Service Analytics

6

Jorge Cardoso, Julia Hoxha, and Hansjörg Fromm

Summary

Service analytics describes the process of capturing, processing, and analyzing the data generated from the execution of a service system to improve, extend, and personalize a service to create value for both providers and customers. This chapter explains how services, especially electronic services, generate a wealth of data which can be used for their analysis. The main tasks and methods, from areas such as data mining and machine learning, which can be used for analysis are identified. To illustrate their application, the data generated from the execution of an IT service is analyzed to extract business insights.

Learning Objectives

1. Understand the concept of service analytics and its importance for service systems.
2. Describe the various tasks and methods associated with analytics and how they can be applied to services.

(continued)

J. Cardoso (✉)
Department of Informatics Engineering, Universidade de Coimbra,
Coimbra, Portugal

Huawei European Research Center (ERC), Munich, Germany
e-mail: jcardoso@dei.uc.pt; jorge.cardoso@huawei.com

J. Hoxha • H. Fromm
Karlsruhe Service Research Institute (KSRI), Karlsruhe Institute
of Technology (KIT), Karlsruhe, Germany
e-mail: hansjoerg.fromm@kit.edu

J. Cardoso et al. (eds.), *Fundamentals of Service Systems*, Service Science: Research and Innovations in the Service Economy, DOI 10.1007/978-3-319-23195-2_6

3. Explain how classification, prediction, and association rules, from the traditional data mining field, can be applied to service systems.
4. Understand how analytics can be used to analyze real-world datasets generated from the execution of IT services.

▶ **Opening Case** Analytics for service improvement

Analyzing Customers' Behavior While Interacting with Service Systems

A couple is traveling with their new car on the expressway. It is cold outside, the heating is on, but they are not pleased with the heat distribution: the upper part of the cabin interior is warm, but the foot space is cold. They try to regulate the temperature using the climate menu of the car's control display—without success.

The co-driver takes the owner's manual out of the glove compartment and finds a section on climate control on page 109. But page 109 does not give the necessary information to solve their problem. From page 109, a reference is made to page 80, where all controls of the center console are described. After browsing backwards and forwards between pages 109 and 80 a number of times, they still have not found the required information.

This simple example is illustrative of many situations that are found in service environments. Consider the car owner's manual as an *information service* that the auto manufacturer provides to customers. After having printed and placed the manual in the glove compartment of every car, the manufacturer knows practically nothing about the usage of this service (the manual).

Possibly hundreds of drivers have been reading page 109 of the car owner's manual. If the car manufacturer would have access to this information, it could reach the conclusion that something is wrong with the temperature control of the car—or at least with the usability of the control display. If the manufacturer would know that many people browse back and forth between pages 109 and 80, it would know that something was wrong with the editorial structure and the content of the manual. Knowledge on the usage could provide important feedback to the engineering department as well as to the department responsible for the owner's manual.

However, if instructions are only printed on paper (Fig. 6.1), there is a "disconnection" between the provider and the user of the service. This disconnection prevents information from flowing from the service consumer to the service provider.

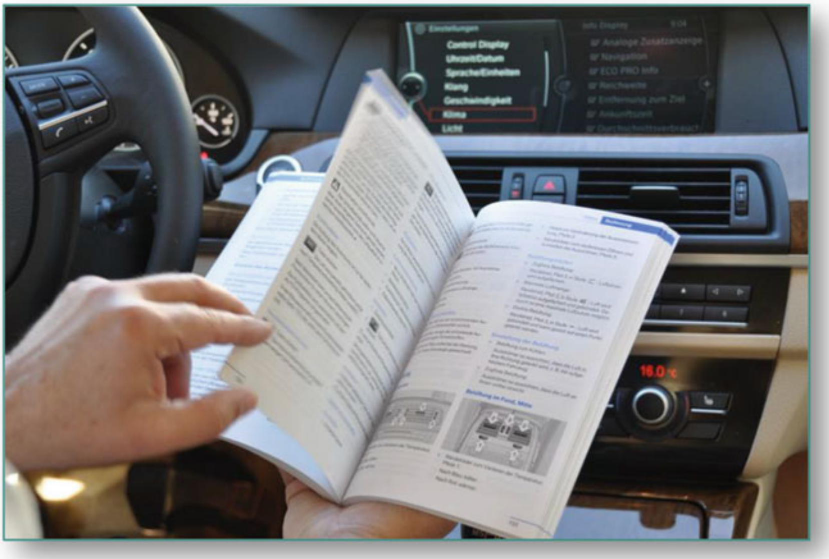
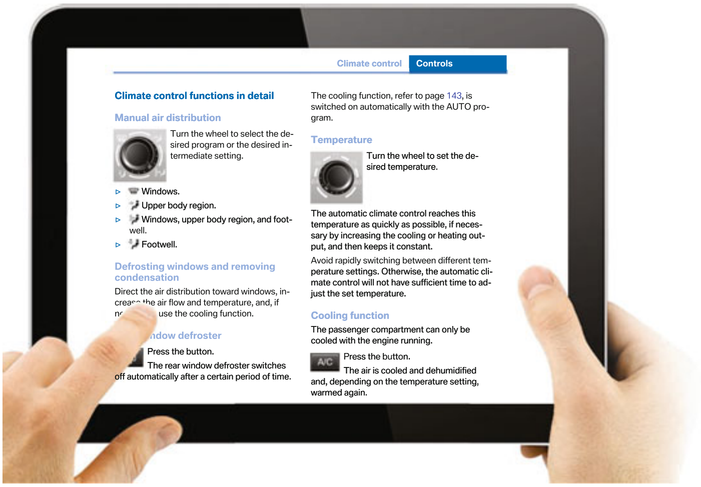


Fig. 6.1 Owner's manuals are typically disconnected from their providers


Climate control | Controls

Climate control functions in detail

Manual air distribution

Turn the wheel to select the desired program or the desired intermediate setting.

- Windows.
- Upper body region.
- Windows, upper body region, and footwell.
- Footwell.

Defrosting windows and removing condensation

Direct the air distribution toward windows, increa the air flow and temperature, and, if n use the cooling function.

ndow defroster

Press the button.

The rear window defroster switches off automatically after a certain period of time.

The cooling function, refer to page 143, is switched on automatically with the AUTO program.

Temperature

Turn the wheel to set the desired temperature.

The automatic climate control reaches this temperature as quickly as possible, if necessary by increasing the cooling or heating output, and then keeps it constant.

Avoid rapidly switching between different temperature settings. Otherwise, the automatic climate control will not have sufficient time to adjust the set temperature.

Cooling function

The passenger compartment can only be cooled with the engine running.

A/C Press the button.

The air is cooled and dehumidified and, depending on the temperature setting, warmed again.

Fig. 6.2 The use of mobile devices enables manufacturers to monitor and analyze customers behavior

If the manual would be electronically accessible in the car and connected to the provider using the internet (Fig. 6.2), the manufacturer could analyze customers' usage patterns. Counting the number of visits to particular pages could easily reveal which dashboard controls the customers have the most difficulties with. Measuring the time intervals between page visits could be an indicator for how well the text is readable and how quickly the customer finds a solution to his problem. The behavior described previously—a user jumping back and forth between two pages—could easily be detected and corrected in the next version of the manual.

► **Opening Case**

6.1 Introduction

Automotive manufacturers deliver an increasing number of services with their cars, such as traffic and weather information systems, and communication, driver assistance, navigation, and entertainment services. But they often do not know much about the usage of these services by their customers.

Electricity service providers typically ask their customers to read their meters once per quarter or once per year. Based on the consumption they bill customers. However, electricity providers miss information about weekly or daily consumption patterns, which would allow them to give a useful feedback to customers or offer specially-priced contracts.

In healthcare, service providers, such as family doctors or hospital staff, typically have enough information about the patient while under their custody. But as soon as the patient leaves their facilities, they lose track of their condition, health status, and behavior.

All these examples are characterized by an information gap between the service provider and the customer. This gap is preventing the provider to get more insight into the customer's service usage. But this disconnection is being overcome with the emergence of new information technologies. An increasing volume of data is being collected either by users (e.g., through smartphones) or by technologies like smart metering in energy services, telematics in automotive services, RFID in logistics, condition sensors in engineering, and healthcare.

R Traditional human-based services are characterized by the personal service encounter. Very often, the provider and the customer know each other well from past service interactions. The provider understands what the customer wants and can deliver an individual, personalized service. This advantage was initially overlooked in technology-enabled services, where the face-to-face contact between customer and provider did not occur anymore.

From the many use case domains available (e.g., automotive, energy, and healthcare), this chapter presents a real-world example from IT service management. Section 6.3 explains how service provisioning can be analyzed with the objective of improving operational performance and increasing productivity. The section analyzes an IT service, more precisely, the incident management service from the Information Technology Infrastructure Library (ITIL) [1]. The objective of this service is to restore normal service operation as quickly as possible and ensure that the best possible levels of service quality and availability are maintained.

The next chapter, Chap. 7, provides a complementary view on service systems by explaining how optimization theories and techniques can be used to improve the efficiency and effectiveness of services.

6.1.1 Sharing Data

If customers are willing to share data with service providers, providers will have the opportunity to study customers' behaviors and preferences to gain insights on the customer almost like it was possible in traditional face-to-face services. But studying service usage does not only provide more focus towards the customer, it can also help to identify weaknesses in the provisioning of the service which in turn gives rise to service improvements. This new focus on studying service usage can be achieved with service analytics.

In the energy scenario above, a move from quarterly or yearly meter readings to almost real-time consumption recordings with smart meters can result in a win-win situation for both customers and providers. With detailed knowledge about the fluctuating energy consumption over the day, energy providers can improve management by more accurately anticipating expensive peak consumption periods. At the same time, the provider can influence demand, e.g., by lowering electricity prices and, thus, incentive consumption in off-peak periods. The provider can give feedback on usage statistics that help the customer to identify power guzzlers in the household to handle energy consumption with greater care. In summary, there are advantages for all parties: lower energy prices for consumers, operational improvements for providers, and environmental benefits for everyone.

6.1.2 Big Data

Reading a smart meter every 15 minutes instead of only once a year generates about 35,000 times more data. Reading it every 15 s generates over 2,000,000 times more data. This is the dilemma of service analytics: when providers and customers were disconnected, there was no data about the service usage available. Now that new technologies like sensors, smart meters, and telematics have been introduced, there are large volumes of data available—often more than desired. This has been recently described with the term *big data* [2].

