Information & Security: An International Journal

Dimitar Kamenov, v.39:3, 2018, 255-262 https://doi.org/10.11610/isij.3921

INTELLIGENT METHODS FOR BIG DATA ANALYTICS AND CYBER SECURITY

Dimitar KAMENOV

Abstract: The article examines some intelligent computational methods for big data analysis which are applicable to issues of cyber security and military science, including the analysis of hybrid threats. It presents and compares big data analysis techniques such as quantitative analysis, qualitative analysis, data mining, statistical analysis, machine learning, semantic analysis, and visual analysis. The importance and prospects of intelligent methods for big data analysis are emphasized.

Keywords: cyber security, human factor, big data, data mining, statistical analysis, machine learning, classification, clustering, outlier detection, filtering, semantic analysis, network graphs, spatial data mapping.

Big data is a term designating any collection of datasets so huge or complex that it becomes difficult to process using traditional data management techniques such as, for example, relational database management systems (RDBMS). Data science uses methods to analyse massive amount of data and extract the knowledge it contains.

The characteristics of big data are often referred to as the three Vs:

- Volume – How much data is to be processed?
- Variety – How varied are different types of data?
- Velocity – At what speed is new data generated?

These characteristics can be complemented with a fourth V, veracity, i.e. how accurate is the data? These four characteristics make big data different from the data processed by the traditional data management techniques.

Traditionally, a typical batch processing is implemented with availability of entire dataset and applying statistical sampling from population. Parallel with this, the mainstream of data processing currently is adding the need for processing of streaming data, where the dataset is growing and ordering over time.[2] Processing is done continuously, in a timely manner. Results of streaming data processing have a time-significant ordering and value. Nowadays, big data analytics is a rich mixture of sta-

tistical and intelligent computational methods. William Agresti (2003) recognised the shift towards computational methods and called it *Discovery Informatics*.[5] Agresti saw the process as a composition (or synthesis) of pattern recognition (data mining), artificial intelligence (machine learning), document and text processing (semantic processing), database system management and information retrieval and storage. With the passing of time we can agree with Agresti's foreseeing the analysis to come (with some adding of composites). Big data and data science evolved from statistics and traditional data management but are now considered to be distinct and very perspective subjects.

As far as practical processing of big data is concerned, statistical analysis is preferably used at the first stages of data analysis. Understanding gained from the initial dataset is used for the subsequent computational and intelligent data techniques. Real-time processing adds the need for efficient time-cost and memory-cost algorithms.

Storing data employs RAM, SSDs (solid-state drives), hard-disk drives which insure analytical flexibility and simultaneously providing enough persistent storage. Intelligent computational methods add additional possibilities for streaming data analysis. Big data analysis requires two-speed processing: immediate processing of streaming data and batch analysis of data already stored for patterns and trends to be found. Finding balance between analytical accuracy and volumes and speed of incoming data is of crucial importance.

Quantitative analysis quantifies patterns and correlations searched for in a dataset. Using statistical methods, analysis is based on large cross sections. Large sample size defines generalization of received results to entire dataset. Results are numerical and therefore could be used for further numerical comparisons and operations to reveal and measure relationships.

Qualitative analysis relies on human word description of data patterns and relations. Analysed data sample is smaller compared to the volume of data analysed with quantitative analysis and here the goal is to dig deeper in the data semantics. Based on relatively small sample size, the analyses cannot generalize conclusions to overall dataset. The results are not numerical but descriptive and cannot produce numerical comparisons. The output of qualitative analysis is a description of the interdependencies expressed in words.

Data mining (also data discovery) deals with very large datasets. Analogous to big data analysis, it depends on massive software-based techniques to traverse huge datasets to discover patterns and trends. It tries to reveal dormant and unknown patterns in data and consequently to identify and make known these patterns and trends. Data mining is the underlying basis for predictive analytics and business intelligence (BI).

Statistical analysis uses statistical methods based on mathematical statistics for data analysis. Statistical analysis can be quantitative (most often) or qualitative. It deals with a dataset using summation, such as finding the mean, median, or statistical mode of this dataset. Further, the calculated regression and correlation deepen the data analysis to extract patterns and trends. For example, types of statistical analysis are A/B testing, correlation, regression.

The combination of the high data throughput and calculating power of computing machines and the human knowledge and intuition for discovering patterns and relationships on the other hand led to the invention of machine learning.[1, 2] Presumably, machine learning depends on the instantiation of human knowledge in intelligent data processing performed by computers with human interference as little as possible. In this article, machine learning and its relation to data mining are studied through the scope of the following types of machine learning techniques: classification, clustering, outlier detection and filtering.

Classification (supervised machine learning) is a supervised learning technique where data is classified into bound, previously learned categories. Classification consists of the following two phases:

1. Categorized or labelled in advance training data is inputted in the system. In this way, the system operates a notion of the classified categories;
2. Subsequently, unidentified but analogous data is inputted for classification and founded on comprehension it developed from the training data; the algorithm will classify the unknown unlabelled data.

