Bestimmung von Datenunsicherheit in einem probabilistischen Datenstrommanagementsystem

Christian Kuka
SCARE-Graduiertenkolleg
Universität Oldenburg
D-26129 Oldenburg
christian.kuka@uni-oldenburg.de

Daniela Nicklas
Universität Bamberg
D-96047 Bamberg
dnicklas@acm.org

Abstract: Für die kontinuierliche Verarbeitung von unsicherheitsbehafteten Daten in einem Datenstrommanagementsystem ist es notwendig das zugrunde liegende stochastische Modell der Daten zu kennen. Zu diesem Zweck existieren mehrere Ansätze, wie etwas das Erwartungswertmaximierungsverfahren oder die Kerndichteschätzung. In dieser Arbeit wird aufgezeigt, wie die genannten Verfahren in ein Datenstrommanagementsystem verwendet werden können, umso eine probabilistische Datenstromverarbeitung zu ermöglichen und wie sich die Bestimmung des stochastischen Modells auf die Latenz der Verarbeitung auswirkt. Zudem wird die Qualität der ermittelten stochastischen Modelle verglichen und aufgezeigt, welches Verfahren unter welchen Bedienungen bei der kontinuierlichen Verarbeitung von unsicherheitsbehafteten Daten am effektivsten ist.

1 Einführung

Für die qualitätssensitive Verarbeitung von Sensordaten ist es notwendig die aktuelle Qualität der Daten zu kennen. Eine der hierbei häufig verwendeten Qualitätsdimensionen ist der statistische Fehler von Sensormessungen. In vielen Fällen wird hierbei die aus dem Datenblatt stammende Kennzahl für die Standardabweichung herangezogen um das stochastische Modell im Sinne einer Normalverteilung zu verwenden. Jedoch kann das Rauschen eines Sensors von vielen Kriterien abhängen und sich vor allem auch dynamisch zur Laufzeit ändern. Eine Form der Qualitätsbestimmung besteht darin, direkt das zugrunde liegende stochastische Modell der Sensormessungen kontinuierlich neu zu ermitteln. Vor allem im Bereich der kontinuierlichen Verarbeitung von hochfrequenten Sensordaten ist es hierbei notwendig die Speicherkapazitäten des Systems zu beachten und die Daten so schnell wie möglich zu verarbeiten. Für diese Form der Verarbeitung existiert mittlerweile eine Vielzahl von Systemen, welche unter dem Begriff Datenstrommanagementsystem zusammengefasst werden können.

Im Rahmen von Datenstrommanagementsystemen hat sich für die Verarbeitung von Unsicherheiten der Begriff der probabilistischen Datenstromverarbeitung [TPD+12, JM07, KD09] etabliert. Ziel der Verarbeitung ist es nicht nur den reinen Messwert, sondern die zugrunde liegende Unsicherheit innerhalb der Verarbeitung in einem Datenstrommanagementsystem zu repräsentieren und zu verarbeiten, so dass der entstehende kontinuierliche

Ausgabestrom einer Anfrage auch immer die aktuelle Ergebnisunsicherheit enthält. Bei der Verarbeitung von Unsicherheiten kann dabei zwischen zwei Klassen unterschieden werden, der Verarbeitung von diskreten Wahrscheinlichkeitsverteilungen und der Verarbeitung von kontinuierlichen Wahrscheinlichkeitsverteilungen. Diskrete Verteilungen werden häufig dazu genutzt die Existenzunsicherheit von möglichen Welten darzustellen. Kontinuierliche Wahrscheinlichkeitsverteilungen dienen dagegen dazu, Unsicherheiten in der Sensorwahrnehmung, welche etwa durch das Messverfahren an sich oder Umwelteinflüsse induziert werden, zu beschreiben. Im Folgenden liegt der Fokus daher auf der Bestimmung von kontinuierlichen stochastischen Modellen.