► **Definition (Big Data)** Big data describes large volumes of both structured and unstructured data that are difficult to process using traditional databases and software techniques. The difficulties are associated with the nature of the data: volume (e.g., terabytes and petabytes), velocity (e.g., streaming and near-real time), and variety (e.g., formats).

In the case of service systems, the volumes of captured data have grown extremely fast. Big data processing now often allows to work with the raw data in situations that were not possible before, both in terms of the statistics (enough samples available) and in terms of technology (enough processing power available).

6.1.3 Knowledge Discovery

The process of creating useful knowledge from large data sets and documents is often described as knowledge extraction or knowledge discovery. It describes the overall process of discovering useful knowledge from data, while analytics and data mining refer to one particular step in this process.

▶ **Definition (Knowledge Discovery (KD))** The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [3].

The KD process starts with data preparation, data selection, data cleaning, and the incorporation of appropriate prior knowledge, before analytics/data mining techniques can be applied. After application of these techniques, the proper interpretation of the results is essential to ensure that useful knowledge is derived from the data.

The blind application of data mining methods can be a dangerous activity, which can easily lead to the discovery of meaningless and invalid patterns.

6.1.4 Analytics for Service Systems

The application of analytics to service systems is progressing in many domains. Some areas like web analytics are more advanced, while other areas like healthcare analytics are making the first steps [4, 5]. Interesting examples include digital content providers (e.g., iTunes, Spotify, and Amazon Kindle) which use analytics to recommend and personalize content; online gaming services that use analytics to improve their games; telco providers that analyze communications behavior to optimize provisioning of network resources; and price comparison websites and intermediaries (e.g., tour operators) that use demand patterns to optimize pricing.

The goal is to apply basic and advanced analytics to an entire service system to generate the highest benefits for all stakeholders according to the value co-creation principle. The challenge is to draw information and insights out of big volumes of data to study customer behavior and characterize service usage. This requires sophisticated methods for capturing, processing, and analyzing data. These methods are subsumed under the term service analytics [6].

▶ **Definition (Service Analytics)** Describes the methods for capturing, processing, and analyzing data taken from a service system—in order to improve, extend, and personalize the service provided. It also describes how new value is created for both the provider and the customer.

When data is available, the potential is clearly visible—as with electronic services (services fully rendered over the internet) since by design these services require connectivity between providers and customers. For example, customers visit the provider's web pages in order to obtain the service. Thus, the provider is able to analyze customers usage characteristics at several levels of detail. This analysis is known under the terms web analytics or web usage mining [7,8]. Typical data of interest are the overall number of page visits, the number of page visits per customer, the time intervals between page visits, the path that customers take through the web site, etc. With this data, the provider can analyze the behavior and preferences of individual customers, make recommendations, assess the general acceptance and attractiveness of web offerings, and discover usability problems related to navigating and finding information on web pages.

The methods and techniques from web analytics can readily be transferred to analyze service systems. It is mainly due to the difficulties in obtaining usage data and in reducing the vast amount of data to a manageable size that these techniques have not been used extensively in the past.

6.2 General Notion of Analytics

There is no single agreed-upon definition of the term *analytics*. Some authors like Kohavi et al. [9] use the terms analytics and *data mining* interchangeably. Others like Davenport and Harris [10] use analytics as a synonym for *business intelligence*, a term which refers to the applications and best practices to analyze information to improve and optimize business decisions and performance.

▶ **Definition (Data Mining)** Generally, data mining (sometimes called data or knowledge discovery) is the process of extracting useful, often previously unknown information, from large databases or data sets.

Opposing opinions are grounded to the doubt if analytics should include or exclude data management and reporting technologies. Davenport and Harris [11] distinguish between access and reporting, and analytics, both are seen as subsets of business intelligence. Data management and reporting are often considered as basic analytics, which are a prerequisite for advanced analytics (see [12]) built on methods from statistics and operations research. Basic analytics include reporting solutions based on data warehouse and data marts like standard and ad hoc reporting, online analytical processing (OLAP), queries, drilldowns, and alerts [13].

6.2.1 Data Preprocessing

This section identifies the most common tasks, which are part of the process executed in the context of a data analytics project. Such a process usually involves data pre-processing, cleaning, integration, transformation, reduction, and discretization.

Data pre-processing is an important first step in the analytics process. Since data gathering is usually performed with loosely controlled methods, they can generate datasets with several missing and out-of-range values, redundant entries, invalid data combinations, etc. It is often the case that datasets exhibit the following three main problems:

1. Data may be incomplete, for example they lack attribute values or attributes of interest.
2. Data may be noisy, containing errors, anomalies, or outliers.
3. Data may also be inconsistent, for example they contain discrepancies in codes or names.

Performing analytics on data that have not been previously checked against these problems may cause misleading results. The data pre-processing activity screens the gathered data for quality problems. Data pre-processing typically includes the following tasks: *cleaning*, *integration*, *transformation*, *reduction*, and *discretization*.

Data Cleaning

Data cleaning consists in the following activities: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies. Missing data is handled in different ways, for example, by manually filling values; using global constants; applying the mean of the attribute values to fill the missing value of a particular attribute; finding the most probable value by applying, e.g., Bayesian rules or decision trees; or simply ignoring the instance whose value is missing. Noisy data are handled by sorting and partitioning the data to detect and remove outliers. This is done by applying automatic binning, clustering methods, or through human inspection.

Data Integration and Transformation

Data integration combines datasets from multiple sources, such as files, databases, or data cubes, into a coherent data store. Integration tasks confront the problem of resolving value conflicts, as well as handling redundant data. Data transformation helps with normalization, i.e., scaling of values to fall within a range, as well as with aggregation techniques.

Data Reduction

Data reduction is a very important step when, as it is usually the case, organizations are dealing with very high volumes of data and need to perform analytics tasks, which are computationally expensive if performed on complete datasets. As such, the data reduction task consists in finding a reduced representation of the dataset that is smaller than the complete set, yet it enables to yield almost the same analytical results. Among the strategies used in data reduction are data aggregation, dimensionality reduction, numerical reduction, and concept hierarchy generation.

Data Discretization

Data discretization, often considered as part of data reduction, consists in reducing the number of values of a particular continuous attribute by dividing the range of the attribute into intervals. Afterwards, the data values of the attribute are replaced by discrete labels.

6.2.2 Data Analysis

Methods subsumed under the title analytics originate from mathematics (statistics, linear algebra, and operations research) and computer science (database systems, data mining, machine learning, artificial intelligence, and computational linguistics).

Table 6.1 gives an overview of the different analytics methods and how they can be applied to service systems. In Section 6.3, a selection of these methods—data preprocessing, classification and prediction, and association rules—will be presented in more detail. Section 6.4 briefly presents cluster analysis, regression analysis, and text mining. For more information on these and other methods, the reader is referred to textbooks on data mining [7, 14], and business forecasting [15, 16].

Classification of Approaches

Analytics comprise various methods from statistics and operations research, which can follow descriptive, predictive, or prescriptive approaches [17].

Descriptive analytics is the simplest form of analytics. It condenses and summarizes big data into smaller, more useful and manageable pieces of information. For example, creating summarized information about the number of customers or about the page views of an electronic service.

Predictive analytics uses several methods to study historical data to forecast future trends or events. A model is created using past data to predict future data.

Prescriptive analytics extends descriptive and predictive analytics by recommending to decision-makers the possible courses of action that can be taken as well as the likely outcome of each decision.

Several existing methods can be used to support service analytics. Examples include text and image mining, clustering, anomaly detection, forecasting algorithms, and visual analysis:

- Data mining and machine learning algorithms like clustering and association can be used to identify similarities between customers (e.g., segmentation).

Table 6.1 Overview of analytics tasks, methods, and applications

Tasks	Methods	Applications
Preprocessing	Preprocessing Techniques (Data Cleaning, Data Transformation) Dimensionality Reduction (Principal Component Analysis (PCA), Support Vector Machines (SVM), Factor Analysis (FA), Eigen decomposition, Latent Variable Analysis)	Reduction of large volumes of data to manageable size (dimensionality reduction) Removal of incorrect, irrelevant and redundant information
Classification	Decision Tree (C4.5), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), logistic regression	Prediction of user behavior; prediction of customer churn/attrition, loyalty, profitability; prediction of failures (resulting in predictive maintenance)
Association	With candidate generation (k-Means, k-Medoids) Without candidate generation (RELIM, FP-GROWTH)	Co-purchase information (Market Basket Analysis)
Cluster Analysis	Partitioning Approaches (k-Means, k-Medoids) Hierarchical Approaches (AGNES, DIANA) Density-Based Approaches (DBSCAN, OPTICS, CLIQUE) Outlier Detection (ABOF)	Customer segmentation; similarity of interest; combinations of problems/complaints; recommendations; identifying customers requiring similar types of assistance; identifying service objects with similar problems (e.g., auto repair)
Regression	Basic Statistical Techniques (nonlinear, multi-variate, logistic regression) Time Series Forecasting (Moving Average, Exponential Smoothing, Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA))	Demand forecasting (electricity consumption, service calls, emergency calls, service parts, etc.); usage trend analysis; next action by customer; customer lifecycle management
Sequence Mining	Apriori-Based Algorithms (AprioriAll, GSP, SPADE) Pattern-Growth Algorithms (FreeSpan, PrefixSpan) Early Pruning Algorithms (SPAM)	Service usage analytics (flow of patients through hospital, user browsing patterns); mobility pattern analysis (mobile phone usage, traffic, sports); repeat visit analysis (repair, hospital); temporal recommendations; service personalization
Text Mining	Preprocessing Techniques (document standardization, tokenization, filtering, lemmatization, stemming) Sentiment Analysis Text Summarization	Customer/patient complaint analysis (from calls or social networks); customer experience analysis; customer emotion and sentiment analysis; analysis of service documentation (e.g., maintenance records and reports, medical records and reports)

- Text mining algorithms are the basis for customer experience, sentiment, and complaint analysis using unstructured data sources like blogs on the internet.
- Temporal data mining algorithms are able to discover sequential usage patterns, e.g., in the browsing behavior of customers on a provider's web pages.
- Visual analytics and visual data mining provide a clearer view and understanding of relationships within a complex service system.
- Time series forecasting and regression analysis methods can be used for prediction and trend analysis (e.g., service demand).

