Data stream clustering and classification

Hai-Long Nguyen, Yew-Kwong Woon, Wee Keong Ng:
A survey on data stream clustering and classification. Knowl. Inf. Syst. 45(3): 535-569 (2015)

Data stream applications

- Mining query streams
  - Learning to cluster web search results
  - Semantic similarity between search engine queries using temporal correlation
- Network monitoring
  - detect and prevent malicious attacks in a large Internet service provider network
  - E.g., classify in real time different kinds of attacks, such as
    - denial-of-service (DOS), unauthorized access from a remote machine (R2L), unauthorized access to local super-user privileges (U2R), surveillance and other probing attacks.
- Sensor networks
  - patient monitoring system to improve healthcare quality and staff productivity
- Social network streams
  - stream clustering methods are used to detect communities and monitor their evolution in social networks

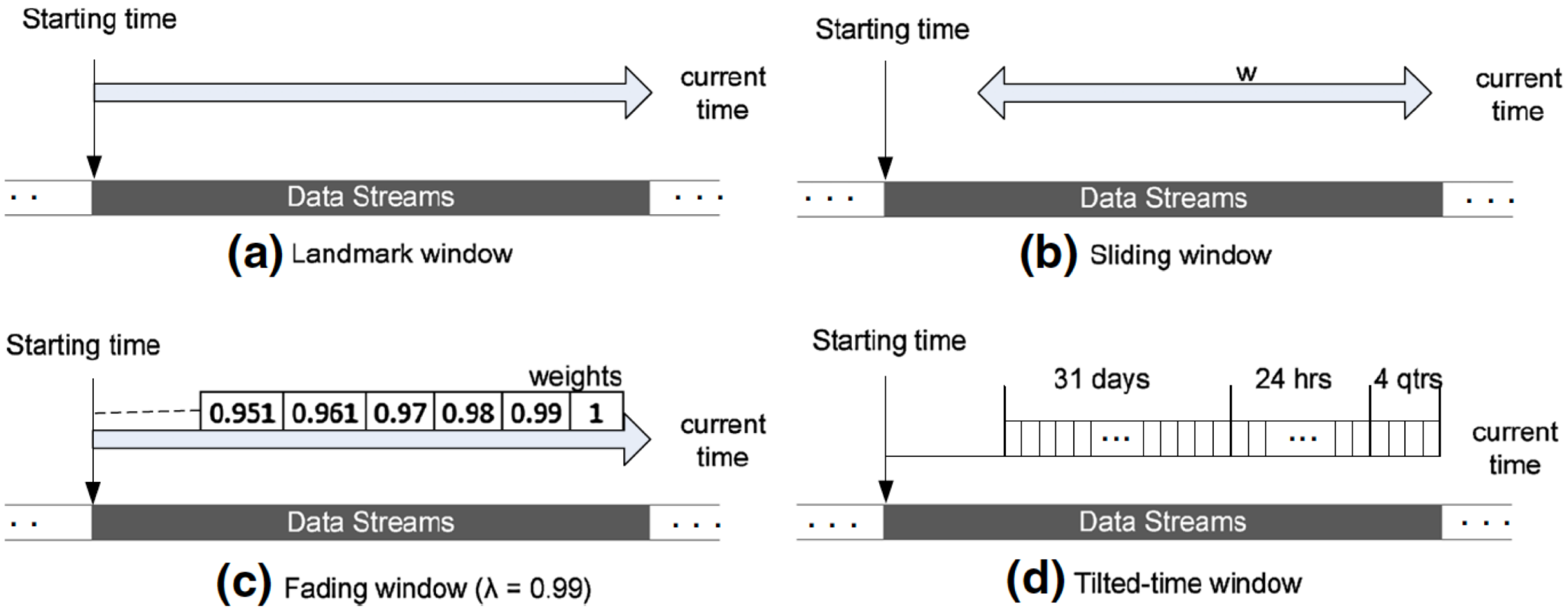
Mining constraints

Characteristics/domains	Traditional data mining	Data stream mining
Number of passes	Multiple	Single
Time	Unlimited	Real-time
Memory	Unlimited	Bounded
Number of concepts	One	Multiple
Result	Accurate	Approximate

