บทที่ 1

บทนำ

การทำเหมืองข้อมูล (Data mining) กำลังเป็นหัวข้อสำคัญและมีผู้ให้ความสนใจอย่างแพร่หลายในหลาย ๆ วงการ ทั้งทางด้านวิชาการศึกษาทั้งเชิงสถิติ การวิจัยและพัฒนา การวางแผน รวมถึงการนำไปใช้จริงในเชิงธุรกิจเพื่อรองรับการตัดสินใจของผู้บริหารซึ่งจะเป็นผู้ใช้ประโยชน์หรือความรู้ที่ได้จากการทำเหมืองข้อมูล

การจัดกลุ่มข้อมูล (Data Clustering) หรือการแบ่งกลุ่ม เพื่อหาตัวแทนของกลุ่มข้อมูลนั้น ๆ หรือหาจุดกึ่งกลางของกลุ่มข้อมูล (Centroids) ก็เป็นกระบวนการหนึ่งที่จำเป็นและมีความสำคัญสำหรับการทำเหมืองข้อมูล การจัดกลุ่มข้อมูลเพื่อแบ่งข้อมูลออกเป็นกลุ่มซึ่งข้อมูลในแต่ละกลุ่มจะต้องมีความเหมือนกัน และกลุ่มของข้อมูลที่ถูกจัดแบ่งจะต้องมีความแตกต่างกันที่ชัดเจน และรวมถึงจะต้องมีการวิเคราะห์ถึงข้อมูลที่ไม่เข้าพวก (Outliner) จึงจะถือได้ว่า การจัดกลุ่มข้อมูลมีความถูกต้องและมีประสิทธิภาพ พร้อมที่จะนำไปใช้ต่อในกระบวนการทำเหมืองข้อมูลต่อไป

การจัดกลุ่มข้อมูล มีทฤษฎีที่เป็นที่ยอมรับได้และนำมาใช้จริงแล้ว เช่น K-means อัลกอริทึม ต่อมาได้มีนักวิจัยนำเสนอวิธีการต่าง ๆ ตามมามากมาย หนึ่งในหลาย ๆ สิ่งที่กำลังได้รับความสนใจและนำเสนอเพื่อมาแก้ปัญหาให้กับกระบวนการทำเหมืองข้อมูลคือ Swarm Intelligence ซึ่งแบ่งได้เป็น 2 อัลกอริทึม คือ (1) Ant Colony Optimization (ACO) (การหาค่าที่เหมาะที่สุดด้วยโคโลนีมด) และ (2) Particle Swarm Optimization (PSO) (การหาค่าเหมาะที่สุดด้วยการเคลื่อนที่ของกลุ่มอนุภาค) โดย ACO ก็มีนักวิจัยให้ความสนใจนำมาประยุกต์ใช้กับกระบวนการในการทำเหมืองข้อมูลอย่างหลากหลาย Parpinelli Lopes และ Freitas (2002) นำเสนออัลกอริทึมสำหรับการจำแนกกฎจากข้อมูลสำหรับการทำเหมืองข้อมูลเรียกว่า Ant-Miner เปรียบเทียบกับอัลกอริทึม CN2 ทดสอบกับชุดข้อมูลหกชุด ได้ผลว่าอัลกอริทึมที่นำเสนอมีความถูกต้องและยังได้กฎที่น้อยกว่า เรียบง่ายกว่าอัลกอริทึม CN2 ที่นำมาทดลองเปรียบเทียบ Zhu และ Li นำเสนอ OAAC สำหรับจำแนกกฎในการทำเหมืองข้อมูลมีประสิทธิภาพที่รวดเร็วและยังมีผลลัพธ์ที่สามารถนำไปใช้ได้จริง ส่วน Ji,Zhang,Liu และ Zhong (2006) เพิ่มความสามารถให้กับ Ant-Miner ได้กฎที่น้อยลงและมีประสิทธิภาพรวดเร็วขึ้นในการทำงาน Wei และ Zhu-hua (2008) นำเสนอการปรับปรุงอัลกอริทึม Ant colony clustering สำหรับการพยากรการสูญเสียกำลังในช่วง