The boundaries between descriptive, predictive, and prescriptive methods are blurred. Association, clustering, sequential pattern analysis, and text mining (also visual data mining which is not mentioned in Table 6.1) are generally classified as descriptive methods, whereas classification, regression, and time series forecasting are clearly predictive methods. Prescriptive methods are in the realm of mathematical optimization and are covered in the Chap. 7 on Service Optimization.

Characterization of Methods

Data mining experts use the following terms to characterize algorithms:

- Supervised learning
- Unsupervised learning

In *supervised learning*, the data records contain attributes (input variables) and a target variable (output variable or label). The objective of learning is to construct a function which relates a given set of attributes with the desired output. If this function has been determined from an existing training data set, it can be used to predict the output for any new combination of attributes. Thus supervised learning methods are clearly predictive.

Unsupervised learning methods do not use labeled data (no target variable)—the goal is rather to find patterns and structures in the data, which were previously unknown [7]. Unsupervised learning methods are therefore descriptive.

6.3 Analyzing IT Services

ITIL is commonly used in the industry for IT service management. It provides a set of best practices, which take often the form of reference models and accepted processes, which are sound and efficient. The adoption of reference models is motivated by the following drivers [18]:

Design. They significantly speed up the design of services by providing reusable and high quality content.

Optimization. They optimize the design as they have been developed over a long period and usually capture the business insight of experts.

Compliance. They ease the compliance with industry regulations and requirements and, thus, mitigate risk.

Alignment. They are an essential mean to align business needs and IT service implementations.

Worldwide, many well-known companies are adopting ITIL for IT service management. Examples include large software providers such as Microsoft, HP, and IBM; financial services societies such as Bank of America, Deutsche Bank, and Barclays Bank; manufacturers such as Boeing, Caterpillar, Toyota, and Bombardier; and departments of defence such as the US Army, US Navy, and US Air Force.

Example

As a concrete example of cost reduction, Proctor and Gamble reduced IT spending in 10 % of their annual IT budget ($125M) by adopting ITIL. The efficiency and optimization of service provisioning were the main reasons behind the savings.

The ITIL Lifecycle

ITIL consists of five main books, which correspond to the five phases of the ITIL lifecycle: service strategy, service design, service transition, service operation, and continual service improvement. An introductory book to ITIL service management is also available. Each of the five main volumes textually describes the best practices that can be followed by a company to manage IT services. Thus, ITIL should not be viewed as a piece of code, system, or software application.

This section will look into the service operation phase, and, more precisely, it will analyze the incident management service.

The Incident Management Service

The primary objective of the incident management (IM) service is to resolve incidents (e.g., application bugs, disks-usage thresholds exceeded, or printers not working) in the quickest and most effective possible way. The incident management service can be characterized as reactive, and a formal working process is put into place to respond as efficiently and effectively as possible in resolving reported incidents.

If a user cannot print, he contacts the service desk for help, which creates a record describing the incident. If the issue cannot be resolved immediately, the service desk manager opens an incident record, which is assigned to a technician. When the technician finds the cause of the incident, he fixes the problem. The service desk manager informs the user to retry to print. If the user can print, the service desk manager closes the incident record. Otherwise, the record remains open and another attempt to resolve the incident is made. Figure 6.3 provides a simple representation of the business process model behind the IM service.

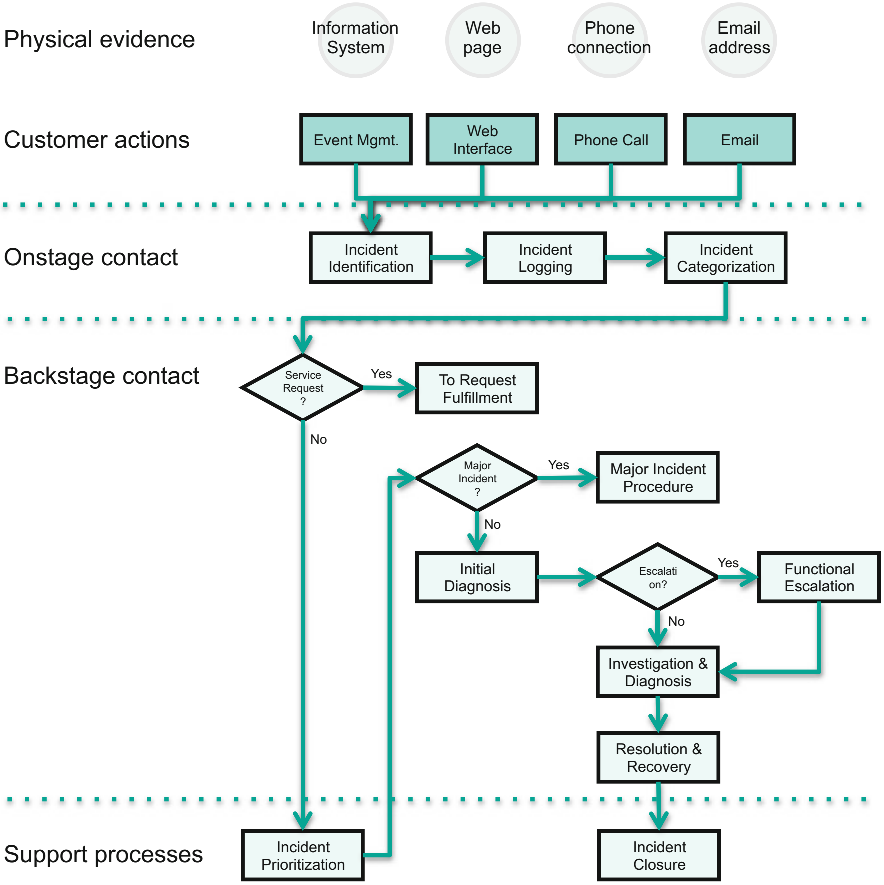
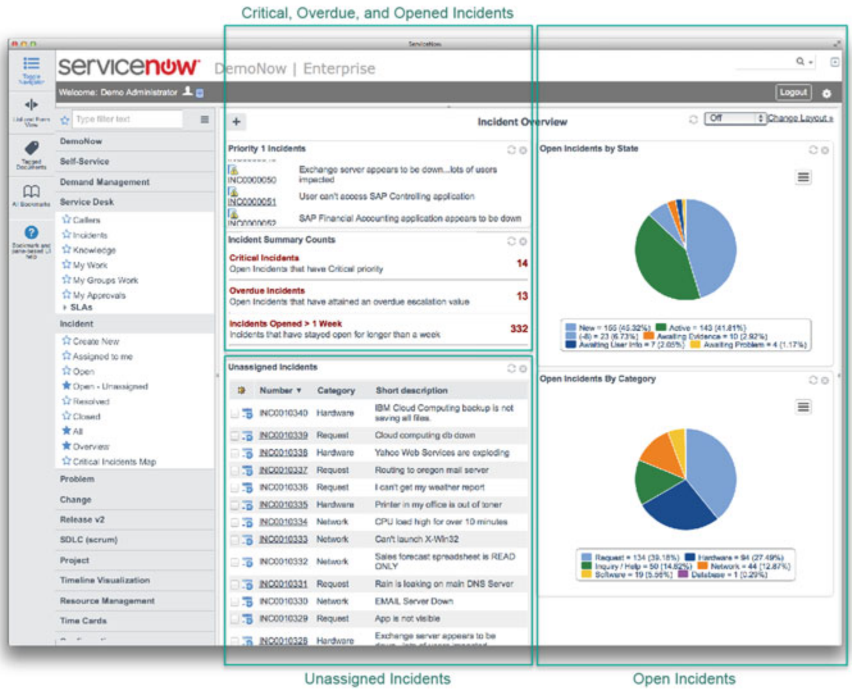


Fig. 6.3 The ITIL incident management service blueprint (adapted from p. 48, [19])

6.3.1 ITIL Software Systems

Nowadays, software solutions to support and manage ITIL services already exist (e.g., ServiceNow, Zendesk, Kayako, UserVoice, and Freshdesk). Depending on the requirements, companies can acquire solutions containing a broad list of features ranging from SLA and escalation management, to the integration with social media platforms, to automated ticket routing, and to graphical forms design.

During service provisioning, incidents are handled using the activity record system. Activities are assigned to appropriate team members, who will deal with the task as appropriate. These systems generate events, which are recorded in a log to provide an audit trail that can be used to understand how services were provisioned. For example, each time an activity of the service illustrated in Fig. 6.3 is executed, an event is generated and stored in a log. Event records typically contain information



Priority 1 Incidents

INC0000050	Exchange server appears to be down...lots of users impacted
INC0000051	User can't access SAP Controlling application
INC0000052	SAP Financial Accounting application appears to be down

Incident Summary Counts

Critical Incidents Open Incidents that have Critical priority	14
Overdue Incidents Open Incidents that have attained an overdue escalation value	13
Incidents Opened > 1 Week Incidents that have stayed open for longer than a week	332

Unassigned Incidents

Number ▾	Category	Short description
INC0010340	Hardware	IBM Cloud Computing backup is not saving all files.
INC0010339	Request	Cloud computing db down
INC0010338	Hardware	Yahoo Web Services are exploding
INC0010337	Request	Routing to oregon mail server
INC0010336	Request	I can't get my weather report
INC0010335	Hardware	Printer in my office is out of toner
INC0010334	Network	CPU load high for over 10 minutes
INC0010333	Network	Can't launch X-Win32
INC0010332	Network	Sales forecast spreadsheet is READ ONLY
INC0010331	Request	Rain is leaking on main DNS Server
INC0010330	Network	EMAIL Server Down
INC0010329	Request	App is not visible
INC0010328	Hardware	Exchange server appears to be down...lots of users impacted

Fig. 6.4 Software application to support the ITIL incident management service illustrated in Fig. 6.3

such as a timestamp, the name of the activity executed, the owner and priority of the incident case, the status of the service, and a description of the incident. Once an incident has been properly dealt with, it is closed.

Figure 6.4 provides an example of an interface that team members use to handle incidents.

These log files need to be integrated into one consistent dataset for analysis. This process includes the detection of errors and inconsistencies to improve the quality of data, the integration of schema, and the consolidation of instances. Afterwards, service analytics are applied to the dataset to extract valuable insights and knowledge about service provisioning patterns.

