การจำแนกมะเร็งเม็ดเลือดขาวโดยใช้เทคนิคการลดมิติข้อมูลด้วย Chi-square

Classification for Leukemia Using Feature Selection based on Chi-square Technique

นรินทร์ พนาวาส[1] และ นิเวศ จิระวิชิตชัย[2]

[1]สาขาวิชาคอมพิวเตอร์ธุรกิจ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม วิทยาเขตชลบุรี

[2]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสวนสุนันทา

Emails: narin.pa@east.spu.ac.th, nivet99@hotmail.com

บทคัดย่อ

งานวิจัยนี้ได้เสนอวิธีการลดมิติของยีน ด้วยค่าสถิติไคแสควร์ (Chisquare) และโดยทำการทดสอบประสิทธิภาพ การจำแนกประเภทของมะเร็งเม็ดเลือดขาวกับ อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ต้นไม้ตัดสินใจ (Decision Tree) เนอีฟเบย์ (Naïve-Bayes) เคเนียเรสเนเบอร์ (K-Nearest Neighbor) จากการทดลองพบว่า การลดมิติของข้อมูลด้วยวิธีไคแสควร์แล้วส่งเข้าประมวลผลด้วยเครื่องจักรการเรียนรู้ โดยวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) จากจำนวนมิติของข้อมูลที่ส่งผลดีที่สุดพบว่า อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน(Support Vector Machine) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ที่จำนวน 300 มิติ ค่าความถูกต้อง (Accuracy) เท่ากับ 98.61 % เท่ากัน อัลกอริทึมเนอีฟเบย์ (Naïve-Bayes) ที่จำนวน 4000 มิติ ค่าความถูกต้อง (Accuracy) เท่ากับ 98.61 % อัลกอริทึม ต้นไม้ตัดสินใจ (Decision Tree) ที่จำนวน 1 มิติ ค่าความถูกต้อง (Accuracy) เท่ากับ 93.06 %

คำสำคัญ— การลดมิติข้อมูล;เครื่องจักรการเรียนรู้;การจำแนกประเภท

Abstract

This research presented dimension reduction of genes with the Chi Square technique.Test performance classification with Support Vector Machine,Decision Tree, Naive-Bayes and K-Nearest Neighbor algorithms. The experimental results showed that reducing the dimension by Chi Square technique and process with machine learning algorithms. The performance of accuracy found Support Vector Machin , Naive - Bayes and K-Nearest Neighbor yielded a very high classification with the accuracy equal to 98.61%. Followed by the Decision Tree yielded with the accuracy 93.06 % respectively.

Keywords*: dimension reduction, machine learning,classification*

1. บทนำ

มะเร็งเม็ดเลือดขาว (Leukemia) จัดเป็นมะเร็งที่เกิดจากความผิดปกติของเซลล์เม็ดเลือด แบ่งออกเป็น 2 ชนิดใหญ่ๆ คือชนิดเฉียบพลัน และชนิดเรื้อรังซึ่งมะเร็งเม็ดเลือดขาวชนิดเฉียบพลันเป็นชนิดที่พบมากที่สุด จัดเป็นโรคมะเร็งทางโลหิตวิทยาที่เกิดจากความผิดปกติของเซลล์ต้นกำเนิด เม็ดเลือดในไขกระดูก ทำให้มีการแบ่งตัวเพิ่มจำนวนของเซลล์ตัวอ่อนของเม็ดโลหิตขาวเพิ่มขึ้นอย่างมาก แต่เซลล์เหล่านั้นไม่สามารถเจริญเป็นตัวแก่ได้ตามปกติ เซลล์ตัวอ่อนของเม็ดโลหิตขาวในไขกระดูกจะเพิ่มจำนวนมากขึ้นเรื่อยๆ เซลล์มะเร็งเหล่านี้ จะแทนที่และยับยั้งการเจริญเติบโตของเซลล์ซึ่งสร้างเม็ดเลือดแดง เม็ดเลือดขาวและเกร็ดเลือด จนเกิดอาการและอาการแสดงในผู้ป่วยในระยะเวลาอันสั้น เช่น มีอาการซีด อ่อนเพลีย มีเลือดออกผิดปกติ หรือจุดเลือดออก เนื่องจากเกร็ดเลือดต่ำ ติดเชื้อง่ายมีไข้สูง นอกจากนี้ยังอาจเกิดอาการผิดปกติจากการที่เซลล์มะเร็ง แทรกตัวเข้าไปในเนื้อเยื่ออื่น ๆ [1]

