

การเรียนรู้รูปแบบรหัสพันธุกรรมเพื่อจำแนกชนิดของโรคมะเร็ง เม็ดเลือดขาวด้วยอัลกอริทึม C5.0

Leukemia Bioinformation Recognition by C5.0

อุบลวรรณ กิจคณะ (Ubonwan Kidkana)* ดร.วนิดา แก่นอากาศ (Dr. Wanida Kanarkard)**
วิโรจน์ ทวีปวรเดช (Wiroj Taweepworadej)**

บทคัดย่อ

บทความนี้นำเสนอการทดสอบการวิเคราะห์ข้อมูลโรคมะเร็งเม็ดเลือดขาวจากปัจจัยเสี่ยงของการเกิดโรคด้วยข้อมูลรหัสพันธุกรรม โดยใช้ข้อมูลโรคมะเร็งเม็ดเลือดขาว 3 ชนิด ได้แก่ ชนิด ALL AML และ MLL ร่วมกับข้อมูลที่เป็นยีนปกติ (Normal) ที่มีคุณสมบัติ 12,583 ลักษณะ จำนวน 85 ตัวอย่าง ทำการประมวลผลด้วยอัลกอริทึม C5.0 อัลกอริทึม Simple k-mean และอัลกอริทึม Cart เพื่อศึกษาพฤติกรรมและเปรียบเทียบประสิทธิภาพของการคัดแยกข้อมูล ผลการศึกษาพบว่าอัลกอริทึม C5.0 ให้ค่าความถูกต้อง 98.82 % และใช้ระยะเวลาในการประมวลผลน้อยกว่า เมื่อเปรียบเทียบกับอัลกอริทึม Simple k-mean และ อัลกอริทึม Cart

ABSTRACT

This paper proposes the testing of leukemia data analysis of the disease risk factor presence in the DNA by using 3 types of leukemia data, which includes ALL AML and MLL, together with the normal genetic data. The experiment has tested on totally 85 genetic data sets which each set contains 12,583 attributes. The C5.0 algorithm, the Simple k-mean and the Cart algorithm are used as the analytical testing methods to study the behavior and compare the efficiency of data classification. The results show that the C5.0 algorithm has the most accuracy at 98.82 % and requires less processing time comparing to the Simple k-mean algorithm and the Cart algorithm.

คำสำคัญ : การคัดแยกประเภท อัลกอริทึม C5.0 โรคมะเร็งเม็ดเลือดขาว

Key Words : Classification, C5.0 algorithm, Leukemia

* มหบัณฑิต หลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น

** รองศาสตราจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น

บทนำ

ในปัจจุบันมีโรคภัยไข้เจ็บต่าง ๆ เกิดขึ้นมากมาย ซึ่งโรคแต่ละชนิดได้พัฒนาความซับซ้อนและเพิ่มระดับความรุนแรงของเชื้อโรค ส่งผลให้การติดเชื้อโรคสามารถเกิดขึ้นได้จากหลายปัจจัย และนำไปสู่การเสียชีวิตในที่สุด แพทย์ผู้ให้การรักษาจะต้องทำการตรวจวินิจฉัยโรคโดยคำนึงถึงความถูกต้องแม่นยำ โรคมะเร็งเม็ดเลือดขาวเป็นอีกโรคหนึ่งที่จำนวนผู้ป่วยมีแนวโน้มเพิ่มขึ้นในแต่ละปี มูลเหตุของโรคมะเร็งเม็ดเลือดขาวมาจากหลายปัจจัยเสี่ยง เช่น ปัจจัยที่เกิดจากความผิดปกติทางพันธุกรรม

ในทางการแพทย์ กระบวนการและขั้นตอนในการตรวจวินิจฉัยปัจจัยที่เกิดจากความผิดปกติทางพันธุกรรมในโรคมะเร็งเม็ดเลือดขาวมีความซับซ้อน ทำได้ยาก และสิ้นเปลืองค่าใช้จ่าย โดยการตรวจวินิจฉัยต้องอาศัยข้อมูลที่ใช้ในการวิเคราะห์จำนวนมาก เพื่อหาความผิดปกติที่เกิดขึ้นของผู้ป่วยแต่ละคน และทำการคัดแยกผู้ป่วยตามประเภทของโรค เพื่อให้ผู้ป่วยได้รับการรักษาอย่างถูกต้องตรงเป้าหมายที่สุด

