การตรวจสอบความผิดปกติของเครื่องจักรจากแฟ้มล็อก ในกระบวนการผลิตฮาร์ดดิสก์ด้วยต้นไม้ตัดสินใจ

ธนาธาร ภัทรกวิน*[1] ประภาส จงสถิตย์วัฒนา[2]

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพฯ 10330

บทคัดย่อ

การผลิตฮาร์ดดิสก์เป็นอุตสาหกรรมที่สำคัญของประเทศไทย กระบวนการผลิตมีความซับซ้อนสูงและประกอบด้วยเครื่องจักรอัตโนมัติจำนวนมากทำงานการผลิตอย่างต่อเนื่อง เมื่อมีเหตุการณ์ผิดปกติเกิดขึ้นสายการผลิตจะหยุดการทำงานและทีมวิศวกรจะวินิจฉัยโดยระบุและค้นหาแหล่งที่มาของข้อผิดพลาดในหมู่เครื่องจักรเหล่านั้น และแก้ไขให้ถูกต้องอย่างรวดเร็ว ในสายการผลิตเครื่องจักรอัตโนมัติทั้งหมดติดต่อกับเครื่องบริการโดยบันทึกเป็นแฟ้มไว้อย่างต่อเนื่อง วิศวกรใช้แฟ้มบันทึกเหล่านี้ในการวินิจฉัยสาเหตุของข้อผิดพลาด งานวิจัยนี้ใช้วิธีการเรียนรู้ของเครื่องจักรเพื่อแจ้งเหตุการณ์ที่เกิดความผิดปกติในแฟ้มล็อก ผลการทดลองแสดงให้เห็นว่าการตรวจสอบมีความถูกต้องมาก และสามารถช่วยให้ทีมงานวิศวกรดำเนินการวินิจฉัยได้รวดเร็วและมีประสิทธิภาพ

คำสำคัญ: การผลิตฮาร์ดดิสก์ ชุดประกอบหัวอ่าน แฟ้มล็อก การเรียนรู้ของเครื่อง ต้นไม้ตัดสินใจ

* Corresponding author. E-mail: thanatarn.pattarakavin@wdc.com

[1] นิสิตปริญญาโท สาขาวิชาวิศวกรรมซอฟต์แวร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

[2] ศาสตราจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

Detection of Machines Anomaly from Log Files in Hard Disk Manufacturing Process using Decision Trees

Thanatarn Pattarakavin*[1] Prabhas Chongstitvatana[2]
Department of Computer Engineering, Faculty of Engineering,
Chulalongkorn University, Bangkok 10330, Thailand

Abstract

Hard disk manufacturing is an important industry in Thailand. The production line of its manufacturing process is highly complex and consists of hundreds of automated machines running a continuous flow production. When an anomaly event occurs, the production line has to be stopped and the diagnosis engineering team must identify and locate the source of error with the machines and correct it quickly. In an automated production line, all of the machines are monitored and their log files are sent to a server continuously; engineers then use these log files to diagnose the causes of the errors. This work proposes to use the machine learning method to identify the anomalous events in the log files. such as an anomaly caused by engineers, anomaly from systems, and anomaly from softwares. The experimental results showed that accuracy was 100% using 10-Fold validation, so it is very accurate and it can help teams of engineers perform diagnosis quickly and effectively.

Keywords: hard disk manufacturing, head stack assembly, log files, machine learning, decision tree

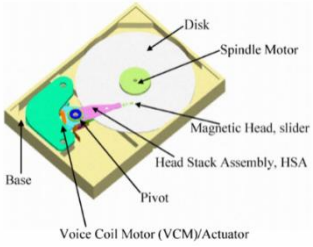
* Corresponding author. E-mail: thanatarn.pattarakavin@wdc.com

[1] Master Student of Software Engineering in Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University