เวลาสั้น ๆ ได้ผลการทดสอบที่มีประสิทธิภาพ และอัลกอริทึมที่นำเสนอยังมีความอ่อนไหวต่ออุณหภูมิและสภาพอากาการที่กำหนดเงื่อนไขไว้อีกด้วย ซึ่งในระยะยาวของการทำงานจะต้องพัฒนาปรับปรุงขึ้นไปอีก ดังที่กล่าวมาแล้ว สิ่งหนึ่งที่ยังคงเป็นปัญหาสำหรับ ACO คือการเพิ่มประสิทธิภาพในการทำงานที่ได้ผลลัพธ์ที่รวดเร็วเพราะ ACO เองก็ต้องยอมรับว่า อัลกอริทึมของตนนั้นดีแต่ไม่ได้ดีที่สุด (good but it not best) เพราะการทำงานของ อัลกอริทึม นี้จะเป็นแบบรวมศูนย์ โดยการแลกเปลี่ยนข้อมูลเพื่อหาแนวทางที่ดีที่สุด ของการจำลองพฤติกรรมของมดที่อยู่ภายในอาณาจักรหรือรังเดียวกันในการออกไปหาแหล่งอาหารก็ยังคงต้องกลับมาแลกเปลี่ยนข้อมูลเส้นทางที่ค้นพบกันที่รังของมัน แต่สำหรับ PSO นั้นจะเป็นการประมวลผลระหว่างตัวแทน (agent) หรืออนุภาค (particle) ซึ่งได้มาจากการจำลองชีวิตของสัตว์อื่น ๆ เช่น นกที่อาศัยรวมกันเป็นฝูง หรือ ฝูงปลาหรือฝูงผึ้งซึ่งสัตว์เหล่านี้ในชีวิตจริงไม่จำเป็นที่จะต้องกลับไปแลกเปลี่ยนข้อมูลเส้นทางในการไปหาแหล่งอาหารถึงที่อยู่หรือรังของตนเอง สามารถที่จะแลกเปลี่ยนข้อมูลได้ระหว่างทางเพื่อย่นระยะเวลาในการค้นหาอาหาร PSO จึงเป็นอัลกอริทึมที่น่าจะมีประสิทธิภาพในด้านเวลาสำหรับการค้นหาข้อมูลมากกว่า ACO ในการนำ PSO มาประยุกต์ใช้ในกระบวนการทำเหมืองข้อมูลนั้นอีกหัวข้อหนึ่งที่นักวิจัยให้ความสนใจ และพยายามที่จะได้เกิดการทำงานที่เป็นไปอย่างอัตโนมัติในการจัดเตรียม วิเคราะห์ข้อมูลที่ต้องการในการค้นหาองค์ความรู้ แต่สิ่งที่ PSO ไม่สามารถแก้ไขปัญหาของตัวเองคือ local optima (ปัญหาในการที่ได้ค่าผลลัพธ์ที่คิดว่าค่าที่ได้เป็นค่าที่ดีที่สุดแล้วแต่ในความเป็นจริงอาจมีค่าอื่นที่ดีกว่า) ซึ่งอาจจะเกิดขึ้นได้ในการใช้อัลกอริทึมนี้

ในงานวิจัยที่ผ่านมา Lu,Luo,Liu และ Long (2008) นำเสนออัลกอริทึม HSPPSO ใช้แนวคิดของการควบคุมโครงสร้างแบบลำดับชั้น เพื่อเพิ่มประสิทธิภาพให้ PSO แบบดั้งเดิมมีความถูกต้องมากยิ่งขึ้น Marco,Montes,Pena,Stutzle,Pinciroli และ Dorigo นำเสนอการนำอนุภาคที่ไม่เข้าพวกกันมาช่วยกันทำงานแต่ได้ผลไม่ดีในทุก ๆ ปัญหาที่นำมาประเมินผลแต่โดยรวมก็มีผลการทดสอบที่ดี ทั้งสองบทความดังกล่าวเป็นความพยายามที่จะแก้ไขปัญหาของ PSO ให้เป็นไปอย่างมีประสิทธิภาพ ได้มีการนำเสนออัลกอริทึมเพื่อปรับปรุงการทำงานของ PSO เดิมให้หลุดพ้นจากปัญหา local optima งานวิจัยหนึ่ง Jiang Huang และ Chan (2009) ซึ่งนำเสนอแนวทางในการใช้ Multi-Swarm และเพิ่มกระบวนการเพิ่มหรือเร่งการทำงานให้มีประสิทธิภาพบนพื้นฐานของ PSO แบบดั้งเดิม และให้นิยามอัลกอริทึมใหม่ว่า MSA-PSO นั้น มาวิเคราะห์ปัญหาการทำงานอย่างต่อเนื่องแบบไม่คงที่และได้ผลดีกว่า ในงานวิจัยฉบับนี้จะนำเอา MSA-PSO มาประยุกต์กับการจัดกลุ่มข้อมูลในกระบวนการทำเหมืองข้อมูล เพื่อให้กระบวนการนี้สามารถรองรับการทำงานกับกลุ่มข้อมูลแบบไม่คงทีได้ดีเพิ่มขึ้นด้วย ซึ่งต้องยอมรับว่าหลาย ๆ สิ่ง

ในชีวิตจริงนั้นเป็นแบบไม่คงที่เช่น การเปลี่ยนแปลงราคาของหุ้นในตลาดหุ้น ความต้องการของลูกค้า ณ เวลาหนึ่งเวลาใด และข้อมูลที่ขึ้นอยู่กับเวลาที่เปลี่ยนแปลง เป็นต้น