

บทที่ 5

การจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอโดยใช้วิธีการเค-เนียร์เรสต์เนเบอร์ บนพื้นฐานของสเปกตรัมเคอร์เนล

ในบทนี้จะเป็นการนำวิธีการจำแนกชนิดข้อมูลสายลำดับ ด้วยวิธีสเปกตรัมเคอร์เนลมารวมกับวิธีเค-เนียร์เรสต์เนเบอร์ ซึ่งจะทำการทดสอบกับกระบวนการวิธีในการจำแนกข้อมูลสายลำดับที่ได้จากบทที่ 3 กรณีศึกษาคือ การจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอ โดยเราจะมุ่งเน้นการแก้ปัญหาในการจำแนกชนิดของโมเลกุลเอชแอลเอ ซึ่งสามารถแบ่งออกได้เป็นโมเลกุลของเอชแอลเอกลุ่มหลักและโมเลกุลของเอชแอลเอกลุ่มย่อย

ตามหน้าที่เฉพาะของโมเลกุลเอชแอลเอแต่ละกลุ่มในระบบภูมิคุ้มกัน (Katz et al., 1973; Han et al., 2000) โดยกลุ่มย่อยของเอชแอลเอที่ I จะประกอบด้วย A, B, Cw, E, F, G ส่วนกลุ่มย่อยของเอชแอลเอกลุ่มที่ II จะประกอบด้วย DMA, DMB, DOA, DOB, DPA1, DPB1, DQA1, DQB1, DRA, DRB โดยรายละเอียดข้อมูลจะแสดงตามตาราง 4.1

ซึ่งประสิทธิภาพในกระบวนการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอนั้น สามารถทำให้เข้าใจถึงระบบภูมิคุ้มกันของมนุษย์ ได้มากขึ้น โดยในปัจจุบันมีวิธีการคำนวณในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอมากมาย ได้แก่ วิธีการที่มีการนำเสนอ โดยใช้คุณลักษณะต่างๆ รวมกับการใช้โคดอน ยูสเอสเป็นคุณลักษณะข้อมูลเข้า (Feature inputs) ในการจำแนกชนิดของยีน ซึ่งแต่ละองค์ประกอบของข้อมูลจะมีการปรับให้อยู่ในรูปแบบเวกเตอร์ ด้วยขั้นตอนวิธีอาเอสซียู และใช้กับวิธีซัพพอร์ตเวกเตอร์แมชชีน ในการจำแนกชนิดข้อมูล (Ma et al., 2009) โดยวิธีการดังกล่าวนี้ในขั้นตอนแรกจะทำการแปลงข้อมูลสายลำดับของโมเลกุลเอชแอลเอให้อยู่ในรูปของเวกเตอร์ ซึ่งมีรูปแบบเท่ากับ 59 รูปแบบ ด้วยวิธี อาเอสซียู ดังที่ได้อธิบายในบทที่ 2 และทำการฝึกสอนและจำแนกชนิดข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน จากนั้นจึงทำการเปรียบเทียบกับวิธีการทันสมัยอื่น เป็นต้น

ต่อมาจึงมีผู้เสนอวิธีการในการใช้โคดอนและไค-โคดอน ในการวิเคราะห์คุณลักษณะเพื่อจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอ และแสดงให้เห็นว่าการใช้ไค-โคดอน ได้ผลลัพธ์ที่มีประสิทธิภาพดีกว่าโคดอน จึงทำให้เห็นว่าไค-โคดอน สามารถเป็นตัวบ่งชี้ที่ดีของคุณลักษณะสำหรับข้อมูลในกลุ่มยีนและโมเลกุล

