัดแย้งกับการกระจายของประเภท นั่นคือสังเคราะห์ให้ข้อมูลที่มีน้ำหนักความสำคัญมากมีจำนวนตัวอย่างน้อย ส่วนข้อมูลที่มีน้ำหนักความสำคัญน้อยมีจำนวนตัวอย่างมาก กล่าวคือเราต้องการให้อัลกอริทึมที่สนใจน้ำหนักความสำคัญของตัวอย่าง มีจำนวนครั้งในการทำนายผิดมาก แต่เสียค่าใช้จ่ายรวมในการทำนายผิดน้อย ส่วนอัลกอริทึมที่ไม่ได้สนใจน้ำหนักความสำคัญของตัวอย่างเสียค่าใช้จ่ายรวมในการทำนายผิดมาก แต่มีจำนวนครั้งในการทำนายผิดน้อย เราทำการสังเคราะห์ข้อมูล 3 ชุดข้อมูล

ข้อมูลสังเคราะห์ชุดที่ 1

ข้อมูลสังเคราะห์ชุดที่ 1 ประกอบไปด้วย 10000 ตัวอย่าง แบ่งเป็นข้อมูล 2 ประเภท ได้แก่ ประเภท -1 และประเภท +1 แต่ละประเภทมีข้อมูล 5000 ตัวอย่าง ซึ่งแบ่งเป็นตัวอย่างที่มีน้ำหนักที่มีน้ำหนักความสำคัญ 10 จำนวน 1000 ตัวอย่าง และตัวอย่างที่มีน้ำหนักที่มีน้ำหนักความสำคัญ 1 อีก 4000 ตัวอย่าง โดยมีการกระจายตัวดังแสดงในภาพที่ 5



ภาพที่ 5 การกระจายตัวของข้อมูลสังเคราะห์ชุดที่ 1

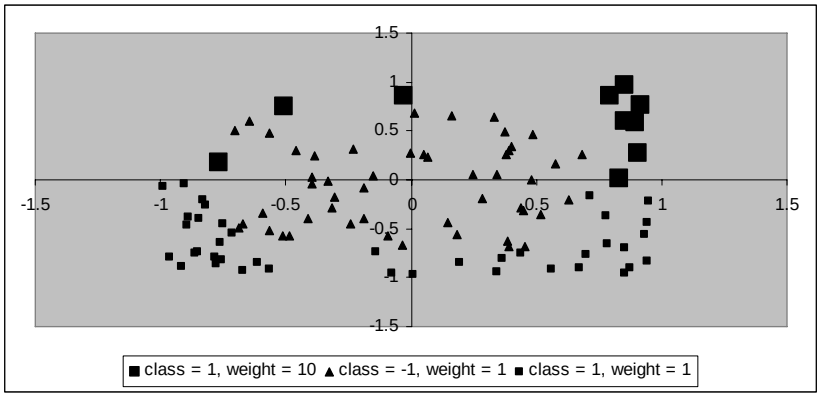
ข้อมูลสังเคราะห์ชุดที่ 2

ข้อมูลสังเคราะห์ชุดที่ 2 ประกอบไปด้วยข้อมูลประกอบไปด้วย 10000 ตัวอย่าง แบ่งเป็น 2 ประเภท ได้แก่ประเภท -1 และประเภท +1 แต่ละประเภทมีข้อมูล 5000 ตัวอย่าง สำหรับประเภท +1 ประกอบด้วยตัวอย่างที่มีน้ำหนักที่มีน้ำหนักความสำคัญ 10 จำนวน 1000 ตัวอย่าง และตัวอย่างที่มีน้ำหนักที่มีน้ำหนักความสำคัญ 1 อีก 4000 ตัวอย่าง ส่วนประเภท -1 แต่ละตัวอย่างมีน้ำหนักความสำคัญเป็น 1 โดยมีการกระจายตัวดังแสดงในภาพที่ 6

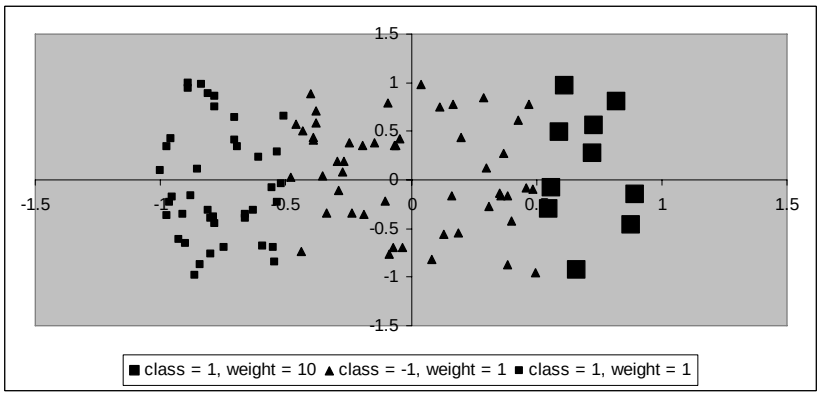
ข้อมูลสังเคราะห์ชุดที่ 3

ข้อมูลสังเคราะห์ชุดที่ 3 ประกอบไปด้วยข้อมูลประกอบไปด้วย 10000 ตัวอย่าง แบ่งเป็น 2 ประเภท ได้แก่ประเภท -1 และประเภท +1 แต่ละประเภทมีข้อมูล 5000 ตัวอย่าง สำหรับประเภท +1 ประกอบด้วยตัวอย่างที่มีน้ำหนักที่มีน้ำหนักความสำคัญ 10 จำนวน 1000 ตัวอย่าง และตัวอย่างที่มี

น้ำหนักที่มีน้ำหนักความสำคัญ 1 อีก 4000 ตัวอย่าง ส่วนประเภท -1 แต่ละตัวอย่างมีน้ำหนักความสำคัญเป็น 1 โดยมีการกระจายตัวดังแสดงในภาพที่ 7



ภาพที่ 6 การกระจายตัวของข้อมูลสังเคราะห์ชุดที่ 2



ภาพที่ 7 การกระจายตัวของข้อมูลสังเคราะห์ชุดที่ 3

ผลการทดลอง

การเปรียบเทียบประสิทธิภาพการทำนายของ 3 อัลกอริทึม ได้แก่ อัลกอริทึมแบบลดรูปที่ใช้เพอร์เซพตรอนเป็นกล่องดำ, อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ (*IWP*) และอัลกอริทึมเพอร์เซพตรอน สำหรับแต่ละชุดข้อมูลที่กล่าวมาในข้างต้น เราทำโดยการนับจำนวนครั้งและคิดค่าใช้จ่ายรวมที่แต่ละวิธีการทำนายผิดได้ผลดังที่แสดงในตารางที่ 1 เมื่อพิจารณาร้อยละของจำนวนที่แต่ละวิธีการ ทำนายผิด (%ตอบผิด) จะเห็นว่าโดยส่วนใหญ่อัลกอริทึมเพอร์เซพตรอนมี

จำนวนครั้งในการทำนายผิดใกล้เคียงกับอัลกอริทึมแบบลดรูปซึ่งน้อยกว่าอัลกอริทึม *IWP* แต่เมื่อพิจารณาร้อยละของค่าใช้จ่ายรวมที่แต่ละวิธีการทำนายผิด (%ค่าใช้จ่าย) จะเห็นว่าอัลกอริทึม *IWP* กลับมีค่าใช้จ่ายรวมในการทำนายผิดน้อยกว่าอีกสองอัลกอริทึมที่ให้ผลการทดลองใกล้เคียงกัน จากผลการทดลองที่กล่าวมานี้ เราสามารถกล่าวได้ว่า อัลกอริทึม *IWP* ทำนายข้อมูลที่ตัวอย่างมีน้ำหนักความสำคัญได้ดีกว่าอัลกอริทึมแบบลดรูป และอัลกอริทึมเพอร์เซพตรอน

ตารางที่ 1 ผลการเปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึมแบบลดรูป, อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ และอัลกอริทึมเพอร์เซพตรอน กับชุดข้อมูลแบบสองประเภท

ชุดข้อมูล	*IWP*		ลดรูป		เพอร์เซพตรอน	
	%ตอบผิด	%ค่าใช้จ่าย	%ตอบผิด	%ค่าใช้จ่าย	%ตอบผิด	%ค่าใช้จ่าย
KDD-98	10.33	54.90	52.66	51.18	**9.59**	54.92
Auto-Mpg	23.21	17.00	**20.15**	**13.02**	26.60	16.37
สังเคราะห์1	35.70	56.25	61.31	**44.04**	**33.71**	57.79
สังเคราะห์2	43.01	56.06	57.17	**40.03**	**41.81**	57.37
สังเคราะห์3	40.64	56.31	56.80	**39.00**	**38.27**	58.57

สังเกตว่า อัลกอริทึมแบบลดรูปมีประสิทธิภาพการทำนายไม่แตกต่างจากอัลกอริทึมเพอร์เซพตรอนเท่าใดนัก ซึ่งประเด็นนี้เราจะทำการวิเคราะห์ในส่วนผลการวิจัยภาคทฤษฎี

ผลการวิจัยภาคทฤษฎี

ในส่วนของผลการวิจัยภาคทฤษฎีเราจะทำการวิเคราะห์ประสิทธิภาพการทำนายของอัลกอริทึมแบบลดรูปที่ใช้เพอร์เซพตรอนเป็นกล่องดำ, อัลกอริทึมเพอร์เซพตรอนเซพตรอนแบบมีน้ำหนักความสำคัญ (*IWP*) และอัลกอริทึมเพอร์เซพตรอน

