technische universität
dortmund

Masterarbeit

Kernel k-means Methoden zur spektralen Clusteranalyse von Graphen

Lukas Pradel
9. Dezember 2014

Gutachter:
Prof. Dr. Christian Sohler
Dipl.-Inf. Melanie Schmidt

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl II - Effiziente Algorithmen und Komplexitätstheorie
http://ls2-www.cs.tu-dortmund.de/

Hier kommt eine Danksagung hin!

Inhaltsverzeichnis

1 Einleitung

2 Grundlegende Definitionen und Algorithmen

In diesem Kapitel definieren wir die für unsere Zwecke relevanten Begriffe im Kontext der Clusteranalyse und führen die populärsten grundlegenden Algorithmen ein, deren Ideen für uns im Folgenden noch von Bedeutung sein werden. Wir gehen dabei nach Themengebieten geordnet vor: In Abschnitt 2.1 skizzieren wir kurz das Themengebiet der Clusteranalyse, definieren die wichtigsten Zielfunktionen und stellen zwei wichtige Algorithmen vor. Abschnitt 2.2 führt kurz in die Graphentheorie sowie die Clusteranalyse von Graphen ein. In diesem Abschnitt werden wir zudem die von der klassischen Clusteranalyse sehr unterschiedlichen Optimierungskritierien für die Clusteranalyse von Graphen herausstellen. Schließlich fassen wir in Abschnitt 2.3 die wichtigsten Methoden und Algorithmen aus dem Bereich der spektralen Clusteranalyse zusammen und stellen zudem die wichtigsten Konzepte von Kernel-Methoden vor.

2.1 Clustering und k-means

Clusteranalyse oder „Clustering“ beschäftigt sich mit der Einteilung von Objekten in Gruppen („Cluster“), sodass sich die Objekte innerhalb eines Clusters gemäß eines bestimmten Optimierungskriteriums ähnlich sind und von Objekten eines anderen Clusters unterscheiden. Es existieren zahlreiche grundsätzlich verschiedene Ansätze, Clusteringprobleme zu lösen. Wir beschränken uns in dieser Arbeit auf *partitionierende* Clusteringprobleme und -verfahren. Bei diesen soll eine Menge von d-dimensionalen Punkten, welche der erste Teil der Eingabe ist, gemäß einer Cluster-Zielfunktion möglichst optimal in genau k Cluster unterteilt werden, wobei k der ganzzahlige zweite Teil der Eingabe ist.

Für die Zielfunktion, welche die Nähe oder Ferne von Punkten zueinander quantifiziert, sind bei Eingabepunkten aus $\mathbb{R}^d$ Metriken naheliegend. Intuitiv ist dabei die euklidische Distanz, welche als Zielfunktion für die beiden bekanntesten Clustering-Problemstellungen dient.

Definition 2.1.1 (k-median und k-means). Sei $P \subset \mathbb{R}^d$ und $k \in \mathbb{N}^+$. Das k-median-Problem besteht darin, eine Menge von k (Cluster-)*Zentren* $C = \{c_1, \ldots, c_k\}$ mit $c_i \in \mathbb{R}^d$ zu finden, sodass der folgende Term minimal wird:

$$\sum_{p \in P} \min_{c_i \in C} \|p - c\|$$

Das k-means-Problem unterscheidet sich nur darin, dass bei diesem die Summe der *quadrierten* euklidischen Distanzen zum jeweils nächstgelegenen Zentrum minimiert

werden soll, das heißt, dass der folgende Term minimiert werden soll:

$$\sum_{p \in P} \min_{c_i \in C} \|p - c\|^2$$

Beim *gewichteten* k-means-Problem werden den Eingabepunkten zusätzlich mit einer Funktion $w : P \to \mathbb{R}$ Gewichte zugewiesen. Die zu minimierende Zielfunktion lautet dann entsprechend

$$\sum_{p \in P} \min_{c_i \in C} w(p)\|p - c\|^2$$

Sowohl das k-Median-Problem [MS84] als auch das k-means-Problem [ADHP09] sind optimal NP-schwer lösbar. Typischerweise werden zur Lösung daher approximative oder heuristische Algorithmen eingesetzt.

2.2 Graphen und Clusteranalyse von Graphen

2.3 Kernel-Methoden und spektrales Clustering

Literatur

[ADHP09] Aloise, Daniel ; Deshpande, Amit ; Hansen, Pierre ; Popat, Preyas: NP-hardness of Euclidean sum-of-squares clustering. In: *Machine Learning* (2009), S. 245–248

[MS84] Megiddo, Nimrod ; Supowit, Kenneth J.: On the Complexity of Some Common Geometric Location Problems. In: *SIAM J. Comput.* 13 (1984), Nr. 1, S. 182–196

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie Zitate kenntlich gemacht habe.

Dortmund, den 9. Dezember 2014

Lukas Pradel