การออกแบบและพัฒนาเทคนิคไฮบริดสำหรับการเติมค่าข้อมูลที่สูญหาย

นางสาวพัชราวรรณ ชินไธสง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2555

THE DESIGN AND DEVELOPMENT OF A HYBRID TECHNIQUE FOR MISSING VALUE IMPUTATION

Phatcharawan Chinthaisong

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Master of Engineering in Computer Engineering

Suranaree University of Technology

Academic Year 2012

การออกแบบและพัฒนาเทคนิคไฮบริดสำหรับการเติมค่าข้อมูลที่สูญหาย

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

_____

(รศ. ดร.กิตติศักดิ์ เกิดประสพ)
ประธานกรรมการ

_____

(รศ. ดร.นิตยา เกิดประสพ)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

_____

(อ. ดร.จิติมนต์ อั่งสกุล)
กรรมการ

_____ _____

(ศ. ดร.ชูกิจ ลิมปิจำนงค์) (รศ. ร.อ. ดร.กนต์ธร ชำนิประศาสน์)
รองอธิการบดีฝ่ายวิชาการ คณบดีสำนักวิชาวิศวกรรมศาสตร์

พัชราวรรณ ชินไชสง : การออกแบบและพัฒนาเทคนิคไฮบริดสำหรับการเติมค่าข้อมูลที่สูญหาย (THE DESIGN AND DEVELOPMENT OF A HYBRID TECHNIQUE FOR MISSING VALUE IMPUTATION) อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.นิตยา เกิดประสพ, 83 หน้า.

ปัจจุบันการทำเหมืองข้อมูลมีความสำคัญและเป็นประโยชน์ในหลาย ๆ ด้านเช่น การสร้างโมเดลเพื่อทำนายด้านการแพทย์ ด้านการศึกษา ด้านธุรกิจ ด้านวิศวกรรม และด้านอื่น ๆ แต่ถ้าชุดข้อมูลที่มีข้อมูลที่สูญหายจะมีผลกระทบต่อโมเดลที่ถูกสร้างขึ้นมาเพื่อการทำนาย ถ้านำโมเดลนั้นไปใช้ในด้านการแพทย์ เช่น การสร้างโมเดลเพื่อทำนายโรคให้กับผู้ป่วยจะทำให้โมเดลมีค่าความถูกต้องต่ำและการทำนายผลอาจผิดพลาดจะทำให้ผู้ป่วยเสียชีวิตได้ ถ้าในด้านธุรกิจอาจทำให้เสียทรัพย์สินถึงขั้นล้มละลายได้ งานวิจัยนี้จึงออกแบบเทคนิคมาเพื่อช่วยเติมค่าให้กับข้อมูลที่สูญหายซึ่งจะทำให้โมเดลที่ได้มีความถูกต้องและแม่นยำขึ้น การพัฒนาและทดสอบอัลกอริทึมจะใช้การเขียนโปรแกรมด้วยภาษาอาร์ซึ่งเป็นภาษาเชิงฟังก์ชันที่นิยมใช้ในด้านสถิติ การออกแบบโปรแกรมเติมค่าให้กับข้อมูลสูญหาย จะให้โปรแกรมทำการวิเคราะห์หาเทคนิคที่เหมาะสมสำหรับการเติมค่าให้กับข้อมูลสูญหายที่จะทำให้ประสิทธิภาพของโมเดลดีขึ้นโดยการตรวจสอบด้วยวิธีการสร้างโมเดลของต้นไม้ตัดสินใจด้วยข้อมูลที่เตรียมไว้สำหรับฝึกสอน และจะตรวจสอบค่าความถูกต้องด้วยข้อมูลที่เตรียมไว้สำหรับทดสอบ ซึ่งงานวิจัยนี้จะทำให้มีเทคนิคในการทำนายค่าสูญหายขึ้นมาใหม่ และยังสามารถใช้เทคนิคนี้กับวิธีการทำเหมืองข้อมูลแบบอื่น ๆ ได้อีกด้วย

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2555 ลายมือชื่ออาจารย์ที่ปรึกษา_____

PHATCHARAWAN CHINTHAISONG : THE DESIGN AND DEVELOPMENT OF A HYBRID TECHNIQUE FOR MISSING VALUE IMPUTATION. THESIS ADVISOR : ASSOC. PROF. NITTAYA KERDPRASOP, Ph.D., 83 PP.

MISSING VALUE/IMPUTATION TECHNIQUE/R PROGRAMMING

At present data mining is important and can be useful in many domains such as medical, education, business, engineering, and others. But if the data are incomplete, they can affect the induced predictive model. If such model has been used to diagnose patients, it can have deadly impact. For business applications, the poor model can cause bankruptcy. This research thus proposes the design of missing-value imputation techniques to help improving the predictive performance of a classification model. An implementation of the proposed techniques has been done with the R language, which is the functional language used in statistics. The program has been designed to automatically select an appropriate imputation techniques for each specific kind of data types. The performance of our imputation methods has been tested through the assessment of the induced classification tree model. The imputation techniques proposed in this research can also support other kinds of data mining tasks.

School of Computer Engineering Student's Signature_____

Academic Year 2012 Advisor's Signature_____

กิตติกรรมประกาศ

การทำวิทยานิพนธ์เล่มนี้สำเร็จไปด้วยดีต้องขอกราบขอบพระคุณบุคคลต่าง ๆ ที่ได้ให้ความกรุณาในการช่วยเหลือให้คำปรึกษาและคำแนะนำ ในด้านงานวิชาการ และด้านการวิจัย ดังต่อไปนี้

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ อาจารย์ที่ปรึกษาวิทยานิพนธ์ และรองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ ที่ให้คำปรึกษาทั้งการดำเนินงานวิจัย การตรวจสอบรูปแบบและความถูกต้องของงานวิจัยในเล่มวิทยานิพนธ์

ผู้ช่วยศาสตราจารย์ ดร.พิชโยทัย มหัทธนาภิวัฒน์ ผู้ช่วยศาสตราจารย์ ดร.คะชา ชาญศิลป์ ผู้ช่วยศาสตราจารย์ สมพันธ์ ชาญศิลป์ อาจารย์ ดร. ชาญวิทย์ แก้วกสิ ผู้ช่วยศาสตราจารย์ ดร. ปรเมศวร์ ห่อแก้ว อาจารย์ วิชัย ศรีสุรักษ์ และอาจารย์ ศรัญญา กาญจนวัฒนา อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

ขอขอบคุณคุณภาสพิชญ์ ชูใจ ที่ช่วยตรวจทานเอกสารและนักศึกษาบัณฑิตสาขาวิชาวิศวกรรมคอมพิวเตอร์ทุกคน ที่ช่วยให้คำปรึกษาและคำแนะนำมาโดยตลอด และขอขอบพระคุณคุณครู อาจารย์ทั้งในปัจจุบันและที่ผ่านมาที่เคยให้ความรู้ต่าง ๆ สุดท้ายขอขอบคุณบิดา มารดาที่อบรมเลี้ยงดูและคอยสนับสนุนในด้านการศึกษามาเป็นอย่างดีโดยตลอด จนสามารถทำให้ผู้วิจัยประสบความสำเร็จในชีวิต

พัชราวรรณ ชินไธสง

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

การทำเหมืองข้อมูล (Data mining) เป็นการค้นหารูปแบบหรือโมเดลจากชุดข้อมูลหรือคลังของข้อมูลที่มีจำนวนของข้อมูลมาก ซึ่งประเภทในการทำเหมืองข้อมูลมีได้หลายประเภทเช่น การค้นหากฎความสัมพันธ์ (Association rule mining) จะเป็นการวิเคราะห์และแสดงความสัมพันธ์ของเหตุการณ์ที่เกิดขึ้นพร้อมกัน ตัวอย่างเช่น ถ้ามีการซื้อขนมปังและซื้อน้ำอัดลมแล้วจะไม่ซื้อนม การจำแนกประเภทข้อมูล (Data classification) จะทำการค้นหากฎเพื่อจัดประเภทของวัตถุโดยดูจากคุณสมบัติของวัตถุ และการแบ่งกลุ่มข้อมูล (Data clustering) ซึ่งจะทำการแบ่งข้อมูลที่มีลักษณะคล้ายกันออกเป็นกลุ่มย่อยเพื่อง่ายในการนำไปวิเคราะห์ข้อมูลในแต่ละกลุ่มย่อยต่อไป เป็นต้น (วิกิพีเดีย สารานุกรมเสรี, 2555ก)

ปัจจุบันการทำเหมืองข้อมูลถูกทำมาใช้ในงานหลาย ๆ ด้านไม่ว่าจะเป็นด้านการศึกษา เช่น การทำเหมืองข้อมูลเพื่อประเมินพฤติกรรมการเรียนของนักเรียนว่าควรจัดกลุ่มการเรียนการสอนอย่างไร ด้านการแพทย์ซึ่งนิยมนำการทำเหมืองข้อมูลเข้ามาช่วยในการทำนายโรคของผู้ป่วยจากอาการที่เกิดขึ้นหรือจากประวัติของผู้ป่วย ด้านธุรกิจมีการนำเทคนิคการทำเหมืองข้อมูลเข้ามาช่วยในการทำนายการขึ้นลงของหุ้น การวางแผนทางการตลาด และประโยชน์ในด้านอื่น ๆ เช่น การพยากรณ์อากาศ การคาดการณ์การเกิดแผ่นดินไหว เป็นต้น

การทำเหมืองข้อมูลส่วนใหญ่จะเป็นการจำแนกประเภทข้อมูลโดยการใช้อัลกอริทึมการสังเคราะห์ต้นไม้ตัดสินใจ (Decision tree induction) ซึ่งเป็นอัลกอริทึมที่ใช้การสร้างโมเดลที่มีลักษณะคล้ายต้นไม้ซึ่งจะมีโหนดราก กิ่งก้าน และส่วนของใบ สำหรับการทำนายจะพิจารณาจากเงื่อนไขมีการกำหนดเป้าหมายที่ต้องการเช่น ตัวอย่างชุดข้อมูลการเล่นกอล์ฟ ถ้ามีการกำหนดเป้าหมายในการทำนายอยากรู้ว่าวันนี้จะออกไปเล่นกอล์ฟหรือไม่ โดยจะพิจารณาจากตัวแปรของข้อมูลต่าง ๆ ที่เกิดขึ้นเช่น ความแรงของลม อากาศ สภาพของความชื้น ซึ่งการสร้างแผนภาพต้นไม้ตัดสินใจจะมี ราก กิ่งก้าน ใบที่สร้างขึ้นมาจากเงื่อนไข การใช้โมเดลต้นไม้ตัดสินใจในการทำนายอาจจะให้ผลลัพธ์ว่า ถ้าวันพรุ่งนี้ลมแรง ฝนตก จะตัดสินใจว่าไม่ออกไปเล่นกอล์ฟ เป็นต้น

การรวบรวมข้อมูลเป็นขั้นตอนแรกก่อนการส่งข้อมูลออกไปยังอัลกอริทึมสังเคราะห์โมเดล ปัญหาที่มักจะเกิดขึ้นในขั้นตอนนี้ได้แก่ การรวบรวมข้อมูลที่นำมาสร้างเป็นชุด

ข้อมูลซึ่งอาจเกิดข้อผิดพลาดกับข้อมูลที่เรียกว่า ข้อมูลที่สูญหาย (Missing value) ข้อมูลที่ผิดพลาดอาจจะเกิดจากการเก็บข้อมูลไม่ครบถ้วนหรือผิดพลาดจากการกรอกข้อมูลลงฐานข้อมูลทำให้ข้อมูลบางส่วนเกิดการสูญหาย ซึ่งการที่ข้อมูลสูญหายนี้จะมีผลต่อการทำเหมืองข้อมูลโดยประสิทธิภาพในการพยากรณ์ข้อมูลมีความถูกต้องต่ำหรืออาจจะทำนายผลผิดพลาด ถ้าในด้านธุรกิจอาจจะทำให้ขาดทุนหรือสูญเสียรายได้ ในส่วนของงานด้านการแพทย์ถ้าการทำนายเกิดข้อผิดพลาดไม่ใช่เพียงแต่เสียทรัพย์แต่อาจจะทำให้ผู้ป่วยถึงกับเสียชีวิตได้ จึงต้องมีการให้ความสำคัญกับข้อมูลที่สูญหาย โดยมีเทคนิคที่จะช่วยในการทำนายค่าของข้อมูลที่สูญหายเพื่อให้การทำเหมืองข้อมูลมีประสิทธิภาพเพิ่มมากยิ่งขึ้น ซึ่งเทคนิคในการจัดการกับข้อมูลสูญหายนั้นมีหลากหลายเทคนิคที่ส่งผลต่อประสิทธิภาพในการทำนายของโมเดลแตกต่างกัน

ดังนั้นการศึกษาเปรียบเทียบเทคนิคการหาค่าให้กับข้อมูลที่สูญหายและการออกแบบเทคนิคใหม่เพื่อให้สามารถทำนายค่าของข้อมูลที่สูญหายให้มีประสิทธิภาพดียิ่งกว่าเทคนิคอื่น ๆ จึงเป็นสิ่งที่น่าสนใจและนำมาพัฒนาเป็นงานวิจัย เพื่อให้การทำเหมืองข้อมูลจากชุดข้อมูลที่มีการทำนายค่าข้อมูลที่สูญหายแล้วมีโมเดลในการทำนายผลที่ดีขึ้นและลดโอกาสการทำนายที่ผิดพลาด สามารถนำโมเดลนั้นมาใช้เพื่อการทำนายและนำผลที่ได้จากการทำนายไปใช้ให้เกิดประโยชน์ได้ดียิ่งขึ้น

1.2 วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบเทคนิคที่มีประสิทธิภาพในการเติมค่าให้กับข้อมูลที่สูญหายให้มีค่าที่ถูกต้องใกล้เคียงกับค่าที่แท้จริงมากที่สุด เพื่อทำให้ประสิทธิภาพของโมเดลในการทำนายเพิ่มขึ้น โดยจำแนกเป็นวัตถุประสงค์ย่อยได้ดังนี้

1.2.1 เพื่อศึกษาเทคนิคการเติมค่าให้กับข้อมูลที่สูญหายในลักษณะไม่มีผู้ฝึกสอน (Unsupervised)

1.2.2 เพื่อศึกษาเทคนิคการเติมค่าให้กับข้อมูลที่สูญหายในลักษณะมีผู้ฝึกสอน (Supervised) โดยพิจารณาจากข้อมูลประกอบอื่น ๆ

1.2.3 เพื่อสร้างเทคนิคใหม่ที่ช่วยให้การเติมค่าให้กับข้อมูลสูญหายและช่วยให้สร้างโมเดลการจำแนกประเภทข้อมูลที่มีประสิทธิภาพในการทำนายที่ดีขึ้นและมีความถูกต้องมากขึ้น

1.3 ขอบเขตของการวิจัย

งานวิจัยนี้เป็นการศึกษาและพัฒนาเทคนิคการเติมค่าให้กับข้อมูลที่สูญหาย โดยเมื่อนำมาใช้เพื่อเติมค่าข้อมูลที่สูญหายแล้วจะสามารถนำชุดข้อมูลไปสร้างโมเดลเพื่อการจำแนก

ประเภท (Classification model) ที่ให้ความถูกต้องแม่นยำเพิ่มขึ้น และการสร้างเทคนิคที่มีประสิทธิภาพในการทำนายค่าข้อมูลที่สูญหายมีขอบเขตการวิจัยดังนี้

1.3.1 งานวิจัยนี้มีการศึกษาและเปรียบเทียบเทคนิคในการเติมค่าให้กับข้อมูลที่สูญหายในลักษณะไม่มีผู้ฝึกสอน (Unsupervised) คือไม่ต้องพิจารณาค่าจากแอททริบิวต์อื่นประกอบและลักษณะมีผู้ฝึกสอน (Supervised) คือใช้ค่าจากแอททริบิวต์อื่นช่วยในการพิจารณาเติมข้อมูลที่สูญหาย

1.3.2 งานวิจัยนี้ได้สร้างเทคนิคในการทำนายค่าให้กับข้อมูลที่สูญหายขึ้นมาใหม่ให้มีประสิทธิภาพที่ดีขึ้นและการวัดประสิทธิภาพจะใช้ค่าความถูกต้องในการทำนายเพื่อเปรียบเทียบของ Classification model ที่สร้างด้วยเทคนิคการสังเคราะห์ต้นไม้ตัดสินใจ

1.3.3 งานวิจัยนี้พัฒนาด้วยภาษา R ซึ่งเป็นภาษาในเชิงฟังก์ชัน โดยโปรแกรมที่พัฒนาขึ้นจะเป็นโปรแกรมช่วยงานในขั้น Pre-process ซึ่งเป็นขั้นตอนการเตรียมข้อมูลของกระบวนการทำเหมืองข้อมูล และขั้นตอนการทำเหมืองข้อมูลที่เป็นการสังเคราะห์ต้นไม้ตัดสินใจจะใช้อัลกอริทึมที่มีอยู่แล้วในไลบรารี่ของภาษา

1.4 ประโยชน์ที่ได้รับ

งานวิจัยนี้เป็นการศึกษาและพัฒนาเทคนิคในการเติมค่าให้กับข้อมูลที่สูญหายให้สามารถสร้างโมเดลที่มีการทำนายที่ดีขึ้น ประโยชน์ที่ได้รับมีดังต่อไปนี้

1.4.1 สามารถเติมค่าให้กับข้อมูลที่สูญหายในชุดข้อมูลหลากหลายรูปแบบ เช่นข้อมูลประเภท ข้อมูลเชิงคุณลักษณะ (Categorical) ประเภทข้อมูลเชิงตัวเลข (Numeric) หรือข้อมูลที่มีทั้ง ข้อมูลเชิงคุณลักษณะและเชิงตัวเลขผสมกัน โดยจะให้ประสิทธิภาพในการทำนายที่ดีกว่าโมเดลที่สร้างจากข้อมูลที่ไม่มีการเติมค่าให้กับข้อมูลที่หาย

1.4.2 สามารถเปรียบเทียบเทคนิคในการเติมค่าข้อมูลที่สูญหายได้ โดยการสร้างเทคนิคขึ้นมาใหม่เพื่อเติมข้อมูลที่สูญหายที่มีประสิทธิภาพในการสร้างโมเดลเพื่อการทำนายที่มีความถูกต้องแม่นยำสูง เมื่อเปรียบเทียบกับเทคนิคต่าง ๆ ที่ได้ศึกษามา

1.4.3 เทคนิคใหม่ในการเติมค่าให้กับข้อมูลที่สูญหายมีความยืดหยุ่นโดยสามารถเลือกเทคนิคที่เหมาะสมกับแต่ละประเภทข้อมูลข้อมูลได้แก่ เชิงคุณลักษณะ(Categorical) ข้อมูลเชิงตัวเลข (Numeric) ข้อมูลแบบผสมผสาน (Mixed) ได้โดยอัตโนมัติ

1.4.4 สามารถนำเทคนิคการเติมค่าให้กับข้อมูลที่สูญหายที่เสนอในงานวิจัยนี้ ไปใช้ร่วมกับอัลกอริทึมอื่น ๆ ในการทำเหมืองข้อมูลได้ เช่น อัลกอริทึมเพื่อค้นหากฎความสัมพันธ์

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

บทนี้จะกล่าวถึงปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง โดยจะอธิบายรายละเอียดที่เกี่ยวข้องกับการเติมค่าให้กับข้อมูลสูญหาย และการเขียนโปรแกรมด้วยภาษาอาร์ที่ใช้ในการทำเหมืองข้อมูล และงานวิจัยที่เกี่ยวข้อง

2.1 การทำเหมืองข้อมูล (Data mining)

การทำเหมืองข้อมูลเป็นกระบวนการค้นหาและสร้างรูปแบบจากฐานข้อมูลออกมาอยู่ในรูปแบบของโมเดลมาเพื่อนำช่วยในการวิเคราะห์และทำนายผลให้กับข้อมูลชุดอื่นเพื่อให้เกิดประโยชน์สูงสุดในการนำไปใช้งานโดยการเลือกรูปแบบที่มีประสิทธิภาพดีที่สุด ซึ่งการทำเหมืองข้อมูลจะอยู่ในกลุ่มของสาขาปัญญาประดิษฐ์ (Artificial intelligence) ผู้ที่นิยมใช้ส่วนมากจะเป็นนักสถิติที่นำไปใช้วิเคราะห์ข้อมูลแล้วยังสามารถนำไปประยุกต์ใช้ในด้านธุรกิจที่ช่วยในการตัดสินใจในการลงทุน ด้านการแพทย์ที่สามารถช่วยในการวินิจฉัยโรคให้กับผู้ป่วย และด้านอื่น ๆ ได้อีกมากมาย ซึ่งการเรียนรู้ที่ได้จากการทำเหมืองข้อมูลมีหลายประเภทแบ่งออกเป็นกลุ่มใหญ่ ๆ ได้คือ

2.1.1 ประเภทการทำเหมืองข้อมูล

1.) ค้นหากฎความสัมพันธ์ (Association rule)

การหาความสัมพันธ์ที่เกิดขึ้นพร้อมกันของวัตถุนั้นแล้วแสดงออกมาในรูปแบบของกฎซึ่งเป็นกระบวนการที่นิยมใช้กันมากในการทำเหมืองข้อมูล และวิธีการที่นิยมใช้กันคือ Apriori เพราะสามารถดึงกฎที่มีประสิทธิภาพออกมาใช้งานได้และลดความซับซ้อนในการเลือกกฎได้ กฎความสัมพันธ์ที่วิเคราะห์จากชุดข้อมูลการซื้อสินค้าในซุปเปอร์มาเกต จะได้กฎที่แสดงพฤติกรรมการซื้อสินค้าเช่น ถ้าลูกค้าซื้อนมก็จะต้องซื้อขนมปังด้วยเสมอเป็นต้น ดังรูปที่ 2.1 จะแสดงตัวอย่างการทำเหมืองข้อมูลแบบหากฎความสัมพันธ์

Transaction ID	Milk	Bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

	Milk	Bread	Butter	Beer
Milk	2*	2	1	0
Bread	2	4*	2	0
Butter	1	1	2*	0
Beer	0	0	0	1*

IF condition Then result

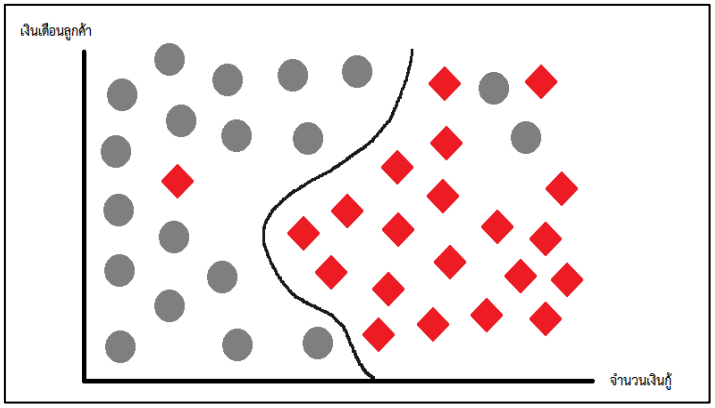
1.If Milk Then Bread

2.If Milk Then Butter

รูปที่ 2.1 ตัวอย่างการทำเหมืองข้อมูลแบบหากฎความสัมพันธ์ (วิกิพีเดีย สารานุกรมเสรี, 2555ข)

2.) การจำแนกประเภทข้อมูล (Classification)

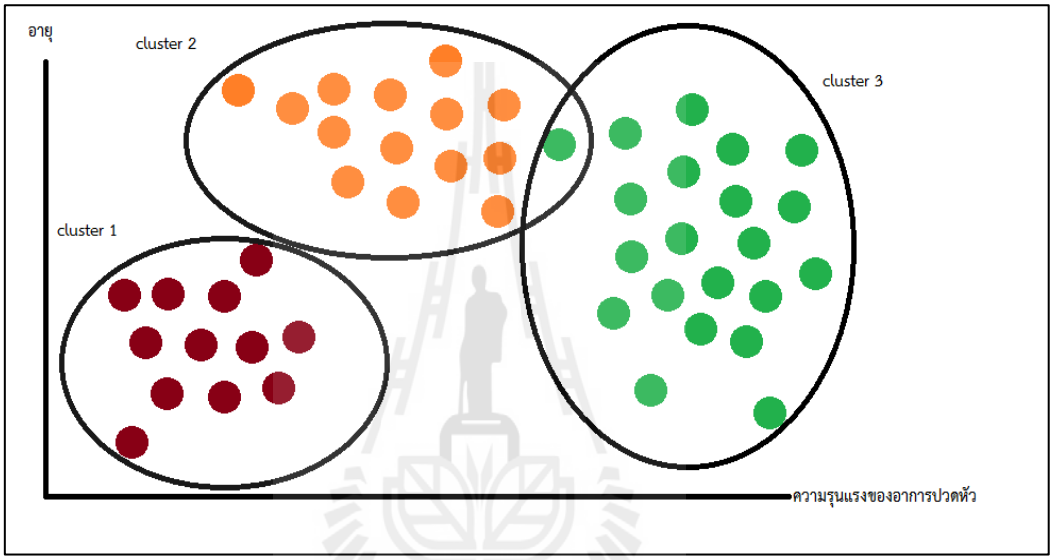
การจำแนกประเภทข้อมูลสามารถจัดข้อมูลออกเป็นแต่ละคลาสได้ (นิตยา เกิดประสพ, 2555ข)เช่น ถ้าชุดข้อมูลเป็นประวัติของลูกค้าธนาคารแห่งหนึ่งข้อมูลแต่ละเรคคอร์ดประกอบด้วย เพศ อายุ เงินเดือน ตำแหน่งงาน และประเภทของลูกค้า (ลูกค้าชั้นดีและลูกค้ามีประวัติผิดนัดชำระหนี้) แล้วเลือกนำประเภทมาหาความสัมพันธ์ระหว่างเงินเดือนและตำแหน่งงาน เพื่อใช้วิเคราะห์คุณสมบัติว่าควรพิจารณาอนุมัติเงินกู้ให้กับลูกค้ารายใหม่ว่าจะก่อหนี้ดีหรือหนี้เสียให้กับธนาคารหรือไม่เป็นต้น ดังรูปที่ 2.2 แสดงตัวอย่างการทำเหมืองข้อมูลแบบการจำแนกประเภทข้อมูล โดยเส้นในภาพแสดงการแบ่งขอบเขตระหว่างลูกค้าชั้นดีและไม่ดี รูปวงกลมแทนลูกค้าชั้นดีและรูปสี่เหลี่ยมแทนลูกค้าที่ไม่ดี



รูปที่ 2.2 ตัวอย่างการทำเหมืองข้อมูลแบบการจำแนกประเภทข้อมูล

3.) การแบ่งกลุ่มข้อมูล (Clustering)

การจัดกลุ่มแบ่งกลุ่มเป็นการให้กับชุดข้อมูลออกเป็นกลุ่มย่อย ๆ เป็นการค้นหาลักษณะที่คล้ายคลึงกันหรือใกล้เคียงกันความสัมพันธ์เพื่อนำมาแบ่งกลุ่มข้อมูล เช่นการนำไปประยุกต์ใช้ในงานด้านการแพทย์ โดยการนำชุดข้อมูลของผู้ป่วยมาวิเคราะห์ ซึ่งจะแบ่งกลุ่มตามอาการของผู้ป่วยที่มีโรคเดียวกัน เพื่อนำไปช่วยวิเคราะห์สาเหตุในการเป็นโรคจากผู้ป่วยที่มีลักษณะอาการใกล้เคียงกัน ซึ่งรูปที่ 2.3 ตัวอย่างการทำเหมืองข้อมูลแบบการแบ่งกลุ่มข้อมูล



รูปที่ 2.3 ตัวอย่างการทำเหมืองข้อมูลแบบการแบ่งกลุ่มข้อมูล

4.) การสร้างมโนภาพ (Visualization)

การสร้างมโนภาพเป็นการนำเสนอข้อมูลด้วยใช้รูปกราฟิก เป็นแผนผังหรือภาพที่เคลื่อนไหว สามารถนำเสนอข้อมูลได้อย่างชัดเจนและครบถ้วน แทนการใช้ข้อความที่นำเสนอข้อมูลมากมาย ทำให้การสร้างข้อมูลเชิงสรุปหรือเชิงเปรียบเทียบแบบใช้รูปจะสามารถอ่านโมเดลง่ายกว่าการแสดงด้วยข้อความ

2.1.2 ขั้นตอนในการทำเหมืองข้อมูล

ประกอบด้วยการทำงานแต่ละขั้นตอนเพื่อเปลี่ยนชุดข้อมูลเป็นโมเดลที่มีประสิทธิภาพ โดยมีรายละเอียดขั้นตอนดังนี้ (นิตยา เกิดประสพ, 2555ก)

1.) Data selection / sampling

ขั้นตอนนี้จะทำการเลือกข้อมูลจากฐานข้อมูลที่มีข้อมูลจำนวนมาก ซึ่งจะเลือกเฉพาะข้อมูลที่สนใจหรืออาจจะใช้การสุ่มข้อมูลสำหรับนำมาใช้วิเคราะห์ในขั้นต่อไป

2.) Data cleaning / preprocessing

ขั้นตอนที่สองเป็นการคัดทิ้งข้อมูลที่มีความผิดพลาดหรือไม่เกี่ยวข้องออก เช่นข้อมูลที่ผิดพลาดจากการเก็บข้อมูล การกรอกข้อมูลไม่ตรงกับหัวข้อของข้อมูล ข้อมูลเกิดการสูญหายไป ซึ่งข้อมูลเหล่านี้ต้องได้รับการแก้ไขหรือตัดออกก่อนจะนำไปใช้วิเคราะห์ในขั้นต่อไป

3.) Data transformation / reduction

ขั้นตอนที่สามจะเป็นการแปลงข้อมูลให้เหมาะสมกับการใช้งานตามรูปแบบของโปรแกรมที่จะใช้ในการทำเหมืองข้อมูล ในกรณีที่ชุดข้อมูลมีรูปแบบไม่ตรงกับรูปแบบของโปรแกรมต้องจัดการปรับแก้ก่อนนำไปใช้วิเคราะห์ และเลือกพิจารณาเฉพาะคอลัมน์ที่ต้องการ

4.) Data mining

ขั้นตอนที่สี่จะเป็นการค้นหารูปแบบจากชุดข้อมูลที่เราได้ดึงและปรับแก้ไขในขั้นตอนก่อนหน้านี้ โดยการทำเหมืองข้อมูลจะมีหลายประเภทซึ่งจะทำการค้นหารูปแบบที่เหมาะสม เพื่อให้เกิดประโยชน์ในการนำไปใช้งาน

5.) Interpretation / evaluation

ขั้นตอนที่ห้าเป็นการประเมินผลของรูปแบบที่ได้จากการทำเหมืองข้อมูลในขั้นตอนที่สี่ว่ามีประสิทธิภาพและความถูกต้องที่จะสามารถนำไปใช้ได้หรือไม่ ถ้าประสิทธิภาพของรูปแบบที่ได้ยังไม่ดีพอ อาจจะต้องเริ่มกระบวนการตั้งแต่การเลือกชุดข้อมูลที่จะนำมาหารูปแบบใหม่

6.) Consolidating discovered knowledge

ขั้นตอนสุดท้ายจะทำการสรุปผลที่ได้จากการทำงานตามขั้นตอนข้างต้นแล้วนำไปเสนอความรู้ที่ได้จากการทำเหมืองข้อมูลด้วยเทคนิคการนำเสนอแบบต่าง ๆ เพื่อให้ผู้ฟังหรือผู้ใช้สามารถเข้าใจผลลัพธ์ได้ง่ายขึ้น

2.1.3 สถาปัตยกรรมของระบบการทำเหมืองข้อมูล

โครงสร้างหรือสถาปัตยกรรมของระบบการทำเหมืองข้อมูล (นิตยา เกิดประสพ, 2555ก) มีส่วนประกอบสำคัญดังนี้

1.) Database, Data Warehouse, World Wide Web, Other Info Repositories

เป็นแหล่งที่มาของข้อมูลหรือฐานข้อมูล ซึ่งจะนำข้อมูลในส่วนนี้มาใช้วิเคราะห์สำหรับการนำไปใช้ทำเหมืองข้อมูล