การวัดประสิทธิภาพของอัลกอริทึมการเรียนรู้แบบออนไลน์

โดยทั่วไปการระบุประสิทธิภาพของอัลกอริทึมการเรียนรู้ออนไลน์ A ใดๆ จะระบุโดยเทียบกับสมมติฐานจากบางเซตสมมติฐาน H กล่าวคือ ให้ X เป็นเซตของตัวอย่างทั้งหมดที่เป็นไปได้ และ Y เป็นเซตของประเภทของตัวอย่างทั้งหมดที่เป็นไปได้ สำหรับบางฟังก์ชัน ℓ ที่ระบุความสูญเสียของ h บนลำดับของข้อมูล $S = ((x_1, y_1), \ldots, (\mathbf{x}_m, y_m)) \in (X \times Y)^m$ ใดๆ และฟังก์ชัน $Loss$ ที่ระบุความสูญเสียของ A มีฟังก์ชัน $g : \Re \to \Re$ ที่ถ้ามีสมมติฐาน $h \in H$ ที่ทำนายได้ดีใน S นั่นคือ $\ell(h,S) \le L$ ค่าความสูญเสีย $Loss(A,S)$ ของอัลกอริทึม A บนลำดับ S จะไม่เกิน $g(L)$ โดยทั่วไปแล้ว ฟังก์ชัน ℓ จะถูกนิยามให้เป็นผลรวมของค่าความสูญเสียบนข้อมูลแต่ละตัว กล่าวคือ $\ell(h,S) = \sum_{t=1}^{m} \ell(h,(x_t, y_t))$ ในงานวิจัยนี้จะทำการวิเคราะห์โดยใช้ความจริงดังกล่าวด้วย

ในงานวิจัยนี้เราจะทำการวิเคราะห์ประสิทธิภาพของอัลกอริทึม โดยใช้ค่าความสูญเสียแบบมีน้ำหนัก $Loss^w$ บนลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก $Q = ((x_1, y_1, c_1), \ldots, (\mathbf{x}_m, y_m, c_m)) \in (X \times Y \times [0,\infty))^m$ ใดๆ เมื่อเทียบกับฟังก์ชันความสูญเสียแบบมีน้ำหนัก $\ell^w(h,Q)$ ของ $h \in H$ บน Q โดยที่ $Loss^w$ มีนิยามเป็นความสูญเสีย $Loss$ คูณด้วยน้ำหนักของตัวอย่าง และ ฟังก์ชันความสูญเสียแบบมีน้ำหนักมีนิยามเป็น

$$\ell^w(h, x_t) = c_t \ell(h, x_t) \tag{4}$$

ต่อไปเราจะทำการวิเคราะห์ค่าความสูญเสียแบบมีน้ำหนักความสำคัญของแต่ละอัลกอริทึม

การรับประกันค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของอัลกอริทึมแบบลดรูป

จากการทำงานของอัลกอริทึมแบบลดรูปที่ได้นำเสนอในหัวข้อวิธีการ เราสามารถพิสูจน์ค่าความสูญเสียแบบมีน้ำหนัก $Loss^w(A^w,Q)$ ของอัลกอริทึมแบบลดรูปซึ่งในที่นี้จะเขียนแทนด้วย A^w บนลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก Q ใดๆ เมื่อเทียบกับค่าความสูญเสียแบบมีน้ำหนักของตัวทำนาย $h \in H$ ใดๆ ได้ด้วยการคิดเทียบค่าความสูญเสียของอัลกอริทึมการเรียนรู้แบบออนไลน์ A ที่ A^w ใช้เป็นกล่องดำ กับค่าความสูญเสียของ h

ให้ฟังก์ชัน $g:\Re \to \Re$ เป็นฟังก์ชันที่ใช้ระบุขีดจำกัดบนของความสูญเสียของ A และสมมติให้สำหรับตัวอย่างใดๆ บนลำดับ Q มีค่าเป็นจำนวนเต็มบวก เราสารมารถรับประกันค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของวิธีการแบบลดรูป A^w ได้ดังนี้

ทฤษฎีบทที่ 3 *สำหรับลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก* $Q=((\mathbf{x}_1,y_1,c_1),(\mathbf{x}_2,y_2,c_2),\ldots,(\mathbf{x}_m,y_m,c_m))\in(X\times Y\times I^+)^m$ *ใดๆ ถ้ามีตัวทำนาย* $h\in H$ *ที่มี* $\ell^w(h,Q)\le L$ *เราจะได้ว่า*

$$Loss^w(A^w,Q)\le c_{\max}\cdot g(L)$$

เมื่อ $c_{\max}=\max_{t\in\{1,\ldots,m\}} c_t$

บทพิสูจน์ พิจารณาลำดับลำดับข้อมูลป้อนเข้าแบบไม่มีน้ำหนักความสำคัญ $S=(x_{11},y_1),\ldots\ \ldots,(x_{1c_1},y_1),\ldots,(x_{m1},y_m),\ldots,(x_{mc_m},y_m)$ ที่ A^w สร้างจาก Q เรียกลำดับย่อย $(x_{t1},y_t),\ldots,(x_{tc_t},y_t)$ บน S ว่า *ลำดับย่อย* X_t สังเกตว่าการทำงานของ A^w ในการปรับค่าตัวทำนายบน Q จะเหมือนกับการทำงานของ A บน S

ในขั้นแรกเราจะเทียบ $\ell(h,S)$ กับ $\ell^w(h,Q)$ จะเห็นว่าถ้าตัวทำนาย h ทำนายตัวอย่างแรกบน X_t ผิด แล้ว h จะทำนายตัวอย่างอื่นๆ บน X_t และตัวอย่าง x_t บน Q ผิดด้วย และถ้า h ทำนายตัวอย่างแรกบน X_t ถูก แล้ว h จะทำนายตัวอย่างอื่นๆ บน X_t และตัวอย่าง x_t บน Q ถูกด้วย ดังนั้นจะได้ว่า

$$\ell(h,X_t)=\sum_{i=1}^{c_t}\ell(h,x_t)=c_t\ell(h,x_t)=\ell^w(h,x_t)$$

นั่นคือ $\ell(h,S)=\ell^w(h,Q)$

ถัดมาเราจะเทียบ $Loss(A,S)$ กับ $Loss^w(A^w,Q)$ พิจารณาตัวอย่างแบบมีน้ำหนัก (x_t,y_t,c_t) และลำดับย่อย X_t เนื่องจาก A^w ใช้คำตอบของ $h_t(x_{t1})$ เป็นคำตอบของการทำนายตัวอย่าง x_t บน Q พิจารณากรณีที่ A^w ทำนายตัวอย่าง x_t ผิด เราจะได้ $Loss^w(A^w,x_t)$ จะมีค่าเป็น c_t อย่างไรก็ตามสังเกตว่าถ้า A^w ทำนายผิดบนตัวอย่างดังกล่าว เราจะได้ว่า A ทำนายผิดในตัวอย่าง x_{t1} ด้วยเสมอ นั่นคือ $Loss(A,X_t)\ge 1$ ดังนั้น จะได้

$$Loss^{w}(A^{w},Q) \le \sum_{t=1}^{m} c_t \cdot Loss(A,X_t) \le c_{\max} \cdot Loss(A,S)$$

โดยการแทนค่าขีดจำกัดของความสูญเสียของ A เราจะได้ว่าทฤษฎีบทถูกต้อง เนื่องจาก $Loss^{w}(A^{w},Q) \le c_{\max} \cdot Loss(A,S) \le c_{\max} \cdot g(L)$ ■

การรับประกันค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของอัลกอริทึมการเรียนรู้แบบออนไลน์ใดๆ

ในส่วนนี้เราจะทำการวิเคราะห์ประเด็นสำคัญในส่วนของผลการวิจัยภาคทดลอง นั่นคือเหตุใดอัลกอริทึมแบบลดรูปซึ่งเรียนรู้โดยการทำซ้ำตัวอย่างที่ทำนายผิดตามค่าน้ำหนักความสำคัญของตัวอย่างนั้น จึงมีประสิทธิภาพการทำนายใกล้เคียงกับ อัลกอริทึมเพอร์เซพตรอนซึ่งไม่ได้สนใจน้ำหนักความสำคัญของตัวอย่าง โดยเราจะทำการวิเคราะห์ค่าขีดจำกัดบนของความผิดพลาดของอัลกอริทึมเพอร์เซพตรอน ซึ่งในที่นี้เราจะวิเคราะห์ในกรณีทั่วไป นั่นคือเราจะทำการวิเคราะห์ค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของอัลกอริทึมการเรียนรู้แบบออนไลน์ใดๆ ที่ไม่สนใจค่าน้ำหนักความสำคัญของตัวอย่าง

สำหรับอัลกอริทึมการเรียนรู้แบบออนไลน์ A ใดๆ ที่ไม่สนใจค่าน้ำหนักความสำคัญของตัวอย่าง ให้ $g:\Re \to \Re$ เป็นฟังก์ชันขีดจำกัดความสูญเสียของ A เมื่อเทียบกับตัวทำนาย $h \in H$ ใดๆ และสำหรับลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก Q ใดๆ สมมติให้สำหรับตัวอย่างใดๆ บนลำดับ Q มีค่าเป็นจำนวนเต็มบวก เราสามารถรับประกันค่าขีดจำกัดบนของความผิดพลาดของ A ได้ดังนี้

ทฤษฎีบทที่ 4 *สำหรับลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก* $Q = ((\mathbf{x}_1, y_1, c_1), (\mathbf{x}_2, y_2, c_2), \ldots, (\mathbf{x}_m, y_m, c_m)) \in (X \times Y \times I^{+})^{m}$ *ใดๆ ถ้ามีตัวทำนาย* $h \in H$ *ที่มี* $\ell^{w}(h,Q) \le L$ *จะได้ว่า*

$$Loss^{w}(A,Q) \le c_{\max} \cdot g(L)$$

เมื่อ $c_{\max} = \max_{t \in \{1,\ldots,m\}} c_t$

บทพิสูจน์ เนื่องจาก A เป็นอัลกอริทึมที่ไม่สนใจน้ำหนักความสำคัญของตัวอย่าง ดังนั้นจะให้ $S=(x_1,y_1),(x_2,y_2),...,(x_m,y_m)$ เป็นลำดับข้อมูลป้อนเข้าแบบไม่มีน้ำหนัก ซึ่งเกิดจากลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก Q ที่ตัดค่าน้ำหนักความสำคัญของแต่ละตัวอย่างออกไป

