DISS. ETH NO. 21628

# INFORMATION-THEORETIC VALIDATION OF CLUSTERING ALGORITHMS

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
MORTEZA HAGHIR CHEHREGHANI
Master of Science in Computer Engineering
Sharif University of Technology, Tehran, Iran
born Feb 23th, 1982
citizen of Iran

accepted on the recommendation of
Prof. Dr. Joachim M. Buhmann, examiner
Prof. Dr. Peter Widmayer , co-examiner
Prof. Dr. Marcello Pelillo , co-examiner

2013

# ABSTRACT

This thesis focuses on an information-theoretic analysis of clustering algorithms. In many real-world applications, because of imprecise or incomplete measurements, the data is contaminated by noise. For instance, gene expression levels might be measured imperfectly due to improper experimental conditions. This renders the empirical output of an algorithm to be unstable. Statistical learning advocates to employ the generalization ability of models as a measure of model quality. Therefore, we introduce the *Minimum Transfer Costs* (MTC) principle for model order selection inside a specific family of models described by a cost function. We employ the principle to compute the number of clusters in several clustering models such as Gaussian Mixture Models, Pairwise Clustering, and Correlation Clustering.

Stability is, however, only one aspect of statistical modeling; the *informativeness* of the solutions is the other side of the modeling trade-off. Maximizing stability without informativeness can yield very simple and useless solutions. An optimal trade-off requires an information-theoretic approach where the uncertainty in the measurements quantizes the solution space and thereby, induces a coarsening in the solution space. Approximation Set Coding (ASC) [Buh10] attempts to address such questions by establishing a conceptual set-based communication scenario. We elaborate *Generalization Capacity* ($\mathcal{GC}$), a context-sensitive principle for model validation based on Approximation Set Coding. An algorithm is assumed as a data processing mechanism that during execution, produces a weight distribution over the solution space. Generalization capacity computes the optimal concentration of the weights, i.e. it measures the maximal rate of reliable information which can be captured by the algorithm.

We establish a principled pipeline to compute and analyze the generalization capacity for model selection and validation in the context of data clustering. This approach provides a framework to address the fundamental learning questions: i) finding the optimal number of clusters, ii) ranking different similarity measures, and iii) validating alternative clustering methods. Efficient approximation schemes such as *mean-field approximation* are utilized to overcome the computational challenges that occur when computing generalization capacity. Furthermore, we propose exploiting a Hamming metric in the solution space to analyze the ad hoc algorithms that do not yield a trajectory of weight distributions over the solution space.

The principle is first exemplified for density estimation in different settings, particularly for learnability phase transitions in the high dimensional limit. Generalization capacity confirms the evolution of the phase transitions detected by order parameters. Moreover, it yields consistent results with other principles such as BIC and MTC.

The principle is then employed to analyze several aspects of well-known graph clustering methods. We particularly investigate the parametrization of the clustering models in an information-theoretic manner. The principle, for example, computes the optimal adaptation of Ratio Cut with respect to different Laplacians, or, determines the optimal termination of Dominant Set clustering. In the same way, we design a prototypical model by augmenting the basic Min Cut model, which is shown to be equivalent to Correlation Clustering, by a *shift* parameter. Its optimal adaptation is obtained through a *context sensitive* search over the space of alternatives, by exploiting the generalization capacity principle. This approach advocates a scientific procedure for validating the optimal model suitable for the specific application at hand, rather than an arbitrary elegant design which might yield bias towards specific types of patterns.

The framework is demonstrated on clustering of experimental gene expression data. For each method and similarity measure, $\mathcal{GC}$ computes the optimal number of clusters. It constitutes a consistent but more general principle than BIC. In particular, we compare different clustering methods and similarity measures and show how properly shifted Correlation Clustering with an appropriate measure extracts the largest amount of reliable information for all algorithms under consideration. In different biological applications, $\mathcal{GC}$ suggests a consistent ranking of similarity measures with respect to the context of the data, e.g. correlation coefficients are preferred for temporal data.

# ZUSAMMENFASSUNG

Diese Arbeit behandelt eine Informations-theoretische Analyse von Cluster-Algorithmen. In vielen praktischen Anwendungen sind die Daten aufgrund unpräziser oder unvollständiger Messungen durch Rauschen verunreinigt. So könnten beispielsweise Gen-Expressions-Levels wegen ungeeigneter experimenteller Bedingungen ungenau gemessen werden. Dies führt dazu, dass die empirische Ausgabe eines Algorithmus instabil wird. Statistisches Lernen plädiert dafür, die Generalisierungsfähigkeit von Modellen als Mass für die Modellqualität einzusetzen. Daher führen wir *Minimum Transfer Cost* (MTC) als Prinzip zur Modell-Ordnungsauswahl innerhalb einer speziellen Familie von Modellen ein, die durch eine Kostenfunktion beschrieben werden. Wir wenden das Prinzip an, um die optimale Anzahl an Clustern für mehrere Clustering-Modelle wie Gauss'sche Mixturmodelle, paarweises Clustern und korrelations-basiertes Clustern zu finden.

