Machine-learning technologies in telecommunications

Operators can employ machine-learning techniques to exploit user, network and traffic data assets to better understand their subscriber base and to analyze network traffic and media files. They may also apply these capabilities to boost services or to identify why users do not adopt them.

MARTIN SVENSSON
JOAKIM SÖDERBERG

Operators have a vast amount of assets in the form of user, network, and traffic data. Daily human interaction with telecommunication networks and services of all kinds contain unconnected gems of information which, when correctly interpreted, reveal how people behave and interact. ML technologies can be put to good use to reveal how all this information fits into relevant patterns.

ML technologies have finally reached a mature stage – the associated costs and learning curves are feasible; therefore, the industry is adopting them in a wide array of applications. Operators who are looking for new and smart ways of sharpening their service offerings and of strengthening their role in the service-provider ecosystem would be wise to make the most of these opportunities.

From a broad perspective, machine learning is about giving software the ability to build knowledge from experience.[1] This ability is derived from the

- patterns and rules extracted from large volumes of data – useful knowledge is extracted from available network data and media; or
- logic or reasoning assigned to a set of general rules in order, for example, to draw conclusions and automate a process.

At the core of ML technologies is a toolbox of algorithms that can be trained to adapt a particular function or model. Specialized algorithms exist to analyze social networks, discover user segments, identify potential churners and fraudsters, analyze media files, and so on. In every instance, the technologies learn by experience. Therefore, to reach a particular goal, they must be trained using prepared data sets. The results are then presented to people or applications that can provide feedback, which leads to further training to refine what has been learned.

Consider the example in Table 1 where the learning process is guided or supervised by providing sample data and associated categories. If we use this training data set to devise a typical ML algorithm that is to learn about the concept of a geometric figure (for example, a square or triangle), the algorithm will probably conclude that triangles have 180 degrees and squares have 360 degrees.

In general, the more samples provided to train an algorithm, the better it will learn. However, overly large input data sets can impede learning because too much data introduces noise or bogs down the process. To make the system learn properly, one should thus emphasize supplying the most significant training samples.

One other observation is that once an algorithm has been trained to learn a concept, it must be tested or verified. This is generally accomplished by having the algorithm classify unseen samples or a test data set (Table 2).

One can determine the performance of an algorithm by measuring how accurately it recognizes the geometric figures from the examples in the test data set. The example above corresponds to the supervised learning (or classification) category, since the correct answer is provided in the learning process.

If, on the other hand, the classes to which the samples belong are not available, the problem falls under the unsupervised learning (or clustering) category. For such problems, the ML algorithm tries to identify the features that samples have in common – in order to group them in different clusters. These features are what describe a sample. The feature in Table 2, for example, is "degrees." Therefore, using Tables 1 and 2 as input, the algorithm would

BOX B
Machine learning

Machine learning (ML) is concerned with the design and development of algorithms and techniques that allow computers to "learn." The major focus of ML research is to extract information from data automatically, by computational and statistical methods. It is thus closely related to data mining and statistics as well as theoretical computer science.

TABLE 1 **Training data set for a supervised learning process.**

Example	Degrees = Feature	Figure = Class
1	180	Triangle
2	360	Square
3	180	Triangle
4	180	Triangle
5	360	Square

TABLE 2 **Test data set for verifying a learned concept.**

Example	Degrees = Feature	Figure = Class
6	180	?
7	360	?

probably identify one group consisting of samples 1, 3, 4 and 6 (180°) and one other group consisting of samples 2, 5 and 7 (360°).

Business opportunities for machine learning

Research in machine learning and innovation in algorithms give operators and suppliers an opportunity to grow existing offerings as well as to find and develop new ones. Three general business scenarios where ML technologies can be used to considerable advantage are recommendations, personalization and media recognition.

Business scenario 1: Recommendations

Recommendation services are based on the analysis of captured network data to provide hints about a user's preference for certain kinds of information. In terms of services, this entails identifying the services that would interest a user either generally or in a specific situation. To date, recommendations have most commonly been used to suggest media, such as music, movies and books.