แทนที่จะใช้คุณลักษณะดังกล่าวข้างต้น วิธีการที่เราจะนำเสนอต่อไปนี้จะพบว่าสามารถใช้งานได้ดีกับโมเดลเอชแอลเอโดยตรง ซึ่งสามารถจำแนกชนิดข้อมูลสายลำดับของโมเดลเอชแอลเอได้เช่นกัน ดังนั้นจึงมีการนำเสนอวิธีการใหม่ ที่เป็นการรวมวิธีการของวิเค-เนียร์เรสต์เนเบอร์กับวิธีการของสเปกตรัมเคอร์เนล โดยวิธีการของสเปกตรัมเคอร์เนล จะเป็นวิธีการที่ง่ายและมีประสิทธิภาพในการจำแนกชนิดของข้อมูลสายลำดับต่างๆ ส่วนขั้นตอนวิธีการของเค-เนียร์เรสต์เนเบอร์นั้น จะเป็นวิธีการที่รู้จักและนิยมใช้ในการจำแนกชนิดข้อมูลกันอย่างแพร่หลาย ซึ่งเราจะทำการรวมวิธีการทั้ง 2 วิธีเข้าด้วยกัน และทำการวัดประสิทธิภาพการจำแนกชนิดข้อมูลกับวิธีการทันสมัยอื่น ซึ่งผลการทดลองที่ได้ จะนำเสนอต่อไป

5.1 การจำแนกข้อมูลสายลำดับโดยวิธีการเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของสเปกตรัมเคอร์เนล

วิธีการเค-เนียร์เรสต์เนเบอร์เป็นวิธีในการจำแนกชนิดข้อมูล โดยอยู่บนแนวคิดของการวัดระยะห่างของความคล้ายของวัตถุใดวัตถุหนึ่ง ซึ่งตัวจำแนกนั้นจะมีขั้นตอนในการเรียนรู้ เพื่อแสดงให้เห็นถึงประสิทธิภาพของการแก้ไขปัญหามากมาย รูปแบบ (Aha et al., 1991) โดยในการจำแนกข้อมูลนั้นจะทำการกำหนดลักษณะของข้อมูลต่างๆ ได้ดังนี้

โดย D คือเซตของข้อมูลสายลำดับทั้งหมด

k คือจำนวนเต็มบวกขนาดเท่ากับ k

x คือข้อมูลสายลำดับตัวใหม่ในการพิจารณา

ซึ่งตัวจำแนกของวิธีการเค-เนียร์เรสต์เนเบอร์นั้นจะทำการหาข้อมูลที่ใกล้เคียงกับ x มากที่สุดในเซตของข้อมูลสายลำดับทั้งหมด ซึ่งสามารถกำหนดได้เป็น $knn(x)$ โดยคำนียามพื้นฐานนั้นจะอยู่ในรูปแบบของการสร้างฟังก์ชันระยะห่างที่เหมาะสม เพื่อเป็นตัวชี้ให้เห็นถึงประสิทธิภาพของตัวจำแนกของเค-เนียร์เรสต์เนเบอร์ โดยเป็นการหาชนิดของข้อมูลด้วยระยะทางแบบยูคลิดเดียน (Euclidean distance) ซึ่งเป็นขั้นตอนในการวัดระยะห่างปกติระหว่างจุดสองจุดในแนวเส้นตรง เพื่อหาฟังก์ชันของระยะห่างที่เหมาะสม ซึ่งเป็นวิธีจะทำการเปลี่ยนข้อมูลสายลำดับของดีเอ็นเอให้อยู่ในรูปข้อมูลที่สามารถคำนวณได้ ก่อนที่จะทำการจำแนกชนิดข้อมูลด้วยวิธีการเค-เนียร์เรสต์เนเบอร์ ซึ่งวิธีการนี้มีใช้กันอย่างแพร่หลาย แต่วิธีการนี้จะไม่เหมาะกับลักษณะข้อมูลที่มีมิติสูง เนื่องจากวิเค-เนียร์เรสต์เนเบอร์นี้จะใช้ได้กับข้อมูลที่มีคุณลักษณะการจัดเรียงที่ไม่ซับซ้อนมากแต่ในเหตุผลหลักคือวิธีการเค-เนียร์เรสต์ เนเบอร์จะเป็นการวัดความคล้ายของข้อมูลในรูปของเวกเตอร์ ซึ่งจะอิงกับฟังก์ชันระยะห่างแต่ข้อมูลที่เราทดลองนั้น อยู่ในรูปของสายลำดับ ดังนั้นเราจึงใช้วิธีสเปกตรัมเข้ามาช่วยในการจำแนกชนิดข้อมูล