Die Bestimmung von stochastischen Modellen auf Basis von Datenströmen bei Filteroperationen wurde unter anderem in [ZCWQ03] behandelt. Hierbei war allerdings das Ziel, das stochastische Modell zu verwenden, um das Rauschen um einen Selektionsbereich innerhalb der Verarbeitung zu bestimmen. Ziel dieser Arbeit ist es aber das mehrdimensionale stochastische Modell der Daten selbst zu bestimmen, um eine probabilistische Verarbeitung der Daten, wie sie in [TPD+12] mit dem Mischtyp-Modell eingeführt wurde, zu ermöglichen. Das Modell hat den Vorteil, dass es sowohl die Unsicherheit über die Existenz einzelner Attribute, sowie auch die Unsicherheit über die Existenz ganzer Tupel repräsentieren kann. Zur Evaluation von verschiedenen Verfahren zur Bestimmung und Verarbeitung der mehrdimensionalen stochastischen Modelle wurde diese probabilistische Verarbeitung mit den Konzepten der deterministischen Verarbeitung mit Zeitintervallen aus [Krä07] kombiniert und in dem Datenstrommanagementsystem Odysseus [AGG+12] implementiert.

2 Verfahren zur Bestimmung von stochastischen Modellen

Für die Bestimmung von mehrdimensionalen stochastischen Modellen, wie sie bei der probabilitischen Datenstromverarbeitung verwendet werden, existieren prinzipiell mehrere Möglichkeiten. Zu diesen Verfahren zählen etwa das Erwartungswertmaximierungsverfahren und die Kerndichteschätzung, welche im Folgenden näher erläutert werden.

2.1 Erwartungsmaximierungsverfahren

Das Erwartungswertmaximierungsverfahren [DLR77] dient dazu die Parameter eines stochastischen Modells durch mehrere Iterationen an die Verteilung von Daten anzunähern. Hierzu wird versucht die Log-Likelihood L zwischen den zu bestimmenden Parametern und den zur Verfügung stehenden Daten in jeder Iteration t des Algorithmus zu maximieren. Als Parameter bieten sich hierfür die Parameter einer multivariaten Mischverteilung aus Gauß-Verteilungen mit Parameter $\theta = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{m}$ an. Eine multivariate Mischverteilung aus Gauß-Verteilungen über eine kontinuierliche Zufallsvariable X ist eine Menge

von m gewichteten Gauß-Verteilungen $X_1, X_2, \ldots, X_m$, wobei X_i die Wahrscheinlichkeitsdichtefunktion

$$f_X(x) = \sum_{i=1}^{m} w_i f_{X_i}(x) \text{ mit } f_{X_i}(x) = \frac{1}{(2\pi)^{k/2} \mid \Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}$$

ist. Dabei gilt, dass $0 \leq w_i \leq 1, \sum_{i=1}^{m} w_i = 1$, k die Größe des Zufallsvektors ist und jede Mischverteilungskomponente X_i eine k-variate Gauß-Verteilung mit Erwartungswert μ_i und Kovarianz-Matrix Σ_i ist. Zur Annäherung einer Gauß-Mischverteilung wird zunächst ein initiales stochastisches Modell mit m Gauß-Verteilungen bestimmt. Auf Basis des aktuellen Modells werden nun im E-Schritt die Erwartungswerte bestimmt, also die Wahrscheinlichkeiten, dass die aktuellen Werte aus dem aktuellen stochastischen Modell generiert wurden.

$$\tau_{ij}^{(t)} = \frac{w_j^{(t)} F_{Xj}(x_i; \theta_j^{(t)})}{\sum_{l=1}^{m} w_l^{(t)} F_{Xj}(x_i; \theta_l^{(t)})}, i = 1, \ldots, n, j = 1, \ldots, m$$

$$\gamma_j^{(t)} = \sum_{i=1}^{n} \tau_{ij}^{(t)}, j = 1, \ldots, m$$

Während des M-Schrittes werden die neuen Parameter für θ anhand der Ergebnisse aus dem E-Schritt bestimmt.