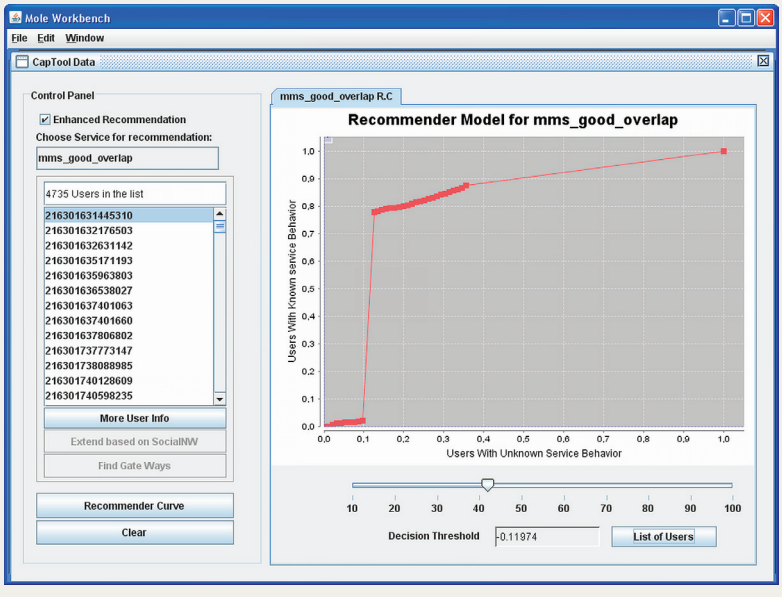
Service recommender

For a given service, a service recommender identifies users with high probability for uptake. It does so by looking at similarities among service-usage patterns **(Figure 1)**.

Following these same principles, service recommenders may also be used to boost existing services or to identify why users do not adopt some services.

In addition, service recommenders can be used to predict churn – using information derived from the usage patterns of past churners, one can detect changes in other usage profiles.

FIGURE 1 **A service recommender for MMS usage. The behavior of users listed on the left is similar to that of MMS users. The curve on the right characterizes the model accuracy for a human expert.**



Business scenario 2: Personalization

Personalization is based on adapting operator services and offerings to the needs and preferences of the customer base. Today this is done by hand, by interviewing a subset of the customer base in order to understand what kinds of customers an operator has. However, by applying ML techniques to customer-generated network data, one can automate the process of grouping customers into segments; for example, by assigning a profile to customers according to calling and messaging behavior **(Figures 2–3)**.

Services can be personalized to have a different appearance to different types of users – even before a user has interacted with the service in question (first-time personalization).

What is more, operators can reduce spam by personalizing their offerings by directing adequate marketing to users with a certain profile. Likewise, they can tailor their overall offerings, such as types of subscriptions, to user profiles.

Personalized ads

A typical business scenario is targeted advertising. The idea is to present (or produce) advertisements that are tailored to an individual, situation and device. Instead of using coarse-grained segmentation models to target the customer base, operators and advertisers can zero in on customers with ads that fit their needs and interests.

One other opportunity arises from the ability to use all available information about a user and his or her social environment to personalize the service experience itself. To be effective, one must successfully capture the "typical users" associated with a given service.

Business scenario 2: Media recognition

Recognition applies to all kinds of patterns in pictures, sound, video and traffic bursts. The goal is to learn to classify

known patterns and then use this information to identify the category to which previously unseen patterns belong.

Photo tagging

Operators can enable a community of users to share image metadata. In return, users get access to an efficient image-classification system that, thanks to annotations by other subscribers, has learned to recognize a large set of different images.

This practice, known as collaborative indexing, describes the process by which many users add metadata (keywords) to shared content. The result is often an instantaneous and rewarding gain in users' ability to find related content.

Mobile tagging

Mobile phones can be equipped with a software function that facilitates photo tagging by detecting images and objects. After a user has taken a photo, the automatic tagging function in the phone predicts the class of the photo – for example, outdoor, family, cityscape – and suggests an annotation keyword from a residential vocabulary. The user can accept the keyword, refine it, or create a new one. The annotation is stored in a common repository together with a visual descriptor that represents the photo.

Photo browsing

In a managed network that enables cross-platform communication, users can upload photos to a private digital media server. A photo-browsing application running on a set-top box can then make use of the annotation repository to let users browse photos by categories on their TV at home, for example.

Architecture entities and technologies

Architecture entities and technologies are sometimes referred to as data mining, even though this term predates machine learning and overlaps with purely statistical techniques.