[2] Professor in Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University

1. บทนำ

การผลิตฮาร์ดดิสก์ (รูปที่ 1) เป็นอุตสาหกรรมที่สำคัญในประเทศไทยสร้างรายได้ถึง 15,000 ล้านดอลลาร์สหรัฐของมูลค่าการส่งออกและทำให้เกิดการจ้างงานกว่า 110,000 คนตั้งแต่ปี ค.ศ. 2007 [1] กระบวนการผลิตฮาร์ดดิสก์ประกอบด้วยเครื่องจักรอัตโนมัตินับพันเครื่องทำงานอยู่ในเครือข่ายประสานแน่นทำงานอย่างต่อเนื่อง [2] เครื่องจักรเหล่านี้จะต้องมีการสื่อสารและการทำงานร่วมกันเพื่อให้สายการผลิตที่ทำงานได้อย่างราบรื่น เพื่อตรวจสอบสถานะของเครื่องเหล่านี้ในทันทีประกอบด้วยรายงานจำนวนมากที่สร้างขึ้นอย่างอัตโนมัติจากเครื่องจักรต่าง ๆ ถูกส่งไปยังเครื่องบริการถูกบันทึกไว้เป็นแฟ้มจำนวนมหาศาล โดยแต่ละระเบียนไหลไปยังเครื่องบริการตลอดเวลา เมื่อมีความผิดปกติเกิดขึ้นและสายการผลิตหยุดถือเป็นเรื่องใหญ่ นำมาซึ่งการหาสาเหตุของการหยุดชะงักและแก้ไขอย่างรวดเร็ว โดยทีมวิศวกรวินิจฉัยสาเหตุจากการวิเคราะห์แฟ้มล็อกจำนวนมากเหล่านั้น



รูปที่ 1 ส่วนประกอบของฮาร์ดดิสก์ [2]

รูปที่ 2 ชุดประกอบหัวอ่าน / เขียน (HSA) [1]

แฟ้มล็อกจากเครื่องจักรต่าง ๆ มีรูปแบบที่แตกต่างกัน [3] และมีข้อความที่ไม่สำคัญจำนวนมาก การวิเคราะห์ปัญหาจากเครื่องจักรที่มีความซับซ้อนจำเป็นต้องอาศัยผู้เชี่ยวชาญ ข้อผิดพลาดมากมายเป็นผลจากพนักงานคุมเครื่องจักร ยกตัวอย่างพนักงานคุมเครื่องที่ละเว้นข้อความที่แสดงผลบนหน้าจอสำหรับเครื่องฉายรังสีเอกซ์เครื่องใหม่เนื่องจากมีบ่อยเกินไปและบางข้อความเป็นเรื่องปกติ แต่ในบางข้อความที่ละเลยไปนั้นเกี่ยวข้องกับการตั้งค่าพลังงานรังสีเอกซ์ ส่งผลให้เครื่องส่งพลังงานเต็มกำลัง แทนการบำบัดกลายเป็นฆ่าผู้ป่วย 4 ราย เนื่องจากการละเลยข้อความที่สำคัญมากเหล่านั้น [4]

เครื่องคอมพิวเตอร์สร้างแฟ้มล็อกขนาดใหญ่ขึ้นมามากมายหลายแฟ้มโดยถูกนำมาใช้ในการวิเคราะห์และเฝ้าสังเกตการจัดการระบบ โดยแฟ้มล็อกเหล่านี้อาจมีที่มาจากตั๋วปัญหาต่าง ๆ (trouble tickets) รายงานสถานภาพระบบต่าง ๆ (system status reports) ความขัดข้องของฮาร์ดแวร์ (hardware faults) หรือความผิดพลาดของซอฟต์แวร์ (software errors) แฟ้มล็อกรกไปด้วยข้อความต่าง ๆ ทำให้บดบังผลของปัญหาที่ควรจะถูกค้นพบ รายงาน [3] นำเสนอการสร้างมโนภาพเพื่อกรองสิ่งรบกวนแฟ้มล็อก การใช้เครื่องมือนี้ทำให้พบกลุ่มปัญหาที่รุนแรงที่สุด แนวโน้มของช่วงเวลาที่เกิดปัญหา ปัญหาที่มีที่มาจากตัวประมวลผลปัญหาที่เกิดต่อกันเป็นลำดับต่อกัน และสาเหตุของปัญหา แต่ก็ยังจำเป็นต้องใช้มนุษย์ในการวิเคราะห์

Tingting Gao และคณะ ในปี ค.ศ. 2016 พบว่ากระบวนการผลิตเต็มไปด้วยสิ่งรบกวน แม้แต่ผลลัพธ์ของกระบวนการผลิตก็อาจมารบกวนกระบวนขาเข้าได้เช่นกัน [5] มีงานวิจัยเสนอการตรวจสอบเหตุการณ์ที่เกิดความผิดพลาดในแฟ้มล็อกโดยอัตโนมัติ ระบบสามารถตรวจสอบแฟ้มล็อกได้อย่างทันทีทันใด และแจ้งทีมวิศวกรเมื่อมีความผิดปกติเกิดขึ้น Schuster และคณะ ในปี ค.ศ. 2013 ประยุกต์ใช้การตรวจสอบความผิดปกติในเครือข่ายระบบควบคุม [6] สามารถระบุตำแหน่งและต้นเหตุที่ก่อให้เกิดความผิดปกติ เครื่องมือนี้สามารถช่วยสนับสนุนวิศวกรในการหาสาเหตุที่ส่งผลกระทบต่อสายการผลิตและแก้ไขได้อย่างรวดเร็ว นอกจากนี้เทคนิคการเรียนรู้ของเครื่องยังสามารถเพิ่มพูนประสิทธิภาพของระบบ โดยอาศัยประสบการณ์จริง [7, 8] และยังมีการใช้การวิเคราะห์โดยใช้ต้นไม้ตัดสินใจเพื่อปรับปรุงประสิทธิภาพการผลิต [9] ดังนั้นการเรียนรู้ของเครื่องสามารถแยกความผิดปกติจากสิ่งรบกวน และช่วยลดกำลังคนในงานที่เกี่ยวข้อง