$$w_j^{(t+1)} = \frac{\gamma_j^{(t)}}{n}, j = 1, \ldots, m, \ \mu_j^{(t+1)} = \frac{1}{\gamma_j^{(t)}} \sum_{i=1}^{n} \tau_{ij}^{(t)}, j = 1, \ldots, m$$

$$\Sigma_j^{(t+1)} = \frac{1}{\gamma_j^{(t)}} \sum_{i=1}^{n} \tau_{ij}^{(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T, j = 1, \ldots, m$$

Nach jedem EM-Schritt wird die Log-Likelihood berechnet und mit einem gegebenen Schwellwert verglichen. Ist die Differenz kleiner als der gegebene Schwellwert oder überschreitet die Anzahl der Iterationen die maximale Anzahl, werden die bestimmten Parameter für die Gewichte (w), den Erwartungswert (μ), sowie die Kovarianz-Matrix (Σ) der Mischverteilung zurückgeliefert.

2.2 Kerndichteschätzung

Im Gegensatz zum EM-Verfahren wird bei der Kerndichteschätzung (KDE) für jeden Messwert eine Komponente in einer Mischverteilung erstellt und eine Bandbreite bestimmt. Die Bandbreite dient dazu eine Varianz/Kovarianz Matrix für alle Komponenten der Mischverteilung zu bilden und so das eigentliche zugrunde liegende Modell möglichst gut wieder zu geben. Zur Bestimmung der Bandbreite B haben sich mehrere Verfahren

etabliert, wie etwa die Scott-Regel [Sco92]. Die Parameter der Komponenten der Mischverteilung lassen sich somit wie folgt berechnen:

$$w_j = \frac{1}{n}, \mu_j = x_j, \Sigma_j = \Sigma(x) * B$$

Wobei $\Sigma(x)$ die Varianz/Kovarianz der zugrunde liegenden Daten repräsentiert. Man sieht bereits, dass das KDE ohne mehrmalige Iterationen über die zugrunde liegenden Daten auskommt, da sowohl der Erwartungswert wie auch die Varianz/Kovarianz inkrementell bestimmt werden können.

Da bei der Kerndichteschätzung die Anzahl an Komponenten der Mischverteilung linear zu der Zahl der Messwerte steigt und somit das Ergebnis generell ungeeignet ist für eine Verarbeitung in einem Datenstrommanagementsystem, benötigt es ein Verfahren zur Reduktion der Komponenten. In [ZCWQ03] stellen die Autoren ein Verfahren vor, welches das KDE-Verfahren auf einen eindimensionalen Strom anwendet und die dabei resultierende Mischverteilung durch ein Kompressionsverfahren auf eine geringere Anzahl von Verteilungen reduziert. Dieses Verfahren ist allerdings nicht für multivariate Verteilungen anwendbar. In [CHM12] wurde eine Selbstorganisierende Merkmalskarten (SOM) verwendet um Cluster zu bilden und diese Cluster durch eine Verteilung darzustellen. SOMs haben allerdings allgemein den Nachteil, dass die Gefahr einer Überanpassung der Gewichtsvektoren besteht. Eine weitere Möglichkeit zur Reduktion der Komponenten besteht in dem Bregman Hard Clustering Verfahren [BMDG05]. Bei dem Bregman Hard Clustering Verfahren wird versucht, ähnliche Verteilungen innerhalb einer Mischverteilung durch die Bildung von Clustern zu vereinfachen. Hierbei werden zunächst Cluster mit je einem Repräsentanten gebildet und anschließend für jedes Cluster eine Minimierung ausgeführt mit dem Ziel den Informationsverlust zwischen den Clusterzentren und den Komponenten zu minimieren. Das Verfahren kann als eine Generalisierung des Euklidischen k-Means Verfahrens angesehen werden, wobei die Kullback-Leibler Divergenz als Minimierungsziel verwendet wird. Um allerdings die Bestimmung des Integrals innerhalb der Kullback-Leibler Divergenz zu umgehen, wird die Kullback-Leibler Divergenz in eine Bregman Divergenz umgewandelt. Die Bregman Divergenz ist dabei definiert als:

$$D_F(\theta_j||\theta_i) = F(\theta_j) - F(\theta_i) - < \theta_j - \theta_i, \nabla F(\theta_i) > \quad (1)$$

