รายงานการวิจัยฉบับสมบูรณ์

การลดขนาดของข้อมูลของการทำเหมืองข้อมูล

Dimensionality Reduction for Data Mining

ดร. กิติ์สุชาต พสุภา

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ ประจำปีงบประมาณ 2555

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อโครงการ (ภาษาไทย) การลดขนาดของข้อมูลของการทำเหมืองข้อมูล
แหล่งเงิน เงินรายได้
ประจำปีงบประมาณ 2555 จำนวนเงินที่ได้รับการสนับสนุน 50,000.00 บาท
ระยะเวลาทำการวิจัย 1 ปี ตั้งแต่ ตุลาคม 2555 ถึง กันยายน 2555
ชื่อ-สกุล หัวหน้าโครงการ
ดร. กิติ์สุชาต พสุภา คณะเทคโนโลยีสารสนเทศ สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

ในฟังก์ชันลงโทษ (Penalty Function) นั้นต้องการการเลือกเรกิวลาไรเซชันพารามิเตอร์ (Regularization Parameter) ที่ใช้ในการควบคุมระดับของความซับซ้อนของโมเดลจำแนกข้อมูลชนิดเคอเนล (Kernel Classifier) ซึ่งในโมเดลจำแนกข้อมูลชนิดเคอเนลนั้นมีพารามิเตอร์ที่ต้องเลือกอยู่แล้วหนึ่งตัว นั่นคือ เคอเนลพารามิเตอร์ (Kernel Parameter) นั่นหมายความว่า การเพิ่มฟังก์ชั่นลงโทษเข้าไปนั้นต้องการใช้พารามิเตอร์ถึงสองตัวในการออพติไมเซชัน (Optimization) ที่สามารถทำได้โดยใช้การตรวจสอบไขว้ (Cross Validation) โครงการนี้ได้นำเสนอ อัลกอริทึมการวิเคราะห์การจำแนกของฟิชเชอร์แบบเคอเนลที่ไม่ซับซ้อน (Parsimonious Kernel Fisher Discriminant Analysis) ที่ไม่ต้องการเรกิวลาไรเซชันพารามิเตอร์ ซึ่งสามารถทำได้โดยการใช้ประโยชน์จากการแจกแจงก่อนที่ไม่ให้ข้อมูลของเจฟฟรี (Jeffrey's Noninformative Hyperprior) ที่ไม่ต้องการพารามิเตอร์ใดๆในฟังก์ชัน และถูกนำมาใช้ผ่านการตีความลำดับชั้นของเบส์ของการแจกแจงลาพลาสเซียน ทำให้ไม่จำเป็นต้องใช้เรกิวลาไรเซชันพารามิเตอร์ อัลกอริทึมที่นำเสนอนี้ได้ถูกเปรียบเทียบกับอัลกอริทึมอื่นๆ บนข้อมูลหลายชนิด และยังเปรียบเทียบอัลกอริทึมกับอัลกอริทึมที่ดีที่สุดในการคัดกรองเสมือน (Virtual Screening) อีกด้วย อย่างไรก็ตามผลการทดลองแสดงให้เห็นว่า ประสิทธิภาพนั้นยังน้อยกว่า แต่สามารถเปรียบเทียบและอยู่ในระดับเดียวกันได้เป็นกรณีไป

คำสำคัญ : การเรียนรู้ของเครื่อง, เคมีสารสนเทศ, การคัดกรองเสมือน, ความซับซ้อนของโมเดล, การวิเคราะห์การจำแนกของฟิชเชอร์, เคอเนล

Research Title: Dimensionality Reduction for Data Mining

Researcher: Dr. Kitsuchart Pasupa

Faculty: Information Technology **Department:** Information Technology

ABSTRACT

The penalty function requires a choice of regularization parameter which controls the degree of parsimony in sparse kernel classifier. This involves an extra parameter apart from kernel parameter in the optimization which must be found via, e.g. cross-validation. This paper introduces a new parsimonious binary kernel Fisher discriminant analysis which does not require a regularization parameter. This can be done by using a Jeffrey's noninformative hyperprior. A Jeffrey's noninformative hyperprior is parameter-free and is adopted through a hierarchical-Bayes interpretation of the Laplacian prior distribution. This leads to a non-requirement of the regularization parameter. The proposed algorithm is compared with other machine learning methods on substantial benchmarks. Moreover, it is also compared with the leading machine learning in virtual screening application. It is found to be less accurate but it is still comparable in a number of cases.

Keywords : Machine learning, chemoinformatics, virtual screening, sparsity control, Fisher discriminant analysis, kernel

กิตติกรรมประกาศ

การวิจัยครั้งนี้ได้รับทุนสนับสนุนการวิจัยจากสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง จากแหล่งทุนเงินรายได้ของคณะเทคโนโลยีสารสนเทศ ประจำปีงบประมาณ พ.ศ. 2555 ที่ได้รับการจัดสรรทุนอุดหนุนการวิจัย

ดร. กิติ์สุชาต พสุภา

สารบัญ

สารบัญตาราง

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์การจำแนกของฟิชเชอร์ (Fisher Discriminant Analysis - FDA) พยายามที่จะฉายข้อมูล (Project) เพื่อแยกกลุ่มข้อมูลโดยใช้ประโยชน์จากสมการเชิงเส้น โดยพยายามที่จะเพิ่มระยะห่างระหว่างข้อมูลของกลุ่มสองกลุ่มออกจากกัน และลดระยะห่างระหว่างข้อมูลที่อยู่ในกลุ่มเดียวกัน คุณสมบัติของอัลกอริทึมนี้ได้ถูกพิสูจน์แล้วว่าเหมาะสมที่สุด (Optimal) [1] แต่จากข้อมูลในโลกแห่งความเป็นจริงนั้น มีนัยสำคัญว่า ข้อมูลมีการกระจายตัวที่ไม่เป็นแบบเกาส์เซียน (Non-Gaussian Distribution) และมีเมทริกของความแปรปรวนรวม (Covariance Matrices) ของแต่ละกลุ่มที่แตกต่างกัน จึงส่งผลให้ประสิทธิภาพของ FDA นั้นลดลงอย่างเห็นได้ชัด เนื่องมาจาก FDA ไม่สามารถหาทิศทางการฉายข้อมูลได้ เพราะว่า FDA นั้นมีสมมติฐานว่า การที่จะได้มาของโมเดลที่เหมาะสมที่สุดนั้น ข้อมูลที่ป้อนเข้ามาจะต้องมีเมทริกของความแปรปรวนรวมเท่ากัน และมีการกระจายตัวแบบเกาส์เซียน อย่างไรก็ตามประสิทธิภาพของ FDA ก็สามารถปรับปรุงและแก้ไขปัญหาดังกล่าวได้โดยการใช้การวิเคราะห์การจำแนกของฟิชเชอร์แบบเคอเนล (Kernel FDA - kFDA) [2] ซึ่งไม่นานมานี้ [3] ได้นำเสนออัลกอริทึมที่เรียกว่า การวิเคราะห์การจำแนกของฟิชเชอร์แบบเคอเนลที่ไม่ซับซ้อน (Parsimonious kFDA) ซึ่งอัลกอริทึมดังกล่าวได้ใช้ความสัมพันธ์ของการวิเคราะห์การจำแนกของฟิชเชอร์และวิธีกำลังสองน้อยที่สุด (Least Squares - LS) โดยที่ความซับซ้อนของโมเดลนั้นถูกควบคุมด้วยฟังก์ชันลงโทษ L_q (L_q - Penalty Function) โดยที่ q มีค่าระหว่าง 0 และ 1 เป็นที่รู้จักกันดีว่าฟังก์ชันลงโทษนั้นมีคุณสมบัติในการทำให้โมเดลซับซ้อนน้อยลงหรือเบาบางลง (Sparsity) อีกทั้งยังทำให้ฟังก์ชันนั้นเกิดปัญหาและการแก้ปัญหาที่ยากขึ้น ซึ่งสามารถแก้ปัญหานี้ได้โดยการใช้วิธีเมเจอไรซ์-มินนิไมซ์ (Majorize-Minimize Principle - MM) มาช่วยในการแก้ปัญหานี้ ซึ่งนำไปสู่อัลกอริทึมการทำซ้ำ (Iterative Algorithm) อย่างง่าย

อย่างไรก็ตาม การใช้ฟังก์ชันลงโทษดังกล่าวนั้นทำให้มีพารามิเตอร์ที่จะต้องเลือกใช้เพื่อขึ้นนั่นคือ ไฮเปอร์พารามิเตอร์ (Hyperparameter) หรือ เรกิวราไรซ์พารามิเตอร์ (Regularize Parameter) ที่ไว้สำหรับควบคุมความซับซ้อนของโมเดล กล่าวคือ ในการหาโมเดลที่ดีที่สุด นอกเหนือจากเคอเนลพารามิเตอร์ที่ต้องการหาด้วยการใช้การตรวจสอบไขว้ (Cross-validation) แล้วนั้นจะต้องหาเรกิวราไรเซชันพารามิเตอร์อีกด้วย ไม่นานมานี้ [4] ประยุกต์ใช้การแจกแจงก่อนที่ไม่ให้ข้อมูลของเจฟฟรี (Jeffrey's Noninformative Hyperprior) ผ่านการตีความลำดับชั้นของเบส์ (Hierarchical-Bayes) ของการแจกแจงลาพลาสเซียน (Laplacian) ทำให้ในการใช้ฟังก์ชันลงโทษไม่จำเป็นจ้องใช้เรกิวลาไรเซชันพารามิเตอร์เนื่องมาจากฟังก์ชันนี้ไม่ต้องการพารามิเตอร์ในฟังก์ชัน

แรงจูงใจในการทำวิจัยเรื่องนี้อยู่ในพื้นฐานของเคมีสารสนเทศ (Chemoinformatics) โดยเน้นในเรื่อง การคัดกรองเสมือน (Virtual Screening - VS) ความสามารถในการจัดเรียงลำดับโมเลกุลในฐานข้อมูลตามจุดประสงค์ต่างๆ (เช่น การค้นพบยาค่าแมลง การค้นพบยารักษาโรคต่าง) นั้นสำคัญเนื่องมาจากค่าใช้จ่ายและเวลาที่ใช้ในการสังเคราะห์สารและการทดสอบสารประกอบ VS นั้นพยายามที่จะทำสิ่งที่กล่าวมานี้ด้วยคอมพิวเตอร์ ซึ่งมีศักยภาพในการลดค่าใช้จ่ายหลายสิบล้านบาทเพื่อเพิ่มผลกำไร อีกทั้งช่วยลดเวลาในการส่งจำหน่ายสู่ท้องตลาด ในสมัยก่อนในขั้นตอนการคิดค้นยานั้น จะต้องสังเคราะห์สารประกอบเคมีในหลอดทดลอง ซึ่งใช้เวลาและค่าใช้จ่ายสูงมาก ดังนั้นบริษัทยาจำนวนมากเริ่มหันมาใช้ VS เพื่อการพิสูจน์และคัดเลือกสารประกอบเคมีที่มีศักยภาพตามที่ต้องการ เพื่อที่จะลดจำนวนของสารที่ต้องการที่จะสังเคราะห์ในหลอดแก้ว ซึ่งสิ่งเหล่านี้สามารถทำได้โดยคอมพิวเตอร์

การเรียนรู้ของเครื่อง (Machine Learning) นั้นเริ่มมีบทบาทใน VS มากขึ้น อย่างไรก็ตามปัญหาก็เกิดขึ้น เมื่อการใช้ลายนิ้วมือ (Fingerprint) เป็นตัวบ่งบอกลักษณะของสารประกอบโมเลกุลเพราะว่าลายนิ้วมือมีมิติที่สูงมาก (High-dimensional) ซึ่งจะส่งผลให้เกิดปัญหาจำนวนข้อมูลที่เล็กเกินไปในการเรียนรู้ (Small-sample-size Problem) ปัญหาที่กล่าวนี้เป็นปัญหาหนึ่งที่ท้าทายมากในด้านเหมืองข้อมูล (Data Mining) และการค้นพบความรู้สำหรับวิศวกร นักวิทยาศาสตร์คอมพิวเตอร์ และนักสถิติ เพราะว่าอัลกอริทึมที่จะพัฒนานั้นจะต้องมีความทนทาน (Robust) และมีประสิทธิภาพที่ดี นอกจากนี้ยังต้องคำนึงถึงชุดข้อมูลที่มีมิติที่สูง และข้อมูลที่มีขนาดใหญ่ เนื่องจากระดับความซับซ้อนของข้อมูลที่คาดไม่ถึง ดังนั้นการควบคุมความซับซ้อนนั้นเป็นสิ่งที่สำคัญและจำเป็นต่อ VS

1.2 วัตถุประสงค์ของการวิจัย

โครงการวิจัยนี้มีจุดประสงค์ที่จะพัฒนาเทคนิคการเรียนรู้ของเครื่องสำหรับข้อมูลที่มีจำนวนของขิอมูลขนาดใหญ่และสำหรับข้อมูลที่มีมิติสูงด้วย เพื่อลดเวลาในการสร้างโมเดล และเวลาในการทำนายข้อมูลที่ได้รับมาใหม่ โดยเน้นไปทางเคมีสารสนเทศ

1.3 ขอบเขตของการวิจัย

โครงการวิจัยนี้ได้พัฒนาอัลกอลิทึมการเรียนรู้ของเครื่อง โดยพัฒนาอัลกอริทึมการวิเคราะห์การจำแนกของฟิชเชอร์โดยใช้ความสัมพันธ์ระหว่างการวิเคราะห์การจำแนกของฟิชเชอร์และวิธีกำลังสองน้อยที่สุด และควบคุมความซับซ้อนของโมเดลโดยใช้ฟังก์ชันการลงโทษ L_1 ที่ถูกประยุกต์ผ่านการใช้การแจกแจงก่อนที่ไม่ให้ข้อมูลของเจฟฟรี ผ่านการตีความลำดับชั้นของเบส์ของการแจกแจงลาพลาสเซียน ทำให้อัลกอริทึมที่นำเสนอไม่จำเป็นต้องใช้เรกิวลาไรเซชันพารามิเตอร์ และเปรียบเทียบอัลกอริทึมที่นำเสนอกับอัลกอริทึมต่างๆ บนข้อมูลต่างๆ และฐานข้อมูลอื่นๆ ด้วย

1.4 วิธีการดำเนินงานวิจัย

โครงงานนี้มีการดำเนินการวิจัยโดยมีขั้นตอน และรายละเอียดต่างๆ ดังต่อไปนี้ คือ

1.4.1 ศึกษาอัลกอริทึมการวิเคราะห์การจำแนกของฟิชเชอร์แบบเชิงเส้น

1.4.2 ศึกษาอัลกอริทึมการวิเคราะห์การจำแนกของฟิชเชอร์แบบเคอเนล

1.4.3 ศึกษาเจฟฟรีไพรเออร์ และการตีความของลาพลาสเชียนไพรเออร์โดยผ่านตามลำดับขั้นของเบส์

1.4.4 ทดสอบอัลกอริทึมที่นำเสนอบนฐานข้อมูล FIRST IDA Repository

1.4.5 ทดสอบอัลกอริทึมที่นำเสนอบนข้อมูลในทางเคมีสารสนเทศ

1.4.6 สรุปผลการทดลองและข้อเสนอแนะ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

อัลกอริทึมใหม่ที่ลดพารามิเตอร์ที่จำเป็นในการหาโมเดลที่ดีที่สุด นั่นคือเรกิวราไรซ์เซชันพารามิเตอร์ ดังนั้นสามารถลดเวลาในการสร้างโมเดลที่ดีที่สุดลงไปได้ นอกจากนี้โมเดลที่ได้ยังเป็นโมเดลที่ไม่ซับซ้อน ดังนั้นในการทดสอบข้อมูลที่เข้ามาใหม่จึงใช้เวลาน้อยลง

โครงการนี้ได้รับการตอบรับให้ตีพิมพ์ในการประชุมวิชาการนานาชาติ โดยมีรายละเอียดดังนี้

- Pasupa, K. (In Press) Sparse Fisher Discriminant Analysis with Jeffrey's Hyperprior, In: Proceeding of the International Conference on Control, Automation & Information Sciences (ICCAIS'2012), 26-29 November 2012, Ho Chi Minh City, Vietnam.