ปัจจุบันมีการศึกษาเกี่ยวกับข้อมูลรหัสพันธุกรรมของโรคอย่างกว้างขวาง เนื่องจากข้อมูลที่ได้จากการแสดงออกในระดับยีนมีให้ศึกษาเป็นจำนวนมาก ผู้ทวิวิจัยได้พยายามนำเทคนิคต่าง ๆ มาปรับปรุงและพัฒนาร่วมกับอัลกอริทึมที่ใช้ในการคัดแยกข้อมูลให้สามารถทำงานร่วมกันได้เป็นอย่างดี ทำให้รูปแบบของอัลกอริทึมที่ใช้ในการประมวลผลข้อมูลมีลักษณะแตกต่างกันทั้งแบบต้นไม้อการตัดสินใจ การใช้โครงข่ายประสาทเทียม และการใช้กฎการคัดแยกข้อมูล

ในศาสตร์ด้านปัญญาประดิษฐ์ อัลกอริทึม C5.0 เป็นวิธีการคัดแยกข้อมูลที่น่าสนใจวิธีหนึ่ง แนวคิดในการคัดแยกข้อมูลของอัลกอริทึม C5.0 อาศัยเกณฑ์การคัดแยกจากคุณลักษณะต่าง ๆ ของข้อมูล และแสดงผลลัพธ์ที่ง่ายต่อการวิเคราะห์ด้วยรูปแบบต้นไม้อการตัดสินใจ ซึ่งเป็นคุณลักษณะเด่นของอัลกอริทึม C5.0 ที่แตกต่างจากการทำงานของอัลกอริทึมอื่น

โดยอัลกอริทึม C5.0 ได้ถูกนำไปเปรียบเทียบกับอัลกอริทึมอื่นในหลายงานวิจัย เช่น งานวิจัยของ Sun et al. (2007) ได้ศึกษาการคัดแยกข้อมูลโรคมะเร็งเม็ดเลือดขาวด้วยอัลกอริทึม C5.0 เพื่อเปรียบเทียบประสิทธิภาพความถูกต้องกับอัลกอริทึม K-NN และอัลกอริทึม Navie bayes เป็นต้น

งานวิจัยนี้จึงมีแนวคิดในการนำอัลกอริทึม C5.0 มาช่วยแก้ปัญหาของกระบวนการวิเคราะห์ข้อมูลรหัสพันธุกรรมของโรคมะเร็งเม็ดเลือดขาว 3 ชนิด คือ Acute Lymphocytic Leukemia (ALL), Acute Myeloid Leukemia (AML) และ Mixed Lineage Leukemia (MLL) โดยวิเคราะห์ร่วมกับข้อมูลยีนปกติ โดยทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึม C5.0 กับอัลกอริทึม Simple k-mean และอัลกอริทึม Cart ซึ่งผลลัพธ์จากการวิเคราะห์ข้อมูลเป็นประโยชน์ในการช่วยคัดกรองผู้ป่วยในเบื้องต้นให้กับแพทย์ และนำไปสู่การรักษาที่ดีต่อไป

วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนากระบวนการวิเคราะห์ข้อมูลรหัสพันธุกรรมในโรคมะเร็งเม็ดเลือดขาวที่มีจำนวนมากให้เกิดความถูกต้องในการคัดแยกข้อมูลด้วยอัลกอริทึม C5.0
2. เพื่อเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึม C5.0 กับวิธีการคัดแยกด้วยอัลกอริทึม Simple k-mean และ อัลกอริทึม Cart

อัลกอริทึม C5.0

อัลกอริทึม C5.0 เป็นอัลกอริทึมที่ใช้เป็นตัวคัดแยกในรูปแบบของการสร้างต้นไม้พัฒนาจากอัลกอริทึม C4.5 ที่ทำการเพิ่มความถูกต้องของตัวคัดแยก ด้วยการใช้วิธี Boosting ซึ่งเป็นการปรับค่าน้ำหนักของตัวอย่างที่เป็นข้อมูลในการฝึกสอน โดยตัวอย่างใดมีความแม่นยำมากอยู่แล้วจะลดค่าน้ำหนักลงและไปเพิ่มค่าน้ำหนักให้ตัวอย่างที่มีความแม่นยำน้อย อัลกอริทึม C5.0 ได้รวมเอาฟังก์ชันต่าง ๆ