Hierbei wird die Dichtefunktion einer Normalverteilung in die kanonische Dekomposition der jeweiligen Exponentialfamilie wie folgt umgeschrieben:

$$\mathcal{N}(x; \mu, \sigma^2) = exp\{< \theta, t(x) > -F(\theta) + C(x)\} \quad (2)$$

Wobei $\theta = (\theta_1 = \frac{\mu}{\sigma^2}, \theta_2 = -\frac{1}{2\sigma^2})$ die natürlichen Parameter, $t(x) = (x, x^2)$ die notwendige Statistik und $F(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\frac{-\pi}{\theta_2}$ die Log-Normalisierung für eine Normalverteilung darstellen.

Unter der Bedingung, dass beide Verteilungen von der gleichen Exponentialfamilie stammen, lässt sich die Kullback-Leibler Divergenz in die Bregman Divergenz umformen

$$KL(\mathcal{N}(x; \mu_i, \sigma_i^2)||\mathcal{N}(x; \mu_j, \sigma_j^2)) = D_F(\theta_j||\theta_i) \quad (3)$$

, so dass nun direkt die Bregman Divergenz als Distanz innerhalb des k-Means Verfahrens zur Clusterbildung angewendet werden kann.

3 Evaluation der Verfahren

Im Folgenden werden die Verfahren zur Bestimmung des stochastischen Modells der Daten eines Datenstrom hinsichtlich ihrer Latenz, aber auch hinsichtlich der Güte des stochastischen Modells evaluiert. Zu diesem Zweck wurden die Verfahren als Verarbeitungsoperatoren innerhalb des Odysseus DSMS realisiert. Die Evaluation wurde dabei sowohl auf synthetischen Daten, wie auch auf Daten aus einem Ultrabreitband-Positionierungssystem [WJKvC12] durchgeführt. Hierzu wurden 10.000 Messwerte aus einer Normalverteilung, sowie aus einer logarithmischen Normalverteilung generiert um einen Datenstrom aus Messwerten zu simulieren. Die Evaluation der Latenz und der Güte des Modells betrachtet dabei drei Szenarien mit Datenfenstern der Größe 10, 100 und 1000. Das Datenfenster definiert dabei die Anzahl an Messwerten auf denen die Operatoren das stochastische Modell bestimmen sollen. Die Güte des Modells betrachtet das aktuell bestimmte stochastische Modell im Hinblick auf alle 10.000 Datensätze. Als Qualitätskriterium wird hierzu das Akaike-Informationskriterium verwendet. Das AIC ist ein Maß für die relative Qualität eines stochastischen Modells für eine gegebene Datenmenge und ist definiert als:

$$AIC = 2k - 2\ln(L) \tag{4}$$

Der Parameter k repräsentiert hierbei die Anzahl der freien Parameter in dem stochastischen Modell und der Parameter L gibt die Log-Likelihood zwischen dem stochastischen Modell und der gegebenen Datenmenge wieder. Dieses Informationskriterium ist für die Evaluation deshalb gut geeignet, da es sowohl die Nähe der generierten Mischverteilung aus den drei Verfahren zu den tatsächlichen Daten bewertet und zudem die Anzahl der Komponenten innerhalb der Mischverteilungen in die Bewertung mit einfließen lässt. Die Nähe zu den tatsächlichen Daten ist wichtig für die Qualität der Verarbeitungsergebnisse und die Komponentenanzahl der Mischverteilung hat eine Auswirkung auf die Latenz der Verarbeitung, da jede Komponente innerhalb einer Mischverteilung bei Operationen wie der Selektion oder dem Verbund mit einem Selektionskriterium bei einer probabilistischen Verarbeitung integriert werden muss.