ในขั้นแรกเราจะเทียบ $\ell(h,S)$ กับ $\ell^w(h,Q)$ เมื่อพิจารณาการทำนายของ h บน S และบน Q จะเห็นว่าถ้า h ทำนายตัวอย่าง (x_t, y_t) บน S ผิด แล้ว h จะทำนาย (x_t, y_t, c_t) บน Q ผิดด้วย และเนื่องจากสำหรับตัวอย่าง (x_t, y_t, c_t) ใดๆ มีค่า c_t เป็นจำนวนเต็มบวก ดังนั้นจะได้ว่า

$$\ell^w(h,Q)=\sum_{t=1}^{m} c_t \ell(h,x_t) \geq \ell(h,S)$$

ถัดมาเราจะเทียบ $Loss(A, S)$ กับ $Loss^w(A, Q)$ พิจารณาการทำนายของ A บน S และบน Q จะเห็นว่าถ้า A ทำนายตัวอย่าง (x_t, y_t) บน S ผิด แล้ว A จะทำนาย (x_t, y_t, c_t) บน Q ผิดด้วย นั่นคือ

$$Loss^w(A,Q)=\sum_{t=1}^{m} c_t \cdot Loss(A,x_t) \leq c_{\max} \cdot Loss(A,S)$$

เมื่อแทนค่าขีดจำกัดของความสูญเสียของ A เราจะได้ว่าทฤษฎีบทถูกต้อง เนื่องจาก $Loss^w(A^w,Q) \leq c_{\max} \cdot Loss(A,S) \leq c_{\max} \cdot g(L)$ ■

จากทฤษฎีบทที่ 4 จะเห็นว่าค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของอัลกอริทึมการเรียนรู้แบบออนไลน์ใดๆ ที่ไม่สนใจค่าน้ำหนักความสำคัญของตัวอย่างมีค่าเท่ากับค่าขีดจำกัดบนของของความสูญเสียแบบมีน้ำหนักของอัลกอริทึมแบบลดรูปในทฤษฎีบทที่ 3 ทำให้เราสามารถสรุปประเด็นที่กล่าวถึงในตอนต้นได้ว่า ไม่ว่าจะนำอัลกอริทึมการเรียนรู้แบบออนไลน์ A ใดๆ ที่ไม่สนใจค่าน้ำหนักความสำคัญของตัวอย่าง ประสิทธิภาพการทำนายอย่างแย่นั้น ไม่ได้แย่ไปกว่าประสิทธิภาพการทำนายของอัลกอริทึมแบบลดรูปที่ใช้ A เป็นกล่องดำ

การรับประกันค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของอัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญแบบง่าย

จากการทำงานของอัลกอริทึม *IWP* ที่ได้นำเสนอไปในหัวข้อวิธีการ จะเห็นว่า *IWP* ทำงานใกล้เคียงกับอัลกอริทึมเพอร์เซพตรอน แต่ปรับปรุงในส่วนของการปรับเวคเตอร์น้ำหนักในรอบที่มี

การทำนายผิด ซึ่งจะปรับด้วย $c\mathbf{x}$ แทนที่จะเป็น $\mathbf{x}$ ดังเช่นเพอร์เซพตรอน ดังนั้นสำหรับลำดับป้อนเข้าแบบมีน้ำหนัก $Q = ((\mathbf{x}_1,y_1,c_1), (\mathbf{x}_2,y_2,c_2), \ldots, (\mathbf{x}_m,y_m,c_m))$ สามารถพิจารณาได้ว่า *IWP* เป็นเพอร์เซพตรอน ที่มีลำดับข้อมูลป้อนเข้าแบบไม่มีน้ำหนักความสำคัญ $S = (c_1\mathbf{x}_1, c_2\mathbf{x}_2,\ldots, c_m\mathbf{x}_m)$ ซึ่งทำให้สามารถพิสูจน์ประสิทธิภาพของ *IWP* ได้ โดยใช้ทฤษฎีบทที่ 5 เป็นกล่องดำในการพิสูจน์ สำหรับทฤษฎีบทที่ 5 นั้นเป็นรูปแบบเฉพาะของทฤษีบทที่ว่าด้วยขีดจำกัดบนของความสูญเสียอัลกอริทึมพีนอร์ม ใน Gentile (2003) ซึ่งทฤษฎีบทดังกล่าว ให้ขีดจำกัดบนของความสูญเสียได้ทั้งกรณีที่มีและไม่มีตัวทำนายเชิงเส้นที่แบ่งกลุ่มตัวอย่างสองกลุ่มออกจากกันได้ การวัดประสิทธิภาพแบบนี้จะคิดเทียบกับสมมติฐาน h ซึ่งนิยามโดยเวคเตอร์ $\mathbf{u}$ โดยใช้ฟังก์ชันความเบี่ยงเบนซึ่งนิยามเป็น

$$D_\gamma(\mathbf{u};(\mathbf{x}_t, y_t)) = \max\{0, \gamma - y_t\langle \mathbf{u}, \mathbf{x}_t\rangle\} \tag{5}$$

โดย γ เป็นค่าคงที่ใดๆ โดยที่ $\gamma > 0$ สำหรับฟังก์ชันความเบี่ยงเบนตามสมการ (5) นี้จะเป็นรูปแบบทั่วไปของฟังก์ชันความสูญเสียแบบฮิงก์ตามสมการ (1) ซึ่งได้กล่าวถึงในหัวข้อการตรวจเอกสาร เมื่อ $\gamma = 1$

ทฤษฎีบทที่ 5 Gentile (2003) *ให้* $S = ((\mathbf{x}_t, y_t),\ldots,(\mathbf{x}_m, y_m)) \in (R^n \times \{1,-1\})^m$ *เป็นลำดับข้อมูลป้อนเข้าของเพอร์เซพตรอน,* M *เป็นเซตของรอบที่เพอร์เซพตรอนตอบผิด และ* $R = \max_{t\in M}\|\mathbf{x}_t\|_2$ *สำหรับเวคเตอร์* $\mathbf{u}$ *ใด ๆ ที่มี* $C = R^2\|\mathbf{u}\|_2^2$ *และ* $L = \sum_{t\in M} D_\gamma(\mathbf{u};(\mathbf{x}_t, y_t))$ *ถ้า* $\mathbf{w}_1 = 0$ *จำนวนรอบที่เพอร์เซพตรอนตอบผิด* $|M| \le \frac{L}{\gamma} + \frac{C}{\gamma^2} + \frac{1}{\gamma^2}\sqrt{\gamma LC}$

จากทฤษฎีบทที่ 5 จะเห็นว่าค่าความเบี่ยงเบนของตัวทำนายเชิงเส้น $\mathbf{u}$ จะคิดเฉพาะครั้งที่เพอร์เซพตรอนตอบผิดเท่านั้น และจำนวนรอบที่เพอร์เซพตรอนตอบผิดก็จะคิดเทียบกับค่าดังกล่าว ในกรณีเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ ประสิทธิภาพของอัลกอริทึมจะคิดเทียบกับสมมติฐาน h ซึ่งนิยามโดยเวคเตอร์ $\mathbf{u}$ จากสมการ (5) จะได้ว่า ค่าความเบี่ยงเบนของ h แบบมีน้ำหนักสำหรับตัวอย่าง $\mathbf{x}_t$ จะมีค่าเป็น

$$D_{c_t}(\mathbf{u};(c_t\mathbf{x}_t, y_t)) = \max\{0, c_t - y_t \cdot \langle c_t\mathbf{x}_t, \mathbf{u}\rangle\} \tag{6}$$

ทฤษฎีบทที่ 6 *สำหรับลำดับป้อนเข้าแบบมีน้ำหนัก* $Q = ((\mathbf{x}_1, y_1, c_1), (\mathbf{x}_2, y_2, c_2), \ldots, (\mathbf{x}_m, y_m, c_m)) \in (\Re^n \times \{-1,+1\} \times [0,\infty))^m$ *ถ้ามีตัวทำนาย* $h \in H$ *ใดๆ ที่นิยามโดย* $\mathbf{u}$ *ที่มี* $\ell^w(h,Q) \le L$ *แล้ว จะได้ว่า*

$$Loss^w(IWP,Q) \le \theta L + c_{\max}\theta^2 C + \theta^2\sqrt{c_{\min} LC}$$

เมื่อ $C = \|\mathbf{u}\|_2^2 \cdot \max_{t\in\{1,..,m\}} \|\mathbf{x}_t\|_2^2$, $c_{\max} = \max_{t\in\{1,...,m\}} c_t$, $c_{\min} = \min_{t\in\{1,...,m\}} c_t$ *และ* $\theta = \frac{c_{\max}}{c_{\min}}$

บทพิสูจน์ ให้ลำดับข้อมูลป้อนเข้าแบบไม่มีน้ำหนัก $S = (c_1\mathbf{x}_1, \ldots, c_m\mathbf{x}_m)$ เป็นลำดับข้อมูลป้อนเข้าของเพอร์เซพตรอนและ ให้ M แทนเซตของรอบที่ *IWP* ตอบผิดพลาด พิจารณาตัวทำนาย $h \in H$ ใด ๆ ที่นิยามโดยเวคเตอร์ $\mathbf{u}$ ให้

$$L' = \sum_{t\in M} D_{c_{\min}}(u;(c_t\mathbf{x}_t, y_t))$$

แทนความเบี่ยงเบนที่ระดับ $\gamma = c_{\min}$ ของ h จากทฤษฎีบทที่ 5 จะได้ว่า

$$|M| \le \frac{L'}{c_{\min}} + \frac{C'}{c_{\min}^2} + \frac{1}{c_{\min}^2}\sqrt{c_{\min} L'C'}$$

เมื่อ $C' = c_{\max}^2 \cdot \|\mathbf{u}\|^2 \cdot \max_{t\in M} \|\mathbf{x}_t\|^2$

อย่างไรก็ตาม แต่ละครั้งที่ *IWP* ตอบผิดบนตัวอย่าง $\mathbf{x}_t$ จะมีค่าความสูญเสียเกิดขึ้น ไม่เกิน $c_{\max}$ นั่นคือ

$$Loss(IWP,Q) \le c_{\max}\left[\frac{L'}{c_{\min}} + \frac{C'}{c_{\min}^2} + \frac{1}{c_{\min}^2}\sqrt{c_{\min} L'C'}\right]$$