เช่น ค่าความแปรปรวนของการคัดแยกผิด (variable misclassification costs) ซึ่งจะมีการแยกค่าสูญหายในการคัดแยกของกลุ่มข้อมูลทำให้ลดการเกิดความแปรปรวนของการคัดแยกผิดได้ ในอัลกอริทึม C5.0 ยังให้ความสำคัญในการกระจายน้ำหนักบนกลุ่มข้อมูล เช่น ในงานวิจัยของ Berrar *et al.* (2001) และ Ménard *et al.* (2006) กล่าวถึงการกำหนดค่าน้ำหนักอัลกอริทึม C5.0 ในการประมวลผลเป็นค่าเดียวกัน แต่ถ้าเกิดกรณีที่มีการคัดแยกผิดก็จะมีค่าเพิ่มค่าน้ำหนักให้สูงขึ้นและทำการคัดแยกข้อมูลใหม่อีกครั้ง นอกจากนี้ อัลกอริทึม C5.0 ได้เพิ่มการรองรับชนิดข้อมูลใหม่ ๆ ด้วย ซึ่งทำให้การประยุกต์ใช้งานอัลกอริทึม C5.0 สามารถทำได้ง่ายขึ้น

ลำดับขั้นตอนการทำงานของอัลกอริทึม C5.0 (Govindarajan, 2007) แสดงได้ดังนี้

1. เตรียมข้อมูลที่จะศึกษา โดยการจัดเรียงข้อมูลเป็นแถวและคอลัมน์ ซึ่งแต่ละแถวเป็นค่าของข้อมูลที่น่ามาพิจารณาและแต่ละคอลัมน์เป็นชื่อคุณสมบัติของข้อมูลนั้น ๆ
2. สำหรับแต่ละคุณสมบัติ หาค่า Information Gain ของแต่ละคุณสมบัติเพื่อเป็นการจัดลำดับของความสำคัญของคุณสมบัติ
3. สร้างไหนดรากจากค่า Information Gain ที่มากที่สุด
4. ทำการสร้างกิ่งจากไหนดรากให้เป็นไหนดลูก (Children Node)
5. เมื่อไม่สามารถสร้างกิ่งต่อไปได้อีกจะได้ไหนดสุดท้ายเป็นผลลัพธ์ หรือ ไหนดใบ

Frey *et al.* (2003) กล่าวถึงการสร้างต้นไม้ด้วยการใช้ Gain Ratio ดังสมการการวัดค่าความไม่บริสุทธิ์ของข้อมูล (Entropy)

$$E(S) = - \sum_{c=1}^N p(S_c) \times \log_2 p(S_c) \quad (1)$$

จากสมการ (1) ค่า c_1 ถึง c_n เป็นกลุ่มข้อมูลย่อยของ S และ $p(S_c)$ คือ ความน่าจะเป็นของกลุ่มข้อมูล C

ที่เกิดขึ้นใน S โดยการวัดค่า Information Gain เพื่อเป็นการสร้างลำดับทำได้ด้วยสมการ

$$Gain(S, V) = E(S) - \sum_{v \in \text{values}(V)} \frac{|S_v|}{|S|} \times E(S_v) \quad (2)$$

ค่า Gain ที่ได้มีค่าอคติ (bias) ที่เป็นส่วนหนึ่งในการแบ่งแยกกลุ่มข้อมูล การลดค่าอคติสามารถทำได้โดยคำนวณดังสมการ

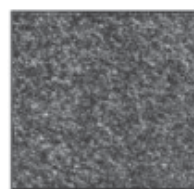
$$SplitInfo(S, V) = \sum_{i=1}^m - \frac{|S_i|}{|S|} \times \log_2 \frac{|S_i|}{|S|} \quad (3)$$

การหาค่า Gain Ratio เป็นการแบ่งค่า Gain ด้วยค่าการ Split Info ดังสมการ

$$GainRatio(S, V) = \frac{Gain(S, V)}{SplitInfo(S, V)} \quad (4)$$

ข้อมูลโรคมะเร็งเม็ดเลือดขาว (Leukemia)