2.) Database Management System หรือ Data Warehouse Server

เป็นส่วนที่ช่วยนำข้อมูลเข้ามาจากฐานข้อมูลตามคำสั่งที่ได้รับ

3.) Data Mining Engine

เป็นส่วนประกอบหลักที่จะรับผิดชอบในขั้นตอนการทำเหมืองข้อมูล ซึ่งจะค้นหารูปแบบที่เหมาะสมจากการทำเหมืองข้อมูลประเภทต่างๆ เช่น การค้นหากฎความสัมพันธ์ (Association rule mining) การจำแนกประเภท (Data classification) และการแบ่งกลุ่มข้อมูล (Data clustering) เป็นต้น

4.) Pattern Evaluation Module

ส่วนนี้จะเป็นตัวที่ใช้วัดและคัดเลือกรูปแบบที่น่าสนใจของผลลัพธ์ที่ได้จาก Data Mining Engine เพื่อให้ได้รูปแบบที่น่าสนใจและมีประสิทธิภาพในการนำไปใช้ประโยชน์

5.) Knowledge Base

ส่วนนี้จะเป็นแหล่งรวบรวมวามรู้จากการทำเหมืองข้อมูลซึ่งมีประโยชน์เมื่อต้องการค้นหาหรือใช้ร่วมกับการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล

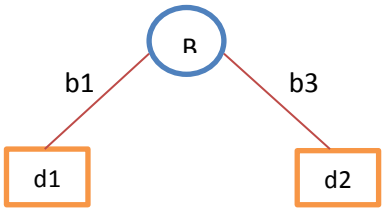
6.) User Interface

ส่วนนี้จะช่วยติดต่อระหว่างผู้ใช้งานระบบกับระบบที่ใช้ในการทำเหมืองข้อมูล ซึ่งจะช่วยให้ผู้ใช้สามารถสั่งงานและดูรูปแบบที่ได้จากการทำเหมืองข้อมูล หรือผลลัพธ์การประมวลประสิทธิภาพของรูปแบบที่ได้ว่าสามารถนำไปใช้งานต่อได้หรือไม่

2.2 การเติมค่าให้กับข้อมูลสูญหาย (Missing value imputation)

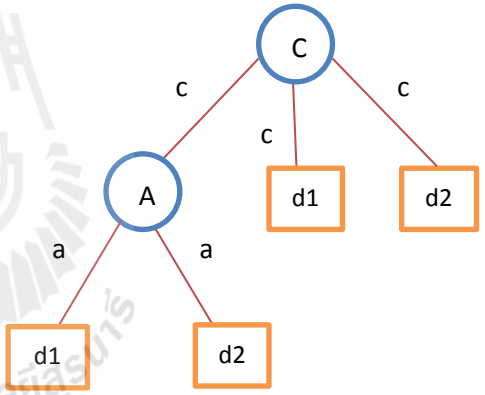
การทำเหมืองข้อมูลกับชุดข้อมูลที่มีข้อมูลสูญหายมักจะทำให้ได้โมเดลที่มีประสิทธิภาพต่ำ โดยการทำนายของโมเดลอาจเกิดข้อผิดพลาดในการทำนายได้ ส่วนใหญ่ข้อมูลสูญหายมักจะเกิดมาจากการเก็บข้อมูลที่ไม่สมบูรณ์หรือการกรอกข้อมูลในระหว่างการจัดเก็บชุดข้อมูลเกิดผิดพลาด ตัวอย่างข้อมูลอย่างง่ายในการเกิดค่าของข้อมูลสูญหายและผลกระทบที่เกิดขึ้นกับโมเดลแสดงได้ดังรูปที่ 2.4

A	B	C	D
a1	b1	c1	**d1**
a1	b3	c3	**d2**
a2	b1	c2	**d1**
a2	b3	c1	**d2**



(ก) กรณีข้อมูลปกติและโมเดลที่ได้จากข้อมูล

A	B	C	D
a1		c1	**d1**
a1	b3	c3	**d2**
a2		c2	**d1**
a2	b3	c1	**d2**



(ข) กรณีที่เกิดข้อมูลสูญหายและโมเดลที่เปลี่ยนไป

รูปที่ 2.4 ตัวอย่างการเกิดข้อมูลสูญหายและผลกระทบถึงโมเดลที่ได้

จากรูปที่ 2.4 (ก) แสดงกรณีที่มีข้อมูลครบถ้วนสมบูรณ์และโมเดลที่ได้อธิบายการจำแนกข้อมูลได้ว่าค่าในแอททริบิวต์ B เพียงแอททริบิวต์เดียวสามารถจำแนกคลาสของแอททริบิวต์ D ซึ่งเป็นแอททริบิวต์เป้าหมายได้ แต่เมื่อเกิดเหตุการณ์ของข้อมูลสูญหายในแอททริบิวต์ B ดังรูปที่ 2.4 (ข) จะส่งผลให้โมเดลที่ได้มีความซับซ้อนมากขึ้น โมเดลที่ซับซ้อนจะมีโอกาสทำนายข้อมูลใน

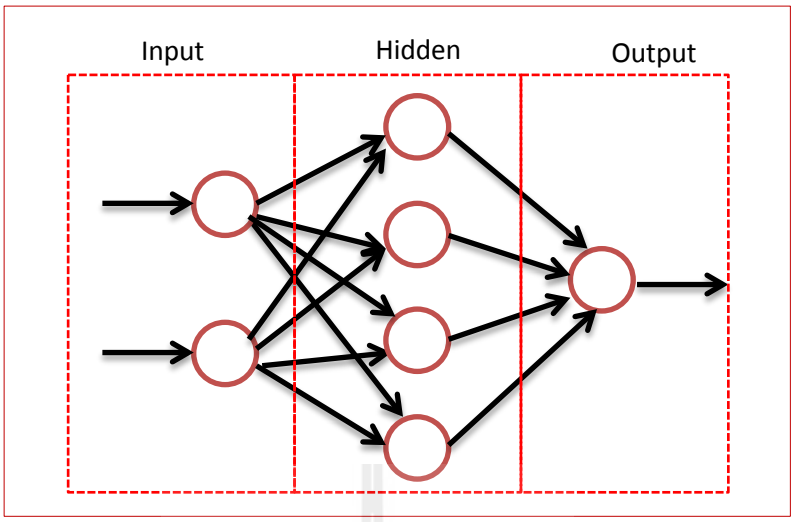
อนาคตได้ผิดพลาดสูง เนื่องจากเป็นโมเดลที่เจาะจงมากเกินไป (Over fitting) การเติมค่าข้อมูลสูญหายจึงช่วยลดลักษณะโมเดลเจาะจงและช่วยเพิ่มโอกาสในการทำนายถูกให้กับโมเดล

ตัวอย่างเทคนิคการทำนายข้อมูลสูญหาย

การเติมค่าให้กับชุดข้อมูลที่มีข้อมูลสูญหายเกิดขึ้น เป็นคิดวิธีที่จะเพิ่มประสิทธิภาพให้กับโมเดลที่ใช้เพื่อทำนาย เทคนิคการทำนายค่าให้กับของข้อมูลสูญหายมีหลากหลายเทคนิค (Jerzy, 2004; Shichao, 2005; Joseph, 2007; Karlien, 2009; Pedro, 2010) ได้แก่

1. การใช้ทฤษฎีราฟเซต โดยจะสร้างเงื่อนไขตามเซตและนำมาสร้างเป็นกฎเพื่อเติมค่าให้กับข้อมูลสูญหาย ทฤษฎีราฟเซต (Rough set) ถูกนำเสนอในปีค.ศ. 1982 โดย Zdzislaw Pawlak ซึ่งเป็นการหาค่าเซตที่เล็กที่สุดออกมาเพื่อใช้งาน โดยจะมีการประมาณค่าเซตออกมาสองแบบคือ R-lower หรือเรียกว่า lower approximation คือเซตของข้อมูลทั้งหมดที่สนใจและสามารถเกิดขึ้นได้แน่นอน R-upper หรือเรียกว่า upper approximation คือเซตของข้อมูลทั้งหมดที่สนใจสามารถมีความเป็นไปได้ ซึ่งเป็นทฤษฎีหนึ่งที่นิยมนำมาใช้เป็นเทคนิคในการเติมค่าให้กับข้อมูลสูญหาย (Jerzy, 2003; Jerzy, 2004)

2. การใช้โครงสร้างเครือข่ายประสาทเทียม โดยการพิจารณาถึงคุณลักษณะอื่นของข้อมูลเพื่อนำมาใช้ทำนายข้อมูลสูญหาย โครงสร้างเครือข่ายประสาทเทียม (Artificial neural network) เป็นแนวคิดให้คอมพิวเตอร์ทำงานเหมือนสมองของมนุษย์ มีการนำไปใช้ในด้านการแพทย์ เช่น หุ่นยนต์ผ่าตัด หุ่นยนต์ช่วยหยิบอุปกรณ์ ด้านอุตสาหกรรมเช่นการให้เครื่องจักรอุตสาหกรรมสามารถวิเคราะห์แยกผลไม้ลงกล่องได้เป็นต้น โดยมีกระบวนการทำงานเช่นการเปลี่ยนตัวเองจากการประมวลผลตามลำดับให้เป็นการประมวลผลแบบคู่อันดับได้ มีลักษณะการทำงานที่แต่ละ Process จะรับ Input เข้าไปคำนวณ และสร้าง Output ออกมาในลักษณะที่ไม่ใช่การทำงานเชิงเส้นตรง เพราะ Input แต่ละตัวจะถูกให้ลำดับความสำคัญของค่าไม่เท่ากัน ค่าของ Output ที่ได้จากการเชื่อมโยงนี้จะถูกนำมาเปรียบเทียบกับ Output ที่ได้ตั้งเอาไว้ ถ้าค่าที่ออกมาเกิดความคลาดเคลื่อน ก็จะนำไปสู่การปรับค่าหรือน้ำหนักของค่าที่ใส่ไว้แต่ละ Input เป็นต้น ดังตัวอย่างในรูปที่ 2.5 ซึ่งโครงสร้างเครือข่ายประสาทเทียมมีหลายประเภท และมีการสร้างโมเดลได้ทั้งแบบมีผู้ฝึกสอน (Supervised) และแบบไม่มีผู้ฝึกสอน (Unsupervised) และมีงานวิจัยที่นิยมใช้เทคนิคนี้ด้วย (George, 2007; Pedro, 2010; Loris, 2012)



รูปที่ 2.5 ตัวอย่างการออกแบบโครงสร้างประสาทเทียม

3. การใช้แผนภาพต้นไม้ตัดสินใจเป็นตัวช่วยในการตัดสินใจคัดเลือกให้กับค่าที่จะนำมาเติม ต้นไม้ตัดสินใจ (Decision tree) เป็นโมเดลทางคณิตศาสตร์และเป็นวิธีการในการทำเหมืองข้อมูลด้วย โดยสร้างโมเดลในลักษณะแผนภาพเป็นลำดับชั้น เพื่อให้อ่านแล้วเข้าใจการตัดสินใจของแผนภาพต้นไม้ได้ง่าย ซึ่งงานวิจัยส่วนมากจะใช้แผนภาพต้นไม้เพื่อสร้างเป็นโมเดลที่จะช่วยทำนายค่าของข้อมูลสูญหาย (Shichao, 2005)

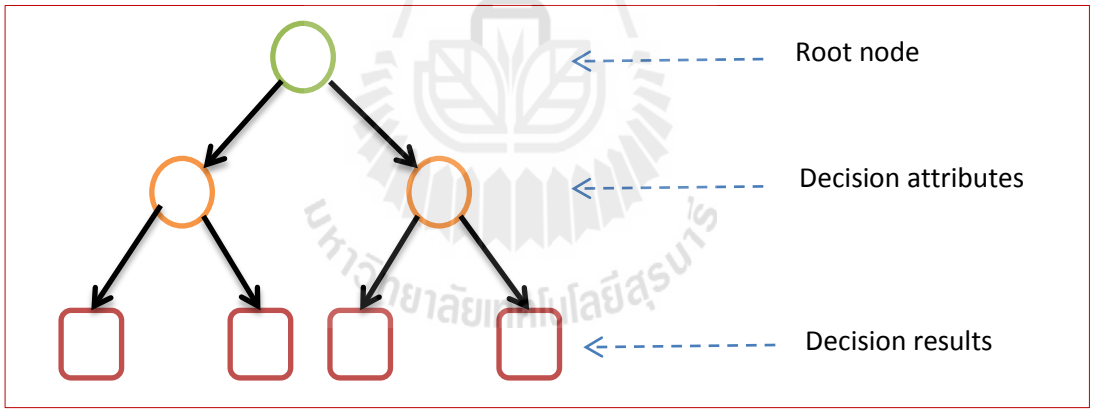
4. การใช้ค่าทางสถิติเข้ามาช่วยในการหาค่าของข้อมูลสูญหายมีงานวิจัยจำนวนมาก (Shmuel, 2005; Xiao, 2011) จะทำการหาค่าให้กับข้อมูลสูญหายด้วยเทคนิคการประมาณค่าเชิงสถิติ เช่น การใช้ ค่าเฉลี่ย ค่ากลาง ค่าสหสัมพันธ์ เป็นต้น โดยส่วนมากจะเน้นทางด้านข้อมูลทางสถิติที่คำนวณค่าได้ ข้อมูลทางด้านสถิติมีหลายประเภทเช่น ข้อมูลเชิงปริมาณ (Quantitativedata) เป็นข้อมูลที่ใช้การวัดค่าได้ ซึ่งจะมีสองแบบคือ ข้อมูลแบบต่อเนื่อง (Continuous data) กับ ข้อมูลแบบไม่ต่อเนื่อง (Discrete Data) และประเภทข้อมูลเชิงคุณภาพ (Qualitative data) เป็นข้อมูลซึ่งไม่สามารถคำนวณค่าได้ ถ้าข้อมูลไม่ใช่ตัวเลขต้องแบ่งประเภทแล้วแทนด้วยตัวเลข ซึ่งการใช้ค่าทางสถิติจะใช้กับข้อมูลสูญหายได้เฉพาะชนิดแบบเป็นตัวเลขเท่านั้น

5. การสร้างกฎความสัมพันธ์ขึ้นมาเพื่อทำนายค่าข้อมูลที่สูญหายโดยพิจารณาจากค่าสนับสนุนและค่าความเชื่อมั่นของกฎ การค้นหากฎความสัมพันธ์ (Association rule mining) เป็นกระบวนการในการทำเหมืองข้อมูล เป็นการเรียนรู้มาจากข้อมูลที่มีอยู่เพื่อประโยชน์ในการทำนายข้อมูลใหม่ที่จะเกิดขึ้นในอนาคต โดยการสร้างกฎมีความสัมพันธ์ของไอเทมจากชุดข้อมูลที่มีอยู่ ตัวอย่างเช่น ถ้าปรากฏไอเทม A และไอเทม B ก็ต้องปรากฏไอเทม D ด้วย ส่วนมากการเลือกกฎที่

จะนำไปใช้ได้จริงจะพิจารณาค่าสนับสนุนและค่าความเชื่อมั่นยิ่งมีค่ามากเท่าไรยิ่งดี ซึ่งเทคนิคนี้ได้มีการวิจัยเพื่อนำไปใช้หาค่าของข้อมูลสูญหายกันอย่างแพร่หลาย (Jianhua, 2007)

2.3 การสร้างโมเดลข้อมูลในลักษณะต้นไม้ตัดสินใจ

การสร้างต้นไม้ตัดสินใจเป็นการทำเหมืองข้อมูลประเภทการจำแนก (Classification) โดยจะสร้างโหนดหรือเงื่อนไขที่ได้จากการเรียนรู้จากชุดข้อมูลที่มีการระบุประเภทข้อมูลหรือคลาสอยู่แล้ว และจะนำไปสร้างโมเดลช่วยในการตัดสินใจเพื่อทำนายประเภทข้อมูลที่จะเกิดขึ้นในอนาคต โมเดลที่ได้จะมีลักษณะเป็นโมเดลที่คล้ายต้นไม้เพราะมีกิ่งก้านสาขาแตกย่อยออกมาก จึงเรียกว่าต้นไม้ตัดสินใจ โดยจะมีโหนดรากทำหน้าที่เป็นจุดเริ่มต้นในการค้นหา และมีโหนดลำดับชั้นถัดมาเรียกว่าโหนดลูก ซึ่งจำนวนโหนดลูกนั้นจะขึ้นอยู่กับเงื่อนไขของโหนดในลำดับบนที่เรียกว่าโหนดแม่ ถ้ามีเงื่อนไขมากก็จะทำให้โหนดลูกมีจำนวนมากเท่ากับจำนวนเงื่อนไข และโหนดสุดท้ายจะเป็นโหนดเป้าหมายในการทำนายคลาสของข้อมูลซึ่งเป็นสิ่งที่เราต้องการทราบว่าโมเดลนี้ทำนายแล้วได้คำตอบอย่างไร โครงสร้างของต้นไม้ตัดสินใจแสดงได้ดังรูปที่ 2.6



รูปที่ 2.6 ตัวอย่างแผนภาพต้นไม้ตัดสินใจ

Root node เรียกว่า รากของต้นไม้ตัดสินใจหรือตำแหน่งเริ่มต้นการค้นหา

Decision attributes เรียกว่า โหนดกิ่งหรือโหนดลูกมีจำนวนขึ้นอยู่กับจำนวนเงื่อนไขที่ใช้ในการตัดสินใจของโหนดแม่

Decision results เรียกว่า โหนดใบซึ่งก็คือเป้าหมายที่ต้องการในการทำนายของโมเดล

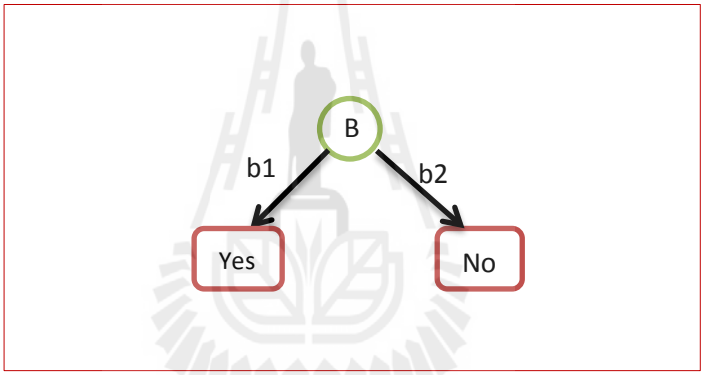
ตัวอย่างในการสร้างแผนภาพต้นไม้ตัดสินใจจากข้อมูลในตารางที่ 2.1 จะสามารถสร้างต้นไม้ตัดสินใจจากข้อมูลนี้ได้ดังรูปที่ 2.7 และสามารถแปลงโมเดลต้นไม้ตัดสินใจเป็นกฎได้ดังนี้

if B = b1 then Yes

if B = b2 then No

ตารางที่ 2.1 ตัวอย่างชุดข้อมูลอย่างง่าย

A	B	C	D
a1	b1	c1	Yes
a1	b2	c2	No
a2	b2	c1	No
a2	b1	c1	Yes

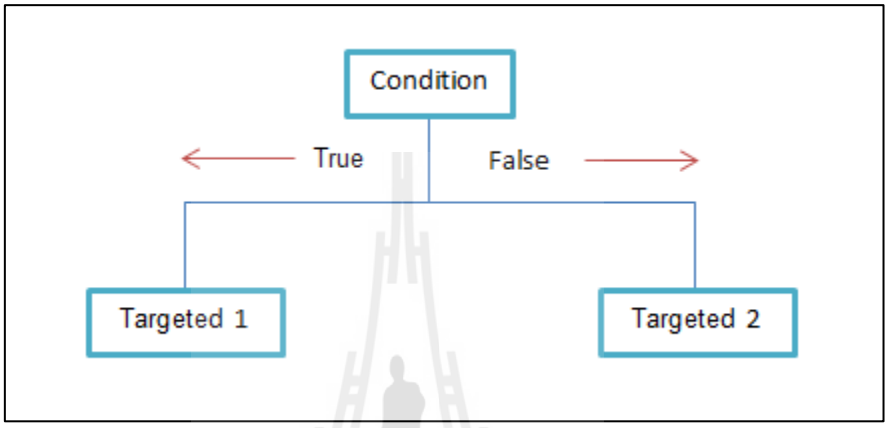


รูปที่ 2.7 ตัวอย่างการสร้างแผนภาพในลักษณะของต้นไม้ตัดสินใจ

ในงานวิจัยนี้หลังจากเติมค่าให้กับข้อมูลสูญหายแล้วจะทดสอบความมีประสิทธิภาพของเทคนิคด้วยการนำข้อมูลที่เติมค่าของข้อมูลสูญหายแล้วมาสร้างโมเดลต้นไม้ตัดสินใจ เทคนิคที่มีคุณภาพดีคือเทคนิคที่ช่วยให้สามารถสร้างต้นไม้ตัดสินใจที่มีความแม่นยำในการทำนายสูง การทดสอบประสิทธิภาพของโมเดลที่สร้างจากแผนภาพต้นไม้จะทำการทดสอบโดยการแบ่งชุดข้อมูลเป็น 2 ชุด ชุดแรกใช้สำหรับทำการฝึกสอนเพื่อสร้างโมเดล และชุดที่สองซึ่งเป็นชุดทดสอบใช้สำหรับการวัดประสิทธิภาพการทำนายของโมเดลนั้น

Recursive partitioning เป็นเทคนิคการสร้างต้นไม้ตัดสินใจแบบการแบ่งเป็นไบนารี่หรือแบบสองเส้นทางในการตัดสินใจของแต่ละเงื่อนไขเท่านั้นตัวอย่างดังรูปที่ 2.8 ถ้าเส้นทางไปทางด้านซ้ายเป็นจริงในเงื่อนไขนั้น แต่ถ้าไปทางด้านขวาจะเป็นเท็จของเงื่อนไขนั้น ซึ่งจะพิจารณา

ความเป็นไปได้ทั้งหมดในการสร้างต้นไม้ตัดสินใจ และสามารถใช้เรียกปัจจัยเดิมนำมาประกอบการตัดสินใจหลายครั้งแต่ใช้เงื่อนไขที่ต่างกันจึงเรียกว่าเป็นการแบ่งไบนารี่แบบเรียกซ้ำ การแบ่งพาร์ทิชันจะใช้สมการ (2.1) ในการคำนวณของน้ำหนักเปรียบเทียบ การใช้ Logic regression เป็นมาตราฐานสำหรับการวิเคราะห์ข้อมูลของข้อมูลแบบไบนารี่ (Heping, 2011)



รูปที่ 2.8 ตัวอย่างการสร้างต้นไม้ตัดสินใจประเภทการแบ่งแบบไบนารี่

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \quad (2.1)$$

n_i = จำนวนเรคคอร์ดหลัง Split เรียกว่า Child node

n = จำนวนเรคคอร์ดก่อน Split เรียกว่า Parent node

$GINI(i) = GINI(t) - GINI_{split}$

$$Splits = 2^{k-1} - 1 \quad (2.2)$$

สมการที่ 2.2 เป็นการคำนวณการแบ่งข้อมูลในแต่ละระดับชั้นของต้นไม้ตัดสินใจ การแบ่งข้อมูลจะหยุดเมื่อข้อมูลทุกตัวในโหนดเป็นคลาสชนิดเดียวกัน หรือมีค่าของแอททริบิวต์ที่เหมือนกัน

$$Gini(t) = 1 - \sum_{i=0}^{C-1} [p(i|t)]^2 \quad (2.3)$$

สมการที่ 2.3 เป็นการคำนวณ Gini Index ซึ่งเป็นค่าที่เป็นเกณฑ์ว่าควรนำแอททริบิวต์นั้นมาเป็นคุณลักษณะที่ใช้แบ่งหรือไม่

$$Entropy(t) = 1 - \sum_{i=0}^{C-1} [p(i|t)] log_2 p(i|t) \quad (2.4)$$

สมการที่ 2.4 เป็นการคำนวณหาค่าความยุ่งเหยิงของกลุ่มข้อมูลหนึ่ง

$$Classsification\ error(t) = 1 - Max[p(i|t)] \quad (2.5)$$

สมการที่ 2.5 เป็นการคำนวณความผิดพลาด (Misclassification error) ที่เกิดกับโหนดของต้นไม้ตัดสินใจ

2.4 การวิเคราะห์ข้อมูลทางสถิติ

2.4.1 ค่าเฉลี่ยหรือมัชฌิมเลขคณิต (Mean)

การหาค่าเฉลี่ยจะแบ่งออกเป็น 6 ชนิด (สายชล สินสมบูรณ์ทอง, 2555) คือ ค่าเฉลี่ยเลขคณิตหรือมัชฌิมเลขคณิต (Arithmetic mean) ค่าเฉลี่ยเรขาคณิตหรือมัชฌิมเรขาคณิต (Geometric mean) ค่าเฉลี่ยฮาร์โมนิค (Harmonic Mean) รากที่สองของค่าเฉลี่ยกำลังสอง (Root Mean Square) ค่าเฉลี่ยตัวอย่างแบบทำให้เรียบ (Trimmed Simple Mean) และค่าเฉลี่ยตัวอย่างแบบวินโสไรซน์ (Winsorized Simple Mean) ซึ่งในงานวิจัยนี้ได้ใช้การหาค่าเฉลี่ยเลขคณิต (Arithmetic mean) ซึ่งการหาค่าเฉลี่ยชนิดเลขคณิตจะใช้ค่าทั้งหมดที่มีในข้อมูลนำมารวมกันแล้วหารด้วยจำนวนของข้อมูลทั้งหมดดังในสมการที่ 2.6

$$Mean = \frac{\sum_{i=1}^{N} X_i}{N} \quad (2.6)$$

N = จำนวนของข้อมูลที่ใช้หาค่าเฉลี่ย

X_i = ค่าของข้อมูลที่นำมาบวกรวมหาค่าเฉลี่ยโดยเริ่มต้นที่จำนวนที่ 1 จนถึงจำนวนที่ N

<u>ตัวอย่างการหาค่าเฉลี่ย</u>

ข้อมูล : 2 13 7 20 12 15 10 25 4 32

การหาค่าเฉลี่ยจากสูตร $Mean = \frac{\sum_{i=1}^{N} X_i}{N}$

$$= \frac{2+13+7+20+12+15+10+25+4+32}{10}$$

$$= \frac{140}{10}$$

$$= 14$$

ค่าเฉลี่ยของข้อมูลมีค่าเท่ากับ 14

ข้อดีและข้อเสียของค่าเฉลี่ยเลขคณิต

การหาค่าเฉลี่ยเลขคณิตสามารถคำนวณได้ง่ายและสะดวก และจะได้ผลลัพธ์เพียงหนึ่งค่า จึงเป็นวิธีที่ง่ายต่อการนำไปใช้และสรุปผลของข้อมูลนั้น แต่การหาค่าเฉลี่ยจะใช้ได้เฉพาะข้อมูลที่เป็นเชิงปริมาณและมีการกระจายตัวที่ปกติคือการที่ไม่มีข้อมูลที่มีค่าสูงมาก ๆ หรือต่ำมาก ๆ กว่าข้อมูลอื่น ๆ

2.4.2 ค่ากลางหรือค่ามัธยฐาน (Median)

การหาค่ากลางมีสองชนิดคือ การใช้กับข้อมูลที่ไม่ได้จัดกลุ่มและการใช้กับกับข้อมูลที่จัดกลุ่มแล้ว (สายชล สินสมบูรณ์ทอง, 2555) ซึ่งในงานวิจัยนี้จะใช้กับข้อมูลที่ยังไม่ได้จัดกลุ่มคือ การนำข้อมูลทั้งหมดที่ต้องการหาค่ากลางมาเรียงจากมากไปน้อยหรือน้อยไปมาก และเลือกตำแหน่งที่อยู่ระหว่างกลางของข้อมูลแล้วนำค่าของตำแหน่งนั้นมาเป็นค่ากลาง ซึ่งการเลือกตำแหน่งจะมีวิธีการเลือก 2 แบบคือ ถ้าจำนวนชุดข้อมูลเป็นจำนวนคี่จะสามารถหาตำแหน่งได้ตามสมการที่ 2.7 และถ้าจำนวนของข้อมูลเป็นจำนวนคู่จะใช้สมการที่ 2.8 และสมการที่ 2.9 เพื่อค้นหาตำแหน่งซึ่งได้มาสองตำแหน่งแล้วนำค่าของทั้งสองตำแหน่งมารวมกันแล้วหารด้วยสอง ซึ่งค่า n = จำนวนของข้อมูลทั้งหมด

$$\frac{n+1}{2} \qquad \text{(2.7)}$$

$$\frac{n}{2} \qquad \text{(2.8)}$$

$$\frac{n+2}{2} \qquad \text{(2.9)}$$

ตัวอย่างการหาค่ากลาง

ข้อมูล : 2 13 7 20 12 15 10 25 4 32

นำข้อมูลมาเรียง : 2 4 7 10 12 13 15 20 25 32

ชุดข้อมูลนี้มีจำนวนข้อมูลเป็นเลขคู่จึ่งต้องใช้สมการ $\frac{n}{2}$ และสมการ $\frac{n+2}{2}$ เพื่อหาค่ากลาง

ตำแหน่ง$_1$ = $\frac{10}{2}$ จะได้ตำแหน่งที่ 5 มีค่าเท่ากับ 12

ตำแหน่ง$_2$ = $\frac{10+2}{2}$ จะได้ตำแหน่งที่ 6 มีค่าเท่ากับ 13

นำค่าของตำแหน่ง$_1$+ค่าของตำแหน่ง$_2$ แล้วหารสอง = $\frac{12+13}{2}$ = 12.5

ค่ากลางของข้อมูลมีค่าเท่ากับ 12.5

ข้อดีและข้อเสียของค่ากลาง

สามารถหาค่ากลางกับข้อมูลที่มีการกระจายแบบผิดปกติหรือมีค่าผิดปกติเกิดขึ้นภายในข้อมูลได้ โดยไม่มีผลกระทบกับค่ากลางที่ได้ แต่การค่าหากลางต้องเสียเวลามาเรียงข้อมูลก่อนการค้นหาและไม่ได้ใช้ข้อมูลทุกค่าในการคำนวณเพียงเลือกค่าของตำแหน่งกลางเท่านั้น

2.4.3 ค่าปรากฏบ่อยหรือค่าฐานนิยม (Mode)

การหาค่าฐานนิยมของข้อมูลคือการเลือกข้อมูลที่ปรากฏบ่อยที่สุดหรือมีความถี่สูงที่สุด(สายชล สินสมบูรณ์ทอง, 2555) ซึ่งการหาค่าฐานนิยมมีสองชนิดคือ การหาค่าฐานนิยมสำหรับข้อมูลที่ไม่ได้จัดกลุ่มจะพิจารณาข้อมูลที่ปรากฏบ่อยที่สุดและการหาค่าฐานนิยมสำหรับข้อมูลที่มีการจัดกลุ่มจะพิจารณาข้อมูลที่มีความถี่สูงสุด ซึ่งงานวิจัยนี้จะใช้การหาค่าฐานนิยมสำหรับข้อมูลที่ไม่ได้มีการจัดกลุ่ม

ตัวอย่างค่าฐานนิยม

ข้อมูล : 0 0 1 2 0 0 1 1 1 2 2 1

ค่าฐานนิยมมีค่าเท่ากับ 1 เพราะ 1 ปรากฏบ่อยที่สุดมากกว่า 0 กับ 2

ข้อดีและข้อเสียของค่าฐานนิยม

สามารถหาค่าฐานนิยมกับชุดข้อมูลที่เป็นเชิงคุณภาพได้และไม่มีผลกระทบต่อการกระจายผิดปกติ แต่ถ้าข้อมูลไม่ปรากฏข้อมูลที่ซ้ำกันเลยจะไม่สามารถหาค่าฐานนิยมได้และถ้าข้อมูลปรากฏซ้ำเท่า ๆ กันอาจจะทำให้มีค่าฐานนิยมมากกว่าหนึ่งค่าได้

2.4.4 สมการถดถอยเชิงเส้น (Linear regression)

การวิเคราะห์สมการถดถอยเชิงเส้นเป็นการศึกษาความสัมพันธ์ของสองตัวแปรขึ้นไป เช่น การใช้ในด้านการขายสินค้าที่จะใช้ความสัมพันธ์ระหว่างราคาต้นทุนและกำไรที่ได้จากการขาย เพื่อทำให้ทราบถึงความสัมพันธ์ ทิศทางความสัมพันธ์และลักษณะความสัมพันธ์ระหว่างตัวแปร (เพื่อง่ายต่อการศึกษาจะใช้ความสัมพันธ์ของตัวแปรในในรูปของเส้นตรง Simple Linear Regression Analysis) เพื่อนำไปทำนายตัวแปรที่ไม่ทราบค่า (ตัวแปร) โดยอาศัยค่าที่ทราบจากตัวแปรหนึ่ง (ตัวแปรต้น) ซึ่งความแม่นยำนั้นจะที่ขึ้นกับจำนวนข้อมูลและความแปรปรวนของข้อมูล (ศิริชัย พงษ์วิชัย, 2555)