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การวิเคราะห์การจำแนกของฟิชเชอร์ (Fisher Discriminant Analysis - FDA)

พิจารณาเมทริกที่ประกอบไปด้วยเวกเตอร์ที่มีขนาด m, $\mathrm{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_N \end{bmatrix}^{\mathrm{T}}$,ที่ประกอบไปด้วยข้อมูลสองกลุ่ม, G_i, ที่มีขนาด N_i, i = 1,2 โดยแสดงอยู่ในรูปของพาร์ทิชัน $\begin{bmatrix} \mathrm{X}_1 & \mathrm{X}_2 \end{bmatrix}^{\mathrm{T}}$ โดยที่สมาชิกของ G_1 ถูกกำหนดด้วย $\hat{y} = +N/N_1$ และ G_2 ด้วย $\hat{y} = -N/N_2$ และสัมประสิทธิ์ของการจำแนกของฟิชเชอร์ถูกหนดด้วย $\mathbf{w}$ ซึ่งทำให้ฟังก์ชันจุดประสงค์ (Objective Function) มีค่าสูงที่สุด

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathrm{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathrm{S}_{\mathrm{W}}\mathbf{w}} \tag{1}$$

S_{B} คือ เมทริกการกระจายตัวระหว่างกลุ่ม ซึ่งคำนวนได้จาก $\mathrm{S}_{\mathrm{B}} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}}$ และ S_{W} คือ เมทริกการกระจายตัวภายในกลุ่ม ซึ่งคำนวนได้จาก $\mathrm{S}_{\mathrm{W}} = \sum_{i=1,2}\sum_{\mathbf{x}\in G_i}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^{\mathrm{T}}$ โดยที่ $\mathbf{m}_i = \frac{1}{N_i}\sum_{\mathbf{x}\in G_i}\mathbf{x}$

คำตอบที่ดีที่สุดแท้จริง (Global Optimal) ของสมการ (1) ซึ่งถูกเรียกว่า ผลลัพธ์ของเรย์ลี (Rayleigh Quotient) นั้นสามารถหาได้จากการแก้ปัญหาค่าเจาะจง (Eigenvalue Problem) [1] ดังนั้น

$$\mathbf{w} = \left(\mathrm{S}_{\mathrm{W}}\right)^{-1}\left(\mathbf{m}_1 - \mathbf{m}_2\right) \tag{2}$$

อีกทั้งเป็นที่รู้กันดีว่า FDA มีความสัมพันธ์กับ LS ในการจำแนกประเภท [1] พิจารณาฟังก์ชันเส้นตรงที่ประกอบด้วยค่าความเอนเอียง b, (Bias Term), $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$ จุดประสงค์ของ LS คือพยายามที่จะทำให้ค่าของผลรวมของความผิดพลาดกำลังสอง (Sum of Square Error - SSE) นั้นน้อยที่สุด

$$\begin{aligned} \mathrm{SSE}(b,\mathbf{w}) &= \sum_{i=1}^{N}\left(\hat{y}_i - f(\mathbf{x}_i)\right)^2 \\ &= \sum_{i=1}^{N}\left(\hat{y}_i - \mathbf{w}^{\mathrm{T}}\mathbf{x} + b\right)^2 \end{aligned} \tag{3}$$

ซึ่งสามารถเขียนให้อยู่ในรูปของเมทริกได้ดังนี้

$$\arg\min_{(b,\mathbf{w})}\left\|\begin{bmatrix} \frac{N}{N_1}\mathbf{1}N_1 \\ \frac{N}{N_2}\mathbf{1}N_2 \end{bmatrix} - \begin{bmatrix} \mathbf{1}N_1 & \mathrm{X}_1 \\ \mathbf{1}N_2 & \mathrm{X}_2 \end{bmatrix}\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}\right\|_2^2 \tag{4}$$

คำตอบของปัญหา LS ของ $\left\|\hat{\mathbf{y}} - \tilde{\mathrm{X}}\boldsymbol{\omega}\right\|_2^2$ สามารถคำนวนได้จากซูโดอินเวอร์ส (Pseudo-inverse) $\tilde{\mathrm{X}}^{\dagger}$ ของ $\tilde{\mathrm{X}}$,

$$\boldsymbol{\omega} = \tilde{\mathrm{X}}^{\dagger}\hat{\mathbf{y}} \tag{5}$$

ที่ซึ่ง $\tilde{\mathbf{X}}^{\dagger}=\left(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}$, $\boldsymbol{\omega}=\begin{bmatrix} b & \mathbf{w}^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}$, $\tilde{\mathbf{X}}=\begin{bmatrix}\mathbf{1}_N & \mathbf{X}\end{bmatrix}$, และ $\mathbf{1}_N$ คือเวกเตอร์คอลัมน์หนึ่งที่มีขนาด N เพราะฉะนั้น $\left(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}\right)^{-1}\boldsymbol{\omega}=\tilde{\mathbf{X}}^{\mathrm{T}}\hat{\mathbf{y}}$ และแทนสมการนี้ด้วยนิยามของเมทริกการกระจายตัวระหว่างกลุ่มและ ภายในกลุ่ม คำตอบที่ได้ของ LS จะไปในทิศทางเดียวกันกับ FDA ดังสมการ (2)

อย่างไรก็ตาม การใช้ฟังก์ชันเส้นตรงนั้นไม่เพียงพอที่จะทำให้อัลกอริทึมสามารถทำงานได้ดีในข้อมูลในปัจจุบัน ดังนั้นการใช้เคอเนลก็เป็นหนึ่งในการแก้ปัญหานี้ [5] ซึ่งได้ถูกนำมาใช้ใน [3] และ [6]

2.2 การวิเคราะห์การจำแนกของฟิชเชอร์แบบเคอเนล (Kernel FDA - kFDA)

ในโครงงานนี้ จะปรับปรุงประสิทธิภาพของ kFDA ที่นำเสนอโดย [3] อัลกอลิทึมนี้ใช้ความสัมพันธ์ระหว่าง FDA และ LS ร่วมกับการควบคุมความซับซ้อนของโมเดลโดยใช้ฟังก์ชันลงโทษ L_q โดยที่ $0 < q < 1$ อีกทั้งเป็นที่รู้จักกันดีว่าฟังก์ชันลงโทษนี้มีความสามารถในการทำให้โมเดลมีความซับซ้อนน้อยลง ยิ่งไปกว่านั้นยังทำให้ฟังก์ชันจุดประสงค์นั้นเกิดปัญหาและ แก้ปัญหาได้ยากขึ้น ซึ่ง [7] สามารถแก้ปัญหานี้ได้โดยการใช้วิธีเมเจอไรซ์-มินนิไมซ์ (Majorize-Minimize Algorithm) ซึ่งจะนำไปสู่อัลกอริทึมวนซ้ำที่ไม่ซับซ้อน ฟังก์ชันลอก-ไลคลีฮูด $\ell(\boldsymbol{\omega})$ (Log-likelihood Function) ของปัญหานี้สามารถเขียนอยู่ในรูปผลบวกของฟังก์ชันสองฟังก์ชัน นั่นคือ $\ell_e(\boldsymbol{\omega})=\frac{1}{2}\left\|\hat{\mathbf{y}}-\tilde{\mathbf{K}}\boldsymbol{\omega}\right\|_2^2$ และ $\ell_p(\boldsymbol{\omega})=\rho N\left\|\boldsymbol{\omega}\right\|_q^q$ โดยที่ $\hat{\mathbf{y}}=\begin{bmatrix}\frac{N}{N_1}\mathbf{1}^{\mathrm{T}}N_1 & \frac{N}{N_2}\mathbf{1}^{\mathrm{T}}N_2\end{bmatrix}^{\mathrm{T}}\in\Re^N$ และ $\tilde{\mathbf{K}}=\begin{bmatrix}\mathbf{1}_N & \mathbf{K}\end{bmatrix}\in\Re^{N\times(N+1)}$

ดังนั้น

$$\ell(\boldsymbol{\omega})=\frac{1}{2}\left\|\hat{\mathbf{y}}-\tilde{\mathbf{K}}\boldsymbol{\omega}\right\|_2^2+\rho N\left\|\boldsymbol{\omega}\right\|_q^q \tag{6}$$

โดยที่ K_i คือ แกรมเมทริก (Gram Matrix) ที่เกิดจากเคอเนล $k(.,.)$ นั่นคือ $k_{jl}=k(\mathbf{x}_j,\mathbf{x}_l)$, $j, l = 1, 2, ..., N$, $\mathbf{x}_l\in G_i$ และ ρ คือเรกิวราไลซ์เซชันพารามิเตอร์, $\boldsymbol{\omega}$ คือสัมประสิทธิ์ของฟังก์ชันการจำแนกแบบเส้นตรงของฟีเจอร์สเปสซ์ของ $k(.,.)$ ซึ่งส่งผลให้เกิดการจำแนกแบบไม่เป็นเส้นตรงในฟีเจอร์สเปสซ์ที่แท้จริง

แก้ปัญหาสมการ (6) โดยใช้วิธีเมเจอไรซ์-มินนิไมซ์จะได้สมการวนซ้ำดังนี้

$$\boldsymbol{\omega}(n+1)=\left(\tilde{\mathbf{K}}^{\mathrm{T}}\tilde{\mathbf{K}}+\rho Nq\mathrm{B}(\boldsymbol{\omega}(n))\right)^{-1}\tilde{\mathbf{K}}^{\mathrm{T}}\hat{\mathbf{y}} \tag{7}$$

โดยที่ $\mathrm{B}(\boldsymbol{\omega}(n))=\mathrm{diag}\left\{\left|\omega_i(n)\right|^{q-2}\right\}$ อย่างไรก็ตามจะมีปัญหาที่ตามมาเวลา $\boldsymbol{\omega}(n)$ ลู่เข้าใกล้ศูนย์ [3] แต่ก็สามารถแก้ไขปัญหาได้โดยการเขียน $\mathrm{B}(\boldsymbol{\omega}(n))=\boldsymbol{\psi}_n^{-2}$ โดย $\boldsymbol{\psi}_n=\mathrm{diag}\left\{\left|\omega_i(n)\right|^{\frac{2-q}{2}}\right\}$ [8] ดังนั้นสมการทำซ้ำ (7) จะเปลี่ยนเป็น

$$\boldsymbol{\omega}(n+1)=\boldsymbol{\psi}_n\left(\boldsymbol{\psi}_n\tilde{\mathbf{K}}^{\mathrm{T}}\tilde{\mathbf{K}}\boldsymbol{\psi}_n+\rho Nq\mathbf{I}_{N+1}\right)^{-1}\boldsymbol{\psi}_n\tilde{\mathbf{K}}^{\mathrm{T}}\hat{\mathbf{y}};\ \boldsymbol{\omega}(0)\neq\mathbf{0} \tag{8}$$

ในการคำนวนของการใช้ฟีเจอร์พื้นฐาน (Primal) นั่นหมายความว่าพิจารณา $\tilde{\mathrm{X}}$ แทน $\tilde{\mathrm{K}}$ ซึ่งจำนวนของฟีเจอร์นั้นมีขนาดใหญ่กว่าจำนวนข้อมูล ($m>N$) ความเป็นเอกลักษณ์ของ $\mathrm{M}_1, \mathrm{M}_2, \mathrm{M}_3$ สามารถหาได้ด้วย $\mathrm{M}_1, \mathrm{M}_3$ ซึ่งมีคุณสมบัติคือ เป็นบวกอย่างแน่นอน (Positive Definite) และสามารถอินเวิร์สได้ดังนี้

$$\left(\mathrm{M}_1^{-1}+\mathrm{M}_2^{\mathrm{T}}\mathrm{M}_3^{-1}\mathrm{M}_2\right)^{-1}\mathrm{M}_2^{\mathrm{T}}\mathrm{M}_3^{-1}=\mathrm{M}_1\mathrm{M}_2^{\mathrm{T}}\left(\mathrm{M}_2\mathrm{M}_1\mathrm{M}_2^{\mathrm{T}}+\mathrm{M}_3\right)^{-1} \tag{9}$$

จากนั้น แทนค่าต่างๆ โดยที่ $\mathrm{M}_1^{-1}=\rho\mathrm{B}(\boldsymbol{\omega}(n))$, $\mathrm{M}_2=\tilde{\mathrm{X}}$, และ $\mathrm{M}_3^{-1}=Nq\mathbf{I}_N$ ซึ่งจะได้

$$\boldsymbol{\omega}(n+1)=\mathrm{B}^{-1}(\boldsymbol{\omega}(n))\tilde{\mathrm{X}}^{\mathrm{T}}\left(\tilde{\mathrm{X}}\mathrm{B}^{-1}(\boldsymbol{\omega}(n))\tilde{\mathrm{X}}^{\mathrm{T}}+\rho Nq\mathbf{I}_N\right)^{-1}\hat{\mathbf{y}} \tag{10}$$