งานของ Schölkopf และคณะ ในปี ค.ศ.2001 [10] ใช้ Supervised Learning สำหรับการรู้จำลายมือว่าเป็นตัวเลขใด ระหว่าง 0 ถึง 9 โดยปรับปรุงผลงานเดิมจากปี ค.ศ. 2000 ในการบอกความผิดปกติหรือ abnormal class และงานของ Volker Roth 2004, 2006 [11, 12] ใช้ Unsupervised Learning ในการรู้จำใบหน้าระหว่างคนใส่แว่นธรรมดากับคนใส่แว่นตาดำ และตรวจความผิดปกติสำหรับ one-class classification

การเรียนรู้ของต้นไม้ตัดสินใจ (Decision Tree) เป็นการเรียนรู้โดยจำแนกประเภท (classification) ข้อมูลออกเป็นคลาส (class) ต่าง ๆ โดยใช้คุณลักษณะ (attribute) ของข้อมูลในการจำแนกประเภท ต้นไม้ตัดสินใจ (Decision Tree) ถูกพัฒนาโดย Ross Quinlan ในชื่อ ID3 algorithm ในปี ค.ศ.1986 [13] พัฒนาต่อเป็น C4.5 ในปี ค.ศ.1993 [14] ต้นไม้ตัดสินใจมีข้อจำกัดคือ ไม่สามารถสร้างต้นไม้ตัดสินใจที่ดีที่สุดให้ครอบคลุมทุกกรณีได้ และไม่สามารถทดสอบผ่านกรรมวิธีกล่องดำ (Black Box Testing) [15]

งานวิจัยที่เรานำเสนอใช้ต้นไม้ตัดสินใจเช่นกัน ข้อมูลที่ใช้ในการสอนนำมาจากสายการผลิตชุดประกอบหัวอ่าน/เขียน (HSA) (รูปที่ 2) โดยใช้แฟ้มล็อกประกอบด้วยข้อมูลจาก 61 ชั่วโมงภายในจากวันที่ 12 ตุลาคม ค.ศ. 2015 ถึงวันที่ 13 มกราคม ค.ศ. 2016 จะขอกล่าวถึงกระบวนการผลิตฮาร์ดิสก์พอเป็นสังเขบ

2. กระบวนการผลิตฮาร์ดดิสก์

กระบวนการผลิตฮาร์ดดิสก์ประกอบไปด้วยเครื่องจักรหลากหลายและสลับซับซ้อน ชุดประกอบหัวอ่าน/เขียน [16, 17] เป็นแขนที่ยึดอยู่ภายในฮาร์ดดิกส์ที่มีการเลื่อนพร้อมกับหัวอ่าน/เขียนที่ติดอยู่ที่ปลาย มีพื้นผิวที่จัดเก็บข้อมูลหลายชุดที่มีหัวอ่าน/เขียนแต่ละชุด ทุกชุดขยับไปพร้อม ๆ กัน การประกอบต้องใช้ทักษะพิเศษโดยผู้ที่ผ่านการฝึกอบรมมาก่อน คนงานจะประจำในแต่ละสถานี ทำการประกอบและตรวจสอบ เครื่องจักรในการประกอบทุกเครื่องเชื่อมต่อกับเซิร์ฟเวอร์กลางที่ตรวจสอบการดำเนินงานทั้งหมด พร้อมกับเก็บบันทึกเหตุการณ์ที่เกิดขึ้นเป็นส่วนหนึ่งของการควบคุมทางสถิติของสายการผลิต

ส่วนประกอบหลักของชุดประกอบหัวอ่าน/เขียน คือ หัวอ่าน/เขียน สไลเดอร์ แอคโชเอเตอร์ พลีเวิร์ด เฟล็คแอสเซมบลี วงจรเฟล็ค แบริ่งและวอยซ์คอยล์ มีเจ็ดขั้นตอนในการประกอบ กระบวนการสามารถอธิบายได้ดังต่อไปนี้ [1]

