

Networked Statistical Inference: Cost-Performance Tradeoff

Animashree Anandkumar

Electrical & Computer Engr.,
Cornell University,
Ithaca, NY 14853, USA.
aa332@cornell.edu

ABSTRACT

The problem of minimum cost in-network fusion of measurements, collected from distributed sensors via multihop routing is considered. A designated fusion center collects all the sensor measurements and performs a statistical-inference test on the measurements, which are spatially correlated according to a Markov random field. Conditioned on the delivery of a sufficient statistic to the fusion center, the fusion scheme minimizing the total routing costs is shown to be a Steiner tree on an expanded graph. This Steiner-tree reduction preserves the approximation ratio, which implies that any Steiner-tree approximation can be employed for minimum cost fusion with the same approximation ratio. The problem is then extended to finding a fusion scheme that achieves optimal tradeoff between the total routing costs and the resulting end inference performance at the fusion center. It is shown that optimal linear tradeoff in a large network has a prize-collecting Steiner tree reduction with the approximation factor preserved. The performance of heuristics for minimum cost fusion are evaluated through theory and simulations, showing a significant saving in routing costs, when compared to routing all the raw measurements to the fusion center.

Categories and Subject Descriptors

H.1.1 [Systems and Information Theory]: Information theory

General Terms

Performance, Theory

Keywords

Detection and Estimation, Performance Evaluation, Wireless Sensor Networks, Statistical Analysis, Routing.

1. INTRODUCTION

In the distributed statistical inference setup, there are many sensors that measure a signal field and make certain local decisions. These decisions are then transported to a designated node, known as the fusion center, in an error-free manner which makes the final (global) decision on the underlying hypothesis. The classical approach to statistical inference considers only two design issues, viz the design of local-decision rule at that sensors and the design of global-decision at the fusion center. The vast array of literature on distributed inference is mostly limited to the case when the measurements are independent conditioned on the underlying hypothesis. See [5, 6].

Such an approach to the inference problem does not deal with the issue of routing the sensor measurements to the fusion center, usually achieved through a generic layered architecture [4]. Here, the design of routing is independent of the actual application implying that application-specific intermediate (also known as in-network) data-processing is not possible. Although this approach simplifies design, it may be inefficient for wireless sensor network facing severe resource constraints. Moreover, since a sensor network is usually deployed for a particular application, an application-specific design is justified in this case. Hence, we pursue a data-centric cross-layer design for a statistical-inference application.

We consider schemes that not only achieve savings in routing costs through in-network processing, but also have a guaranteed end-detection performance. We achieve this through the use of a *sufficient statistic* for inference which can result in reduction of data dimensionality without any loss in the inference performance. Hence, through the in-network processing of the sufficient statistic, routing costs can be reduced without sacrificing the end performance. The maximum reduction is obtained when we use the minimal sufficient statistic. The extent of such data reduction is, of course, affected by the underlying statistical model of the sensor measurements.

The presence of correlation among the sensor measurements crucially affects the design of fusion scheme through the minimal sufficient statistic. We use a *Markov random field* model which incorporates correlation in terms of a graph, known as the *dependency graph*, and leads to a highly structured minimal sufficient statistic. This enables us to obtain a tractable solution to the optimal fusion problem. Also interestingly, such a scheme consists of a sequence of simple operations, viz., data forwarding, and sum-function computation.

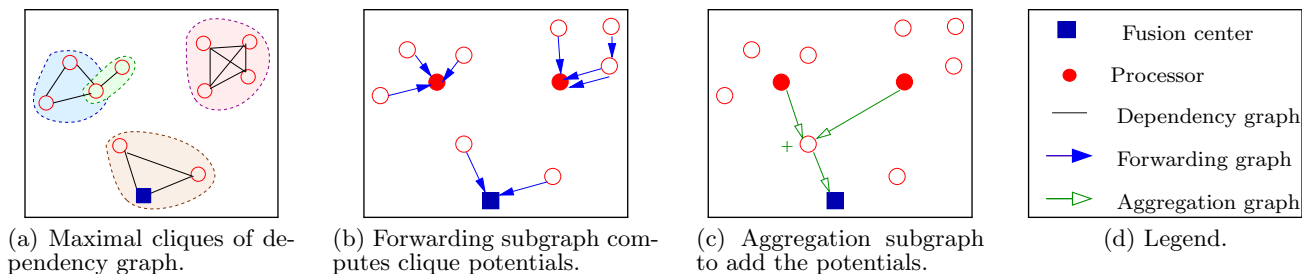


Figure 1: Schematic of dependency graph of Markov random field and stages of data aggregation. The cliques of Markov random field arise due to spatial dependence of data. The forwarding and aggregation subgraphs consist of links transporting raw data and aggregated values. The delivery of likelihood ratio to the fusion center needs to be ensured.