ซึ่งหนึ่งวิธีในการวินิจฉัยในการตรวจโรคมะเร็งเม็ดเลือดขาว คือการตรวจระดับยีน การตรวจวิธีนี้นอกจากจะมีประโยชน์ในการช่วยแยกชนิดของมะเร็งเม็ดเลือดขาวชนิดเฉียบพลันแล้ว ยังสามารถพยากรณ์โรคได้ ซึ่งมีประโยชน์ในการติดตามและการรักษาโรคได้อย่างทันท่วงที ดังนั้นงานด้านการจำแนกประเภทของมะเร็งเม็ดเลือดขาว จึงเป็นสิ่งจำเป็นสำหรับการตรวจวินิจฉัยอาการผู้ป่วย เพื่อที่จะทำให้การดำเนินดูแลการรักษาได้ตรงตามอาการของโรค ในทางการแพทย์การวินิจฉัยมะเร็งเม็ดเลือดขาวแต่ละชนิดจึงมีความสำคัญอย่างยิ่ง เนื่องจากการตอบสนองต่อการรักษาด้วยยาหรือวิธีการรักษาของมะเร็งเม็ดเลือดขาวแต่ละชนิดแตกต่างกัน [2-3]

แต่ปัญหาในการจำแนกประเภทของมะเร็งเม็ดเลือดขาว ซึ่งประกอบด้วยมิติของยีนอยู่เป็นจำนวนมากนั้น จะอาศัยความสามารถของมนุษย์เพียงอย่างเดียว จะทำให้การจำแนกมีประสิทธิภาพอย่างจำกัด จากความของปัญหาดังกล่าว งานวิจัยนี้จึงมุ่งเน้นศึกษาวิธีการเพื่อเพิ่มประสิทธิภาพในการจำแนกประเภทของมะเร็งเม็ดเลือดขาวให้เป็นแบบอัตโนมัติ โดยประสิทธิภาพที่ได้ จะใกล้เคียงกับการจำแนกมะเร็งเม็ดเลือดขาวที่ทำโดยมนุษย์ ซึ่งประโยชน์จากงานวิจัยนี้จะทำให้ประหยัดแรงงานมนุษย์เป็นอย่างมาก เพราะทดแทนผู้เชี่ยวชาญในการจำแนกประเภทมะเร็งเม็ดเลือดขาว โดยงานวิจัยนี้ได้เสนอวิธีการลดมิติของยีนด้วยค่าสถิติไคแสควร์ (Chisquare) [4] และโดยทำการทดสอบประสิทธิภาพการจำแนกประเภทของมะเร็งเม็ดเลือดขาว กับอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree)

ซัพพอร์ทเวกเตอร์แมชชีน(Support Vector Machine) เนอีฟเบย์ (Naïve-Bayes) เคเนียเรสเนเบอร์ (K-Nearest Neighbor)

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การเลือกคุณลักษณะ (Feature Selection)

จากปัญหาที่กล่าวมา การจำแนกประเภทของมะเร็งเม็ดเลือดขาว ซึ่งมีจำนวนของยีนปริมาณมากนั้น จำเป็นต้องค้นหากลุ่มของยีนที่มีคุณลักษณะที่มีอำนาจในจำแนกประเภทของมะเร็งเม็ดเลือดขาวที่มีประสิทธิภาพ แต่ลักษณะของข้อมูลมะเร็งเม็ดเลือดขาว ข้อมูลมีจำนวนมิติมากและมิติข้อมูลเหล่านั้นไม่ได้ไปในทิศทางเดียวกัน ทำให้ข้อมูลยีนมะเร็งเม็ดเลือดขาว เกิดการกระจาย (data sparse) ซึ่งจะทำให้บางจุดอาจจะไม่มีข้อมูลอยู่เลย ซึ่งสิ่งนี้ถูกเรียกว่าเป็นปัญหาของมิติข้อมูล ดังนั้นการขจัดปัญหาของมิติข้อมูลนั้นนอกจากจะช่วยประหยัดทรัพยากรแล้ว ยังเพิ่มประสิทธิภาพในการแยกประเภทข้อมูลของมะเร็งเม็ดเลือดขาวได้แม่นยำขึ้นอีกด้วย ซึ่งเทคนิคการลดมิติเป็นขั้นตอนหนึ่งของการเตรียมข้อมูล ที่สามารถค้นหากลุ่มย่อยของยีนที่มีอำนาจจำแนกได้อย่างมีระสิทธิภาพ [2-6]

2.2 ค่าสถิติไคสแควร์ (Chisquare)