คุณสมบัติของข้อมูลที่ใช้การทดสอบได้จากการใช้ Affymetrix GeneChip ทำให้ได้ข้อมูลเป็น Microarray ที่มีค่าตัวเลขแปลงจากภาพแสดงดังภาพที่ 1 ภาพ 1 (ก) เป็นตัวอย่างของข้อมูลที่ได้จาก Affymetrix ความสว่างของจุดแสดงถึงความหนาแน่นของการแสดงออกจากรหัส แล้วถูกแปลงด้วย Microarray ไปเป็นตัวเลขได้ดังภาพ 1 (ข)



(ก)

Gene	Value
D26018_at	193
D26067_at	-70
D26068_at	144
D26129_at	33

(ข)

ภาพที่ 1 ค่าที่ได้จาก Affymetrix GeneChip (Piatetsky-Shapiro, 2003)

ข้อมูลโรคมะเร็งเม็ดเลือดขาวมีลักษณะของข้อมูลที่ใช้ในการทดสอบ ดังนี้

1. ชนิดข้อมูลเป็นแบบ Multivariate
2. จำนวนคุณสมบัติ (attribute) ทั้งหมดที่ใช้มี 12,583 คุณสมบัติ
3. คุณสมบัติที่ใช้เป็นคุณสมบัติของการแสดงออกในระดับยีนมีลักษณะเหมือนกับตัวอย่างภาพที่ 1 เช่น Gene D26018_at ใช้เป็นรหัสยีนที่แทนรายละเอียดของยีนนั้น
4. ลักษณะของข้อมูลเป็น Real Data
5. เป็นข้อมูลที่ถูกนำไปใช้ในการคัดแยก

วิธีการทดลอง

แบ่งออกเป็น 2 ขั้นตอนหลัก ดังนี้

1. การทดสอบข้อมูลโรคมะเร็งเต้านม (ข้อมูลมาตรฐาน)

การทดสอบข้อมูลมาตรฐานใช้ข้อมูลของโรคมะเร็งเต้านมซึ่งเป็นข้อมูลที่ถูกยอมรับและนำมาใช้ในการศึกษาจากหลายงานวิจัย และเป็นข้อมูลที่มีคุณสมบัติต่าง ๆ ใกล้เคียงกันกับข้อมูลโรคมะเร็งเม็ดเลือดขาว ในขั้นตอนนี้ นำข้อมูลมาตรฐานมาทดสอบกับอัลกอริทึมทั้งสามวิธี คือ อัลกอริทึม C5.0 อัลกอริทึม Simple k-mean และ อัลกอริทึม Cart โดยปรับค่าสภาวะของอัลกอริทึมในแต่ละวิธี เพื่อให้ได้ค่าที่ถูกต้องเหมาะสมที่สุด เพื่อนำไปใช้ในการทดสอบกับข้อมูลโรคมะเร็งเม็ดเลือดขาวในขั้นตอนที่ 2

2. การทดสอบข้อมูลโรคมะเร็งเม็ดเลือดขาว (ข้อมูลจริง)

ข้อมูลโรคมะเร็งเม็ดเลือดขาวที่ใช้ทดสอบเป็นข้อมูลรหัสพันธุกรรมที่มีคุณสมบัติ 12,583 ลักษณะ จำนวน 72 ตัวอย่าง ซึ่งประกอบด้วยโรคมะเร็งเม็ดเลือดขาวชนิด ALL AML และ MLL นำมารวมกับข้อมูลรหัสพันธุกรรมที่มีสภาพเป็นปกติจำนวน 13 ตัวอย่าง รวมทั้งสิ้น 85 ตัวอย่าง ข้อมูลทั้ง 85 ตัวอย่างถูกปรับรูปแบบการจัดเรียงข้อมูลให้ถูกต้องตามรูปแบบของเครื่องมือที่ใช้ทดสอบ การประมวลผลข้อมูลด้วยอัลกอริทึมทั้งสามวิธีใช้สภาวะของการปรับค่าต่าง ๆ ตามที่ได้จากการทดสอบกับข้อมูล