บรรจุ – ประกอบเฟล็คแอสเซมบลีกับหัวอ่าน/เขียน ใส่รถเข็นรับส่งของโดยพนักงาน 2 คน

ตอกทำแบบ - ยิงลูกเหล็กเพื่อผูกเฟล็คแอสเซมบลีกับหัวอ่าน/เขียนเข้าด้วยกัน

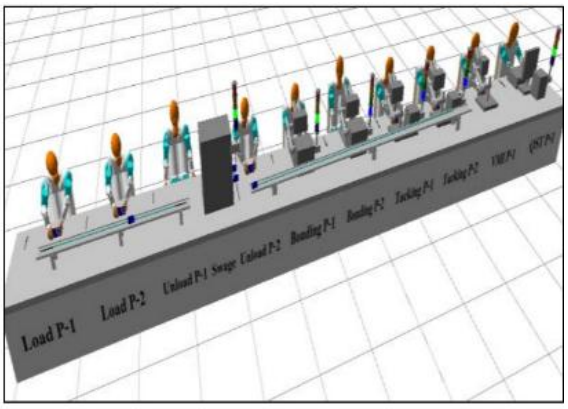
ถ่ายโอน – นำหัวอ่าน/เขียนออกจากรถเข็นรับส่งของ หลังจากนั้นนำหัวอ่าน/เขียนใหม่เข้าประจำที่ เพื่อไหลต่อไปโดยพนักงาน 2 คน

เชื่อม – เชื่อมต่อสายไฟฟ้าระหว่างเฟล็คแอสเซมบลีกับหัวอ่าน/เขียน โดยพนักงาน 2 คน

ตรึง – หยอดกาวอีพ็อกซี่ไปตามแนวยาวของหางหัวอ่าน/เขียน และช่องเสียบ โดยพนักงาน 2 คน

วีเอ็มไอ – การตรวจสอบคุณภาพเพื่อหาข้อบกพร่องทางกายภาพ โดยพนักงาน 1 คน

การทดสอบ - การทดสอบประสิทธิภาพการทำงานของไฟฟ้า ดำเนินการโดยพนักงาน 1 คน

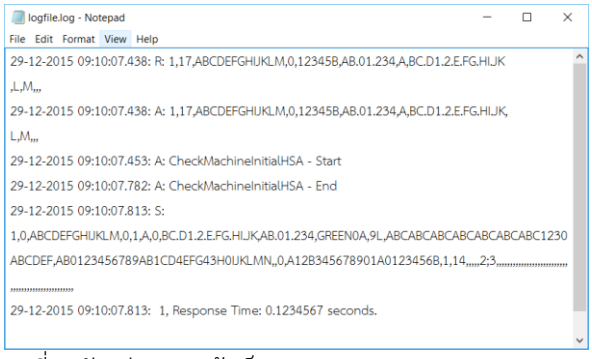


รูปที่ 3 แบบจำลองกระบวนการประกอบชุดประกอบหัวอ่าน/เขียน [1]

3. วิธีวิจัย

มีหลายสาเหตุที่แสดงถึงความผิดปกติของเครื่องจักรในสายการผลิต ที่มีข้อผิดพลาดบางอย่างที่เกิดจากมนุษย์ เช่น พนักงานป้อนรหัสเครื่องที่ไม่ถูกต้องหรือ

สิทธิ์การทำงานหมดอายุ ข้อผิดพลาดส่วนมากจะเกิดจากการเปลี่ยนแปลงในการกำหนดค่าซอฟต์แวร์บนเครื่องบริการและเครื่องจักรควบคุมการผลิต ตัวอย่างเช่น การปรับปรุงซอฟต์แวร์และเฟิร์มแวร์ ข้อผิดพลาดอีกอย่างคือข้อผิดพลาดทางกายภาพ เป็นต้นว่าไฟฟ้าดับและความล้มเหลวทางกลของเครื่องจักรในสายการผลิต ข้อผิดพลาดเหล่านี้สามารถตรวจพบได้ง่ายดาย โดยพนักงานที่เกี่ยวข้อง ส่วนใหญ่ของข้อผิดพลาดที่สำคัญมักถูกส่งออกมาจากซอฟต์แวร์ที่ไม่สามารถเห็นด้วยตา เป็นผลให้ไม่สามารถแก้ไขได้ทันที