ปัจจุบันมีหลายเทคนิคสำหรับการค้นหากลุ่มย่อยของยีน ที่มีอำนาจ ในการจำแนกประเภทของมะเร็งเม็ดเลือดขาว มีวิธีการพื้นฐานโดยจัดอันดับการแสดงออกของยีน (Ranking Gene Expression) ซึ่งงานวิจัยนี้ได้มุ่งเน้นศึกษาการลดคุณลักษณะด้วยวิธี Chi-square ผลที่ได้จากการทดลองคือกลุ่มข้อมูลการแสดงออกของยีนในระดับที่แตกต่างกัน ทั้งนี้มีความเชื่อว่าระดับการแสดงออกที่สูงของยีนในตัวอย่างทดสอบใด ๆ ย่อมแสดงว่ายีนนั้นมีอิทธิพลต่อการจำแนกประเภทของมะเร็งเม็ดเลือดขาว ค่าสถิติไคสแควร์ วัดความเป็นอิสระต่อกันระหว่างตัวแปรและกลุ่ม โดยค่าสถิติไคสแควร์ที่ใช้ในงานวิจัยนี้จะพิจารณาจากค่าสูงสุด สามารถนิยามได้โดย [4]

$$\chi^2(w,c_j)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

A คือ จำนวนตัวอย่างจากกลุ่ม C_j ซึ่งมียีน w

B คือ จำนวนตัวอย่างซึ่งมียีน w แต่ไม่ได้อยู่ในกลุ่ม C_j

C คือ จำนวนตัวอย่างจากกลุ่ม C_j ซึ่งไม่มียีน w

D คือ จำนวนตัวอย่างที่ไม่ได้อยู่ในกลุ่ม C_j หรือไม่มียีน w

N คือ จำนวนกลุ่มตัวอย่างทั้งหมด

C_j คือ กลุ่มประเภทของมะเร็งเม็ดเลือดขาว

2.3 อัลกอริทึมการจำแนกประเภท (Classifier Algorithm)

อัลกอริทึมในการจำแนกประเภทเรียนรู้แบบมีผลเฉลย (Supervised Learning) สามารถแบ่งขั้นตอนวิธีการจำแนกประเภทได้เป็น 2 ขั้นตอนคือ การเรียนรู้เพื่อสร้างกลุ่มต้นแบบและ จำแนกประเภทของกลุ่มตัวอย่างที่สนใจ โดยการตรวจสอบหาความคล้ายกับกลุ่มตัวอย่างต้นแบบ [6-9]

2.3.1 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้จะประกอบด้วยโหนดแทนคุณลักษณะ และโหนดล่างสุดแทนหมวดหมู่ การสร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าความจริงที่ใช้จะมาจากการคำนวณหาค่า Entropy และค่า Information Gain คุณลักษณะใดที่มีค่า Information Gain มากที่สุดจะถูกเลือกเป็นโหนดลูก หากมีค่า Entropy เป็น 0 จะได้เป็นโหนดล่างสุด

2.3.2 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine)

หลักการของ SVM คือการสร้างสมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกันโดย SVM จะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มี ระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด SVM จะใช้ฟังก์ชั่นแม็ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่า (Kernel Function) บน Feature Space

2.3.3 เนอีฟเบย์ (Naïve-Bayes)

ใช้ทฤษฎี Bayes Theorem ในการคำนวณความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล Naïve-Bayes เป็นเทคนิคในการแก้ปัญหาแบบจำแนก (classification) ที่ทั้งสามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย มันจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์

2.3.4 เคเนียเรสเนเบอร์ (K-Nearest Neighbor)

หลักการของวิธีการนี้ จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุดK ตัวจากข้อมูลบนชุดข้อมูลตัวอย่างเป็นอัลกอริทึ่มที่เรียบง่าย ทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่ หรือข้อมูลที่ป้อนถาม (input query instance) กับข้อมูลตัวอย่างฝึกฝน จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นเราจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่

2.4 การวัดประสิทธิภาพ

วิธีการทดสอบ เพื่อเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมในการจำแนกประเภทของมะเร็งเม็ดเลือดขาวนั้น สามารถวัดที่ประสิทธิภาพของการจำแนกข้อมูล ตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) และค่า F-Measurement ซึ่งคำนวณได้ดังตาราง Confusion matrix ดังภาพ [6-8]

		actual value		
		p	*n*	total
prediction outcome	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

ภาพที่ 1 Confusion matrix

TP = จำนวนตัวอย่างที่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าอยู่กลุ่ม C_j

FP = จำนวนตัวอย่างที่ไม่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าอยู่กลุ่ม C_j

FN = จำนวนตัวอย่างที่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่กลุ่ม C_j

TN = จำนวนตัวอย่างที่ไม่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่กลุ่ม C_j