Simple Linear Regression Analysis

รูปแบบของสมการคือ $Y_i = \beta_0+\beta_1 X_i + \varepsilon$

รูปแบบสมการที่ใช้ในการทำนายคือ $\hat{y}_i = a+ bX_i$

การประมาณค่า β_0 ด้วย a และประมาณค่า β_1 ด้วย b โดยวิธีกำลังสองน้อยที่สุดจะได้ว่า

$a = \overline{Y} - b\overline{X}$ และ b คำนวณในค่าได้ตามสมการ (2.10)

$$\mathbf{b} = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum X^2 - \frac{(\sum X)^2}{n}} \qquad (2.10)$$

ซึ่งสมการที่ได้จากการประมาณค่า $\beta_0 =$ a และ $\beta_1 =$ b ยังไม่สามารถตัดสินใจได้ว่าความสัมพันธ์ของตัวแปรต้น (X_i) และตัวแปรตาม ($\hat{y}_i$) ว่ามีค่าใกล้เคียง สมการ $Y_i = \beta_0+\beta_1 X_i + \varepsilon$ จึงต้องมีสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination, R^2) ซึ่งคำนวณได้ตามต้องการ (2.11) และ (2.12)

$$R^2 = \frac{\text{ความแปรปรวนของ } Y \text{ เนื่องจากอิทธิพลของ } X, (SSR)}{\text{ความแปรปรวนของ } Y \text{ ทั้งหมด}, (SST)} \qquad (2.11)$$

$$\mathbf{R^2} = \frac{b*(Sxy^2)}{Sy^2} \qquad (2.12)$$

เมื่อ R^2 มีค่าเข้าใกล้ 1 หมายถึง เปอร์เซ็นต์ที่ X สามารถอธิบายการเปลี่ยนแปลงของ Y ได้มาก หรือ X และ Y มีความสัมพันธ์กันมาก

R^2 มีค่าเข้าใกล้ 0 หมายถึง เปอร์เซ็นต์ที่ X สามารถอธิบายการเปลี่ยนแปลงของ Y ได้น้อย หรือ X และ Y มีความสัมพันธ์กันน้อย

R^2 มีอยู่ระหว่าง 0.3 และ 0.7 หมายถึง เปอร์เซ็นต์ที่ X สามารถอธิบายการเปลี่ยนแปลง Y ได้ปานกลาง หรือ X และ Y มีความสัมพันธ์กันปานกลาง

2.5 การเขียนโปรแกรมด้วยภาษา R

ภาษาอาร์คือภาษาเชิงฟังก์ชันสำหรับใช้เขียนโปรแกรมกับงานประยุกต์ทางด้านสถิติ (กิตติ ศักดิ์ เกิดประสพ, 2012) เช่นการวิเคราะห์ข้อมูล การสร้างกราฟ สามารถทำงานกับข้อมูลได้ทั้งแบบเวกเตอร์และเมตริก และยังมีคำสั่งการใช้งานที่ง่าย สะดวก เหมาะสมสำหรับการนำมาเขียน

โปรแกรมในการทำเหมืองข้อมูล ลักษณะการทำงานของภาษาอาร์จะอำนวยความสะดวกให้ผู้ใช้สามารถเขียนโปรแกรมด้วย Command line ผ่านหน้าจอของเทอร์มินอล คำสั่งการเขียนโปรแกรมด้วยภาษาอาร์ที่เกี่ยวข้องกับการจัดการข้อมูลสูญหายมีหลักการเขียนและรูปแบบที่ง่าย ดังนี้

การตั้งชื่อตัวแปรหรือชื่อฟังก์ชัน การตั้งชื่อด้วยตัวพิมพ์ใหญ่กับตัวพิมพ์เล็กจะให้ตัวแปรที่แตกต่างกันแม้จะเป็นคำเดียวกันก็ตาม โดยถ้าผู้ใช้ตั้งชื่อให้กับฟังก์ชันซ้ำกับฟังก์ชันที่มีอยู่เดิม ระบบจะเลือกใช้ฟังก์ชันที่ผู้ใช้สร้างขึ้นใหม่แทนฟังก์ชันที่มีอยู่เดิม การกำหนดค่าให้กับตัวแปรสามารถใช้เครื่องหมาย “=” หรือ “<-” ตัวอย่าง “A=5” จะหมายถึงกำหนดให้ตัวแปร A มีค่าเท่ากับ 5 หรืออาจจะเขียนอีกแบบได้คือ “A<-5”

การสร้างฟังก์ชันขึ้นมาใช้งาน สามารถสร้างฟังก์ชันขึ้นมาได้ด้วยจะทำการเขียนชื่อฟังก์ชันไว้ด้านซ้ายมือโดยมีสัญลักษณ์ “<-” กั้นระหว่างชื่อฟังก์ชันกับการใช้งานในฟังก์ชันไว้ แล้วตามด้วยคำว่า function() ซึ่งข้างในวงเล็บจะเป็นการกำหนดพารามิเตอร์ที่ต้องการรับเข้ามาแล้วตามด้วย “{” และเขียนการทำงานของฟังก์ชันนั้นลงไป เมื่อเสร็จแล้วให้ปิดฟังก์ชันด้วย “}”

การเรียกใช้งานฟังก์ชันที่มีอยู่แล้ว โดยแต่ละฟังก์ชันจะแบ่งเก็บไว้ในแต่ละไลบรารี่ซึ่งก่อนจะเรียกใช้งานฟังก์ชันนั้นได้ ต้องมีการเรียกใช้งานไลบรารี่นั้นเสียก่อน การเรียกใช้ library() ภายในวงเล็บจะใส่ชื่อของไลบรารี่ที่ต้องการเรียกใช้ ถ้าไม่มีไลบรารี่นั้นต้องทำการติดตั้งไลบารี่นั้นก่อน จากนั้นจึงจะสามารถใช้งานฟังก์ชันที่ต้องการได้

ตัวอย่าง library(rpart)

ฟังก์ชัน na.omit() เป็นฟังก์ชันสำหรับใช้ตัดแถวที่มีค่าของข้อมูลสูญหายออกจากชุดข้อมูลชุดนั้น โดยการใส่ชื่อชุดข้อมูลในวงเล็บ แล้วฟังก์ชันจะตัดแถวที่มีค่าข้อมูลสูญหายออกให้ดังรูปที่ 2.9

ตัวอย่าง Dataset<-na.omit(dataset)

	number	size1	size1.1	have	class
1	1	1.50	13.64	NA	1
2	2	1.01	13.89	0	2
3	3	2.00	13.53	NA	1
4	4	1.50	13.21	1	1
5	5	1.50	13.27	1	2
6	6	1.00	13.64	1	1
7	7	1.50	13.89	0	2
8	8	2.00	13.53	NA	1
9	9	1.00	13.21	NA	1
10	10	1.00	13.27	NA	2

(ก) ตัวอย่างตารางข้อมูลที่มีค่าของข้อมูลสูญหาย

	row.names	number	size1	size1.1	have	class
1	2	2	1.01	13.89	0	2
2	4	4	1.50	13.21	1	1
3	5	5	1.50	13.27	1	2
4	6	6	1.00	13.64	1	1
5	7	7	1.50	13.89	0	2

(ข) ตัวอย่างตารางข้อมูลเมื่อใช้คำสั่ง na.omit

รูปที่ 2.9 ตัวอย่างผลลัพธ์การใช้งานฟังก์ชัน na.omit

ฟังก์ชัน mean() เป็นฟังก์ชันสำหรับการหาค่าเฉลี่ยในคอลัมน์ที่ต้องการ โดยการใส่ชื่อของชุดข้อมูลและคอลัมน์ที่ต้องการลงไป และกำหนดให้ค่าข้อมูลสูญหายไม่ต้องนำมาคิดเฉลี่ย ด้วยการใส่ค่าพารามิเตอร์ na.rm=T และนำค่าไปเติมในค่าของข้อมูลที่สูญหายดังในรูปที่ 2.10

ตัวอย่าง dataset[is.na(dataset[[Column]]),Column] <-mean(dataset[[Column]],na.rm=T)

	number	size1	size2	high	class
1	1	1.50	13.64	20	1
2	2	1.01	13.89	NA	2
3	3	2.00	13.53	22	1
4	4	1.50	13.21	25	2
5	5	1.50	13.27	28	2
6	6	1.00	13.64	NA	1
7	7	1.50	13.89	NA	1
8	8	2.00	13.53	23	2
9	9	1.00	13.21	30	1
10	10	1.00	13.27	35	1

(ก) ตัวอย่างตารางข้อมูลที่มีค่าของข้อมูลสูญหายที่มีการกระจายตัวแบบปกติ

	number	size1	size2	high	class
1	1	1.50	13.64	20	1
2	2	1.01	13.89	26	2
3	3	2.00	13.53	22	1
4	4	1.50	13.21	25	2
5	5	1.50	13.27	28	2
6	6	1.00	13.64	26	1
7	7	1.50	13.89	26	1
8	8	2.00	13.53	23	2
9	9	1.00	13.21	30	1
10	10	1.00	13.27	35	1

(ข) ตัวอย่างตารางข้อมูลเมื่อใช้คำสั่งในการหาค่าเฉลี่ย (mean)

รูปที่ 2.10 ตัวอย่างผลลัพธ์การใช้งานฟังก์ชันในการหาค่าเฉลี่ย

ฟังก์ชัน median() เป็นฟังก์ชันสำหรับการหาค่ากลางในคอลัมน์ที่ต้องการ โดยการใส่ชื่อของชุดข้อมูลและคอลัมน์ที่ต้องการลงไป และกำหนดให้ค่าข้อมูลสูญหายไม่ต้องนำมาคิดเฉลี่ยด้วยการใส่ na.rm=T และนำค่าไปเติมในค่าของข้อมูลที่ที่สูญหายดังรูปที่ 2.11

ตัวอย่าง dataset[is.na(dataset[[Column]])Column]<- median(dataset[[Column]],na.rm=T)

	number	size1	size2	high	class
1	1	1.50	13.64	100	1
2	2	1.01	13.89	NA	2
3	3	2.00	13.53	22	1
4	4	1.50	13.21	25	2
5	5	1.50	13.27	2	2
6	6	1.00	13.64	NA	1
7	7	1.50	13.89	NA	1
8	8	2.00	13.53	23	2
9	9	1.00	13.21	30	1
10	10	1.00	13.27	35	1

(ก) ตัวอย่างตารางข้อมูลที่มีค่าของข้อมูลสูญหายที่มีการกระจายตัวแบบเอียง

	number	size1	size2	high	class
1	1	1.50	13.64	100	1
2	2	1.01	13.89	25	2
3	3	2.00	13.53	22	1
4	4	1.50	13.21	25	2
5	5	1.50	13.27	2	2
6	6	1.00	13.64	25	1
7	7	1.50	13.89	25	1
8	8	2.00	13.53	23	2
9	9	1.00	13.21	30	1
10	10	1.00	13.27	35	1

(ข) ตัวอย่างตารางข้อมูลเมื่อใช้คำสั่งในการหาค่ากลาง (median)

รูปที่ 2.11 ตัวอย่างผลลัพธ์การใช้งานฟังก์ชันในการหาค่ากลาง

ฟังก์ชัน cor() เป็นฟังก์ชันสำหรับหาค่าสหสัมพันธ์ระหว่างคอลัมน์ที่มีค่าของข้อมูลสูญหายที่สัมพันธ์กับคอลัมน์อื่นมากที่สุด โดยจะแสดงผลออกมาเป็นตารางเมตริกความสัมพันธ์ให้เห็นว่าคอลัมน์ใดมีความสัมพันธ์กันมากที่สุดแล้วเลือกค่าของสองคอลัมน์นั้นนำมาแทนในสมการแล้วแทนค่าให้กับข้อมูลสูญหายดังรูปที่ 2.12

ตัวอย่าง cor (dataset[number:number] , use= ‘complete.obs’)

```
> symnum ( cor (iris[ , 1:4 ] , use= 'complete.obs') )
             S.L S.W P.L P.W
Sepal.Length 1
Sepal.Width      1
Petal.Length +   .   1
Petal.Width  +   .   B   1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

รูปที่ 2.12 ตัวอย่างผลลัพธ์การใช้งานฟังก์ชันในการหาค่าสหสัมพันธ์

ฟังก์ชัน predict() เป็นฟังก์ชันสำหรับการใช้ทำนาย โดยจะรับพารามิเตอร์ที่เป็นโมเดลและพารามิเตอร์ที่สองจะเป็นชุดข้อมูลที่ต้องการนำมาใช้ในการทำนาย

ตัวอย่าง predict(model_ctree, newdata = testData)

ฟังก์ชัน table() เป็นฟังก์ชันที่ใช้แสดงการนับข้อมูล แล้วแสดงตารางเมตริกดังรูปที่ 2.13

ตัวอย่าง table(predict(rf), Dataset$Goal)

```
> table(predict(iris_ctree), trainData$Species)

             setosa versicolor virginica
  setosa         34          0         0
  versicolor      0         38         3
  virginica       0          1        33
```

รูปที่ 2.13 ตัวอย่างผลลัพธ์การใช้งานฟังก์ชัน table

2.6 งานวิจัยที่เกี่ยวข้อง

การเติมค่าให้กับข้อมูลสูญหาย เป็นขั้นตอนเตรียมข้อมูลที่สำคัญขั้นตอนหนึ่งของการทำเหมืองข้อมูล ได้มีนักวิจัยเสนอเทคนิคต่าง ๆ เพื่อเติมหรือคาดเดาข้อมูลสูญหายเหล่านั้น

งานวิจัยของ Jianhua Wu et al. (2007) ได้เสนอเทคนิคในการเติมค่าให้กับข้อมูลสูญหาย โดยใช้ Association rule mining เข้ามาช่วยสร้างเป็นกฎในการทำนายค่าที่สูญหายไป แล้วนำมาเปรียบเทียบกับเทคนิค K-NN (k-nearest neighbors) โดยมีการอธิบายถึงกระบวนการในการเลือกกฎคือเลือกจากกฎที่มีความยาวมากที่สุด ถ้าความยาวเท่ากันจะพิจารณาจากค่าความเชื่อมั่น และถ้าค่าความเชื่อมั่นเท่ากันจะดูจากค่าสนับสนุน แล้วถ้าเท่ากันทุกอย่างก็จะเรียงตามตัวอักษรมาเป็นกฎ

ที่จะช่วยในการทำนายขึ้นมา ซึ่งผลที่ได้จากการวิจัยพบว่าการใช้เทคนิค Association rule mining จะมีการเติมค่าได้มีความถูกต้องมากกว่าเทคนิค K-NN

งานวิจัยของ George Ssali and Tshilidzi Marwala (2007) ได้นำเสนอเทคนิคการสร้างโมเดลสำหรับมาช่วยทำนายค่าของข้อมูลสูญหาย โดยการสร้างต้นไม้ตัดสินใจ (Decision Trees) เป็นโมเดลร่วมกับวิธีเครือข่ายประสาทเทียมแบบเครือข่ายเทียมอัตโนมัติ (AANN) และการวิเคราะห์องค์ประกอบหลักของโครงสร้างเครือข่ายเทียม (PCA-NN) ซึ่งต้นไม้ตัดสินใจที่ใช้ในการทำนายจะมีอัลกอริทึมที่เลือกขอบเขตในการค้นหาเป็นแบบลำดับและลดข้อผิดพลาดในการทำนาย ทำให้การทำนายค่าของข้อมูลสูญหายมีประสิทธิภาพดียิ่งขึ้น และวิธีการทั้งสองแบบมีประสิทธิภาพที่จะสามารถนำไปใช้ได้ ซึ่งผลการวิจัยพบว่าการใช้เทคนิคในการเติมค่าให้กับข้อมูลสูญหายด้วยเทคนิคที่มีการใช้ต้นไม้ตัดสินใจจะมีประสิทธิภาพในการหาค่าข้อมูลสูญหายดีกว่าการไม่ใช้ต้นไม้ตัดสินใจร่วมด้วยประมาณ 13 %

งานวิจัยของ Fulufhelo Vincent Nelwamondo and Tshilidzi Marwala (2008) ได้นำเสนอเทคนิคในการเติมค่าให้กับข้อมูลสูญหายโดยการใช้ทฤษฎีราฟเซต (Rough Sets) ซึ่งจะใช้วิธีการหาความสัมพันธ์ระหว่างแต่ละคอลัมน์มาสร้างเป็นเซตเพื่อนำมาใช้เป็นกฎที่ช่วยในการทำนาย ชุดข้อมูลที่ใช้ในการวิจัยเป็นชุดข้อมูลของผู้ป่วยโรคเอดส์ซึ่งข้อมูลส่วนมากจะเป็นข้อมูลของตัวเลขที่กระจัดกระจายกันอยู่ ถ้าข้อมูลที่เป็นตัวเลขจะทำการจัดกลุ่มเป็นช่วงข้อมูล (Discretized) เพื่อง่ายต่อการทำการวิจัยในการหาค่าของข้อมูลสูญหาย ซึ่งเทคนิคการเติมค่าให้กับข้อมูลสูญหายมีค่าความถูกต้องในการทำนาย 74.7%-100%

งานวิจัยของ Jianhua Dai et al. (2011) ได้นำเสนอเทคนิคการเติมค่าให้กับข้อมูลสูญหายโดยการใช้ทฤษฎีราฟเซต (Rough Sets) และได้เพิ่มเทคนิคอื่นเพื่อนำมาเปรียบเทียบ 3 วิธี คือวิธีการตัดแถวข้อมูลที่มีค่าข้อมูลสูญหายออกแล้วจึงทำเหมืองข้อมูล (IO) วิธีการเลือกค่าที่จะนำมาเติมให้กับข้อมูลสูญหายจากข้อมูลที่มีค่าที่ปรากฏบ่อยที่สุดในคอลัมน์นั้น (MC) และวิธีการแปลงข้อมูลทั้งชุดข้อมูลให้เป็นข้อมูลแบบเมตริกดิสเซอร์นิบิลิตี้ (Discernibility matrix) แล้วนำมาสร้างเป็นกฎเพื่อมาทำนายค่าที่สูญหายไป (EDM) การวิจัยของคณะผู้วิจัยนี้ได้ใช้ชุดข้อมูลหกชุดข้อมูลเพื่อนำมาเปรียบเทียบวิธีการในการหาค่าของข้อมูลสูญหายทั้งสามวิธี ซึ่งการผลการทดลองวิธี IO จะมีประสิทธิภาพในการทำนายผลที่แย่ที่สุด ถ้ายิ่งมีค่าของข้อมูลสูญหายมากจะทำให้ค่าความถูกต้องในการทำนายต่ำ เทคนิค MC จะมีประสิทธิภาพที่เสถียรกว่าเทคนิค EDM ซึ่งเทคนิค EDM จะขึ้นอยู่กับประเภทชุดข้อมูลที่ไม่สมบูรณ์

งานวิจัยของ Loris Nanniet et al. (2012) ได้นำเสนอเทคนิคการเติมค่าให้กับข้อมูลสูญหายโดยมีวิธีการในการจัดกลุ่มข้อมูล โดยการจัดกลุ่มข้อมูลจะพิจารณาจากการคำนวณค่าความห่างจากจุดศูนย์กลางของกลุ่ม ซึ่งถ้ามีค่าของข้อมูลสูญหายปรากฏจะพิจารณาค่าใกล้เคียงแล้วใช้ค่า

ใกล้เคียงนั้นทำนายค่าให้กับข้อมูลสูญหายแล้วมาสร้างโมเดลจำแนกข้อมูล (Classifiers) โดยข้อมูลที่ใช้ในการทำงานวิจัยจะใช้ชุดข้อมูลทางด้านการแพทย์เนื่องจากเป็นชุดข้อมูลที่ต้องการความถูกต้องในการทำนายสูง เพื่อความปลอดภัยของผู้ป่วยในการทำนาย ซึ่งงานวิจัยนี้ได้มีการเปรียบเทียบเทคนิคในการเติมค่าให้กับข้อมูลสูญหายหลายเทคนิคทั้งเทคนิค MIE (Multiple imputation ensemble), Mean, NN(Neural network), EM(Expectation–maximization) และการผสมเทคนิคเป็นต้น ซึ่งเทคนิคที่ให้ประสิทธิภาพในการทำนายดีที่สุดจะเป็นเทคนิคที่มีการผสมผสานระหว่างเทคนิค MIE และ EM

จากการศึกษางานวิจัยที่เกี่ยวข้องกับการทำนายค่าของข้อมูลสูญหายพบว่า งานวิจัยส่วนใหญ่นิยมใช้วิธีการทำเหมืองข้อมูล เช่นการสร้างกฎขึ้นมาเพื่อทำนายค่าของข้อมูลสูญหาย แต่อาจจะมีการเพิ่มกระบวนการในการเลือกกฎหรือโมเดลที่จะนำมาใช้งานในการทำนายข้อมูลสูญหาย เมื่อเลือกกฎหรือโมเดลที่เหมาะสมเรียบร้อยแล้วก็จะทำการเติมค่าให้กับข้อมูลสูญหาย โดยแต่ละงานวิจัยก็จะมีวิธีการเปรียบเทียบว่าข้อมูลสูญหายเมื่อเติมเข้าไปมีความถูกต้องมากเพียงใด บางงานวิจัยจะใช้ค่าตัวเลขทางด้านสถิติมาเป็นเกณฑ์ช่วยในการเปรียบเทียบ ซึ่งการทดลองที่ปรากฏในทุก ๆ งานวิจัยจะใช้ชุดข้อมูลที่ถูกแบ่งไว้สำหรับมาใช้ในการทดสอบ ส่วนมากจะแบ่งชุดข้อมูล 70% สำหรับการฝึกสอนและ 30% สำหรับการทดสอบ โดยชุดข้อมูลที่ใช้ในการฝึกสอนจะมีค่าของข้อมูลสูญหายอยู่ บางงานวิจัยจะใช้ข้อมูลจริง บางงานวิจัยจะจำลองให้มีการเกิดค่าของข้อมูลสูญหายขึ้นมาเองเป็นระดับในการเกิดค่าของข้อมูลสูญหาย แต่ชุดข้อมูลที่ใช้ในการตรวจสอบประสิทธิภาพจะไม่มีค่าของข้อมูลสูญหาย งานวิจัยบางงานจะเน้นเจาะจงชุดข้อมูลที่เลือกใช้เพื่อให้เกิดความน่าสนใจ ซึ่งงานวิจัยอื่นที่เกี่ยวข้องจะเป็นการเลือกใช้เทคนิคที่คงที่เพียงหนึ่งหรือสองเทคนิค แต่ในงานวิจัยของวิทยานิพนธ์นี้จะใช้เทคนิคที่ยืดหยุ่นโดยสร้างเป็นโปรแกรมจัดการกับข้อมูลสูญหายซึ่งการจะเลือกใช้เทคนิคใดนั้นจะขึ้นกับประเภทข้อมูล เช่น ข้อมูลที่เป็นข้อความหรือ ข้อมูลเชิงคุณลักษณะ (Categorical) จะใช้การจัดการที่แตกต่างกับข้อมูลที่เป็นตัวเลขหรือ Numeric สาระสำคัญของงานวิจัยนี้สามารถเปรียบเทียบกับงานวิจัยอื่นที่เกี่ยวข้องได้ดังตารางที่ 2.2 ส่วนกระบวนการในการวิจัยจะอธิบายในบทถัดไป

ตารางที่ 2.2 สรุปเปรียบเทียบงานวิจัยที่เกี่ยวข้องกับการเติมค่าให้กับข้อมูลสูญหาย

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง*					
	ก	ข	ค	ง	จ	ฉ
เทคนิคที่จัดการกับข้อมูลสูญหาย						
ตัดแถวที่มีค่าของข้อมูลสูญหายออก			√	√		√
การสร้าง Association Rule มาทำนายค่าข้อมูลสูญหาย	√					
สร้างต้นไม้ตัดสินใจมาทำนายค่าข้อมูลสูญหาย		√				
การใช้ทฤษฎีราฟเซตมาทำนายค่าข้อมูลสูญหาย			√	√		
การเลือกค่าจากข้อมูลที่มีค่าที่ปรากฏบ่อยที่สุด				√		√
การจัดกลุ่มเพื่อหาค่ากึ่งกลางนำมาใช้เติมค่า					√	
การใช้เทคนิคการผสมผสาน - ค่าเฉลี่ยหรือค่ากลางผสมกับค่าปรากฏบ่อย - สมการถดถอยเชิงเส้นผสมกับค่าปรากฏบ่อย						√
การทดสอบประสิทธิภาพในการทำนาย						
การใช้ชุดข้อมูลที่เตรียมสำหรับการทดสอบ	√	√	√	√	√	√
เกณฑ์การวัดประสิทธิภาพ						
วัดจากค่าความถูกต้องความแม่นยำในการทำนาย	√	√	√	√		√
วัดค่าทางด้านสถิติ					√	
การจัดการกับชุดข้อมูล						
การใช้ชุดข้อมูลจริงที่มีค่าบางส่วนสูญหาย		√	√	√		√
การจำลองโดยกำหนดค่าของข้อมูลสูญหายขึ้นมา	√				√	√
การจัดช่วงชุดข้อมูลที่เป็นตัวเลข			√			

*“ก” แทนงานวิจัยของ Jianhua Wu et al. (2007)

“ข” แทนงานวิจัยของ George Ssali and Tshilidzi Marwala (2007)

“ค” แทนงานวิจัยของ Fulufhelo Vincent Nelwamondo and Tshilidzi Marwala (2008)

“ง” แทนงานวิจัยของ Jianhua Dai et al. (2011)

“จ” แทนงานวิจัยของ Loris Nanniet et al. (2012)

“ฉ” แทนงานวิจัยเรื่องการออกแบบและพัฒนาเทคนิคไฮบริดสำหรับการเติมค่าข้อมูลที่สูญหาย (งานวิจัยของวิทยานิพนธ์ฉบับนี้)

บทที่ 3
วิธีการดำเนินการวิจัย

งานวิจัยในบทนี้จะอธิบายถึงวิธีการวิจัย เครื่องมือที่ใช้ และการเปรียบเทียบเทคนิคในการเติมค่าให้กับข้อมูลสูญหายดังนี้

3.1 กรอบแนวคิดของงานวิจัย

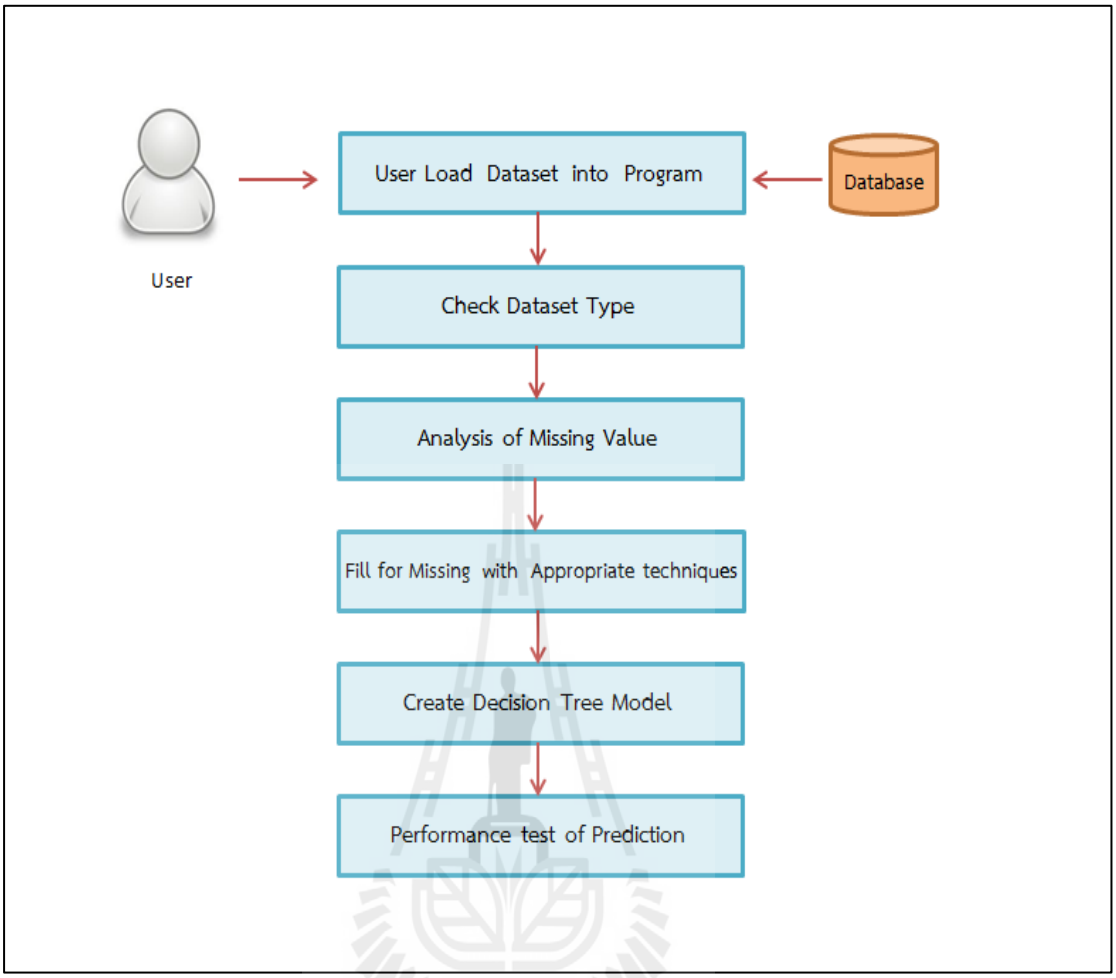
การพัฒนาเทคนิคการทำนายข้อมูลสูญหายด้วยภาษาอาร์ซึ่งเป็นภาษาเชิงฟังก์ชัน มีขั้นตอนในการพัฒนาและศึกษาเปรียบเทียบเทคนิคดังนี้

3.1.1 กรอบแนวคิดของงานวิจัยที่ 1 : การออกแบบระบบการเติมค่าให้กับข้อมูลสูญหาย

เทคนิคที่ใช้เติมค่าให้กับข้อมูลสูญหายมีหลากหลายเทคนิค เช่น การใช้เทคนิคทางการทำเหมืองข้อมูลต่าง ๆ เข้ามาช่วย การหาค่าเฉลี่ย โดยแต่ละเทคนิคจะมีค่าความถูกต้องในการทำนายไม่เท่ากัน งานวิจัยนี้มีเป้าหมายในการออกแบบเทคนิคสำหรับช่วยในการเติมค่าให้กับข้อมูลสูญหาย โดยจะใช้เทคนิคที่นำค่าเฉลี่ย (Mean) ค่ากลาง (Median) และการหาค่าจากสมการถดถอยเชิงเส้น (Linear regression) มาผสมกับเทคนิคค่าที่ปรากฏบ่อย (Mode) หรือเทคนิคการแทนค่าให้ทราบว่าเป็นข้อมูลสูญหาย (Replacement) แล้วให้โปรแกรมสามารถหาค่าของข้อมูลสูญหายได้เองเพื่อหาค่าของข้อมูลสูญหายที่ดีที่สุดดังรูปที่ 3.1

โดยจะอธิบายขั้นตอนการออกแบบเทคนิคที่จะใช้เติมค่าให้กับข้อมูลสูญหายดังนี้

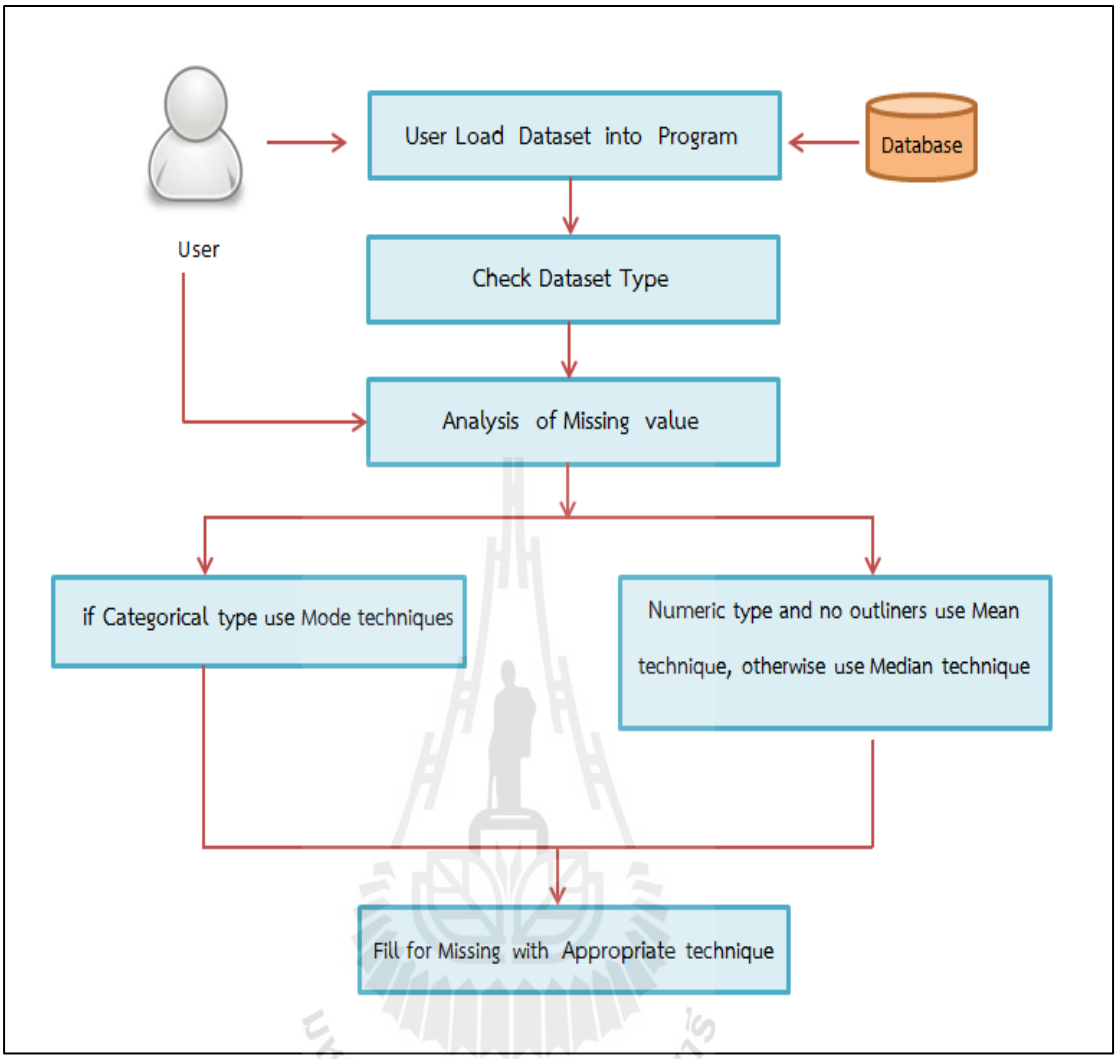
1. ผู้ใช้จะป้อนชุดข้อมูลเข้ามาในโปรแกรมพร้อมกำหนดชนิดแต่ละแอททริบิวต์
2. โปรแกรมจะตรวจสอบชนิดของข้อมูลแต่ละแอททริบิวต์ตามที่ผู้ใช้ป้อนเข้ามา
3. วิเคราะห์แอททริบิวต์ที่ปรากฏข้อมูลสูญหายว่าเป็นประเภทข้อมูลเชิงลักษณะ (Categorical) หรือข้อมูลเชิงตัวเลข (Numerical) และในแอททริบิวต์ที่เป็นข้อมูลเชิงตัวเลขมีข้อมูลผิดปกติ (Outlier) หรือไม่
4. ระบบการเติมค่าอัตโนมัติจะเลือกเทคนิคที่เหมาะสมมาใช้ในการเติมค่าข้อมูลสูญหายภายในคอลัมน์นั้น และระบบนี้ผู้ใช้สามารถเลือกได้ว่าต้องการเลือกเทคนิคอะไรมาเติมค่าให้ข้อมูล
5. สร้างโมเดลแผนภาพต้นไม้ตัดสินใจ (Decision tree) สำหรับใช้เป็นโมเดลทดสอบการเติมค่าให้กับข้อมูลสูญหาย
6. ทดสอบประสิทธิภาพในการทำนายค่าของข้อมูลสูญหาย โดยชุดข้อมูลที่ได้เตรียมไว้สำหรับการทดสอบความถูกต้องในการทำนาย



รูปที่ 3.1 แผนภาพการออกแบบระบบ

3.1.2 กรอบแนวคิดของงานวิจัยที่ 2 : การออกแบบเทคนิคการเติมค่าข้อมูลสูญหายแบบอัตโนมัติ

การออกแบบเทคนิคที่ระบบจะสามารถเติมค่าให้กับข้อมูลสูญหายได้แบบอัตโนมัติจะเป็นการออกแบบการใช้เทคนิคเติมค่าให้กับข้อมูลที่เป็นประเภทข้อมูลเชิงลักษณะ (Categorical) ซึ่งระบบจะกำหนดเทคนิคการเติมค่าที่ปรากฏบ่อยที่สุด (Mode) และถ้าข้อมูลเป็นประเภทข้อมูลเชิงตัวเลข (Numerical) ระบบจะทำการเลือกเทคนิคโดยจะทำการตรวจสอบก่อนว่าข้อมูลมีข้อมูลผิดปกติ (Outlier) หรือไม่ ถ้าข้อมูลมีข้อมูลผิดปกติจะทำการเลือกเทคนิคการเติมค่ากลาง (Median) แต่ถ้าไม่มีข้อมูลผิดปกติจะเลือกเทคนิคในการเติมค่าเฉลี่ย (Mean) ให้กับข้อมูลสูญหาย ซึ่งระบบจะมีกระบวนการทำงานดังรูปที่ 3.2 และสามารถเขียนเป็นขั้นตอนวิธีได้ดังรูปที่ 3.3



รูปที่ 3.2 แผนภาพการออกแบบเทคนิคการเติมค่าข้อมูลสูญหายแบบอัตโนมัติ

Algorithm Automatic Hybrid Technique

Input : Dataset

Output : NEWDataset

(1) Read Dataset

(2) Call function missing-amount ( ) to check missing value in each column

(3) If missing value > 40 % Then remove that column

(4) Check data type of each column

(5) If (data type is numeric) AND (there is outlier) Then call function median_value_impute( )

(6) If (data type is numeric) AND (there is no outlier) Then call function mean_value_impute( )

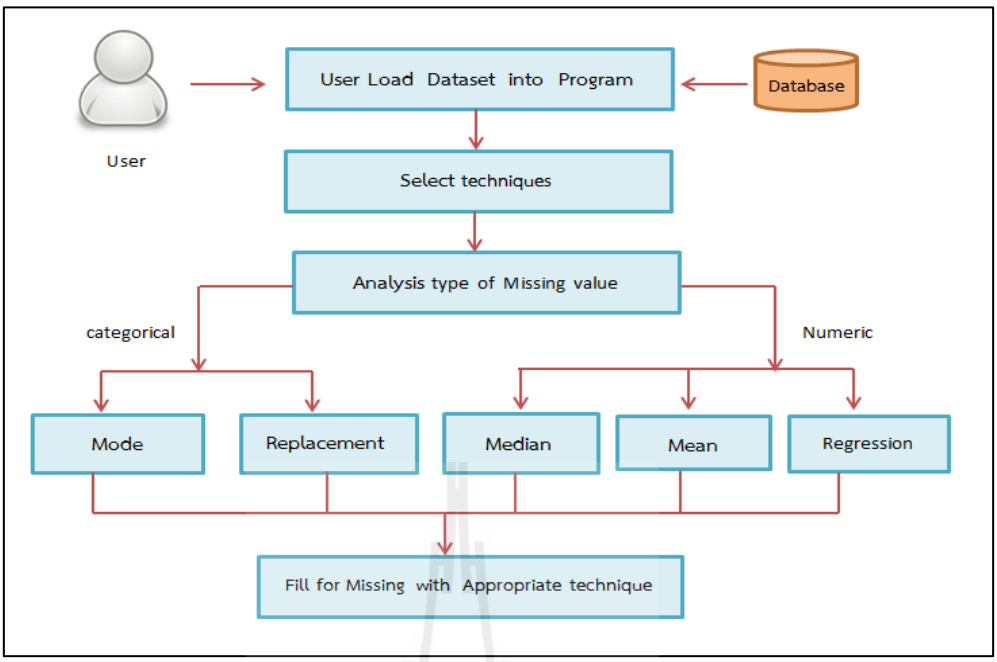
(7) If (data type is categorical) Then call function mode_value_impute ( )

(8) Return NEWDataset

รูปที่ 3.3 ขั้นตอนวิธีเทคนิคการเติมค่าข้อมูลสูญหายแบบอัตโนมัติ

3.1.3 กรอบแนวคิดของงานวิจัยที่ 3 : การออกแบบเทคนิคการเติมค่าข้อมูลสูญหายแบบผู้ใช้กำหนดเอง

การออกแบบเทคนิคที่ระบบจะสามารถให้ผู้ใช้กำหนดเองได้ว่าต้องการเทคนิคชนิดใดมาเติมค่าข้อมูลสูญหาย ซึ่งการออกแบบให้ใช้เทคนิคการเติมค่าให้กับข้อมูลที่เป็นประเภทข้อมูลเชิงลักษณะ (Categorical) ผู้ใช้จะสามารถเลือกเทคนิคการเติมค่าที่ปรากฏบ่อยที่สุด (Mode) หรือเทคนิคการแทนค่ากำกับ และถ้าข้อมูลเป็นประเภทข้อมูลเชิงตัวเลข (Numeric) ผู้ใช้จะสามารถเลือกเทคนิคการเติมค่ากลาง (Median) เทคนิคการเติมค่าเฉลี่ย (Mean) หรือเทคนิคการใช้สมการถดถอยเชิงเส้นในการเติมค่าข้อมูลสูญหายได้ ซึ่งระบบที่ผู้ใช้สามารถเลือกเทคนิคได้ด้วยตนเองจะมีกระบวนการทำงานดังรูปที่ 3.4 และสามารถเขียนเป็นขั้นตอนวิธีได้ดังรูปที่ 3.5



รูปที่ 3.4 แผนภาพการออกแบบเทคนิคการเติมค่าข้อมูลสูญหายแบบผู้ใช้กำหนดเอง

Algorithm User-Specified Technique

Input : Dataset

Output : NEWDataset

(1) Read Dataset

(2) Check Technique that user selects for missing value imputation

(3) Check Type Attribute

(4) Call corresponding function for each selection code

1 : Call function mode_value_impute ( )

2 : Call function replacement _value_impute ( )

3 : Call function median_value_impute ( )

4 : Call function mean_value_impute ( )

5 : Call function regression_value_impute ( )

(5) Return NEWDataset

รูปที่ 3.5 ขั้นตอนวิธีเทคนิคการเติมค่าข้อมูลสูญหายแบบผู้ใช้กำหนด

3.2 การออกแบบเทคนิค

การออกแบบเทคนิคที่จะใช้เป็นเทคนิคในการทำนายค่าของข้อมูลสูญหาย โดยการออกแบบจะใช้เทคนิคที่จะสามารถทำนายค่าของข้อมูลสูญหายได้ทั้งประเภทข้อมูลเชิงลักษณะ (Categorical) และ ประเภทข้อมูลเชิงตัวเลข (Numeric) โดยเทคนิคที่ใช้จะพัฒนามาจากกระบวนการดังต่อไปนี้

3.2.1 การหาค่าเฉลี่ยและค่ากลาง

เป็นการหาค่าของข้อมูลสูญหายที่เป็นข้อมูลประเภทตัวเลข โดยจะอาศัยการหาค่าเฉลี่ยและค่ากลางในคอลัมน์นั้น ๆ โดยจะไม่สนใจพิจารณาคอลัมน์อื่น เทคนิคในการพิจารณาแบบนี้จะเรียกว่าแบบ Unsupervised เป็นแบบไม่มีการฝึกสอน การเลือกว่าในคอลัมน์นั้นควรใช้เทคนิคการหาค่าเฉลี่ยหรือค่ากลางจะพิจารณาการเกิดข้อมูลสูญหาย ซึ่งการตรวจสอบการกระจายโปรแกรมจะเรียกใช้ฟังก์ชัน boxplot.stats ถ้าข้อมูลมีการกระจายแบบปกติจะส่งค่า 0 แต่ถ้าข้อมูลมีการกระจายแบบเอียงจะส่งค่า 1 มายังฟังก์ชันนี้ ถ้าข้อมูลสูญหายมีการกระจายแบบปกติจะใช้การหาค่าเฉลี่ย แต่ถ้าข้อมูลที่มีการเกิดข้อมูลสูญหายมีการกระจายแบบเอียงจะใช้เทคนิคการหาค่ากลาง สามารถเขียนเป็นขั้นตอนวิธีดังรูปที่ 3.6 และตัวอย่างคำสั่งภาษาอาร์ที่ใช้สำหรับเทคนิคการหาค่าเฉลี่ยและค่ากลางดังรูปที่ 3.7

Algorithm MeanMedian

Input : Dataset

Output : NEWDataset.

(1) Read Dataset

(2) column = column contains missing value

(3) Check distribution of colM

(4) If distribution is nomal Then call function mean ( )

(5) If distribution is not nomal Then call function mean ( )

(6) return NEWDataset

รูปที่ 3.6 ขั้นตอนวิธีที่ใช้สำหรับเทคนิคการหาค่าเฉลี่ยและค่ากลาง

