NASA SBIR 2011 Phase I Solicitation

A1.17 Data Mining and Knowledge Discovery

Lead Center: ARC

The fulfillment of the SSAT project's goal requires the ability to transform the vast amount of data produced by the aircraft and associated systems and people into actionable knowledge that will aid in detection, causal analysis, and prediction at levels ranging from the aircraft-level, to the fleet-level, and ultimately to the level of the national airspace. The vastness of this data means that data mining methods must be efficient and scalable so that they can return results quickly. Additionally, much of this data will be distributed among multiple systems. Data mining methods that can operate on the distributed data directly are critical because centralizing large volumes of data is typically impractical. However, these methods must be provably able to return the same results as what a comparable method would return if the data could be centralized because this is a critical part of verifying and validating these algorithms, which is important for aviation safety applications. Additionally, algorithms that can learn in an online fashion---can learn from new data in incremental fashion without having to re-learn from the old data---will be important to allow deployed algorithms to update themselves as the national airspace evolves. The data is also heterogeneous: it consists of text data (e.g., aviation safety reports), discrete sequences (e.g., pilot switches, phases of flight), continuous time-series data (e.g., flight-recorded data), radar track data, and others. Data mining methods that can operate on such diverse data are needed because no one data source is likely to be sufficient for anomaly detection, causal analysis, and prediction.

This topic will yield efficient and scalable data-driven algorithms for anomaly detection, causal analysis, and prediction that are able to operate at levels ranging from the aircraft level to the fleet level. To that end, the methods must be able to efficiently learn from vast historical time-series datasets (at least 10 TB) that are heterogeneous (contain continuous, discrete, and/or text data). Distributed data-driven algorithms that provably return the same results as a comparable method that requires data to be centralized are also of great interest. Online algorithms that can update their models in incremental fashion are also of great interest for this subtopic.