Literature Survey

A Novel Method For Handwritten Digit Recognition System

Data Analytics: A Literature Review Paperwork

Authors : 1. Nada Elgendy , University of Oulu.

2.Ahmed Elragal, Luleå University of Technology .

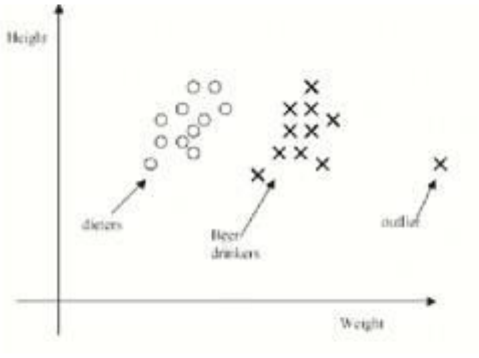
Data Analytics :

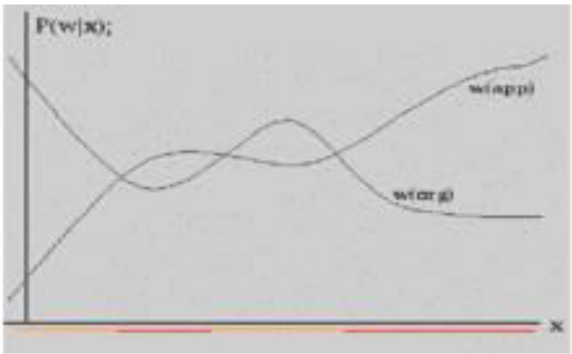
The term “Data Analytics ” has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn’t know before Hence, big data analytics is where advanced analytic techniques are applied on big data sets. Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage

Abstract

An enormous number of CNN classification algorithms have been proposed in the literature. Nevertheless, in these algorithms, appropriate filter size selection, data preparation, limitations in datasets, and noise have not been taken into consideration. As a consequence, most of the algorithms have failed to make a noticeable improvement in classification accuracy. To address the shortcomings of these algorithms, our paper presents the following contributions: Firstly, after taking the domain knowledge into consideration, the size of the effective receptive field (ERF) is calculated. Calculating the size of the ERF helps us to select a typical filter size which leads to enhancing the classification accuracy of our CNN. Secondly, unnecessary data leads to misleading results and this, in turn, negatively affects classification accuracy. To guarantee the dataset is free from any redundant or irrelevant variables to the target variable, data preparation is applied before implementing the data classification mission. Thirdly, to decrease the errors of training and validation, and avoid the limitation of datasets, data augmentation has been proposed. Fourthly, to simulate the real-world natural influences that can affect image quality, we propose to add an additive white Gaussian noise with = 0.5 to the MNIST dataset. As a result, our CNN algorithm achieves state-of-the-art results in handwritten digit recognition, with a recognition accuracy of 99.98%, and 99.40% with 50% noise in handwritten digit recognition, with a recognition accuracy of 99.98%, and 99.40% with 50% noise.

Keywords: data augmentation, Root Mean Square Propagation (RMSprop), batch normalization, MNIST handwritten digit database, receptive field





Data Storage and Management :

One of the first things organizations have to manage when dealing with big data, is where and how this data will be stored once it is acquired. The traditional methods of structured data storage and retrieval include relational databases, data marts, and data warehouses. The data is uploaded to the storage from operational data stores using Extract, Transform, Load (ETL), or Extract, Load, Transform (ELT), tools which extract the data from outside sources, transform the data to fit operational needs, and finally load the data into the database or data warehouse. Thus, the data is cleaned, transformed, and catalogued before being made available for data mining and online analytical functions . However, the big data environment calls for Magnetic, Agile, Deep (MAD) analysis skills, which differ from the aspects of a traditional Enterprise Data Warehouse (EDW) environment. First of all, traditional EDW approaches discourage the incorporation of new data sources until they are cleansed and integrated. Due to the ubiquity of data nowadays, big data environments need to be magnetic, thus attracting all the data sources, regardless of the data quality [5]. Furthermore, given the growing numbers of data sources, as well as the sophistication of the data analyses, big data storage should allow analysts to easily produce and adapt data rapidly. This requires an agile database, whose logical and physical contents can adapt in sync with rapid data evolution [. Finally, since current data analyses use complex statistical methods, and analysts need to be able to study enormous datasets by drilling up and down, a big data repository also needs to be deep, and serve as a sophisticated algorithmic runtime engine. lel Processing (MPP) databases for providing high query performance and platform scalability, to non-relational or in-memory databases, have been used for big data. Non-relational databases, such as Not Only SQL (NoSQL), were developed for storing and managing unstructured, or non-relational, data. NoSQL databases aim for massive scaling, data model flexibility, and simplified application development and deployment. Contrary to relational databases, NoSQL databases separate data management and data storage. Such databases rather focus on the high - performance scalable data storage , and allow data management tasks to be written in the application layer instead of having it written in databases specific languages. Alternatively, Hadoop is a framework for performing big data analytics which provides reliability, scalability, and manageability by providing an implementation for the MapReduce paradigm, which is discussed in the following section, as well as gluing the storage and analytics together. Hadoop consists of two main components: the HDFS for the big data storage, and MapReduce for big data analytics [9]. The HDFS storage function provides a redundant and reliable distributed file system, which is optimized for large files, where a single file is split into blocks and distributed across cluster nodes. Additionally, the data is protected among the nodes by a replication mechanism, which ensures availability and reliability despite any node failures [3]. There are two types of HDFS nodes: the Data Nodes and the Name Nodes. Data is stored in replicated file blocks across the multiple Data Nodes, and the Name Node acts as a regulator between the client and the Data Node, directing the client to the particular Data Node which contains the requested data.