```
> imputation.mm<- function(dataM,colM){
+     more=boxplot.stats(dataM[[colM]])$out
+     if(length(more)==0){
+         dataM[is.na(dataM[[colM]]),colM]<-round(mean(dataM[[colM]],na.rm=T))
+     }else{
+         dataM[is.na(dataM[[colM]]),colM]<-round(median(dataM[[colM]],na.rm=T))
+     }
+     return(dataM)
+ }
```

รูปที่ 3.7 คำสั่งภาษาอาร์ที่ใช้สำหรับเทคนิคการหาค่าเฉลี่ยและค่ากลาง

3.2.2 การหาค่าสูญหายด้วยสมการถดถอยเชิงเส้น

เป็นเทคนิคช่วยในการทำนายค่าของข้อมูลสูญหาย โดยจะพิจารณาความน่าจะเป็นของคอลัมน์ที่เกิดค่าความสัมพันธ์ของข้อมูลสูญหายกับคอลัมน์อื่นๆ ที่มีค่ามากที่สุด โดยการเรียกใช้ฟังก์ชัน cor ในภาษาอาร์แล้วจะปรากฏตารางค่าความสัมพันธ์ของแต่ละแอททริบิวต์ ซึ่งในตารางจะปรากฏเป็นสัญลักษณ์ (‘1’ มีค่า 0, ‘ ’ มีค่า 0.3, ‘.’ มีค่า 0.6, ‘+’ มีค่า 0.9, ‘*’ มีค่า 0.95, ‘B’ มีค่า 1) แล้วเลือกคอลัมน์ที่มีความสัมพันธ์มากมาใช้แทนในสมการการถดถอยเชิงเส้นและนำค่าผลลัพธ์ที่ได้จากสมการมาใช้ช่วยในการเติมค่าให้กับข้อมูลสูญหาย ซึ่งเป็นเทคนิคที่เรียกว่าแบบ Supervised หรือแบบมีการฝึกสอน เพราะมีการพิจารณาคอลัมน์อื่นเป็นองค์ประกอบด้วย สามารถเขียนเป็นขั้นตอนวิธีดังรูปที่ 3.8 และตัวอย่างคำสั่งภาษาอาร์ที่ใช้สำหรับเทคนิคการหาค่าด้วยสมการถดถอยเชิงเส้นแสดงดังรูปที่ 3.9

Algorithm Linear regression

Input : Dataset

Output : NEWDataset

(1) Read Dataset

(2) column = a column contains missing value

(3) Find the column C that most correlates to column

(4) call function mean linear regression ( ) to find coefficients in the equation

y =a+bx

(5) Use the results to fill missing values in colM

(6) return NEWDataset

รูปที่ 3.8 ขั้นตอนวิธีที่ใช้สำหรับเทคนิคการหาค่าสมการถดถอยเชิงเส้น

```
> creXY<- function(colM,dataM,NN){
+     mM<-lm(colM,data=dataM)$coefficients[NN]
+     mN<-mM[1][[1]]
+     return(mN)
+ }
> inputf<- function(oP){
+     if ( is.na(oP) ) return(NA)
+     else return ( (oP+(mY))/mX )
+ }
> line.input<- function(colA,colB,dataM){
+     dataM[ is.na ( dataM[[colA]] ),colA ] <-
+         round(sapply ( dataM[ is.na (dataM[[colA]]),colB],inputf))
+     return(dataM)
+ }
```

รูปที่ 3.9 คำสั่งภาษาอาร์ที่ใช้สำหรับเทคนิคการหาค่าสูญหายที่คำนวณได้จากสมการถดถอยเชิงเส้น

3.2.3 การหาค่าที่ปรากฏซ้ำบ่อยที่สุด

เป็นเทคนิคที่หาค่าของข้อมูลสูญหายที่เป็นข้อมูลประเภทข้อมูลเชิงลักษณะ (Categorical) โดยจะพิจารณาว่าข้อความใดปรากฏบ่อยที่สุดในคอลัมน์นั้นจะใช้ค่านั้นมาช่วยเติมค่าที่สูญหาย และเป็นเทคนิคที่เรียกว่าแบบ Unsupervised หรือเป็นแบบไม่มีการฝึกสอนซึ่งจะไม่

สนใจค่าในคอลัมน์อื่น สามารถเขียนเป็นขั้นตอนวิธีดังรูปที่ 3.10 และคำสั่งภาษาอาร์ที่ใช้สำหรับเทคนิคการหาค่าที่ปรากฏซ้ำบ่อยสุดดังรูปที่ 3.11

Algorithm Mode

Input : Dataset, Column

Output : NEWDataset

(1) DataM=Dataset ,ColM= Column

(2) If column contains missing value {

(3) Call function to find the value that appears most often in column

(4) Use the most frequent value to fill missing value in dataset }

(5) return NEWDataset

รูปที่ 3.10 ขั้นตอนวิธีที่ใช้สำหรับเทคนิคการหาค่าปรากฏซ้ำบ่อยสุด

```
> val.mode<- function (x) {
+     if (is.factor(x)) levels(x) [which.max(table(x))]
+     else {
+         f <- as.factor(x)
+         levels(f) [ which.max(table(f)) ]
+     }
+ }
>
> imputation.mode<- function(dataM,colM) {
+     v.mode<-val.mode(dataM[[colM]])
+     dataM[is.na(dataM[[colM]]),colM]<-v.mode
+     return(dataM)
+ }
>
```

รูปที่ 3.11 คำสั่งภาษาอาร์ที่ใช้สำหรับเทคนิคการหาค่าที่ปรากฏซ้ำบ่อยสุด

3.2.4 การแทนค่ากำกับให้ข้อมูลสูญหาย

เป็นการนำค่ากำกับให้ข้อมูลสูญหายเช่นคำว่า (Missing, Not, NN) เพื่อให้กระบวนการนำไปใช้งานไม่เกิดปัญหาสามารถทราบได้ว่านี้คือข้อมูลสูญหายที่เกิดขึ้นกับชุดข้อมูล และเมื่อนำไปสร้างโมเดลก็จะทราบทันทีว่าข้อมูลในจุดนั้นเกิดข้อผิดพลาดในชุดข้อมูลนี้มีผลกับ

การสร้างโมเดลเพียงใด ซึ่งเขียนเป็นขั้นตอนวิธีดังรูปที่ 3.12 และคำสั่งภาษาอาร์ที่ใช้สำหรับเทคนิคการแทนค่ากำกับข้อมูลให้กับข้อมูลสูญหายดังรูปที่ 3.13

Algorithm Replacement

Input : Dataset, Colum

Output : NEWDataset

(1) DataM = Dataset, ColM = Columm

(2) If variable of column is missing value {

(3) Call function to add factor that is name missing in column

(4) Use variable is missing to replacement missing value in dataset }

(5) return Dataset is complete

รูปที่ 3.12 ขั้นตอนวิธีที่ใช้สำหรับเทคนิคการหาค่าปรากฏซ้ำบ่อยสุด

```
> aLevel<- function(dataM,colM) {
+     levels(dataM[[colM]])<-c(levels(dataM[[colM]]),"missing")
+     return(levels(dataM[[colM]]))
+ }
>
> imputation.mis<- function(dataM,colM) {
+     var="missing"
+     dataM[is.na(dataM[[colM]]),colM]<-var
+     return(dataM)
+ }
```

รูปที่ 3.13 คำสั่งภาษาอาร์ที่ใช้สำหรับเทคนิคการแทนค่ากำกับข้อมูลให้กับข้อมูลสูญหาย

เทคนิคที่ได้ออกแบบขึ้นมาเพื่อทำนายค่าของข้อมูลสูญหาย จะเป็นการผสมผสานหลายเทคนิคซึ่งจะอธิบายการออกแบบเทคนิคดังนี้

3.2.5 เทคนิคแบบผสมผสาน

การออกแบบเทคนิคแบบผสมผสานนี้จะใช้เทคนิคสองแบบผสมกันทั้งการหาค่า

แบบ Unsupervised โดยใช้ค่าเฉลี่ย ค่ากลาง เมื่อเป็นประเภทข้อมูลเชิงตัวเลข (Numeric) แต่ถ้าเป็นประเภทข้อมูลเชิงลักษณะ (Categorical) จะใช้การหาค่าที่ปรากฏบ่อยที่สุด ถ้าการหาค่าแบบ Supervised ซึ่งเป็นประเภทข้อมูลเชิงตัวเลขจะใช้การหาค่าจากสมการถดถอยเชิงเส้น (Linear regression) ของคอลัมน์ที่เกิดข้อมูลสูญหายกับคอลัมน์อื่นที่มีค่าความน่าจะเป็นมากที่สุด และถ้าเป็นประเภทข้อมูลเชิงลักษณะจะใช้การแทนค่ากำกับให้ข้อมูลสูญหาย โดยการออกแบบเทคนิคจะให้สามารถเลือกได้ว่าถ้าเกิดข้อมูลสูญหายในคอลัมน์นี้จะเลือกใช้เทคนิคอะไรที่เหมาะสมในการหาค่าของข้อมูลสูญหายและผู้ใช้สามารถเลือกเทคนิคที่ต้องการในการแทนค่าให้กับข้อมูลสูญหายด้วยตนเองได้อีกด้วย เพื่อให้การหาค่าของข้อมูลสูญหายมีประสิทธิภาพที่ดีขึ้น

3.3 การใช้งานระบบเพื่อเติมค่าให้ข้อมูลสูญหาย

ส่วนนี้จะอธิบายการใช้งานระบบสำหรับเติมค่าที่สูญหายด้วยเทคนิคการใช้ค่าเฉลี่ย (Mean) ค่ากลาง (Median) ค่าที่ปรากฏบ่อย (Mode) ค่าสมการถดถอยเชิงเส้น (Linear regression) และการแทนค่ากับกำข้อมูลสูญหาย เพื่อให้เข้าใจถึงกระบวนการการทำงานของระบบในการเติมค่าให้กับข้อมูลสูญหาย

3.3.1 การเตรียมชุดข้อมูลมาใช้เพื่อเติมค่าข้อมูลสูญหาย

ข้อมูลตัวอย่างจะเป็นข้อมูลสำหรับการวินิจฉัยเนื้องอกว่าสามารถเป็นโรคมะเร็งหรือไม่ ข้อมูลคอลัมน์มี Gender ระบุเพศของผู้ป่วย BI-RADS ระบุตัวเลขระดับของ BI-RADS ซึ่งมี 5 ระดับ(1-5) Age ระบุอายุของผู้ป่วย Shape ระบุรูปร่างของเนื้องอก (round=1, oval=2, lobular=3, irregular=4) Margin ระบุบริเวณโดยรอบของเนื้องอก (circumscribed=1, microlobulated=2, obscured=3, ill-defined=4, spiculated =5) Density ระบุความหนาแน่นของเนื้องอก (high=1, iso=2, low=3, fat-containing=4) Severity ระบุผลการวินิจฉัยการเป็นมะเร็ง (benign=0 or malignant=1) โดยการนำชุดข้อมูลเข้ามาใช้งานในโปรแกรมจะใช้คำสั่งนำชุดข้อมูลมาใช้งานในโปรแกรม ตามตัวอย่างในรูปที่ 3.14 สัญลักษณ์ ? แทนค่าของข้อมูลสูญหาย เมื่อข้อมูลถูกนำเข้าในโปรแกรมแล้วจะแสดงรายละเอียดข้อมูลดังรูปที่ 3.15 ค่าที่ปรากฏ NA ในคอลัมน์จะแทนข้อมูลสูญหาย

```
@RELATION dataex
@ATTRIBUTE Gender {F,M}
@ATTRIBUTE BI-RADS   NUMERIC
@ATTRIBUTE Age       NUMERIC
@ATTRIBUTE Shape     NUMERIC
@ATTRIBUTE Margin    NUMERIC
@ATTRIBUTE Density   NUMERIC
@ATTRIBUTE Severity  {0,1}

@DATA
F,4,40,1,?,?,0
M,?,66,?,?,1,1
?,4,08,4,3,1,1
M,4,43,1,?,?,0
F,5,?,4,4,3,1
F,4,59,2,4,3,1
M,2,42,?,?,4,0
F,5,67,4,5,3,1
M,4,74,2,1,2,0
F,5,80,3,5,3,1
```

รูปที่ 3.14 ชุดข้อมูลตัวอย่างชนิดไฟล์.arff

<u>คำสั่งในการนำชุดข้อมูลเข้ามาใช้งานในโปรแกรม</u>

```
> dataex<-read.arff("dataex.arff")
```

เป็นการอ่านไฟล์นามสกุล.arff ซึ่งในวงเล็บจะเป็นการใส่ชื่อไฟล์ของชุดข้อมูล และจัดเก็บไฟล์ของชุดข้อมูลที่อ่านแล้วมาเก็บในตัวแปรชื่อ dataex

	Gender	BI-RADS	Age	Shape	Margin	Density	Severity
1	F	4	40	1	NA	NA	0
2	M	NA	66	NA	NA	1	1
3	NA	4	8	4	3	1	1
4	M	4	43	1	NA	NA	0
5	F	5	NA	4	4	3	1
6	F	4	59	2	4	3	1
7	M	2	42	NA	NA	4	0
8	F	5	67	4	5	3	1
9	M	4	74	2	1	2	0
10	F	5	80	3	5	3	1

รูปที่ 3.15 ตัวอย่างชุดข้อมูลถูกแปลงเพื่อนำมาใช้งานในโปรแกรม

3.3.2 การนำชุดข้อมูลที่เตรียมมาใช้งานในระบบ

1. การใช้งานระบบแบบอัตโนมัติ

เมื่อเตรียมชุดข้อมูลที่จะใช้ในการเติมค่าให้กับข้อมูลสูญหายแล้ว จะนำชุดข้อมูลเข้ามาในระบบโดยจะใช้คำสั่งเรียกใช้งานระบบเลือกแบบอัตโนมัติ ส่วนนี้โปรแกรมจะกำหนดเทคนิคการเติมค่าให้กับข้อมูลสูญหายแบบอัตโนมัติ ระบบจะตรวจสอบประเภทของข้อมูลและตรวจสอบการกระจายตัวของข้อมูล เพื่อเติมค่าที่เหมาะสมแล้วแสดงผลลัพธ์ของข้อมูลดังรูปที่ 3.15

<u>คำสั่งในการนำชุดข้อมูลมาใช้งานในระบบเลือกแบบอัตโนมัติ</u>

```
> hybrid.Technique(Dataset)
```

จะเป็นการเรียกใช้งานระบบผ่านฟังก์ชันชื่อว่า hybrid.Technique มีพารามิเตอร์เป็นชื่อชุดข้อมูลที่ต้องการนำมาเติมค่าให้กับข้อมูลสูญหาย

	Gender	BI-RADS	Age	Shape	Margin	Density	Severity
1	F	4	40	1	4	2	0
2	M	4	66	3	4	1	1
3	F	4	8	4	3	1	1
4	M	4	43	1	4	2	0
5	F	5	59	4	4	3	1
6	F	4	59	2	4	3	1
7	M	2	42	3	4	4	0
8	F	5	67	4	5	3	1
9	M	4	74	2	1	2	0
10	F	5	80	3	5	3	1

รูปที่ 3.16 ตัวอย่างชุดข้อมูลที่ระบบมีการเติมค่าของมูลที่สูญหายแบบอัตโนมัติ

จากรูปที่ 3.15 ชุดข้อมูลที่ถูกเติมค่าข้อมูลสูญหายให้แบบอัตโนมัติระบบจะวิเคราะห์ชนิดของข้อมูลเช่น ค่าข้อมูลสูญหายในคอลัมน์ Gender ซึ่งเป็นประเภทข้อมูลเชิงลักษณะ (Categorical) ระบบอัตโนมัติจะเลือกใช้ค่า Mode ส่วนข้อมูลสูญหายในคอลัมน์อื่นจะเป็นประเภทข้อมูลเชิงตัวเลข (Numeric) ระบบจะตรวจสอบหาข้อมูลผิดปกติ (Outlier) ถ้ามีข้อมูลผิดปกติ (Outlier) ปรากฏเช่นที่เกิดในคอลัมน์ Age ระบบจะเลือกใช้ค่า Median แต่ถ้าข้อมูลมีการกระจายตัวปกติจะใช้ค่าเฉลี่ย (Mean) เติมค่าให้กับข้อมูลสูญหาย

2. การใช้งานระบบแบบผู้ใช้กำหนดเอง

การใช้งานแบบผู้ใช้กำหนดเองจะทำการเติมค่าให้ข้อมูลสูญหายเฉพาะคอลัมน์ที่ผู้ใช้กำหนด และผู้ใช้ต้องรู้ว่าคอลัมน์ที่เกิดข้อมูลสูญหายเป็นประเภทข้อมูลเชิงลักษณะ (Categorical) หรือ ประเภท ข้อมูลเชิงตัวเลข (Numeric) ถ้าข้อมูลสูญหายเป็นประเภทข้อมูลเชิงลักษณะ (Categorical) ก็สามารถกำหนดได้ว่าต้องการใช้ค่าที่ปรากฏบ่อย (Mode) หรือ การแทนค่ากำกับลงไปในข้อมูลสูญหาย แต่ถ้าเป็นประเภทข้อมูลเชิงตัวเลข (Numeric) ก็สามารถกำหนดได้ว่าต้องการใช้ค่าเฉลี่ย (Mean) ค่ากลาง (Median) หรือการใช้สมการถดถอยเชิงเส้น (Linear regression) การเติมค่าให้กับข้อมูลประเภทข้อมูลเชิงลักษณะ (Categorical) ด้วยค่าที่ปรากฏบ่อยจะใช้คำสั่งการใช้งานแบบเลือกเทคนิคค่าที่ปรากฏบ่อยแล้วแสดงผลลัพธ์ดังรูปที่ 3.17 หรือถ้าต้องการกำกับค่าให้ข้อมูลสูญหายใช้คำสั่งการใช้งานแบบเลือกเทคนิคกำกับค่าให้ข้อมูลสูญหายแล้วแสดงผลลัพธ์ดังรูปที่ 3.18 การเติมค่าให้กับข้อมูลประเภทข้อมูลเชิงตัวเลข (Numeric) ด้วยค่าเฉลี่ยจะใช้คำสั่งการใช้งานซึ่งเลือกเทคนิคค่าเฉลี่ยแล้วแสดงผลลัพธ์ดังรูปที่ 3.19 ถ้าต้องการเติมด้วยค่ากลางให้ข้อมูลสูญหายใช้คำสั่งการใช้งานแบบเลือกเทคนิคการใช้ค่ากลาง แล้วแสดงผลลัพธ์ดังรูปที่ 3.21 และถ้าต้องการเติมค่าให้กับข้อมูลสูญหายด้วยการใช้สมการถดถอยเชิงเส้น เพื่อดูความสัมพันธ์ของคอลัมน์ที่ต้องการเติมค่าให้กับข้อมูลสูญหาย แล้วผลลัพธ์แสดงดังรูปที่ 3.20 และผู้ใช้ต้องพิจารณาความสัมพันธ์แล้วสร้างตัวแปรที่ใช้ในสมการถดถอยเชิงเส้น เพื่อเลือกคอลัมน์เป้าหมายและคอลัมน์ที่มีความสัมพันธ์กันมากที่สุดดังรูปที่ 3.21 การเรียกใช้คำสั่งจะเรียกชื่อฟังก์ชัน hybrid.User ซึ่งมีพารามิเตอร์ 4 ตัว ตัวแรกจะเป็นพารามิเตอร์ของชุดข้อมูล ตัวที่สองจะเป็นเทคนิคที่ผู้ใช้ต้องการ (1 = Mode, 2 = Replacement, 3 = Mean, 4 = Median, 5 = Linear regression) พารามิเตอร์ที่สามจะเป็นคอลัมน์เป้าหมายในการเติมค่าข้อมูลสูญหาย พารามิเตอร์ที่สี่จะเป็นคอลัมน์ที่ช่วยในการเติมค่าข้อมูลสูญหายของเทคนิคการใช้สมการเชิงเส้น แต่ถ้าผู้ใช้เลือกเทคนิคไม่เหมาะสมกับชนิดของแต่ละแอททริบิวต์แล้วโปรแกรมจะไม่เติมค่าให้กับข้อมูลที่สูญหายให้กับชุดข้อมูลนั้น

<u>คำสั่งการใช้งานแบบเลือกเทคนิคค่าที่ปรากฏบ่อยที่สุด</u>