รูปที่ 4 ตัวอย่างของแฟ้มล็อก

เนื่องจากข้อมูลในงานจริงมีข้อมูลไม่สมบูรณ์ (incomplete) ข้อมูลรบกวน (noise) ข้อมูลไม่สอดคล้อง (inconsistent) จึงต้องมีการประมวลผลข้อมูลก่อน (preprocessing) เพื่อดึงลักษณะสำคัญ ก่อนจะป้อนเข้าสู่โปรแกรมเรียนรู้ต่อไป แฟ้มล็อกตั้งแต่วันที่ 12 ตุลาคม ค.ศ. 2014 ถึงวันที่ 13 มกราคม ค.ศ.2015 จำนวน 69 แฟ้ม รวมขนาด 74.7 MB ตัดเป็นระเบียนละหนึ่งเหตุการณ์ ข้อมูลหลังทำการประมวลผล ได้ข้อมูล 104,500 ระเบียน ประกอบด้วยลักษณะสำคัญ ทั้งหมด 29 อย่าง ประกอบด้วยข้อมูลชนิด Discrete จำนวน 20 อย่าง และข้อมูลชนิด Numeric จำนวน 9 อย่าง เมตาเดตาอีก 7 อย่าง ใช้ในการวิเคราะห์สาเหตุของการผิดพลาด ข้อมูลแบ่งเป็นสองกลุ่ม เป็นกลุ่มข้อมูลปกติ 104,381 ระเบียนและข้อมูลผิดปกติ 119 ระเบียน ซึ่งถูกระบุโดยผู้เชี่ยวชาญที่รับผิดชอบในการวินิจฉัยของสายการผลิต คุณลักษณะเหล่านี้สามารถครอบคลุมความผิดปกติที่หลากหลาย แต่ไม่ใช่ทั้งหมด การศึกษาถูกจำกัดด้วยปริมาณข้อมูลผิดปกติที่หายาก เพราะเกิดขึ้นน้อยในโรงงานผลิตระดับหกซิกมา นั่นคือมีของเสียไม่เกิน 3.4 ชิ้น ต่อการผลิตหนึ่งล้านชิ้น เช่นโรงงานผลิตฮาร์ดดิสก์

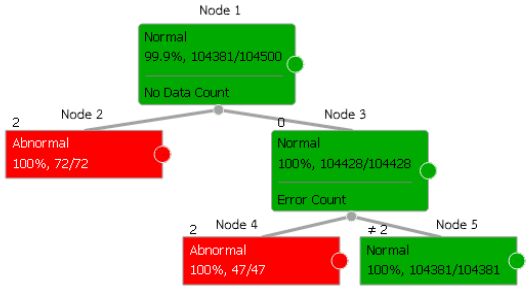
ลักษณะสำคัญประกอบด้วย 3 กลุ่มใหญ่ ๆ กลุ่มที่หนึ่งคือการการบูรณาการข้อมูล (Data Integration) ได้แก่ นับคำ ต่อหนึ่งหน่วยการเปลี่ยนแปลง (Transaction) จากการแจ้งเวลามีสิ่งผิดปกติ ด้วยคำว่า “No Data” “Error” “R: ” “A: ” “S: ” “Response Time” “Cert” “Expire” “Fail” “Abort” “Block” “Validate” “TX” และ “BR” กลุ่มที่สองคือการแปลงข้อมูล (Data Transformation) ค่าต่าง ๆ ที่ตัดออกมาเป็นความหมายที่มนุษย์สามารถเข้าใจได้ ได้แก่ เวลาการตอบสนองต่อหนึ่งหน่วยการเปลี่ยนแปลงข้อความที่ตอบกลับจากระบบจะบันทึกเวลาไว้ด้วย ชั่วโมง นาที วินาที ที่เกิดล็อกดังกล่าว ขนาดแฟ้มล็อก เวลาเริ่มต้น เวลาสิ้นสุด ของแต่ละรายการการเปลี่ยนแปลง เวลาที่ปรับปรุงแฟ้มล็อกครั้งล่าสุด หมายเลขข้อความ ความถูกต้องของลำดับการบันทึก จำนวนบรรทัดต่อรายการเปลี่ยนแปลง เวลาเริ่มต้นที่เริ่มเขียนแฟ้มล็อก กลุ่มที่สามคือการลดข้อมูล (Data Reduction) เช่น ค่าการคำนวนช่วงเวลาของเวลาเริ่มต้นและสิ้นสุดของแต่ละหน่วยการเปลี่ยนแปลง และช่วงเวลาของเวลาสุดท้ายที่แฟ้มล็อกถูกปรับปรุงกับเวลาสุดท้ายของแต่ละการเปลี่ยนแปลง