A typical application of this technique for filtering input data (e-mail spam, for example, or any unwanted type of data input as a whole, as a cyber security measure). Classification is performed for two or more categories based on labelled input data. At the first stage, the training forms understanding of the classification and so the machine learning is used to automatically classify datasets. Then, at the second level, the machine receives unlabelled data and classifies them by itself.

Clustering (unsupervised machine learning) is an unsupervised learning technique by which data classification (division) is handled into different groups based on similarity of data properties as a division criterion. There is no need for previous learning of categories. Categories are formed tacitly as a consequence of data grouping instead of this. Data classification algorithm uses forms in a way in which data is grouped or separate algorithms use different cluster identification techniques.[2,3]

Altogether, clustering is used in data mining to get a conception of the properties of a given dataset. After forming this conception, classification can be used for better predictions about the characteristics of similar data to come.

Clustering can be applied to the categorization of unknown documents and to group together humans with similar behaviour, or different kinds of threats. A scatter graph can be used to summarize visually the results of clustering.

For example, groups so formed are then treated with decisions most suitable to the characteristics of the generic profile of the group, including groups which are threatening cyber security.

Outlier detection is the process of discovering data which is significantly different from or conflicting with the majority of the data within the dataset under consideration. This machine learning technique is used to identify abnormalities, anomalies, irregularities, viciousness and deviations. They can result in advantages (opportunities) or can be unfavourable (risks, threats).

Outlier detection is closely associated with the notion of classification and clustering. The point here is that its algorithms concentrate on finding abnormal values that differ significantly from expected values. It can be solved using supervised or unsupervised learning.

Applications for outlier detection involve fraud detection, network data analysis, sensor data analysis and medical diagnosis. A scatter graph depicts the clusters and data points that are outliers as distant points compared to the clustering.

For example, in order to find out if a user request is likely to be a knavish attack, the system can apply an outlier detection technique that is based on supervised learning. A set of known knavish attacks is initially fed into the outlier detection algorithm. After training the system, the unknown user requests are then let into the outlier detection algorithm to foretell if they are attacks or not.

Filtering is the process of discovering relevant items from a pool of items. Filtering of the items can be made based on single (separate) user's behaviour or the behaviour of multiple users. Filtering is based practically on two main approaches: collaborative filtering and content-based filtering.

Unifying medium basis for filtering implementation is a recommender system.[2,3] Collaborative filtering uses collaboration (merging) a concrete user's behaviour with behaviours of other users; the way behaviours have developed as a past experience. The user's behaviour under consideration gives selected behavioural characteristics, which are collaborated with the same behavioural characteristics of the others in the group. Filtering for target user is achieved based on the similarity of the collective behaviour with the selected behavioural characteristics. Collaborative filtering is grounded on the similarity found in finite number of behavioural characteristics. So, there is a need for large amount of collected user behaviour data in order to precisely filter the items. It applies the law of large numbers.

Content-based filtering concentrates on item filtering based on convergence between users and items (content). Grounded in the users' past behaviour, for example, their likes and demonstrated interests, a user profile is constructed. The convergences identified between the user profile and the characteristics of various items define the items being filtered for the target user. Content-based filtering is concentrated on the separate user preferences and needs no data about other users opposite to collaborative filtering.

A recommender system prognosticates user preferences and produces suggestions for the user respectively. Suggestions generally regard recommending items of interest for the user. A recommender system uses either collaborative filtering or content-based filtering to produce recommendations. It is possible to assemble a hybrid of both collaborative filtering and content-based filtering to increase accuracy and effectiveness of generated recommendations.[3] For example, a recommender system can be built using content-based filtering. Based on similarities found between items searched by users and the characteristics of similar items, the recommender system prepares suggestions that users may also be interested in. The recommender system can be built as a negative limitary system with no output recommendations in the cases of rules' violations and cyber threats.

Text or speech data excerption can possess concrete meanings in concrete contexts. Overall meaning of the complete sentence may retain definite meaning even if built in different variants. Text and speech data need to be comprehended by computing machines in the same way as humans do for the machines to extract meaningful information. Semantic analysis represents methods for obtaining meaningful information from textual and speech data. This article explores the following types of semantic analysis: natural language processing, text analytics and sentiment analysis.[1]

Natural language processing is a preparedness of a computer system to understand human speech and text as they are understood by humans.[4] This allows for performing full-text searches and other manipulations of speech and text in a human-like manner. For example, natural language processing can be used to transcribe speech into textual data that are subsequently mined for recurring patterns of human-computer interactions.

Supervised or unsupervised machine learning can be applied to evolve computer's ability to comprehend natural language instead of using hard-coded learning rules. As a rule, the increased volume of learning data gathered in computer defines higher correctness in human language comprehension. Speech recognition unifies two phases: the computer tries to comprehend the speech and then writes out the text.

Text analytics is a specialized text analysis to extract value out of unstructured text. Unstructured text compared to structured text is significantly more difficult to analyse

and search in general. Text analytics applies data mining, machine learning and natural language processing. Text analytics ensures the possibility of discovering text, and not just searching it.

For example, the transcribed textual data from the example above is further analysed by the means of text analytics to extract meaningful information about the reasons of a user's behaviour, including dangerous types of behaviour.