6.3.2 Dataset Description

A dataset from the automotive industry is used to explain how service analytics can be operationalized. It is a real-life dataset, which contains event records from operational business processes. The dataset was generated by an incident management system called VINST and has the following characteristics.

Table 6.2 The attributes of the dataset (attributes are written using the `typewriter` font)

Attribute	Count	Description
`SR number`	7.554	A unique ticket number for each incident reported.
`Change date+time`	many	The timestamp indicating when the status of an incident case changed.
`Status`	4	The status of the incident management service: queued, accepted, completed, closed.
`Substatus`	13	The substatus of an incident case: assigned, awaiting assignment, cancelled, closed, in progress, wait, unmatched.
`Involved ST function div`	24	The IT organization that provides the service. It is divided into functions (mostly technology wise).
`Involved organization`	25	The business area of the user reporting the incident to the service desk.
`Involved ST`	649	The team responsible for resolving the incident.
`SR latest impact`	4	The impact of an incident for the customer: major, high, medium, low.
`Product`	704	The identification of the product which originated the incident.
`Country`	23	The country of the support team that takes ownership of the incident record.
`Owner country`	32	The country of the owner.
`Owner first name`	1.440	The person of the support team that is the owner of the reported incident.

- It is available from the Business Process Intelligence (BPI) Challenge 2013 web site.[1]
- It contains event records from a three-week period from the 1st of May 2012 up to and including the 23rd of May 2012.
- It contains 65.533 events pertaining to 7.554 incident records.
- It was not preprocessed or filtered, other than anonymized.

Table 6.2 provides the description of the attributes that are part of the log schema and Table 6.3 shows an extract of the dataset.

6.3.3 Preprocessing and Cleaning

It was necessary to preprocess the dataset before conducting the analysis. The basic tool for data munging used for our analyses was Microsoft Excel. Its feature to import CSV files has allowed the original dataset to be imported to a tabular format. The filtering and sorting capabilities of Excel have eased the identification of missing data and have allowed for raw estimations of attributes' count (Table 6.2 shows the count for each attribute).

[1] www.win.tue.nl/bpi/2013/challenge.

Table 6.3 Examples of instances of the dataset

Instances
`1-364285768;2010-03-31T15:59:42+01:00;Accepted;In Progress;A2_4;Org line A2;V30;Medium;PROD582;fr;France;Frederic`
`1-364285768;2010-03-31T16:00:56+01:00;Accepted;In Progress;A2_4;Org line A2;V30;Medium;PROD582;fr;France;Frederic`
`1-364285768;2010-03-31T16:45:48+01:00;Queued;Awaiting Assignment;A2_5;Org line A2;V5 3rd;Medium;PROD582;fr;France;Frederic`
`1-364285768;2010-04-06T15:44:07+01:00;Accepted;In Progress;A2_5;Org line A2;V5 3rd;Medium;PROD582;fr;France;Anne Claire`
`1-364285768;2010-04-06T15:44:38+01:00;Queued;Awaiting Assignment;A2_4;Org line A2;V30;Medium;PROD582;fr;France;Anne Claire`
`1-364285768;2010-04-06T15:44:47+01:00;Accepted;In Progress;A2_5;Org line A2;V13 2nd 3rd;Medium;PROD582;fr;France;Anne Claire`
`1-364285768;2010-04-06T15:44:51+01:00;Completed;Resolved;A2_5;Org line A2;V13 2nd 3rd;Medium;PROD582;fr;France;Anne Claire`

The Python language was used during the loosely process of manually converting the dataset from its original format into another format that allowed for a more convenient analysis. The objective of the conversion was to obtain a better understanding of support tiers, owner names, and to handle missing values.

Support Tiers

The incident management service aims to restore normal service operations after the occurrence of a specific incident. Incident reports are first handled by the 1st line (the service desk) and escalated to the 2nd and 3rd line teams when 1st line engineers are not able to resolve the incident.

To explore the behavior of service provisioning, it was required to know which support tier (1st, 2nd, or 3rd line) handled an incident. Unfortunately, the attribute `Involved ST` combined, into one field, the support team name and the support tier. For example, `S2 2nd` combined the name of the support team `S2` and the support tier `2nd` indicating the 2nd support tier. A 1st line involvement was assumed for activities for which no explicit line number was present in the `Involved ST` attribute. For example, `S2` indicated implicitly a 1st level support tier. Therefore, a new attribute was created to solely indicate the support tier. The `Involved ST` was divided into two attributes: `Involved ST` and `Support line`. The values for `Support line` were `1st`, `2nd`, `3rd`, and `2nd-3rd`.

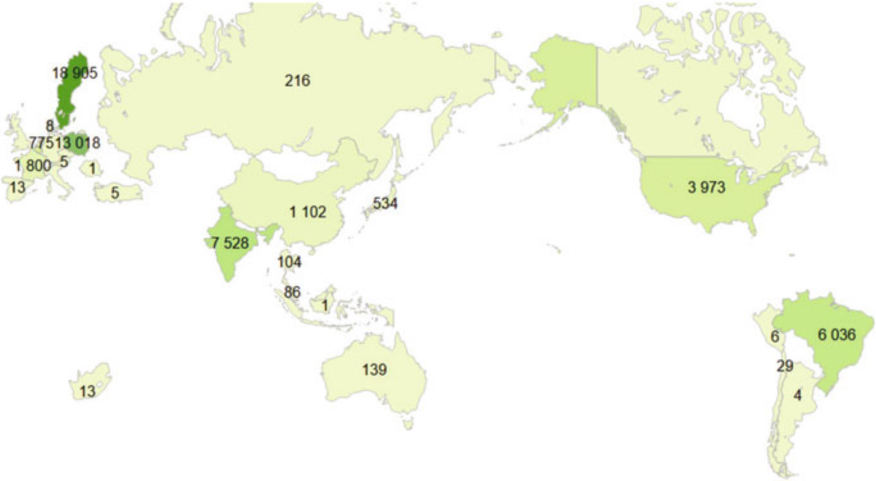


Fig. 6.5 Worldwide distribution of the volume of activities associated with the incident management process [20]

Basemap, a library for plotting 2D data on maps in Python, was used to visualize the results of converting the original dataset. The map in Fig. 6.5 shows that 2.911 (4.44 %) activities are performed in the 3rd line mainly in Sweden, Poland, France, and India. Incidents that require the intervention of the `2nd-3rd` support line were handled in France (39 (0.06 %) activities). More than 46.000 (70.26 %) activities are performed by the 1st tier mainly in Sweden, Poland, Brazil, and India. The most active countries that involve the 2nd line to solve incidents are Sweden, India, Poland, and Brazil with 16.541 (25.24 %) activities handled.

Owner Names

The attribute `Owner first name` has 1.440 unique values (e.g., `Frederic` and `Anne Claire`). These names do not map uniquely to the `Owner country` attribute. The most probable explanation is that several people using the system and located in different countries have the same name. To solve this problem, the `Owner country` was concatenated with `Owner first name` to create a new attribute. This generated 1.688 distinct new entities for the incident dataset.

Missing Values

Several values of the attributes `Status` and `Substatus` were not specified in the dataset. For example, `wait-customer`, `in-call`, `unmatched`, and `resolved`. For some activities, no organization or function division were indicated. Instances without these values were removed from the dataset. The format of timestamps was modified to enable their processing by various tools.

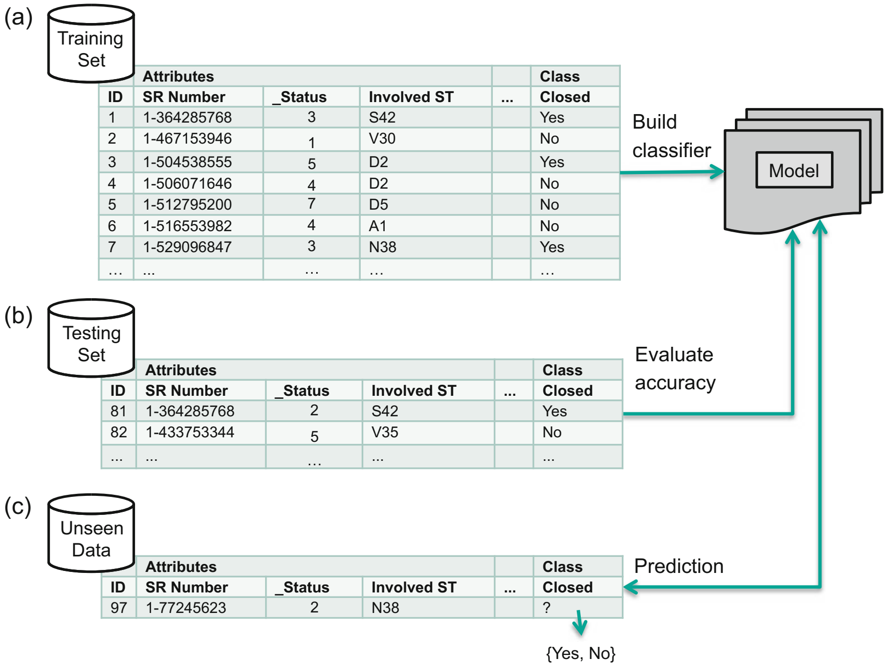
6.3.4 Predicting Incident Closure

The goal of this section is to construct a predictive model that will enable to understand the factors that influence an incident to be closed or not. More precisely, the objective is to understand the influence that the number of functional divisions, support teams, and organizations involved in the resolution of an incident has on the closing of incidents.

Classification and Prediction

Classification and prediction are forms of data analysis used to build models that capture important data patterns or predict future data trends. The question that will be answered is: *how to predict whether an IT service incident report submitted will be resolved or not*. The prediction is based on the attributes that describe an incident, e.g., support teams involved, country owning the incident, and functional departments involved. Organizations can make use of classification techniques to answer this question.

Classification is a two-step process (Fig. 6.6). In the first step, a model is built to describe a predetermined set of data classes. A collection of data records is used, such that each record contains a set of attributes. One of the attributes is the target



	Attributes				Class
ID	SR Number	_Status	Involved ST	...	Closed
1	1-364285768	3	S42		Yes
2	1-467153946	1	V30		No
3	1-504538555	5	D2		Yes
4	1-506071646	4	D2		No
5	1-512795200	7	D5		No
6	1-516553982	4	A1		No
7	1-529096847	3	N38		Yes
…	...	…	…		…

	Attributes				Class
ID	SR Number	_Status	Involved ST	...	Closed
81	1-364285768	2	S42		Yes
82	1-433753344	5	V35		No
...	...	…	...		...