4. การทดลอง

การทดลองอาศัยเครื่องมือ Orange canvas [18, 19, 20, 21] อันเป็นชุดเครื่องมือสำหรับวิเคราะห์ข้อมูลด้วยการเรียนรู้ของเครื่องและการทำเหมืองข้อมูล และเป็นโอเพนซอร์สพัฒนาจากโปรแกรมภาษา C++ ในปี ค.ศ. 1996 โดยมหาวิทยาลัยสโลวีเนียและสถาบันเทคโนโลยีลูบลิยานาแล้วต่อยอดด้วยภาษาไพตอนจนถึงปัจจุบัน แล้วกลายเป็นเครื่องมือหนึ่งที่กูเกิ้ลสนับสนุนให้นิสิตนักศึกษามากกว่า 10,900 คนจาก 103 ประเทศทั่วโลกเข้ามาพัฒนาต่อภายใต้โครงการ “Google Summer of Code” หรือ “GSoC” [22] ด้วยชื่อว่า “Orange at Google Summer of Code 2016” [23] โดยก่อนนำข้อมูลเข้าเครื่องมือดังกล่าว ผู้วิจัยได้เขียนโปรแกรมประมวลผลก่อน ด้วยภาษาไพตอนเพื่อสกัดสาระในข้อมูลจากแฟ้มล็อกออกมาเป็นแต่ละรายการการเปลี่ยนแปลง โดยแต่ละรายการการ

เปลี่ยนแปลงสะท้อนถึงเหตุการณ์หนึ่ง ๆ ด้วยวิธีการเรียนรู้ของเครื่องแบบต้นไม้ตัดสินใจเรียนรู้รูปแบบและแยกแยะข้อมูลที่ปกติและผิดปกติออกมา

การสอน (Train) และทดสอบ (Test) ใช้ 10-Fold Validation กล่าวคือแบ่งข้อมูลออกเป็น 10 ชุดแล้วในชุดข้อมูลแต่ละชุดแบ่งออกเป็น 10 ส่วน 9 ส่วนเพื่อสอน และอีก 1 ส่วนเพื่อทดสอบ 10 รอบ แล้วนำค่าความแม่นยำมาคำนวณได้จากอัตราส่วนระหว่างจำนวนตัวอย่างในชุดข้อมูลทดสอบที่ทำนายกลุ่มได้อย่างถูกต้องทั้งหมด ได้แผนภาพต้นไม้ดัง รูปที่ 5



รูปที่ 5 ต้นไม้ตัดสินใจเพื่อแยกกลุ่มข้อมูลปกติและผิดปกติ

ต้นไม้ตัดสินใจ รูปที่ 5 ด้วยการเลือกคุณลักษณะโดยใช้เอนโทรปีด้วยการแต่งกิ่ง (Pruning) แตกใบด้วยค่าอย่างน้อย 2 ค่า โดยหยุดแยกโหนดด้วยค่าน้อยกว่า 5 ค่า โดยจำกัดความลึกไม่เกิน 100 ชั้น ด้วยสีเขียวคือกลุ่มข้อมูลที่ปกติ และสีแดงคือกลุ่มข้อมูลที่ผิดปกติ

Node 1 ซึ่งเป็นราก ประกอบด้วยข้อมูลปกติ 104,381 ระเบียน และผิดปกติ 119 ระเบียน รวมเป็น 104,500 ระเบียน

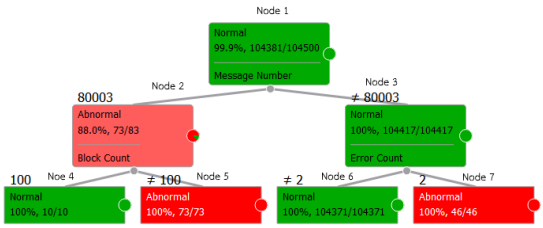
Node 2 เป็นกลุ่มข้อมูลที่ผิดปกติ 72 ระเบียน แทนด้วยสีแดงจากการตั้งเงื่อนไขการนับจำนวนคำว่า “No Data” เท่ากับสองคำ

Node 3 เป็นกลุ่มข้อมูลปกติจำนวน 104,428 ระเบียน เงื่อนไขว่าไม่พบคำว่า “No Data” ในรายการการเปลี่ยนแปลง

Node 4 เป็นกลุ่มข้อมูลที่ผิดปกติ 47 ระเบียน จากเงื่อนไขจำนวนคำว่า “Error” ต่อรายการเปลี่ยนแปลงเท่ากับสองคำ

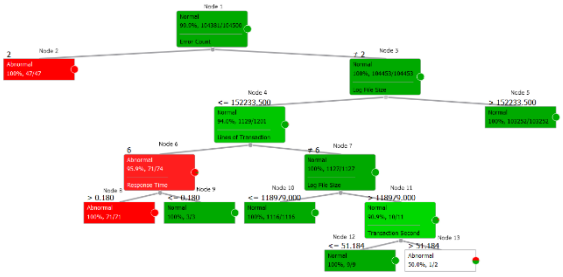
Node 5 เป็นกลุ่มข้อมูลปกติ 104,381 ระเบียน จากเงื่อนไขจำนวนคำว่า “Error” ไม่ใช่สองคำ

