Implementation of DM Techniques In Data Science For Big Data

V. Sree Rekha
Department of Computer Science
Sri Durga Malleswara Siddhartha Mahila Kalasala
Vijayawada, India
rekha.sdmsmk@gmail.com

Dr. R. Padmavathy
Department of Commerce
Montessori Mahila Kalasala
Vijayawada, India
padmavathi.raavi@gmail.com

Dr. B. Syam Sundar Raju
Department of History
Montessori Mahila Kalasala
Vijayawada, India
syamsundarraju82@gmail.com

Abstract:

With the increase of social media sites and proliferation of digital computing devices and Internet access, massive amounts of public data is being generated on a regular basis. Efficient techniques are analysed on amount of data can provide near real-time information about emerging trends and provide early warning in case of an imminent emergency. In addition, mining of these data can reveal many useful indicators of socioeconomic and political events, which can help in establishment of public policies. The focus of this study is to review the application of big data analytics for the purpose of human development. The emerging ability to use big data techniques promises to revolutionalize healthcare, education, and agriculture; facilitate the alleviation of poverty; and help to deal with humanitarian crises and violent conflicts. Size is the first, and at times, the only dimension that leaps out at the mention of big data. This paper attempts to offer a broader definition of big data that captures its other unique and defining characteristics. The rapid evolution and adoption of big data by industry has leapfrogged the discourse to popular outlets, forcing the academic press to catch up. Academic journals in numerous disciplines, which will benefit from a relevant discussion of big data, have yet to cover the topic. In addition to reviewing the implementation of DM techniques in Data Science for Big Data are also introduced.

Keywords: **proliferation, imminent, revolutionalize, humanitarian, leapfrogged, Data Science.**

I. Introduction

Data is everywhere, and is found in huge and exponentially increasing quantities. Information management is an important role in today world. Huge amounts of data is gathered, stored, secured and collected data is interpreted. In order to apply effective analysis on data organizations need to gain competitive advantage while dealing with big data. The big data is the latest technology which is about data and its proper usage. Big data requires military-grade encryption keys to keep information safe and confidential. Data science as a whole reflects the ways in which data is discovered, conditioned, extracted, compiled, processed, analyzed, interpreted, modeled, visualized, reported on, and presented regardless of the size of the data being processed. Big data is a special application of data science.It is a very complex field, which is largely due to the diversity and number of academic disciplines and technologies it draws upon. Data science incorporates mathematics, statistics, computer science and programming, statistical modeling, database technologies, signal processing, data modeling, artificial intelligence and learning, natural language processing, visualization, predictive analytics, and so on.

The presence of “big data”, or this massive amount of increasing data, offers both an opportunity as well as a challenge to researchers. A lot of progress has been made in developing the capability to process, store, and analyze big data: In addition to the big data computing capability, the rapid advances in using intelligent data analytics techniques—drawn from the emerging areas of artificial intelligence (AI) and machine learning (ML)—provide the ability to process massive amounts of diverse unstructured data that is now being generated daily to extract valuable *actionable* knowledge. This provides a great opportunity to researchers to use this data for developing useful knowledge and insights. This is where data science comes in. Many organizations, faced with the problem of being able to measure, filter, and analyze data, are turning to data science for solutions – hiring data scientists, people who are specialists in making sense out of a huge amount of data. Generally, this means making use of statistical models to create algorithms to sort, classify, and process data.

Data Science

Data science is complex and involves many specific domains and skills, but the general definition is that data science encompasses all the ways in which information and knowledge is extracted from data.

Data science is highly applicable to many fields including social media, medicine, security, health care, social sciences, biological sciences, engineering, defense, business, economics, finance, marketing, geolocation, and many more.

In the modern world, companies such as Google and Facebook dealing with petabytes of data. Google processes more than 24 petabytes of data per day, while Facebook, a company founded a decade ago, gets more than 10 million of postings per hour. The lot of data in advancing technology is exponentially increasing due to increased digitization like IoT which uses sensors, for example in the shape of wearable devices, to provide data related to human activities and different behavioral patterns. It is estimated that we are generating 10^{18} bytes per day in our daily life to process our basic needs using technology.

Big Data

Big Data is essentially a special application of data science, in which the data sets are enormous and require overcoming logistical challenges to deal with them. The primary concern is efficiently capturing, storing, extracting, processing, and analyzing information from these enormous data sets. Processing and analysis of these huge data sets is often not feasible or achievable due to physical and/or computational constraints. Special techniques and tools (e.g., software, algorithms, parallel programming, etc.) are therefore required.

Big Data is the term that is used to encompass these large data sets, specialized techniques, and customized tools. It is often applied to large data sets in order to perform general data analysis and find trends, or to create predictive models. A primary component of big data is the so-called *Three Vs* (3Vs) model. This model represents the characteristics and challenges of big data as dealing with volume, variety, and velocity. Companies such as IBM include a fourth “V”, veracity, while Wikipedia also notes variability.

Big data essentially aims to solve the problem of dealing with enormous amounts of varying-quality data, often of many different types, that is being captured and processed sometimes at tremendous (real-time) speeds. No easy task to say the least!

So in summary, Big Data can be thought of being a relative term that applies to huge data sets that require an entity (person, company, etc.) to leverage specialized hardware, software, processing techniques, visualization, and database technologies in order to solve the problems associated with the *3Vs* and similar characteristic models.

Data is everywhere. In fact, the amount of digital data that exists is growing at a rapid rate—in fact, more than 2.7 zettabytes of data exist in today’s digital universe, and that is projected to grow to 180 zettabytes in 2025.

All this data—from your photos to the Fortune 500’s financials—has only recently begun to be analyzed to tease out insights that can help organizations improve their business. That’s why more organizations are seeking professionals who can make sense of all the data.

It is easy enough to become a data scientist. