POSTER ABSTRACT

Wavelet Density Estimators over Data Streams *

Christoph Heinz
Department of Computer Science
Philipps University Marburg
heinzch@mathematik.uni-marburg.de

Bernhard Seeger
Department of Computer Science
Philipps University Marburg
seeger@mathematik.uni-marburg.de

ABSTRACT

Density estimation is a building block of many data analysis techniques. A recently examined approach based on wavelets promises to be superior to traditional density estimation techniques. For possibly infinite data streams, however, this approach is not feasible due to the limited resources, e.g. memory. In this paper, we propose a new technique for computing wavelet density estimators over data streams that only requires a fixed amount of memory. Our estimators are updated in an online manner such that a continuous analysis of data streams is supported during runtime.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Nonparametric statistics

Keywords

Data Streams, Density Estimation, Wavelets

1. INTRODUCTION

Many applications require an immediate processing and analysis of transient data streams [1]. The development of appropriate analysis algorithms for data streams is of utmost importance since their rigid processing requirements prevent the usage of traditional analysis and mining techniques. A suitable approach to data streams requires an online analysis with a fixed amount of memory [3]. In this paper, we address the problem of estimating the probability density function of a data stream while satisfying the processing requirements of data streams [3]. Density estimation is an important building block for data analysis since it provides knowledge about the entire data distribution. A popular class of density estimates for continuous distributions is based on kernel functions [4]. Different estimates, which are built on top of wavelets, can achieve optimal error rates for a wide range of smooth densities [5]. In this paper, we consider the problem of computing wavelet density estimators over data streams. Our technique is derived from a framework for maintaining nonparametric estimators over data streams [2]. This framework requires the specification of two functions, where the first merges two independent estimators and the second allows to dynamically adapt the allocated memory of an estimator.

*This work has been supported by the German Research Society (DFG) under grant no. SE 553/4-1.

SAC'05 March 13-17, 2005, Santa Fe, New Mexico, USA

2. DATA STREAMS AND STATISTICS

We consider a data stream as an unbounded sequence of numerical elements $X_1, ..., X_n$ that are assumed to be an *iid*-sample of a continuous random variable X for each $n, n \geq 1$. The random variable X in turn is completely characterized by its probability density function f. In general, f is unknown for an arbitrary data stream. A density estimate $\hat{f}$ of f can be computed by using kernel-based density estimates [4]. These estimates are accurate under certain assumptions, e.g. a sufficiently smooth f. Recent studies revealed that wavelet-based density estimates promise to be superior due to the local nature of wavelets [5].

Wavelets form an orthogonal basis of a function vector space. The **discrete wavelet transform (DWT)** separates a function into two parts, where one part covers the general shape and the other represents local features. This allows compressing a function efficiently by simply dropping the least important local details. The **wavelet series expansion** of a function f is a sum of basis functions

$$f(x) = \sum_{k \in \mathbb{Z}} c_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k}(x) \qquad (1)$$

with appropriate basis coefficients $c_{j_0,k}, d_{j,k}$. The scaling functions $\phi_{j_0,k}$ create the general shape of f, while the wavelets $\psi_{j,k}(x)$ provide the local details.

The wavelet series expansion of f is the point of origin for a **wavelet density estimator (WDE)** [5]. Since f is unknown, the sample $X_1, ..., X_n$ is used for estimating the coefficients of the wavelet series expansion. Since n increases continuously over time, the resulting number of coefficients of a WDE may be infinite. It is neither theoretically nor practically reasonable to keep all of them due to convergence properties and memory constraints. Thus, the wavelet series expansion of $\hat{f}$ has to be appropriately truncated. It is crucial to the quality of a WDE which coefficients are

kept. Different strategies exist for a finite selection including an appropriate choice for the initial resolution j. Two commonly used WDEs are the linear WDE which is well-suited for smooth densities and the thresholded WDE with its proper ability to identify local features.

3. WAVELET DENSITY ESTIMATORS OVER DATA STREAMS

A first approach to create a WDE over a data stream would be to update the coefficients for each arriving element. Unfortunately, their computation employs the parameter j that in turn depends on the number of elements. Moreover, the recomputation of the coefficients would require to access all previous elements. Our approach to computing a WDE over a data stream is derived from our framework for maintaining nonparametric estimators over data streams [2]. The basic idea of this framework is to partition the data stream into contiguous blocks and to process the data stream block by block. For each block B_i, a separate estimator $\hat{f}_i$ is computed. The convex linear combination of the first m block estimators $\hat{g}(x) = \sum_{i=1}^{m} \omega_i \hat{f}_i(x)$ with $0 \leq \omega_i \leq 1$, $i = 1, ..., m$ returns our m^{th} **overall estimator**. This allows an online computation of the overall estimators

$$\hat{g}_i(x) = \begin{cases} \hat{f}_1(x), & i = 1 \\ (1 - \tilde{\omega}_i)\hat{g}_{i-1}(x) + \tilde{\omega}_i \hat{f}_i(x), & i \geq 2 \end{cases} \quad (2)$$

where the weighting sequence $\tilde{\omega}_i$ has to be set appropriately.