C_j = กลุ่มประเภทของมะเร็งเม็ดเลือดขาวที่สนใจวัดประสิทธิภาพ

$$Accuracy = (TP + TN)/(TP + FP + FN + TN)$$

$$Precision = TP /(TP + FP)$$

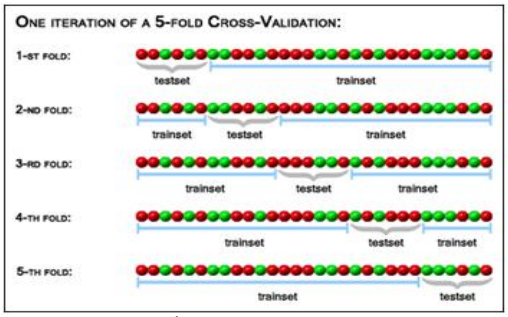
$$Recall = TP /(TP + FN)$$

$$F\text{-}Measure = 2*Precision*Recall/(Precision+Recall)$$

2.5 ครอสวาลิเดชั่น (Cross Validation)

วิธีการในการคาดการค่าความผิดพลาดของโมเดลหรือ วิธีการที่นำเสนอโดยพื้นฐานของวิธีการครอสวาลิเดชั่นคือ การสุ่มตัวอย่าง (Resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆและนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำครอสวาลิเดชั่นมักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (Network architecture), โมเดลในการคัดแยกประเภท(Classification model) เช่นในการทำ Classify ข้อมูลโดยใช้เทคนิคของ Data mining เช่น Neural Network หรือ Decision Tree นั้นจะต้องมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ แต่ในบางครั้งอาจเกิดปัญหาจากการเลือกข้อมูลที่ดีและง่ายมาเป็นข้อมูลชุดทดสอบ ทำให้ผลการจำแนกประเภทนั้นดีเกินจริง ดังนั้นจะมีการคิดวิธี k-fold cross validation ขึ้นมาแก้ปัญหา

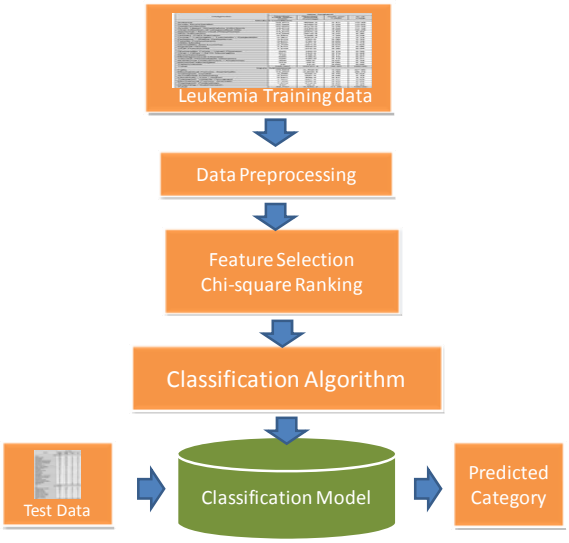
การแบ่งข้อมูลแบบ K - fold cross-validation คือการแบ่งข้อมูลออกเป็น K ชุดเท่าๆกัน และทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก K - 1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ดังตัวอย่างต่อไปนี้ K - fold Cross Validation (K = 5) ชุดข้อมูลหลังจากทำการแบ่งออกเป็น 5 ชุดข้อมูลย่อยเท่าๆกัน โดยแต่ละกล่องคือชุดข้อมูลย่อย 1 ชุด [6-8] ดังภาพ



ภาพที่ 2 K - fold cross-validation

3.วิธีการดำเนินการวิจัย

งานวิจัยนี้ ทำการทดลองการลดมิติของยีนร่วมกับอัลกอริทึมในการจำแนกประเภทเรียนรู้แบบมีผลเฉลย (Supervised Learning) โดยมุ่งเน้นจำแนกประเภทของมะเร็งเม็ดเลือดขาว โดยทดสอบกับกลุ่มตัวอย่างข้อมูลยีนของโรคมะเร็งเม็ดเลือดขาวแบบเฉียบพลัน (Acute Leukemia) จาก Pablo de Olavide University of Seville ซึ่งข้อมูลชุดนี้ประกอบด้วย 2 ประเภท คือ ALL และ AML ซึ่งมีจำนวนมิติของข้อมูล 7,130 มิติ จำนวน 72 ตัวอย่าง เป็นกลุ่มตัวอย่างเรียนรู้ และทำการทดสอบด้วยวิธี 10-fold cross validation โดยแบบจำลองที่นำเสนอในงานวิจัยนี้ จะทำการลดขนาดมิติของข้อมูลด้วยค่าสถิติไคแสควร์ (Chisquare) โดยพิจารณาเรียงลำดับจากค่าสูงสุด ซึ่งผลที่ได้จากการลดขนาดของมิติดังกล่าว จะถูกส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลยด้วยอัลกอริทึม ต้นไม้ตัดสินใจ (Decision Tree) เนอีฟเบย์ (Naïve-Bayes) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) เคเนียเรสเนเบอร์ (K-Nearest Neighbor) มาเรียนรู้ ทำการทดสอบเปรียบเทียบประสิทธิภาพด้านความถูกต้อง ความแม่นยำ โดยวัดที่ประสิทธิผลของการจำแนกตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) และค่า F-Measurement