อัลกอริทึมจะหยุดทำงานเมื่อมีการลู่เข้าของการเปลี่ยนแปลงของขนาดของเวกเตอร์สัมประสิทธิ์ นั่นหมายความว่าการเปลี่ยนแปลงของขนาดของเวกเตอร์สัมประสิทธิ์จะต้องน้อยกว่าค่าคงที่เมื่อ $\varepsilon \ll 1$ จากนั้นค่าสัมประสิทธิ์จะถูกปรับให้เท่ากับศูนย์ ในกรณีที่ $\dfrac{|\omega|}{\arg\max|\omega_i|}<\eta;\eta\ll 1$ และค่าขอบเขตการแบ่งแยก คือ ค่ากลางของ y นั่นคือ

$$c=\frac{1}{2}\left(\frac{N}{N_1}-\frac{N}{N_2}\right) \tag{11}$$

ดังนั้น กฎของการแบ่งแยกกลุ่มคือ เป็น G_1 ถ้า $b+\sum_{i=1}^{N}\alpha_i k(\mathbf{x},\mathbf{x}_i)\geq c$ และถ้าไม่ใช่ให้เป็น G_2 อัลกอริทึมนี้จะถูกเรียกว่า kFDA_q และ FDA_q สำหรับ เคอเนล และ สมการเส้นตรง ตามลำดับ

2.3 การกำจัดเรกิวราไรซ์พารามิเตอร์

จากที่ได้อธิบายไปข้างต้น เพื่อที่จะกำจัดเรกิวราไรซ์พารามิเตอร์ จำเป็นต้องประยุกต์ใช้การแจกแจงก่อนที่ไม่ให้ข้อมูลของเจฟฟรี ในฟังก์ชันลงโทษ ผ่านการตีความลำดับชั้นของเบส์ ของการแจกแจงลาพลาสเซียน (q=1) [4]

พิจารณาลาพลาสเซียนไพรเออร์ (Laplacian Prior)

$$\begin{aligned}p(\boldsymbol{\omega}\mid\rho)&=\prod_{i=1}^{k}\frac{\rho}{2}\exp\{-\rho|\omega_i|\}\\&=\left\{\frac{\rho}{2}\right\}^{k}\exp\{-\rho|\omega|_1^1\}\end{aligned} \tag{12}$$

เนื่องจากไพรเออร์นี้ไม่สามารถหาอนุพันธ์ได้ (Non-differentiable) ดังนั้นจึงต้องใช้การตีความลำดับชั้นของเบส์มาช่วยแก้ปัญหาโดยพิจารณาจากเกาส์เซียนไพรเออร์ (Gaussian Prior) ที่มีค่าเฉลี่ยเท่ากับศูนย์และค่าความแปรปรวน τ_i

$$p(\boldsymbol{\omega}_i\mid\tau_i)=N(\boldsymbol{\omega}_i\mid 0,\tau_i) \tag{13}$$

และเอกซ์โพเนนเชียลไพรเออร์

$$p(\tau_i\mid\rho)=\frac{\rho}{2}\exp\{-\frac{\rho}{2}\tau_i\} \tag{14}$$

อินทิเกรตทั้งสองฟังก์ชันเทียบกับ τ_i

$$p(\omega_i \mid \rho) = \int_0^\infty p(\omega_i \mid \tau_i) p(\tau_i \mid \rho) d\tau_i$$
$$= \frac{\sqrt{\rho}}{2} \exp\{-\sqrt{\rho}|\omega|\} \quad (15)$$

สมการข้างต้นได้แสดงถึงลาพลาสเชียนไพรเออร์ที่มาจากการรวมโดยการตีความลำดับชั้นของเบส์สองระดับนั่นคือ เกาส์เชียนไพรเออร์ที่มีค่าเฉลี่ยเท่ากับศูนย์และค่าความแปรปรวน τ_i กับเอกซ์โพเนนเชียลไพรเออร์ จะเห็นได้ว่าสมการดังกล่าวยังมีเรกิวราไรซ์พารามิเตอร์ ρ อยู่ ดังนั้น [4] จึงได้แทนเอกซ์โพเนนเชียลไพรเออร์ด้วยเจฟฟรีไพรเออร์สำหรับ τ_i

$$p(\tau_i) \propto \frac{1}{\tau_i} \quad (16)$$

อินทิเกรตเกาส์เชียนไพรเออร์และเจฟฟรีไพรเออร์เทียบกับ τ_i จะได้

$$p(\omega_i) = \int_0^\infty p(\omega_i \mid \tau_i) p(\tau_i) d\tau_i = |\omega_i|^{-1} \quad (17)$$

[4] ได้แก้ปัญหานี้ผ่านอัลกอริทึมอีเอ็ม (Expectation–maximization Algorithm - EM) ซึ่งจะได้อัลกอริทึมสองขั้นตอนที่มีลักษณะคล้ายสมการ (8) คือ

$$\sigma^2(n+1) = \frac{1}{N} \left\| \hat{\mathbf{y}} - \tilde{\mathbf{K}}\boldsymbol{\omega} \right\|_2^2 \quad (18)$$

และ

$$\boldsymbol{\omega}(n+1) = \boldsymbol{\psi}_n \left(\boldsymbol{\psi}_n \tilde{\mathbf{K}}^{\mathrm{T}} \tilde{\mathbf{K}} \boldsymbol{\psi}_n + \sigma^2(n+1) \mathbf{I}_{N+1} \right)^{-1} \boldsymbol{\psi}_n \tilde{\mathbf{K}}^{\mathrm{T}} \hat{\mathbf{y}}; \boldsymbol{\omega}(0) \neq \mathbf{0} \quad (19)$$

โดยที่ $\boldsymbol{\psi}_n = \mathrm{diag}\{|\omega_i(n)|\}$

ในการคำนวนของการใช้ฟีเจอร์พื้นฐาน เมื่อ $m > N$ อัลกอริทึมจะเปลี่ยนเป็น

$$\sigma^2(n+1) = \frac{1}{N} \left\| \hat{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\omega} \right\|_2^2 \quad (20)$$

และ

$$\boldsymbol{\omega}(n+1) = \mathrm{B}^{-1}(\boldsymbol{\omega}(n)) \tilde{\mathbf{X}}^{\mathrm{T}} \left(\tilde{\mathbf{X}} \mathrm{B}^{-1}(\boldsymbol{\omega}(n)) \tilde{\mathbf{X}}^{\mathrm{T}} + \sigma^2(n+1) \mathbf{I}_N \right)^{-1} \hat{\mathbf{y}} \quad (21)$$

อัลกอริทึมนี้จะถูกเรียกว่า $\mathrm{kFDA}_{\mathrm{Jef}}$ และ $\mathrm{FDA}_{\mathrm{Jef}}$ สำหรับเคอเนลและสมการเส้นตรง ตามลำดับ

บทที่ 3
วิธีดำเนินงานวิจัย

ประสิทธิภาพของอัลกอริทึมที่นำเสนอได้ถูกทดสอบและเปรียบเทียบกับข้อมูล 2 ฐานข้อมูลดังนี้

3.1 FIRST IDA Repository Database

เพื่อประเมินประสิทธิภาพในการทำงานของอัลกอริทึมที่นำเสนอ $kFDA_{Jef}$ ได้ถูกนำมาทดลองบนชุดข้อมูลทั้ง 13 ชุดจากฐานข้อมูล FIRST IDA [9] การทดลองนั้นได้ถูกวางรูปแบบให้เหมือนกับ [3] เพื่อที่จะสามารถเปรียบเทียบได้โดยตรงกับวิธีต่างๆ ที่ใช้ในบทความอ้างอิง ซึ่งผลการทดลองของวิธีต่างๆ นั้นได้ถูกรวบรวมและรายงานอยู่ใน [3]

ในการทดลองนั้นได้ใช้เลือกใช้เรเดียลเบซิสฟังก์ชัน (Radial Basis Function) เป็นเคอเนลฟังก์ชั่น การตรวจสอบไขว้จำนวน 5 พับ (5-fold Cross Validation) ได้ถูกใช้ในการหาโมเดลที่ดีที่สุดนั่นคือ เคอเนลพารามิเตอร์เพียงพารามิเตอร์เดียว จากที่กล่าวถึงในหัวข้อก่อนหน้านี้ว่า การใช้การกระจายของเจฟฟรีย์ในฟังก์ชันลงโทษนั้นไม่จำเป็นต้องใช้เรกิวลาไรเซชันพารามิเตอร์ ส่วนเกณฑ์ในการเลือกโมเดลนั้น เลือกจากอัตราความผิดพลาด (Misclassification Rate - MCR) ที่ดีที่สุดจาก 5 พาร์ทิชันแรก และใช้โมเดลนั้นทดสอบกับทุกๆ 100 พาร์ทิชัน ยกเว้นแต่ Splice และ Image ที่มีเพียง 20 พาร์ทิชัน การทดลองนั้นถูกทดสอบบน Matlab [10]

3.2 MDL Drug Data Report (MDDR)

ในการประยุกต์กับการคัดกรองเสมือนนั้น ได้ทดลองกับข้อมูลจำนวน 11 แอคทิวิตี้คลาส (Activity Class) จากฐานข้อมูล MDL Drug Data Report (MDDR) [11] ในการเลือกแอคทิวิตี้คลาสนั้นได้เลือกจากแอคทิวิตี้คลาสที่ใช้ในกระบวนการค้นพบยาของบริษัทผลิตยาซึ่งถูกแสดงอยู่ในตารางที่ 3.1 ฐานข้อมูล MDDR ประกอบไปด้วยชุดของยาและโมเลกุลที่รู้รายละเอียดโครงสร้างที่เก็บรวบรวมมาจากสิทธิบัตร วารสาร และการประชุม ฐานข้อมูล MDDR นั้นถูกแสดงด้วยลายนิ้วมือ (Fingerprint) ชนิด ECFP_4 ที่แสดงผลด้วยเลขฐานสองที่มีขนาด 1024 บิท [12]

การทดลองได้ถูกแบ่งออกเป็น 2 ส่วน คือ การวิเคราะห์การจำแนกของฟิชเชอร์แบบเส้นตรงและเคอเนล การทดลองได้ถูกสุ่มแยกข้อมูลที่ใช้ในการเรียนรู้และทดสอบต่างๆกันจำนวนทั้งหมด 5 ครั้ง จำนวนข้อมูลที่ใช้ในการเรียนรู้คือ 20% ของจำนวนโมเลกุลที่มีฤทธิ์ (Active Molecule) ในฐานข้อมูล ข้อมูลที่ใช้ในการทดสอบจะถูกจัดอันดับโดยเรียงจากค่าที่มีผลบวกมากที่สุด (มีแน้วโน้มมากที่จะออกฤทธิ์) จนถึงค่าที่มีผลลบมากที่สุด (มีแน้วโน้มมากที่จะไม่ออกฤทธิ์) กล่าวคือผลลัพธ์ที่ออกมานั้นคือระยะทางจากขอบเขตการตัดสินใจ (Decision Boundary) ในการทดลองได้ใช้การตรวจสอบไขว้จำนวน

5 พับ บนพื้นฐานของค่าที่น้อยที่สุดของผลรวมของอันดับของโมเลกุลที่มีฤทธิ์ในการปรับค่าพารามิเตอร์ของโมเดลนั้นๆ

ตารางที่ 3.1 11 แอกทิวิตีคลาสทั้ง 11 คลาสจากฐานข้อมูล MDDR

Activity Class	Index	No. of Actives	No. of Training Sample
Renin Inhibitors	1	1130	226
Angiotensin II AT1 Antagonists	2	943	190
HIV Protease Inhibitor	3	750	150
Thrombin Inhibitor	4	803	162
Substance P Antagonists	5	1246	250
5HT3 Antagonists	6	752	150
D2 Antagonists	7	395	80
5HT1A Agonists	8	827	166
5HT Reuptake Inhibitors	9	359	72
Protein Kinase C Inhibitor	10	453	92
Cyclo-oxygenase Inhibitor	11	636	128

บทที่ 4
ผลการวิจัย

4.1 FIRST IDA Repository Database

จากตารางที่ 4.1 $kFDA_{Jef}$ ได้ผลดีที่สุด 5/13 ชุดข้อมูล (เท่ากับ $kFDA_q$ เมื่อ $q \in \{1,0.5\}$ ใน Titanic) นอกจากนั้นโมเดลจาก $kFDA_{Jef}$ นั้นให้ผลของค่าความซับซ้อนน้อยที่สุดใน 3/13 ชุดข้อมูล ถึงแม้ว่าค่าเฉลี่ยของความซับซ้อนของ $kFDA_{Jef}$ ยังมากกว่า $kFDA_{0.5}$ แต่ในทางกลับกันค่าเฉลี่ยของอัตราความผิดพลาดนั้นน้อยกว่า $kFDA_{0.5}$ ส่วน $kFDA_1$ นั้นยังคงได้ค่าความผิดพลาดเฉลี่ยน้อยที่สุด

4.2 MDL Drug Data Report (MDDR)

ผลการทดลองได้ถูกแสดงอยู่ในตารางที่ 4.2 และ 4.3 สำหรับการวิเคราะห์การจำแนกของฟิชเชอร์แบบเส้นตรงและเคอเนลตามลำดับ ในการรายงานประสิทธิ์ภาพของโมเดลนั้น ถูกรายงานด้วยร้อยละของจำนวนของสารประกอบที่ออกฤทธิ์ที่อยู่ใน 5% แรกของจำนวนสารประกอบที่ถูกเรียงลำดับในฐานข้อมูลทั้งหมด ตามด้วยร้อยละของลักษณะเด่นที่ใช้ ผลการทดสอบนั้นได้ถูกเปรียบเทียบกับ Binary Kernel Discrimination (BKD) ที่ถูกพัฒนาโดย [20] ที่เป็นอัลกอริทึมการเรียนรู้ของเครื่องที่ดีที่สุดในเคมีสารสนเทศในปัจจุบัน นอกจากนี้ในตารางผลการทดลองได้แสดงถึงค่าเฉลี่ยของค่าความคล้ายคลึง (Mean Self-similarity) ที่วัดจากความเป็นเอกพันธ์ (Homogeneity) ของแต่ละแอกทิวิตีคลาส ซึ่งเป็นค่าที่ใช้ในการเปรียบเทียบการกระจายและครอบคลุมของข้อมูล