A conceptual architecture for data mining

A data-mining system architecture can be modeled in many different ways, but in essence most suggestions follow the same basic scheme, see for instance the CRISP-DM model (cross-industry

FIGURE 2 Understanding your customer base: A user profiler increases the probability that a user ID will belong to one of many predefined user categories.

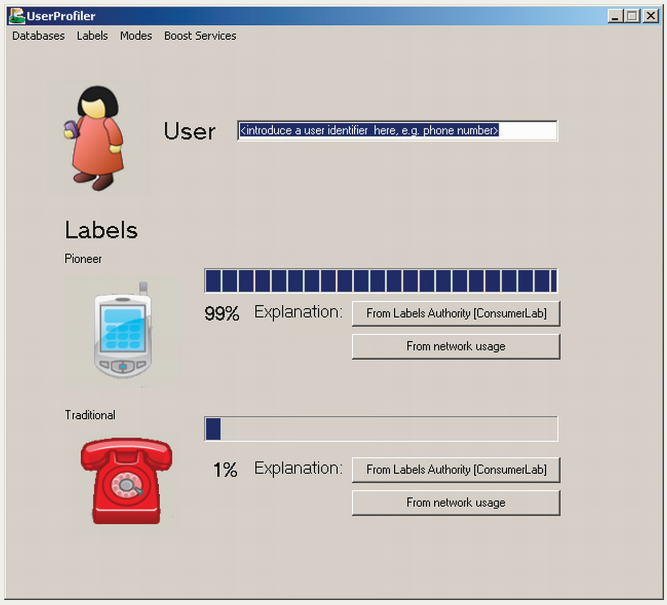


FIGURE 3 A user category can be inferred from network data. An explanation in marketing terms might not always exist but can be provided after a new category has been discovered.

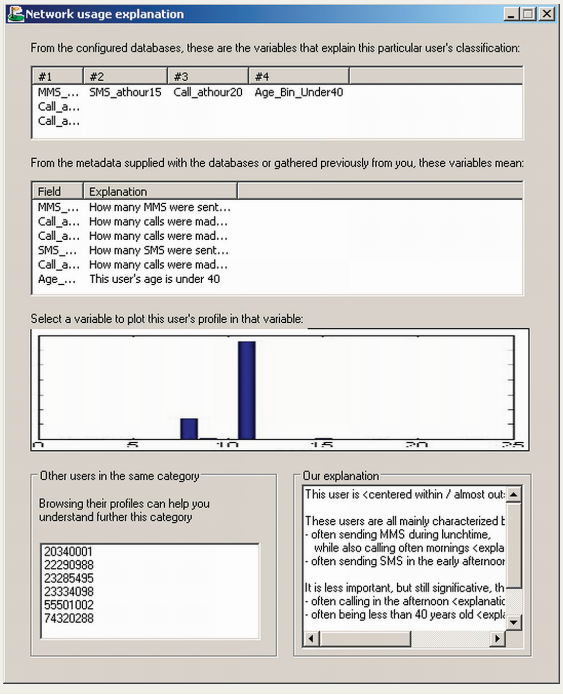
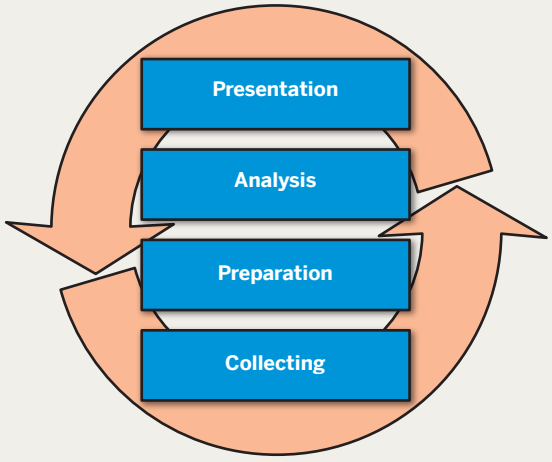


FIGURE 4 Conceptual architecture. Note how the data-mining process is highly iterative.



standard process for data mining, www.crisp-dm.org).