ทฤษฎีบทเป็นจริง โดยสังเกตว่า

$$L \ge \ell^w(h,Q)$$

$$= \sum_{t=1}^{m} D_{c_t}(\mathbf{u};(c_t\mathbf{x}_t, y_t)) \qquad \text{จากสมการ (6)}$$

$$\geq \sum_{t\in M} D_{c_t}(\mathbf{u};(c_t\mathbf{x}_t, y_t)) \qquad \blacksquare$$

ถ้าใช้อัลกอริทึมแบบลดรูปทำนายบนลำดับ Q โดยใช้เพอร์เซพตรอนเป็นกล่องดำ จะได้ขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักเป็น $c_{\max}(L+C+\sqrt{LC})$ สังเกตว่าอัลกอริทึม *IWP* จะให้ผลลัพธ์ที่ดีกว่า ในกรณีที่ $c_{\max}$ มีค่าใกล้เคียงกับ $c_{\min}$ อย่างไรก็ตาม เราสามารถปรับปรุงค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของ *IWP* ได้อีกดังจะกล่าวถึงในส่วนถัดไป

การปรับปรุงค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของอัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ

จากทฤษฎีบทที่ 6 จะเห็นว่าค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของอัลกอริทึม *IWP* ขึ้นอยู่กับค่า $c_{\max}/c_{\min}$ ซึ่งเราได้ทำการปรับปรุงค่าขีดจำกัดบนดังกล่าวและสามารถกำจัด $c_{\max}/c_{\min}$ ออกไปได้ โดยการปรับปรุงบทพิสูจน์ของทฤษีบทที่ว่าด้วยขีดจำกัดบนของความสูญเสียอัลกอริทึมพีนอร์ม ใน Gentile (2003) ซึ่งได้ผลออกมาดังทฤษฎีบทต่อไปนี้

ทฤษฎีบทที่ 7 *สำหรับลำดับข้อมูลป้อนเข้าแบบมีน้ำหนักความ* $Q=((\mathbf{x}_1,y_1,c_1),(\mathbf{x}_2,y_2,c_2),\ldots,(\mathbf{x}_m,y_m,c_m))\in(\Re^n\times\{-1,+1\}\times[0,\infty))^m$ *ใดๆ ถ้า* $R=\max_t\|\mathbf{x}_t\|_2$ *และมีตัวทำนายเชิงเส้น* $h\in H$ *ใดๆ ที่นิยามด้วยเวคเตอร์* $\mathbf{u}$ *โดยที่* $\ell^w(h,Q)\leq L$ *และให้* $C=R^2\|\mathbf{u}\|_2^2$ *แล้วจะได้ว่า*

$$Loss^w(IWP,Q)\leq L+c_{\max}C+\sqrt{c_{\max}LC}$$

เมื่อ $c_{\max}=\max_{t\in\{1,\ldots,m\}}c_t$

สำหรับการพิสูจน์ทฤษฎีบทที่ 7 เราจะพิสูจน์ผ่านทางอีกทฤษฎีบทหนึ่งซึ่งเป็นการณีทั่วไปของทฤษฎีบทดังกล่าวโดยจะใช้วิธีการพิสูจน์ในแนวทางเดียวกับบทพิสูจน์ใน Gentile (2003) โดยทำการนิยามฟังก์ชันความเบี่ยนเบนแบบมีน้ำหนักใหม่ดังนี้

สำหรับตัวอย่าง $(\mathbf{x}, y, c)$ ใดๆ, เวคเตอร์ $\mathbf{u}$ ใดๆ และค่าคงที่ $\gamma > 0$ ใดๆ ค่าความเบี่ยงเบนแบบมีน้ำหนักความสำคัญ

$$D_{\gamma}(\mathbf{u};(\mathbf{x},y,c)) = c\big(0, \gamma - y\langle \mathbf{u}, \mathbf{x}\rangle\big)_{+} \tag{7}$$

ฟังก์ชันความเบี่ยงเบนแบบมีน้ำหนักในสมการ (7) นี้ เป็นกรณีทั่วไปของฟังก์ชันความเบี่ยงเบนแบบมีน้ำหนักในสมการ (6)

ในการพิสูจน์ทฤษฎีบทที่จะกล่าวต่อไป เราจะวัดค่าระยะห่างของเวคเตอร์ $\mathbf{u}$ และ $\mathbf{w}$ ใดๆ ดังนี้

$$d(\mathbf{u},\mathbf{w}) = \frac{1}{2}\|\mathbf{u}\|_2^2 + \frac{1}{2}\|\mathbf{w}\|_2^2 - \langle \mathbf{u}, \mathbf{w}\rangle \tag{8}$$

ฟังก์ชันนี้เป็นรูปแบบเฉพาะของการลู่ออกของเบรกแมนในสมการ (2) ซึ่งมี $p = 2$ และใช้ทฤษฎีบทย่อย 2 ทฤษฎีดังนี้

ทฤษฎีบทย่อยที่ 3 *ให้* $\mathbf{u}, \mathbf{w}, \mathbf{x} \in \Re^n$, $a \in \Re$ *และ* $\mathbf{w}' = \mathbf{w} + a\mathbf{x}$ *แล้ว จะได้ว่า*

$$a\big(\langle \mathbf{u}, \mathbf{x}\rangle - \langle \mathbf{w}, \mathbf{x}\rangle\big) = d(\mathbf{u},\mathbf{w}) - d(\mathbf{u},\mathbf{w}') + d(\mathbf{w},\mathbf{w}')$$

ทฤษฎีบทย่อยที่ 4 *ให้* $\mathbf{w}, \mathbf{x} \in \Re^n$, $a \in \Re$ *และ* $\mathbf{w}' = \mathbf{w} + a\mathbf{x}$ *แล้ว จะได้ว่า*

$$d(\mathbf{w},\mathbf{w}') \le \frac{a^2}{2}\|\mathbf{x}\|_2^2$$

สังเกตุว่าทฤษฎีบทย่อยทั้งสองข้างต้นเป็นรูปแบบเฉพาะ เมื่อ $p = q = 2$ ของทฤษฎีบทย่อยที่ 1 และทฤษฎีบทย่อยที่ 2 ซึ่งกล่าวไว้ในส่วนการตรวจเอกสารตามลำดับ ต่อไปเราจะทำการพิสูจน์ทฤษฎีบทซึ่งเป็นกรณีทั่วไปของทฤษฎีบทที่ 7 โดยในทฤษฎีบทนี้เราจะทำการพิสูจน์ค่าขีดจำกัดบนของความความสูญเสียแบบมีน้ำหนักของอัลกอริทึม *IWP* ที่มีเวคเตอร์น้ำหนักเริ่มต้น $\mathbf{w}_1 \in \Re^n$ และ

สำหรับตัวอย่าง $\mathbf{x}_t$ ใดๆ มีน้ำหนักความสำคัญ $0 \le c_t \le 1$ โดยเราจะเรียกอัลกอริทึมดังกล่าวนี้ว่า *IWP'* ในการพิสูจน์เราจะใช้ฟังก์ชันความเบี่ยนเบนแบบมีน้ำหนักดังที่นิยามในสมการ (7)

ทฤษฎีบทที่ 8 *สำหรับลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก* $Q = ((\mathbf{x}_1, y_1, c_1), (\mathbf{x}_2, y_2, c_2), \ldots, (\mathbf{x}_m, y_m, c_m)) \in (\Re^n \times \{-1,+1\} \times [0,1])^m$ *ใดๆ, ให้* M *เป็นเซตของรอบที่ IWP' ทำนายผิดบน* Q *ถ้า* $R = \max_{t \in M} \|\mathbf{x}_t\|_2$ *และสำหรับตัวทำนายเชิงเส้น* $h \in H$ *ใดๆ ที่นิยามด้วยเวคเตอร์* $\mathbf{u} \in \Re^n$ *แล้วจะได้ว่า*

$$\begin{aligned} Loss^{w}(IWP', Q) \quad &\le \quad \frac{1}{\gamma} L_{\mathbf{u},\gamma}(M) - \frac{\langle \mathbf{u}, \mathbf{w}_1 \rangle}{\gamma} + \frac{C}{2\gamma^2} \\ &+ \frac{1}{2\gamma^2} \sqrt{4\gamma C L_{\mathbf{u},\gamma}(M) + C^2 - 4\gamma C \langle \mathbf{u}, \mathbf{w}_1 \rangle + 4\gamma^2 \|\mathbf{u}\|_2^2 \|\mathbf{w}_1\|_2^2} \end{aligned}$$

เมื่อ $L_{\mathbf{u},\gamma}(M) = \sum_{t \in M} D_\gamma(\mathbf{u}; (\mathbf{x}_t, y_t, c_t))$ *และ* $C = R^2 \|\mathbf{u}\|_2^2$

บทพิสูจน์ พิจารณาพจน์ในเครื่องหมายรากที่สองเราสามารถเขียนใหม่ได้เป็น

$$4\gamma C L_{\mathbf{u},\gamma}(M) + \left(C - 2\gamma \langle \mathbf{u}, \mathbf{w}_1 \rangle\right)^2 + 4\gamma^2 \left(\|\mathbf{u}\|_2^2 \|\mathbf{w}_1\|_2^2 - \left(\langle \mathbf{u}, \mathbf{w}_1 \rangle\right)^2 \right)$$

เนื่องจาก $\|\mathbf{u}\|_2 \|\mathbf{w}_1\|_2 \ge \langle \mathbf{u}, \mathbf{w}_1 \rangle$ จึงทำให้ค่าข้างต้นมีค่าอย่างต่ำเท่ากับศูนย์ สังเกตว่า ถ้า $Loss^{w}(IWP', Q)$ มีค่าไม่เกิน $\frac{1}{\gamma} L_{\mathbf{u},\gamma}(M) - \frac{c_{\max} \langle \mathbf{u}, \mathbf{w}_1 \rangle}{\gamma}$ แล้ว ทฤษฎีบทจะเป็นจริง เพราะฉะนั้นในการพิสูจน์ต่อจากนี้เราจะสมมติให้ $Loss^{w}(IWP', Q) > \frac{1}{\gamma} L_{\mathbf{u},\gamma}(M) - \frac{c_{\max} \langle \mathbf{u}, \mathbf{w}_1 \rangle}{\gamma}$