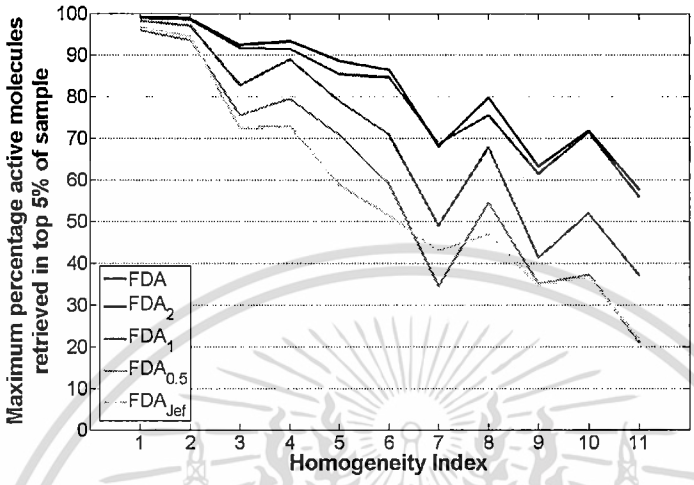
4.2.1 ผลการทดลองของ FDA, FDA_q และ FDA_{Jef} แสดงอยู่ในตารางที่ 4.2 ประสิทธิภาพของ FDA นั้นดีที่สุดใน 1/11 คลาสแต่ใช้ลักษณะเด่นทั้งหมด ขณะที่ FDA_2 แสดงค่าความถูกต้องดีที่สุดและใช้ลักษณะเด่น 86.00% โดยเฉลี่ย อย่างไรก็ตาม ลดค่า q เหลือ 1 ส่งผลให้ค่าเฉลี่ยของขนาดโมเดลลดลงเหลือ 7.37% ขณะที่สูญเสียความถูกต้องเพียงเล็กน้อย (<3%) ในแอกทิวิตีที่มีความเอกพันธ์สูงสุดใน 2 คลาส แต่สูญเสียความถูกต้องมากใน 9/11 ที่เหลือ โดยรวมแล้ว ค่าเฉลี่ยลดลงจาก 81.76% เหลือเพียง 69.45% ยิ่งไปกว่านั้น ลด q เหลือเพียง 0.5 ส่งผลให้โมเดลมีขนาดลดลงอีก ขณะที่สูญเสียค่าความถูกต้องมากใน 10/11 คลาส สำหรับการใช้ FDA_{Jef} นั้นขนาดของโมเดลลดลงเหลือ 9.38% และแน่นอนความถูกต้องลดลงและน้อยกว่า $kFDA_q$ เมื่อ $q \in \{1,0.5\}$ อย่างไรก็ตาม ค่าความถูกต้องในสองคลาสแรกที่มีความเป็นเอกพันธ์สูงสุดยังคงระดับอยู่ ผลการทดลองของทุกโมเดลได้ถูกสรุปอยู่ในรูปที่ 4.1 และ 4.2 สำหรับความถูกต้องและขนาดของโมเดลตามลำดับ กล่าวสรุปคืออัลกอริทึมที่นำเสนอนั้นมีประสิทธิภาพอยู่ในระดับเดียวกันกับ BKD ในคลาสที่มีความเอกพันธ์สูง แต่โดยรวมแล้วประสิทธิภาพยังคงน้อยกว่า BKD [19] อย่างไรก็ตามประสิทธิภาพนั้นยังสามารถพัฒนาได้อีกโดยใช้เคอเนล ซึ่งจะนำไปสู่การจำแนกประเภทที่ซับซ้อนยิ่งขึ้น

ตารางที่ 4.1 FIRST IDA Repository Database: เปรียบเทียบค่าเฉลี่ยของค่าความผิดพลาด, ค่าความแปรปรวนของค่าความผิดพลาด, และค่าความซับซ้อนของโมเดล (ข้างล่าง) สำหรับอัลกอริทึม $kFDA_{Jef}$, $kFDA_q$, และ อัลกอริทึมที่ดีที่สุดจากการทบทวนวรรณกรรมที่ใช้ข้อมูลเดียวกัน: Import Vector Machine [13] (★), Conventional kFDA [14] (△), Reformative kFDA [14] (▲), Naive Kernel-based Nonlinear Method [15] (▽), Fast Kernel-Based Nonlinear Method [15] (▼), Kernel Logistic Regression [16] (□), Sparse kFDA with Linear Loss [17] (■), Linear Programming Adaboost [18] (◇), และ Suppressed Kernel Sample Space Projection [19] (◆) ตัวเลขใต้ชื่อข้อมูลคือ จำนวนตัวอย่าง, ตัวหนาคือ ค่าที่ดีที่สุด

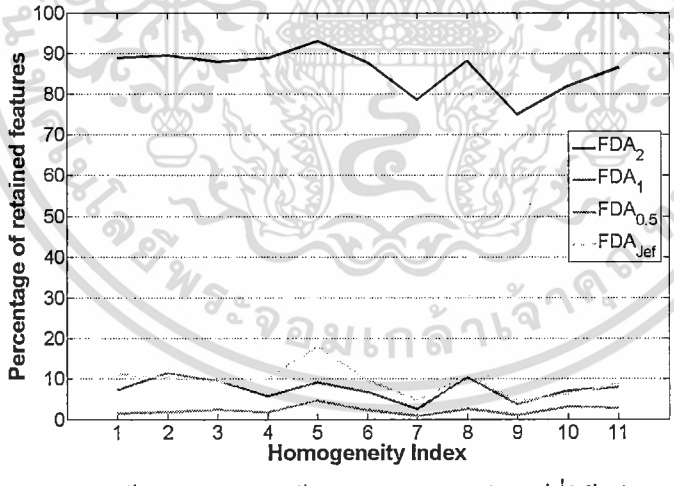
Database	Published Best	$kFDA_1$	$kFDA_{0.5}$	$kFDA_{Jef}$
Banana	10.34 ± 0.46	9.74 ± 0.10	**9.63 ± 0.10**	11.76 ± 0.10
400	5.25 ± 1.75 ★	14.00	**4.50**	5.00
B. Cancer	22.70 ± 4.40	21.26 ± 3.70	20.55 ± 3.96	**20.09 ± 3.81**
200	100.00 △	13.00	**3.50**	10.50
Diabetes	22.10 ± 1.90	21.57 ± 1.41	21.58 ± 1.63	**21.18 ± 1.55**
468	100.00 △	5.13	**1.28**	2.14
German	21.30 ± 2.10	21.19 ± 1.81	23.51 ± 2.03	**20.77 ± 1.70**
700	100.00 △	8.86	**1.14**	5.14
Heart	**10.80 ± 2.60**	14.56 ± 3.04	14.84 ± 3.31	13.75 ± 2.54
170	16.00 ▲	7.06	**2.35**	5.29
Image	1.78 ± N/A	**1.56 ± 0.46**	1.98 ± 0.32	2.52 ± 0.77
1300	100.00 □	24.77	**15.46**	38.15
Ringnorm	1.50 ± 0.10	**1.42 ± 0.04**	1.48 ± 0.03	1.74 ± 0.04
400	6.00 ■	4.75	**2.75**	22.25
S. Flare	31.60 ± 1.90	32.98 ± 1.74	31.64 ± 1.85	**30.95 ± 1.87**
666	100.00 ▽	27.03	**2.58**	10.45
Splice	9.30 ± 0.70	7.07 ± 0.71	**7.02 ± 0.83**	7.72 ± 0.64
1000	100.00 ◇	85.00	75.90	**51.70**
Thyroid	1.40 ± 0.90	**0.99 ± 0.90**	1.05 ± 0.93	1.83 ± 1.15
140	16.40 ▼	22.86	12.86	**6.43**
Titanic	21.70 ± 0.30	**21.10 ± 0.23**	**21.10 ± 0.23**	**21.10 ± 0.23**
150	100.00 ▽	64.67	4.67	**4.00**
Twonorm	**2.30 ± 0.1**	2.36 ± 0.04	2.58 ± 0.04	2.60 ± 0.04
400	100.00 ◆	7.75	**1.25**	4.25
Waveform	9.30 ± 0.4	**9.26 ± 0.13**	9.97 ± 0.13	9.56 ± 0.13
400	100.00 ◇	7.00	**3.00**	7.25
Average	**12.78**	**12.70**	12.84	12.74
	72.59	22.45	**10.10**	13.27

ตารางที่ 4.2 แสดงร้อยละของจำนวนของสารประกอบที่ออกฤทธิ์ที่อยู่ใน 5% แรกของจำนวนสารประกอบที่ถูกเรียงลำดับในฐานข้อมูลทั้งหมด โดย FDA แบบเชิงเส้นตรง

Index	Self-similarity		FDA	FDA_2	FDA_1	$FDA_{0.5}$	FDA_{Jef}
	Mean	S.D.	(%)	(%)	(%)	(%)	(%)
1	0.337	0.105	98.99	**99.13**	98.29	96.01	96.72
				88.87	7.15	**1.50**	11.07
2	0.296	0.100	98.47	**98.84**	97.03	93.61	94.48
				89.45	11.41	**1.80**	10.37
3	0.226	0.101	91.67	**92.41**	82.76	75.56	72.39
				87.95	9.43	**2.42**	9.28
4	0.212	0.098	91.55	**93.38**	89.03	79.47	72.96
				88.79	5.63	**1.74**	10.14
5	0.179	0.082	85.51	**88.60**	78.93	70.78	58.79
				92.99	9.14	**4.71**	17.97
6	0.175	0.090	84.67	**86.53**	70.90	59.03	51.37
				87.79	6.76	**2.38**	9.59
7	0.173	0.089	**68.68**	68.00	48.90	34.37	42.87
				78.59	2.54	**0.86**	4.71
8	0.166	0.086	75.56	**79.81**	67.90	54.46	46.75
				88.14	10.35	**2.66**	10.84
9	0.153	0.092	61.49	**63.22**	41.30	35.05	34.67
				75.00	3.69	**1.07**	4.36
10	0.141	0.103	71.40	**71.79**	51.89	37.30	36.71
				81.91	7.03	**3.13**	6.17
11	0.130	0.073	55.91	**57.62**	37.06	21.05	21.68
				86.50	7.91	**2.83**	8.71
Average			80.34	**81.76**	69.45	59.70	57.22
				86.00	7.37	**2.28**	9.38



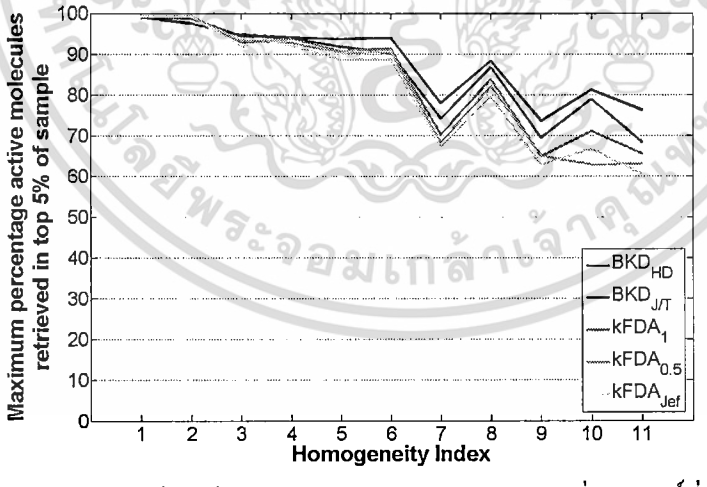
ภาพที่ 4.1 แสดงค่าความถูกต้อง (ร้อยละของจำนวนของสารประกอบที่ออกฤทธิ์ที่อยู่ใน 5% แรกของจำนวนสารประกอบที่ถูกเรียงลำดับในฐานข้อมูลทั้งหมด) เรียงลำดับตามค่าความเอกพันธ์ของ 11 แอกทิวิตีคลาสจากฐานข้อมูล MDDR สำหรับ FDA, FDA_q และ FDA_{Jef}



ภาพที่ 4.2 แสดงค่าความซับซ้อนของโมเดล (ร้อยละของจำนวนฟีเจอร์ที่ใช้) เรียงลำดับตามค่าความเอกพันธ์ของ 11 แอกทิวิตีคลาสจากฐานข้อมูล MDDR สำหรับ FDA_q และ FDA_{Jef}

4.2.2 ในส่วนนี้การแจกแจงทวินาม (Binomial distribution) ได้ถูกใช้เป็นฟังก์ชันเคอเนล $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \lambda^{m-d(\mathbf{x}_i,\mathbf{x}_j)}(1-\lambda)^{d(\mathbf{x}_i,\mathbf{x}_j)}$ โดยที่ $\lambda \in (0.5,1)$ เป็นความกว้างของฟังก์ชัน, m = 1024, และ $d(\mathbf{x}_i, \mathbf{x}_j)$ คือ ระดับความแตกต่าง (ระยะทาง) ระหว่างโมเลกุล i และ j เคอเนลนี้เป็นฟังก์ชันชนิดเรเดียลเบสิส (Radial Basis) [5, p. 46] ในการทดลองของ [20] ได้แสดงผลว่าฟังก์ชันระยะทางแจคคาด/ทานิโมโตะ (Jaccard/Tanimoto Distance - J/T) มีประสิทธิภาพที่ดีกว่าฟังก์ชันระยะทางแฮมมิง (Hamming Distance - HD) เมื่อใช้ใน BKD ฟังก์ชันระยะทาง J/T นั้นมีคุณสมบัติเป็นเมทริก [21] ดังนั้นการแทนใช้ฟังก์ชันระยะทาง J/T ในฟังก์ชันเคอเนลการแจกแจงทวินามนั้น สามารถทำได้โดยถูกต้อง ในอดีตมีงานวิจัยหลายงานที่พยายามที่จะการใช้ฟังก์ชันระยะทางอื่นๆ แทนในฟังก์ชันเคอเนลต่างๆ เช่น [22, 23] ในการทดลองนี้จึงใช้ฟังก์ชันเคอเนลแบบทวินามโดยฟังก์ชันระยะทาง J/T ในการทดสอบเปรียบเทียบอัลกอริทึม $kFDA_{Jef}$ กับ $kFDA_q$ และ BKD ที่ซึ่งถูกรายงานใน [24] จากผลการทดลองจะเห็นได้ว่า $BKD_{J/T}$ นั้นยังคงเป็นอัลกอริทึมที่ทำนายได้ถูกต้องที่สุด ซึ่งได้ผลดีที่สุดใน 8/11 คลาส แต่โมเดลใช้ทุกฟีเจอร์ (100%) ขณะที่ $BKD_{H/D}$ นั้นได้ผลดีที่สุด 1/11 ส่วน $kFDA_{Jef}$ นั้นได้โมเดลที่มีความซับซ้อนน้อยที่สุดถึง 10/11 คลาส โดยที่สูญเสียความถูกต้องเพียงเล็กน้อย (<3%) เมื่อเทียบกับ $kFDA_1$ โดยเฉลี่ยอีกทั้งยังถูกต้องมากกว่า $BKD_{J/T}$ ในแอกทิวิตีคลาสที่มีความเอกพันธ์สูงสุด 2 คลาส และอยู่ในระดับเดียวกันกับอีก 2 คลาส (<±3%)