จากต้นไม้นี้จะพบว่าใช้เพียงสองคุณลักษณะคือการนับจำนวนคำว่า “No Data” และคำว่า “Error” เพื่อให้เข้าใจสาเหตุอย่างอื่นที่ทำให้เกิดความผิดพลาด เราจะตัดคุณลักษณะนับจำนวนคำ “No Data” ออกจากระบบเรียนรู้ต้นไม้ตัดสินใจจะมีลักษณะเป็นดังรูปที่ 6



รูปที่ 6 ต้นไม้สินใจที่ตัดคุณลักษณะการนับคำว่า “No Data” ออก

ต้นไม้นี้แสดงว่าคุณลักษณะแรกที่ใช้ตัดสินใจคือหมายเลขข้อความเป็น 80003 จากนั้นเป็น นับคำว่า “Block” และนับคำว่า “Error” เนื่องจากหมายเลขข้อความไม่สามารถนำมาใช้ในทางปฏิบัติเนื่องจากหมายเลขจะเปลี่ยนไปตามเวลาที่เครื่องจักรทำงาน เราจะตัดคุณลักษณะนี้ออก เพื่อดูให้ลึกลงไปอีกว่า มีคุณลักษณะใดอีกที่เป็นสาเหตุของความผิดพลาด ได้ต้นไม้ตัดสินใจตามรูปที่ 7



รูปที่ 7 ต้นไม้สินใจที่ตัดหมายเลขข้อความออก

จะเห็นได้ว่า คุณลักษณะที่บ่งบอกความผิดพลาดคือ error count, log file size, line of transaction, response time และ transaction time

หลังจากได้แบบจำลองแล้วทดลองนำแบบจำลองจากรูปที่ 5 การวิเคราะห์แฟ้มล็อกของวันที่ 2 มกราคม ค.ศ. 2016 จำนวน 5,640 ระเบียน ได้ผลถูกต้องทั้งหมดโดยใช้เวลาไม่ถึงหนึ่งนาที

5. สรุปและวิจารณ์

ต้นไม้ตัดสินใจ ซึ่งได้จากการเรียนรู้ข้อมูลจริงในสายการผลิต สามารถนำมาใช้ตรวจสอบความผิดปกติของเครื่องจักรจากแฟ้มล็อก ได้ถูกต้อง 100% คุณลักษณะที่สำคัญ คือ การปรากฏคำว่า “No data,” “Error,” ขนาดของแฟ้มล็อก จำนวนบรรทัดของเหตุการณ์ เวลาตอบสนอง และเวลาที่ใช้ในการลงบันทึก คุณลักษณะเหล่านี้สามารถอธิบายเกี่ยวข้องกับการเกิดความผิดพลาดของเครื่องจักรผลิตได้ เช่นคำว่า “No data” สะท้อนถึงความผิดปกติที่มาจากวิศวกร และคำว่า “Error” สะท้อนถึงความผิดปกติที่มาจากระบบ ขนาดของแฟ้มล็อกที่ใหญ่สะท้อนถึงความผิดปกติของเครื่องจักร เวลาตอบสนอง และระยะเวลาที่เกิดเหตุการณ์มีส่วนบ่งบอกความผิดพลาดเช่นกัน

การใช้โปรแกรมเพื่อตรวจสอบความผิดปกติของเครื่องจักรจากแฟ้มล็อกโดยอัตโนมัติช่วยให้การหาสาเหตุที่เครื่องจักรทำงานผิดพลาด และการแก้ไขโดยทีมวิศวกรผู้ดูแลการผลิตให้สายการผลิตดำเนินการต่อไป เป็นไปอย่างรวดเร็วและแม่นยำ ช่วยลดเวลาที่สายการผลิตต้องหยุดลง นอกจากนี้ การวิเคราะห์ต้นไม้ตัดสินใจสามารถทำให้ทีมวิศวกรเข้าใจสาเหตุที่เครื่องจักรในการผลิตทำงานผิดพลาดได้ดียิ่งขึ้น ซึ่งจะนำไปสู่การปรับปรุงกระบวนการผลิตให้ลดข้อผิดพลาดลงอีกด้วย

6. กิตติกรรมประกาศ

โครงงานนี้ได้รับความอนุเคราะห์จาก Western Digital เพื่อสนับสนุนโครงการด้านการศึกษาและโครงการให้ความช่วยเหลือของมหาวิทยาลัย อย่างไรก็ตามรายงานฉบับนี้จัดทำขึ้นโดยผู้วิจัยเป็นผู้สร้างสรรค์ขึ้นเองและไม่ได้จัดพิมพ์เผยแพร่โดยหรือในนามของ Western Digital

7. เอกสารอ้างอิง

[1] D. Amornthep and K. Athakorn, "The improvement for optimization of Head Stack Assembly (HSA) assembling process by using the virtual reality 3D simulation model," *AIJSTPME.,* Vol. 3(4), pp. 47-56, 2010.