	Attributes				Class
ID	SR Number	_Status	Involved ST	...	Closed
97	1-77245623	2	N38		?

Fig. 6.6 The classification process: (**a**) the training dataset is used to build a classification model, (**b**) the test data is used to evaluate the accuracy of the model, and (**c**) unseen incident reports are classified using the model

class, often referred to as the class label attribute. The goal is to build a model for the class label attribute as a function of the values of other attributes.

The second step is the classification (prediction). The inferred model is used to classify previously unseen records by assigning a class to each of them as accurately as possible. Data records are also referred to as instances, tuples, samples, examples, or observations. The data records analyzed to build the model collectively form the training data set. A test set is used to determine the accuracy of the model. Usually, a given dataset is divided into training and test sets. The training set is used to build the model and the test set is used to validate the model.

Decision Trees

A *decision tree* is a structure resembling a graph, which looks like an upside down tree. It is composed of internal nodes, which contain attribute test conditions, branches to child nodes that correspond to the possible values of the respective attribute, as well as leaf (terminal) nodes representing classes. The starting node in a decision tree is the root node.

Figure 6.7 illustrates a typical decision tree build to classify incident records. This tree can be used to predict whether or not an incident record will be closed. Internal nodes are represented by rectangles, whereas circles represent leaf nodes. To classify an unknown incident instance, the attribute values of the instance are tested along the decision tree, tracing a path from the root node to the leaf node that predicts its class.

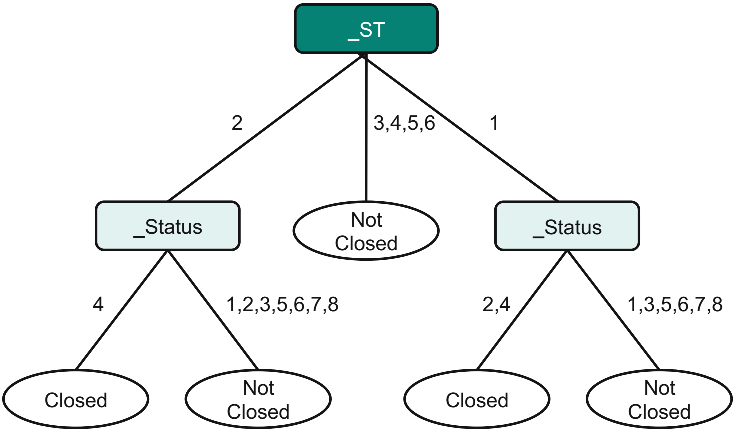


Fig. 6.7 A decision tree for IT service provisioning analysis, which can be used when an IT provider needs to quickly decide whether or not an incident report is likely to be resolved (it should be noted that this tree generated by Weka can be optimized). The underscore (_) in front of `ST` and `Status` indicates that the attribute captures the number of times the value of the attribute has changed in the log file

The process of constructing a decision tree out of known samples (observations/-training data) is known as decision tree induction. The basic algorithm uses a greedy strategy, which builds a decision tree in a top-down, recursive, divide-and-conquer way. The basic approach is the following:

1. Create the root node representing all instance records.
2. Choose the best attribute to split the remaining instances and make that attribute a decision node.
3. Repeat this process recursively for each child.
4. Stop when:
   (a) All the instances have the same class attribute label.
   (b) There are no more attributes.
   (c) There are no more instances.

The basic procedure to build a decision tree is the same for most decision tree-based algorithms and only differs on the method used to select the best attribute (node) to split the instances. The best attribute provides most information based on an *attribute selection measure*. It is the best one to split the samples into individual classes. A popular algorithm is the ID3 that uses the measure of *information gain* to select the best attribute for splitting.

Training and Test Set

To build a more accurate predictive model, new attributes were added to the original dataset that was shown in Tables 6.2 and 6.3. This involved changes at the schema level and instance level. The original dataset was parsed and for each incident record, a new instance was created with new attributes.

A Python application was created to read the original dataset (see Table 6.2) and transform its content to contain five additional attributes[2] (see Table 6.4):

`_Status, _FuncDiv, _Org, _ST, _Closed`

The transformation involved reading the attributes `_Status`, `_FuncDiv`, `_Org`, and `_ST` from the original dataset, and adding a new attribute indicating how many times the original attribute values changed for each incident reported. For example, `_ST` = 5 indicated that an incident was handled by 5 support teams. Another new attribute was created to represent the target class: the attribute `_Closed`. It was set to `true` to indicate that an incident report was closed.

Attribute Selection

Since many of the attributes present in the original dataset were possibly not relevant for the objective of this study, and, thus, did not contributed to creating a good predictive model, the less significant attributes were removed. The correlation-based

[2] New attribute names start with an underscore _.

Table 6.4 The new attributes added to the original dataset (the column *count* identifies how many different values exist for each attribute)

Attribute	Count	Description
`_Status`	8	The number of distinct states associated with an incident report.
`_FuncDiv`	4	The number of distinct functional divisions involved in resolving an incident.
`_Org`	13	The number of organizational divisions involved in resolving an incident.
`_ST`	24	The number of support teams involved in resolving an incident.
`_Closed`	2	A Boolean variable indicating if a incident was closed or not.

feature subset selection algorithm [21] was used to evaluate the worth of a subset of attributes by considering the individual predictive ability of each attribute along with the degree of redundancy between them. Subsets of attributes that are highly correlated with the class while having low inter-correlation were preferred. The application of the selection algorithm identified `_Status`, `_FuncDiv`, `_Org`, `_ST`, and `_Closed` as the most relevant attributes.

Classifier Construction

The J48 method from Weka (see Sect. 6.5) was used to build a decision tree. Figure 6.8 shows the Weka application and the decision tree constructed. Figure 6.7 shows a simplified view of the decision tree constructed to enable a better interpretation of the results. The most important attribute (the root node) was `_ST` and represents the number of supporting teams involved in resolving an incident. When `_ST` is 3, 4, 5, or 6, incidents are typically not closed. On the other hand, when `_ST` has values 1 or 2, the attribute `_Status` is used to predict if an incident will be closed or not. In both cases, if the number is equal to 4, incidents are likely closed. Furthermore, for `_ST` = 1 and `_Status` = 2, incidents are also closed.

Accuracy Evaluation

Table 6.5 shows a summary of the evaluation of the classification tree created. A high number of instances (incidents) were correctly classified: 92.5 %. Precision and the F-measure were higher for incidents that were not closed; precision was almost 100 %. For incidents that were closed, the precision was of approximately 80 %.

The F-measure (or F-score) is calculated based on the precision and recall. The calculation is as follows:

$$Precision = t_p/(t_p + f_p)$$
$$Recall = t_p/(t_p + f_n)$$
$$F - score = 2 * Precision * Recall/(Precision + Recall)$$

Where t_p is the number of true positives, f_p the number of false positives, and f_n the number of false negatives. Precision is defined as the fraction of elements correctly classified as

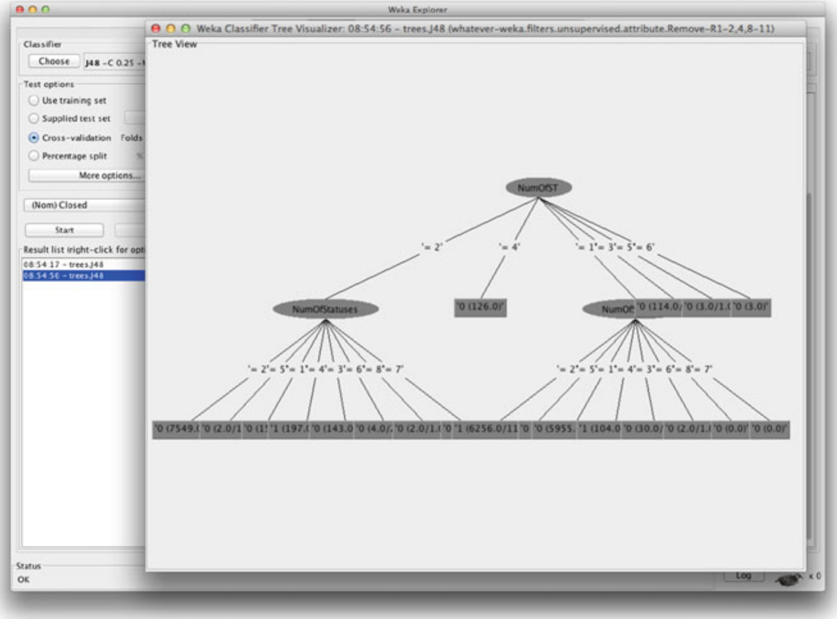


Fig. 6.8 Visualization of the decision tree in Weka

Table 6.5 Summary of the evaluation of the decision tree

Stratified cross-validation (summary)

```
Correctly Classified Instances       19119               92.5904 %
Incorrectly Classified Instances      1530                7.4096 %
Kappa statistic                          0.8207
Mean absolute error                      0.1205
Root mean squared error                  0.2462
Relative absolute error                 30.9035 %
Root relative squared error             55.7619 %
Total Number of Instances            20649

=== Detailed Accuracy By Class ===

              TP Rate   FP Rate   Precision   Recall  F-Measure  ROC Area   Class
              0.916     0.046     0.982       0.916   0.948      0.944      0
              0.954     0.084     0.804       0.954   0.872      0.944      1
Weighted Avg. 0.926     0.056     0.935       0.926   0.928      0.944

=== Confusion Matrix ===

    a     b   <-- classified as
13891  1276 |     a = 0
  254  5228 |     b = 1
```

positive out of all the elements the algorithm classified as positive, whereas recall is the fraction of elements correctly classified as positive out of all the positive elements.

Interpretation of the Model

An interesting operational insight from the constructed model, shown in Fig. 6.7, is that the most relevant factor predicting that an incident will be closed is the number of support teams (`_ST`) involved in its resolution. In other words, a lower number of support teams involved in solving an incident increases its likeliness to be closed. This insight is supported by analyzing the `_Status`. From the 10.699 instances analyzed, more than 6200 incidents were closed when only one support team was involved. Two hypotheses can be raised:

- Complex incidents require several support teams to be resolved and, thus, are more likely to remain open.
- Incidents that require several support teams introduce communication noise and overhead, which make their resolution less likely.