In order to obtain a specific nonparametric estimator, two functions have to be provided. The first performs a convex merging of the current overall estimator and a new block estimator according to equation (2). The second function compresses the estimator being returned by the merge step such that it fits into the available main memory.

Let us discuss the implementation of these two functions for WDEs in our approach. For the convex merging of the current overall estimator and the new block WDE, we employ the wavelet series expansion of both functions. The convex linear combination of both expansions is reduced to the convex linear combination of the coefficients of their associated basis functions. Thus, the resulting coefficients $c_{j,k}^{new}$ of the new overall estimator are weighted sums, i.e. $c_{j,k}^{new} = (1 - \tilde{\omega}_i)c_{j,k}^1 + \tilde{\omega}_i c_{j,k}^2$ where $c_{j,k}^1$ and $c_{j,k}^2$ are coefficients of the current overall estimator and the new block WDE, respectively. Note that the new overall estimator $\hat{g}$ is again represented by its wavelet series expansion.

While this allows an efficient online computation of a WDE over a data stream, we still face the problem that the convex merge may result in a new $\hat{g}$ whose number of coefficients exceeds the available memory. If a memory overflow occurs, we remove those coefficients of the wavelet series expansion that represent the least important local details. In case of a system where resources like memory are shared among different users, it might occur that the estimator receives a larger fraction of memory that allows improving its accuracy. The additional memory is then shared by the coefficients of the overall estimator and the next data block. First, we try to keep as many coefficients as possible. Then, if memory is still available, we assign it to the next data block. This is beneficial since the larger data blocks are, the more accurate the corresponding block estimators are.

4. EXPERIMENTAL RESULTS

In our experiments, we examined how the quality of our WDEs over data streams depends on the number of processed elements. We used two synthetic data sets of 100k elements drawn from the Claw density and the CP2 density [2], respectively. The WDEs over these streams used data blocks of 500 elements and a maximum of 100 coefficients for the overall estimator. We conducted experiments with the linear WDE and the thresholded WDE, where both of them used Daubechies5 wavelets. In order to compare the efficiency of our approach with other techniques, we also considered kernel density estimates over data streams as proposed in [2]. The accuracy of each technique was measured during runtime in terms of the mean squared error (MSE). All techniques provided consistent estimates as the MSE decreased for an increasing number of elements. The MSE first decreased rapidly and then slew down, while approaching to a constant. We also considered traditional offline WDEs obtained from the entire data set. Our stream-based WDEs performed very close to their offline counterparts. For the Claw density, the thresholded WDE significantly outperformed both, the linear WDE and the kernel estimator. The latter estimation techniques generated very coarse approximations where important details of the Claw density are not covered. For the CP2 density, the linear WDE was slightly superior to the thresholded WDE and the kernel-based estimate. The thresholded WDE consisted of many artificial peaks that did not occur in the smooth parts of the CP2 density. For both data distributions, the WDEs required only a very small number of coefficients, e.g. about 30 for the linear WDE on Claw data.

5. CONCLUSIONS

In this work, we presented our approach to computing wavelet density estimators under the rigid requirements of the data stream model. The estimators are derived from our framework for maintaining nonparametric estimators over data streams. Our technique generates separate wavelet density estimators for contiguous blocks of stream elements and successively merges these blocks into an estimator while consuming only a constant amount of memory. The results of our experimental performance study validates the feasibility of our wavelet-based estimators. In future work, we will consider wavelet density estimators as underlying technology for advanced analysis tasks on data streams.

6. REFERENCES

[1] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and Issues in Data Stream Systems. In *Symp. on Principles of Database Systems*, 2002.

[2] B. Blohsfeld, C. Heinz, and B. Seeger. Maintaining Nonparametric Estimators over Data Streams. Technical Report No. 39, Philipps-University Marburg, 2004.

[3] P. Domingos and G. Hulten. A general framework for mining massive data streams. *Journal of Computational and Graphical Statistics*, 2003.

[4] D. W. Scott. *Multivariate Density Estimation. Theory, Practice, and Visualization.* John Wiley & Sons, New York, 1992.

[5] B. Vidakovic. *Statistical Modeling by Wavelets.* John Wiley & Sons, New York, 1999.