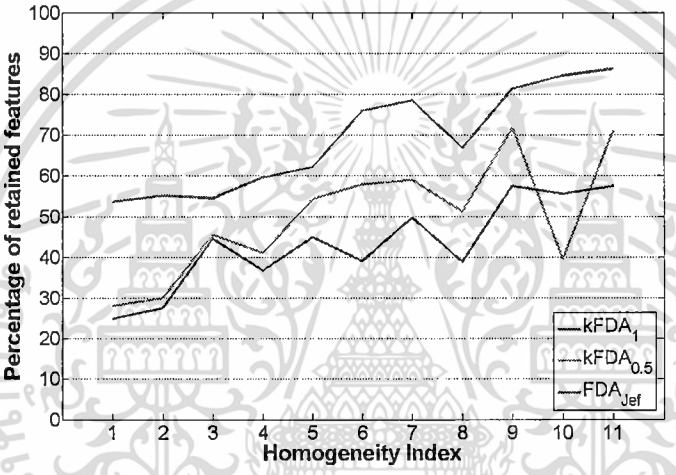
จะเห็นได้ว่า $kFDA_q$ และ $kFDA_{Jef}$ นั้นได้ผลดีที่สุดในข้อมูลที่มีความเอกพันธ์สูงมาก ผลการทดลองของทุกโมเดลได้ถูกสรุปอยู่ในรูปที่ 4.3 และ 4.4 สำหรับความถูกต้องและขนาดของโมเดลตามลำดับ



ภาพที่ 4.3 แสดงค่าความถูกต้อง (ร้อยละของจำนวนของสารประกอบที่ออกฤทธิ์ที่อยู่ใน 5% แรกของจำนวนสารประกอบที่ถูกเรียงลำดับในฐานข้อมูลทั้งหมด) เรียงลำดับตามค่าความเอกพันธ์ของ 11 แอกทิวิตีคลาสจากฐานข้อมูล MDDR สำหรับ $kFDA_q$, $kFDA_{Jef}$และ BKD

ตารางที่ 4.3 แสดงร้อยละของจำนวนของสารประกอบที่ออกฤทธิ์ที่อยู่ใน 5% แรกของจำนวนสารประกอบที่ถูกเรียงลำดับในฐานข้อมูลทั้งหมด โดย kFDA และ BKD

Index	Self-similarity		BKD		$kFDA_1$	$kFDA_{0.5}$	$kFDA_{Jef}$
	Mean	S.D.	HD (%)	J/T (%)	J/T (%)	J/T (%)	J/T (%)
1	0.337	0.105	98.84	99.10	**99.25**	99.23	99.17
					53.63	28.14	**24.96**
2	0.296	0.100	98.77	97.43	**99.27**	99.22	99.13
					55.16	30.11	**27.58**
3	0.226	0.101	94.37	**94.70**	92.89	93.45	92.03
					54.53	45.47	**44.53**
4	0.212	0.098	**94.04**	94.02	93.77	92.74	92.02
					59.63	41.11	**36.79**
5	0.179	0.082	91.86	**93.70**	90.74	90.38	88.64
					62.16	54.32	**44.88**
6	0.175	0.090	90.19	**93.88**	91.32	90.61	88.83
					76.00	57.87	**39.07**
7	0.173	0.089	74.25	**77.97**	70.25	68.34	67.32
					78.50	59.00	**49.75**
8	0.166	0.086	86.77	**88.28**	83.98	82.10	79.46
					66.87	51.21	**38.80**
9	0.153	0.092	69.47	**73.62**	64.89	65.08	62.85
					81.39	71.67	**57.50**
10	0.141	0.103	78.92	**81.23**	71.15	62.95	66.78
					84.57	39.57	**55.65**
11	0.130	0.073	68.43	**76.26**	65.52	63.11	60.49
					86.25	71.09	**57.50**
Average			85.99	**88.20**	83.91	82.47	81.52
					68.97	49.96	**43.36**



ภาพที่ 4.4 แสดงค่าความซับซ้อนของโมเดล (ร้อยละของจำนวนฟีเจอร์ที่ใช้) เรียงลำดับตามค่าความเอกพันธ์ของ 11 แอกทิวิตีคลาสจากฐานข้อมูล MDDR สำหรับ $kFDA_q$ และ $kFDA_{Jef}$

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

โครงงานวิจัยนี้ได้นำเสนออัลกอริทึมใหม่ที่เรียกว่าการวิเคราะห์การจำแนกของฟิชเชอร์แบบเคอเนลที่ไม่ซับซ้อนด้วยเจฟฟรีไพรเออร์ โดยอัลกอริทึมนี้ไม่ต้องการเรกิวราไรซ์พารามิเตอร์ ดังนั้นมีเพียงเคอเนลพารามิเตอร์เพียงพารามิเตอร์เดียวเท่านั้นที่ต้องการการเลือกเพื่อให้ได้โมเดลที่ดีที่สุด สิ่งนี้ทำให้ลดต้นทุนในการสร้างโมเดลได้อย่างมาก อัลกอริทึมที่นำเสนอนั้นมีประสิทธิภาพอยู่ในระดับเดียวกับอัลกอริทึมชั้นนำอื่นๆ ทั่วไป ซึ่งได้ทดสอบบนข้อมูลต่างๆ

อัลกอริทึมที่นำเสนอได้ทดสอบกับข้อมูลทางเคมีสารสนเทศ ผลการทดลองของอัลกอริทึมการวิเคราะห์การจำแนกของฟิชเชอร์แบบเชิงเส้นแบบไม่ซับซ้อนนั้นมีประสิทธิภาพต่ำกว่า BKD ที่นำเสนอโดย [19] อย่างไรก็ตาม ก็สามารถปรับปรุงประสิทธิภาพได้โดยการใช้เทคนิคเคอเนล แม้ว่าการใช้เคอเนลสามารถเพิ่มประสิทธิภาพได้นั้น ก็ยังได้ผลไม่ดีเท่ากับ $BKD_{\mu T}$ ในข้อมูลที่มีความเอกพันธ์น้อยมาก เนื่องมาจากการทำให้โมเดลซับซ้อนน้อยนั้นอาจจะทำให้เกิดการสูญเสียข้อมูลที่สำคัญไปได้ โดยเฉพาะในข้อมูลที่มีความเอกพันธ์น้อย แต่การทำให้โมเดลมีความซับซ้อนน้อยนั้นก็ยังมีข้อดี โดยเฉพาะข้อมูลในเชิงพาณิชย์ที่มีจำนวนข้อมูลมาก $O(10^6)$ ที่สามารถเพิ่มความเร็วในการเรียกค้นได้ ความแตกต่างของข้อมูลจาก FIRST IDA Repository และ MDDR ที่เห็นได้ชัดคือ ปัญหาข้อมูลที่เล็กเกินไปในการเรียนรู้ (Small-sample-size Problem) ซึ่งส่งผลให้ประสิทธิภาพในการเรียนรู้ตกลง จากผลของโครงการนี้สรุปได้ว่า BKD ยังเป็นอัลกอริทึมที่มีประสิทธิภาพมาก และทนกับข้อมูลที่มีสัญญาณรบกวน เช่น ลายนิ้วมือ เป็นต้น

บรรณานุกรม

[1] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2nd ed. New York: Wiley-Interscience, 2001.

[2] S. Mika, G. R¨atsch, J. Weston, B. Sch¨olkopf, and K.-R. M¨uller, "Fisher discriminant analysis with kernels," in Neural Networks for Signal Processing IX, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. IEEE Press, 1999, pp. 41–48.

[3] R. F. Harrison and K. Pasupa, "A simple iterative algorithm for parsimonious binary kernel Fisher discrimination," Pattern Analysis Applications, vol. 13, no. 1, pp. 15–22, 2010.

[4] M. A. Figueiredo, "Adaptive sparseness for supervised learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, pp. 1150–1159, 2003.

[5] B. Sch¨olkopf and A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge: MIT Press, 2002.

[6] S. Mika, G. R¨atsch, J. Weston, B. Sch¨olkopf, A. Smola, and K. M¨uller, "Constructing descriptive and discriminative nonlinear features: Rayleigh Coefficients in kernel feature spaces," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, pp. 623–628, 2003.

[7] D. R. Hunter and R. Li, "Variable selection using MM algorithms," The Annals of Statistics, vol. 33, pp. 1617–1642, 2005.

[8] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemeink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, pp. 957–968, 2005.

[9] G. R¨atsch, "FIRST IDA Benchmark Repository," 2001. [Online]. Available: http://theoval.cmp.uea.ac.uk/matlab/benchmarks/benchmarks.mat

[10] MathWorks, "Matlab version 7.10," 2010, the MathWorks Inc., Natick, MA.

[11] MDL Information Systems Inc., "The MDL drug data report database," 2006. [Online]. Available: http://www.mdli.com

[12] Scitegic Inc., "ECFP 4 fingerprints," 2006, [Online]. Available: http://www.scitegic.com

[13] J. Zhu and T. Hastie, "Kernel logistic regression and the import vector machine," Journal of Computational and Graphical Statistic, vol. 14, no. 1, pp. 185–205, 2005.

[14] Y. Xu, J.-Y. Yang, and J. Yang, "A reformative kernel fisher discriminant analysis," Pattern Recognition, vol. 37, no. 6, pp. 1299–1302, 2004.

[15] Y. Xu, D. Zhang, Z. Jin, M. Li, and J.-Y. Yang, "A fast kernel-based nonlinear discriminant analysis for multi-class problems," Pattern Recognition, vol. 39, no. 6, pp. 1026–1033, 2006.

[16] S. S. Keerthi, K. Duan, S. K. Shevade, and A. N. Poo, "A fast dual algorithm for kernel logistic regression," Machine Learning, vol. 61, no. 1–3, pp. 151–165, 2005.

[17] S. Mika, G. R¨atsch, and M. K.-R., "A mathematical programming approach to the kernel fisher algorithm," in Advances in Neural Information Processing Systems, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., vol. 13. London, England: MIT Press, 2001, pp. 591–597.

[18] Y. Sun, S. Todorovic, and J. Li, "Increasing the robustness of boosting algorithms within the linear-programming framework," Journal of VLSI Signal Processing, vol. 48, no. 1–2, pp. 5–20, 2007.

[19] Y. Washizawa and Y. Yamashita, "Kernel projection classifiers with suppressing features of other classes," Neural Computation, vol. 18, no. 8, pp. 1932–1950, 2006.

[20] B. Chen, R. F. Harrison, K. Pasupa, P. Willett, D. J. Wilton, D. J. Wood, and X. Q. Lewell, "Virtual screening using binary kernel discrimination: Effect of noisy training data and the optimization of performance." Journal of Chemical Information and Modeling, vol. 46, no. 2, pp. 478–486, 2006.

[21] M. Brena and V. Batagelj, "The metric index," Croatica Chemica Acta, vol. 790, pp. 399–410, 2006.

[22] C. Bahlmann, B. Haasdonk, and H. Burkhardt, "On-line handwriting recognition with support vector machines: A kernel approach," in Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02). IEEE Computer Society, 2002, pp. 49–54.

[23] Q. Yong and Y. Jie, "Modified kernel functions by geodesic distance," EURASIP Journal on Applied Signal Processing, vol. 16, pp. 2515–2521, 2004.

[24] K. Pasupa, R. F. Harrison, and P. Willett, "Parsimonious kernel Fisher discrimination," in Pattern Recognition and Image Analysis, ser. Lecture Notes in Computer Science, vol. 4477. Springer, 2007, pp. 531–538

ภาคผนวก

ภาคผนวก ก

บทความวิจัยในการประชุมวิชาการนานาชาติ ICCAIS'2012

Sparse Fisher Discriminant Analysis with Jeffrey's Hyperprior

Kitsuchart Pasupa

Abstract—**The penalty function requires a choice of regularization parameter which controls the degree of parsimony in sparse kernel classifier. This involves an extra parameter apart from kernel parameter in the optimization which must be found via, e.g. cross-validation. This paper introduces a new parsimonious binary kernel Fisher discriminant analysis which does not require a regularization parameter. This can be done by using a Jeffrey's noninformative hyperprior. A Jeffrey's noninformative hyperprior is parameter-free and is adopted through a hierarchical-Bayes interpretation of the Laplacian prior distribution. This leads to a non-requirement of the regularization parameter. The proposed algorithm is compared with other machine learning methods on substantial benchmarks. Moreover, it is also compared with the leading machine learning in virtual screening application. It is found to be less accurate but it is still comparable in a number of cases.**

I. Introduction

Fisher discriminant analysis (FDA) seeks a linear projection that maximizes the separation between data belonging to two classes while minimizing the separation between those of the same class. Its properties are under certain circumstances prove optimal [1]. In real world data, when the data samples are significantly non-Gaussian and have different covariance matrices, the performance of FDA might be degraded dramatically. It might fail to find the optimal projection direction for classification because, for optimality, FDA has an assumption that the input patterns have equal covariance matrix and are Gaussian. The performance of FDA can be improved by kernel FDA [2]. Recently, [3] introduced a new algorithm into kernel machine family the so-called "parsimonious binary kernel Fisher discriminant analysis", a kernel-based extension of FDA. The algorithm utilizes a connection between Fisher discriminant analysis and least squares problem. The complexity is controlled by L_q penalty function where $0 < q \leq 1$. This penalty function is well-known to have a sparsity-inducing property, and it leads to a non-smooth formulation. The problem is solved by the majorize-minimize principle, which gives a very simple iterative algorithm.

However, the above penalty function requires a choice of hyperparameter or regularization parameter which controls the degree of parsimony. This involves an extra parameter apart from kernel parameter in the optimization that must be found via, e.g. cross- validation. Alternatively, [4] adopted a Jeffrey's noninformative hyperprior through a hierarchical-Bayes interpretation of the Laplacian prior. This approach is more convenient because it does not require the hyperparameter but solving a slightly different problem.

K. Pasupa is with the Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, 1 Soi Chalongkrung 1, Ladkrabang, Ladkrabang, Bangkok 10520, Thailand. kitsuchart@it.kmitl.ac.th

The motivations of this work lie in the area of Chemoinformatics in particular in virtual screening (VS). The ability to rank molecules according to their effectiveness in some domains (e.g. pesticide and drug discovery) is important, owing to the cost of synthesising and testing chemical compounds. VS seeks to do this computationally with potential savings of millions of pounds and large profits associated with reduced time to market. Traditional methods of drug discovery requires chemists to synthesise chemical compounds in test tubes. This is time consuming and expensive. Thus many pharmaceutical companies are now using VS to carry out the early identification of potential drugs in order to reduce the size of the samples to be synthesised. This can be performed using a computer.

Machine learning methods are becoming popular in this domain. In VS tasks, the problems arise when 2D fingerprints are used as descriptors, i.e. high-dimensional, small-sample-size problem. These problems are one of the grand challenges in the field of data mining and knowledge discovery for engineer, computer scientist, and statistician. Developed algorithms are needed to be robust and efficient. Moreover, they should handle very large set of high-dimensional data because of an unprecedented level of complexity in data model. Thus, complexity control is particular relevant and indispensable in VS tasks.