สำหรับการทำนายรอบที่ $t \in M$ ใดๆ เวคเตอร์น้ำหนัก $\mathbf{w}_t$ จะถูกปรับปรุง เมื่อใช้ทฤษฎีบทย่อยที่ 3 ในรอบนั้นๆ โดยให้ $a = c_t y_t$ และแทน ด้วย $\mathbf{u}$ ด้วย $\lambda \mathbf{u}$ เมื่อ $\lambda > 0$ จะได้ว่า

$$c_t y_t \left(\langle \lambda \mathbf{u}, \mathbf{x}_t \rangle - \langle \mathbf{w}_t, \mathbf{x}_t \rangle \right) \quad = \quad d(\lambda \mathbf{u}, \mathbf{w}_t) - d(\lambda \mathbf{u}, \mathbf{w}_{t+1}) + d(\mathbf{w}_t, \mathbf{w}_{t+1}) \qquad (9)$$

สังเกตว่า ค่าของสมการข้างต้นแปรผันตรงกับค่าความแตกต่างของคำทำนายของ h และ *IWP'* กล่าวคือ หากค่าดังกล่าวมีค่าน้อย แสดงว่า h ทำนายไม่ต่างจาก *IWP'* มากนักในรอบที่ *IWP'*

ทำนายผิด ดังนั้น เราต้องการให้ค่าดังกล่าวมีค่าน้อยๆ, จากสมการ (9) เนื่องจาก $y_t \cdot \langle \mathbf{w}_t, \mathbf{x}_t \rangle \le 0$ ดังนั้นจะได้ว่า

$$\lambda c_t y_t \langle \mathbf{u}, \mathbf{x}_t \rangle \quad \le \quad d(\lambda \mathbf{u}, \mathbf{w}_t) - d(\lambda \mathbf{u}, \mathbf{w}_{t+1}) + d(\mathbf{w}_t, \mathbf{w}_{t+1})$$

เมื่อทำการหาผลรวมของอสมการข้างต้น สำหรับทุกๆ $t \in M$ จะได้ว่า

$$\lambda \hat{\gamma}_{\mathbf{u},M} Loss^w(IWP', Q) \quad \le \quad d(\lambda \mathbf{u}, \mathbf{w}_1) - d(\lambda \mathbf{u}, \mathbf{w}_{m+1}) + \sum_{t \in M} d(\mathbf{w}_t, \mathbf{w}_{t+1}) \qquad (10)$$

เมื่อ $\hat{\gamma}_{\mathbf{u},M} = \frac{1}{Loss^w(IWP', Q)} \sum_{t \in M} c_t y_t \langle u, \mathbf{x}_t \rangle$ และเมื่อใช้ทฤษฎีบทย่อยที่ 4 กับพจน์สุดท้ายของอสมการ (10) จะได้ว่า

$$\sum_{t \in M} d(\mathbf{w}_t, \mathbf{w}_{t+1}) \quad \le \quad \sum_{t \in M} \frac{c_t^2}{2} \|\mathbf{x}_t\|_2^2$$

อย่างไรก็ตาม เนื่องจาก $c_t \le 1$ ดังนั้น $c_t^2 \le c_t$ เพราะฉะนั้นจะสามารถกำหนดขีดจำกัดของพจน์สุดท้ายของอสมการ (10) ได้ดังนี้

$$\begin{aligned} \sum_{t \in M} \frac{c_t^2}{2} \|\mathbf{x}_t\|_2^2 \le & \quad \sum_{t \in M} \frac{c_t}{2} \|\mathbf{x}_t\|_2^2 \\ = & \quad \frac{R^2}{2} \sum_{t \in M} c_t \\ = & \quad \frac{R^2}{2} Loss^w(IWP', Q) \end{aligned} \qquad (11)$$

จากอสมการ (10) เนื่องจาก $d(\lambda \mathbf{u}, \mathbf{w}_{t+1}) \ge 0$ จากนิยามการลู่เข้าของเบรกแมน ดังนั้นเมื่อแทน (8) และ (11) ในอสมการ (10) จะได้ว่า

$$\lambda \hat{\gamma}_{\mathbf{u},M} Loss^w(IWP', Q) \quad \le \quad \frac{\lambda^2 \|\mathbf{u}\|_2^2 + \|\mathbf{w}_1\|_2^2}{2} - \lambda \langle \mathbf{u}, \mathbf{w}_1 \rangle + \frac{R^2}{2} Loss^w(IWP', Q) \qquad (12)$$

พิจารณาอสมการ $\gamma - \frac{L_{\mathbf{u},\gamma}(M)}{Loss^{w}(IWP',Q)} \le \hat{\gamma}_{\mathbf{u},M}$ ซึ่งเป็นจริง จากนิยามของความเบี่ยงเบนแบบมีน้ำหนักความสำคัญ $D_{\gamma}(\mathbf{u};(\mathbf{x}_t,y_t,c_t))$ และ $\hat{\gamma}_{\mathbf{u},M}$ จะเห็นว่า หากค่าทางซ้ายของอสมการมีค่ามาก แสดงว่า ตัวทำนายเชิงเส้น h ทำนายได้ดีบน Q เมื่อพิจารณาเฉพาะครั้งที่ *IWP'* ทำนายผิด หรือกล่าวได้ว่า เราเปรียบเทียบประสิทธิภาพการทำนายของ *IWP*' กับตัวทำนายที่ดี, เมื่อนำไปรวมกับอสมการ (12) จะได้

$$\frac{\lambda^2\|\mathbf{u}\|_2^2+\|\mathbf{w}_1\|_2^2}{2}-\lambda\langle\mathbf{u},\mathbf{w}_1\rangle+\frac{R^2}{2}Loss^{w}(IWP',Q)-\lambda\gamma Loss^{w}(IWP',Q)+\lambda L_{\mathbf{u},\gamma}(M)\ge 0 \qquad (13)$$

พิจารณาค่าอสมการ (13) มีค่าน้อยที่สุดเมื่อ

$$\lambda \;=\; \frac{\langle\mathbf{u},\mathbf{w}_1\rangle - L_{\mathbf{u},\gamma}(M)+\gamma Loss^{w}(IWP',Q)}{\|\mathbf{u}\|_2^2}$$

ซึ่งมีค่าเป็นบวก ตามที่ข้อสมมติตอนต้นที่ว่า $Loss^{w}(IWP',Q) > \frac{1}{\gamma}L_{\mathbf{u},\gamma}(M) - \frac{\langle\mathbf{u},\mathbf{w}_1\rangle}{\gamma}$ นั่นคือ อสมการ (13) มีค่าน้อยที่สุดเป็น

$$\|\mathbf{w}_1\|_2^2 + R^2 Loss^{w}(IWP',Q) - \frac{\left(\langle\mathbf{u},\mathbf{w}_1\rangle - L_{\mathbf{u},\gamma}(M)+\gamma Loss^{w}(IWP',Q)\right)^2}{\|\mathbf{u}\|_2^2} \qquad (14)$$

เราจะแทน $Loss^{w}(IWP', Q)$ ด้วย L_w ใน (14) และเรียกว่า (14') ซึ่งเป็นฟังก์ชันกำลังสองของ L_w ที่มีค่าสัมประสิทธิ์ของพจน์นำเป็นลบ และที่ $L_w = 0$ ค่าของ (14') มีค่าเป็นบวก เนื่องจาก $\|\mathbf{u}\|_2\|\mathbf{w}_1\|_2 \ge \langle\mathbf{u},\mathbf{w}_1\rangle$ ดังนั้นค่าของ $Loss^{w}(IWP', Q)$ จะมีขีดจำกัดบนเป็นค่าขอบเขตบนที่น้อยที่สุดของ $L_w > 0$ ที่ทำให้ (14') เป็นบวก นั่นคือค่ารากที่สองที่มากที่สุดของ (14') ซึ่งมีค่าเป็น

$$\begin{aligned} Loss^{w}(IWP',Q) \;\le\; & \frac{1}{\gamma}L_{\mathbf{u},\gamma}(M)-\frac{\langle\mathbf{u},\mathbf{w}_1\rangle}{\gamma}+\frac{C}{2\gamma^2} \\ & +\frac{1}{2\gamma^2}\sqrt{4\gamma C L_{\mathbf{u},\gamma}(M)+C^2-4\gamma C\langle\mathbf{u},\mathbf{w}_1\rangle+4\gamma^2\|\mathbf{u}\|_2^2\|\mathbf{w}_1\|_2^2} \end{aligned}$$

■

ทฤษฎีบทที่ 8 มีการกำหนดว่า $\mathbf{w}_1 \in \Re^n$ และ $c_t \leq 1$ ส่วนกรณีอัลกอริทึม *IWP* ที่มี $\mathbf{w}_1$=0 และ $c_t \geq 0$ สามารถเขียนอสมการข้างต้นใหม่ได้เป็น

$$\begin{aligned}
\frac{Loss^{w}(IWP,Q)}{c_{\max}} &\leq \frac{L_{\mathbf{u},\gamma}(M)}{\gamma c_{\max}} + \frac{C}{2\gamma^2} + \frac{1}{2\gamma^2}\sqrt{4\gamma C\frac{L_{\mathbf{u},\gamma}(M)}{c_{\max}} + C^2} \\
&\leq \frac{L_{\mathbf{u},\gamma}(M)}{\gamma c_{\max}} + \frac{C}{2\gamma^2} + \frac{1}{2\gamma^2}\sqrt{\left(\sqrt{4\gamma C\frac{L_{\mathbf{u},\gamma}(M)}{c_{\max}}} + C\right)^2} \\
&= \frac{L_{\mathbf{u},\gamma}(M)}{\gamma c_{\max}} + \frac{C}{\gamma^2} + \frac{1}{\gamma^2}\sqrt{\gamma C\frac{L_{\mathbf{u},\gamma}(M)}{c_{\max}}}
\end{aligned}$$

ทฤษฎีบทที่ 7 เป็นจริง เมื่อ $\gamma = 1$

ผลการวิจัยการแก้ปัญหาการเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญ

สำหรับผลการวิจัยการแก้ปัญหาการเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญแบ่งออกเป็น 2 ส่วน ได้แก่ ผลการวิจัยภาคทดลองและผลการวิจัยภาคทฤษฎี