[2] W. Yu, M. Qiang and P. Michael, "Health monitoring of hard disk drive based on Mahalanobis distance," in *2011 Prognostics and System Health Management Conference*, 24-25 May. Shenzhen: pp. 1-8, 2011.

[3] S. G. Eick, M. C. Nelson and J. D. Schmidt, "Graphical analysis of computer log files," *Communications of the ACM December 1994.,* Vol. 37(12), pp. 50-56, 1994.

[4] F. Adam (26 October 2015). Killed by a Machine: The Therac-25, HACKADAY. [Online], Available: http://hackaday.com /2015/10/26/killed-by-a-machine-the-therac-25/

[5] G. Tingting, D. Chunling, S. Weijie and X. Lihua, "An integrated plant/control design method and application in hard disk drives," *International Journal of Systems Science*., Vol. 47(3), pp. 644-651, 2016.

[6] F. Schuster, A. Paul and H. König, "Towards learning normality for anomaly detection in industrial control networks". *Emerging Management Mechanisms for the Future Internet.,* Vol.7943, pp. 61-72, 2013.

[7] M. M. Tom, *Machine learning*. Boston: WCB/McGraw-Hill, 1997.

[8] M. M. Tom, *The Discipline of Machine Learning*. Technical Report CMU-ML-06-108, Carnegie Mellon Univerisy, 2006.

[9] S. Anotai, S. Sukree and C. Prabhas, "Improving quality of products in Hard Drive manufacturing by decision tree technique," *International Journal of Computer Science* Issues., Vol. 9(3), pp. 29-34, 2012.

[10] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*., Vol. 13(7), pp. 1443-1471, 2001.

[11] R. Volker, "Kernel fisher discriminants for outlier detection," *Neural Computation*., Vol. 18(4), pp. 942-960, 2006.

[12] R. Volker, "Outlier detection with one-class kernel fisher discriminants," *Advances in Neural Information Processing Systems., Vol. 17,* pp. 1169-1176, 2004.

[13] J. R. Quainlan, "Induction of decision trees," *Machine Learning.*, Vol 1(1), pp. 81-106, 1986.

[14] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Franciso, CA, USA: Morgan Kaufmann, 1993.

[15] L. Victor and G. Nigel. (19 January, 2014). *Decicion Tree, School of Informatics*, [Online]. Available: https://www.youtube.com/watch?v=eKD5gxPPeY0&list=PLBv09BD7ez_4temBw7vLA19p3tdQH6FYO

[16] K. Weerapol and T. Kiatfa, "The characterization in the Head Stack Assembly (HSA) during the swaging process: optimzation of actuator arm material," *KKU Res J.,* Vol. 15(10), pp. 910-918, 2010.

[17] T. Kiatfa and K. Weerapol, "A study of the Head Stacks Assembly (HSA) during the swaging process: optimization of the ball velocity," *KKU Research Journal.,* Vol. 15(2), pp. 104-112, 2010.

[18] B. Pooja and G. Kanwal, "An imperial learning of data mining classification algorithms in intrusion detection dataset," *International Journal of Scientific & Engineering Research.,* Vol. 4(6), pp. 2394-2399, 2013.

[19] D. Janez, Z. Blaž, L. Gregor and C. Tomaz, "Orange: From Experimental Machine Learning to Interactive Data Mining," *Knowledge Discovery in Databases: PKDD 2004.,* Vol. 3202, pp. 537-539, 2004.

[20] D. Janez, C. Tomaž, E. Aleš, G. Črt, H. Tomaž, M. Mitar, M. Martin, P. Matija, T. Marko, S. Anže, Š. Miha, U. Lan, Ž. Lan, Ž. Jure, Ž Marinka and Z Blaž, "Orange: Data Mining Toolbox in Python," *Journal of Machine Learning Research.,* Vol. 14(Aug), pp. 2349-2353, 2013.

[21] T. Miha and D. Janez, "Interactive Network Exploration with Orange," *Journal of Statistical Software.,* Vol. 53(6), pp. 1-24, 2013.

[22] Google Inc. (3 March, 2016). *Google Summer of Code.* [Online] Available: https://summerofcode.withgoogle.com/.

[23] P. Ajda. (3 March 2016). *Orange at Google Summer of Code 2016*. [Online] Available: http://blog.biolab.si/