ดังนั้นวิธีการที่เรานำเสนอจะทำการนำวิธีการเค-เนียร์เรสต์เนเบอร์มาทำงานร่วมกับสเปกตรัมเคอร์เนล เพื่อจำแนกชนิดของโมเลกุลเอชแอลเอ ซึ่งวิธีการนี้มีขั้นตอนการทำงานแบบเฉพาะเจาะจง โดยเน้นถึงฟังก์ชันของระยะห่างของข้อมูลเป็นสำคัญ โดยเราจะนำเสนอวิธีการของสเปกตรัมเคอร์เนล ซึ่งสามารถใช้กับชุดข้อมูลสายลำดับได้โดยตรง และตามเงื่อนไขทางคณิตศาสตร์ (Schölkopf et al., 2004) ซึ่งวิธีการที่นำเสนอนี้จะใช้เทคนิคการเรียนรู้ด้วยเครื่องที่เข้าใจง่ายและสามารถใช้กับวิธีการวิเคราะห์แบบอื่นได้อย่างหลากหลาย

5.2 ผลการทดลอง

ในกระบวนการทดลองนั้นจะทำ 10-ครอสวาริเคชัน กับชุดข้อมูลสายลำดับของโมเลกุลเอชแอลเอเท่ากับ 4,403 สายลำดับโดยใช้วิธีการนี้ในการจำแนกข้อมูลของโมเลกุลเอชแอลเอ ในแต่ละกลุ่มได้แก่ เอชแอลเอกลุ่มที่ I และเอชแอลเอกลุ่มที่ II และทำการจำแนกชนิดของโมเลกุลเอชแอลเอ ในกลุ่มย่อยต่อไป

5.2.1 วิธีการเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของสเปกตรัมเคอร์เนล

ในตาราง 5.1 มีรูปแบบการพิจารณาค้างนี้คือในหลักของแถวจะแสดงถึง เค-สเปกตรัม โดยจะพิจารณาตามขนาดหรือช่วงพิจารณาที่แตกต่างกันออกไปและในหลักของสดมภ์ จะแสดงถึงเค-เอ็นเอ็น โดยจะพิจารณาจากค่าของเค-เอ็นเอ็น แต่ละตัว ซึ่งเราจะกำหนดความยาวของ เค ของวิธีสเปกตรัมเคอร์เนลและวิธีเค-เนียร์เรสต์เนเบอร์ เพื่อถ่ายทอดการเปรียบเทียบประสิทธิภาพของการทำนาย ซึ่งผลการทดลองจะแสดงได้ในตาราง 5.1

ตาราง 5.1 การวัดประสิทธิภาพของการทำนายในวิธีสเปกตรัมเคอร์เนล
ร่วมกับวิธีเค-เนียร์เรสต์เนเบอร์

เค-สเปกตรัม	กลุ่ม	เค-เอ็นเอ็น	
		3-เอ็นเอ็น	5-เอ็นเอ็น
เอสเค = 3	กลุ่มทั้งหมด	99.72	99.65
	กลุ่มที่ I	99.90	99.81
	กลุ่มที่ II	95.50	95.66
เอสเค = 4	กลุ่มทั้งหมด	99.74	99.74
	กลุ่มที่ I	99.90	99.88
	กลุ่มที่ II	95.42	95.40
เอสเค = 5	กลุ่มทั้งหมด	99.74	99.74
	กลุ่มที่ I	99.90	99.88
	กลุ่มที่ II	96.41	96.80
เอสเค = 6	กลุ่มทั้งหมด	99.74	99.74
	กลุ่มที่ I	99.90	99.88
	กลุ่มที่ II	96.20	96.85