ผลการวิจัยภาคทดลอง

งานวิจัยนี้ได้การทดลองนี้ได้ทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพการทำนายแบบหลายประเภทแบบมีน้ำหนักความสำคัญของ 3 อัลกอริทึม ได้แก่ อัลกอริทึมแบบลดรูปที่ใช้เพอร์เซพตรอนเป็นกล่องดำ, อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญ (*IWP*) และอัลกอริทึมเพอร์เซพตรอน เมื่อทำการลดรูปจากปัญหาการเรียนรู้แบบหลายประเภทไปเป็นปัญหาเรียนรู้สองประเภทด้วยวิธีการหลายเวคเตอร์และวิธีการทุกคู่ โดยทดลองกับชุดข้อมูลมาตรฐาน 4 ชุดข้อมูลใน Asuncion and Newman (2007) ดังนี้

ชุดข้อมูล Letter Recognition

ชุดข้อมูล Letter Recognition ประกอบด้วยข้อมูล 26 ประเภท และมีตัวอย่างทั้งสิ้น 20000 รายการ แต่ละรายการมี 16 คุณลักษณะ ซึ่งเป็นข้อมูลของภาพตัวอักษรพิมพ์ใหญ่ภาษาอังกฤษ แต่ละตัวอักษรจะแบ่งเป็น 20 รูปแบบตัวพิมพ์ (font) ในชุดข้อมูลนี้แต่ละตัวอย่างไม่มีน้ำหนัก

ความสำคัญกำกับ เราจึงทำการกำหนดค่าน้ำหนักความสำคัญของตัวอย่างใดๆ โดยใช้ค่าสุ่มจากศูนย์ถึงค่าความถี่ของตัวอักษรภาษาอังกฤษของตัวอย่างนั้นๆ เป็นน้ำหนักความสำคัญของตัวอย่างดังกล่าว

ชุดข้อมูล Isolet

ชุดข้อมูล Isolet ประกอบด้วยข้อมูล 26 ประเภท และมีตัวอย่างทั้งสิ้น 6238 รายการ แต่ละรายการมี 617 คุณลักษณะ ซึ่งเป็นข้อมูลของเสียงของการอ่านตัวอักษรภาษาอังกฤษ โดยผู้อ่าน 50 คน ในชุดข้อมูลนี้แต่ละตัวอย่างไม่มีน้ำหนักความสำคัญกำกับ เราจึงทำการกำหนดค่าน้ำหนักความสำคัญของตัวอย่าง เช่นเดียวกับชุดข้อมูล Letter Recognition นั่นคือ ค่าน้ำหนักความสำคัญของตัวอย่างใดๆ มีค่าเป็นค่าสุ่มจากศูนย์ถึงค่าความถี่ของตัวอักษรภาษาอังกฤษของตัวอย่างนั้นๆ

ชุดข้อมูล Car Evaluation

ชุดข้อมูล Car Evaluation ประกอบด้วยข้อมูล 4 ประเภท และมีตัวอย่างทั้งสิ้น 1728 รายการ แต่ละรายการมี 8 คุณลักษณะ ซึ่งเป็นข้อมูลการประเมินคุณภาพของรถยนต์ ในชุดข้อมูลนี้แต่ละตัวอย่างไม่มีน้ำหนักความสำคัญกำกับ เราจึงทำการกำหนดค่าน้ำหนักความสำคัญของตัวอย่างใดๆ เป็นค่าสุ่มจาก 0 ถึง $\frac{1}{P(r)}$ โดยที่ $P(r)$ เป็นค่าความน่าจะเป็นที่จะสุ่มได้ตัวอย่างในประเภท $r \in Y$

ในชุดข้อมูลนี้มีบางคุณลักษณะที่มีชนิดข้อมูลเป็นข้อความ เราได้ทำการแปลงคุณลักษณะเหล่านั้นเป็นกลุ่มของคุณลักษณะของ 0 หรือ 1ในแนวทางเดียวกับ Rifkin and Klautau (2004)

ชุดข้อมูล Abalone

ชุดข้อมูล Abalone ประกอบด้วยข้อมูล 28 ประเภท และมีตัวอย่างทั้งสิ้น 4177 รายการ แต่ละรายการมี 6 คุณลักษณะ ซึ่งเป็นข้อมูลที่ใช้ในการทำนายอายุของหอยทากจากข้อมูลทางกายภาพ ในชุดข้อมูลนี้แต่ละตัวอย่างไม่มีน้ำหนักความสำคัญกำกับ เราจึงทำการกำหนดค่าน้ำหนัก

ความสำคัญของตัวอย่างใดๆ เป็นค่าสุ่มจาก 0 ถึง $\frac{1}{P(r)}$ โดยที่ $P(r)$ เป็นค่าความน่าจะเป็นที่จะสุ่มได้ตัวอย่างในประเภท $r \in Y$

ในชุดข้อมูลนี้มีบางคุณลักษณะที่มีชนิดข้อมูลเป็นข้อความ เราได้ทำการแปลงคุณลักษณะเหล่านั้นเป็นกลุ่มของคุณลักษณะของ 0 หรือ 1ในแนวทางเดียวกับ Rifkin and Klautau (2004)

ผลการทดลอง

เราได้ทำการเปรียบเทียบประสิทธิภาพการทำนายแต่ละอัลกอริทึมโดยใช้วิธีการเดียวกับในส่วนของการทดลองกับข้อมูลแบบสองประเภท นั่นคือทำการนับจำนวนครั้งและคิดค่าใช้จ่ายรวมที่แต่ละวิธีการทำนายผิด สำหรับการทดลองกับชุดข้อมูลแบบหลายประเภทโดยการลดรูปด้วยวิธีการหลายเวคเตอร์ได้ผลตามที่แสดงในตารางที่ 2 เมื่อพิจารณาร้อยละของจำนวนที่แต่ละวิธีการทำนายผิด (%ตอบผิด) จะเห็นว่าโดยส่วนใหญ่เพอร์เซพตรอนมีจำนวนครั้งในการทำนายผิดใกล้เคียงกับวิธีการแบบลดรูปซึ่งน้อยกว่าวิธีการ *IWP* แต่เมื่อพิจารณาร้อยละของค่าใช้จ่ายรวมที่แต่ละวิธีการทำนายผิด (%ค่าใช้จ่าย) จะเห็นว่าอัลกอริทึม *IWP* กลับมีค่าใช้จ่ายรวมในการทำนายผิดน้อยกว่าอีกสองวิธีการที่ให้ผลการทดลองใกล้เคียงกัน ยกเว้นชุดข้อมูล Abalone ซึ่งทุกวิธีการผลการทดลองใกล้เคียงกัน นั่นคือทำนายผิดในเกือบทุกตัวอย่าง อย่างไรก็ตามในงานวิจัยของ Rifkin and Klautau (2004) ได้ชี้ให้เห็นว่าชุดข้อมูล Abalone เป็นชุดข้อมูลที่ยากต่อการทำนายให้ถูก เนื่องจากอัลกอริทึมการเรียนรู้แบบออฟไลน์ที่สามารถทำนายได้ดีบนตัวอย่างที่ไม่มีน้ำหนักความสำคัญยังทำนายผิดบนชุดข้อมูลดังกล่าวเกินร้อยละ 70

ส่วนการทดลองโดยการลดรูปด้วยวิธีการทุกคู่ได้ผลตามที่แสดงในตารางที่ 3 เมื่อเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมที่ลดรูปด้วยวิธีการหลายเวคเตอร์ (ตารางที่ 2) และวีธีการทุกคู่จาก (ตารางที่ 3) จะเห็นว่าการลดรูปด้วยวิธีการหลายเวคเตอร์และวีธีการทุกคู่นั้นมีผลลัพธ์เท่ากัน ซึ่งเป็นประเด็นที่น่าสนใจว่าเหตุใดการลดรูป 2 วิธีการจึงให้ผลลัพธ์เท่ากัน โดยเราจะทำการวิเคราะห์ประเด็นนี้ในส่วนของผลการวิจัยภาคทฤษฎี

ตารางที่ 2 ผลการเปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึมแบบลดรูป, อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญและอัลกอริทึมเพอร์เซพตรอน เมื่อทำการลดรูปจากปัญหาการทำนายหลายประเภทเป็นปัญหาการทำนายหลายประเภทด้วยวิธีการหลายเวคเตอร์ กับชุดข้อมูลแบบหลายประเภท

ชุดข้อมูล	*IWP*		ลดรูป		เพอร์เซพตรอน	
	%ตอบผิด	%ค่าใช้จ่าย	%ตอบผิด	%ค่าใช้จ่าย	%ตอบผิด	%ค่าใช้จ่าย
Letter	56.59	58.52	64.94	53.23	**55.74**	57.34
Isolet	**25.76**	26.56	43.52	**25.97**	25.92	26.67
Car	**20.27**	43.21	33.56	**39.84**	20.33	43.77
Abalone	85.10	93.60	92.99	94.41	**84.61**	**93.47**

ตารางที่ 3 ผลการเปรียบเทียบประสิทธิภาพการทำนายของอัลกอริทึมแบบลดรูป, อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักความสำคัญและอัลกอริทึมเพอร์เซพตรอน เมื่อทำการลดรูปจากปัญหาการทำนายหลายประเภทเป็นปัญหาการทำนายหลายประเภทด้วยวิธีการทุกคู่กับชุดข้อมูลแบบหลายประเภท

ชุดข้อมูล	*IWP*		ลดรูป		เพอร์เซพตรอน	
	%ตอบผิด	%ค่าใช้จ่าย	%ตอบผิด	%ค่าใช้จ่าย	%ตอบผิด	%ค่าใช้จ่าย
Letter	56.59	58.52	64.94	53.23	**55.74**	57.34
Isolet	**25.76**	26.56	43.52	**25.97**	25.92	26.67
Car	**20.27**	43.21	33.56	**39.84**	20.33	43.77
Abalone	85.10	93.60	92.99	94.41	**84.61**	**93.47**

ผลการวิจัยภาคทฤษฎี

ในส่วนของผลการวิจัยภาคทฤษฎีของการแก้ปัญหาการเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญ เราได้ทำการวิเคราะห์ประเด็นที่กล่าวถึงไว้ในส่วนของผลการวิจัยภาคทฤษฎีที่ว่า เหตุใดการลดรูปด้วยวิธีการหลายเวคเตอร์จึงให้ผลลัพธ์เท่ากับวิธีการทุกคู่ ซึ่งเราจะทำการพิสูจน์ว่าการ