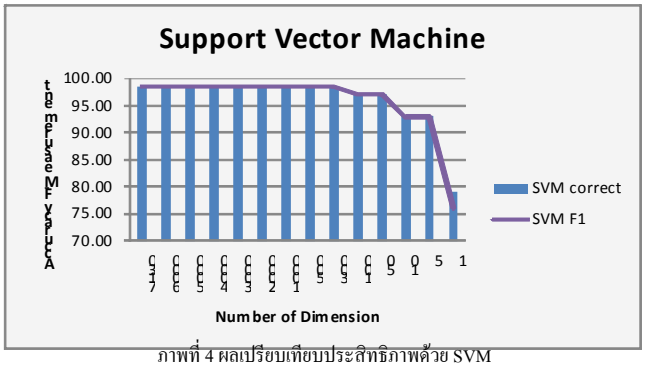
ภาพที่ 3 ขั้นตอนวิธีจำแนกประเภทของมะเร็งเม็ดเลือดขาว

4.ผลการทดลอง

การทดลองค้นหากลุ่มย่อยของยีน ที่มีอำนาจในการจำแนกประเภทของมะเร็งเม็ดเลือดขาว ด้วยการลดคุณลักษณะด้วยวิธี Chi-square ร่วมกับอัลกอริทึมเครื่องจักรเรียนรู้ (Machine learning) เครื่องมือที่ใช้ในการทดสอบอัลกอริทึมในการสังเคราะห์โมเดลโดยใช้ซอฟต์แวร์ Weka version 3.6 ซึ่งประกอบด้วยอัลกอริทึม ซัพพอร์ทเวกเตอร์แมชชีน(Support Vector Machine) ใช้เคอร์เนลฟังก์ชั่นแบบ Linear ต้นไม้ตัดสินใจ (Decision Tree) และเนอีฟเบย์ (Naive-Bayes) ใช้พารามิเตอร์ดั้งเดิม (Default parameter) เคเนียเรสเนเบอร์ (K-Nearest Neighbor)ใช้ 3-Nearest Neighbor ซึ่งได้ผลการทดลองดังนี้

ตารางที่ 1 ผลเปรียบเทียบประสิทธิภาพด้วย SVM

	Support Vector Machine			
Dimension	Correct	Precision	Recall	F1
7130	98.61	98.60	98.60	98.60
6000	98.61	98.60	98.60	98.60
5000	98.61	98.60	98.60	98.60
4000	98.61	98.60	98.60	98.60
3000	98.61	98.60	98.60	98.60
2000	98.61	98.60	98.60	98.60
1000	98.61	98.60	98.60	98.60
500	98.61	98.60	98.60	98.60
300	**98.61**	**98.60**	**98.60**	**98.60**
100	97.22	97.20	97.20	97.20
50	97.22	97.20	97.20	97.20
10	93.06	93.20	93.10	92.90
5	93.06	93.20	93.10	92.90
1	79.17	84.20	79.20	76.10