In this paper, we extend the work from [3]. The algorithm uses a connection between FDA and least squares problem but the complexity is controlled by L_1 penalty function which is interpreted by a hierarchical-Bayes with a Jeffrey's noninformative hyperprior. Therefore, there is no regularization parameter in the algorithm.

The paper is organized as follows. Section II outlines the FDA, kernel FDA and introduces our proposed algorithm. Section III explains experimental framework which includes application to VS and compares the proposed algorithm with other leading machine learning methods in the area.

II. Methodologies

A. Fisher discriminant analysis

Consider the matrix of m-dimensional sample vectors $\mathrm{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^\mathrm{T}$ comprising two groups, $\mathscr{G}_i$ of size N_i, $i = 1, 2$, represented by the partition, $\begin{bmatrix} \mathrm{X}_1 & \mathrm{X}_2 \end{bmatrix}^\mathrm{T}$. Membership of $\mathscr{G}_1$ is denoted by $\hat{y} = +N/N_1$ and of $\mathscr{G}_2$ by $\hat{y} = -N/N_2$. Fisher discriminant coefficients are given by $\boldsymbol{w}$ which maximises the following objective function:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\mathrm{T} \mathrm{S}_\mathrm{B} \boldsymbol{w}}{\boldsymbol{w}^\mathrm{T} \mathrm{S}_\mathrm{W} \boldsymbol{w}} \tag{1}$$

S_B is the between-class scatter matrix which is calculated from, $S_B = (m_1 - m_2)(m_1 - m_2)^T$. The within-class scatter matrix is defined by $S_W = \sum_{i=1,2} \sum_{x \in \mathcal{G}_i} (x - m_i)(x - m_i)^T$ where $m_i = \frac{1}{N_i} \sum_{x \in \mathcal{G}_i} x$.

A global optimal solution of (1), the so-called "Rayleigh quotient", can easily be found by solving an eigenvalue problem [1]. Hence, this gives

$$w = (S_W)^{-1}(m_1 - m_2) \tag{2}$$

It is well-known that FDA has strong connections to the least squares (LS) approach for classification [1]. Consider a linear function including a bias term, b, $f(x) = w^T x + b$. The LS approach is to minimise the sum of square error (SSE),

$$\begin{aligned} \mathrm{SSE}(b, w) &= \sum_{i=1}^{N} (\hat{y}_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} (\hat{y}_i - w^T x - b)^2 \end{aligned} \tag{3}$$

This can be represented in matrix form as,

$$\arg\min_{(b,w)} \left\| \begin{bmatrix} \frac{N}{N_1} 1_{N_1} \\ -\frac{N}{N_2} 1_{N_2} \end{bmatrix} - \begin{bmatrix} 1_{N_1} & X_1 \\ 1_{N_2} & X_2 \end{bmatrix} \begin{bmatrix} b \\ w \end{bmatrix} \right\|_2^2 \tag{4}$$

The solution of a LS problem of $\| \hat{y} - \tilde{X}\omega \|_2^2$ can be computed by using pseudo-inverse $\tilde{X}^\dagger$ of $\tilde{X}$,

$$\omega = \tilde{X}^\dagger \hat{y} \tag{5}$$

where $\tilde{X}^\dagger = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T$, $\omega = [b \quad w]$, $\tilde{X} = [1_N \quad X]$, and 1_N denotes an N-vector of ones. Hence, $(\tilde{X}^T \tilde{X})\omega = \tilde{X}^T \hat{y}$. Then substituting this equation by using the definition of the sample means and within-class scatter. Hence, the solution of the LS problem is in the same direction as the coefficients of Fisher discriminant in (2).

However, the linearity of the approach is normally insufficient to allow the required level of performance in real-world applications. While explicit expansion of data in basis functions can resolve this problem for low dimensional data, the combinatorial increase in the number of coefficients to be estimated may make this impractical [3]. Kernel machines can address this problem via "kernel trick" [5] and a number solutions have been provided e.g. [6], [3].

B. Kernel Fisher Discriminant Analysis

In this paper, we use the kernel Fisher discriminant analysis (KFDA) proposed by [3]. The algorithm uses a connection between FDA and LS problem. The complexity is controlled by Lq penalty function where $0 < q \leq 1$. This penalty function is well-known to have a sparsity-inducing property, and it leads to a non-smooth formulation. The problem is solved by the majorize-minimize principle, which gives a very simple iterative algorithm.

The derivation of a very simple, Newton-like algorithm for the penalized minimum likelihood estimation of the KFDA coefficients is outlined, c.f. [7]. The log-likelihood function, $\ell(\omega)$, is written as the sum of two functions, $\ell_e(\omega) = \frac{1}{2}\|\hat{y} - \tilde{K}\omega\|_2^2$, and $\ell_p(\omega) = \rho N \|\omega\|_q^q$, where $\hat{y} = \begin{bmatrix} \frac{N}{N_1} 1_{N_1} \\ -\frac{N}{N_2} 1_{N_2} \end{bmatrix} \in \mathbb{R}^N$ and $\tilde{K} = [1_N \quad K] \in \mathbb{R}^{N \times (N+1)}$, giving:

$$\ell(\omega) = \frac{1}{2}\|\hat{y} - \tilde{K}\omega\|_2^2 + \rho N \|\omega\|_q^q \tag{6}$$

K_i denotes the Gram matrix associated with the kernel, $k(.,.)$, i.e. $k_{jl} = k(x_j, x_l)$, $j, l = 1, 2, \ldots, N$, $x_l \in \mathcal{G}_i$. ρ is a regularization parameter. The solution, ω, denotes the coefficients (b, α) of a linear discriminant function in the feature space associated with $k(.,.)$ hence non-linear discrimination in the original data space.

Solving the above equation using the majorize-minimize principle leads to the following iterative equation:

$$\omega(n+1) = \left(\tilde{K}^T \tilde{K} + \rho N q \mathrm{B}(\omega(n)) \right)^{-1} \tilde{K}^T \hat{y} \tag{7}$$

where $\mathrm{B}(\omega(n)) = \mathrm{diag}\left\{ |\omega_i(n)|^{q-2} \right\}$. A problem arises when the elements of $\omega(n)$ approach zero [3]. However, this can be avoided the difficulty by re-writing $\mathrm{B}(\omega(n)) = \Psi_n^{-2}$ with $\Psi_n = \mathrm{diag}\left\{ |\omega_i(n)|^{\frac{2-q}{2}} \right\}$ [8]. This gives the following iterative equation:

$$\omega(n+1) = \Psi_n \left(\Psi_n \tilde{K}^T \tilde{K} \Psi_n + \rho N q I_{N+1} \right)^{-1} \Psi_n \tilde{K}^T \hat{y} \tag{8}$$

where $\omega(0) \neq \mathbf{0}$.

In *primal*[1] formulations, when the dimension of the input vector is greater than the number of samples ($n > N$), the identity for matrices M_1, M_2, and M_3 can be exploited with M_1, M_3 positive definite and invertible.

$$\begin{aligned} &\left(M_1^{-1} + M_2^T M_3^{-1} M_2 \right)^{-1} M_2^T M_3^{-1} = \\ &\qquad M_1 M_2^T \left(M_2 M_1 M_2^T + M_3 \right)^{-1} \end{aligned} \tag{9}$$

Then making the substitutions as follow: $M_1^{-1} = \rho \mathrm{B}(\omega(n))$, $M_2 = \tilde{X}$, and $M_3^{-1} = N q I_N$, gives the following iteration

$$\begin{aligned} &\omega(n+1) = \\ &\quad \mathrm{B}^{-1}(\omega(n)) \tilde{X}^T \left(\tilde{X} \mathrm{B}^{-1}(\omega(n)) \tilde{X}^T + \rho N q I_N \right)^{-1} \hat{y} \end{aligned} \tag{10}$$

When the relative change in the norm of the coefficient vectors is less than some threshold, $\varepsilon \ll 1$, convergence is declared. Then, a coefficient is deemed to equal zero if $\frac{|\omega|}{\arg\max |\omega_i|} < \eta$ when $\eta \ll 1$. The decision threshold is defined as the mid-value of target labels,

$$c = \frac{1}{2} \left(\frac{N}{N_1} - \frac{N}{N_2} \right) \tag{11}$$

The classification rule is as follows: decide $\mathcal{G}_1$ if $b + \sum_{i=1}^{N} \alpha_i k(x, x_i) \geq c$; otherwise decide $\mathcal{G}_2$.

The resulting classifiers kFDA_q and FDA_q denote for the kernel and the linear version, respectively.

[1] We now consider $\tilde{X}$ instead of $\tilde{K}$ because $\tilde{K}$ is used for *dual* formulations.

C. Removing of the Hyperparameter

As described in section I, in order to remove the regularization parameter, a Jeffreys' prior is used in the penalty function. It is adopted through a hierarchical-Bayes interpretation of the Laplacian ($q = 1$) prior [4].

Consider the Laplacian prior,

$$p(\omega|\rho) = \prod_{i=1}^{k} \frac{\rho}{2} \exp\{-\rho|\omega_i|\} = \left\{\frac{\rho}{2}\right\}^k \exp\{-\rho\|\omega\|_1^1\} \quad (12)$$

As this prior is non-differentiable, hence, a hierarchical-Bayes interpretation is introduced to solved this problem. Considering a zero-mean Gaussian prior with variance τ_i

$$p(\omega_i|\tau_i) = \mathcal{N}(\omega_i|0, \tau_i) \quad (13)$$

and an exponential hyperprior,

$$p(\tau_i|\rho) = \frac{\rho}{2} \exp\left\{-\frac{\rho}{2}\tau_i\right\} \quad (14)$$

Integrating both functions with respect to τ_i gives,

$$p(\omega_i|\rho) = \int_0^\infty p(\omega_i|\tau_i)p(\tau_i|\rho)d\tau_i = \frac{\sqrt{\rho}}{2} \exp\{-\sqrt{\rho}|\omega_i|\} \quad (15)$$

The above equation shows that the Laplacian prior is a combination of a two-level hierarchical-Bayes model: a zero-mean Gaussian priors with variance τ and an exponential hyperprior. However, it involves a hyperparameter, ρ, hence, [4] replaced the exponential hyperprior by a Jeffrey's noninformative hyperprior for τ_i,

$$p(\tau_i) \propto \frac{1}{\tau_i} \quad (16)$$

Integrating the zero-mean Gaussian prior with variance τ_i and the Jeffrey's noninformative hyperprior with respect to τ_i gives,

$$p(\omega_i) = \int_0^\infty p(\omega_i|\tau_i)p(\tau_i)d\tau_i = |\omega_i|^{-1} \quad (17)$$

[4] carried out this implementation through an EM algorithm. This gives the following two-step iteration which is similar to the iteration in (8),

$$\sigma^2(n+1) = \frac{1}{N}\|\hat{y} - \tilde{K}\omega\|_2^2 \quad (18)$$

and

$$\omega(n+1) = \Psi_n\left(\Psi_n\tilde{K}^T\tilde{K}\Psi_n + \sigma^2(n+1)I_{N+1}\right)^{-1}\Psi_n\tilde{K}^T\hat{y} \quad (19)$$

where $\omega(0) \neq 0$, $\Psi_n = \text{diag}\{|\omega_i(n)|\}$.

In primal formulations, when $m > N$, the iteration becomes,

$$\sigma^2(n+1) = \frac{1}{N}\|\hat{y} - \tilde{X}\omega\|_2^2 \quad (20)$$

and

$$\omega(n+1) = B^{-1}(\omega(n))\tilde{X}^T\left(\tilde{X}B^{-1}(\omega(n))\tilde{X}^T + \sigma^2(n+1)I_N\right)^{-1}\hat{y} \quad (21)$$

The resulting classifiers kFDA_{Jef} and FDA_{Jef} denote for the kernel and the linear version, respectively.

III. Experiments

A. FIRST IDA Repository Database

To evaluate the performance of kFDA_{Jef} extensive experimentation has been carried out on 13 datasets of the FIRST IDA repository database [9]. The methodology outlined in [3] was followed to allow direct comparisons with a number of techniques used in the articles revealed in the citation search and also kFDA_q. These results are reported in [3].

In the experiments, a radial basis function is used as a kernel function. Five-fold cross-validation is applied to obtain the optimal model (kernel parameter). As mentioned in the previous section, using Jeffrey's prior in the penalty function removes the requirement for a search for the regularization parameter. We examined classifiers selected based on minimum misclassification rate (MCR) from amongst the first five realisations. Each is then applied to all 100 test partitions except "Splice" and "Image" which are split into 20 partitions. All experiments are carried out using the Matlab environment [10].

From Table I, selecting on minimum MCR, 5/13 cases achieve the best MCR by kFDA_{Jef} (equal to kFDA_q for $q \in \{1, 0.5\}$ in "Titanic"). kFDA_{Jef} exhibits best sparsity in 3 out of 13 cases. Its sparsity on average across all domains is worse than $\text{kFDA}_{0.5}$ but the situation is reversed when considering accuracy. kFDA_1 is still the most accurate on average across all datasets.

B. Application to Virtual Screening

Here 11 different activity classes from the MDL Drug Data Report (MDDR) database are used [18]. The selected activity classes were selected to reflect typical drug discovery projects for pharmaceutical companies as shown in tables II. MDDR database is a set of 102,514 known drugs and biologically relevant molecules collected from patent literature, journals, meetings and congresses. The MDDR database represented by 1,024-dimensional (ECFP_4) fingerprint [19]. We examine two situations: linear and kernel Fisher discrimination. The experiment was run five times with different random data splits. The number of training samples is equal to 20% of the number of active molecules in the database. New data are ranked on the predicted output value from most positive (most likely to be active) to most negative. The predicted output value is equal to distance from decision boundary. Tuning parameters are identified by five-fold cross-validation on the basis of *sum of active rank position*. The experimental results are shown in table III, and IV for linear and kernel Fisher discrimination, respectively. It is usual in chemoinformatics applications to report the percentage of the maximum possible number

TABLE I

FIRST IDA REPOSITORY DATABASE: COMPARISON OF MEAN MISCLASSIFICATION RATE, ITS STANDARD DEVIATION AND SPARSITY (BELOW) FOR THE PROPOSED ALGORITHM $kFDA_{Jef}$, $kFDA_q$, AND THE BEST PUBLISHED ALGORITHM: IMPORT VECTOR MACHINE [11] (★), CONVENTIONAL KFDA (△) [12], REFORMATIVE KFDA (▲) [12], NAIVE KERNEL-BASED NONLINEAR METHOD [13] (▽), FAST KERNEL-BASED NONLINEAR METHOD [13] (▼), KERNEL LOGISTIC REGRESSION [14] (□), SPARSE KFDA WITH LINEAR LOSS [15] (■), LINEAR PROGRAMMING ADABOOST [16] (◇), AND SUPPRESSED KERNEL SAMPLE SPACE PROJECTION [17] (◆). EACH NUMBER BELOW DATABASE'S NAME IS SAMPLE SIZE. BOLD TYPE – BEST PERFORMANCE/SPARSITY.

