Section 3: Week 8: Data Mining IoT

Nate Bachmeier

TIM-8130: Data Mining

March 28, 2020

North Central University

Data Mining IoT

With ubiquitous access to high-bandwidth wireless networking and cloud computing, physical devices are evolving business capabilities to both monitor and react to changes across their supply chains and operational footprint. These devices produce enormous volumes of unstructured telemetry that require curation processes to transform raw data into business intelligence, enabling data-driven decisions that make the organization more competitive. During this transformation, data mining strategies extract patterns and statistical inferences through regression, clustering, classification, and rule association algorithms (Barua & Mondal, 2019). Each of these algorithmic categories has distinct objectives for scenario-specific applications. Consider the difference between asking (1) how much a customer will spend versus (2) which customers are most alike. Point of sales (PoS) records can answer either question, though the question structure creates a dependency on which sales information becomes relevant facts into (1) regression versus (2) clustering solutions. Using the wrong approach or not defining the outcome goals upfront always leads to nonsensical results (Snee, 2015). Instead, a formal data mining lifecycle begins with a specific question, then collects relevant facts to derive a conclusion. Next, an evaluation method confirms these conclusions are scientifically sound and not wishful thinking through some statistical variance or cross-validation testing. After constructing a probabilistic model of the scenario, the company needs to deploy it into its production environment and begin collecting a return on investment (ROI). Measuring the amount of return depends on Key Performance Indicators (KPI) that typically align with high-level corporate mandates, such as increasing sales per customer or reducing inventory carry times. Despite alignment challenges across data producers, business questions, relevant facts, conclusions, operationalization, and KPIs—planning and methodical approaches lead to success.

# Section I: Business Structure

## Who is Black Bean

Black Bean started as a small ‘mom and pops restaurant,’ and over the last several years, expanded operations to include hundreds of international locations. The company prides itself on delivering consistently high-quality fruit desserts, regionally sourced from local farms. This decentralized approach to supply chain management allows each site to reduce shipping times and ensure the freshest produce. However, it creates challenges for the corporate office, as purchase invoices and inventory management reports do not follow consistent schemas. These discrepancies are not limited to format and also include different units (e.g., pounds versus kilograms), which makes direct value comparisons impossible.

## What types of data artifacts exist

Guidelines published by the executive leadership also require several Key Performance Indicators (KPI) that describe sales, health and safety, and customer satisfaction. For instance, auditing the temperature of all refrigeration must occur hourly. Many restaurant locations have embraced IoT sensors for automating these collections, though some values still come from manual entry. Each location uploads these data points into a data lake hosted in the public cloud.

## What business goals use these artifacts

Snee (2015) misconception that lots of data + analysis = magic.

A recent outbreak of food poisoning at some locations has damaged the brand’s image and caused a significant decrease in sales. The leadership team wants to restore consumer confidence by operationalizing their data lake to answer targeted questions about the incident. Which sites are likely to have an outbreak next? Are food handling procedures being followed? Who should promotional material target?

# Section II: Collecting and Enhancement

## Preprocessing Data Resources

According to Gibert et al. (2016), nearly 70% of all data mining occurs during the cleaning phase.

Before analysis can begin, the data analyst needs to normalize the incoming data through an extract-transform-load (ETL) process. This process needs to perform column renaming and reordering, adjusting quantity units, filtering erroneous values, populating missing values, and similar cleanup actions. When analysis does not handle these aspects upfront, it creates a garbage-in/garbage-out scenario. For example, a temperature reading of 55 degrees could be manually entered as 555 degrees, causing later analysis to become skewed.

After cleaning and schematizing the incoming data, the next analyst needs to determine which aspects are relevant for their data mining objective. Having large amounts of unrelated information does not improve results, and for many scenarios, it only slows down model training times.

Another critical challenge is handling missing values (Rawal et al., 2017) as they need to be normalized or removed. These decisions become scenarios specific.

## Required Collection Resources

* Personal
* Hardware
* Future Capacity or needs exist

## What logical components or assumptions exist

* Logical Component 1 of 2
* Logical Component 2 of 2

# Section III: Evaluation Procedures

## What statistical techniques can measure process ROI

# Section IV: Applications

## What data mining strategies can apply to this information

# Conclusions and Future Studies