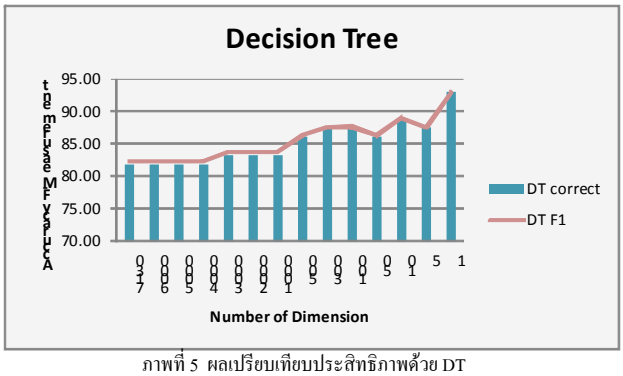


ภาพที่ 4 ผลเปรียบเทียบประสิทธิภาพด้วย SVM

ผลทดลองโดยใช้วิธี Chi-square ในการลดคุณลักษณะและทำการเรียนรู้ด้วยอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน(Support Vector Machine) พบว่าเมื่อใช้จำนวนยีนของข้อมูลทั้งหมด 7,130 มิติ มาทดสอบให้ค่าความถูกต้อง (Accuracy) เท่ากับ 98.61 % และค่า F-Measurement เท่ากับ 98.60 % แต่เมื่อทดสอบลดมิติข้อมูลลงเหลือ 300 มิติ ยังคงให้ค่าความถูกต้อง (Accuracy) เท่ากับ 98.61 % และค่า F-Measurement เท่ากับ 98.60 % เท่ากับข้อมูลต้นฉบับที่ยังไม่ลดมิติข้อมูล สามารถลดมิติของข้อมูลได้ถึง 95.79% (7130-300 / 7130) โดยไม่กระทบต่อค่าความถูกต้องแต่อย่างใด แต่เมื่อลดมิติลงมากกว่าระดับดังกล่าวพบว่า ค่าความถูกต้องและค่า F-Measurement เริ่มถดถอยลงตามจำนวนของมิติข้อมูลที่ลดลงตามลำดับ

ตารางที่ 2 ผลเปรียบเทียบประสิทธิภาพด้วย DT

	Decision Tree			
Dimension	Correct	Precision	Recall	F1
7130	81.94	83.40	81.90	82.30
6000	81.94	83.40	81.90	82.30
5000	81.94	83.40	81.90	82.30
4000	81.94	83.40	81.90	82.30
3000	83.33	85.10	83.30	83.70
2000	83.33	85.10	83.30	83.70
1000	83.33	85.10	83.30	83.70
500	86.11	87.10	86.10	86.30
300	87.50	88.10	87.50	87.60
100	87.50	88.80	87.50	87.70
50	86.11	87.10	86.10	86.30
10	88.89	89.80	88.90	89.10
5	87.50	88.10	87.50	87.60
1	**93.06**	**93.10**	**93.10**	**93.00**

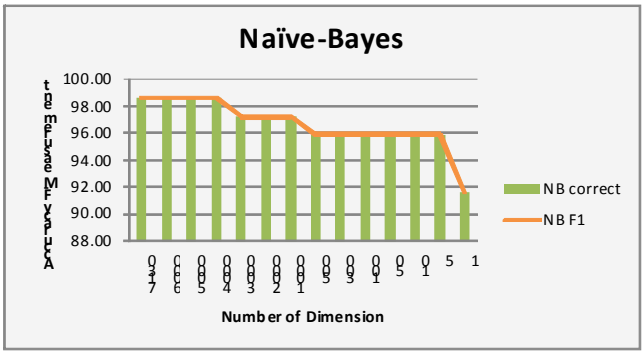


ภาพที่ 5 ผลเปรียบเทียบประสิทธิภาพด้วย DT

ผลทดลองโดยใช้วิธี Chi-square ในการลดคุณลักษณะและทำการเรียนรู้ด้วยอัลกอริทึม ต้นไม้ตัดสินใจ (Decision Tree) พบว่า เมื่อใช้จำนวนยีนของข้อมูลทั้งหมด 7,130 มิติ มาทดสอบให้ค่าความถูกต้อง (Accuracy) เท่ากับ 81.94 % และค่า F-Measurement เท่ากับ 82.30 % แต่ เมื่อทดสอบการลดมิติข้อมูลลงเหลือ 1 มิติ กลับให้ค่าความถูกต้อง (Accuracy) สูงขึ้นเท่ากับ 93.06 % และค่า F-Measurement สูงขึ้นเท่ากับ 93.00 % ซึ่งค่าความถูกต้องดังกล่าวสูงกว่าเรียนรู้ด้วยมิติของข้อมูลต้นฉบับทั้งหมดถึง 11.12% และค่า F-Measurement สูงกว่าถึง 10.7% และสามารถลดมิติของข้อมูลได้ถึง 99.98% (7130-1 / 7130)

ตารางที่ 3 ผลเปรียบเทียบประสิทธิภาพด้วย NB

	Naïve-Bayes			
Dimension	Correct	Precision	Recall	F1
7130	98.61	98.60	98.60	98.60
6000	98.61	98.70	98.60	98.60
5000	98.61	98.70	98.60	98.60
4000	**98.61**	**98.70**	**98.60**	**98.60**
3000	97.22	97.20	97.20	97.20
2000	97.22	97.20	97.20	97.20
1000	97.22	97.20	97.20	97.20
500	95.83	95.90	95.80	95.90
300	95.83	95.90	95.80	95.90
100	95.83	95.90	95.80	95.90
50	95.83	95.90	95.80	95.90
10	95.83	95.90	95.80	95.90
5	95.83	95.90	95.80	95.90
1	91.67	91.70	91.70	91.60