Database	Published Best	$kFDA_1$	$kFDA_{0.5}$	$kFDA_{Jef}$
Banana	10.34 ± 0.46	9.74 ± 0.10	9.63 ± 0.10	11.76±0.10
400	5.25 ± 1.75 ★	14.00	4.50	5.00
B. Cancer	22.70 ± 4.40	21.26 ± 3.70	20.55 ± 3.96	**20.09±3.81**
200	100.00 △	13.00	3.50	10.50
Diabetes	22.10 ± 1.90	21.57 ± 1.41	21.58 ± 1.63	**21.18±1.55**
468	100.00 △	5.13	1.28	2.14
German	21.30 ± 2.10	21.19 ± 1.81	**23.51** ± 2.03	**20.77±1.70**
700	100.00 △	8.86	**1.14**	5.14
Heart	**10.80 ± 2.60**	14.56 ± 3.04	14.84 ± 3.31	13.75±2.54
170	16.00 ▲	7.06	2.35	5.29
Image	1.78 ± N/A	**1.56 ± 0.46**	1.98 ± 0.32	2.52 ±0.77
1300	100.00 □	24.77	**15.46**	38.15
Ringnorm	1.50 ± 0.10	**1.42 ± 0.04**	1.48 ± 0.03	1.74 ±0.04
400	6.00 ■	**4.75**	**2.75**	**22.25**
S. Flare	31.60 ± 1.90	32.98±1.74	31.64 ± 1.85	**30.95±1.87**
666	100.00 ▽	27.03	**2.58**	10.45
Splice	9.30 ± 0.70	7.07 ±0.71	**7.02 ±0.83**	7.72 ±0.64
1000	100.00 ◇	85.00	75.90	**51.70**
Thyroid	1.40 ± 0.90	**0.99 ±0.90**	1.05 ± 0.93	1.83 ±1.15
140	16.40 ▼	22.86	12.86	6.43
Titanic	21.70 ± 0.30	**21.10±0.23**	**21.10±0.23**	**21.10±0.23**
150	100.00 ▽	64.67	4.67	4.00
Twonorm	**2.30 ± 0.1**	2.36 ±0.04	2.58 ± 0.04	2.60 ±0.04
400	100.00 ◆	7.75	**1.25**	4.25
Waveform	9.30 ± 0.4	9.26 ±0.13	9.97 ± 0.13	9.56 ±0.13
400	100.00 ◇	7.00	3.00	7.25
Average	12.78	12.70	12.84	12.74
	72.59	22.45	10.10	13.27

TABLE II

THE 11 ACTIVITY CLASSES FROM MDDR DATABASE.

Activity Class	Index	No. of Actives	No. of Training Samples
Renin Inhibitors	1	1130	226
Angiotensin II AT1 Antagonists	2	943	190
HIV Protease Inhibitor	3	750	150
Thrombin Inhibitor	4	803	162
Substance P Antagonists	5	1246	250
5HT3 Antagonists	6	752	150
D2 Antagonists	7	395	80
5HT1A Agonists	8	827	166
5HT Reuptake Inhibitors	9	359	72
Protein Kinase C Inhibitor	10	453	92
Cyclo-oxygenase Inhibitor	11	636	128

TABLE III

COMPARISON OF MAXIMUM PERCENTAGE ACTIVES RETRIEVED IN TOP 5% OF RANKED DATABASE USING LINEAR FISHER DISCRIMINATION.

Index	Self-Similarity		FDA	FDA_2	FDA_1	$FDA_{0.5}$	FDA_{Jef}
	Mean	S.D.	(%)	(%)	(%)	(%)	(%)
1	0.337	0.105	98.88	**99.13**	98.29	96.01	96.72
				88.87	7.15	**1.50**	11.07
2	0.269	0.100	98.47	**98.84**	97.03	93.61	94.48
				89.45	11.41	**1.80**	10.37
3	0.226	0.101	91.67	**92.41**	82.76	75.56	72.39
				87.95	9.43	**2.42**	9.28
4	0.212	0.098	91.55	**93.38**	89.03	79.47	72.96
				88.79	5.63	**1.74**	10.14
5	0.179	0.082	85.51	**88.60**	78.93	70.78	58.79
				92.99	9.14	**4.71**	17.97
6	0.175	0.090	84.67	**86.53**	70.90	59.03	51.37
				87.79	6.76	**2.38**	9.59
7	0.173	0.089	68.68	**68.00**	48.90	34.37	42.87
				78.59	2.54	**0.86**	4.71
8	0.166	0.086	75.56	**79.81**	67.90	54.46	46.75
				88.14	10.35	**2.66**	10.84
9	0.153	0.092	61.49	**63.22**	41.30	35.05	34.67
				75.00	3.69	**1.07**	4.36
10	0.141	0.103	71.40	**71.79**	51.89	37.30	36.71
				81.91	7.03	**3.13**	6.17
11	0.130	0.073	55.91	**57.62**	37.06	21.05	21.68
				86.50	7.91	**2.83**	8.71
Average			80.34	81.76	69.45	59.70	57.22
				86.00	7.37	**2.28**	9.38

of active compounds ranked in the top 5% of the ranked database along with the percentage of retained features (below). The results are compared with the current leading machine learning, the modification of BKD [20]. In both tables, the mean self-similarity provides a measure of the homogeneity of each of the activity class. It is also a useful way to compare design spreads and coverage.

1) Linear Fisher Discrimination: Results from FDA, FDA_q, and FDA_{Jef} are presented in table III. FDA is a leading classifier in 1/11 cases but delivers no sparsity while FDA_2 displays best accuracy in the other cases with sparse model at 86.00% on average[2]. However, reducing q to 1 leads to a dramatic reduction in number of retained samples (NRS) which is equal to 7.37% on average across all cases, but with a small loss of accuracy ($< \pm 3\%$) in the first two homogeneous classes (but an excessive loss of accuracy in the other 9/11 cases). Overall, the average accuracy is dramatically reduce from 81.76% to 69.45%. Setting q equal to 0.5 leads to a simpler model but again with too much loss of accuracy in 10/11 cases.

Using FDA_{Jef} gives 9.38% of NRS and 57.22% of accuracy on average which is worse than using $FDA_q, q \in \{1, 0.5\}$. However, the first two highest homogeneity classes are in line with other proposed classifiers.

These results are illustrated in Fig. 1 and Fig. 2 for accuracy and sparsity, respectively. To summarise, the proposed algorithm is competitive for the classes that are most homogeneous. Overall, it is still behind the modification of BKD [20]. This can be improved by introducing the "kernel trick" in order to accommodate more complex discriminant functions.

2) Kernel Fisher Discrimination: A binomial distribution is used, $k_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \lambda^{m-d(\boldsymbol{x}_i,\boldsymbol{x}_j)}(1-\lambda)^{d(\boldsymbol{x}_i,\boldsymbol{x}_j)}$ where $\lambda \in (0.5, 1)$ denotes the "width" and $m = 1024$. $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a (metric) measure of the degree of dis-similarity between molecules i and j. This kernel is of radial basis function type, see [5, p. 46]. In [20] the Jaccard/Tanimoto (J/T) distance was found to offer substantial gains over the conventional

[2]A coefficient is deemed to be zero if its magnitude, relative to the largest, is less than η.

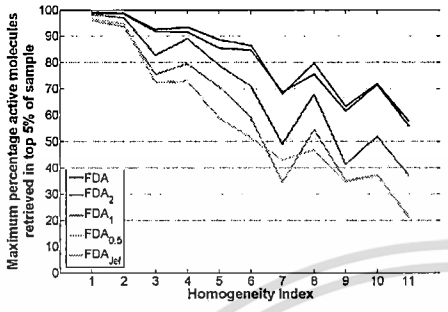


Fig. 1. Accuracy (maximum percentage active molecules retrieved in first 5% of database) as a function of homogeneity index for 11 activity classes from the MDDR database for FDA and its variants.

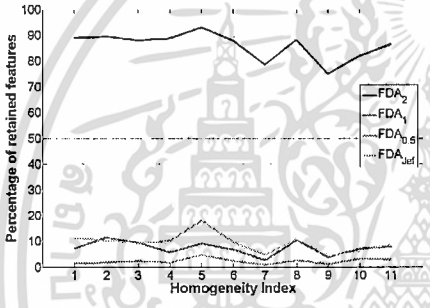


Fig. 2. Sparsity (percentage NRS) as a function of homogeneity index for 11 activity classes from the MDDR database for FDA.

Hamming distance (HD) when used in BKD. It should be noted that J/T distance is a metric for binary vectors [21] so its use with the binomial kernel leads to a valid kernel. There have been many attempts in the past to substitute alternative distances in the kernel function for machine learning methods e.g. [22], [23], etc. Experiments show that this is also the case for $kFDA_q$ so only results using this function are reported. Results from $kFDA_q$, $kFDA_{Jef}$, and from BKD are presented. It should be noted that results from $kFDA_q$ and from BKD are previously reported in [24]. It is still clear that $BKD_{J/T}$ is the leading contender in eight out of 11 cases but delivers no sparsity, while BKD_{HD} is most accurate in one. $kFDA_{Jef}$ displays best sparsity in 10/11 cases, again, with a small loss of accuracy on average (comparing to $kFDA_1$). It is better than $BKD_{J/T}$ in the two most homogeneous cases and is comparable in two others ($< \pm 3\%$).

It is worth noting that $kFDA_q$ and $kFDA_{Jef}$ deliver their best accuracy in the classes that are most homogeneous. These results are illustrated in Fig. 3 and Fig. 4 that show the relative performance and sparsity of the candidate methods,

TABLE IV

COMPARISON OF MAXIMUM PERCENTAGE ACTIVES RETRIEVED IN TOP 5% OF SAMPLE USING KERNEL FISHER DISCRIMINATION.

Index	Self-Similarity		BKD		$kFDA_1$	$kFDA_{0.5}$	$kFDA_{Jef}$
	Mean	S.D.	HD (%)	J/T (%)	J/T (%)	J/T (%)	J/T (%)
1	0.337	0.105	98.84	99.10	**99.25**	**99.23**	**99.17**
					53.63	**28.14**	**24.96**
2	0.269	0.100	98.77	97.43	**99.27**	**99.22**	**99.13**
					55.16	**30.11**	**27.58**
3	0.226	0.101	94.37	94.70	92.89	93.45	92.03
					54.53	45.47	**44.53**
4	0.212	0.098	94.04	94.02	93.77	92.74	92.02
					59.63	41.11	**36.79**
5	0.179	0.082	91.86	93.70	90.74	90.38	88.64
					62.16	54.32	**44.88**
6	0.175	0.090	90.19	93.88	91.32	90.61	88.83
					76.00	57.87	**39.07**
7	0.173	0.089	74.25	77.97	70.25	68.34	67.32
					78.50	59.00	**49.75**
8	0.166	0.086	86.77	88.28	83.98	82.10	79.46
					66.87	51.21	**38.80**
9	0.153	0.092	69.47	73.62	64.89	65.08	62.85
					81.39	71.67	**57.50**
10	0.141	0.103	78.92	81.23	71.15	62.95	66.78
					84.57	39.57	**55.65**
11	0.130	0.073	68.43	76.26	65.52	63.11	60.49
					86.25	71.09	**57.50**
Average			85.99	88.20	83.91	82.47	81.52
					68.97	49.96	43.36

respectively, as functions of homogeneity.

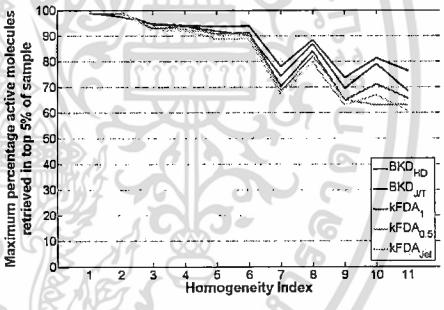


Fig. 3. Accuracy (maximum percentage active molecules retrieved in first 5% of database) as a function of homogeneity index for 11 activity classes from the MDDR database for kFDA and two variants of BKD.

IV. Discussion & Conclusions

A new algorithm the so-called sparse Fisher discriminant analysis with Jeffrey's Hyperprior has been introduced to machine learning technique family. The algorithm does not require the regularization parameter. Hence, there is only one parameter (kernel parameter) which is needed to be search in the optimization. This can reduce the optimization cost. The proposed algorithm is shown to be competitive to other leading machine learning algorithms in a substantial benchmark. The methods linear and kernel Fisher discrimination are applied to a problem in VS. The performance of linear Fisher

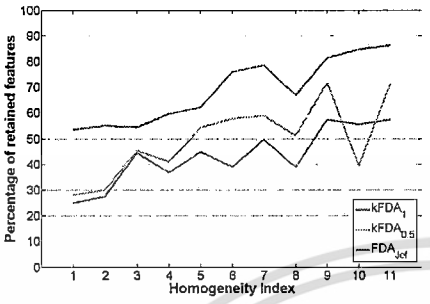


Fig. 4. Sparsity (percentage NRS) as a function of homogeneity index for 11 activity classes from the MDDR database for proposed kFDA.

discrimination is generally worse than the modification of an important recent development in this field, BKD [20]. However, the performance is improved by using kernel trick. However, the results are still worse than the modification of BKD. Particularly, it is worse than the modification of BKD in the heterogeneous classes, this leads to a drop in overall performance. The reason is, a sparse solution might loss some important information in the Gram matrix in heterogeneous data. However, operationally, a sparse solution may still be of value since many commercial databases contain $\mathcal{O}\left(10^6\right)$ samples and speed of recall can be an issue. There is a significant difference between the benchmark and the molecular data the fingerprint samples suffer extremely from the small-sample-size problem and this fact may account for the drop in performance. However, BKD is still the most robust and effective for noisy data e.g. 2D fingerprints.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2001.

[2] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX*, Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. IEEE Press, 1999, pp. 41–48.

[3] R. F. Harrison and K. Pasupa, "A simple iterative algorithm for parsimonious binary kernel Fisher discrimination," *Pattern Analysis Applications*, vol. 13, no. 1, pp. 15–22, 2010.

[4] M. A. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1150–1159, 2003.

[5] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge: MIT Press, 2002.

[6] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, and K. Müller, "Constructing descriptive and discriminative nonlinear features: Rayleigh Coefficients in kernel feature spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 623–628, 2003.

[7] D. R. Hunter and R. Li, "Variable selection using MM algorithms," *The Annals of Statistics*, vol. 33, pp. 1617–1642, 2005.