ในการทดลอง เราเริ่มต้นด้วยขนาดหรือช่วงการพิจารณาของสเปกตรัมเคอร์เนลเท่ากับ 3 เนื่องจากแนวคิดนี้ตรงกับวิธีการของโคดอนซึ่งเป็นคุณสมบัติพื้นฐานในการศึกษาข้อมูลระดับโมเลกุล ซึ่งกลุ่มงานวิจัยส่วนใหญ่ (Milhon et al., 1995 และ Mitreva et al., 2007) ได้เสนอว่า การใช้เค-สเปกตรัมเท่ากับ 3 นั้นสามารถแสดงถึงคุณลักษณะโดยตรงในระดับโมเลกุลแต่ละพันธุกรรมของสิ่งมีชีวิตแต่ละสิ่งมีชีวิต ซึ่งเมื่อใช้ เค-สเปกตรัมเท่ากับ 3, 4, 5, 6 ร่วมกับ 3-เอ็นเอ็นและ 5-เอ็นเอ็น ตามตาราง 5.1 จะเห็นได้ว่าค่าความถูกต้องที่ได้มีค่าใกล้เคียง 99.7% ในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มทั้งหมด และกลุ่มของเอชแอลเอกลุ่มที่ I เท่ากับ 99.8-99.9% ส่วนกลุ่มของเอชแอลเอกลุ่มที่ II เท่ากับ 95.4-96.8% ซึ่งผลลัพธ์ดังกล่าวได้จากการใช้ขนาดของเค-สเปกตรัมที่ต่างกันแต่ผลลัพธ์ที่ได้มีค่าไม่แตกต่างกัน ดังนั้นการเลือกใช้ขนาดของเค-สเปกตรัมนั้น จะขึ้นอยู่กับการใช้งานและปัจจัยอื่นที่เราสนใจศึกษา ดังนั้นเราจึงเลือกใช้เค-สเปกตรัมเท่ากับ 6 ร่วมกับ 5-เอ็นเอ็น เนื่องจากระดับการพิจารณานี้มีความสัมพันธ์กับการถอดรหัสพันธุกรรมและการศึกษาในระดับโมเลกุล (Nguyen et al., 2009) นอกจากนี้การใช้โคดอน นี้เป็นตัวบ่งชี้ที่ดีในการ

แสดงออกของยีนและการศึกษาระดับโมเลกุล ซึ่งยังมีประโยชน์ต่อการจัดหมวดหมู่ของยีนเป็นต้น

นอกจากนี้ เราได้ทำการวัดประสิทธิภาพในการทำนายของเค-สเปกตรัมที่มีขนาด เค เท่ากับ 7, 8 และ 9 ซึ่งผลลัพธ์ที่ได้สามารถแสดงให้เห็นว่าการเพิ่มความยาวของขนาดหรือช่วงการพิจารณาขึ้นนั้น ก็ผลลัพธ์ใกล้เคียงกับการใช้เค-สเปกตรัมเท่ากับ 6 โดยสามารถสรุปได้ว่าการที่จะได้ผลลัพธ์ในการทำนายที่ได้นั้นไม่ได้ขึ้นอยู่กับค่า เค ที่มีขนาดใหญ่ ซึ่ง 6-สเปกตรัมนั้นมีความยาวเพียงพอต่อการแสดงออกของคุณลักษณะสำคัญของข้อมูลเอชแอลได้

5.2.2 วิธีการเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของสเปกตรัมเคอร์เนลเปรียบเทียบกับวิธีการทันสมัยอื่น ในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มหลัก

จากที่ได้กล่าวมาในข้างต้น การทดลองของเราจะแสดงให้เห็นถึงค่าของพารามิเตอร์ที่ดีที่สุดในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอ คือ การใช้เค-สเปกตรัมเท่ากับ 6 ร่วมกับ 5-เอ็นเอ็น โดยในหัวข้อนี้จะทำการเปรียบเทียบประสิทธิภาพในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอ ด้วยวิธีการเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของสเปกตรัมเคอร์เนลกับวิธีการทันสมัยอื่น ซึ่งใช้หลักเกณฑ์เดียวกันในการวิเคราะห์ข้อมูล ได้แก่ เคเอสวีเอ็มและโครงข่ายประสาทเทียม (Ripley et al., 2009) เพราะว่าวิธีการทันสมัยเหล่านี้เป็นตัวแทนจำลองที่เป็นที่รู้จักและใช้กันอย่างแพร่หลายสำหรับการจำแนกชนิดข้อมูล ซึ่งประสิทธิภาพของการทำนายด้วยวิธีเคเอสวีเอ็ม โดยใช้เคอร์เนลแบบเชิงเส้น, แบบโพลิโนเมียลที่มีมิติ เท่ากับ 2, 3, 4 มิติ และแบบเกาส์เซียนในวัดประสิทธิภาพการทำนาย แต่ในวิธีการของเคเอสวีเอ็มและโครงข่ายประสาทเทียมนั้น ขั้นตอนแรกจะต้องทำการแปลงโมเลกุลของเอชแอลเอให้อยู่ในรูปแบบเวกเตอร์ ซึ่งมีมิติเท่ากับ 59 มิติ ก่อนที่จะทำการจำแนกชนิดข้อมูลต่อไป ส่วนในตาราง 5.2 จะแสดงให้เห็นถึงประสิทธิภาพในการทำนายของวิธีการแบบเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของสเปกตรัมเคอร์เนล โดยเปรียบเทียบกับวิธีการทันสมัยอื่นซึ่งได้ผลลัพธ์คือค่าความถูกต้องเท่ากับ 99.6%, ค่าความไวในการทำนายโมเลกุลของเอชแอลเอกลุ่มที่ I เท่ากับ 100% และค่าความแม่นยำในการทำนายโมเลกุลของเอชแอลเอกลุ่มที่ II เท่ากับ 100% ซึ่งผลลัพธ์ที่ได้แสดงให้เห็นว่าวิธีการที่เรานำเสนอนั้นดีกว่าวิธีการทันสมัยอื่น ยกเว้นค่าสัมประสิทธิ์ความสัมพันธ์แมทริวและค่าความไวของวิธีเคเอสวีเอ็มในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มที่ II