Um mögliche Ausreißer zu minimieren wurde jede Evaluation wurde dabei 10-mal wiederholt. Als Testsystem diente ein Lenovo Thinkpad X240 mit Intel Core i7 und 8GB RAM. Die verwendete Java Laufzeitumgebung war ein OpenJDK Runtime Environment (IcedTea 2.5.2) (7u65-2.5.2-2) mit einer OpenJDK 64-Bit Server VM (build 24.65-b04, mixed mode). Bei dem Betriebssystem handelte es sich um ein Debian GNU/Linux mit einem 3.14 Kernel.

Das EM-Verfahren versucht ein stochastisches Modell an die eingehenden Daten anzupassen. Dabei spielen neben der Datenfenstergröße die Anzahl der Iterationen, der Konvergenzschwellwert für die Veränderung der Log-Likelihood in jeder Iteration, sowie die Komponentenanzahl der Mischverteilungen eine Rolle für die Latenz dieses Operators. Für die Evaluation wurde der Konvergenzschwellwert auf 0.001 gesetzt, die Anzahl an Iterationen auf 30 und die Zahl der Komponenten auf 2. Die gleiche Anzahl an Iterationen wird ebenfalls in der von V. Garcia bereitgestellten Java Bibliothek jMEF[1] verwendet.

[1] http://vincentfpgarcia.github.io/jMEF/

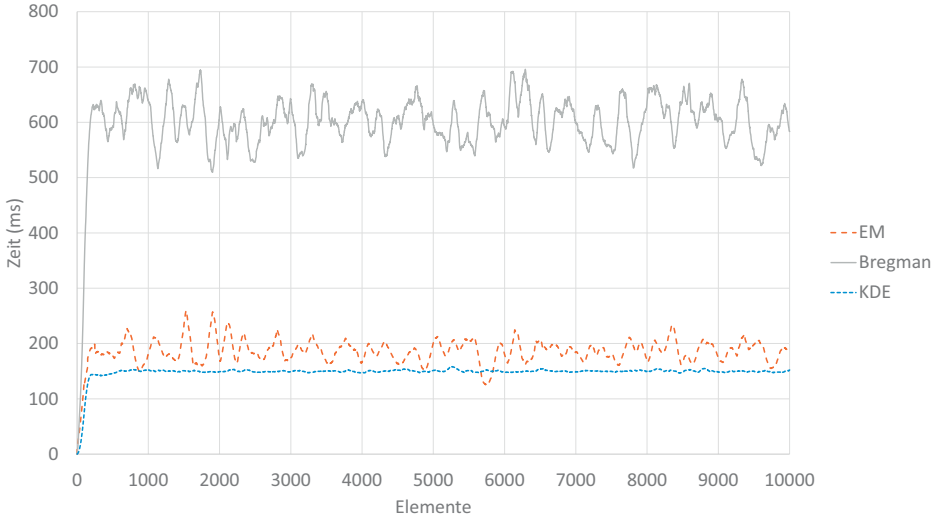


Abbildung 1: Latenz der Operatoren bei einem Datenfenster der Größe 100 für Daten aus einer logarithmischen Normalverteilung

Das KDE-Verfahren bestimmt für jeden Datenwert eine eigene Komponente in der resultierenden Mischverteilung. Der entwickelte Operator verwendet hierzu die Scott-Regel zur Bestimmung der Bandbreite der Kovarianzmatrix der Komponenten. Das Bregman Hard Clustering, welches in einem weiteren Schritt verwendet wird um die Anzahl an Komponenten auf die gewünschte Zahl zu minimieren wurde mit einer maximalen Anzahl von 30 Iterationen konfiguriert. Um die Resultate vergleichbar zu halten wurde der Operator so konfiguriert, dass er ebenfalls eine 2-komponentige Mischverteilung ermittelt, also zwei Cluster bildet.

Der hier verwendete Konvergenzschwellwert für das Erwartungswertmaximierungsverfahren liegt oberhalb des, in der verwendeten Apache Commons Math3 Bibliothek[2] als Standardwert festgelegten, Wertes von 0.00001, da sich in den Versuchen zeigte, dass bereits ein höherer Konvergenzschwellwert ausreichte um die Verfahren hinsichtlich der Güte des stochastischen Modells und der gemessenen Latenz miteinander zu vergleichen.