มาตรฐาน ผลการทดลองที่ได้ถูกนำมาวิเคราะห์โดยเปรียบเทียบประสิทธิภาพในการคัดแยกในแต่ละอัลกอริทึม

ผลการวิจัยและอภิปรายผล

ในการทดสอบข้อมูลมาตรฐานทำการปรับค่าสภาวะต่าง ๆ ในแต่ละอัลกอริทึมเพื่อให้ได้ค่าสภาวะที่เหมาะสมก่อนนำอัลกอริทึมนั้นไปทดสอบกับข้อมูลโรคมะเร็งเม็ดเลือดขาวแสดงค่าสภาวะที่ปรับแล้วดังตารางที่ 1

ตารางที่ 1 การปรับค่าสภาวะต่าง ๆ ของแต่ละอัลกอริทึม

อัลกอริทึม	ค่าสภาวะที่ใช้
C5.0	- Boosting = 10 - Cross Validation = 10 - Expected noise = 30%
K-means	- Euclidean Distance - Max Iterations = 500
Cart	- Cross Validation = 10

โดยการทดสอบข้อมูลมาตรฐานจากค่าสภาวะที่เหมาะสมในแต่ละอัลกอริทึมได้ผลลัพธ์แสดงดังตารางที่ 2 ซึ่งแสดงให้เห็นถึงประสิทธิภาพของการประมวลผลข้อมูลด้วยอัลกอริทึม C5.0 ที่มีความถูกต้องมากที่สุด

ตารางที่ 2 ผลลัพธ์การประมวลผลข้อมูลมาตรฐานในแต่ละอัลกอริทึม

อัลกอริทึม	ความถูกต้อง (%)	ความผิดพลาด (%)
C5.0	99.65	0.35
K-means	92.79	7.21
Cart	93.32	6.68

เมื่อนำอัลกอริทึมทั้งสามแบบที่ได้ปรับค่าสถานะให้ถูกต้องเหมาะสมแล้ว มาประมวลผลข้อมูลโรคมะเร็งเม็ดเลือดขาวที่มีทั้งหมด 12,583 ลักษณะ จำนวน 85 ตัวอย่างด้วยอัลกอริทึม C5.0 จะได้ผลลัพธ์ดังภาพที่ 2

Results for output field CLASS

Comparing \$C-CLASS with CLASS

Correct	84	98.82%
Wrong	1	1.18%
Total	85	

Coincidence Matrix for \$C-CLASS (rows show actuals)

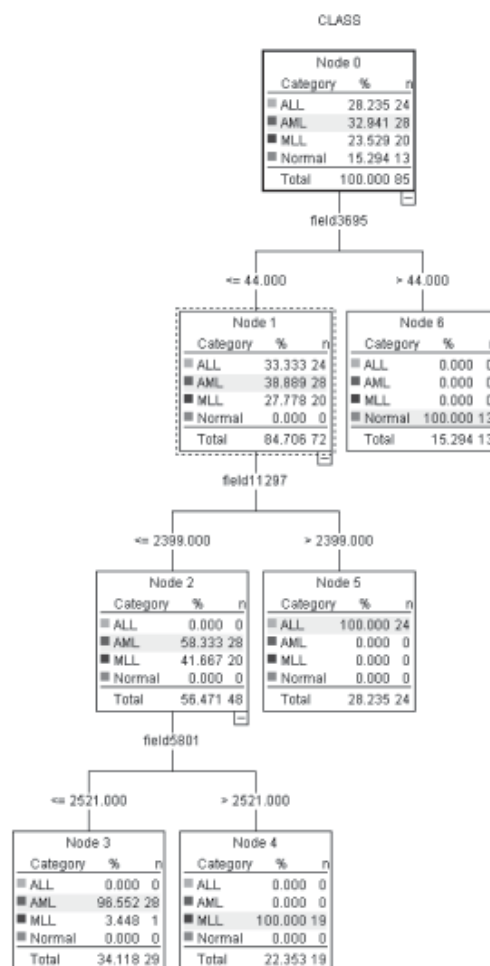
	ALL	AML	MLL	Normal
ALL	24	0	0	0
AML	0	28	0	0
MLL	0	1	19	0
Normal	0	0	0	13