```
> hybrid.Technique(Dataset,1,"Gender",not)
```

	Gender	BI-RADS	Age	Shape	Margin	Density	Severity
1	F	4	40	1	NA	NA	0
2	M	NA	66	NA	NA	1	1
3	F	4	8	4	3	1	1
4	M	4	43	1	NA	NA	0
5	F	5	NA	4	4	3	1
6	F	4	59	2	4	3	1
7	M	2	42	NA	NA	4	0
8	F	5	67	4	5	3	1
9	M	4	74	2	1	2	0
10	F	5	80	3	5	3	1

รูปที่ 3.17 ผลลัพธ์จากการเลือกเทคนิคค่าที่ปรากฏบ่อยที่สุด

คำสั่งการใช้งานแบบเลือกเทคนิคกำกับค่าให้ข้อมูลสูญหาย

```
> hybrid.Technique(Dataset,2,"Gender",not)
```

	Gender	BI-RADS	Age	Shape	Margin	Density	Severity
1	F	4	40	1	NA	NA	0
2	M	NA	66	NA	NA	1	1
3	missing	4	8	4	3	1	1
4	M	4	43	1	NA	NA	0
5	F	5	NA	4	4	3	1
6	F	4	59	2	4	3	1
7	M	2	42	NA	NA	4	0
8	F	5	67	4	5	3	1
9	M	4	74	2	1	2	0
10	F	5	80	3	5	3	1

รูปที่ 3.18 ผลลัพธ์จากการเลือกเทคนิคกำกับค่าให้ข้อมูลสูญหาย

คำสั่งการใช้งานแบบเลือกเทคนิคการใช้ค่ากลาง

```
> hybrid.Technique(Dataset,3,"Age",not)
```

	Gender	BI-RADS	Age	Shape	Margin	Density	Severity
1	F	4	40	1	NA	NA	0
2	M	NA	66	NA	NA	1	1
3	NA	4	8	4	3	1	1
4	M	4	43	1	NA	NA	0
5	F	5	59	4	4	3	1
6	F	4	59	2	4	3	1
7	M	2	42	NA	NA	4	0
8	F	5	67	4	5	3	1
9	M	4	74	2	1	2	0
10	F	5	80	3	5	3	1

รูปที่ 3.19 ผลลัพธ์จากการเลือกเทคนิคการใช้ค่ากลาง

คำสั่งการใช้งานแบบเลือกเทคนิคการใช้ค่าเฉลี่ย

```
> hybrid.Technique(Dataset,4,"Age",not)
```

	Gender	BI-RADS	Age	Shape	Margin	Density	Severity
1	F	4	40	1	NA	NA	0
2	M	NA	66	NA	NA	1	1
3	NA	4	8	4	3	1	1
4	M	4	43	1	NA	NA	0
5	F	5	53	4	4	3	1
6	F	4	59	2	4	3	1
7	M	2	42	NA	NA	4	0
8	F	5	67	4	5	3	1
9	M	4	74	2	1	2	0
10	F	5	80	3	5	3	1

รูปที่ 3.20 ผลลัพธ์จากการเลือกเทคนิคการใช้ค่าเฉลี่ย

```
> lookCor(datatest[,2:6])
        B A S M D
BI-RADS 1
Age     . 1
Shape   . . 1
Margin  ,   . 1
Density , +   , 1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

รูปที่ 3.21 คำสั่งนำข้อมูลมาตรวจสอบความสัมพันธ์ระหว่างคอลัมน์ที่มีค่าสูญหายกับคอลัมน์อื่น

คำสั่งการกำหนดตัวแปรที่ใช้ในสมการถดถอยเชิงเส้น

```
> mX<-creXY(Age ~ Density,datatest,2)
> mY<-creXY(Age ~ Density,datatest,1)
```

คำสั่งการใช้งานแบบเลือกเทคนิคการใช้สมการถดถอยเชิงเส้น

```
> hybrid.Technique(Dataset,5,"Age","Density")
```

	Gender	BI-RADS	Age	Shape	Margin	Density	Severity
1	F	4	40	1	NA	NA	0
2	M	NA	66	NA	NA	1	1
3	NA	4	8	4	3	1	1
4	M	4	43	1	NA	NA	0
5	F	5	60	4	4	3	1
6	F	4	59	2	4	3	1
7	M	2	42	NA	NA	4	0
8	F	5	67	4	5	3	1
9	M	4	74	2	1	2	0
10	F	5	80	3	5	3	1

รูปที่ 3.22 ผลลัพธ์จากการเลือกเทคนิคการใช้สมการถดถอยเชิงเส้น

จากรูปที่ 3.21 จะปรากฏตารางความสัมพันธ์ซึ่งมีสัญลักษณ์ปรากฏเค่าแทนด้วยสัญลักษณ์ (‘1’ มีค่า 0, ‘ ’ มีค่า 0.3, ‘.’ มีค่า 0.6, ‘+’ มีค่า 0.9, ‘*’ มีค่า 0.95, ‘B’ มีค่า 1) สังเกตได้ว่าคอลัมน์ Density มีความสัมพันธ์กับคอลัมน์ Age สูงที่สุดประมาณ 0.9 จึงเป็นสาเหตุที่ทำให้ต้องสร้างสมการถดถอยเชิงเส้นโดยการใช้คอลัมน์ Density เข้ามาช่วยในการคำนวณค่าให้กับข้อมูลสูญหายในคอลัมน์ของ Age และการใช้คำสั่งเรียกฟังก์ชัน creXY เพื่อการกำหนดตัวแปร a, b ที่ใช้ในสมการ y=a+bx ส่วน x ระบบจะทำการเลือกค่าจากแถวเดียวกันที่เกิดข้อมูลสูญหายใสคอลัมน์ที่ผู้ใช้เลือกมาสร้างความสัมพันธ์ ซึ่งในข้อมูลตัวอย่างที่ใช้จะเท่ากับ y = 42+6(5)

3.4 เครื่องมือที่ใช้ในการวิจัย

3.4.1 เครื่องมือที่ใช้พัฒนาระบบ

1. เครื่องคอมพิวเตอร์ที่ใช้สำหรับการพัฒนาเทคนิคในการทำนายค่าของข้อมูลสูญหาย โดยมีรายละเอียดดังนี้

- หน่วยประมวลผลกลาง : Intel(R) Core(TM) i7-3612QM CPU @2.10GHz

- ฮาร์ดดิสก์ : 750 GB

- หน่วยความจำหลัก : 4.00 GB

- อุปกรณ์เสริมอื่นๆ เช่น เมาส์ แป้นพิมพ์ เป็นต้น

2. ระบบปฏิบัติการและโปรแกรมที่ใช้ในการเขียนโปรแกรมให้กับเทคนิคการเติมข้อมูลสูญหายที่ออกแบบ

- ระบบปฏิบัติการ Windows 7

- โปรแกรมสำหรับการใช้เขียนโปรแกรมภาษาอาร์ RStudio V.0.97.311

3.4.2 เครื่องมือที่ใช้วัดประสิทธิภาพ

เครื่องมือที่ใช้วัดประสิทธิภาพในงานวิจัยนี้จะประกอบด้วยไฟล์โปรแกรม ไฟล์ชุดข้อมูลสำหรับการเปรียบเทียบเทคนิคการเติมข้อมูลว่าเทคนิคใดมีประสิทธิภาพที่ดีกว่า โดยรายละเอียดในการวัดประสิทธิภาพ จะใช้ค่าความถูกต้องในการทำนายของโมเดลเป็นตัวชี้วัดประสิทธิภาพว่า โมเดลที่สร้างจากเทคนิคการเติมค่าแบบใดให้ค่าความถูกต้องมากที่สุดซึ่งจะแสดงว่าเทคนิคที่ใช้เติมค่านั้นดี โมเดลจะสร้างโดยใช้ชุดข้อมูลในการฝึกสอนที่ผ่านการเติมค่าสูญหายแล้ว และในการทดสอบประสิทธิภาพของโมเดลจะใช้ชุดข้อมูลสำหรับการทดสอบที่ได้แบ่งไว้ (วิกิพีเดีย สารานุกรมเสรี, 2555ค) ในส่วนของมาตรวัดความถูกต้อง (Relative Error) คำนวณจากสมการต่อไปนี้

$$\text{ค่าความถูกต้อง} = \frac{\text{ค่าที่ทำนายถูก}}{\text{ค่าทำนายทั้งหมด}} \times 100 \qquad (3.1)$$

บทที่ 4
การทดสอบและอภิปรายผล

การทดสอบประสิทธิภาพของเทคนิคการทำนายข้อมูลที่สูญหายในระบบนั้น จะทดสอบประสิทธิภาพประเภทการจำแนกข้อมูล (Classification) ด้วยการนำข้อมูลที่ทำการเติมข้อมูลสูญหายแล้วไปสร้างเป็นโมเดลต้นไม้ตัดสินใจ (Decision Tree) และตรวจสอบความถูกต้องในการทำนายของโมเดล เพื่อนำความถูกต้องของแต่ละโมเดลมาทำการเปรียบเทียบประสิทธิภาพ

4.1 ข้อมูลสำหรับการทดสอบประสิทธิภาพ

4.1.1 ชุดข้อมูลที่มีข้อมูลสูญหายเกิดขึ้นจริง

การทดสอบเทคนิคในการเติมค่าข้อมูลที่สูญหายในระบบ ชุดข้อมูลที่ใช้ในการทดสอบประสิทธิภาพคือชุดข้อมูลของโรคหัวใจ (Heart Disease) ซึ่งสามารถดาวน์โหลดชุดข้อมูลได้จากเว็บไซต์ของ UCI ที่ http://archive.ics.uci.edu/ml/datasets/Heart+Disease มีจำนวนแอททริบิวต์ 14 แอททริบิวต์ แต่ชุดข้อมูลโรคหัวใจนี้จะแบ่งข้อมูลเป็นชุดข้อมูลย่อย ๆ 4 ชุดข้อมูล โดยสามารถสรุปรายละเอียดของชุดข้อมูลนี้ได้ดังในตารางที่ 4.1 ซึ่งลักษณะของชุดข้อมูลแบบผสมผสาน (Mixed) นั้นจะมีข้อมูลทั้งเป็นประเภทข้อมูลเชิงลักษณะ (Categorical) และข้อมูลเชิงตัวเลข (Numerical) อยู่ภายในชุดข้อมูล และยังมีข้อมูลที่สูญหายเกิดขึ้นในชุดข้อมูลโรคหัวใจนี้ด้วย จึงเป็นชุดข้อมูลที่เหมาะสมในการนำมาใช้ทดสอบประสิทธิภาพ ดังรูปตัวอย่างของข้อมูลภายในชุดข้อมูลจะแสดงได้ดังรูปที่ 4.1 และจะใช้ข้อมูลทั้งหมด 70 %ในการฝึกสอน (Train data) และจะสุ่มข้อมูล 30% เพื่อใช้เป็นข้อมูลทดสอบ (Test data) โมเดลที่สร้างขึ้นภายหลังจากการทำนายข้อมูลที่สูญหาย

ตารางที่ 4.1 รายละเอียดของชุดข้อมูลย่อยของโรคหัวใจ

ชุดข้อมูลย่อย	จำนวนอินสแตนซ์	จำนวนเรคคอร์ดของข้อมูลสูญหาย	จำนวนข้อมูลฝึกสอน	จำนวนข้อมูลทดสอบ
Cleveland	303	6	213	90
Hungary	294	293	206	88
Switzerland	123	123	86	37
Va	200	199	140	60

```
@RELATION HeartDiseaseVa
@ATTRIBUTE Age        NUMERIC
@ATTRIBUTE Sex        {0,1}
@ATTRIBUTE Cp         {1,2,3,4}
@ATTRIBUTE Trestbps   NUMERIC
@ATTRIBUTE Chol       NUMERIC
@ATTRIBUTE Fbs        {0,1}
@ATTRIBUTE Restecg    {0,1,2}
@ATTRIBUTE Thalach    NUMERIC
@ATTRIBUTE Exang      {0,1}
@ATTRIBUTE Oldpeak    NUMERIC
@ATTRIBUTE Slope      {1,2,3}
@ATTRIBUTE Ca         {0,1,2,3}
@ATTRIBUTE Thal       {3,6,7}
@ATTRIBUTE Num        {0,1,2,3,4}
@DATA
48,1,3,132,220,1,1,162,0,0,?,?,6,1
61,1,1,142,200,1,1,100,0,1.5,3,?,?,3
66,1,4,112,261,0,0,140,0,1.5,1,?,?,1
68,1,1,?,181,1,1,?,?,?,?,?,?,0
55,1,4,172,260,0,0,73,0,2,?,?,?,3
62,1,3,120,220,0,2,86,0,0,?,?,?,0
71,1,3,?,221,0,0,?,?,?,?,?,?,3
53,1,3,155,175,1,1,160,0,?,?,?,6,0
58,1,3,150,219,0,1,118,1,0,?,?,?,2
75,1,4,160,310,1,0,112,1,2,3,?,7,0
56,1,3,?,208,1,1,?,?,?,?,?,?,4
58,1,3,?,232,0,1,?,?,?,?,?,?,2
64,1,4,134,273,0,0,102,1,4,3,?,?,4
54,1,3,?,203,0,1,?,?,?,?,?,?,0
54,1,2,?,182,0,1,?,?,?,?,?,?,0
59,1,4,140,274,0,0,154,1,2,2,?,?,0
55,1,4,?,204,1,1,?,?,?,?,?,?,1
57,1,4,144,270,1,1,160,1,2,2,?,?,3
61,1,4,?,292,0,1,?,?,?,?,?,?,3
41,1,4,150,171,0,0,128,1,1.5,2,?,?,0
71,1,4,130,221,0,1,115,1,0,?,?,?,3
38,1,4,110,289,0,0,105,1,1.5,3,?,?,1
57,1,4,122,264,0,2,100,0,0,?,?,?,1
56,1,4,128,223,0,1,119,1,2,3,?,?,2
64,1,4,150,193,0,1,135,1,0.5,2,?,?,2
72,1,4,160,?,1,2,130,0,1.5,?,?,?,2
69,1,4,122,216,1,2,84,1,0,?,?,7,2
61,1,4,190,287,1,2,150,1,2,3,?,?,4
64,1,4,130,258,1,2,130,0,0,?,?,6,2
58,1,4,160,256,1,2,113,1,1,1,?,?,3
```

รูปที่ 4.1 ตัวอย่างชุดข้อมูลโรคหัวใจที่ใช้ในการทดสอบประสิทธิภาพเทคนิคการเติมข้อมูลสูญหาย

4.1.2 ชุดข้อมูลที่นำมาสำหรับเพิ่มจำนวนข้อมูลสูญหาย

ชุดข้อมูลสำหรับการทดสอบประสิทธิภาพด้วยการเพิ่มข้อมูลสูญหายให้กับชุดข้อมูลจริง โดยจะแบ่งการสุ่มเพิ่มข้อมูลสูญหายของเรคคอร์ดออกเป็น 10%, 20%, 30%, 40% และ 50% ด้วยได้กำหนดมีข้อมูลที่สูญหายเกิดขึ้น 5 แอททริบิวต์ ชุดข้อมูลที่เลือกใช้ทดสอบเป็นชุดข้อมูลที่เกี่ยวกับโรคผิวหนัง (Dermatology) และสามารถดาวน์โหลดชุดข้อมูลได้จากเว็บไซต์ http://archive.ics.uci.edu/ml/datasets/Dermatology มีจำนวนอินสแตนซ์ทั้งหมด 366 เรคคอร์ด และมีจำนวนแอททริบิวต์ 35 แอททริบิวต์ ซึ่งตัวอย่างของชุดข้อมูลโรคผิวหนังจะแสดงดังรูปที่ 4.2 และชุดข้อมูลนี้มีลักษณะข้อมูลแบบผสมของข้อมูลเชิงลักษณะ(Categorical) และข้อมูลเชิงตัวเลข (Numerical) โดยชุดข้อมูลชุดนี้จะมีค่าสูญหายที่เกิดขึ้นจริงอยู่ก่อน 8 เรคคอร์ด และการทดลองในส่วนนี้จะเพิ่มข้อมูลสูญหายให้กับชุดข้อมูลเป็นขนาดต่าง ๆ ตามที่กำหนดไว้ การเพิ่มข้อมูลสูญหายให้กับชุดข้อมูลจะใช้การสุ่มเรคคอร์ดและสุ่มคอลัมน์แล้วแทนค่าข้อมูลที่สูญหายเข้าไป การแบ่งข้อมูลในการทดลองจะสุ่มข้อมูลสำหรับการฝึกสอน (Train data) 70% และชุดข้อมูลสำหรับการทดสอบ (Test data) 30%

```
@DATA
2,2,0,3,0,0,0,0,1,0,0,0,0,0,0,3,2,0,0,0,0,0,0,0,0,0,0,3,0,0,0,1,0,55,2
3,3,3,2,1,0,0,0,1,1,1,0,0,1,0,1,2,0,2,2,2,2,2,1,0,0,0,0,0,0,0,1,0,8,1
2,1,2,3,1,3,0,3,0,0,0,1,0,0,0,1,2,0,2,0,0,0,0,0,2,0,2,3,2,0,0,2,3,26,3
2,2,2,0,0,0,0,0,3,2,0,0,0,3,0,0,2,0,3,2,2,2,2,0,0,3,0,0,0,0,0,3,0,40,1
2,3,2,2,2,2,0,2,0,0,0,1,0,0,0,1,2,0,0,0,0,0,0,0,2,2,3,2,3,0,0,2,3,45,3
2,3,2,0,0,0,0,0,0,0,0,0,2,1,0,2,2,0,2,0,0,0,1,0,0,0,0,0,2,0,0,0,1,0,41,2
2,1,0,2,0,0,0,0,0,0,0,0,0,0,3,1,3,0,0,0,2,0,0,0,0,0,0,0,0,0,0,2,0,18,5
2,2,3,3,3,3,0,2,0,0,0,2,0,0,0,2,3,0,0,0,0,0,0,0,0,2,2,3,2,0,0,3,3,57,3
2,2,1,0,2,0,0,0,0,0,0,0,0,0,0,2,1,0,1,0,0,0,0,0,0,0,0,0,2,0,0,0,2,0,22,4
2,2,1,0,1,0,0,0,0,0,0,0,0,0,0,3,2,0,2,0,0,0,0,0,0,0,0,0,2,0,0,0,2,0,30,4
3,3,2,1,1,0,0,0,2,2,1,0,0,0,0,0,3,2,3,2,2,2,1,1,0,0,0,0,0,0,0,1,0,20,1
2,2,0,3,0,0,0,0,0,0,0,0,0,0,2,0,2,2,0,0,0,0,0,1,0,0,0,0,0,3,0,0,0,1,0,21,2
3,3,1,2,0,0,0,0,0,1,0,0,0,2,0,3,1,0,1,0,0,0,0,0,0,0,0,0,2,0,0,0,1,0,22,2
2,3,3,0,0,0,0,0,1,1,1,0,0,1,0,0,2,1,2,1,2,3,0,2,0,0,0,0,0,0,0,0,2,0,10,1
2,2,3,3,0,3,0,2,0,0,0,2,0,0,0,1,1,1,1,0,0,0,0,0,2,0,3,0,3,0,0,1,3,65,3
1,1,0,1,3,0,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,2,0,0,0,2,0,40,4
2,2,1,3,0,0,0,0,0,0,0,0,0,2,0,2,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,30,2
3,3,3,0,0,0,0,0,3,3,1,0,0,2,0,0,2,0,2,3,3,3,2,3,0,3,0,0,0,0,0,2,0,38,1
2,1,3,3,3,3,0,0,2,0,0,3,0,0,0,3,2,0,1,0,0,0,0,0,3,0,2,0,3,0,0,2,3,23,3
1,1,0,3,0,0,0,0,0,0,0,0,0,0,3,0,3,2,2,0,3,0,0,0,0,0,0,1,0,0,0,2,0,17,5
2,1,1,2,0,0,3,0,1,2,0,0,0,1,0,0,1,2,2,0,1,0,1,0,0,0,0,0,0,0,1,2,1,0,8,6
3,2,2,0,0,0,0,0,0,0,0,0,0,2,0,2,2,1,2,0,2,1,2,0,0,0,0,3,0,0,0,2,0,51,2
2,2,0,2,0,0,0,0,0,0,0,0,0,0,1,1,3,1,2,0,2,1,0,0,0,0,0,0,1,0,1,0,2,0,42,5
2,2,2,3,2,2,0,2,0,0,0,3,2,0,0,0,2,1,1,0,0,0,0,0,3,0,3,0,2,0,0,2,3,44,3
2,0,0,3,0,0,0,0,0,0,0,0,0,0,2,2,2,0,0,0,3,0,0,0,0,0,0,0,0,0,0,2,0,22,5
2,1,1,0,1,0,0,0,2,0,0,0,0,0,0,0,2,2,2,2,2,2,1,2,0,2,0,0,0,0,0,2,0,33,1
```

รูปที่ 4.2 ตัวอย่างชุดข้อมูลโรคผิวหนังที่ใช้ทดสอบประสิทธิภาพเมื่อเพิ่มปริมาณข้อมูลสูญหาย

4.2 การออกแบบการทดสอบประสิทธิภาพ

การออกแบบการทดสอบประสิทธิภาพโมเดลของแต่ละเทคนิคในการเติมค่าให้กับข้อมูลที่สูญหายดังรูปที่ 4.3 โดยเทคนิคที่ใช้ในการทดสอบประสิทธิภาพจะมีทั้งหมด 4 เทคนิคดังนี้คือ

เทคนิคที่ 1 ใช้ค่าเฉลี่ย (Mean) หรือค่ากลาง (Median) ผสมกับค่าปรากฏบ่อย (Mode)

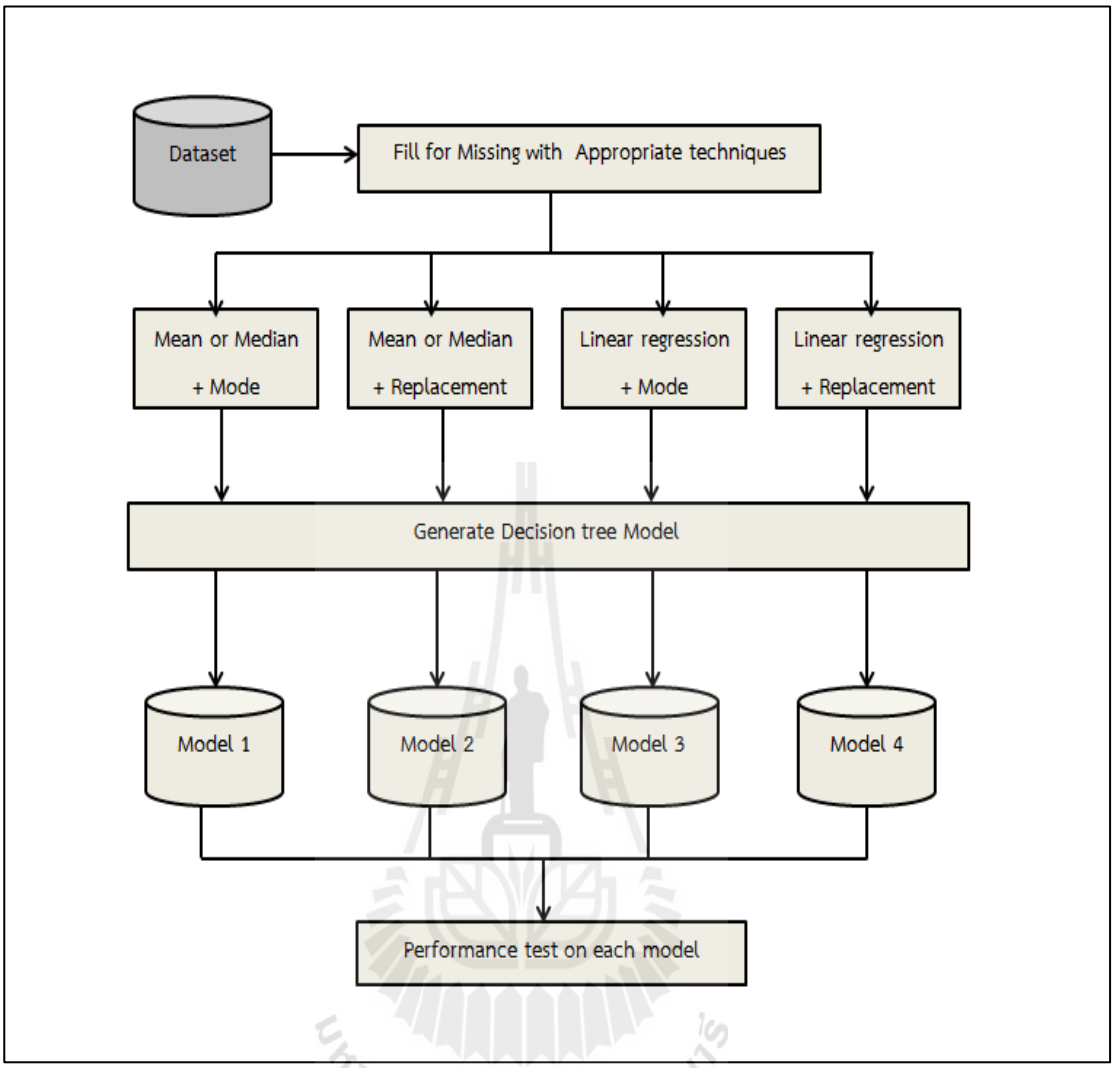
เทคนิคที่ 2 ใช้ค่าเฉลี่ย (Mean) หรือค่ากลาง (Median) ผสมกับการแทนค่า (Replacement)

เทคนิคที่ 3 ใช้สมการถดถอยเชิงเส้น (Linear regression) ผสมกับค่าปรากฏบ่อย (Mode)

เทคนิคที่ 4 ใช้สมการถดถอยเชิงเส้น (Linear regression) ผสมกับการแทนค่า (Replacement)

โดยการออกแบบจะมีกระบวนการขั้นตอนดังนี้

1. นำชุดข้อมูลที่ต้องการใช้ทดสอบประสิทธิภาพเข้ามาในระบบ
2. เลือกเทคนิคแบบผสมผสาน (Hybrid Technique) ที่ต้องการจะทดสอบประสิทธิภาพ
3. เมื่อเติมค่าให้กับข้อมูลที่สูญหายด้วยแต่ละเทคนิคแล้ว จะนำชุดข้อมูลที่ผ่านกระบวนการเติมข้อมูลที่สูญหายไปสร้างโมเดลต้นไม้ตัดสินใจ (Decision Tree)
4. จากขั้นตอนที่ 3 จะได้โมเดลของเทคนิคแบบผสมผสานทั้งหมด 4 โมเดล
5. นำชุดข้อมูลที่แบ่งไว้สำหรับการทดสอบมาทดสอบประสิทธิภาพของโมเดล
6. การทดสอบประสิทธิภาพของแต่ละโมเดล โดยจะนำค่าความถูกต้องในการทำนายของแต่ละโมเดลมาเปรียบเทียบเพื่อประเมินประสิทธิภาพ
7. สรุปผลการทดสอบประสิทธิภาพในการเติมค่าข้อมูลที่สูญหายของแต่ละโมเดล



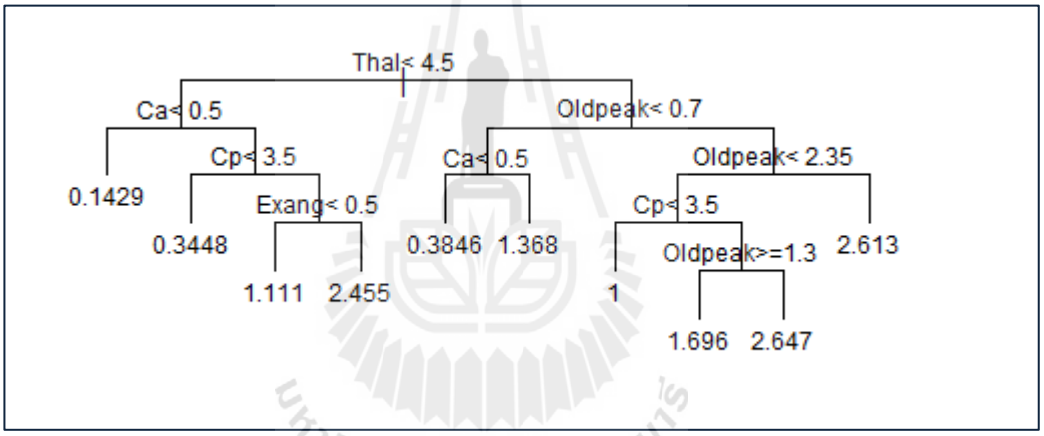
รูปที่ 4.3 การออกแบบการทดสอบประสิทธิภาพโมเดลของแต่ละเทคนิค

4.3 ผลการทดสอบประสิทธิภาพ

4.3.1 การทดสอบประสิทธิภาพโดยใช้ชุดข้อมูลจริงที่มีข้อมูลสูญหาย

ผลที่ได้จากการทดสอบประสิทธิภาพของโมเดลที่สร้างจากข้อมูลที่มีการเติมค่าให้กับข้อมูลที่สูญหาย การใช้ค่าความถูกต้องมาเปรียบเทียบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคการผสมผสาน โดยการทดสอบประสิทธิภาพการทำนายของโมเดลที่ผสมผสานกำกับแทนค่าให้ข้อมูลที่สูญหาย มีค่ากำกับไว้ว่านี้คือข้อมูลที่บกพร่องคือ เทคนิคที่ 2, 4 ไม่สามารถหาค่าความถูกต้องในการทำนายได้ เพราะระบบเข้าใจว่าการแทนค่าด้วยวิธีนี้คือค่าข้อมูลที่สูญหายจึงไม่สามารถนำวิธีนี้มาเปรียบเทียบประสิทธิภาพได้ และผลการทดสอบประสิทธิภาพการเติมค่าของ

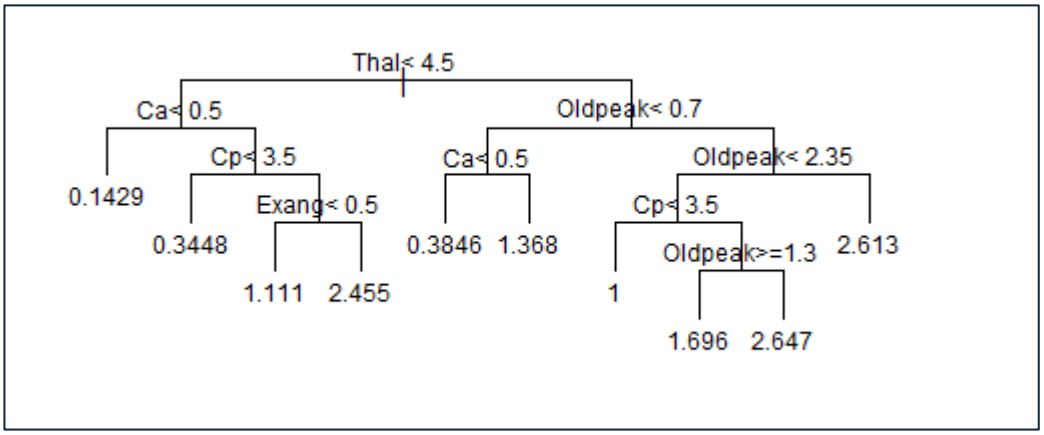
โมเดลที่สร้างจากเทคนิคที่ 1 กับ 3 แล้วนำชุดข้อมูลที่เติมค่าข้อมูลสูญหายด้วยเทคนิคที่ 1 กับ 3 มาสร้างเป็นโมเดลต้นไม้ตัดสินใจได้ดังรูปที่ 4.4 - 4.11 สามารถเปรียบเทียบประสิทธิภาพโดยการทำนายในแอททริบิวต์เป้าหมาย (0, 1, 2, 3, 4) ได้ดังตารางที่ 4.2, 4.4, 4.6, 4.8 เป็นการทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 1 และตารางที่ 4.3, 4.5, 4.7, 4.9 เป็นการทดสอบประสิทธิภาพที่ของโมเดลที่สร้างจากเทคนิคแบบ 3 ซึ่งในชุดข้อมูลโรคหัวใจ (Heart Disease) จะแบ่งข้อมูลเป็นชุดข้อมูลย่อยอีก 4 ชุดข้อมูลคือ Cleveland, Hungary, Switzerland และ Va โดยการทดสอบความถูกต้องจะสามารถคำนวณได้จากการรวมค่าในแนวทแยงมุมของตารางแล้วหารด้วยค่าที่ทำนายได้ทั้งหมดแล้วคูณด้วย 100 จะได้ค่าของความถูกต้อง ซึ่งสามารถสรุปดังตารางที่ 4.10 และแสดงกราฟค่าความถูกต้องของโมเดลที่สร้างจากเทคนิคแบบที่ 1 และเทคนิคแบบที่ 3 ในรูปที่ 4.12