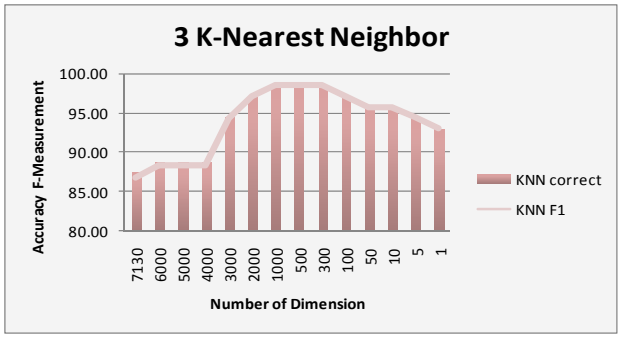
ภาพที่ 6 ผลเปรียบเทียบประสิทธิภาพด้วย NB

ผลทดลองโดยใช้วิธี Chi-square ในการลดคุณลักษณะและทำการเรียนรู้ด้วยอัลกอริทึม เนอีฟเบย์ (Naïve-Bayes) พบว่าเมื่อใช้จำนวนยีนของข้อมูล

ทั้งหมด 7,130 มิติ มาทดสอบให้ค่าความถูกต้อง (Accuracy) เท่ากับ 98.61 % และค่า F-Measurement เท่ากับ 98.60 % แต่เมื่อทดสอบลดมิติข้อมูลลงเหลือ 4000 มิติ ยังคงให้ค่าความถูกต้อง (Accuracy) เท่ากับ 98.61 % และค่า F-Measurement เท่ากับ 98.60 % เท่ากับข้อมูลต้นฉบับที่ยังไม่ลดมิติข้อมูล โดยสามารถลดมิติของข้อมูลได้ถึง 43.89% (7130-4000 / 7130) โดยไม่กระทบต่อค่าความถูกต้องแต่อย่างใด แต่เมื่อลดมิติลงมากกว่าระดับดังกล่าวพบว่า ค่าความถูกต้องและค่า F-Measurement เริ่มลดถอยลงตามจำนวนของมิติข้อมูลที่ลดลงตามลำดับ

ตารางที่ 4 ผลเปรียบเทียบประสิทธิภาพด้วย KNN

	K-Nearest Neighbor			
Dimension	**Correct**	**Precision**	**Recall**	**F1**
7130	87.50	89.50	87.50	86.70
6000	88.89	90.50	88.90	88.30
5000	88.89	90.50	88.90	88.30
4000	88.89	90.50	88.90	88.30
3000	94.44	94.90	94.40	94.30
2000	97.22	97.30	97.20	97.20
1000	98.61	98.60	98.60	98.60
500	98.61	98.60	98.60	98.60
300	**98.61**	**98.60**	**98.60**	**98.60**
100	97.22	97.20	97.20	97.20
50	95.83	95.80	95.80	95.80
10	95.83	95.80	95.80	95.80
5	94.44	94.44	94.44	94.44
1	93.06	93.10	93.10	93.10



ภาพที่ 7 ผลเปรียบเทียบประสิทธิภาพด้วย KNN

ผลทดลองโดยใช้วิธี Chi-square ในการลดคุณลักษณะและทำการเรียนรู้ด้วยอัลกอริทึม เคเนียเรสเนเบอร์ (K-Nearest Neighbor) พบว่า เมื่อใช้จำนวนยีนของข้อมูลทั้งหมด 7,130 มิติ มาทดสอบให้ค่าความถูกต้อง (Accuracy) เท่ากับ 87.50 % และค่า F-Measurement เท่ากับ 86.70 % แต่เมื่อทดสอบการลดมิติข้อมูลลงเหลือ 300 มิติ กลับให้ค่าความถูกต้อง (Accuracy) สูงขึ้นเท่ากับ 98.61 % และค่า F-Measurement สูงขึ้นเท่ากับ 98.60 % ซึ่งค่าความถูกต้องดังกล่าวสูงกว่าเรียนรู้ด้วยมิติของข้อมูลต้นฉบับทั้งหมดถึง 11.11% และค่า F-Measurement สูงกว่าถึง 11.9% และสามารถลดมิติของข้อมูลได้ถึง 95.79% (7130-300 / 7130)