ภาพที่ 2 ผลการคัดแยกข้อมูลโรคมะเร็งเม็ดเลือดขาวจำนวน 85 ตัวอย่าง ด้วยอัลกอริทึม C5.0

จากการทดลองพบว่าจำนวนตัวอย่าง 85 ตัวอย่าง ซึ่งประกอบด้วย ข้อมูลชนิด ALL 24 ตัวอย่าง ชนิด AML 28 ตัวอย่าง ชนิด MLL 20 ตัวอย่าง และ ชนิด Normal 13 ตัวอย่าง อัลกอริทึม C5.0 สามารถคัดแยกกลุ่มข้อมูลได้ผลถูกต้องรวมทั้งหมด 84 ตัวอย่าง คิดเป็นเปอร์เซ็นต์ของความถูกต้อง 98.82 เปอร์เซ็นต์ โดยมีการคัดแยกผิด 1 ตัวอย่าง คิดเป็น 1.18 เปอร์เซ็นต์ ซึ่งเป็นข้อมูลชนิด MLL ที่ถูกคัดแยกไปเป็นชนิด AML

ต้นไม้การตัดสินใจที่ได้จากการทำงานของอัลกอริทึม C5.0 แสดงดังภาพที่ 3 โดยในโหนดที่ 0 แสดงถึงการใช้คุณสมบัติจำนวน 3,695 ลักษณะ เพื่อแยกกลุ่มข้อมูลยีนปกติ (โหนดที่ 6) ออกจากกลุ่มข้อมูลมะเร็งเม็ดเลือดขาว (โหนดที่ 1) ได้ถูกต้องทั้ง 100 เปอร์เซ็นต์ และใช้คุณสมบัติ จำนวน 11,297 ลักษณะในโหนดที่ 1 เพื่อคัดแยกข้อมูลชนิด ALL (โหนดที่ 5) ออกจากกลุ่มข้อมูลโรคมะเร็งเม็ดเลือดขาว ได้ถูกต้อง 100 เปอร์เซ็นต์ และในโหนดที่ 2 ใช้คุณสมบัติจำนวน 5,801 ลักษณะในการแยกข้อมูล

ชนิด MLL (โหนดที่ 4) ได้ถูกต้อง 100 เปอร์เซ็นต์ กับ ชนิด AML (โหนดที่ 3) ที่มีความถูกต้อง 96.55 เปอร์เซ็นต์ โดยมีข้อผิดพลาด 3.45 เปอร์เซ็นต์ ที่เกิดจากการจำแนกข้อมูลชนิด MLL เป็น AML จำนวน 1 ตัวอย่าง



ภาพที่ 3 ต้นไม้การตัดสินใจของการคัดแยกข้อมูลโรคมะเร็งเม็ดเลือดขาวด้วยอัลกอริทึม C5.0

การทดสอบข้อมูลโรคมะเร็งเม็ดเลือดขาวเมื่อนำไปประมวลผลกับอัลกอริทึม Simple k-mean ในแบบ Matrix ได้ผลลัพธ์ดังภาพที่ 4 โดย Cluster 0 เป็นชนิด ALL Cluster 1 เป็นชนิด Normal Cluster 2

เป็นชนิด MLL และ Cluster 3 เป็นชนิด AML สามารถจัดกลุ่มจำนวนตัวอย่างได้ถูกต้องในชนิด ALL ได้ 24 ตัวอย่าง ชนิด Normal ได้ 13 ตัวอย่าง ชนิด MLL ได้ 4 ตัวอย่าง ชนิด AML ได้ 27 ตัวอย่าง และการคัดแยกกลุ่มข้อมูลที่ไม่ถูกต้องอีก 17 ตัวอย่าง โดยคิดเป็นเปอร์เซ็นต์ความถูกต้องได้ 80 เปอร์เซ็นต์ และเปอร์เซ็นต์ความผิดพลาด 20 เปอร์เซ็นต์

```

Class attribute: CLASS
Classes to Clusters:

 0  1  2  3  <-- assigned to cluster
24  0  0  0  | ALL
15  0  4  1  | MLL
 1  0  0 27  | AML
 0 13  0  0  | Normal

Cluster 0 <-- ALL
Cluster 1 <-- Normal
Cluster 2 <-- MLL
Cluster 3 <-- AML

Incorrectly clustered instances :      17.0      20