3.1 Synthetische Sensordaten

Das Latenzverhalten der einzelnen Verfahren ist in Abb. 1 für Daten aus einer log. Normalverteilung für ein Datenfenster der Größe 100 dargestellt. Das EM-Verfahren weist hierbei eine ähnliche und stabile Latenz von durchschnittlich ca. 200 Millisekunden auf. Dies ist durch die mehrmalige Iteration über die aktuell gültigen Daten zur Bestimmung der

[2] http://commons.apache.org/proper/commons-math

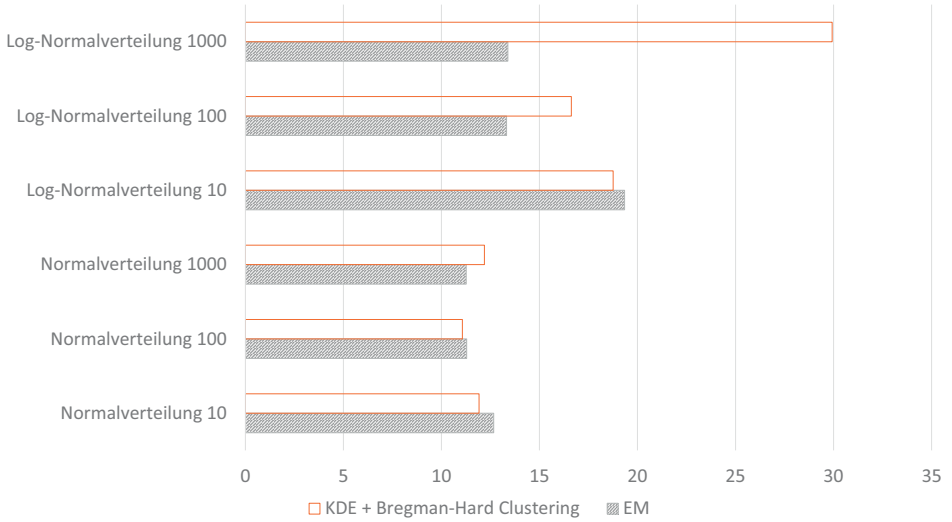


Abbildung 2: Vergleich des AIC zwischen EM-Verfahren und KDE mit Bregman Hard Clustering bei unterschiedlichen Datensatzfenstergrößen für Werte aus einer Normalverteilung und einer logarithmischen Normalverteilung

Log-Likelihood zwischen dem jeweils temporären stochastischen Modell und den Daten geschuldet. Im Gegensatz zum EM-Verfahren kann das Band bei der Kerndichteschätzung kontinuierlich bestimmt werden. Allerdings fällt auf, dass trotz mehrmaliger Wiederholung der Messung das Verfahren zum Bregman Hard Clustering eine deutlich höhere Latenz aufweist. Dieses Verhalten ist dabei unabhängig von der Art der Verteilung. Dies ist vor allem auf die Tatsache zurück zu führen, dass das Bregman Hard Clustering Verfahren in jeder Iteration die Bregman Divergenz zwischen den Clusterzentren und den einzelnen Komponenten bestimmen muss und zusätzlich noch den Zentroiden aus jedem Cluster in jeder Iteration neu ermitteln muss. Beim Vergleich der durchschnittlichen Latenz bei unterschiedlichen Größen von Datenfenstern zeigt sich, dass die Latenz des EM-Verfahrens konstant bleibt, während die Latenz des Bregman Hard Clusterings stark ansteigt. Bei der Qualitätsbetrachtung des ermittelten stochastischen Modells fällt auf, dass die Qualität des EM-Verfahrens im Sinne des AIC bei Werten aus einer logarithmischen Normalverteilung deutlich besser abschneidet als das KDE-Verfahren in Kombination mit dem Bregman Hard Clustering. Bei Werten aus einer Normalverteilung dagegen unterscheidet sich der AIC-Wert nur geringfügig bei den beiden Verfahren. Ein gleiches Verhalten lässt sich auch bei Datensatzfenstern der Größe 1.000 beobachten.