The basic principle of text analytics is to transform unstructured text into data convenient for searching and analysis. There is a growing necessity to use any value that can be acquired from these kinds of semi-structured and unstructured data. Merely analysing structured data is not considered reliable nowadays when its relative volume minimizes constantly compared to that of semi-structured and unstructured data. Applications involve document classification and search, as well as building a user's profile.

Text analytics generally is applied in two steps:

1. Parsing text within documents to obtain:
   - Named entities – person, group, institution, place, city, country;
   - Pattern-based entities – phone number, identification number, zip code, URL address;
   - Concepts – abstract representations of entities;
   - Facts – relationship between entities.
2. Categorization of documents using these extracted entities and facts.

The obtained information can be used to perform context-specific search on entities, based on the type of relationship discovered between the entities. Text analytics uses semantic rules to extract entities from text files and to structure them so that they can be searched.

Sentiment analysis is a specialized form of text analysis that concentrates on revealing the inclination or emotions of individuals. It determines the attitude of the text author by text analysis made within the context of the natural language. Sentiment analysis not only provides information about how persons feel, but also the intensity of their feelings. This information subsequently can be included into the decision-making process for identifying and taking precautions about negative tendencies for destructive emotions.

Visual analysis is such kind of data analysis that utilises graphic presentation of data to enable or enhance its visual comprehension. The value of visual analysis as a discovery instrument in the area of big data is grounded on the understanding that human nature can comprehend and analyse graphic data more easily than text data. The in-

tention is to use imagery and graphic representation to ease the understanding of the analysed data. More specifically, it allows to reveal patterns, dependencies and anomalies that are not so obvious otherwise. Visual analysis stimulates seeing the data from different angles.[1]

The following types of visual analysis are widely used: heat maps, time series plots, network graphs, spatial data mapping.

Network graphs as a form of visual analysis that represents a set or sets of interconnected entities (objects). Entity can be a person, a group, or some object from the domain under consideration. Entities may be connected with each other directly or indirectly. Traversing the graph may be one-way (when the graph is directed) or may be possible in both directions (when the graph is undirected).

Network analysis is a technique that analyses interrelations between entities (vertices) within the network depicted by a graph. The plotting represents entities as nodes (vertices) and connections as it edges between nodes. There are profiled versions of network analysis including:

- route optimization
- social network analysis
- spread prediction.

The following is a simple example for the application of network analysis for a typical logistic task – route optimization. On some weather conditions supplies delivered from the central warehouse to the remote warehouses can be very sensitive. Network analysis is used to find the shortest paths between the central warehouse and the remote warehouses to minimize the durations of deliveries. An example of social network analysis is the probability of relationship between persons based on their common friends.[2] An example of spread prediction is the prediction of infectious disease.

Spatial data mapping is used to determine the geographic location of individual entities that can then be mapped. Spatial data analysis focuses on analysing spatial or geospatial data in order to discover different location-based relations and patterns between entities.

Spatial data is processed by a Geographic Information System (GIS) that depicts spatial data on a map using as a rule longitude and latitude coordinates. The GIS enables interactive examination of the spatial data, for example geometrical measures – distances, areas, etc. The sensor and social media data catalyse the process of gathering of location-based data, and spatial data can be used for deepening the location knowledge. For example, spatial data is used in logistics to draw existing supply loca-

tions and then to determine optimal locations for new supply points under certain geometrically expressed limitations.

Applications of spatial data analysis include operations and logistic optimization, environmental sciences and infrastructure planning. Input data for spatial data analysis can be expressed either as precise locations (longitude and latitude), or the data needed to calculate locations, such as zip codes or IP addresses. Moreover, spatial data analysis can be used to identify the number of entities that are in the scope of a certain distance from another entity and can be manipulated from there.

In conclusion, there are number of limits of the traditional batch processing and statistical methods for big datasets analysis. The mainstream of data processing currently is adding the need for processing of streaming data, parallelism and intelligent methods.

Bibliography

1. Davy Cielen, Arno D. B. Meysman, and Mohamed Ali, *Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools* (Shelter Island, NY: Manning, 2016).
2. Jure Lescovec, Anand Rajaraman, and Jeffrey D. Ullman, *Mining of Massive Datasets*, Second edition (Cambridge: Cambridge University Press, 2014).
3. Douglas G. McIlwraith, Haralambos Marmanis, and Dmitry Babenko, *Algorithms of the Intelligent Web*, Second edition (Shelter Island, NY: Manning, 2016).
4. Stuart Russell, and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Third edition (Harlow, Essex, UK: Pearson Education, 2014).
5. William W. Agresti, "Discovery Informatics," *Communications of the ACM* 46, no. 8 (2003): 25-28, https://doi.org/10.1145/859670.859691.

About the Author

Dimitar KAMENOV is senior expert at "G.S. Rakovski" National Defence College in Sofia, Bulgaria, with interests in cyber security, hybrid warfare, communication and information systems, systems analysis and design, system integration, intelligent systems, artificial intelligence, multi-agent systems, neural networks, machine learning, big data, data science, streaming and parallel data processing, and knowledge engineering.