```

ภาพที่ 4 ผลการคัดแยกข้อมูลโรคมะเร็งเม็ดเลือดขาวจำนวน 85 ตัวอย่าง ด้วยอัลกอริทึม Simple k-mean

```

Correctly Classified Instances      70          82.3529 %
Incorrectly Classified Instances    15          17.6471 %

```

=== Confusion Matrix ===

```

 a  b  c  d  <-- classified as
24  0  0  0  | a = ALL
 0 14  5  1  | b = MLL
 0  9 19  0  | c = AML
 0  0  0 13  | d = Normal

```

ภาพที่ 5 ผลการคัดแยกข้อมูลโรคมะเร็งเม็ดเลือดขาวจำนวน 85 ตัวอย่าง ด้วยอัลกอริทึม Cart

การทดสอบข้อมูลโรคมะเร็งเม็ดเลือดขาวเมื่อนำไปประมวลผลกับอัลกอริทึม Cart ได้ผลลัพธ์ดังภาพที่ 5 โดยอัลกอริทึม Cart สามารถคัดแยกกลุ่มข้อมูลได้ถูกต้อง 70 ตัวอย่าง คิดเป็น 82.35 เปอร์เซ็นต์ และไม่ถูกต้อง 15 ตัวอย่าง คิดเป็น 17.65 เปอร์เซ็นต์ จำนวนตัวอย่างที่คัดแยกได้ถูกต้อง ได้แก่ ชนิด ALL

24 ตัวอย่างชนิด MLL 14 ตัวอย่าง ชนิด AML 19 ตัวอย่าง และชนิด Normal 13 ตัวอย่าง

จากการทดสอบโดยใช้อัลกอริทึมทั้งสามแบบพบว่า ข้อมูลสองชนิดคือ AML และ MLL มีจำนวนในการคัดแยกผิดพลาดมากที่สุด เนื่องจากข้อมูลชนิด MLL เป็นโรคมะเร็งเม็ดเลือดขาวที่เป็นชนิดย่อยของข้อมูลชนิด AML ทำให้รูปแบบของคุณสมบัติของข้อมูลทั้งสองชนิดมีความใกล้เคียงกันมาก

การเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึมทั้งสามแบบคำนึงถึงเวลาที่ใช้ในการวิเคราะห์ข้อมูล ซึ่งเริ่มพิจารณาตั้งแต่เริ่มการประมวลผลข้อมูลด้วยโปรแกรมในแต่ละอัลกอริทึม จนกระทั่งสิ้นสุดการประมวลผล โดยค่าประมาณการทำงานคิดจากค่าเฉลี่ยที่เกิดขึ้นในการทดลอง ผลการทดลองแสดงได้ดังตารางที่ 3

ตารางที่ 3 ค่าประมาณเวลาและผลลัพธ์การประมวลผลข้อมูลโรคมะเร็งเม็ดเลือดขาวด้วยอัลกอริทึมทั้งสามแบบ

อัลกอริทึม	เวลา ประมวลผล (≈ นาที)	ความ ถูกต้อง (%)	ความ ผิดพลาด (%)
C5.0	20	98.82	1.18
k-mean	25	80	20
Cart	30	82.35	17.65

ตารางที่ 3 แสดงให้เห็นว่า การคัดแยกข้อมูลโรคมะเร็งเม็ดเลือดขาวด้วยอัลกอริทึม C5.0 มีความถูกต้องมากที่สุด และใช้เวลาในการประมวลผลน้อยที่สุด เมื่อเทียบกับสองอัลกอริทึม โดยที่อัลกอริทึม Simple k-mean ใช้เวลาในการคัดแยกข้อมูลโดยเฉลี่ยน้อยกว่าอัลกอริทึม Cart แต่อัลกอริทึม Cart ได้ผลความถูกต้องในการคัดแยกข้อมูลมากกว่าอัลกอริทึม Simple k-mean เล็กน้อย