Data Analytic Processing :

After the big data storage, comes the analytic processing. According to [10], there are four critical requirements for big data processing. The first requirement is fast data loading. Since the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time. The second requirement is fast query processing. In order to satisfy the requirements of heavy workloads and real-time requests, many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increase. Additionally, the third requirement for big data processing is the highly efficient utilization of storage space. Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing, and issues on how to store the data so that space utilization is maximized be addressed. Finally, the fourth requirement is the strong adaptivity to highly dynamic workload patterns. As big data sets are analyzed by different applications and users, for different purposes, and in various ways, the underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns

Map Reduce is a parallel programming model, inspired by the “Map” and “Reduce” of functional languages, which is suitable for big data processing. It is the core of Hadoop, and performs the data processing and analytics functions [6]. According to EMC, the MapReduce paradigm is based on adding more computers or resources, rather than increasing the power or storage capacity of a single computer; in other words, scaling out rather than scaling up [9]. The fundamental idea of MapReduce is breaking a task down into stages and executing the stages in parallel in order to reduce the time needed to complete the task

Data Analytics and Decision Making :

Data is becoming an increasingly important asset for decision makers. Large volumes of highly detailed data from various sources such as scanners, mobile phones, loyalty cards, the web, and social media platforms provide the opportunity to deliver significant benefits to organizations. This is possible only if the data is properly analyzed to reveal valuable insights, allowing for decision makers to capitalize upon the resulting opportunities from the wealth of historic and real-time data generated through supply chains, production processes, customer behaviors, etc. Moreover, organizations are currently accustomed to analyzing internal data, such as sales, shipments, and inventory. However, the need for analyzing external data, such as customer markets and supply chains, has arisen, and the use of big data can provide cumulative value and knowledge. With the increasing sizes and types of unstructured data on hand, it becomes necessary to make more informed decisions based on drawing meaningful inferences from the data Accordingly, [8] developed the B-DAD framework which maps big data tools and techniques, into the decision making process [8]. Such a framework is intended to enhance the quality of the decision making process in regards to dealing with big data. The first phase of the decision making process is the intelligence phase, where data which can be used to identify problems and opportunities is collected from internal and external data sources. In this phase, the sources of big data need to be identified and the data needs to be gathered from different sources, processed, stored, and migrated to the end user. Such big data needs to be treated accordingly, so after the data sources and types of data required for the analysis are defined, the chosen data is acquired and stored in any of the big data storage and management tools previously discussed After the big data is acquired and stored, it is then organized, prepared, and processed, This is achieved across a high-speed network using ETL/ELT or big data processing tools, which have been covered in the previous sections.

Consequently, the following phase in the decision making process is the choice phase, where methods are used to evaluate the impacts of the proposed solutions, or courses of action, from the design phase. Finally, the last phase in the decision making process is the implementation phase, where the proposed solution from the previous phase is implemented. As the amount of big data continues to exponentially grow, organizations throughout the different sectors are becoming more interested in how to manage and analyze such data. Thus, they are rushing to seize the opportunities offered by big data, and gain the most benefit and insight possible, consequently adopting big data analytics in order to unlock economic value and make better and faster decisions. Therefore, organizations are turning towards big data analytics in order to analyze huge amounts of data faster, and reveal previously unseen patterns, sentiments, and customer intelligence. This section focuses on some of the different applications, both proposed and implemented, of big data analytics, and how these applications can aid organizations across different sectors to gain valuable insights and enhance decision making. According to Manyika et al.'s research, big data can enable companies to create new products and services, enhance existing ones, as well as invent entirely new business models. Such benefits can be gained by applying big data analytics in different areas, such as customer intelligence, supply chain intelligence, performance, quality and risk management and fraud detection . Furthermore, Cebr's study highlighted the main industries that can benefit from big data analytics, such as the manufacturing, retail, central government, healthcare, telecom, and banking industries.