2. OUR APPROACH AND CONTRIBUTIONS

The main contribution of this work is three fold. Details can be found in [1–3]. First, conditioned on the requirement that the sufficient statistic for statistical inference is delivered to the fusion center, we obtain the minimum cost routing and aggregation scheme, when sensor measurements are drawn from a Markov random field (MRF). We exploit the fact that the likelihood ratio is the minimal sufficient statistic for inference and hence, maximum reduction in data is achieved by in-network processing of the likelihood ratio. For a MRF, the likelihood function consists of components, each depending on a subset of the measurements and these components can be computed independently at various nodes. Therefore, a fusion scheme involves the following considerations, viz., each component is assigned a computation site or a processor; measurements of the component members are then transported to its processor to enable computation of the component values. These values are then combined and delivered to the fusion center. See Fig.1. Hence, a fusion scheme consists of a sequence of three simple operations, viz., data forwarding, processing each component of the likelihood ratio and computing a sum function.

Second, we show that the Steiner tree on an expanded communication graph minimizes the sum costs of routing for the above tasks. The specific Steiner-tree reduction preserves the approximation factor. Hence, our approximation-factor preserving reduction implies that any Steiner-tree approximation algorithm can be used for the problem of optimal fusion with the same approximation ratio. The expansion of the communication graph involves adding component-representative nodes, as selectors of the processors for each component of the likelihood function and connecting them to the component members through edges incorporating the local routing costs.

In contrast to the Steiner-tree approach, we propose a simpler heuristic based on the minimum spanning tree (MST) that ensures optimal inference at the fusion center and has an approximation ratio of two for the special case of the nearest-neighbor dependency graph. Our simulations show that in-network processing achieves significant savings compared to forwarding all the raw data to the fusion center, especially for sparse spatial dependencies.

Thirdly, we consider the problem of selection and fusion of a subset of measurements such that optimal tradeoff between the total routing cost and the resulting end-inference

performance is achieved. We use the *Kullback-Leibler distance* over each component of the MRF as a decentralized performance measure. For a special case of MRF and node selection, we show that the fusion scheme achieving optimal tradeoff is a prize-collecting Steiner tree (PCST) and preserves the approximation-factor.

The substantial reduction in the routing costs comes from the exploitation of the Markovian correlation structure, the use of which is both a contribution and a limitation. To the best of our knowledge, there has been no study of strategies that guarantee optimal statistical inference at the fusion center, while minimizing multihop routing costs. We are able to address this fundamental problem analytically by exploiting the Markov random field structure. On the other hand, the assumed structure raises the practical issues of accuracy and overhead in learning the dependency structure. Within the limitations of our model-based assumptions, we hope to provide insights applicable to more general structures.

Acknowledgement

The author thanks her advisor Prof. L. Tong, and her collaborators Dr. A. Swami and Prof. A. Ephremides, for their research inputs to the thesis.

3. REFERENCES

- [1] A. Anandkumar, A. Ephremides, L. Tong, and A. Swami. Minimum cost routing with local processing for distributed statistical inference. In S. Haykin and R. Liu, editors, *Handbook on Array Processing and Sensor Networks*, chapter 25. IEEE-Wiley, 2008.
- [2] A. Anandkumar, L. Tong, A. Swami, and A. Ephremides. Cost-Performance Tradeoff in Multi-hop Aggregation for Statistical Inference. In *Proc. of IEEE International Symposium on Information Theory*, Toronto, Canada, July 2008.
- [3] A. Anandkumar, L. Tong, A. Swami, and A. Ephremides. Minimum Cost Data Aggregation with Localized Processing for Statistical Inference. In *Proc. of IEEE INFOCOM*, Phoenix, USA, April 2008.
- [4] D. P. Bertsekas and R. Gallager. *Data Networks*. Prentice Hall, 1992.
- [5] P. K. Varshney. *Distributed Detection and Data Fusion*. Springer, New York, NY, 1997.
- [6] R. Viswanathan and P.K.Varshney. Distributed Detection with Multiple Sensors: Part I-Fundamentals. *Proceedings of the IEEE*, 85(1):54–63, Jan. 1997.