Multi-Label Classification and Clustering for Acoustics and Computer Security

Author(s):
Streich, Andreas P.

Publication Date:
2010

Permanent Link:

https://doi.org/10.3929/ethz-a-006212098 →

Diss. ETH No. 19229

Multi-Label Classification and Clustering for Acoustics and Computer Security

A dissertation submitted to

Swiss Federal Institute of Technology

Zurich

for the degree of

Doctor of Sciences

presented by

Andreas Peter Streich

Dipl. Ing. (ETH Zurich)

born 29 April 1980

citizen of Basel, Switzerland

accepted on the recommendation of

Prof. Dr. Joachim M. Buhmann, examiner

Prof. Dr. David Basin, co-examiner

Dr. Stefan Launer, co-examiner

2010

Abstract

This thesis focuses on classification and clustering of data where a part of the data items are jointly emitted by several sources. We design an abstract generative model which offers a clear semantic interpretation for such data. Based on this model, we derive algorithms for multi-label classification and multi-assignment clustering.

For the task of multi-label classification, we show that the presented algorithms estimate source parameters more accurately and classify data items more reliably than previously proposed methods. We apply our method to classify acoustic streams in hearing instrument. Most modern hearing instruments rely on such classification to adapt to acoustic scenes encountered in daily live. In this setting, a correct detection of the present sources is essential to provide comfortable listening in spite of a hearing impairment. We propose a novel set of features for this classification task and show that our generative multi-label classification algorithm outperforms current techniques.

The generality of our model formulation allows us to describe prior work in the same framework. Starting from this unified specification, we derive the asymptotic distribution of the parameter estimators obtained by several algorithms. Furthermore, we prove that a class of popular model assumptions implies a mismatch to the assumed generative process and therefore causes an inconsistency of the parameter estimators and, consequently, suboptimal classification results.

The generative algorithms for multi-assignment clustering are applied to Boolean data. Also in this unsupervised setting, the parameters estimated by the proposed algorithms are more precise and the obtained clustering solution attains higher stability, both compared to state-of-the-art methods. We apply our method to solve an important problem in computer security

known as role mining. The Permissions of new users can be specified more precisely with the roles obtained by our generative methods than with roles detected by other multi-assignment clustering techniques.

To compare the quality of different clustering techniques independently of particular assumptions, we apply the framework of approximation set coding for cluster validation. We observe that the model selection based on this general framework is in agreement with the selection based on specific quality measures for multi-assignment clustering. According to both criteria, the proposed algorithms are identified as the best method for the given clustering task. We thus show for the first time that approximation set coding correctly regularizes the model complexity for a real-world learning task.

Zusammenfassung

Diese Dissertation behandelt die Klassifikation und das Gruppieren von Daten unter der Annahme, dass mindestens ein Teil der Datenpunkte von mehreren Quellen gemeinsam generiert wird. Wir entwerfen ein allgemeines generatives Modell mit einer klaren Semantik für derartige Daten. Basierend auf diesem Modell entwickeln wir Algorithmen für die Klassifikation mit Mehrfachzugehörigkeiten und für die Gruppierung mit Mehrfachzuweisungen.

Im ersten Teil der Arbeit gehen wir detailliert auf Klassifikationsprobleme ein, in denen ein Datenelement gleichzeitig zu mehreren Klassen gehören kann. Die von uns vorgestellten Algorithmen schätzen Quellenparameter genauer und klassifizieren synthetische Daten präziser als bisher bekannte Methoden. Anschliessend wenden wir unsere Algorithmen auf die Klassifikation von akustischen Daten an. Die meisten modernen Hörgeräte teilen die akustischen Signale in verschiedene Klassen ein und wählen anschliessend, entsprechend der geschätzten Klasseneinteilung, die der Situation angepasste Verarbeitung des Signals. Dementsprechend hängt die Gesamtleistung des Hörgerätes grundlegend von der korrekten Identifikation der vorhandenen Geräuschquellen ab. Wir präsentieren neue Kenngrössen für diese Klassifikationsaufgabe. In verschiedenen Experimenten ergeben sowohl die vorgestellten Merkmale als auch der generative Ansatz verbesserte Resultate gegenüber dem aktuellen Stand der Technik.

Die Allgemeinheit unserer Formulierung ermöglicht uns ausserdem, die bisherigen Klassifikationsmethoden als Spezialfälle unseres Modells darzustellen. Ausgehend von dieser einheitlichen Beschreibung leiten wir die asymptotische Verteilung der Parameterschätzer verschiedener Methoden her. Wir beweisen ausserdem, dass eine gängige Modellannahme eine Fehlanpassung des Modells an die Daten impliziert und dadurch zu inkonsisten-

ten Parameterschätzern sowie sub-optimalen Klassifikationsresultaten führt.

Im zweiten Teil untersuchen wir die unüberwachte Gruppierung von Daten unter der Verallgemeinerung, dass ein Datenelement gleichzeitig zu mehreren Gruppen gehören kann. Auch in diesem unüberwachten Szenario liefern die vorgeschlagenen generativen Algorithmen präzisere Schätzer der Quellenparameter und ermöglichen eine genauere Beschreibung von neuen Datenelementen, beides im Vergleich zu bisherigen Methoden. Wir wenden den generativen Gruppierungsalgorithmus auf ein wichtiges Problem aus der Computersicherheit an, nämlich dem automatischen Ermitteln einer Menge von Rollen für rollenbasierte Zugangskontrolle. Die von den vorgeschlagenen Algorithmen gefundenen Rollen beschreiben die Zugriffsrechte neuer Benutzer akkurater als die Rollen, welche von bisherigen Methoden mit Mehrfachzuweisungen gefunden werden.

Die Bewertung der Qualität einer Datengruppierung basiert häufig auf Annahmen über die Natur der Datengruppen. Wir verwenden die Methode der Codierung mittels Näherungsmengen um die Qualität der Lösungen verschiedener Gruppierungsalgorithmen zu beurteilen. Die Modellpräferenzen dieser allgemeinen Methode stimmen mit der Auswahl auf Grund von problemspezifischen Kenngrössen überein. Dieses Modellselektionsprinzip identifiziert den vorgeschlagenen Algorithmus als für die vorliegende Gruppierungsaufgabe am besten geeignet. Damit wurde zum ersten Mal an Realweltdaten bestätigt, dass Codierung mittels Näherungsmengen die Modellkomplexität korrekt kontrolliert.