Customer Intelligence :

Big data analytics holds much potential for customer intelligence, and can highly benefit industries such as retail, banking, and telecommunications. Big data can create transparency, and make relevant data more easily accessible to stakeholders in a timely manner. Big data analytics can provide organizations with the ability to profile and segment customers based on different socioeconomic characteristics, as well as increase levels of customer satisfaction and retention . This can allow them to make more informed marketing decisions, and market to different segments based on their preferences along with the recognition of sales and marketing opportunities . Moreover, social media can be used to inform companies what their customers like, as Additionally, using SNAs to monitor customer sentiments towards brands, and identify influential individuals, can help organizations react to trends and perform direct marketing. Big data analytics can also enable the construction of predictive models for customer behavior and purchase patterns, therefore raising overall profitability [4]. Even organizations which have used segmentation for many years are beginning to deploy more sophisticated big data techniques, such as real-time microsegmentation of customers, in order to target promotions and advertising . Consequently, big data analytics can benefit organizations by enabling better targeted social influencer marketing, defining and predicting trends from market sentiments, as well as analyzing and understanding churn and other customer behaviors.

Quality Management and Improvement :

Especially for the manufacturing, energy and utilities, and telecommunications industries, big data can be used for quality management, in order to increase profitability and reduce costs by improving the quality of goods and services provided. For example, in the manufacturing process, predictive analytics on big data can be used to minimize the performance variability, as well as prevent quality issues by providing early warning alerts. This can reduce scrap rates, and decrease the time to market, since identifying any disruptions to the production process before they occur can save significant expenditures . Additionally, big data analytics can result in manufacturing lead improvements . Furthermore, real-time data analyses and monitoring of machine logs can enable managers to make swifter decisions for quality management. Also, big data analytics can allow for the real-time monitoring of network demand, in addition to the forecasting of bandwidth in response to customer behavior. Additionally, the quality of citizens' lives can be improved through the utilization of big data. For healthcare, sensors can be used in hospitals and homes to provide the continuous monitoring of patients, and perform real-time analyses on the patient data streaming in. This can be used to alert individuals and their health care providers if any health anomalies are detected in the analysis, requiring the patient to seek medical help [22]. Patients can also be monitored remotely to analyze their adherence to their prescriptions, and improve drug and treatment options.

Risk Management and Fraud Detection:

Industries such as investment or retail banking, as well as insurance, can benefit from big data analytics in the area of risk management. Since the evaluation and bearing of risk is a critical aspect for the financial services sector, big data analytics can help in selecting investments by analysing the likelihood of gains against the likelihood of losses. Additionally, internal and external big data can be analysed for the full and dynamic appraisal of risk exposures . Accordingly, big data can benefit organizations by enabling the quantification of risks . High-performance analytics can also be used to integrate the risk profiles managed in isolation across separate departments, into enterprise wide risk profiles.

This can aid in risk mitigation, since a comprehensive view of the different risk types and their interrelations is provided to decision makers.

Conclusion :

In this research, we have examined the innovative topic of big data, which has recently gained lots of interest due to its perceived unprecedented opportunities and benefits. In the information era we are currently living in, voluminous varieties of high velocity data are being produced daily, and within them lay intrinsic details and patterns of hidden knowledge which should be extracted and utilized. Hence, big data analytics can be applied to leverage business change and enhance decision making, by applying advanced analytic techniques on big data, and revealing hidden insights and valuable knowledge. Accordingly, the literature was reviewed in order to provide an analysis of the big data analytics concepts which are being researched, as well as their importance to decision making. Consequently, big data was discussed, as well as its characteristics and importance. Moreover, some of the big data analytics tools and methods in particular were examined. Thus, big data storage and management, as well as big data analytics processing were detailed. In addition, some of the different advanced data analytics techniques were further discussed. By applying such analytics to big data, valuable information can be extracted and exploited to enhance decision making and support informed decisions. Consequently, some of the different areas where big data analytics can support and aid in decision making were examined. It was found that big data analytics can provide vast horizons of opportunities in various applications and areas, such as customer intelligence, fraud detection, and supply chain management. Additionally, its benefits can serve different sectors and industries, such as healthcare, retail, telecom, manufacturing, etc. Accordingly, this research has provided the people and the organizations with examples of the various big data tools, methods, and technologies which can be applied. This gives users an idea of the necessary technologies required, as well as developers an idea of what they can do to provide more enhanced solutions for big data analytics in support of decision.

A Novel Method For Handwritten Digit Recognition System

Authors : 1. Johannes Habel Associate Professor of Marketing University of Houston

2.Sascha Alavi Professor of Sales Management and Chair of the Sales & Marketing Department University of Bochum

3.Nicolas Heinitz Research Associate University of Bochum

Introduction :

Due to the pervasive data ubiquity, sales practice is moving rapidly into an era of predictive analytics, using quantitative methods including machine learning algorithms to reveal unknown information such as customers' personality, value, or churn probabilities. However, many sales organizations face severe difficulties when implementing predictive analytics applications. This article elucidates these difficulties by developing the PSAA Model—a conceptual framework that explains how predictive sales analytics applications support sales employees' job performance. Specifically, the PSAA Model posits that predictive sales analytics applications only improve job performance if (1) sales employees adopt these applications to revise their decision-making, and (2) these updates inherently improve the decision outcome. The mechanisms underlying these two preconditions fundamentally differ. While the former is explained by well-established technology adoption theories, the extent to which adoption improves decision-making is determined by the value potential in the PSA application and the decision-making environment. Thereby, this paper provides a theoretical frame for future studies on predictive sales analytics.