[8] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemeink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 957–968, 2005.

[9] G. Rätsch, "FIRST IDA Benchmark Repository," 2001. [Online]. Available: http://theoval.cmp.uea.ac.uk/matlab/benchmarks/benchmarks.mat

[10] MathWorks, "Matlab version 7.10," 2010, the MathWorks Inc., Natick, MA.

[11] J. Zhu and T. Hastie, "Kernel logistic regression and the import vector machine," *Journal of Computational and Graphical Statistic*, vol. 14, no. 1, pp. 185–205, 2005.

[12] Y. Xu, J.-Y. Yang, and J. Yang, "A reformative kernel fisher discriminant analysis," *Pattern Recognition*, vol. 37, no. 6, pp. 1299–1302, 2004.

[13] Y. Xu, D. Zhang, Z. Jin, M. Li, and J.-Y. Yang, "A fast kernel-based nonlinear discriminant analysis for multi-class problems," *Pattern Recognition*, vol. 39, no. 6, pp. 1026–1033, 2006.

[14] S. S. Keerthi, K. Duan, S. K. Shevade, and A. N. Poo, "A fast dual algorithm for kernel logistic regression," *Machine Learning*, vol. 61, no. 1–3, pp. 151–165, 2005.

[15] S. Mika, G. Rätsch, and M. K.-R., "A mathematical programming approach to the kernel fisher algorithm," in *Advances in Neural Information Processing Systems*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., vol. 13. London, England: MIT Press, 2001, pp. 591–597.

[16] Y. Sun, S. Todorovic, and J. Li, "Increasing the robustness of boosting algorithms within the linear-programming framework," *Journal of VLSI Signal Processing*, vol. 48, no. 1–2, pp. 5–20, 2007.

[17] Y. Washizawa and Y. Yamashita, "Kernel projection classifiers with suppressing features of other classes," *Neural Computation*, vol. 18, no. 8, pp. 1932–1950, 2006.

[18] MDL Information Systems Inc., "The MDL drug data report database," 2006. [Online]. Available: http://www.mdli.com

[19] Scitegic Inc., "ECFP_4 fingerprints." 2006, world Wide Web, http://www.scitegic.com/. [Online]. Available: http://www.scitegic.com/

[20] B. Chen, R. F. Harrison, K. Pasupa, P. Willett, D. J. Wilton, D. J. Wood, and X. Q. Lewell, "Virtual screening using binary kernel discrimination: Effect of noisy training data and the optimization of performance," *Journal of Chemical Information and Modeling*, vol. 46, no. 2, pp. 478–486, 2006.

[21] M. Brena and V. Batagelj, "The metric index," *Croatica Chemica Acta*, vol. 790, pp. 399–410, 2006.

[22] C. Bahlmann, B. Haasdonk, and H. Burkhardt, "On-line handwriting recognition with support vector machines: A kernel approach," in *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)*. IEEE Computer Society, 2002, pp. 49–54.

[23] Q. Yong and Y. Jie, "Modified kernel functions by geodesic distance," *EURASIP Journal on Applied Signal Processing*, vol. 16, pp. 2515–2521, 2004.

[24] K. Pasupa, R. F. Harrison, and P. Willett, "Parsimonious kernel Fisher discrimination," in *Pattern Recognition and Image Analysis*, ser. Lecture Notes in Computer Science, vol. 4477. Springer, 2007, pp. 531–538.

ข้อมูลประวัติผู้วิจัย

ประวัติส่วนตัว

ชื่อ-สกุล ดร. กิติ์สุชาต พสุภา

ตำแหน่งปัจจุบัน อาจารย์ คณะเทคโนโลยีสารสนเทศ สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ประวัติการศึกษา

ชื่อย่อปริญญา	สาขา	สถาบันที่จบ	ปีที่จบ
Ph.D.	Automatic Control & Systems Engineering	University of Sheffield, UK	2008
M.Sc. (Eng.)	Control System	University of Sheffield, UK	2004
B.Eng.	Electrical Engineering	Sirindhorn International Institute of Technology, Thammasat University	2003
มัธยมศึกษาตอนปลาย	วิทย์-คณิต	โรงเรียนเซนต์คาเบรียล	1999

สาขาวิจัยที่มีความชำนาญพิเศษ

Machine Learning, Artificial Intelligence, Data Mining, Computational Data Modeling, Virtual Screening, Information Retrieval, Ranking Algorithm, Eye Movements

ทุนการศึกษาและทุนวิจัยที่เคยได้รับ

ปี พ.ศ.	ทุนการศึกษาและทุนวิจัย	สถาบันที่ให้
2556	ทุนอุดหนุนการวิจัย, Emotional Speech Synthesis for Visibility Impaired: From-Book-to-Speech (ESSVIBS)	สำนักงานคณะกรรมการวิจัยแห่งชาติ
2556	ทุนวิจัยเงินรายได้, The Development of Online Submission System for Computer Programming Teaching Tool, จำนวน 50,000 บาท	คณะเทคโนโลยีสารสนเทศ, สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
2555	ทุนวิจัยเงินรายได้, The Development of Submission System for Computer Programming Teaching Tool, จำนวน 20,000 บาท	คณะเทคโนโลยีสารสนเทศ, สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
2555	ทุนวิจัยเงินรายได้, Dimensionality Reduction for Data Mining, จำนวน 50,000 บาท, เลขที่ 2555-0206003	คณะเทคโนโลยีสารสนเทศ, สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
2551-2553	Research Fellow, Personal Information Navigator Adapting Through Viewing (PinView), จำนวน 668,558.58 ยูโร, เลขที่ 216529	University of Southampton, EU FP7
2550	Research Associate	University of Sheffield, UK
2547-2549	Alumni Scholarship	University of Sheffield, UK

ผลงานวิจัยที่ได้รับการเผยแพร่

บทความวิชาการ

[1] Pasupa, K. (2012) The Review of Virtual Screening Techniques. KMITL Journal of Information Technology, 1(1), pp. 60-82.

[2] Janiam, P. and Pasupa, K. (2012) The Possibilities of Broadcasting TV Shows in 3D. (In Thai) KMITL Journal of Information Technology, 1(1), pp. 14-23.

บทความวิจัยตีพิมพ์ในวารสารนานาชาติ

[1] Harrison, R. F. and Pasupa, K. (2010) A Simple Iterative Algorithm for Parsimonious Binary Kernel Fisher Discrimination. Pattern Analysis & Applications, 13(1), pp 15-22.

[2] Harrison, R. F. and Pasupa, K. (2009) Sparse Multinomial Kernel Discriminant Analysis (sMKDA). Pattern Recognition. 42(9), pp 1795-1802

[3] Chen, B., Harrison, R. F., Pasupa, K., Willett, P., Wilton, D. J. and Wood, D. J. (2006) Virtual Screening Using Binary Kernel Discrimination: Effect of Noisy Training Data and the Optimization of Performance. Journal of Chemical Information and Modeling, 46 (2), pp 478-486.

บทความวิจัยตีพิมพ์ในการประชุมวิชาการนานาชาติ

[1] Pasupa, K. (In Press) Sparse Fisher Discriminant Analysis with Jeffrey's Hyperprior, In: Proceeding of the International Conference on Control, Automation & Information Sciences (ICCAIS'2012), 26-29 November 2012, Ho Chi Minh City, Vietnam.

[2] Pasupa, K. (In Press) Prediction by Nonparametric Posterior Estimation in Virtual Screening, In: Proceeding of the 2nd International Conference on Engineering, Applied Sciences, and Technology (ICEAST'2012) 21-24 November 2012, Bangkok, Thailand.

[3] Pasupa, K., Netisopakul, P. (2011) Thai Paragraph Shortening Based on Binary Classification Model, In: Proceeding of the Joint International Symposium on Natural Language Processing and Agricultural Ontology Service (SNLP-AOS'2011), 9-10 February 2012, Bangkok, Thailand, pp 181-185.

[4] Hussain, Z., Leung, A.P., Pasupa, K., Hardoon, D.R., Auer, P., Shawe-Taylor, J. (2010) Exploration-Exploitation of Eye Movement Enriched Multiple Feature Spaces for Content-Based Image Retrieval, In: Proceeding of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'2010), Lecture Notes in Artificial Intelligence, 6321, 20-24 September 2010, Barcelona, Spain, pp 554-569.

[5] Auer, P., Hussain, Z., Kaski, S., Klami, A., Kujala, J., Laaksonen, J., Leung, A.P., Pasupa, K., Shawe-Taylor, J. (2010) Pinview: Implicit Feedback in Content-Based Image Retrieval, In: Proceeding of Workshop on Applications of Pattern Analysis (WAPA'2010), JMLR Proceedings Series, 11, 1-2 September 2010, Cumberland Lodge, UK, pp 51-57.

[6] Auer, P., Hussain, Z., Kaski, S., Klami, A., Kujala, J., Laaksonen, J., Leung, A.P., Pasupa, K., Shawe-Taylor, J. (2010) Pinview: Implicit Feedback in Content-Based Image Retrieval, In: Proceeding of International Conference on Machine Learning (ICML'2010) Workshop on Reinforcement Learning and Search in Very Large Spaces, 25 June 2010, Haifa, Israel, available on-line at http://institute.unileoben.ac.at/infotech/research/workshops/icml2010-RLsearch/pages/auer.pdf

[7] Hussain, Z., Pasupa, K., Shawe-Taylor, J. (2010) Learning Relevant Eye Movement Feature Spaces Across Users In: Proceedings of the 6th Biennial Symposium on Eye Tracking Research & Applications (ETRA'2010), 22-24 March 2010, Austin, USA, pp. 181-185.

[8] Hardoon, D.R. and Pasupa, K. (2010) Image Ranking with Implicit Feedback from Eye Movements In: Proceedings of the 6th Biennial Symposium on Eye Tracking Research & Applications (ETRA'2010), 22-24 March 2010, Austin, USA, pp. 291-298.

[9] Pasupa, K., Szedmak, S. and Hardoon, D.R. (2009) Image Ranking with Eye Movements In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS'2009) Workshop on Advance in Rankings, 11 December, Whistler, Canada, pp. 37-42, available on-line at http://drona.csa.iisc.ernet.in/~shivani/Events/Ranking-NIPS-09/Proceedin...

[10] Pasupa, K., Saunders, C. J., Szedmak, S., Klami, A., Kaski, S. and Gunn, S. (2009) Learning to Rank Images from Eye Movements In: Proceeding of 2009 IEEE 12th International Conference on Computer Vision (ICCV'2009) Workshops on Human-Computer Interaction (HCI'2009), 27 September-4 October, Kyoto Japan, pp. 2009-2016.

[11] Pasupa, K., Harrison, R. F. and Willett, P. (2007) Parsimonious Kernel Fisher Discrimination. In: Proceeding of Iberian Conference on Pattern Recognition and Image Analysis, Lecture Notes in Computer Science, 4477 (1), 6-8 June 2007, Girona, Spain, pp. 531-538.

[12] Dechanupaprittha, S., Ngamroo, I., Pasupa, K., Tippayachai, J., Hongesombut, K. and Mitani, Y. (2004) New Heuristic-based Design of Robust Power System Stabilizers. In: Proceedings of 2004 International Conference on Power System Technology (POWERCON 2004), 21-24 November 2004, Singapore, pp. 618-623.

[13] Hongesombut, K., Mitani, Y., Dechanupaprittha, S., Ngamroo, I., Pasupa, K. and Tippayachai, J. (2004) Power System Stabilizer Tuning Based on Multiobjective Design Using Hierarchical and Parallel Micro Genetic Algorithm. In: Proceedings of 2004 International Conference on Power System Technology (POWERCON 2004), 21-24 November 2004, Singapore, pp. 402-407.

การนำเสนอผลงานในงานประชุมวิชาการนานาชาติ (บทคัดย่อ)

[1] Pasupa, K., Klami, A., Saunders, C.J., de Campos, T., Kaski, S. (2009) Can relevance of images be inferred from eye movements? In: 15th European Conference on Eye Movements, 23 - 27 August 2009, Southampton, UK, pp. 50

[2] Pasupa, K. (2007) Prediction by Nonparametric Posterior Estimation in Virtual Screening. In: The 2007 University of Sheffield Symposium On Data Modelling for New Researchers, 27 March 2007, University of Sheffield.

รายงานวิจัย

[1] Szedmak, S., Gunn, S.R., Pasupa, K., and Hardoon, D.R. (2010) An invariant approach to multi-view learning with application to missing data estimation. Technical Report, School of Electronics & Computer Science, University of Southampton, available on-line at http://users.ecs.soton.ac.uk/ss03v/papers/multiview_094.pdf.

[2] Pasupa, K., Saunders, C., Szedmak, S., Gunn, S.R., Hardoon, D.R., Klami, A., Kaski, S., Leung, A. and Auer, P. (2009) Ranking algorithms for implicit feedback. Pinview Deliverable D.5.1, available on-line at http://www.pinview.eu/files/pinview-d5-1-final.pdf.

[3] Klami, A., Kaski, S., Pasupa, K., Szedmak, S., Gunn, S., Hardoon, D. and Csurka, G. (2009) Predicting relevance of parts of an image. Pinview Deliverable D.6.3, available on-line at http://www.pinview.eu/files/pinview-d2-2-final.pdf.

[4] de Campos, T., Csurka, G., Perronnin, F., McAuley, J., Antenreiter, M., Ortner, R., Auer, P., Viitaniemi, V., Laaksonen, J., Pasupa, K., Saunders, C., Hussain, Z., Shawe-Taylor, J. (2009) Description and evaluation of techniques for transfer learning across sub-categories. Pinview Deliverable D.6.3, available on-line at http://www.pinview.eu/files/pinview-d6-3-final.pdf.

[5] de Campos, T., Csurka, G., Perronnin, F., Hussain, Z., Shawe-Taylor, J., Pasupa, K., Saunders, C.J., Ali, H., Antenreiter, M., Ortner, R., Auer, P., Viitaniemi, V., Laaksonen, J. (2008) Description, analysis and evaluation of confidence estimation procedures for sub-categorisation. Pinview Deliverable D.6.2.1, available on-line at http://www.pinview.eu/files/pinview-d6-2-1-final.pdf .

[6] Klami, A., Kaski, S., Pasupa, K., Saunders, C.J., de Campos, T. (2008) Prediction of relevance of an image from a scan pattern. Pinview Deliverable D.2.1, available on-line at http://www.pinview.eu/files/pinview-d2-1-final.pdf .

[7] Hussain, Z., Shawe-Taylor, J., Saunders, C.J., Pasupa, K. (2008) Basic metric learning. Pinview Deliverable D.3.1, available on-line at http://www.pinview.eu/files/pinview-d2-1-final.pdf .