Stabilität ist jedoch nur ein Aspekt der statistischen Modellierung; die andere Seite im Modellierungs-Tradeoff ist die Informativität von Lösungen. Das Maximieren der Stabilität - ohne die Informativität zu beachten - kann zu sehr einfachen und nutzlosen Lösungen führen. Ein optimaler Tradeoff verlangt nach einem Informations-theoretischen Ansatz, wobei die Ungenauigkeit in den Messungen den Lösungsraum quantisieren und damit eine Vergröberung im Lösungsraum veranlassen. *Approximation Set Coding* (ASC) [Buh10] versucht diese Fragen anzugehen, indem ein konzeptionelles, Mengen-basiertes Kommunikations-Szenario etabliert wird. Wir arbeiten die sog. *Generalisierungskapazität* ($\mathcal{GC}$) aus, ein Kontext-sensitives Prinzip für die Modellvalidierung basierend auf Approximation Set Coding. Ein Algorithmus wird als datenverarbeitender Mechanismus interpretiert, der im Verlauf seiner Ausführung eine gewichtete Verteilung im Ausgaberaum produziert. Die Generalisierungskapazität berechnet die optimale Konzentration der Gewichte, d.h. sie misst die maximale Rate zuverlässiger Information, die vom Algorithmus erfassst werden kann.

Wir etablieren einen prinzipientreuen Rahmen, um die Generalisierungsfähigkeit zur Modellselektion und -validierung im Kontext von Daten-Clustering zu berechnen und analysieren. Der Rahmen bietet eine wohldefinierte Methodik, um die fundamentalen Fragen der Lerntheorie anzugehen: i) die Bestimmung der optimalen Anzahl an Clustern, ii) ein Ranking der unterschiedlichen Ähnlichkeitsmasse und iii) alternative Clustering-Methoden zu validieren. Wir verwenden effiziente Approximations-Schemen wie die Mean-Field-

Approximation, um die berechnungstechnischen Herausforderungen bei der Berechnung der Generalisierungsfähigkeit zu überkommnen. Wir schlagen die Nutzung einer Hamming-Metrik im Lösungsraum vor, um diejenigen ad-hoc Algorithmen analysieren zu können, die keine Trajektorie von Gewichtsverteilungen über dem Lösungsraum bieten.

Wir veranschaulichen das Prinzip zuerst für die Dichteschätzung in unterschiedlichen Situationen, vor allem der Phasenübergang der Maximum-Likelihood-Schätzung im hochdimensionalen Limit. Die Generalisierungskapazität bestätigt die Entwicklung der Phasenübergänge, die von den Ordnungsparametern detektiert werden. Darüber hinaus sind die Ergebnisse mit anderen Prinzipien wie BIC und MTC konsistent.

Dann wenden wir das Prinzip an, um mehrere Aspekte von wohlbekannten Graph-Clustering Methoden zu analysieren. Speziell untersuchen wir die Parametrisierung der Clustering-Modelle in einer Informations-theoretischen Art. Wir benutzen das Prinzip zum Beispiel, um die optimale Anpassung von Ratio Cut inbezug auf unterschiedliche Laplacians zu berechnen, oder um die optimale Beendigung von Domimant Set Clustering zu bestimmen. In der gleichen Art designen wir ein prototypisches Modell, indem wir das grundlegende Min Cut Modell (welches, wie gezeigt wird, äquivalent zu Correlation Clustering ist) mit einem Shift-Parameter erweitern. Seine optimale Anpassung wird durch eine kontext-sensitive Suche über den Raum der Alternativen erzielt, wobei das Prinzip der Generalisierungs-Kapazität ausgenutzt wird. Dieser Ansatz steht für ein wissenschaftliches Vorgehen, um das optimale Modell - passend für die spezifische, vorliegende Anwendung - zu validieren, anstatt ein beliebiges, elegantes Design zu wählen, welches einen Bias für spezifische Typen von Mustern haben könnte.

Wir demonstrieren den Rahmen für das Clustering von experimentellen Gen-Expressionsdaten. Für jede Methode und jedes Ähnlichkeitsmass bestimmt $\mathcal{GC}$ die optimale Zahl an Clustern. Es stellt ein konsistentes, aber generelleres Prinzip als BIC dar. Wir vergleichen speziell verschiedene Clustering-Methoden und Ähnlichkeitsmasse und zeigen, wie ein angemessen verschobenes Correlation Clustering mit einem angemessenen Mass die grösste Menge an verlässlicher Information unter allen betrachteten Algorithmen extrahiert. In jeder biologischen Anwendung schlägt $\mathcal{GC}$ ein konsistentes Ranking von Ähnlichkeitsmassen bezogen auf den Kontext der Daten vor, z.B. Korrelations-Koeffizienten für zeitliche Daten.