In both situations, the IT service provider can be recommended to foster the communication and coordination across teams. A notification mechanism can be developed to alert the service desk when an incident is involving more than two teams to resolve an incident as it is likely that the incident will remain unsolved. Management actions such as training and active monitoring can be taken to improve incident closure rate.

6.3.5 Identifying Behavioral Patterns

The next study conducted explored three behavioral patterns that are typically associated with incident management service provisioning: wait-user, push-to-front, and ping-pong.

- Exaggerated use of the status *wait-user*. Incidents must be resolved as soon as possible. Support teams are evaluated based on their timely response. However, when it is necessary to contact and wait for an answer from a customer, the waiting period (labeled as `Status = wait-user`) is not accounted to evaluate the efficiency of a team. As a result, teams can abuse this phenomenon to increase their evaluation.
- Low occurrence of *push-to-front*. Three lines of support exist for resolving incidents. Each has a higher degree of specialization from the 1st to the 3rd. Thus, 1st line teams can easily be tempted to transfer incidents to 2nd or 3rd line teams to reduce their workload, even though they could be handled at the 1st level. This phenomenon lowers the overall efficiency of the IM service.
- Occurrence of the *ping-pong* effect. Sometimes, because of the difficulties associated with handling certain incidents, the teams responsible for their resolution can transfer incidents to other units or levels, and these, in turn, can return incident in a vicious cycle that does not allow for progress in their resolution.

This phenomenon should be identified and avoided to increase the efficiency of the IM service.

All these phenomena have an impact on service provisioning. Association rules can be used to analyze the incident management dataset to determine if these behaviors occur.

Since good literature on data mining exists, the reader is refereed to the work from Witten et al. [14] and Han et al. [7] to acquire the necessary knowledge on basic concepts underlying association rules—such as the definition of a rule, its support and confidence, and how patterns are mined—to understand this section.

Association Rules

Association rule mining is an approach that finds frequent co-occurring associations among a collection of elements. The goal is to find associations of elements that occur together more often when compared to randomly sampling of all possibilities. An example of this is: when an incident report involves product `PROD34`, the support team `G45` is most often involved in its resolution.

Suppose that the dataset includes the following information:

- There are 60,000 incident records in total.
- 750 records contain `PROD34` (1.25 %).
- 6000 records contain `G45` (10 %).
- 600 records contain both `PROD34` and `G45` (1 %).

If there was no association between `PROD34` and `G45` (i.e., they are statistically independent), then it is expect that only 10 % of `PROD34` incidents to be handled by `G45` (since 10 % of all records have `G45`). However, 80 % (600/750) of `PROD34` incidents are handled by `G45`. This is a factor of 8 increase over what was expected—that is called *lift*, which is the ratio of the observed frequency of co-occurrence to the expected frequency. This is determined by counting the records in the dataset. So, in this case, the association rule would state that incidents involving `PROD34` also involve `G45` with a *lift* factor of 8.

In statistics, *lift* is estimated by the ratio of the joint probability of two items x and y (i.e., $P(x, y)$), divided by the product of their individual probabilities (i.e., $P(x)$ and $P(y)$):

$$lift = \frac{P(x, y)}{[P(x)P(y)]}$$

If the two items are statistically independent, then $P(x, y) = P(x)P(y)$, corresponding to $lift = 1$. Note that anti-correlation yields $lift < 1$, which is also an interesting discovery—corresponding to mutually exclusive items that rarely co-occur together. For the incident management service, the objective is to find instance records with a $lift > 1$.

Table 6.6 The new attributes added to the original dataset

Attribute	Description
`_Substatus`	The number of different substates an incident was.
`_InvolvedST`	The number of support teams involved in resolving an incident.
`_OrgLines`	The number of fields of the organization involved in resolving an incident.
`_OrgLineA2`	Binary attribute indicating if the organizational line A2 was involved in resolving the incident.
`_OrgLineC`	Binary attribute indicating if the organizational line C was involved in resolving the incident.
`_Wait-User`	Binary attribute indicating if the incident was placed in the wait-user state.
`_Push-to-Front`	Binary attribute indicating if the incident was pushed to front.
`_Ping-Pong`	Binary attribute indicating if teams resolving the incident report exhibited a ping-pong behavior.

Training and Test Set

To extract more meaningful business rules, new attributes were added to the original dataset. A Matlab application was created to read the original dataset and transform its content to contain eight additional attributes (see Table 6.6):

`_Substatus`, `_ST`, `_OrgLines`, `_OrgLineC`, `_OrgLineA2`,
`_Wait-User`, `_PushFront`, `_PingPong`

To account for the valuable information associated with support teams, organizational lines, and substatus of each record, these original attributes were replaced by quantitatively evaluating their value, instead of qualitatively as they were in the original dataset (this approach was already followed in the study carried out in the previous section).

Furthermore, several attributes present in the original dataset were removed. `Status` was removed since the `Substatus` attribute includes the needed knowledge of the incidents' states. `Owner country` was also removed since it provides the same information as the information conveyed by the `Country` attribute. Finally, attribute `Owner first name` refers only to the owners of the incidents which was too specific for our analysis.

The main organizational lines, `A2` and `C`, were captured by creating two new attributes, `_OrgLineC` and `_OrgLineA2`, respectively, to indicate if an incident was handled by these organizational lines.

All records were analyzed individually to determine if the attribute Substatus had the value wait-user. When this condition was verified, the new attribute _Wait-User was set to true, otherwise it was set to false.

The records of the dataset were also scanned to determine if an incident was handled by the 2nd or 3rd line by analyzing the attribute Involved ST of the original dataset. When this condition was verified, the new attribute _Push-to-Front was set to true, otherwise it was set to false.

Finally, the ping-pong phenomenon was identified by setting the value of the new attribute _Push-to-Front to true. This occurred when an incident was initially handled by a particular support team, then sent to another support group, which would send it again to the initial support group. The ping-pong behavior was only identified when the incident report transferred in such a cycle retained its value for the attribute Substatus. If this attribute value changed in a cycle, it meant that productive work was performed.

Model Construction

To find association rules in the incident dataset, the *apriori* algorithm [14] was used. It is an influential algorithm for mining frequent itemsets to construct Boolean association rules. The algorithm uses a bottom up approach, where frequent subsets are extended one item at a time. It was designed to operate on datasets containing records, instances, or transactions.

The apriori algorithm implemented in Weka was used with a minimum support of 55 % and a minimum confidence of 90 %. The results generated various rules shown in Table 6.7.

Interpretation of the Model

The results show that a strong relationship exists between the phenomenon push-to-front (handling as many incidents in the 1st line as possible) and ping-pong (sending incidents back and forth between support teams before they are resolved). The following rule expresses this relationship:

$$\_Push-to-Front = 0(4545) \implies \_Ping-Pong = 1(4393)$$

It identifies that the occurrence of push-to-front generally implies a ping-pong phenomenon with a confidence of 97 %. This confidence is the ratio between the number of incidents in which the rule is correct (4.393) and the number of incidents in which the rules is applicable (4.545).

One recommendation can be made based of the association rules found. If the IT service provider concentrates its efforts in alleviating push-to-front, it will also be able to alleviate the ping-pong. This will optimize IT service operations.

Table 6.7 Examples of instances of the dataset

Associator model (full training set)

```
Apriori
=======

Minimum support: 0.55 (4154 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 9

Generated sets of large itemsets:
Size of set of large itemsets L(1): 6
Size of set of large itemsets L(2): 10
Size of set of large itemsets L(3): 5
Size of set of large itemsets L(4): 1

Best rules found:

1. _OrgLines=1 _OrgLineC=1 4416 ==> _OrgLineA2=0 4416              conf:(1)
2. _OrgLines=1 _OrgLineC=1 _Ping-Pong=0 4246 ==> _OrgLineA2=0 4246 conf:(1)
3. _OrgLineA2=0 _Push-to-Front=0 4270 ==> _Ping-Pong=0 4175       conf:(0.98)
4. _Push-to-Front=0 4545 ==> _Ping-Pong=1 4393                    conf:(0.97)
5. _OrgLines=1 _OrgLineA2=0 4966 ==> _Ping-Pong=0 4779            conf:(0.96)
6. _OrgLines=1 _OrgLineC=1 4416 ==> _Ping-Pong=0 4246             conf:(0.96)
7. _OrgLines=1 _OrgLineA2=0 _OrgLineC=1 4416 ==> _Ping-Pong=0 4246
                                                                  conf:(0.96)
8. _OrgLines=1 _OrgLineC=1 4416 ==> _OrgLineA2=0 _Ping-Pong=0 4246
                                                                  conf:(0.96)
9. _OrgLines=1 5519 ==> _Ping-Pong=0 5284                         conf:(0.96)
10._Push-to-Front=0 _Ping-Pong=0 4393 ==> _OrgLineA2=0 4175       conf:(0.95)
```

6.4 Clustering, Regression, and Text Mining

This section reviews three additional methods that can be used to study and improve service delivery. Namely, cluster analysis, regression analysis, and text mining.

6.4.1 Cluster Analysis

Cluster analysis is the process of finding groups of objects such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups. The goal is to have high homogeneity within clusters and high heterogeneity between clusters. For example, cluster analysis is useful to group similar customers when preparing advertisement campaigns. This is often called market segmentation.

► **Definition (Clustering)** An important task in data analysis that has the objective to group a set of objects so that objects in the same group are more similar to each other than to those in other groups.

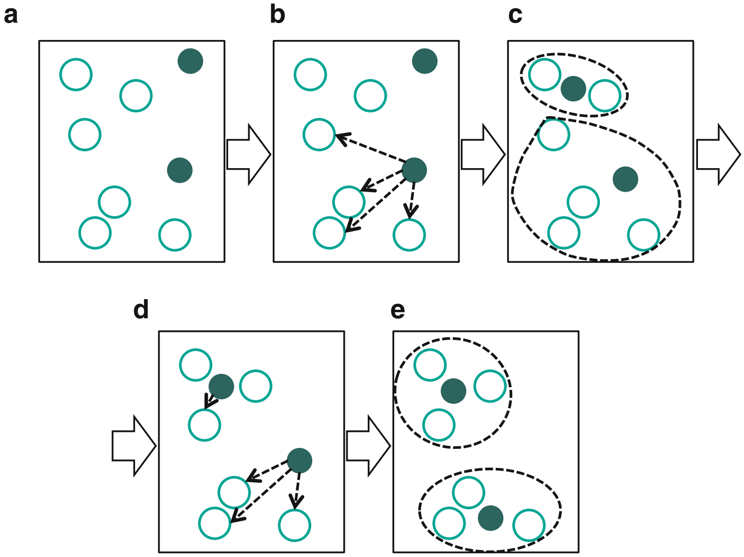


Fig. 6.9 Clustering a set of objects using k-means **(a)** Choose objects as the initial cluster centers. **(b, d)** (Re)assign each object to the closest cluster. **(c)** Update the centroids of the clusters. **(e)** Repeat the procedure until the cluster centroids do not change

Clustering methods are generally classified into the following categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. In this chapter, the focus is on the most popular methods: partitioning and hierarchical methods.