ลดรูป 2 วิธีการดังกล่าวนั้นสมมูลกัน และจากการสมมูลกันนี้ทำให้เราสามารถวิเคราะห์ค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของวิธีการหลายเวคเตอร์ที่ใช้ *IWP* เป็นฐาน และวิธีการทุกคู่ที่ใช้ *IWP* เป็นฐานที่ได้นำเสนอไปในส่วนของวิธีการได้ในคราวเดียว

การสมมูลกันระหว่างวิธีการหลายเวคเตอร์และวิธีการทุกคู่

ในส่วนนี้จะทำการพิสูจน์ว่าการลดรูปจากปัญหาการเรียนรู้หลายประเภทไปเป็นปัญหาการเรียนรู้สองประเภทด้วยวิธีการหลายเวคเตอร์และวิธีการทุกคู่นั้นสมมูลกัน ในการพิสูจน์เพื่อให้ง่ายขึ้นเราจะพิสูจน์กรณีที่ตัวอย่างไม่มีน้ำหนักความสำคัญ นั่นคือ $c_t = 1$ สำหรับทุกๆ t ส่วนในกรณีที่ตัวอย่างมีน้ำหนักความสำคัญ สามารถพิสูจน์ได้ด้วยวิธีการใกล้เคียงกัน

สำหรับปัญหาการเรียนรู้ k ประเภท ให้ $Y=\{1,2,\ldots,k\}$ สำหรับการทำนายรอบที่ t ใดๆ ให้ $h_t^M(\mathbf{x},r)$ แทนคะแนนของประเภท r สำหรับวิธีการหลายเวคเตอร์ และ $h_t^A(\mathbf{x},r)$ แทนคะแนนของประเภท r สำหรับวิธีการทุกคู่ สังเกตว่า สำหรับ $\mathbf{x} \in \Re^n$ ใดๆ และสำหรับประเภท r ใดๆ ถ้า $h_t^A(\mathbf{x},r) = k \cdot h_t^M(\mathbf{x},r)$ แล้วทั้งสองวิธีการจะทำนายเหมือนกันในทุกๆ ตัวอย่าง $\mathbf{x}$ เนื่องจาก ถ้า $h_t^M(\mathbf{x},i) > h_t^M(\mathbf{x},j)$ แล้ว $h_t^A(\mathbf{x},i) > h_t^A(\mathbf{x},j)$ ต่อไปจะเป็นการพิสูจน์ทฤษฎีบทย่อยที่สำคัญสำหรับการพิสูจน์ความสมมูลของทั้งสองวิธีการ

ทฤษฎีบทย่อยที่ 5 *ถ้าการทำนายรอบที่ t,* $h_t^A(\mathbf{x},r) = k \cdot h_t^M(\mathbf{x},r)$ *สำหรับ* $\mathbf{x} \in \Re^n$ *ใดๆ และสำหรับประเภท r ใดๆ แล้ว*

$$h_{t+1}^A(\mathbf{x},r) = k \cdot h_{t+1}^M(\mathbf{x},r)$$

สำหรับ $\mathbf{x} \in \Re^n$ *ใดๆ และสำหรับ r ใดๆ*

บทพิสูจน์ กรณีที่การทำนายรอบที่ t ถูกต้องเวคเตอร์น้ำหนักจะไม่ถูกปรับปรุง ดังนั้นทฤษฎีบทตั้งเป็นจริง ส่วนในกรณีที่ทำนายผิดในรอบที่ t สำหรับตัวอย่าง $\mathbf{x}_t$ เนื่องจากทั้งสองวิธีการมีการทำนายเหมือนกัน นั่นคือทำนายประเภทที่ให้ค่าคะแนนมากที่สุด ดังนั้นทั้งสองวิธีการจะต้องปรับปรุงเวคเตอร์น้ำหนักดังนี้

ให้ Y เป็นเซตของประเภททั้งหมดที่เป็นไปได้ และสำหรับการทำนายรอบที่ t ให้ $\hat{y}_t \in Y$ เป็นประเภทที่ถูกทำนาย และ $y_t \in Y$ เป็นประเภทที่ถูกต้อง พิจารณาการปรับปรุงเวคเตอร์น้ำหนักของวิธีการหลายเวคเตอร์ซึ่งจะปรับปรุงเวคเตอร์น้ำหนักดังนี้

$$\mathbf{w}_{t+1}^{y_t} \leftarrow \mathbf{w}_t^{y_t} + \mathbf{x}_t, \quad \mathbf{w}_{t+1}^{\hat{y}_t} \leftarrow \mathbf{w}_t^{\hat{y}_t} - \mathbf{x}_t$$

ส่วน $\mathbf{w}_{t+1}^{r}$ สำหรับ $r \in Y \setminus \{y_t, \hat{y}_t\}$ จะไม่ถูกปรับปรุง ดังนั้นสำหรับตัวอย่าง $\mathbf{x}_t$ ใดๆ

$$\begin{aligned} h_{t+1}^{M}(\mathbf{x}, y_t) &= \left\langle \mathbf{w}_t^{y_t} + \mathbf{x}_t, \mathbf{x} \right\rangle \\ &= \left\langle \mathbf{w}_t^{y_t}, \mathbf{x} \right\rangle + \left\langle \mathbf{x}_t, \mathbf{x} \right\rangle \\ &= h_t^{M}(\mathbf{x}, y_t) + \left\langle \mathbf{x}_t, \mathbf{x} \right\rangle \end{aligned} \tag{15}$$

ในทำนองเดียวกัน

$$h_{t+1}^{M}(\mathbf{x}, \hat{y}_t) = h_t^{M}(\mathbf{x}, \hat{y}_t) - \left\langle \mathbf{x}_t, \mathbf{x} \right\rangle \tag{16}$$

ในส่วนของการปรับปรุงเวคเตอร์น้ำหนักของวิธีการทุกคู่ พิจารณากรณี $r \in Y \setminus \{y_t, \hat{y}_t\}$

$$\begin{aligned} h_{t+1}^{A}(\mathbf{x}, y_t) &= \sum_{j \in Y \setminus \{r\}} \left\langle \mathbf{w}_{t+1}^{rj}, \mathbf{x} \right\rangle \\ &= \left\langle \sum_{j \in Y \setminus \{r\}} \mathbf{w}_{t+1}^{rj}, \mathbf{x} \right\rangle \end{aligned}$$

เวคเตอร์น้ำหนักที่ถูกปรับปรุงจะมีเพียง $\mathbf{w}_{t+1}^{ry_t}$ และ $\mathbf{w}_{t+1}^{r\hat{y}_t}$ นั่นคือ

$$\mathbf{w}_{t+1}^{ry_t} \leftarrow \mathbf{w}_t^{ry_t} - \mathbf{x}_t, \quad \mathbf{w}_{t+1}^{r\hat{y}_t} \leftarrow \mathbf{w}_t^{r\hat{y}_t} + \mathbf{x}_t$$

ซึ่งจะทำให้ผลรวมของเวคเตอร์น้ำหนักไม่เปลี่ยนแปลง นั่นคือ

$$\sum_{j \in Y \setminus \{r\}} \mathbf{w}_{t+1}^{rj} = \sum_{j \in Y \setminus \{r\}} \mathbf{w}_t^{rj}$$

ดังนั้นจะได้ว่า สำหรับ $\mathbf{x}$ ใดๆ และ สำหรับ $r \in Y \setminus \{y_t, \hat{y}_t\}$ ใดๆ

$$h^A_{t+1}(\mathbf{x}, r) \quad = \quad h^A_t(\mathbf{x}, r) \tag{17}$$

พิจารณากรณีประเภท y_t ซึ่งมี $h^A_{t+1}(\mathbf{x}, y_t) = \sum_{j \in Y \setminus \{y_t\}} \left\langle \mathbf{w}^{y_t j}_{t+1}, \mathbf{x} \right\rangle$ ดังนั้น สำหรับ $r \in Y \setminus \{y_t, \hat{y}_t\}$ จะทำการปรับปรุงเวคเตอร์น้ำหนักดังนี้

$$\mathbf{w}^{y_t r}_{t+1} \leftarrow \mathbf{w}^{y_t r}_t + \mathbf{x}_t$$

และสำหรับประเภท $\hat{y}_t$ เวคเตอร์น้ำหนัก $\mathbf{w}^{y_t \hat{y}_t}_t$ จะถูกปรับปรุงสองครั้ง นั่นคือ

$$\mathbf{w}^{y_t \hat{y}_t}_{t+1} \leftarrow \mathbf{w}^{y_t r}_t + 2 \cdot \mathbf{x}_t$$

ผลรวมของเวคเตอร์น้ำหนักสำหรับประเภท y_t ของการทำนายรอบที่ t+1 มีค่าเป็น

$$\sum_{j \in Y \setminus \{y_t\}} \mathbf{w}^{y_t j}_{t+1} \quad = \quad \sum_{j \in Y \setminus \{y_t\}} \left(\mathbf{w}^{y_t j}_t + k \cdot \mathbf{x}_t \right)$$

ดังนั้นจะได้ว่า

$$\begin{aligned} h^A_{t+1}(\mathbf{x}, y_t) &= \sum_{j \in Y \setminus \{y_t\}} \left\langle \mathbf{w}^{y_t j}_{t+1}, \mathbf{x} \right\rangle \\ &= \left\langle \sum_{j \in Y \setminus \{y_t\}} \mathbf{w}^{y_t j}_t + k \cdot \mathbf{x}_t, \mathbf{x} \right\rangle \\ &= \left\langle \sum_{j \in Y \setminus \{y_t\}} \mathbf{w}^{y_t j}_t, \mathbf{x} \right\rangle + k \cdot \left\langle \mathbf{x}_t, \mathbf{x} \right\rangle \\ &= h^A_t(\mathbf{x}, y_t) + k \cdot \left\langle \mathbf{x}_t, \mathbf{x} \right\rangle \end{aligned} \tag{18}$$