5.สรุปผลการวิจัย

งานวิจัยนี้ได้เสนอวิธีการลดมิติของยีนด้วยค่าสถิติไคแสควร์ (Chisquare) และทำการทดสอบประสิทธิภาพการจำแนกประเภทของมะเร็งเม็ดเลือดขาวกับอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ต้นไม้ตัดสินใจ (Decision Tree) เนอีฟเบย์ (Naïve-Bayes) เคเนียเรสเนเบอร์ (K-Nearest Neighbor) โดยใช้วิธีการลดคุณลักษณะร่วมกับอัลกอริทึมเครื่องจักรการเรียนรู้ เพื่อศึกษาวิธีการที่มีประสิทธิภาพในการจำแนกประเภทของมะเร็งเม็ดเลือดขาว จากการทดลองพบว่าการลดมิติของข้อมูลด้วยวิธีไคแสควร์ แล้วส่งเข้าเครื่องจักรการเรียนรู้และวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) และจำนวนมิติของข้อมูลที่ดีที่สุดพบว่า อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน(Support Vector Machine) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ที่จำนวน 300 มิติ ค่าความถูกต้อง (Accuracy) เท่ากับ 98.61 % เท่ากัน อัลกอริทึม เนอีฟเบย์ (Naïve-Bayes) ที่จำนวน 4000 มิติ ค่าความถูกต้อง (Accuracy) เท่ากับ 98.61 % อัลกอริทึม ต้นไม้ตัดสินใจ (Decision Tree) ที่จำนวน 1 มิติ ค่าความถูกต้อง (Accuracy) เท่ากับ 93.06 %

เมื่อวัดประสิทธิภาพจากค่า F-Measurement ที่ดีที่สุด พบว่าอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน(Support Vector Machine) เนอีฟเบย์ (Naïve-Bayes) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor)ให้ประสิทธิภาพเท่ากันคือ 98.60 % ส่วนต้นไม้ตัดสินใจ (Decision Tree) ประสิทธิภาพเท่ากับ 93.00 % ตามลำดับ

ผลการทดลองจากงานวิจัยนี้ พบว่าเราสามารถลดมิติของข้อมูลของยีนมะเร็งเม็ดเลือดขาวด้วยค่าสถิติไคแสควร์ (Chisquare) ลงได้มากถึง 90% ขึ้นไป โดยไม่กระทบต่อประสิทธิภาพในการจำแนกข้อมูลแต่อย่างใด แต่สามารถลดทรัพยากรของระบบและลดระยะเวลาในการประมวลผลได้เป็นอย่างมาก ผลจากการวิจัยนี้สามารถนำขั้นตอนวิธีที่นำเสนอไปประยุกต์ใช้ในการลดมิติของข้อมูลกับฐานข้อมูลทางการแพทย์อื่นๆ เพื่อค้นหามิติของข้อมูลที่มีอำนาจจำแนกสูง และแก้ปัญหาข้อมูลที่รบกวนหรือตัวแปรที่ไม่เกี่ยวข้องในการจำแนกข้อมูลได้

เอกสารอ้างอิง

[1] มะเร็งเม็ดเลือดขาวชนิดเฉียบพลัน ศูนย์มะเร็งโรงพยาบาลสงขลานครินทร์ http://medinfo2.psu.ac.th/cancer

[2] ภัทราวุฒิ แสงศิริ ศจีมาจ ณ วิเชียร และพยุง มีสัจ “การคัดแยกประเภทของมะเร็งเม็ดเลือดขาว โดยใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน”. การประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาครั้งที่ 11 ปี 2553

[3] มะลิกา วัฒนะ. “การศึกษาคุณลักษณะทางกายภาพของมะเร็งเม็ดเลือดขาวเฉียบพลัน”.วิทยานิพนธ์ปริญญาวิทยาศาสตร์มหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ บัณฑิตวิทยาลัยมหาวิทยาลัยขอนแก่น. ปี 2549

[4] Xin Jin, Anbang Xu, Rongfang Bie and Ping Guo “Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles”. Lecture Notes in Computer

Science,Springer Berlin / Heidelberg,Volume 3916/2006
[5] Cheng-San, Y., C. Li-Yeh, et al. “A hybrid approach for selecting gene subsets using gene expression data”. Soft Computing in Industrial Applications, SMCia '08. IEEE Conference 2008.
[6] นิเวศ จิระวิชิตชัย ปริญญา สงวนสัตย์ และพยุง มีสัจ. “การศึกษาทดลองเทคนิคการลดคุณลักษณะและอัลกอริทึมการจัดหมวดหมู่ของเอกสารภาษาไทย ”. **วารสารวิทยาศาสตร์ลาดกระบัง** ปี 2553.
[7] Jiawei Han and Micheline Kamber. “Data Mining Concepts and Techniques”. San Francisco. Morgan Kaufmann Publishers. 2006.
[8] Ian H. Witten, Eibe Frank. “Data mining : practical machine learning tools and techniques”. Amsterdam , Boston : Morgan Kaufman, 2005.
[9] นิเวศ จิระวิชิตชัย ปริญญา สงวนสัตย์ และพยุง มีสัจ “การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ” การประชุมวิชาการระดับชาติ สถาบันบัณฑิตพัฒนบริหารศาสตร์ ปี 2553