รูปที่ 4.4 การสร้างต้นไม้ตัดสินใจที่ได้จากโมเดลที่ 1 ของชุดข้อมูล Cleveland

ตารางที่ 4.2 การทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 1 ของชุดข้อมูล Cleveland มีค่าความถูกต้อง 62%

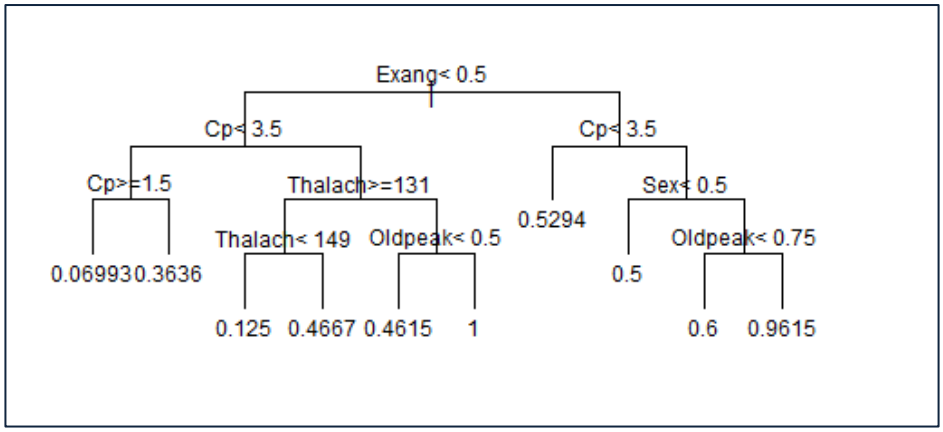
Cleveland	0	1	2	3	4
0	44	7	0	0	0
1	2	5	2	4	0
2	0	4	1	2	1
3	0	1	7	6	5
4	0	0	0	0	0



รูปที่ 4.5 การสร้างต้นไม้ตัดสินใจที่ได้จากโมเดลที่ 3 ของชุดข้อมูล Cleveland

ตารางที่ 4.3 การทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 3 ของชุดข้อมูล Cleveland มีค่าความถูกต้อง 62%

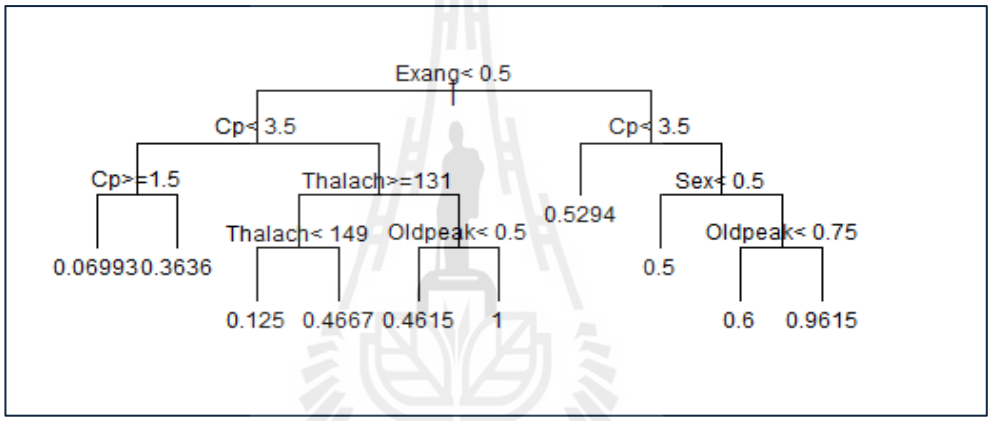
Cleveland	0	1	2	3	4
0	3	2	0	0	0
1	5	6	3	1	0
2	3	4	8	6	0
3	0	3	5	6	3
4	0	0	0	0	0



รูปที่ 4.6 การสร้างต้นไม้ตัดสินใจที่ได้จากโมเดลที่ 1 ของชุดข้อมูล Hungary

ตารางที่ 4.4 การทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 1 ของชุดข้อมูล Hungary มีค่าความถูกต้อง 78.40%

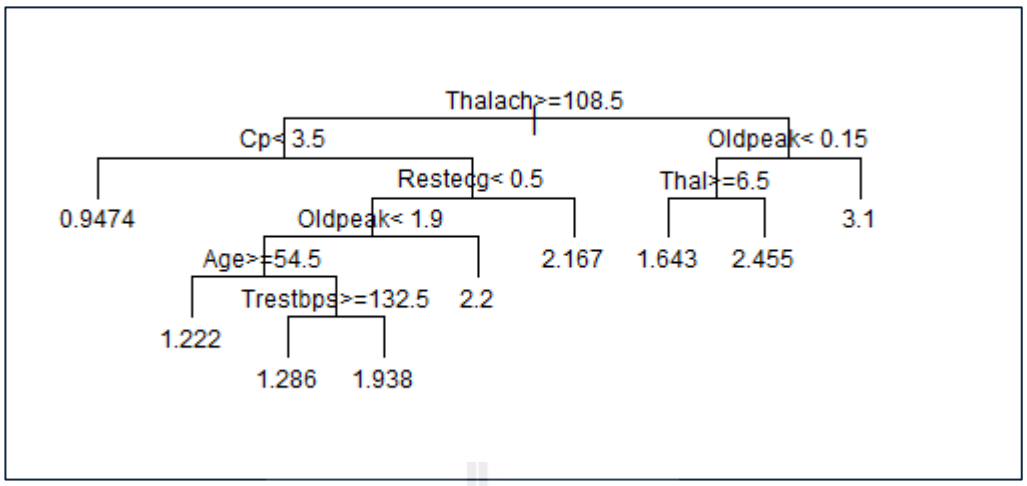
Hungary	0	1	2	3	4
0	27	12	0	0	0
1	2	42	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0



รูปที่ 4.7 การสร้างต้นไม้ตัดสินใจที่ได้จากโมเดลที่ 3 ของชุดข้อมูล Hungary

ตารางที่ 4.5 การทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 3 ของชุดข้อมูล Hungary มีค่าความถูกต้อง 78.40%

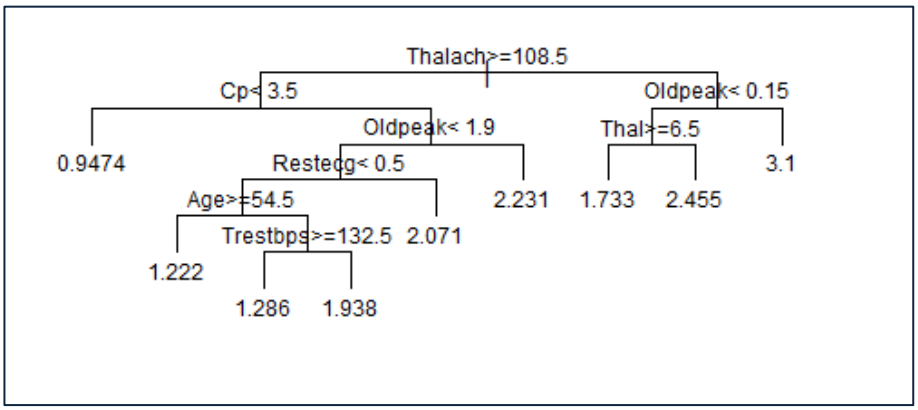
Hungary	0	1	2	3	4
0	27	12	0	0	0
1	2	42	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0



รูปที่ 4.8 การสร้างต้นไม้ตัดสินใจที่ได้จากโมเดลที่ 1 ของชุดข้อมูล Switzerland

ตารางที่ 4.6 การทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 1 ของชุดข้อมูล Switzerland มีค่าความถูกต้อง 51.35%

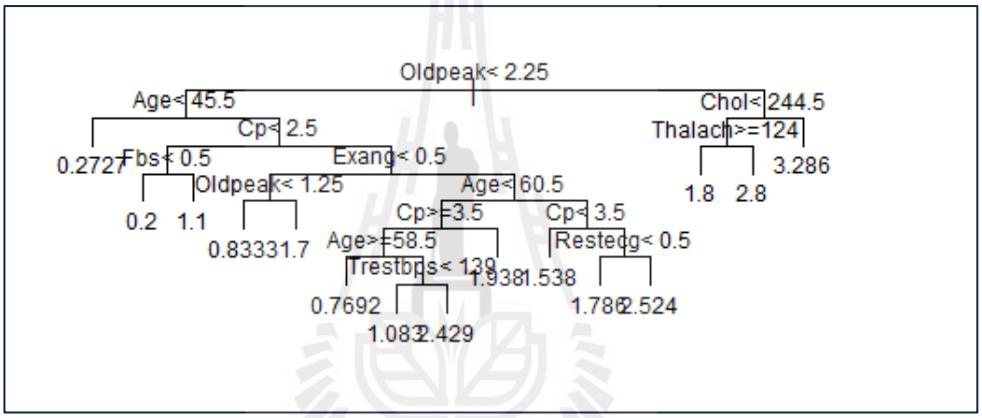
Switzerland	0	1	2	3	4
0	0	0	0	0	0
1	3	9	1	1	0
2	1	3	9	5	2
3	0	0	1	1	1
4	0	0	0	0	0



รูปที่ 4.9 การสร้างต้นไม้ตัดสินใจที่ได้จากโมเดลที่ 3 ของชุดข้อมูล Switzerland

ตารางที่ 4.7 การทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 3 ของชุดข้อมูล Switzerland มีค่าความถูกต้อง 43.24%

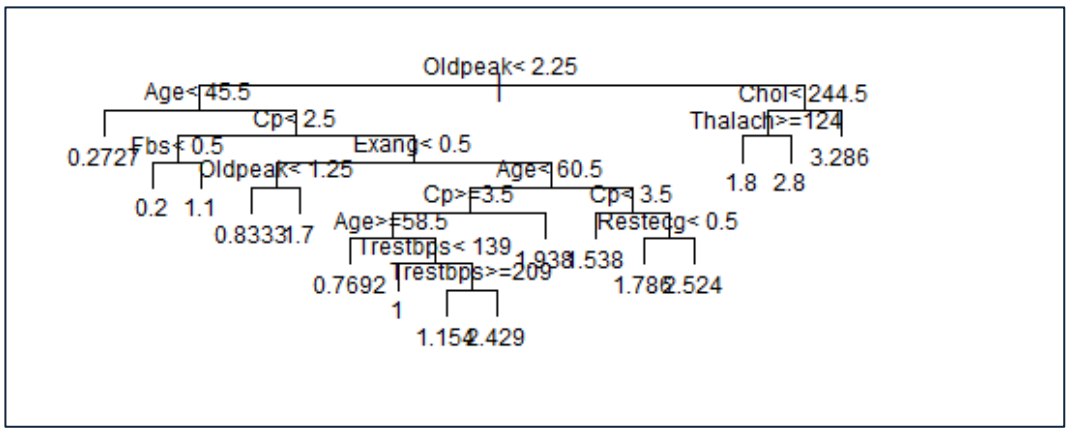
Switzerland	0	1	2	3	4
0	0	0	0	0	0
1	3	4	1	1	0
2	1	3	9	3	2
3	0	0	1	3	1
4	0	0	0	0	0



รูปที่ 4.10 การสร้างต้นไม้ตัดสินใจที่ได้จากโมเดลที่ 1 ของชุดข้อมูล Va

ตารางที่ 4.8 การทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 1 ของชุดข้อมูล Va มีค่าความถูกต้อง 40.00%

Va	0	1	2	3	4
0	3	2	0	0	0
1	5	6	3	1	0
2	3	5	9	8	0
3	0	3	5	6	3
4	0	0	0	0	0



รูปที่ 4.11 การสร้างต้นไม้ตัดสินใจที่ได้จากโมเดลที่ 3 ของชุดข้อมูล Va

ตารางที่ 4.9 การทดสอบประสิทธิภาพของโมเดลที่สร้างจากเทคนิคแบบ 3 ของชุดข้อมูล Va มีความถูกต้อง 38.33%

Va	0	1	2	3	4
0	3	2	0	0	0
1	5	6	3	1	0
2	3	4	8	6	0
3	0	3	5	6	3
4	0	0	0	0	0

ตารางที่ 4.10 ค่าความถูกต้องของโมเดลที่สร้างจากชุดข้อมูลโรคหัวใจ

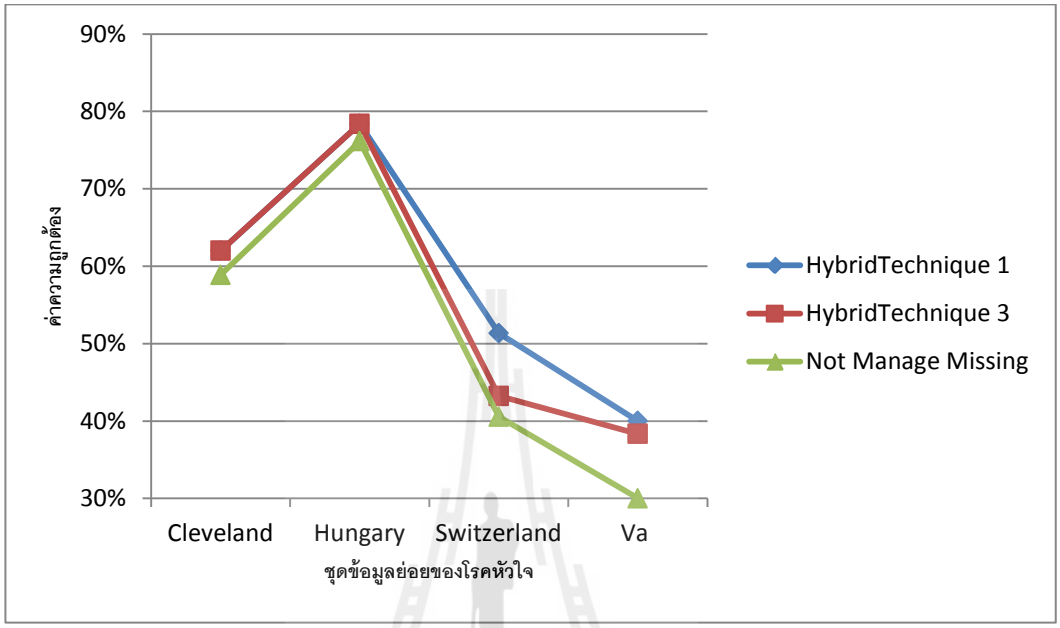
ชุดข้อมูล	ค่าความถูกต้อง				
	Technique 1	Technique 2	Technique 3	Technique 4	ไม่มีการเติม
Cleveland	62.22%	NA	62.22%	NA	58.88%
Hungary	78.40%	NA	78.40%	NA	76.13%
Switzerland	51.35%	NA	51.35%	NA	40.54%
Va	40.00%	NA	38.33%	NA	30.00%

หมายเหตุ Technique 1 คือ Hybrid Technique (Mean or Median + Mode)

Technique 2 คือ Hybrid Technique (Mean or Median + Replacement)

Technique 3 คือ Hybrid Technique (Linear Regression + Mode)

Technique 4 คือ Hybrid Technique (Linear Regression + Replacement)



รูปที่ 4.12 กราฟค่าความถูกต้องของโมเดลที่สร้างจากเทคนิคแบบที่ 1 และเทคนิคแบบที่ 3

4.3.2 การทดสอบประสิทธิภาพโดยเพิ่มข้อมูลสูญหายให้กับชุดข้อมูล

การทดสอบประสิทธิภาพจากการเพิ่มข้อมูลสูญหายด้วยการสุ่ม จะแบ่งการเพิ่มข้อมูลสูญหายออกเป็น 10%, 20%, 30%, 40% และ 50% แล้วเปรียบเทียบโดยใช้เทคนิคแบบผสมผสานของโมเดลที่ 1 และ 3 เปรียบเทียบกับการลบเรคคอร์ดที่มีข้อมูลสูญหายออกด้วย ซึ่งข้อมูลที่ใช้คือชุดข้อมูลโรคผิวหนังที่มี 366 เรคคอร์ด แบ่งเป็นจำนวนข้อมูลที่ใช้ฝึกสอน 70 % หรือ 256 เรคคอร์ดและข้อมูลที่ใช้ทดสอบ 30% หรือ 110 เรคคอร์ด ผลการทดลองแสดงในตารางที่ 4.11

ตารางที่ 4.11 ค่าความถูกต้องของโมเดลที่สร้างจากชุดข้อมูลโรคผิวหนัง

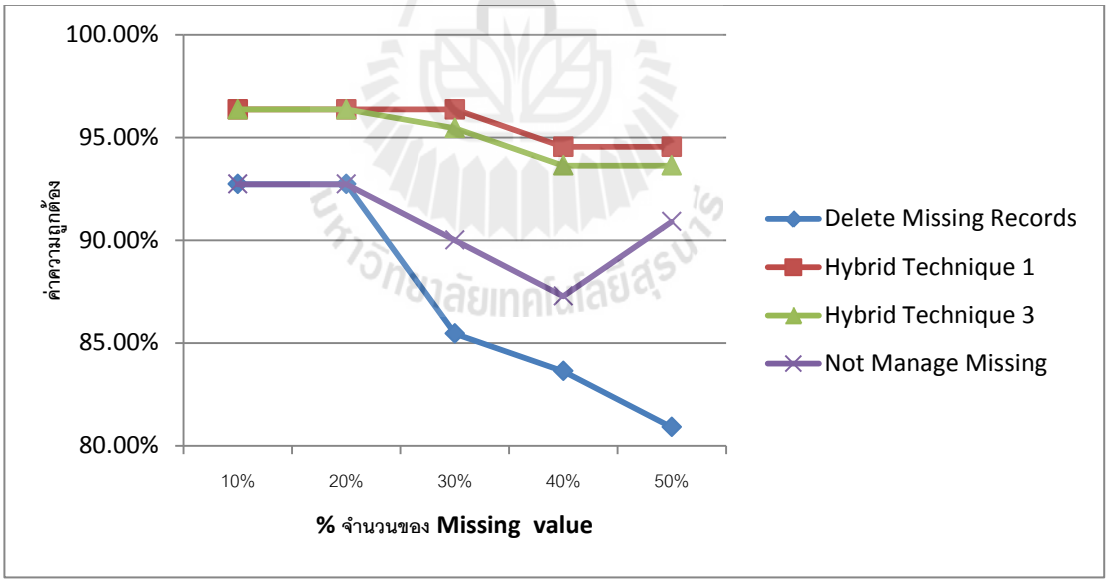
จำนวนการเพิ่ม Missing value	ค่าความถูกต้อง					
	ไม่มีการเติม	ลบเรคคอร์ดข้อมูลสูญหาย	Hybrid Technique			
			1	2	3	4
10%	92.72%	92.72%	96.36%	NA	96.36%	NA
20%	92.72%	92.72%	96.36%	NA	96.36%	NA
30%	90.00%	85.45%	96.36%	NA	95.45%	NA
40%	87.27%	83.63%	94.54%	NA	93.63%	NA
50%	90.90%	80.9%	94.54%	NA	93.63%	NA

หมายเหตุ Hybrid Technique 1 คือ Mean or Median + Mode

Hybrid Technique 2 คือ Mean or Median + Replacement

Hybrid Technique 3 คือ Linear Regression + Mode

Hybrid Technique 4 คือ Linear Regression + Replacement



รูปที่ 4.13 กราฟเปรียบเทียบค่าความถูกต้องของการทดสอบเพิ่มค่าข้อมูลสูญหาย

4.4 การอภิปรายผลการทดสอบประสิทธิภาพ

จากการทดสอบประสิทธิภาพของเทคนิคในการเติมค่าให้กับข้อมูลที่สูญหายแบบผสมผสาน ซึ่งในการทดสอบประสิทธิภาพได้แบ่งเทคนิคการเติมค่าออกเป็น 4 วิธีการ แต่ในการวัดประสิทธิภาพวิธีการที่ 2 และ 4 ไม่สามารถวัดประสิทธิภาพได้เพราะเทคนิคการกำกับค่าให้ข้อมูลที่สูญหายระบบมองเห็นข้อมูลนั้นเป็นข้อมูลผิดพลาด ซึ่งในการทดสอบประสิทธิภาพการเติมค่าที่สูญหายให้กับชุดข้อมูลโรคหัวใจทั้ง 4 ชุดข้อมูลย่อย ผลการทดสอบปรากฏว่าชุดข้อมูลโรคหัวใจ Cleveland กับ Hungary การเติมค่าให้ข้อมูลสูญหายด้วยการใช้วิธีการที่ 1 คือการใช้ค่าเฉลี่ย (Mean) หรือค่ากลาง (Median) ผสมกับค่าปรากฏบ่อย (Mode) และวิธีการที่ 3 คือการใช้สมการถดถอยเชิงเส้น (Linear regression) ผสมกับค่าปรากฏบ่อย (Mode) จะได้ค่าความถูกต้องที่เท่ากัน แต่การทดสอบประสิทธิภาพกับชุดข้อมูลโรคหัวใจ Switzerland กับ Va จะได้ค่าความถูกต้องของวิธีการที่ 1 มากกว่าวิธีการที่ 3 เพราะเทคนิคการเติมค่าของข้อมูลที่สูญหายของวิธีการที่ 1 มีการตรวจสอบการกระจายที่ผิดปกติทำให้มีความทนทานต่อค่าที่ผิดปกติ (Outlier) แต่เทคนิคการเติมค่าของข้อมูลที่สูญหายของวิธีการที่ 3 ไม่มีการตรวจสอบข้อมูลที่มีการกระจายผิดปกติทำให้ไม่ทนทานต่อค่าที่ผิดปกติ (Outlier) ส่วนผลการทดลองที่ได้สุ่มเพิ่มจำนวนข้อมูลสูญหายให้กับชุดข้อมูลโรคผิวหนังแบ่งเป็นช่วง ๆ จะสังเกตได้ว่าถ้ามีจำนวนข้อมูลสูญหายจะมีผลกระทบต่อค่าความถูกต้อง ซึ่งถ้าจำนวนของข้อมูลที่สูญหายในช่วง 10-20% การใช้เทคนิคแบบผสมผสานที่ 1 กับ 3 จะได้ค่าความถูกต้องที่เท่ากัน แต่ถ้ามีจำนวนข้อมูลที่สูญหายมีค่า 40%-50% ค่าความถูกต้องจะน้อยลง และทำให้เทคนิคแบบผสมผสานที่ 1 ดีกว่าที่ 3 เพราะถ้ามีข้อมูลที่สูญหายเพิ่มขึ้นอาจมีผลกระทบต่อข้อมูลที่เป็นชนิดเชิงตัวเลขทำให้ข้อมูลในคอลัมน์นั้นเกิดการกระจายตัวที่ผิดปกติได้ ซึ่งเทคนิคแบบผสมผสานแบบที่ 1 จะทนทานต่อการกระจายตัวผิดปกติส่วนผลการทดลองที่เปรียบเทียบการใช้เทคนิคแบบผสมผสานมาเติมค่าข้อมูลสูญหายกับไม่ใช้เทคนิคการเติมค่าและการจัดการตัดเรคคอร์ดที่มีข้อมูลที่สูญหายทิ้ง ผลที่ได้จะสังเกตได้ว่าการใช้เทคนิคแบบผสมผสานจะให้ค่าความถูกต้องมากกว่าประมาณ 3.64% - 13.64%

สรุปได้ว่าถ้าข้อมูลมีการกระจายตัวปกติวิธีการเติมข้อมูลที่สูญหายทั้ง 2 วิธีการจะให้ค่าความถูกต้องที่เท่ากัน แต่ถ้ามีการกระจายตัวผิดปกติวิธีการที่ 1 จะได้ค่าความถูกต้องมากกว่าวิธีการที่ 3 และถ้ามีกระจายที่ผิดปกติการใช้เทคนิคการเติมค่าด้วยสมการถดถอยเชิงเส้นตรง (Linear regression) ไม่เหมาะสมในการนำมาใช้เติมค่าให้ข้อมูลที่สูญหาย ถ้ามีจำนวนของข้อมูลที่สูญหายใกล้เคียงกันค่าความถูกต้องที่ได้จะมีค่าเท่ากัน แต่ถ้ามีจำนวนข้อมูลที่สูญหายมีค่ามากจะทำให้ค่าความถูกต้องลดน้อยลง และผลการทดสอบปรากฏว่าการใช้เทคนิคแบบผสมผสานเข้ามาช่วยเติมค่าให้กับข้อมูลสูญหายจะช่วยเพิ่มประสิทธิภาพของโมเดลได้มากขึ้น

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

การทำเหมืองข้อมูลโดยนำชุดข้อมูลจริงจากการเก็บสำรวจข้อมูลมาวิเคราะห์เพื่อสร้างโมเดล จะต้องใช้ความระมัดระวังในการเก็บรวบรวมข้อมูล ซึ่งการเก็บข้อมูลอาจมีข้อมูลที่ผิดพลาดเกิดขึ้นเนื่องจากข้อมูลจริงอาจจะมีความผิดปกติเองจริงหรือเกิดข้อผิดพลาดของผู้เก็บข้อมูล ทำให้การนำชุดข้อมูลที่มีข้อมูลสูญหายมาสร้างโมเดลในการทำนาย จะทำให้โมเดลที่ได้มีความผิดพลาดตามไปด้วยหรือมีค่าความถูกต้องในการทำนายมีค่าน้อย และการนำโมเดลที่สร้างขึ้นจากชุดข้อมูลที่ไม่สมบูรณ์ไปใช้งานจริง ถ้าในด้านการแพทย์อาจจะทำให้ผู้ป่วยไม่ได้รับการรักษาที่ถูกต้องจนถึงขั้นเสียชีวิตได้ หรือในด้านธุรกิจอาจจะขาดทุนจนถึงขั้นล้มละลายได้ ดังนั้นชุดข้อมูลที่มีค่าข้อมูลสูญหายควรได้รับการปรับปรุงแก้ไขให้ชุดข้อมูลมีความสมบรูณ์ เพื่อให้มีประสิทธิภาพในการนำไปใช้งานและนำไปใช้การสร้างโมเดลสำหรับการทำนายให้มีความถูกต้องมากยิ่งขึ้น

งานวิจัยนี้จึงมุ่งเน้นการพัฒนาและออกแบบเทคนิคในการเติมค่าให้กับชุดข้อมูลที่มีค่าข้อมูลสูญหายเกิดขึ้น ซึ่งการออกแบบระบบการเติมค่าให้กับข้อมูลสูญหายสามารถใช้งานได้ทั้งสองระบบ คือระบบที่สามารถเติมค่าข้อมูลสูญหายแบบอัตโนมัติโดยจะเลือกเทคนิคที่ดีที่สุดจากการวิจัยครั้งนี้ และระบบที่ผู้ใช้สามารถกำหนดเทคนิคที่ต้องการเติมค่าได้ ส่วนการเปรียบเทียบประสิทธิภาพของเทคนิคในการเติมค่าข้อมูลสูญหายทำได้ด้วยการวัดค่าของความถูกต้องในการทำนายของโมเดล และขั้นตอนในการวิจัยมีดังนี้

1. การศึกษางานวิจัยที่เกี่ยวข้องกับการจัดการและการเติมค่าให้กับข้อมูลสูญหาย (Missing value) เพื่อศึกษาหาเทคนิคการเติมค่าที่มีประสิทธิภาพสามารถแก้ไขปัญหาที่สนใจในงานวิจัยได้ และศึกษาการเขียนโปรแกรมด้วยภาษาอาร์ (R Language) ซึ่งเป็นภาษาเชิงฟังก์ชันที่นิยมใช้ในด้านสถิติและทางคณิตศาสตร์

2. การออกแบบเทคนิคการเติมค่าให้กับข้อมูลสูญหายที่จะสามารถใช้เติมค่าให้กับข้อมูลได้ทั้งข้อมูลเชิงตัวเลข (Numeric) และข้อมูลเชิงลักษณะ (Categorical) การออกแบบเทคนิคการเติมค่าจะมีทั้งเทคนิคที่มีผู้ฝึกสอน (Supervised) และเทคนิคที่ไม่มีผู้ฝึกสอน (Unsupervised)

3. การพัฒนาโปรแกรมจะใช้ภาษาอาร์ในการพัฒนาโปรแกรมทั้งสองระบบคือ ระบบอัตโนมัติที่สามารถช่วยเติมค่าข้อมูลสูญหายด้วยการเลือกเทคนิคที่เหมาะสมให้อัตโนมัติ และระบบที่ผู้ใช้สามารถเลือกเทคนิคการเติมค่าข้อมูลสูญหายตามต้องการได้

4. การทดสอบประสิทธิภาพการทำงานของโปรแกรม โดยใช้ค่าความถูกต้องในการเปรียบเทียบประสิทธิภาพของเทคนิคการเติมข้อมูลสูญหาย ซึ่งการทดสอบโปรแกรมจะแบ่งชุดข้อมูลออกเป็นสองชุดคือมีข้อมูลในการฝึกสอนและชุดข้อมูลที่ใช้ในการทดสอบประสิทธิภาพการทำนายของโมเดล

5.1 สรุปผลการวิจัย

จากการเปรียบเทียบประสิทธิภาพของเทคนิคการเติมค่าให้กับข้อมูลสูญหายโดยใช้ค่าความถูกต้องในการทำนายของโมเดล ซึ่งผลการทดลองที่เปรียบเทียบเทคนิคการเติมค่าด้วยค่าเฉลี่ยในกรณีข้อมูลมีการกระจายตัวปกติ หรือถ้ามีข้อมูลมีการกระจายแบบเอียงจะใช้เทคนิคการหาค่ากลางเติมค่าให้กับข้อมูลที่สูญหายประเภทข้อมูลเชิงตัวเลข จะดีกว่าการใช้เทคนิคการเติมค่าข้อมูลสูญหายด้วยเทคนิคการใช้สมการถดถอยเชิงเส้น เพราะสมการถดถอยเชิงเส้นไม่ทนทานต่อข้อมูลที่มีการกระจายแบบเอียง และเทคนิคของการกำกับค่าให้กับข้อมูลสูญหายโปรแกรมไม่สามารถวัดประสิทธิภาพของเทคนิคนี้ได้ เพราะโปรแกรมจะเห็นข้อมูลผิดพลาดไปจากชนิดเดิม ซึ่งการทดสอบประสิทธิภาพของข้อมูลที่ผ่านการใช้เทคนิคแบบผสมผสานเติมค่าให้ข้อมูลสูญหายแล้วจะดีกว่าการใช้ข้อมูลที่มีค่าสูญหายอยู่ จากผลการทดลองจึงกำหนดเทคนิคในการเลือกแบบระบบอัตโนมัติให้ใช้เทคนิคดีที่สุดในการเติมมูลสูญหายคือ ถ้าเป็นข้อมูลเชิงปริมาณจะเลือกเทคนิคระหว่างใช้ค่าเฉลี่ยหรือค่ากลางด้วยพิจารณาจากการเกิดข้อมูลการกระจายแบบปกติหรือการกระจายแบบเอียง และถ้าข้อมูลเชิงคุณภาพจะใช้ค่าปรากฏบ่อยที่สุดมาเติมค่าให้กับข้อมูลสูญหาย

5.2 ปัญหาและข้อเสนอแนะ

จากผลการทดสอบเปรียบเทียบเทคนิคการเติมค่าให้กับข้อมูลสูญหายจะพบว่าปัญหาที่เกิดขึ้นคือโปรแกรมไม่สามารถวัดประสิทธิภาพของเทคนิคการกำกับค่าได้ เพราะโปรแกรมมองข้อมูลที่ถูกกำกับว่าเป็นข้อมูลสูญหายอยู่ ซึ่งถ้าผู้ใช้ต้องการเลือกเทคนิคการเติมค่าด้วยวิธีนี้ก็ยังสามารถทำได้ด้วยการเลือกระบบส่วนที่ให้ผู้ใช้กำหนดเอง ข้อเสนอแนะสำหรับการปรับปรุงต่อยอดจากงานวิจัยนี้อาจจะปรับปรุงให้มีการใช้เทคนิคในการเติมค่าข้อมูลที่มากขึ้นหรือใช้เทคนิคการเติมค่าให้กับข้อมูลสูญหายร่วมกับเทคนิคอื่นเพื่อเพิ่มประสิทธิภาพให้กับโมเดลเพื่อการจำแนกให้มีค่าความถูกต้องเพิ่มมากขึ้น เช่นการใช้เทคนิคร่วมกับการแบ่งช่วงให้กับข้อมูลเชิงปริมาณ หรือการคัดเลือกเฉพาะคุณสมบัติที่เหมาะสมที่จะนำมาเป็นตัวช่วยในการเติมค่าให้กับข้อมูลสูญหายเป็นต้น

รายการอ้างอิง

กิตติศักดิ์ เกิดประสพ.(2012). **Data Mining Methodology and Development**. Retrieved November 20, 2012, from https://sites.google.com/site/kittisakthailand55/home/datamining2-55

นิตยา เกิดประสพ.(2555ก) .การทำเหมืองข้อมูล. **เอกสารประกอบการสอนวิชาการค้นหาความรู้และการขุดค้นข้อมูล (Knowledge Discovery and Data Mining)**.:3-25

นิตยา เกิดประสพ.(2555ข).**Data Mining Algorithms**. Retrieved January 15, 2013, from http://www.csitvru.com /wiwat/mining/chapter7.ppt

สายชล สินสมบูรณ์ทอง.(2555).**สถิติเบื้องต้น**.พิมพ์ครั้งที่ 10.กรุงเทพฯ:จามจุรีโปรดักท์.:17-39

ศิริชัย พงษ์วิชัย.(2555).**สถิติเพื่อการวิจัยด้วยโปรแกรม R**.พิมพ์ครั้งที่ 2.กรุงเทพฯ:สุพีเรียพริ้นติ้งเฮาส์.:363-414

วิกิพีเดีย สารานุกรมเสรี.(2555ก).**การทำเหมืองข้อมูล**. ที่มา http://th.wikipedia.org/wiki/การทำเหมืองข้อมูล

วิกิพีเดีย สารานุกรมเสรี.(2555ข).**กฎความสัมพันธ์**. ที่มา http://th.wikipedia.org/wiki/กฎความสัมพันธ์

วิกิพีเดีย สารานุกรมเสรี.(2555ค).Accuracy and precision.ที่มา http://en.wikipedia.org/wiki/Accuracy_and_precisi0on

วิกิพีเดีย สารานุกรมเสรี.(2555ง). **Regression analysis**. ที่มา http://en.wikipedia.org/wiki/Regression_analysis

Jianhua Dai, Qing Xu, Wentao Wang. (2011). **A Comparative Study on Strategies of Rule Induction for Incomplete Data Based on Rough Set Approach**, International Journal of Advancements in Computing Technology vol 3 no. 3 .:176-183

Shmuel Friedland, Amir Niknejad, Mostafa Kaveh, Hossein Zare.(2005).**An Algorithm for Missing Value Estimation for DNA Microarray Data**. Retrieved January 23, 2012 from http://arxiv.org/pdf/q-bio.GN/0510047.pdf

Pedro J.Garcıa, Laencina.Jose, Luis.Sancho Gomez.(2010).**Pattern classification with missing data**.Neural Comput & Applic 19th .:263–282

Loris Nannia, Alessandra Luminib, Sheryl Brahnamc.(2012).**A classifier ensemble approach for the missing feature problem**. Artificial Intelligence in Medicine 55.:37–50

Fulufhelo V. Nelwamondo and Tshilidzi Marwala. (2007). **Rough Sets Computations to Impute Missing Data**, CoRR abs/0704.3635.