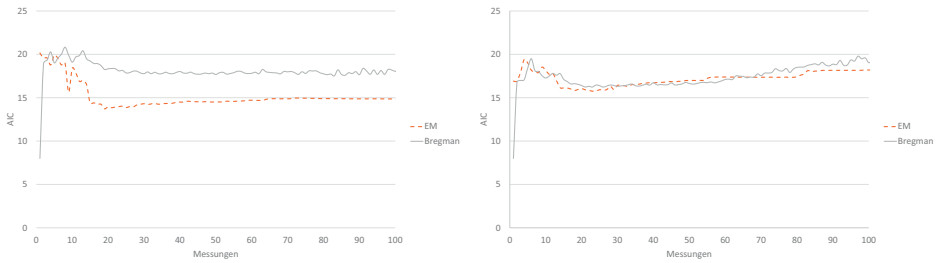
Ist allerdings die Anzahl an Datensätzen gering, ändert sich dieses Verhalten. Bei einem Datensatzfenster der Größe 10 zeigt sich unabhängig von dem zugrunde liegenden stochastischen Modell der Daten, dass die Kombination aus KDE und Bregman Hard Clustering das bessere Modell liefert. Zudem unterscheiden sich die Latenzen bei dieser Datenmenge zwischen den beiden Verfahren nur gering.

Abbildung 3: Messwerte der Positionsbestimmung für die Positionen 1–8

3.2 Reale Sensordaten

Um zu zeigen, dass die Verfahren auch stochastische Modelle von echten Sensordaten erstellen können, wurden die Operatoren auf Sensordatenaufzeichnungen eines Ultrabreitband-Positionierungssystem [WJKvC12] angewendet. Insgesamt wurden 8 Positionen (vgl. Abbildung 3) bestimmt, von denen im Folgenden die Positionen 6 und 7 als repräsentative Positionen näher betrachtet werden. Hierbei wurde das stochastische Modell jeder Position mit dem EM-Verfahren und der Kombination aus KDE und Bregman Hard Clustering auf einem Datensatzfenster der Größe 10 und einem Datensatzfenster der Größe 100 bestimmt.

Bei der Betrachtung der zeitlichen Bestimmung des stochastischen Modells in Abb. 4 fallen zunächst für die Position 6 anfängliche Ausreißer bei der Nähe zum Modell auf. Dies deutet auf eine anfängliche Anpassung der Positionierungsknoten der Anwendung hin. In den darauf folgenden Messungen bleiben sowohl die Modellqualität des EM-Verfahrens, wie auch das resultierende Modell des Bregman Hard Clustering stabil. Wie bereits bei den synthetischen Daten ist auch bei realen Sensordaten das Phänomen erkennbar, dass die Kombination aus KDE mit Bregman Hard Clustering bei kleinen Datensatzfenstern im Vergleich zum EM-Verfahren bessere stochastische Modelle ermittelt. Dagegen ist bei größeren Datensatzfenstern das EM-Verfahren besser geeignet um gute stochastische Modell im Sinne des AIC zu bestimmen.



(a) Position 6

(b) Position 7

Abbildung 4: Qualität des stochastischen Modells über die Zeit bei einem Datensatzfenster der Größe 100 von Position 6 und 7