The k-means Clustering

The *k-means* method is a partitioning approach, where each cluster is associated with a centroid (center point); each point is assigned to the cluster with the closest centroid, and the number of clusters k is pre-defined. Given a set of objects and a number of clusters k, the algorithm is very simple (Fig. 6.9):

- Step 1. Choose arbitrarily k objects as the initial cluster centers (a).
- Step 2. (Re)assign each object to the closest cluster based on the mean value of the objects in the cluster (b, d).
- Step 3. Update the centroids of the clusters (c).
- Repeat Steps 2 and 3 until the cluster centroids do not change (e).

The initial centroids are often chosen randomly. The clusters are prone to vary from one run to another. The centroid is (typically) the mean of the points in the cluster. The *closeness* is computed by Euclidean distance, cosine similarity, or other related measures. K-means converges for common similarity measures, where most of the convergence usually happens in the first few iterations. It is often the case that the stopping condition is replaced by “until relatively few points do not change clusters”.

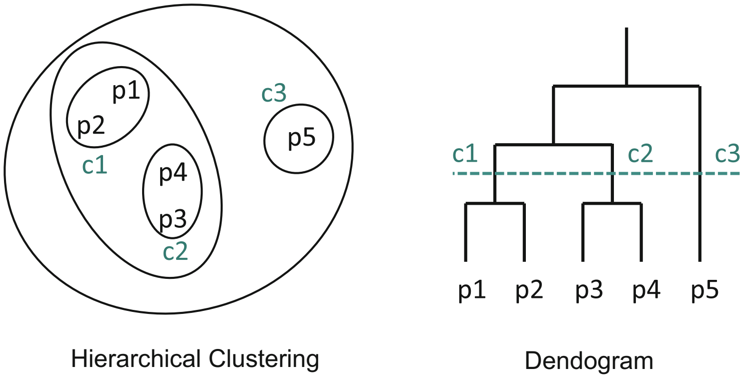


Fig. 6.10 Clustering of objects based on hierarchical methods

Hierarchical Clustering

Hierarchical clustering methods work by grouping data objects into a set of nested clusters organized as a hierarchical tree. The clusters can be visualized as a dendogram: a tree like diagram that records the sequences of merges or splits. Hierarchical clustering operates top-down (divisive) or bottom-up (agglomerative). In a top-down fashion, one starts with a large cluster, which is divided until each element is assigned to a cluster. In a bottom-up fashion, one starts with one element per cluster and aggregates clusters together until a global cluster is constructed. Figure 6.10 shows representations of these clustering methods.

Agglomerative clustering algorithms are one of the most popular hierarchical clustering techniques. The basic procedure is the following:

- Compute the distance matrix between the input data points.
- Let each data point be a cluster.
- Repeat
  – Merge the two closest clusters.
  – Update the distance matrix.
- Until only a single cluster remains.

The key operation of this method is the computation of the distance between two clusters. Different distances can be used: maximum distance between elements of each cluster (*complete-linkage* clustering); minimum distance between elements of each cluster (*single-linkage* clustering); and mean distance between elements of each cluster (*average-linkage* clustering).

The final result is a grouping of all the objects in a single cluster, and, of course, the constructed tree hierarchy (dendogram). To generate partitioning clusters, the tree is cut at a given level.

6.4.2 Regression Analysis

The prediction of continuous values can be performed with models that are based on statistical *regression*. Consider the following goal: Determining how fast (time) incident are handled successfully is very important to assess the efficiency of IT service providers. The variables in this case are time and incidents. The goal is to predict the duration of repairs for a given number of incidents (see [22] for a detailed discussion of this type of IT service).

The statistical task is to predict a value of a given continuous-valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency. Regression can be used. It is the model explaining the variation in a dependent variable (Y) using the variation in independent variables (X). If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction. The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

► **Definition (Regression)** Regression is a data analysis task that can be used to predict a number. It creates a linear model from existing data (e.g., from sales, house values, or temperature), and uses the model to predicted unknown occurrences.

A linear regression model defines a line that denotes the relationship between X and Y. The variables may have different levels of correlation, thereby X and Y might be related to each other in many ways: linear or curved. As such, there are distinct methods for linear regression and nonlinear regression, accordingly.

Linear regression fits a straight line to the data. A linear regression model related to the example is depicted in Fig. 6.11. It shows the relation between the dependent variable *time* (in minutes) and the number of incidents, which is the independent variable. Even though the points of the graph do not fall on a straight line, the pattern suggests a linear correlation between the variables.

Other examples of real-world applications of regression include predicting sales of new products based on advertising expenditure, predicting the number of patients with high risk of heart stroke expected to come in the hospital over the weekend to schedule staff, or predicting number of IT incidents expected during the weekend to plan for support technicians.

6.4.3 Text Mining

Text mining is the analysis of data represented in natural language text. Text analytics consists in mining data contained in unstructured, natural language text in order to discover new knowledge and, accordingly, make decisions to solve problems.

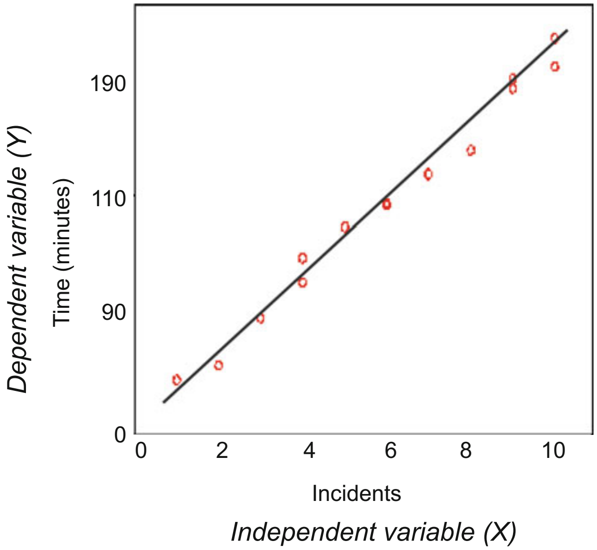


Fig. 6.11 The scatter plot of data suggests a linear relationship between time and the number of incidents

Through an iterative approach, organizations can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, relevance, and intensity.

► **Definition (Text Mining)** A process by which software applications extract specific high-quality information from a large amount of text, usually by recognizing specific terms.

Text mining methods are typically grouped into two main categories: text analysis and document retrieval. Retrieval methods consist in finding from a large corpus of text documents those that are more relevant to a specified query, i.e. user demand. Most prominent examples are web search and retrieval in library catalogs. For example, by analyzing the incident reports submitted to an IT service provider as a corpus of textual documents, text retrieval becomes useful when a technician is searching with a particular query, e.g., incidents related to printers, and the most relevant reports to that query are to be given back as results.

Software that provide text analytics typically transpose words and phrases of the unstructured data representation into numerical values, which can then be analyzed with traditional mining techniques. Important tasks usually performed are document topic detection, text categorization, document clustering, sentiment analysis, and text summarization.

6.5 Tools and Software

The following is a list of tools and software that contain a variety of analytics and machine learning algorithms:

- R and RStudio. A software environment for statistical computing and data mining. It has support for a wide variety of data mining and statistical algorithms, which includes classification, time series analysis, clustering, and linear and non-linear modeling.
- SPSS. A software package suitable for beginners since it is easy to use. It has a "point and click" user interface with commands available via drop-down menus. It performs most general statistical analyses (regression, logistic regression, survival analysis, analysis of variance, factor analysis, and multivariate analysis).
- SAS. In contrast to SPSS, SAS targets power users which can use the application programmability. Users typically write SAS programs that manipulate data and perform data analyses. As with SPSS, it also provides general statistical analyses.
- Mahout. a scalable machine learning and data mining library suitable for Big Data processing. It supports different algorithms such as classification, regression, and recommendation.
- MALLET. A Java based machine learning package for document classification, topic modeling, information extraction, and sequence labeling applications.
- Weka. An open source machine learning software library in Java which includes pre-processing, regression, classification, clustering, and association rules. Figure 6.8 illustrates the visualization of a decision tree generated in Weka.
- Graphlab. A scalable and distributed machine learning library that supports graph-based machine learning algorithms.
- Encog. An advanced machine learning framework that supports a variety of advanced algorithms, as well as methods to normalize and process data. It is available in Java, C++, and .NET.
- Rapid Miner. An open source machine learning in Java, which supports a wide range of machine learning algorithms.
- d3js. Chord diagrams for visualization based on the D3.js javascript library.
- Gephi. An open-source software for visualizing and analyzing network graphs.
- Tableau. A commercial software used for data visualization. The software is suitable for improving the process of finding patterns in data.

Other, more common, tools which can be used include Microsoft Excel for processing of the raw datasets and to explore processed data. Excel is especially helpful for implementing basic and intermediate mathematical functions.

6.6 Conclusions

The large volumes of data available today, generated from service operations, hold implicit information that can be translated into useful knowledge for making intelligent business decisions. In fact, IT service provisioning is often managed by sophisticated information systems, which monitor and log all the activities needed to deliver services with agreed service levels.

Service analytics provides enterprises with powerful mechanisms to convert these logs into in consistent datasets to be analyzed. Extracted insights are important to understand all aspects of service operations to take actions to improve organizational performance and increase customer satisfaction. Service provisioning can be studied to identify human behavior in service intensive organizations at the individual, work group, and organizational levels. Delivery can identify how staff, information systems, and customers arrange themselves to co-create services. Finally, consumption looks into the role of time and place in service delivery.