Karlien Vanden Branden, Sabine Verboven.(2009).**Robust data imputation**.Computational Biology and Chemistry 33.:7-13

George Ssali and Tshilidzi Marwala. (2007). **Estimation of Missing Data Using Computational Intelligence and Decision Trees**, ArXiv e-prints, Volume 709.

UC Irvine Machine Learning Repository, (1998). **Dermatology Data Set**. Retrieved February 15, 2012 from http://archive.ics.uci.edu/ml/datasets/Dermatology

UC Irvine Machine Learning Repository, (1988). **Heart Disease Data Set**. Retrieved January 20, 2012 from http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Joseph DePasquale, Robi Polikar.(2007).**Random Feature Subset Selection for Ensemble Based Classification of Data with Missing Features**.Lecture Notes in Computer Science Volume 4472.:251-260

Jerzy W., Grzymala-Busse.(2003).**Rough Set Strategies to Data with Missing Attribute Values**.IEEE International Conference on Data Mining 2003.:56-63

Jerzy W., Grzymala-Busse.(2004).**Three Approaches to Missing Attribute Values A Rough Set Perspective**.IEEE International Conference on Data Mining 2004.:139-152

Jianhua Wu, Qinbao Song, Junyi Shen.(2007).**An Novel Association Rule Mining Based Missing Nominal Data Imputation Method**.Eighth ACIS International Conference on Software Engineering.:244-249

Xiao-Yong Pan, Ye Tian, Yan Huang, Hong-Bin Shen.(2011).**Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach**.Genomics 97.:257-264

Shichao Zhang, Senior Member, IEEE, Zhenxing Qin, Charles X. Ling, and Shengli Sheng.(2005)."Missing Is Useful": Missing Values in Cost-Sensitive Decision Trees.IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, v.17 N.12,:1689-1693

Heping Zhang.(2011).**Recursive Partitioning and Applications**. Retrieved February 1, 2013, from http://epub.ub.uni-muenchen.de/10589/1/partitioning.pdf

ภาคผนวก ก

รหัสต้นฉบับของโปรแกรมเทคนิคเติมค่าให้กับข้อมูลสูญหาย

โปรแกรมเทคนิคการเติมค่าข้อมูลสูญหายแบบผสมผสานระบบอัตโนมัติ

```
hybrid.Technique <- function(dataset) { #ฟังก์ชันหลักในการทำงานแบบอัตโนมัติ
cc=1
for(i  in seq(ncol(dataset))){
        cn<-check.NA(dataset,cc)  #เรียกฟังก์ชันเพื่อตรวจสอบจำนวนข้อมูลที่สูญหาย
ของคอลัมน์เกิน40% หรือไม่
        if(cn==T){
                dataset=dataset[,c(-cc)]
                cc<-cc-1
        }  cc<-cc+1
}
        colM=1
        for(i in seq(ncol(dataset))){
                dataset<-hybrid.Col(dataset,colM) #เรียกฟังก์ชันเติมค่าข้อมูลสูญหาย
                ในคอลัมน์
        colM<-colM+1
        }  return(dataset)
}
hybrid.Col<- function(dataset,colM) {
        if(is.numeric(dataset[[colM]])){ #ตรวจสอบข้อมูลว่าเป็น Numerical ใช่หรือไม่
                dComplet<-imputation.mm(dataset,colM) #เรียกฟังก์ชันสำหรับเติมค่าข้อมูล
ประเภท Numerical
        }else{
                dComplet<-imputation.mode(dataset,colM) #เรียกฟังก์ชันสำหรับเติมค่า
ข้อมูลประเภท Categorical
        }
return(dComplet)
}
```

```
imputation.mm<- function(dataM,colM){ # ฟังก์ชันสำหรับเติมค่าข้อมูล ประเภท Numerical
        more=boxplot.stats(dataM[[colM]])$out  #ตรวจสอบการกระจายตัวของข้อมูล
        if(length(more)==0){
        dataM[is.na(dataM[[colM]]),colM]<-round(mean(dataM[[colM]],na.rm=T))
                #ใช้การเติมค่าเฉลี่ยให้กับข้อมูลที่สูญหาย
        }else{
        dataM[is.na(dataM[[colM]]),colM]<-round(median(dataM[[colM]],na.rm=T))
                #ใช้การเติมค่ากลางให้กับข้อมูลที่สูญหาย
        } return(dataM)
}
val.mode<- function (x) {
        if (is.factor(x)) levels(x) [which.max(table(x))]
        else {  f <- as.factor(x)
        levels(f) [ which.max(table(f)) ]  }
}
#ฟังก์ชันสำหรับเติมค่าข้อมูลประเภทเชิงลักษณะ (Categorical)
imputation.mode<- function(dataM,colM) {
        v.mode<-val.mode(dataM[[colM]])
        dataM[is.na(dataM[[colM]]),colM]<-v.mode  #เรียกฟังก์ชันค้นหาค่าปรากฏบ่อยซ้ำ
return(dataM)
}
check.NA<- function(dataset,colM) {  #ฟังก์ชันตรวจสอบจำนวนข้อมูลสูญหายเกินกำหนด
num=1
ck=0
        for(i in seq(nrow(dataset))){
                if(is.na(dataset[[colM]][num])){  ck<-ck+1 }
        num<-num+1  }  dataC=F
per  <-round(nrow(dataset)*0.4)  #กำหนดข้อมูลสูญหายไม่เกิน 40%
        if(ck>per){  dataC=T   }
return(dataC)  }
```

โปรแกรมเทคนิคการเติมค่าข้อมูลสูญหายแบบผสมผสานระบบผู้ใช้กำหนดเอง

```
hybrid.User<- function(dataset,tech,colM,colN) { #ฟังก์ชันหลักแบบผู้ใช้เลือกเทคนิคได้
if(is.numeric(dataset[[colM]])){ #ตรวจสอบข้อมูลว่าเป็น Numerical ใช่หรือไม่
if(tech==3){ #ถ้าผู้ใช้ต้องการเติมค่าข้อมูลที่สูญหายด้วยเทคนิคค่ากลาง
                        dComplet<-imputation.mu(dataset,colM,T)}
if(tech==4){ #ถ้าผู้ใช้ต้องการเติมค่าข้อมูลที่สูญหายด้วยเทคนิคค่าเฉลี่ย
                        dComplet<-imputation.mu(dataset,colM,F)}
if(tech==5){ #ถ้าผู้ใช้ต้องการเติมค่าข้อมูลที่สูญหายด้วยเทคนิคสมการถดถอยเชิงเส้น
                        dComplet<-line.input(colM,colN, dataset)  }
}else{
if(tech==1){ #ถ้าผู้ใช้ต้องการเติมค่าข้อมูลที่สูญหายด้วยเทคนิคค่าที่ปรากฏซ้ำ
                        dComplet<-imputation.mode(dataset,colM)}
if(tech==2){ #ถ้าผู้ใช้ต้องการเติมค่าข้อมูลที่สูญหายด้วยเทคนิคค่าการกำกับค่า
                        dComplet<-imputation.mis(dataset,colM)}
return(dComplet)}

val.mode<- function (x) { #ฟังก์ชันสำหรับหาค่า Mode
        if (is.factor(x)) levels(x) [which.max(table(x))]
        else { f <- as.factor(x)
        levels(f) [ which.max(table(f)) ]  }  }
imputation.mode<- function(dataM,colM) {
v.mode<-val.mode(dataM[[colM]])
dataM[is.na(dataM[[colM]]),colM]<-v.mode
return(dataM)
}
```

```
imputation.mis<- function(dataM,colM) { #ฟังก์ชันสำหรับการกำกับค่า  Missing
levels(dataM[[colM]])<-c(levels(dataM[[colM]]),"missing")
var="missing"
dataM[is.na(dataM[[colM]]),colM]<-var
return(dataM)  }
imputation.mu<- function(dataM,colM,more=T){ #ฟังก์ชันสำหรับหาค่า Mean และ Median
        if(more){
        dataM[is.na(dataM[[colM]]),colM]<-round(median(dataM[[colM]],na.rm=T))
        }else{
        dataM[is.na(dataM[[colM]]),colM]<-round(mean(dataM[[colM]],na.rm=T)) }
return(dataM)  }
lookCor<- function(crn){ #ฟังก์ชันสำหรับหาค่า Regression
        gg<-(cor(crn,use='complete.obs') )
        gp<-symnum(gg)
        return(gp)  }
creXY<- function(colM,dataM,NN){
        mM<-lm(colM,data=dataM)$coefficients[NN]
        mN<-mM[1][[1]]
        return(mN)  }
inputf<- function(oP){
        if ( is.na(oP) ) return(NA)
        else return ( (oP*(mX))+mY ) }
line.input<- function(colA,colB,dataM){
        dataM[ is.na ( dataM[[colA]] ),colA ] <-
        round(sapply ( dataM[ is.na (dataM[[colA]]),colB],inputf))
        return(dataM)  }
```

ภาคผนวก ข

บทความวิชาการที่ได้รับการตีพิมพ์เผยแพร่

รายชื่อบทความวิจัยที่ได้รับการตีพิมพ์เผยแพร่

พัชราวรรณ ชินไชสง, กิตติศักดิ์ เกิดประสพ และนิตยา เกิดประสพ. 2555**. การเปรียบเทียบเทคนิคการทำนายให้กับข้อมูลสูญหายด้วยภาษาอาร์ (Comparison of techniques to impute missing data by R language).** การประชุมวิชาการระดับชาติเพื่อการพัฒนาด้านวิจัยอย่างยั่งยืน, มหาวิทยาลัยศรีนครินทรวิโรฒ, ประเทศไทย. 25 – 26 ธันวาคม 2555.

Nuntawut Kaoungka, Phatcharawan Chinthaisong, Kittisak Kerdprasop, and Nittaya Kerdprasop. 2013 **Discretization and Imputation Techniques for Quantitative Data Mining.** Proceedings of the 2013 IAENG International Conference on Data Mining and Application, Hong Kong, 13-15 march 2013. 318-321.

การเปรียบเทียบเทคนิคการทำนายค่าให้กับข้อมูลที่สูญหายด้วยภาษาอาร์

Comparison of techniques to impute missing data by R language.

พัชราวรรณ ชินไธสง*, กิตติศักดิ์ เกิดประสพ, นิดยา เกิดประสพ

Phatcharawan Chinthaisong*, Kittisak Kerdprasop, Nittaya Kerdprasop

สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

Department of Computer Engineering, Faculty of Engineering, Suranaree University of Technology, Thailand.

*Corresponding author, E-mail: b5102730@gmail.com

บทคัดย่อ

งานวิจัยนี้เป็นการนำเสนอการเปรียบเทียบเทคนิคในการทำนายค่าให้กับค่าของข้อมูลที่สูญหาย โดยการใช้ชุดข้อมูลที่มีข้อมูลที่สูญหายมาทำเหมืองข้อมูลจะทำให้โมเดลที่ได้อาจมีประสิทธิภาพการทำนายต่ำหรือการทำนายผลที่ผิดพลาด เพื่อต้องการเพิ่มเพิ่มประสิทธิภาพให้กับโมเดลจึงใช้เทคนิคต่างๆเพื่อเติมค่าให้กับข้อมูลที่สูญหาย ซึ่งเทคนิคที่เลือกใช้มีวิธีการหาค่าเฉลี่ยถ้าข้อมูลที่สูญหายเป็นแบบการกระจายปกติหรือถ้าเป็นแบบเอียงจะใช้วิธีหาค่ากึ่งกลาง ถ้าข้อมูลของคอลัมน์นั้นไม่ใช่ตัวเลขจะเลือกค่าที่ปรากฏบ่อยที่สุดในคอลัมน์นั้น และวิธีการใช้ค่าจากตารางความสัมพันธ์ระหว่างคอลัมน์ที่มีค่าข้อมูลที่สูญหายกับคอลัมน์อื่นที่มีความสัมพันธ์กันมากที่สุด โดยนำเทคนิคที่ใช้มาเขียนเป็นโปรแกรมเชิงฟังก์ชันด้วยภาษาอาร์ในการทำนายและการทดลองผล

คำสำคัญ : การทำนายค่าข้อมูลที่สูญหาย, การทำเหมืองข้อมูล, ภาษาอาร์

Abstract

This paper presents a comparison of techniques to impute and predict values that are missing in the data set. Missing value is an important problem because mining from data with missing values can incur an inefficient model that might predict future data with a high error rate. To improve the model efficiency, we study several techniques to handle missing value cases. These techniques are to impute with the mean value if data in that column distribute normally, impute with the median if data are skew, impute with the mode if data are nominal, and impute with correlated value. We will implement these imputation techniques with R language.

Keyword : Missing value imputation, Data mining, R language

บทนำ

ปัจจุบันการทำเหมืองข้อมูลมีส่วนสำคัญอย่างมากในการนำมาใช้ช่วยทำนายผลทางด้านเศรษฐศาสตร์ ด้านการศึกษา และด้านการแพทย์ โดยการใช้ชุดข้อมูลที่ได้จากการเก็บสำรวจมาสร้างเป็นโมเดลที่ช่วยในการทำนาย ถ้าชุดข้อมูลนั้นมีค่าข้อมูลที่สูญหายส่วนมากจะทำการตัดแถวของข้อมูลที่มีค่าข้อมูลที่สูญหายก่อนการสร้างโมเดลในการทำนาย โดยแถวข้อมูลที่ถูกตัดอาจมีข้อมูลที่สำคัญเป็นส่วนที่มีผลช่วยในการทำนาย เพื่อจะช่วยในการเพิ่มประสิทธิภาพให้กับโมเดลในการทำนายจึงต้องการเติมค่าให้กับข้อมูลที่สูญหาย ซึ่งเทคนิคการเติมค่าให้กับข้อมูลที่สูญหายมีหลากหลายวิธีเพื่อให้ได้ค่าข้อมูลที่สูญหายไปมีความถูกต้องมากที่สุด งานวิจัยนี้ได้ศึกษาเปรียบเทียบเทคนิคต่าง ๆ ที่ใช้ในการเติมค่าข้อมูลที่สูญหาย และพัฒนาเป็นโปรแกรมด้วยภาษาอาร์ โดยการทดสอบประสิทธิภาพการทำนายใช้วิธีการโมเดลในลักษณะต้นไม้ตัดสินใจ

ภาษาอาร์เป็นภาษาเชิงฟังก์ชันที่มีการนิยมใช้กันด้านสถิติ ภาษาอาร์สามารถสร้างเวกเตอร์และเมตริกได้และยังมีคำสั่งการใช้งานที่ง่ายสะดวกเหมาะสมสำหรับการนำมาเขียนโปรแกรมในการทำเหมืองข้อมูล ภาษาอาร์เป็นแบบโอเพนซอร์ส สามารถใช้งานได้ด้วยไม่ต้องเสียค่าใช้จ่าย

ต้นไม้ตัดสินใจเป็นการสร้างโมเดลสำหรับการทำนาย โดยการใช้แผนภาพต้นไม้เพื่อทำนายชุดข้อมูลนั้นโดยมีการกำหนดเป้าหมายที่ต้องการ โดยการทำนายจะเริ่มจากรากของต้นไม้แล้วแตกกิ่งก้านออกเป็นใบซึ่งเป็นค่าการทำนาย โดยมีงานวิจัยของ [4] มีการใช้เทคนิคสร้างใช้ต้นไม้ตัดสินใจมาทำนายค่าของข้อมูลที่สูญหายซึ่งใช้เทคนิคต้นไม้ตัดสินใจร่วมกับวิธีการเครือข่ายประสาทเทียม (Associative neural network) และมีงานวิจัยที่ใช้เทคนิคด้านอื่นมาช่วยในการทำนายเช่น และยังมีงายวิจัยอื่นที่ใช้การใช้ทฤษฎี Rough Set ได้แก่ [3] โดยจะใช้วิธีการหาความสัมพันธ์ระหว่างแต่ละคอลัมน์มาสร้างเป็นเซตเพื่อนำมาเป็นกฎที่ช่วยในการทำนาย ชุดข้อมูลที่เขาใช้ในการวิจัยเป็นชุดข้อมูลของผู้ป่วยโรคเอดส์ซึ่งข้อมูลส่วนมากจะเป็นข้อมูลของตัวเลขที่กระจัดกระจายกันอยู่ ด้วยส่วนที่ข้อมูลเป็นตัวเลขจะทำการจัดกลุ่มเป็นช่วงข้อมูล (Discretized) เพื่อง่ายต่อการทำการวิจัยในการหาค่าของข้อมูลที่สูญหาย และงานวิจัยของ [2] ได้นำเสนอเทคนิคการเติมค่าให้กับข้อมูลที่สูญหายโดยการใช้ทฤษฎีราฟเซต (Rough Sets) และได้เพิ่มเทคนิคเพื่อนำมาเปรียบเทียบ 3 วิธี คือวิธีการตัดแถวข้อมูลที่มีค่าข้อมูลที่สูญหายออกแล้วจึงทำเหมืองข้อมูล วิธีการเลือกค่าที่จะนำมาเติมให้กับข้อมูลที่สูญหายจากข้อมูลที่มีค่าที่ปรากฏบ่อยที่สุดในคอลัมน์นั้น และวิธีการแปลงข้อมูลทั้งชุดข้อมูลให้เป็นข้อมูลแบบเมตริกดิสเซอร์นิไบลิตี้ (Discernibility matrix) แล้วนำมาสร้างเป็นกฎเพื่อนำมาทำนายค่าที่สูญหายไป โดยได้ใช้ชุดข้อมูลหกชุดข้อมูลเพื่อนำมาเปรียบเทียบประสิทธิภาพวิธีการในการหาค่าของข้อมูลที่สูญหายทั้งสามวิธี

จากงานวิจัยที่กล่าวมาข้างต้น จะมีเทคนิคการทำนายค่าของข้อมูลที่สูญหายที่หลากหลายและมีวิธีการเปรียบเทียบประสิทธิภาพที่แตกต่างกัน โดยบางงานวิจัยจะเน้นไปทางชุดข้อมูลที่มีความน่าสนใจ และการแสดงให้เห็นผลงานวิจัยได้ชัดเจนทั้งการแสดงตาราง แสดงกราฟ และการสรุปผลได้อย่างชัดเจน ซึ่งงานวิจัยนี้ก็ได้ศึกษาเทคนิคการเติมค่าให้กับข้อมูลที่สูญหาย และวิธีการเปรียบเทียบประสิทธิภาพของเทคนิคการทำนายแต่ละเทคนิค เพื่อหาเทคนิคที่สามารถทำนายข้อมูลที่สูญหายได้ดีที่สุด ในการวิจัยนี้ได้เพิ่มเติมขั้นตอนของวิเคราะห์ลักษณะการกระจายของข้อมูลของข้อมูลและชนิดของข้อมูล เพื่อให้ในหนึ่งชุดข้อมูลสามารถใช้เทคนิคการเติมค่าข้อมูลสูญหายที่แตกต่างกันได้

วัตถุประสงค์ของการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบเทคนิคในการทำนายค่าของข้อมูลที่สูญหายที่มีค่าใกล้เคียงกับค่าที่เป็นไปได้มากที่สุด โดยการเปรียบเทียบเทคนิคการตัดค่าของข้อมูลที่สูญหายออกก่อน การใช้ค่าเฉลี่ยในคอลัมน์นั้น ถ้าข้อมูลที่สูญหายเป็นแบบการกระจายหรือถ้าเป็นแบบเอียงจะใช้ค่ากึ่งกลาง ถ้าข้อมูลในคอลัมน์นั้นไม่ใช่ตัวเลขจะใช้ค่าที่ปรากฏบ่อยที่สุดในคอลัมน์ และวิธีการใช้ค่าของความสัมพันธ์ระหว่างคอลัมน์ที่มีค่าข้อมูลที่สูญหายกับคอลัมน์อื่นที่มีความสัมพันธ์มากที่สุด แล้วนำมาสร้างเป็นต้นไม้ตัดสินใจเพื่อทดสอบกับข้อมูลที่เตรียมไว้สำหรับการทดสอบ ว่าเทคนิคใดให้ประสิทธิภาพในการทำนายดีที่สุด

วิธีดำเนินการวิจัย

ในงานวิจัยนี้จะมีวิธีการดำเนินงานวิจัย แบ่งออกเป็นขั้นตอนดังนี้

ขั้นตอนที่ 1 การศึกษาวิธีการใช้งานฟังก์ชันที่เกี่ยวข้องกับงานวิจัยในภาษาอาร์

ภาษาอาร์คือภาษาใช้เขียนโปรแกรมเชิงฟังก์ชันนิยมใช้ทางด้านสถิติเช่นการวิเคราะห์ข้อมูล การสร้างกราฟ สามารถทำงานกับข้อมูลได้ทั้งแบบเวกเตอร์และเมตริก สามารถนำมาเขียนโปรแกรมในการทำเหมืองข้อมูลได้สะดวก เพราะมีฟังก์ชันที่สามารถเรียกใช้งานได้ง่าย โดยสามารถเขียนโปรแกรมเรียกใช้คำสั่งผ่านหน้าจอของเทอร์มินอลของหน้าโปรแกรม

ฟังก์ชันสำคัญที่มีการเรียกใช้ในงานวิจัยนี้

ฟังก์ชัน na.omit() เป็นฟังก์ชันสำหรับใช้ตัดแถวที่มีค่าของข้อมูลที่สูญหายออกจากชุดข้อมูลชุดนั้น โดยการใส่ชื่อชุดข้อมูลลงไปในวงเล็บ แล้วฟังก์ชันจะตัดแถวที่มีค่าข้อมูลที่สูญหายออกให้

ฟังก์ชัน mean() เป็นฟังก์ชันสำหรับการหาค่าเฉลี่ยในคอลัมน์ที่ต้องการ โดยการใส่ชื่อของชุดข้อมูลและคอลัมน์ที่ต้องการลงไป และกำหนดให้ค่าข้อมูลที่สูญหายไม่ต้องนำมาคิดเฉลี่ยด้วยการใส่ na.rm=T และนำค่าไปเติมในค่าของข้อมูลที่ที่สูญหาย

ฟังก์ชัน median() เป็นฟังก์ชันสำหรับการหาค่ากึ่งกลางในคอลัมน์ที่ต้องการ โดยการใส่ชื่อของชุดข้อมูลและคอลัมน์ที่ต้องการลงไป และกำหนดให้ค่าข้อมูลที่สูญหายไม่ต้องนำมาคิดเฉลี่ยด้วยการใส่ na.rm=T และนำค่าไปเติมในค่าของข้อมูลที่ที่สูญหาย

ฟังก์ชัน cor() เป็นฟังก์ชันสำหรับหาค่าความสัมพันธ์ระหว่างคอลัมน์ที่มีค่าของข้อมูลที่สูญหายที่สัมพันธ์กับคอลัมน์อื่นมากที่สุด โดยจะแสดงผลออกมาเป็นตารางเมตริกความสัมพันธ์ให้เห็นว่าคอลัมน์ใดมีความสัมพันธ์กันมากที่สุดแล้วเลือกค่าของสองคอลัมน์นั้นนำมาแทนในสมการ แล้วแทนค่าให้กับข้อมูลที่สูญหาย

ฟังก์ชัน predict() เป็นฟังก์ชันสำหรับการใช้ทำนาย โดยจะรับพารามิเตอร์ที่เป็นโมเดลและพารามิเตอร์ที่สองจะเป็นชุดข้อมูลที่ต้องการนำมาใช้ในการทำนาย

ฟังก์ชัน table() เป็นฟังก์ชันที่ใช้แสดงการนับข้อมูล แล้วแสดงออกมาเป็นตารางเมตริก

ขั้นตอนที่ 2 การเตรียมข้อมูลสำหรับการวิจัย

ชุดข้อมูลที่ใช้ในการทำวิจัยคือข้อมูล hepatitis, adult, glass identification, และ furest fires เป็นชุดข้อมูลที่มีสองลักษณะคือมีทั้งแบบมีค่าข้อมูลที่สูญหายเกิดขึ้นจริง และชุดข้อมูลที่ไม่มีค่าของข้อมูลที่สูญหายเพื่อนำมากำหนดค่าข้อมูลที่สูญหายออกเป็นช่วงๆให้เห็นได้ชัดเจน และการแบ่งชุดข้อมูลส่วนสำหรับการฝึกสอนเป็น 70% และข้อมูลสำหรับการทดสอบ 30% การเรียกชุดข้อมูลมาใช้งานจะใช้คำสั่ง read.csv เพื่ออ่านชุดข้อมูลที่ต้องการ col.names การกำหนดชื่อให้กับคอลัมน์ na.strings การกำหนดว่าค่าข้อมูลที่สูญหายในชุดข้อมูลนั้นแทนด้วยสัญลักษณ์อะไร

```
hepatitis<- read.csv("hepatitis.txt", header=F, dec=".",
col.names=c("Class","AGE","SEX","STEROID","ANTIVIRALS","FATIGUE","MALAISE","ANOREXIA",
"LIVER_BIG","LIVERFIRM","SPLEENPALPABLE","SPIDERS","ASCITES","VARICES",
"BILIRUBIN","ALKPHOSPHATE","SGOT","ALBUMIN","PROTIME","HISTOLOGY"),
na.strings=c("?") )
```

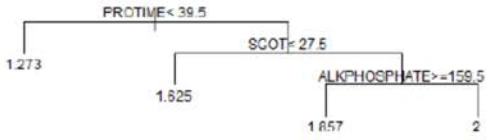
ขั้นตอนที่ 3 การทำนายค่าข้อมูลที่สูญหาย

โดยเทคนิคที่ใช้จัดการกับข้อมูลที่สูญหายจะใช้วิธีการตัดข้อมูลที่สูญหายประกอบด้วยหลายวิธีได้แก่ออกก่อนสร้างโมเดลการทำนาย วิธีการใช้ค่าเฉลี่ยและค่ากึ่งกลางเติมค่าของข้อมูลที่สูญหายหรือถ้าข้อมูลในคอลัมน์นั้นไม่ใช่ตัวเลขจะใช้ค่าที่ปรากฏซ้ำบ่อยที่สุด และวิธีการใช้ค่าความสัมพันธ์ของคอลัมน์ที่มีค่าข้อมูลที่สูญหายกับคอลัมน์อื่นที่มีค่าความสัมพันธ์มากที่สุด ถ้าข้อมูลในแถวใดมีค่าของข้อมูลที่สูญหายมากเกินไปจนไม่สามารถใช้ทำนายข้อมูลที่สูญหายได้จะใช้คำสั่ง hepatitis<-hepatitis[-c(10,26,101),] เพราะแถวที่ 10 26 101 มีข้อมูลที่สูญหายมากเกินไปไม่เหมาะสมมาใช้ทำนาย

1. เทคนิคการตัดค่าข้อมูลที่สูญหายออกก่อนและสร้างโมเดลเพื่อการทำนายในลักษณะของต้นไม้ตัดสินใจ โดยการใช้ฟังก์ชัน na.omit() เพื่อตัดค่าข้อมูลที่สูญหายออก ตัวอย่าง dataset1<- na.omit(hepatitis) จะได้ชุดข้อมูลที่ถูกตัดแถวที่มีค่าข้อมูลที่สูญหายออก แล้วใช้คำสั่งเรียก library(rpart) ขึ้นมาเพื่อใช้งานในการสร้างต้นไม้ตัดสินใจ แล้วใช้คำสั่งเพื่อสร้างต้นไม้ตัดสินใจ

```
rt.a1<-rpart(Class~.,data=dataset1[,1:20])
plot(rt.a1,uniform=T,branch=1, margin=0.1, cex=0.9)
text(rt.a1,cex=0.75)
```

ผลลัพธ์ที่ได้ตามภาพประกอบที่ 1



ภาพประกอบที่ 1 โมเดลที่ได้จากข้อมูลที่มีการตัดทิ้งเรคคอร์ดที่ข้อมูลสูญหาย

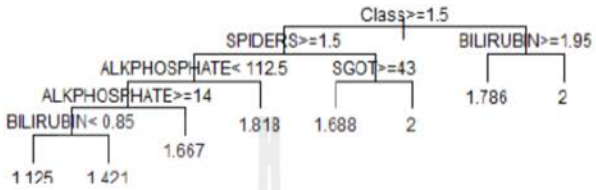
แล้วใช้คำสั่ง lookC(hepatitis) เพื่อดูเมตริกของค่าความสัมพันธ์ แล้วเลือกคอลัมน์ที่ใช้มาคำนวณในสมการโดยใช้คำสั่ง

```
mX<-creXY(MALAISE~FATIGUE,hepatitis,2)
mY<-creXY(MALAISE~FATIGUE,hepatitis,1)
```