สรุปผลการวิจัย

ค่าความถูกต้องของการประมวลผลข้อมูลโรคมะเร็งเม็ดเลือดขาวที่เป็นข้อมูลจริง มีความสอดคล้องกับค่าความถูกต้องที่ได้จากการประมวลผลข้อมูลโรคมะเร็งเต้านม ที่ใช้เป็นข้อมูลมาตรฐานในการปรับค่าสภาวะของแต่ละอัลกอริทึม โดยการตัดแยกข้อมูลด้วยอัลกอริทึม C5.0 มีประสิทธิภาพในการตัดแยกข้อมูลได้ดีกว่าอัลกอริทึม Simple k-mean และอัลกอริทึม Cart เนื่องจาก อัลกอริทึม C5.0 มีความสามารถในการเรียนรู้ลักษณะของข้อมูลในรูปแบบข้อมูลโรคมะเร็งเม็ดเลือดขาวได้ดีกว่า และอัลกอริทึม C5.0 ยังมีส่วนของการปรับค่าให้อัลกอริทึมด้วยฟังก์ชันการทำงานต่าง ๆ ที่ช่วยเพิ่มความสามารถในการจำแนกคุณสมบัติของข้อมูล

นอกจากนี้ อัลกอริทึม C5.0 ยังมีการแสดงผลลัพธ์ในรูปแบบของต้นไม้การตัดสินใจ ซึ่งแสดงให้เห็นลำดับชั้นในการจำแนกข้อมูล ทำให้จำแนกข้อมูลที่เข้ามาใหม่ได้ง่ายและรวดเร็วขึ้นด้วยการพิจารณาข้อมูลตามลำดับชั้นที่ได้จำแนกข้อมูลไว้แล้ว ซึ่งอัลกอริทึมอื่นไม่สามารถทำได้ และจากการทดลองยังพบว่าอัลกอริทึม C5.0 มีประสิทธิภาพในเรื่องของเวลา โดยอัลกอริทึม C5.0 จะใช้ระยะเวลาในการประมวลผลน้อยกว่าอัลกอริทึม Simple k-mean และอัลกอริทึม Cart อย่างเห็นได้ชัด

ข้อเสนอแนะ

ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลที่มีจำนวนคุณสมบัติมาก การคัดกรองเพื่อคัดเลือกเฉพาะคุณสมบัติที่มีนัยสำคัญมาทำการวิเคราะห์ผลด้วยอัลกอริทึม C5.0 อาจทำให้ผลการตัดแยกข้อมูลด้วยอัลกอริทึม C5.0 มีประสิทธิภาพเพิ่มขึ้น และอาจช่วยลดระยะเวลาในการประมวลผลข้อมูลได้ เช่น นำกระบวนการคัดเลือกยีน ที่พิจารณาหาความสัมพันธ์ของคุณสมบัติข้อมูลมาใช้ในการคัดกรองคุณสมบัติข้อมูลก่อนการประมวลผล เป็นต้น

เอกสารอ้างอิง

- Berrar, D., Granzow, M., Dubitzky, W., Stilgenbauer S., Wilgenbus, K., Döhner, H., Lichter, P., and Eils, R. 2001. New insights in clinical impact of molecular genetic data by knowledge-driven data mining. Proceeding of Second International Conference on Systems Biology. 275–281.
- Frey, L., Fisher, D., Tsamardinos, I., Aliferis, CF., and Statnikov, A. 2003. Identifying Markov Blankets with Decision tree Induction. Proceeding of the third IEEE International Conference on Data Mining. 59–66.
- Govindarajan, M. 2007. Text mining technique for data mining application. World Academy of Science, Engineering and Technology. 264–269.
- M?nard, C., Johann, D., Lowenthal, M., Muanza, T., Sproull, M., Ross, S., Gulley, J., Petricoin, E., Coleman, CN., Whiteley, G., Liotta, L., and Camphausen, K. 2006. Discovering clinical biomarkers of ionizing radiation exposure with serum proteomic analysis. Cancer Research 2006. 66(3):1844–1850.
- Piatetsky-Shapiro, G., Khabaza, T., and Ramaswamy, S. 2003. Capturing Best Practice for microarray gene expression data analysis. Proceeding of the ninth ACM SIGKDD International Conference on Knowledge Discovery and data mining. 407–415.

Sun, L., Miao, D., and Zhang, H. 2007. Gene selection with rough sets for cancer classification. Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007. 3:167-172.