Existing algorithms from the field of data mining, e.g., classification, association rules, regression, and clustering, can be used with little adaptation and effort. Naturally, in many situations new algorithms must be developed when specific insights need to be extracted.

Review Section

Review Questions

1. Briefly describe the main characteristics that distinguish service analytics from the general data analytics paradigm.
2. This chapter defines three levels under which service analytics methods fall based on the action performed with the discovered knowledge. Briefly describe these levels and give example of analytics methods that fall under each of them.
3. What are the differences between classification and prediction methods?
4. What is cluster analysis? What is a cluster? List the main differences between partitioning methods and hierarchical methods for cluster analysis. Give examples in each method. Describe a scenario in which application of clustering to service systems.
5. Describe the characteristics of the regression method. Define a scenario in which regression can be applied to service systems. How is regression different from cluster analysis?
6. Describe the main types of text mining methods relevant for services. Give one example for each of the following cases: (a) an application that uses document categorization techniques on the data generated by the consumption of a particular service and (b) an application that uses sentiment analysis methods on the data generated by the consumption of a particular service.

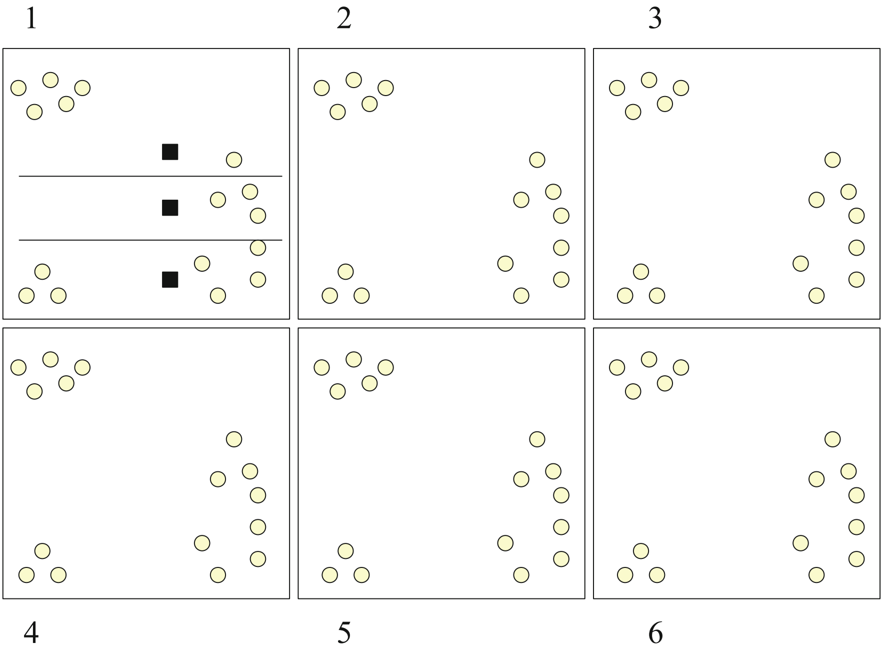


Fig. 6.12 Step-by-step simulation of the k-means clustering algorithm

Project

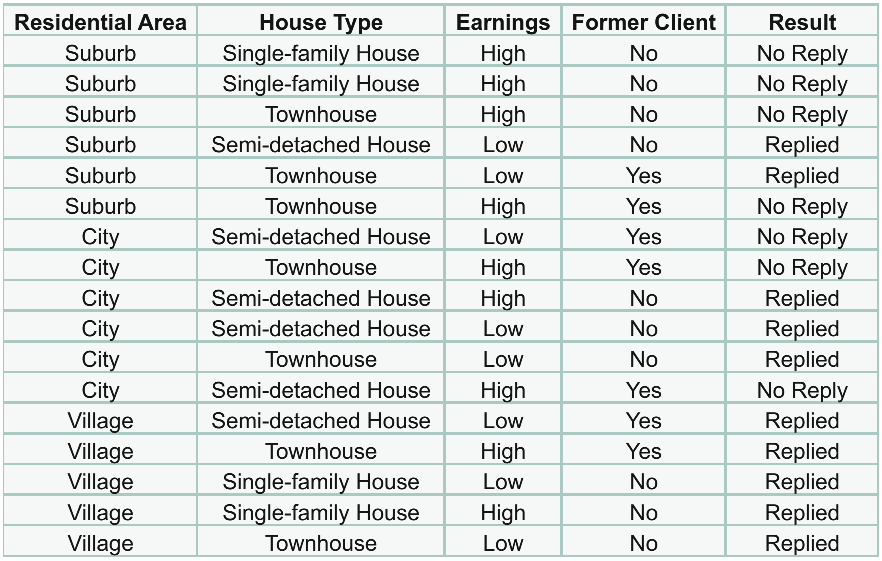
Clustering

Consider the dataset from Fig. 6.12 as a graphical representation of IT service incident (i.e., each point represents an incident). The company needs to cluster incidents by applying the k-means method in this dataset for the first six iterations. The circles are data points. The initial centroids are given as black rectangles. The separating lines show the clusters boundaries in the first iteration.

1. How do the centroids and boundaries look like in the following iterations? Does the method terminate?
   Hint: this is a quality sketch to demonstrate the main principle of clustering.
2. Show graphically an example that applies k-means algorithm, which for different initializations gives different results.

Decision Tree

Figure 6.13 describes the results of a marketing campaign for a real-estate service. It contains information for each potential client regarding the residential area where he or she lives, the type of household, his earnings, and whether he was a former



Residential Area	House Type	Earnings	Former Client	Result
Suburb	Single-family House	High	No	No Reply
Suburb	Single-family House	High	No	No Reply
Suburb	Townhouse	High	No	No Reply
Suburb	Semi-detached House	Low	No	Replied
Suburb	Townhouse	Low	Yes	Replied
Suburb	Townhouse	High	Yes	No Reply
City	Semi-detached House	Low	Yes	No Reply
City	Townhouse	High	Yes	No Reply
City	Semi-detached House	High	No	Replied
City	Semi-detached House	Low	No	Replied
City	Townhouse	Low	No	Replied
City	Semi-detached House	High	Yes	No Reply
Village	Semi-detached House	Low	Yes	Replied
Village	Townhouse	High	Yes	Replied
Village	Single-family House	Low	No	Replied
Village	Single-family House	High	No	Replied
Village	Townhouse	Low	No	Replied

Fig. 6.13 Dataset with the results of a marketing campaign

client of the company. Use this data to predict whether a person is going to react positively (i.e., send a reply) to an advertisement of the service. Build a decision tree in order to make this prediction.

Key Terms

Analytics Data management and reporting methods which are a prerequisite for advanced analytics built on methods from statistics and operations research.

Service Analytics Analytics applied to service systems with the goal to generate the highest benefits for all stakeholders according to the value co-creation principle.

Cluster Analysis The process of finding groups of objects such that the objects in a group are similar (or related) and different from (or unrelated to) objects in other groups.

Classification and Prediction Forms of data analysis that can be applied to build models that describe important data patterns or predict future data trends.

Decision Trees Supervised learning method used for classification. A decision tree is an upside down tree structure, which is used to classify an unknown instance by testing its attribute values along the path from the root node to leaf nodes.

Regression Analysis A statistical technique used to predict a continuous dependent variable from a number of independent variables.

Text Mining The analysis of unstructured natural language text to obtain insights.

Further Reading

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning.* Springer, 2013.

Ian Witten, Eibe Frank, and Mark Hall. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2011.

Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly, 2013.

References

1. Van Bon J, de Jong A, Kolthof A (2007) Foundations of IT Service Management based on ITIL. Van Haren Publishing, Zaltbommel. ISBN 9789087530570
2. Fromm H, Bloehdorn S (2014) Big data - technologies and potential. In: Enterprise integration, Chap 9. Springer, Berlin, pp 107–124
3. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17(3):37
4. Terry K (2013) Analytics: the nervous system of IT-enabled healthcare. Institute for Health Technology Transformation. Report http://ihealthtran.com/analyticsreport.html
5. Groves P et al (2013) The Big Data revolution in healthcare. McKinsey & Company
6. Fromm H, Habryn F, Satzger G (2012) Service analytics: leveraging data across enterprise boundaries for competitive advantage. In: Globalization of professional services. Springer, Berlin, pp 139–149
7. Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques. The Morgan Kaufmann series in data management systems. Morgan Kaufmann, Los Altos, CA
8. Srivastava J et al (2000) Web usage mining: discovery and applications of usage patterns from web data. ACM SIGKDD Explor Newsl 1(2):12–23
9. Kohavi R, Rothleder N, Simoudis E (2002) Emerging trends in business analytics. Comm ACM 45(8):45–48
10. Davenport T (2006) Competing on analytics. Harv Bus Rev 84(1):98
11. Davenport T, Harris J (2007) Competing on analytics: the new science of winning. Harvard Business Press, Watertown, MA
12. Kobielus J (2010) The Forrester wave predictive analytics and data mining solutions, Q1 2010. Forrester Research, Cambridge, MA
13. Chaudhuri S, Dayal U (1997) An overview of data warehousing and OLAP technology. ACM SIGMOD Rec 26(1):65–74
14. Witten I, Frank E, Hall M (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Los Altos, CA
15. Wilson H, Keating B (2008) Business forecasting with business ForecastX, 6th edn. McGraw-Hill/Irwin, New York, 513 pp
16. Hanke JE, Wichern D (2008) Business forecasting, 9th edn. Prentice Hall, Englewood Cliffs, 576 pp
17. Analytics (2015) INFORMS online. http://www.informs.org/Community/Analytics. Accessed: 2015-10-17

18. Gerke K, Cardoso J, Claus A (2009) Measuring the compliance of processes with reference models. In: 17th international conference on cooperative information systems (CoopIS 2009). Springer, Algarve
19. OGC (2007) ITIL service operation. ITIL Series. Stationery Office ISBN: 978-0113310463
20. Paszkiewicz Z, Picard W (2013) Analysis of the Volvo IT incident and problem handling processes using process mining and social network analysis. In: van Dongen B et al (eds) CEUR online proceedings, 2013. Proceedings of the 3rd business process intelligence challenge co-located with 9th international business process intelligence workshop (BPI 2013)
21. Hall M (1998) Correlation-based feature subset selection for machine learning. Ph.D. thesis. University of Waikato, Hamilton
22. Cardoso J, Lopes R, Poels G (2014) Service systems: concepts, modeling, and programming. Springer, Berlin