179202

ปัจจุบันแบบจำลองภาษาสโตแคสติกหรือภาษาเชิงน่าจะเป็นมีอยู่หลายชนิด ได้แก่ แบบจำลองมาร์คอฟแบบซ่อน แบบจำลองลูกโซ่มาร์คอฟ ออโตมาตาอิงความน่าจะเป็น และอื่นๆ งานวิจัยเกี่ยวกับแบบจำลองเหล่านี้ส่วนมากเน้นที่การสร้างแบบจำลองเพื่อการเรียนรู้เชิงสถิต นั่นคือไม่เน้นการเปลี่ยนแปลงภายหลังการเรียนรู้ เนื่องจากเทคโนโลยีทางคอมพิวเตอร์ในปัจจุบันทำให้เกิดข้อมูลรูปแบบใหม่เกิดขึ้น จำเป็นต้องอาศัยการเรียนรู้ที่สามารถเรียนรู้ได้ตลอดเวลาและสามารถตัดสินใจได้ตามสถานการณ์ที่แปรเปลี่ยนไป เพื่อรองรับกรณีที่ข้อมูลมีการแปรผันอยู่ตลอดเวลา ทำให้เกิดอุปสงค์ต่อแบบจำลองการเรียนรู้เชิงพลวัต ที่สามารถเรียนรู้และพัฒนาตัวเองได้ตลอดเวลาที่มีข้อมูลใหม่เข้ามา ในงานวิจัยนี้สนใจแบบจำลองที่มีความสามารถดังกล่าวซึ่งสามารถเรียนรู้ได้จากแหล่งข้อมูลขนาดใหญ่ที่เป็นสายข้อมูลแบบต่อเนื่อง และปราศจากการกำหนดข้อกำหนดเบื้องต้นจากมนุษย์ งานวิจัยนี้ได้เสนอแบบจำลองชนิดใหม่และวิธีการเรียนรู้ของแบบจำลองดังกล่าวเพื่อให้มีความสามารถตามที่กล่าวไว้ข้างต้น โดยอิงกระบวนทัศน์ในการเรียนรู้แบบ การจำแนกภายในจำกัด เรียกแบบจำลองชนิดใหม่นี้ว่า สโทแคสติกออโตมาตาอิงอักขระนำหน้าแบบจำกัด โดยที่อัลกอริทึมในการเรียนรู้ของแบบจำลองชนิดนี้มีฟังก์ชันการเติบโตของเวลาในการเรียนรู้ไม่เกินฟังก์ชันพหุนามและขนาดของแบบจำลองขึ้นอยู่กับบริบทของข้อมูล ซึ่งมีอัตราการเติบโตของขนาดไม่เกินฟังก์ชันพหุนามเช่นกัน

Nowadays, there are several probabilistic (stochastic) language models such as Hidden Markov Model, Markov Chain, Probabilistic Automaton, etc. Most of recent researches on probabilistic models are to improve how to construct a static learning model. However, some work use dynamic mechanism to learn and update. In practice, we need a dynamic model which should be easier and faster to induce, and it should be able to update in real time. We are interested in a model which is able to learn from continuous data stream without prior knowledge or finite length of input. In this research, Identification in the limit is a learning criterion. We propose a novel probabilistic automaton named stochastic finite precedent automaton and its induction algorithm. Particularly, its induction or update algorithm can perform in an on-line mode within polynomial time complexity while the size of the model is dependent on the context of data.