ในทำนองเดียวกัน

$$h^A_{t+1}(\mathbf{x}, \hat{y}_t) \quad = \quad h^A_t(\mathbf{x}, \hat{y}_t) - k \cdot \left\langle \mathbf{x}_t, \mathbf{x} \right\rangle \tag{19}$$

ทฤษฎีบทตั้งเป็นจริงตามสมการ (15) - (19) ■

จากทฤษฎีบทย่อยที่ 5 สังเกตว่า ถ้า $h_1^A(\mathbf{x},r) = k \cdot h_1^M(\mathbf{x},r)$ แล้ววิธีการหลายเวคเตอร์และวิธีการทุกคู่จะทำนายเหมือนกันในทุกๆ รอบ ดังนั้นจะได้ว่า ทั้งสองวิธีการสมมูลกัน อย่างไรก็ตาม เมื่อพิจารณาแง่ของเวลาในการทำงาน จะเห็นว่าวิธีการหลายเวคเตอร์ทำงานที่ใช้เวลาในการทำงานเป็น $O(k)$ นั้นทำงานได้เร็วกว่าวิธีการทุกคู่ที่ใช้เวลาในการทำงานเป็น $O(k^2)$

การรับประกันค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของวิธีการหลายเวคเตอร์และวิธีการทุกคู่ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐาน

สำหรับการพิสูจน์ค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของวิธีการหลายเวคเตอร์ที่ใช้ *IWP* เป็นฐานนั้น สามารถใช้แนวทางการพิสูจน์เดียวกับการพิสูจน์ทฤษฎีบทที่ว่าด้วยค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของวิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนเป็นฐานในงานวิจัยของ Fink *et al*. (2006) ซึ่งมีการนิยามฟังก์ชันความสูญเสียแบบฮิงก์ดังนี้

ให้ $Y = \{1,\ldots,k\}$ เป็นเซตของประเภทของตัวอย่างและสมมติฐาน h_t ใดๆ ซึ่งมีนิยามตามสมการ (3) สำหรับตัวอย่าง $(\mathbf{x}_t,y_t,c_t)$ ใดๆ ฟังก์ชันความสูญเสียแบบฮิงก์สำหรับสมมติฐาน h_t มีค่าเป็น

$$\ell(h_t,\mathbf{x}_t) = \left(1 - h_t(\mathbf{x}_t,y_t) + \max_{r \in Y \setminus \{y_t\}} h_t(\mathbf{x}_t,r)\right)_+ \tag{20}$$

เมื่อ $y_t \in Y$ และ $(\cdot)_+ = \max\{0,\cdot\}$ จากฟังก์ชันความสูญเสียแบบฮิงก์ในสมการ (20) และนิยามฟังก์ชันความสูญเสียแบบมีน้ำหนักในสมการ (4) เราสามารถนิยามฟังก์ชันความสูญเสียแบบฮิงก์แบบมีน้ำหนัก สำหรับการทำนายรอบที่ t ได้เป็น

$$\ell^w(h_t,\mathbf{x}_t) = c_t\left(1 - h_t(\mathbf{x}_t,y_t) + \max_{r \in Y \setminus \{y_t\}} h_t(\mathbf{x}_t,r)\right)_+ \tag{21}$$

เมื่อ $y_t \in Y$ และ $(\cdot)_+ = \max\{0,\cdot\}$

เราสามารถรับประกันค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของวิธีการหลายเวคเตอร์ที่ใช้ *IWP* เป็นฐานได้ โดยใช้ทฤษฎีบทที่ 7 เป็นบทตั้งต้นและใช้ความสูญเสียแบบมีน้ำหนักตามที่นิยามในสมการ (21) ได้ดังบทสืบเนื่องต่อไปนี้

บทสืบเนื่องที่ 1 *สำหรับลำดับข้อมูลป้อนเข้าแบบมีน้ำหนัก* $Q = ((\mathbf{x}_1, y_1, c_1), (\mathbf{x}_2, y_2, c_2), \ldots, (\mathbf{x}_m, y_m, c_m)) \in (\Re^n \times Y \times [0,\infty))^m$ *ใดๆ ถ้ามีตัวทำนายเชิงเส้น* $h \in H$ *ใดๆ ที่นิยามด้วยเวคเตอร์* $U = \{\mathbf{u}^r : r \in Y\}$ *โดยที่* $\ell^w(h,Q) \le L$ *แล้ว วิธีการหลายเวคเตอร์ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐาน จะมีค่าความสูญเสียแบบมีน้ำหนักไม่เกิน*

$$L + c_{\max} C + \sqrt{c_{\max} LC}$$

เมื่อ $R = 2 \cdot \max_{t \in \{1,\ldots,m\}} \|\mathbf{x}_t\|_2$, $C = R^2 \sum_{r \in Y} \|\mathbf{u}^r\|_2^2$ *และ* $c_{\max} = \max_{t \in \{1,\ldots,m\}} c_t$

เนื่องจากวิธีการหลายเวคเตอร์และวิธีการทุกคู่สมมูลกัน ดังนั้นเราสามารถใช้บทสืบเนื่องข้างต้นในการรับประกันค่าขีดจำกัดบนของความสูญเสียแบบมีน้ำหนักของวิธีการทุกคู่ที่ใช้เพอร์เซพตรอนแบบมีน้ำหนักความสำคัญเป็นฐานได้ด้วย

สรุป

งานวิจัยนี้เสนอวิธีการแก้ปัญหาการเรียนรู้แบบมีน้ำหนักความสำคัญทั้งแบบสองประเภทและแบบหลายประเภท สำหรับปัญหาการเรียนรู้สองประเภทแบบมีน้ำหนักความสำคัญนำเสนอ 2 อัลกอริทึม ได้แก่ อัลกอริทึมแบบลดรูป และ อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักควาสำคัญ ซึ่งอัลกอริทึมที่สองให้ผลทั้งเชิงทฤษฎีและเชิงทดลองดีกว่าอัลกอริทึมแรก ส่วนปัญหาการเรียนรู้หลายประเภทแบบมีน้ำหนักความสำคัญเสนอให้ใช้อัลกอริทึมเพอร์เซพตรอนแบบมีน้ำหนักควาสำคัญเป็นฐานและใช้การลดรูปจากปัญหาการเรียนรู้แบบหลายประเภทไปเป็นปัญหาการเรียนรู้แบบสองประเภทด้วย 2 วิธีการ ได้แก่ วิธีการหลายเวคเตอร์ และ วิธีการทุกคู่ และได้ทำการพิสูจน์ว่าการลดรูปสองวิธีดังกล่าวสมมูลกัน

อย่างไรก็ตาม งานวิจัยนี้ยังไม่ได้พิจารณากรณีการทำนายหลายประเภท ที่ความสูญเสียที่เกิดขึ้นจากการทำนายประเภทของข้อมูลผิดพลาด อาจไม่เท่ากันในทุกประเภท การเรียนรู้ในรูปแบบดังกล่าว เรียกว่าการเรียนรู้ที่มีความไวต่อค่าใช้จ่าย ผู้วิจัยหวังว่าแนวทางสร้างอัลกอริทึมและการพิสูจน์ที่นำเสนออาจนำไปใช้ในกรณีดังกล่าวได้ด้วย

เอกสารและสิ่งอ้างอิง

Asuncion, A. and D.J. Newman. 2007. **UCI Machine Learning Repository**. UCI Machine Learning Repository: Data Sets. Available Source: http://www.ics.uci.edu/~mlearn/MLRepository.html, February 11, 2008.

Abe, N., Zadrozny, B. and J. Langford. 2004. An Iterative Method for Multi-Class Cost-Sensitive Learning, pp. 3-11. *In* **Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM, New York, NY, USA.

Cammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. and Y. Singer. 2006 . Online Passive-Aggressive Algorithms. **The Journal of Machine Learning Research** (7): 551-585.

Cesa-Bianchi, N. and G. Lugosi. 2006. **Prediction, Learning, and Games**. Cambridge University Press, New York, NY, USA.

Domingos, P. 1999. MetaCost: A General Method for Making Classifiers Cost-Sensitive, pp. 155-164. *In* **Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM, New York, NY, USA.

Elkan, C. 2001. The Foundations of Cost-Sensitive Learning, vol. 2, pp. 973-973. *In* **Proceedings of the 7th International Joint Conference on Artificial Intelligence**, Seattle.

Fink, M., Shalev-Shwartz, S., Singer, Y. and S. Ullman. 2006. Online Multiclass Learning by Interclass Hypothesis Sharing, pp. 313-320. *In* **Proceedings of the 23rd International Conference on Machine Learning**. ACM, New York, NY, USA.

Gentile, C. 2003. The Robustness of the *p*-Norm Algorithms. **Machine Learning** 3 (10): 265-299.

Novikoff, A.B. 1962**.** On Convergence Proofs on Perceptrons, pp. 615-622. *In* **Proceedings of the Symposium on the Mathematical Theory of Automata**. Polytechnic Institute of Brooklyn.

Rifkin, R. and A. Klautau. 2004. In Defense of One-Vs-All Classification. **The Journal of Machine Learning Research** (5): 101–141.

Rosenblatt, F. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. **Psychological Review** 6 (65): 386-408.

X. Liu. 2006. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. **IEEE Transactions on Knowledge and Data Engineering** 1 (18): pp. 63-77.

Zadrozny, B. and C. Elkan. 2001. Learning and Making Decisions when Costs and Probabilities are both Unknown, pp. 204-213. *In* **Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM, San Francisco, California.

Zadrozny, B., Langford, J. and N. Abe. 2003. Cost-Sensitive Learning by Cost-Proportionate Example Weighting, pp. 435-442. *In* **Proceedings of the 3rd IEEE International Conference on Data Mining**. IEEE Computer Society, Washington, DC, USA.

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	อดิศักดิ์ สุภีสุน
วัน เดือน ปี ที่เกิด	1 มีนาคม พ.ศ. 2524
สถานที่เกิด	จังหวัดสระบุรี
ประวัติการศึกษา	วศ.บ. (วิศวกรรมคอมพิวเตอร์) มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา (พ.ศ. 2546)
ตำแหน่งหน้าที่การงานปัจจุบัน	-
สถานที่ทำงานปัจจุบัน	-
ผลงานดีเด่นและรางวัลทางวิชาการ	-
ทุนการศึกษาที่ได้รับ	-