ตาราง 5.2 ประสิทธิภาพในการทำนายของวิธีการแบบเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของสเปกตรัมเคอร์เนลโดยเปรียบเทียบกับวิธีการทันสมัยอื่น (เอชแอลเอกลุ่มหลัก)

วิธีการ และ พารามิเตอร์				เอชแอลเอ-I		เอชแอลเอ-II	
		Acc	MCC	Prec	Sens	Prec	Sens
โครงข่ายประสาท	size=2	98.5	96.6	87.6	99.0	99.2	95.4
เชิงเส้น	c=20	91.4	98.3	97.7	99.1	99.2	98.9
โพลีโนเมียล	d=2	99.0	98.9	98.8	99.2	99.3	99.3
โพลีโนเมียล	d=3	99.3	99.2	99.1	99.4	99.7	99.5
โพลีโนเมียล	d=4	99.0	98.8	98.8	99.0	99.3	99.4
เกาส์เซียน	sig=0.01	99.5	99.5	99.5	99.8	99.8	99.7
สเปกตรัมเคอร์เนล+เคเอ็นเอ็น	sk=6 และ 5-NN	99.6	99.4	99.7	100.0	100.0	99.0

5.2.3 วิธีการเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของสเปกตรัมเคอร์เนลเปรียบเทียบกับวิธีการทันสมัยอื่น ในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มย่อย

นอกจากนี้เรายังนำวิธีการแบบเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของสเปกตรัมเคอร์เนลจำแนกชนิดข้อมูลในกลุ่มย่อยของโมเลกุลเอชแอลเอกลุ่มที่ I และเอชแอลเอกลุ่มที่ II บนหลักเกณฑ์เดียวกันที่ใช้ก่อนหน้านี้ โดยตาราง 5.3 จะแสดงถึงประสิทธิภาพในการทำนายของวิธีที่เราแนะนำ โดยเปรียบเทียบกับวิธีการทันสมัยอื่นซึ่งได้ผลลัพธ์คือค่าความถูกต้องเท่ากับ 99.9% สำหรับการทำนายกลุ่มย่อยของโมเลกุลเอชแอลเอกลุ่มที่ I แต่วิธีเคอร์เนลฟังก์ชันแบบเกาส์เซียนจะได้ผลลัพธ์ที่ดีที่สุดในการทำนายกลุ่มย่อยของโมเลกุลเอชแอลเอกลุ่มที่ II เท่ากับ 99.3%

ตาราง 5.3 ประสิทธิภาพในการทำนายของวิธีการแบบเค-เนียร์เรสต์เนเบอร์บนพื้นฐานของ
สเปกตรัมเคอร์เนลโดยเปรียบเทียบกับวิธีการทันสมัยอื่น (เอชแอลเอกลุ่มย่อย)

การจำแนกชนิดข้อมูล	เคอร์เนลฟังก์ชัน และ พารามิเตอร์						สเปกตรัมเคอร์เนล sk=6,5-NN
	NN	Linear	Poly d=2	Poly d=3	Poly d=4	Gaussian	
กลุ่มย่อยของ เอชแอลเอกลุ่มที่ I	65.9	98.1	98.6	98.4	98.1	99.6	99.9
กลุ่มย่อยของ เอชแอลเอกลุ่มที่ II	71.4	98.8	98.5	98.9	98.5	99.3	96.9