A conceptual architecture has the following components:

- A data source collection tier is responsible for accessing data sources, storing raw data, and cleaning data. This component, which Ericsson has studied in depth for handling call detail records (CDR) and network logs, has its roots in the traditional extraction, transformation and load (ETL) tools developed by the IT industry.
- A data preparation tier is responsible for preparing (aggregating) data for analysis by the ML component. This tier is also responsible for removing redundant information and applying domain-specific knowledge. In other words, preparation goes a step beyond mere cleaning, by using domain knowledge that is directly related to the business issues that need to be resolved. For example, the data must be aggregated in one way or other, depending on how the system classifies the use of different services (hourly, daily, weekly, and so on).
- An analysis tier is responsible for applying the appropriate ML algorithms to data in order to train models that will be used to solve concrete business questions.
- A presentation tier is responsible for outputting derived knowledge in a suitable manner; for example, using advanced visualization tools.

Although the conceptual architecture is tiered, the data-mining process is highly iterative. Therefore, before arriving at a satisfactory answer to a use case many of the steps outlined above must be revisited countless times **(Figure 4)**.

Telecom-grade data mining

Machine learning has been used for some time in academia and various industries, such as the financial and retail sectors. The methods and algorithms are generic, but their choice and adaptation is often specific to a given domain. Telecommunications is no exception. To get the most from available data assets, suppliers must design their systems specifically for the domain, and operators must become very adept at using them. In telecoms, in particular, the sources of raw data are heterogeneous, very detailed, and produce huge amounts of data. What is more, the networks are constantly evolving with new protocols, new nodes, and new services, all of which add to the amount of raw data (and give rise to new opportunities). Making the most of this mass of data in the collection, the preparation, the analysis and the presentation tiers would be a challenge for anyone and all but impossible for people or organizations that lack expert telecoms knowledge.

For the operator analyst it is also a matter of trust. In order to trust output from the ML system, operators must be certain that the correct domain-specific adaptation has been made. Even for a large telecoms vendor with in-house expertise, it takes years of experience to efficiently manage the delicate interplay between a generic ML technology and its domain-specific version.

Ericsson activities

Ericsson is complementing its numerous statistics-reporting products and services with ML technologies.

Product development

As the telecommunications industry continues to mature, operators will increasingly focus on consumer behavior. In addition, the introduction of new IP-based services will lead to a new way of delivering services. With this in mind, Ericsson is developing an all-new data-mining product with a long-term focus. The solution, called Consumer Information Management (CIM), will enable operators to analyze consumers from a multimedia viewpoint.

Two key advantages that Ericsson has are its installed base and important customer contacts. Ericsson's Multi Mediation and Data Warehousing products are among the best in class. And by complementing these with CIM products, Ericsson will be able to deliver an even more complete solution.

Professional services

Ericsson offers a wide range of services including Managed services, Systems Integration, and Consulting. The Consumer Information family of services includes several consultancy and software services that support customer marketing organizations in launching and operating services. This entails monitoring end-user behavior and experience, creating relevant user profiles, and, for targeting purposes, identifying user segments as well as key individuals in social networks.

Research

Ericsson Research, in close cooperation with academia, is exploring the core technologies behind machine learning, the theoretical aspects of various methods and algorithms, their performance for the telecoms domain, and how they can best be modeled onto com-

puting systems. This research is further exploited in application prototypes with real users. Some examples are personalized user screens with content and ads, a mobile billboard community for matching member ads, mobile phone-based marketing, consumer profiling together with grocery retailers, and automated tagging of media files.

Martin Svensson

joined Ericsson Research in 2007 in the Service Layer Technologies research area in Stockholm, Sweden. He is the project manager of the MOLE project, investigating machine-learning technologies and opportunities for telecom network data. His main research focus is in the areas of information filtering, recommender systems, social network analysis and machine learning. Martin holds a Ph.D. in computer science from the University of Stockholm.

Joakim Söderberg

joined Ericsson in 2007 in the Multimedia Technology research area in Stockholm, Sweden. He holds an M.Sc. in computer science from the Royal Institute of Technology (Sweden) and a Ph.D. in multimedia indexing from ENST (Paris, France). His published work includes conference articles on semantic feature extraction, video modeling and image understanding. The current focus of his work is on image analysis and metadata description.

Acknowledgements

Stefan Brodin, Luis Osa Gomez del Campo and Carl-Johan Nilsson

References

1. Mitchell, T.: Machine Learning: McGraw Hill, 1997.
2. Drucker, S., Wong, C., Roseway, A., Glenner, S. and De Mar, S.: MediaBrowser: Reclaiming the shoebox. AVI 2004: 433-436