4 Zusammenfassung und Ausblick

In dieser Arbeit wurden Verfahren zur kontinuierlichen Bestimmung des zugrunde liegenden mehrdimensionalen stochastischen Modells von Messwerten aus aktiven Datenquellen vorgestellt. Ziel ist es, diese mehrdimensionalen stochastischen Modelle in einem probabilistischen Datenstrommanagementsystem zu verarbeiten. Bei den Verfahren handelt es sich um das Erwartungsmaximierungsverfahren und die Kerndichteschätzung in Kombinationen mit dem Bregman Hard Clustering Ansatz. Zunächst wurden die Grundlagen der jeweiligen Verfahren aufgezeigt. Zur Repräsentation der Unsicherheiten wurde das in [Krä07] entwickelte Modell durch das Mischtyp Modell [TPD+12] erweitert und in dem Odysseus DSMS realisiert. Bei der Evaluation der Verfahren wurde zunächst auf Basis von synthetischen Daten die Latenz der einzelnen Verfahren ermittelt. Hierbei zeigte sich, dass die Kombination aus Kerndichteschätzung und Bregman Hard Clustering aufgrund der mehrmaligen Iterationen über die Komponenten einer Mischverteilung eine wesentlich höhere Latenz als das Erwartungsmaximierungsverfahren aufweist. Zudem sind die resultierenden stochastischen Modelle im Sinne des Akaike Informationskriterium in den meisten Fällen schlechter als die angenäherten Modelle des Erwartungsmaximierungsverfahrens. Aus Sicht der Latenzoptimierung und angesichts der Qualität der bestimmten Modelle sollte daher das Erwartungsmaximierungsverfahren bei der Datenstromverarbeitung bevorzugen werden. Einzige Ausnahme sind Anwendungen in denen nur geringe Mengen an Daten zur Verfügung stehen. Hier konnte die Kombination aus Kerndichteschätzung und Bregman Hard Clustering die besseren stochastischen Modelle bestimmen. Eine Evaluation auf Basis von Sensoraufzeichnungen von Ultrabreitband-Lokalisierungssensoren bestätigten die Resultate aus der Evaluation mit synthetischen Daten.

Danksagung

Die Autoren möchten Herrn Prof. Huibiao Zhu von der East China Normal University für seine Unterstützung danken. Diese Arbeit wurde durch die Deutsche Forschungsgesellschaft im Rahmen des Graduiertenkollegs (DFG GRK 1765) SCARE (www.scare.uni-oldenburg.de) gefördert.

Literatur

[AGG+12] H.-J. Appelrath, Dennis Geesen, Marco Grawunder, Timo Michelsen und Daniela Nicklas. Odysseus: a highly customizable framework for creating efficient event stream management systems. In *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, DEBS '12, Seiten 367–368, New York, NY, USA, 2012. ACM Press.

[BMDG05] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon und Joydeep Ghosh. Clustering with Bregman Divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[CHM12] Yuan Cao, Haibo He und Hong Man. SOMKE: Kernel density estimation over data streams by sequences of self-organizing maps. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1254–1268, 2012.

[DLR77] Arthur P. Dempster, Nan M. Laird und Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977.

[JM07] T. S. Jayram und S. Muthukrishan. Estimating statistical aggregates on probabilistic data streams. In *In ACM Symposium on Principles of Database Systems*, Seiten 243–252, New York, NY, USA, 2007. ACM Press.

[KD09] Bhargav Kanagal und Amol Deshpande. Efficient query evaluation over temporally correlated probabilistic streams. In *International Conference on Data Engineering*, 2009.

[Krä07] Jürgen Krämer. *Continuous Queries over Data Streams-Semantics and Implementation*. Dissertation, Philipps-Universität Marburg, 2007.

[Sco92] D.W Scott. Multivariate Density Estimation: Theory, Practice, and Visualization. 1992.

[TPD+12] Thanh T. L. Tran, Liping Peng, Yanlei Diao, Andrew McGregor und Anna Liu. CLARO: modeling and processing uncertain data streams. *The VLDB Journal*, 21(5):651–676, Oktober 2012.

[WJKvC12] Thorsten Wehs, Manuel Janssen, Carsten Koch und Gerd von Cölln. System architecture for data communication and localization under harsh environmental conditions in maritime automation. In *Proceedings of the 10th IEEE International Conference on Industrial Informatics (INDIN)*, Seiten 1252–1257, Los Alamitos, CA, USA, 2012. IEEE Computer Society.

[ZCWQ03] Aoying Zhou, Zhiyuan Cai, Li Wei und Weining Qian. M-kernel merging: Towards density estimation over data streams. In *8th International Conference on Database Systems for Advanced Applications*, Seiten 285–292. IEEE, 2003.