Time windows

- Landmark window
  - Using the landmark window, all transactions in the window are equally important; there is no difference between past and present data
- Sliding window
  - We are only interested in the *w* most recent transactions; the others are eliminated
- Fading window
  - Assigns a different weight according to its arrival time so that new transactions receive higher weights than old ones
- Tilted time window
  - Between the fading window and sliding window variants

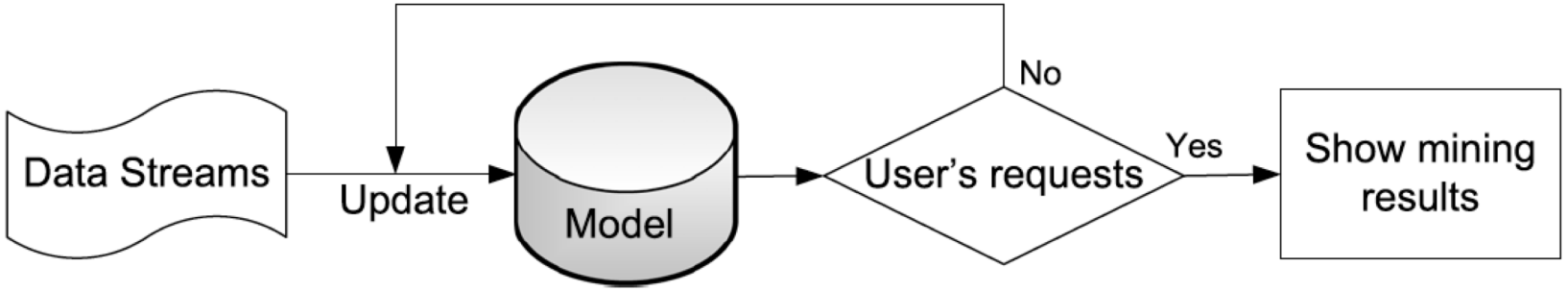
Time windows



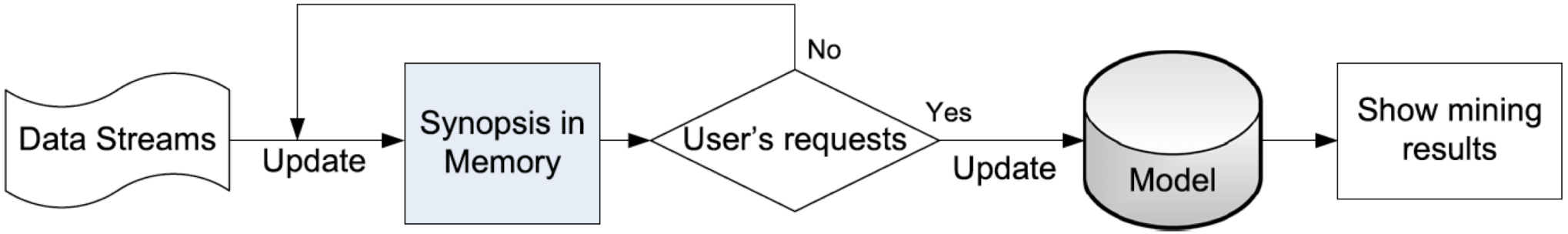
Computational approaches

- Incremental learning
  - The model incrementally evolves to adapt to changes in incoming data
- Two-phase Learning
  - To divide the mining process into two phases
    - Online phase: a synopsis of data is updated in a real-time manner
    - Offline phase: the mining process is performed on the stored synopsis whenever a user sends a request
  - Also known as online–offline learning

Computational approaches



(a) Incremental Learning



(b) Two-phase Learning