

Abstract

Project Code : MRG5980235

Project Title : Enhanced label noise model for classification in the presence of non-random labelling errors

Investigator : Jakramate Bootkrajang
Department of Computer Science, Faculty of Science, Chiang Mai University

E-mail Address : jakramate.b@cmu.ac.th

Project Period : One year (2 May 2016 – 1 May 2017)

Learning from labelled data is becoming more and more challenging due to inherent imperfection of training labels. Existing label-noise tolerant learning machines were primarily designed to tackle class-conditional noise which occurs at random, independently from input instances. However, relatively less attention was given to a more general type of label noise which is influenced by input features.

In this work, we will address the problem of learning a classifier in the presence of instance-dependent label noise by developing a novel label noise model which is expected to capture the variation of label noise rate within a class. This is accomplished by adopting a probability density function of a mixture of Gaussians to approximate the label flipping probabilities. Experimental results demonstrate the effectiveness of the proposed method over existing approaches.

Keywords : Classification, Logistic regression, Instance-dependent label noise

บทคัดย่อ

รหัสโครงการ : MRG5980235

ชื่อโครงการ : การปรับปรุงตัวต้นแบบสำหรับการจำแนกข้อมูลที่ถูกรบกวนด้วยสัญญาณรบกวนในฉลากแบบไม่สุ่ม

ชื่อนักวิจัย : จักรเมธ บุตรกระจำ
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

E-mail Address : jakramate.b@cmu.ac.th

ระยะเวลาโครงการ : 1 ปี (2 พฤษภาคม 2559 – 1 พฤษภาคม 2560)

การพัฒนาตัวจำแนกจากข้อมูลที่มีฉลากกำกับในปัจจุบันมีความท้าทายเพิ่มมากขึ้น สืบเนื่องมาจากความไม่สมบูรณ์ของฉลากสำหรับฝึกหัด ตัวจำแนกที่มีความทนทานต่อสัญญาณรบกวนในฉลากที่มีใช้กันอยู่ ถูกออกแบบและพัฒนาเพื่อต่อกรกับสัญญาณรบกวนที่ขึ้นกับชั้นข้อมูล หรือที่รู้จักกันทั่วไปว่า สัญญาณรบกวนในฉลากแบบสุ่ม ซึ่งเกิดขึ้นโดยอิสระจากข้อมูลขาเข้า ทว่าความเข้าใจเกี่ยวกับสัญญาณรบกวนในฉลากอีกประเภทที่เกิดขึ้นจากอิทธิพลของข้อมูลขาเข้ายังมีอยู่น้อย

ในงานวิจัยชิ้นนี้ ผู้วิจัยมุ่งที่จะศึกษาการพัฒนาตัวจำแนกในสถานการณ์ที่ข้อมูลฝึกหัดมีสัญญาณรบกวนที่ขึ้นกับข้อมูลขาเข้าปะปนอยู่ด้วย โดยจะพัฒนา ตัวต้นแบบแบบใหม่ที่สามารถจะตรวจจับและจำลองการเกิดขึ้นของสัญญาณรบกวนดังกล่าวได้หลากหลายมากขึ้น ในการนี้ ผู้วิจัยเลือกใช้การแจกแจงแบบ Mixture of Gaussians มาใช้ในการประมาณการกระจายตัวของสัญญาณรบกวนในฉลากแบบไม่สุ่มในชุดข้อมูล

ผลการทดลองแสดงให้เห็นว่าวิธีที่นำเสนอไปนั้นมีประสิทธิภาพในการต่อกรกับสัญญาณรบกวนในฉลากแบบไม่สุ่มมากกว่าวิธีการเดิมที่มีอยู่แล้ว

Keywords : Classification, Logistic regression, Instance-dependent label noise