และจะเรียกใช้ฟังก์ชันคำนวณสมการและการทำนายเติมค่าให้กับข้อมูลที่สูญหายโดยใช้คำสั่ง

```
cor.input <- cor.input ("LIVER_BIG","MALAISE",hepatitis)
```

แล้วเมื่อใช้คำสั่งในการสร้างโมเดลต้นไม้ในตัดสินใจจะได้ผลลัพธ์ดังภาพประกอบที่ 3



ภาพประกอบที่ 3 โมเดลที่สร้างจากข้อมูลที่ถูกเติมค่าสูญหายด้วยค่าที่สัมพันธ์กันมากที่สุด

ขั้นตอนที่ 4 การทดสอบประสิทธิภาพของการทำนาย

การทดสอบประสิทธิภาพของโมเดลต้นไม้ตัดสินใจในการทำนายจะใช้ชุดข้อมูลที่แบ่งไว้สำหรับการวัดประสิทธิภาพของโมเดลโดยจะใช้ฟังก์ชัน predict() เข้ามาช่วยในการทำนายโมเดลกับชุดข้อมูลสำหรับการทดสอบ และใช้ฟังก์ชันtable()แสดงตารางนับ โดยการพิมพ์คำสั่ง

```
pred1 <- predict(rt.a1, newdata = hepatitis.test)
table(pred1, hepatitis.test$Class)
```

ผลการวิจัย

ผลการวิจัยในแต่ละการทดลองเพื่อเปรียบเทียบประสิทธิภาพของเทคนิคในการทำนายค่าของข้อมูลที่สูญหายจะแบ่งลักษณะของข้อมูลที่สูญหายออกเป็น 2 แบบคือชุดข้อมูลที่เกิดค่าข้อมูลที่สูญหายขึ้นจริง กับแบบที่ชุดข้อมูลถูกจำลองให้มีค่าข้อมูลที่สูญหายเกิดขึ้นโดยจะแบ่งออกเป็นเปอร์เซ็นต์ในการเกิดข้อมูลที่สูญหาย และการวิจัยจะใช้โมเดลต้นไม้ทั้ง 3 เทคนิคเอามาทดสอบกับข้อมูลที่เตรียมไว้สำหรับการทดสอบ โดยโมเดลที่ 1 หมายถึงโมเดลที่สร้างจากข้อมูลฝึกที่ไม่มีการเติมค่าข้อมูลที่สูญหาย แต่ใช้วิธีตัดเรคคอร์ดที่มีข้อมูลที่สูญหายทิ้ง โมเดลที่ 2 หมายถึงโมเดลที่สร้างจากข้อมูลที่ใช้เทคนิคการเติมค่าที่สูญหายด้วยค่าเฉลี่ยและค่ากึ่งกลาง(พิจารณาตามการกระจายของข้อมูลในคอลัมน์) และค่าที่ปรากฏซ้ำบ่อยที่สุดหรือค่าโหมด และโมเดลที่ 3 หมายถึงโมเดลที่สร้างจากข้อมูลฝึกที่ใช้การเติมค่าสูญหายด้วยการหาความสัมพันธ์กับคอลัมน์อื่น ชุดข้อมูลที่นำมาใช้ในงานวิจัยใช้ชุดข้อมูลมาจากเว็บไซต์ของ UCI ซึ่งจะมีชุดข้อมูลจริงที่สามารถนำมาใช้ทดลองในงานวิจัยได้

โดยการวิจัยนี้ได้เลือกชุดข้อมูล 4 ชุดข้อมูลซึ่งชุดข้อมูลแต่ละชุดมีความแตกต่างกันที่ชุดข้อมูลแรก Hepatitis เป็นชุดข้อมูลเกี่ยวกับไวรัสตับอักเสบซึ่งมีค่าของข้อมูลที่สูญหายเกิดขึ้นในลักษณะข้อมูลเป็นตัวเลขซึ่งผลการทดสอบประสิทธิภาพความถูกต้องจะได้ตามตารางที่ 1 และสามารถสร้างเป็นกราฟแสดงผลดังภาพประกอบที่ 4 ชุดข้อมูลที่ 2 คือ Adult เป็นชุดข้อมูลเกี่ยวกับการทำนายรายได้ของประชากร และมีแอททริบิวต์ที่มีค่าของข้อมูลที่สูญหายเกิดขึ้นในลักษณะเป็นข้อความมีผลการทดลองตามตารางที่ 2 และสร้างเป็นกราฟแสดงผลดังภาพประกอบที่ 5 โดยทั้งสองชุด

ข้อมูลแรกจะเป็นชุดข้อมูลที่มีค่าของข้อมูลที่สูญหายเกิดขึ้นจริงและเพื่อต้องการเปรียบกับข้อมูลที่สูญหายเกิดขึ้นอยู่กับจำนวนของการเกิดข้อมูลที่หายไปได้ชัดเจน จึงได้นำชุดข้อมูลที่ไม่มีค่าของข้อมูลที่สูญหายนำมากำหนดให้เกิดข้อมูลที่สูญหายโดยใช้ชุดข้อมูล Glass Identification เป็นชุดข้อมูลเกี่ยวกับประเภทของแก้วซึ่งมีข้อมูลที่สูญหายเกิดขึ้นในลักษณะเป็นตัวเลข โดยผลการทดลองตามตารางที่ 3 และสร้างเป็นกราฟแสดงผลดังภาพประกอบที่ 6 ชุดข้อมูลที่ 4 คือ Forest Fires เป็นชุดข้อมูลที่ทำนายเกี่ยวกับพื้นที่เกิดไฟป่า ซึ่งเป็นชุดข้อมูลที่มีค่าของข้อมูลที่สูญหายเกิดขึ้นลักษณะเป็นข้อความ โดยผลการทดลองตามตารางที่ 1 และสร้างเป็นกราฟแสดงผลดังภาพประกอบที่ 7

ตารางที่ 1 ความถูกต้องของโมเดลที่สร้างจากชุดข้อมูลที่มีค่าของข้อมูลที่สูญหายเกิดขึ้นจริงลักษณะข้อมูลเป็นตัวเลข

ชุดของข้อมูล	โมเดลที่ 1	โมเดลที่ 2	โมเดลที่ 3
Hepatitis	65.95%	74.46%	80.85%

ตารางที่ 2 ความถูกต้องของโมเดลที่สร้างจากชุดข้อมูลที่มีค่าของข้อมูลที่สูญหายเกิดขึ้นจริงลักษณะข้อมูลเป็นข้อความ

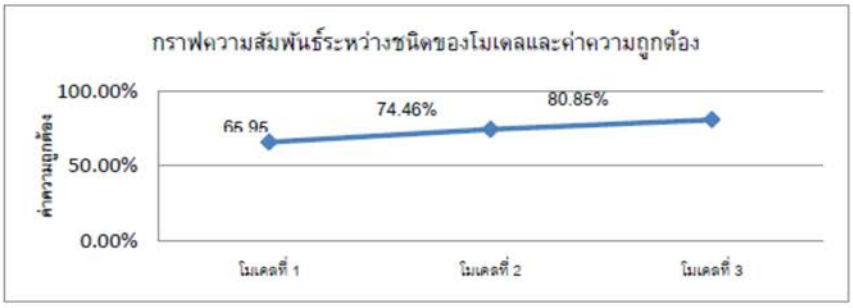
ชุดของข้อมูล	โมเดลที่ 1	โมเดลที่ 2
Adult	59.23%	71.25%

ตารางที่ 3 ความถูกต้องของโมเดลที่สร้างจากชุดข้อมูลที่มีค่าของข้อมูลที่สูญหายเกิดขึ้นเป็นเปอร์เซ็นต์สำหรับข้อมูลที่เป็นตัวเลข

ชุดของข้อมูล **Glass Identification**	โมเดลที่ 1	โมเดลที่ 2	โมเดลที่ 3
Missing value 10%	44.44%	51.23%	62.31%
Missing value 15%	44.44%	54.54%	61.84%
Missing value 20%	40.62%	50.12%	54.22%
Missing value 30%	38.59%	45.86%	52.81%

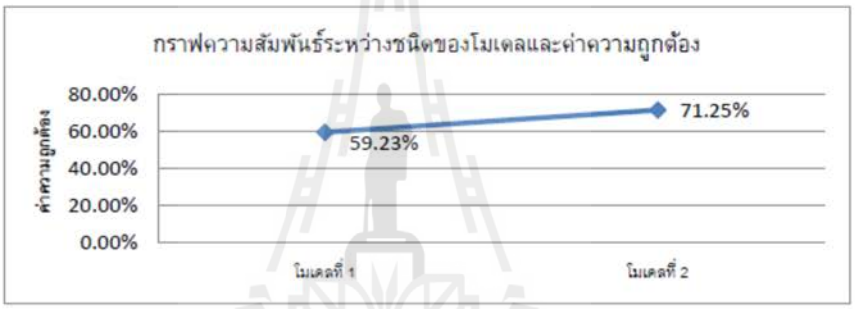
ตารางที่ 4 ความถูกต้องของโมเดลที่สร้างจากชุดข้อมูลที่มีค่าของข้อมูลที่สูญหายเกิดขึ้นเป็นช่วงๆ สำหรับข้อมูลที่เป็นข้อความ

ชุดของข้อมูล **Forest Fires**	โมเดลที่ 1	โมเดลที่ 2
Missing value 10%	52.27%	62.33%
Missing value 15%	50.45%	60.42%
Missing value 20%	48.21%	55.75%
Missing value 30%	45.03%	53.46%



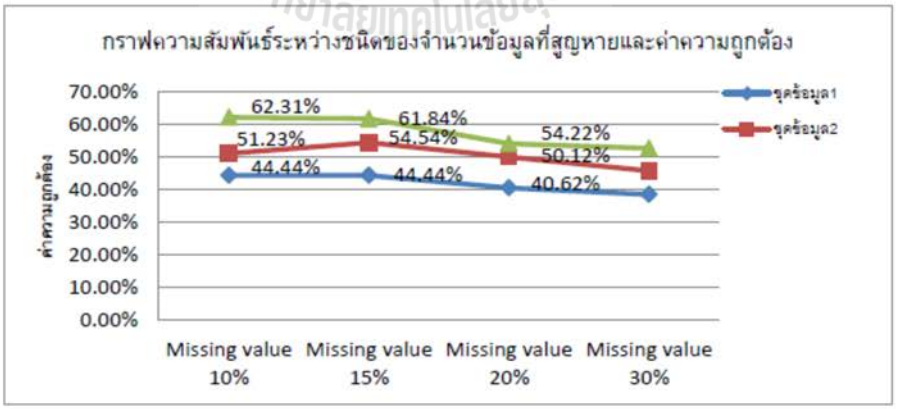
ภาพประกอบที่ 4 กราฟแสดงความสัมพันธ์ระหว่างโมเดลของชุดข้อมูล Hepatitis กับค่าความถูกต้องของโมเดล

จากภาพประกอบที่ 4 การทดลองชุดข้อมูล Hepatitis ซึ่งมีข้อมูลที่สูญหายเกิดขึ้นจริงในลักษณะเป็นตัวเลข ซึ่งความถูกต้องในการทำนายของโมเดลที่ 3 การหาค่าความสัมพันธ์ได้ 80.85% ดีกว่าอีกสองโมเดล



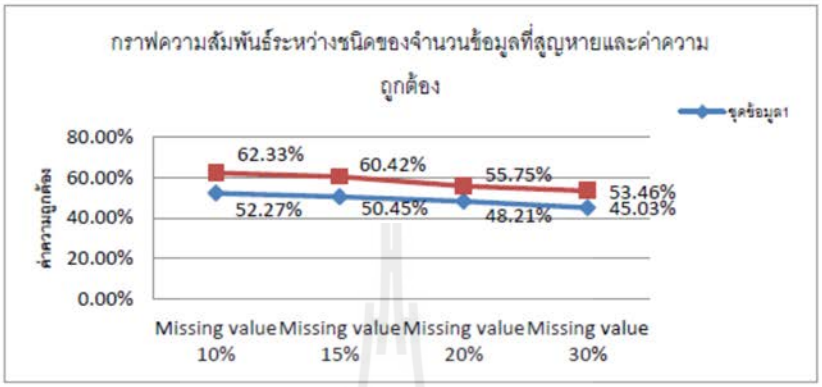
ภาพประกอบที่ 5 กราฟแสดงความสัมพันธ์ระหว่างโมเดลของชุดข้อมูล Adult กับค่าความถูกต้องของโมเดล

จากภาพประกอบที่ 5 การทดลองชุดข้อมูล Adult ซึ่งมีข้อมูลที่สูญหายเกิดขึ้นจริงในลักษณะเป็นข้อความ ซึ่งความถูกต้องในการทำนายของโมเดลที่ 2 จะได้กว่าโมเดลที่ 1



ภาพประกอบที่ 6 กราฟแสดงความสัมพันธ์ระหว่างค่าความถูกต้องของโมเดลกับระดับความสูญหายของชุดข้อมูล Glass Identification

จากภาพประกอบที่ 6 การทดลองชุดข้อมูล Glass Identification ซึ่งได้สมมุติการเกิดข้อมูลที่สูญหายในลักษณะเป็นตัวเลข ซึ่งความถูกต้องในการทำนายของโมเดลที่ 3 การหาค่าความสัมพันธ์ได้ดีกว่าอีกสองโมเดลทุกช่วงที่เกิดข้อมูลที่สูญหาย และถ้ามีการเกิดข้อมูลที่สูญหายมากจะได้ทำให้ความถูกต้องในการทำนายลดลง



ภาพประกอบที่ 7 กราฟแสดงความสัมพันธ์ระหว่างค่าความถูกต้องของโมเดลกับระดับความสูญหาย Forest Fires

จากภาพประกอบที่ 7 การทดลองชุดข้อมูล Forest Fires ซึ่งได้สมมุติการเกิดข้อมูลที่สูญหายในลักษณะเป็นข้อความ ซึ่งความถูกต้องในการทำนายของโมเดลที่ 2 ได้ดีกว่าอีกโมเดลที่ 1 ทุกช่วงที่เกิดข้อมูลที่สูญหาย

สรุปและอภิปรายผล

จากผลการวิจัยที่แสดงการเปรียบเทียบเทคนิคการเติมค่าให้กับข้อมูลที่สูญหาย จะปรากฏว่าเทคนิคการทำนายค่าของข้อมูลที่สูญหายจากตารางการทดลอง สำหรับชุดข้อมูลที่ค่าของข้อมูลสูญหายซึ่งทั้งสามเทคนิคจะสังเกตว่าเทคนิคที่ใช้ค่าความสัมพันธ์จะมีค่าความถูกต้องที่สูงกว่าเทคนิคการตัดข้อมูลที่สูญหายกับเทคนิคการใช้ค่าเฉลี่ยหรือค่ากึ่งกลาง ถ้าข้อมูลที่สูญหายเป็นข้อความเทคนิคการใช้ค่าที่ปรากฏซ้ำบ่อยที่สุดจะดีกว่าเทคนิคการตัดแถวข้อมูลที่มีค่าที่สูญหาย และถ้าการแบ่งข้อมูลที่สูญหายออกเป็นช่วงๆ ที่เกิดข้อมูลที่สูญหายซึ่งเทคนิคการใช้ค่าของความสัมพันธ์จะได้ประสิทธิภาพในการทำนายดีกว่าเทคนิคอื่น ๆ และถ้าข้อมูลที่เป็นข้อความเทคนิคการใช้ค่าที่ปรากฏซ้ำจะได้ประสิทธิภาพการทำนายผลที่ดีกว่าเทคนิคการตัดข้อมูลที่มีค่าสูญหายออก

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับทุนสนับสนุนจากมหาวิทยาลัยเทคโนโลยีสุรนารี

Discretization and Imputation Techniques for Quantitative Data Mining

Nuntawut Kaoungku*, Phatcharawan Chinthaisong, Kittisak Kerdprasop, and Nittaya Kerdprasop

Abstract—**Association rule mining from numerical datasets has been known inefficient because the number of discovered rules is superfluous and sometimes the induced rules are inapplicable. In this paper, we propose the discretization technique based on the Chi2 algorithm to categorize numeric values. We also handle missing values in the dataset with statistical methods. The discovered association rules are then evaluated with the four measurement metrics, that is, confidence, support, lift, and coverage. The dataset imputed with various missing value handling techniques has also been evaluated with the tree-based data classification method to assess predictive accuracy.**

Index Terms—**Association rule analysis, Data mining, Discretization, Missing value imputation**

I. Introduction

Current adoption of data mining technology can be seen in various fields such as economics, education, engineering, life science, medicine, and many more. The models automatically learned from data can facilitate future event prediction, as well as can explain current relations. Models built from datasets with some missing values can, however, cause error in the prediction. Efficient predictive model building, thus, requires the imputation of missing data.

In this research, we comparatively perform three schemes of missing value handling. These schemes are (1) removing record that show missing data, (2) imputation with average attribute value, and (3) imputation with the most correlated value. After data imputation, we investigate these missing value handling scheme through the decision tree induction technique. The decision tree induction is a data classification technique. This data mining task aims at inducing a model in a form of decision tree. This kind of model can be used to predict class of data that may occur in the future. We can call this kind of task as predictive data mining.

Manuscript received December 8, 2012; revised January 10, 2013. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

N. Kaoungku is a master student with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: b5111299@gmail.com).

P. Chinthaisong is a master student with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: killuakaara@gmail.com).

K. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand.

N. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand.

Another kind of learning task is explanatory data mining. The purpose of such task is to explain existing relationships among various data attributes. Association rule mining is a type of data mining that will find the association among data objects and create a set of rules to model relationships. To perform association rule mining, data to be mined have to be categorical. Discretization of numerical values is thus an essential data preparation step for association rule mining.

In this paper, we propose a framework of missing value imputation and numerical data discretization as two major preparation steps for classification and association mining tasks. We also present evaluation results of classification and association rule mining using afferent benchmarks.

This research solves the problem by preparing dataset appropriately before association and classification of discretization methods for numeric in association rule and predicts of data missing that is closest to the most possible value.

II. Preliminaries and Related Work

Data can be in a variety of formats. For example, numeric data, nominal data, and a mixed type of numeric and nominal data. But data mining in some categories is not possible. For instance, to find the association rules from dataset with numeric attributes is impossible for some algorithms. Therefore, methods for managing numeric attribute data is essential. The common method to handle categorizing numeric values is discretization. Many current researches on how to divide the discretization in a variety of ways. For example in [3], Chi 2 algorithm was used as discretization method for handling numeric attributes. The discretization methods for numeric attributes in association rule analysis [7] had been applied with R language [3]. The algorithms used to discretization are, The Chi2 algorithm formed by x^2 they are often used in statistics and discretization methods for numeric attributes.

The predictive value of the data missing is another important problem. We comparatively study the value of data missing technique, lost out in praise. The average value in that column if data missing is disrupted data or skew data. We are used median value. If the data in column aren't numeric. We used value that appears most often in the column. And how to use the value of the correlation between a column that has a data missing value with another column that is associated with the most. Other research also has using Rough Set theory [5] Include is used to determine the association between each column is set to create a rule that allows predicting. Datasets were used in this study was a series of patients that most data is dispersed across numeric data. With the numerical data will be grouped into ranges

(Discretized) are so easy to do the research to find the value of the data missing. And Jianhua's research [1] propose a technique to fill up the missing data by using Rough Sets theoretical and add a technique to compare the 3 methods: how to cut data rows that contain data values that are missing out, and data mining. How to select values that are come to missing data from data that contains values that appear most frequently in the column, and how to convert the entire datasets as a Discernibility matrix and create a rule for predicting the missing value. By using a series of six sets of data to compare efficiency, how to find the value of the data that is missing all three methods and data sets through the technique of value for the information that is missing, and the range, and then create a decision tree to test the data prepared for the test of efficiency technique to predict the best. The rules for an association with the four measurements the effectiveness and value of the gauge is decision of each of the algorithm for discretization methods for numeric attributes.

III. METHODOLOGY

A. Framework

This research proposed discretization and imputation techniques for quantitative data mining. Figure 1 shows conceptual framework of the research. First, the missing value imputation has been applied. Second, the discretization has been performed on numeric attributes. Third, apply the association rule mining. Finally, the benchmarks on association rule mining result are to be evaluated.

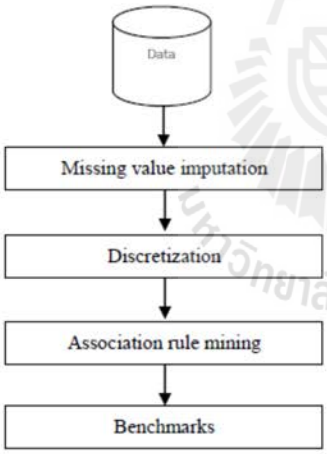


Fig. 1 Conceptual framework of the research

B. Predict the missing value

Techniques to handle missing values in our study are as follows:

1) Remove record that some values are missing.
2) Impute missing values with the average value of the attribute, if the data is normally distributed.
3) Use the correlation of column with missing values to another column, and impute with that column's value.

C. Algorithm Chi2

Chi2 algorithm that is based on the x^2 statistics was used to perform discretization the numerical data [4]. The computation for x2 is as follows.

$$x^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\left(A_{ij} - E_{ij}\right)^2 / E_{ij} \qquad (1)$$

where:

k = number of classes,

A_{ij} = number of patterns in the ith interval, jth class,

Eij = expected frequency of $A_{ij} = R_i * C_j / N$

R_i = number of patterns in the ith interval = $\sum_{j=1}^{k} A_{ij}$

C_j = number of patterns in the ith class = $\sum_{i=1}^{2} A_{ij}$

N = total number of patterns = $\sum_{i=1}^{2} R_i$

The Chi2 algorithm is divided into two parts. The first part starts with a high level of significance, that is 0.5 (sigLevel = 0.5), for all numerical data. After that, it will sort all the numbers continuously.

Part 2 will be on the sideline of the first start of sigLevel0 as set forth in Part 1, then the consistency check after performing an individual attribute the inconsistency rate cannot exceed the assigned sigLevel [i] for inclusion attributes in the next round. This process stops when there is no value left in the attribute.

D. Benchmarks

The benchmarks in this study are the four measurements: support, confidence, lift, and coverage.

1) Support is the frequency of the event occurring, Compute support of equation (2).

$$Support(A \rightarrow B) = P(A \wedge B) \qquad (2)$$

2) Confidence is the frequency of the incident with other events occurring together, Compute confidence of equation (3).

$$Confidence(A \rightarrow B) = Supp(A \rightarrow B) / Supp(A) \qquad (3)$$

3) Lift is the influence of the association rule mining, Compute lift of equation (4).

$$Lift(A \rightarrow B) = Conf(A \rightarrow B) / Supp(A) \qquad (4)$$

4) Coverage is considered the frequency of the association rules mining, Compute coverage of equation (5).

$$Coverage(A \rightarrow B) = Supp(A) = P(A) \qquad (5)$$

IV. Experimental Results

This research experimentation used Hepatitis dataset from the UCI Machine Learning Repository [7]. Hepatitis dataset has 20 attributes and 103 data instances.

For discretization and imputation techniques for quantitative data mining, we used classification and association mining for experimental result assessment. Table 1 and Fig.2 show comparative accuracy of classification both algorithm missing value and missing value + discretization of three models. Model 1 is removing records that contain missing values. Model 2 is missing value imputation with the attribute mean. Model 3 is missing value imputation with correlated value.

TABLE 1
COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY

Algorithm	Model 1	Model 2	Model 3
Missing value	65.95%	74.46%	80.85%
Missing value + Discretize	85.13%	89.36%	87.23%

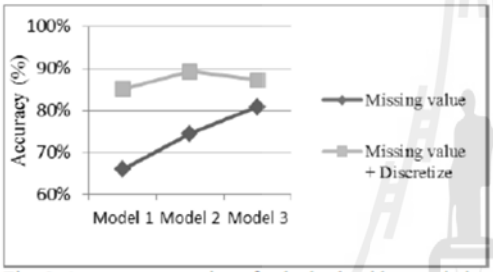


Fig. 2 Accuracy comparison for both algorithms: missing value and missing value + discretization

Table 2 show comparative results of association rule mining using the average of support, confidence, lift, and coverage values to measure performance.

TABLE 2
COMPARATIVE RESULTS OF ASSOCIATION RULE MINING

Models	The average of support	The average of confidence	The average of lift	The average of coverage
Model 1	60.99%	97.66%	103.63%	62.65%
Model 2	62.56%	98.37%	102.94%	63.78%
Model 3	62.02%	98.33%	103.07%	63.27%

Fig. 3 compares the average of confidence and lift for three models. It can be seen from the result that model 3 is the highest compared to the other models.

Fig. 4 compares the average of support and coverage values for three models. It can be seen from the result that model 2 is the highest compared to the other models.

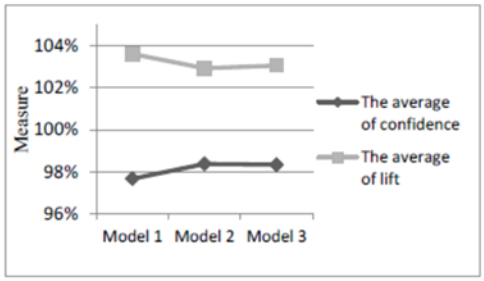


Fig. 3 Comparative the average of confidence and lift both three models

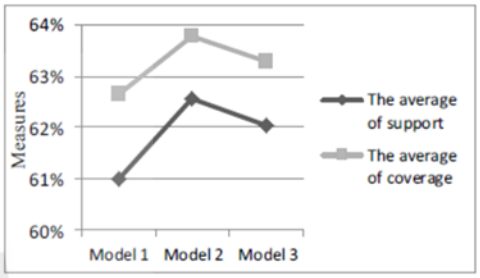


Fig. 4 Comparative the average of support and coverage both three models

V. Conclusion

This research aims to study discretization and imputation techniques for quantitative data mining. The results show that the best model of classification is model 2 that used missing value imputation with the average value if the data is normally distributed and used chi2 for discretization. The results also show that the best model of association rule mining is model 2. Therefore, it can be concluded that the model 2 that imputes missing values by attributes means gives the best result.

References

[1] Jianhua Dai, Qing Xu, and Wentao Wang (2011). "A Comparative Study on Strategies of Rule Induction for Incomplete Data Based on Rough Set Approach," *International Journal of Advancements in Computing Technology*, vol 3, no. 3, pp.176-183

[2] Kittisak Kerdprasop (2012). "Data Mining Methodology and Development," Retrieved November 1, 2012, from https://sites.google.com/site/kittisakthailand55/home/datamining2-55

[3] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash (2002). "Discretization: An Enabling Technique," *Data Mining and Knowledge Discovery*, vol.6, no.4, pp. 393-423

[4] Huan Liu and Rudy Setiono (1995). "Chi2: Feature Selection and Discretization of Numeric Attributes," *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pp. 388-391

[5] Fulufhelo V. Nelwamondo and Tshilidzi Marwala. (2007). "Rough Sets Computations to Impute Missing Data," CoRRabs/0704.3635

[6] UC Irvine Machine Learning Repository, (1988) Hepatitis Data Set. Retrieved October 5, 2012 from http://archive.ics.uci.edu/ml/datasets/Hepatitis

[7] Mohammed J. Zaki (2002). "Scalable Algorithms for Association Mining," *IEEE Transaction on Knowledge and Data Engineering*, vol.12, no.3, pp. 372-390

APPENDIX

Source code in R language to perform missing value imputation and discretization is presented as follows:

```
# Missing value imputation
  hepatitis<-read.csv("hepatitis.csv", fill = TRUE)
  hepatitis <- hepatitis [-c(62,199) , ]
  predict1<- function(dataM){
     cutMissing <-na.omit(dataM)
     return(cutMissing)
  }

  predict2<- function(dataM,colM,more=F){
     if (more){
        dataM[is.na(dataM[[colM]]),colM]<-
     mean(dataM[[colM]],na.rm=T)
     }else{
        dataM[is.na(dataM[[colM]]),colM]<-
    median(dataM[[colM]],na.rm=T)
     }
     return(dataM)
  }

  lookCor<- function(crn){
     gg<-(cor(crn,use='complete.obs') )
     gp<-symnum(gg)
     return(gp)
  }
  creXY<- function(colM,dataM,NN){
     mM<-lm(colM,data=dataM)$coefficients[NN]
     mN<-mM[1][[1]]
     return(mN)
  }
  inputf<- function(oP){
     if ( is.na(oP) ) return(NA)
     else return ( (oP+(-mY))/mX )
  }
  cor.input<- function(colA,colB,dataM){
     dataM[ is.na ( dataM[[colA]] ),colA ] <-
     sapply ( dataM[ is.na (dataM[[colA]]),colB],inputf)
     return(dataM)
  }

  dataset1<-predict1(hepatitis)
  dataset2<-predict2(hepatitis,"Chla",T)
  dataset2<-predict2(dataset2,"Cl",T)
  dataset2<-predict2(dataset2,"PO4",F)

  mX<-creXY(oPO4~PO4,hepatitis,2)
  mY<-creXY(oPO4~PO4,hepatitis,1)
  dataset3<-predict3("PO4","oPO4",hepatitis)
  dataset3<-predict3("Chla","oPO4",dataset3)

  library(rpart)

  rt.a1<-rpart(a1~.,data=dataset1[,1:12])
  plot(rt.a1,uniform=T,branch=1, margin=0.1, cex=0.9)
  text(rt.a1,cex=0.75)

  rt.a2<-rpart(a1~.,data=dataset2[,1:12])
  plot(rt.a2,uniform=T,branch=1, margin=0.1, cex=0.9)
  text(rt.a2,cex=0.75)

  rt.a3<-rpart(a1~.,data=dataset3[,1:12])
  plot(rt.a3,uniform=T,branch=1, margin=0.1, cex=0.9)
  text(rt.a3,cex=0.75)

  testPred <- predict(rt.a1, newdata = test.hepatitis)
  print(testPred)
  table(testPred, test.hepatitis$a1)

#Discretization
  hepatitisM<-read.csv("hepatitis.csv", fill = TRUE)
  new.dataset<-chi2(hepatitisM,0.5,0.05)$Disc.data

#Association rules mining
  rules <- apriori(new.dataset, parameter= list(supp=0.5,
  conf=0.8))

# Benchmarks
  quality(rules) <- cbind(quality(rules), coverage =
  interestMeasure(rules, method = "coverage", tr))
  WRITE(rules, file = "data_disc.csv", quote=TRUE, sep =
  ",", col.names = NA)
```

ประวัติผู้เขียน

นางสาวพัชราวรรณ ชินไธสง เกิดเมื่อวันที่ 9 สิงหาคม พ.ศ.2533 เริ่มเข้าอนุบาลชั้นที่ 1 จนถึงสำเร็จการศึกษาในระดับชั้นประถมศึกษาที่ 6 ที่โรงเรียนชุมชนชวนวิทยา อำเภอบำเหน็จณรงค์ จังหวัดชัยภูมิ หลังจากนั้นได้เข้าศึกษาต่อในระดับชั้นมัธยมศึกษาตอนต้นจนสำเร็จการศึกษาในระดับชั้นมัธยมศึกษาตอนปลายที่โรงเรียนบำเหน็จณรงค์วิทยาคม อำเภอบำเหน็จณรงค์ จังหวัดชัยภูมิ ต่อมาในปีการศึกษา 2551 ได้เข้าศึกษาต่อในระดับปริญญาตรีและสำเร็จการศึกษาระดับปริญญาตรีในปีการศึกษา 2554 ในสาขาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรานารี ภายหลังจากนั้นได้เข้าศึกษาต่อในระดับบัณฑิตศึกษาหลักสูตรปริญญาโท สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรานารี ในปี 2555

ระหว่างการศึกษาในระดับบัณฑิตศึกษาได้รับความอนุเคราะห์จากอาจารย์ประจำวิชา Knowledge Discovery And Data Mining และ Database System ซึ่งได้รับความไว้วางใจให้เป็นผู้ช่วยสอนปฏิบัติการ และงานวิจัยที่ได้ศึกษาก็ได้รับการตีพิมพ์ในเอกสารการประชุมวิชาการซึ่งรายละเอียดสามารถดูได้ที่ภาคผนวก ก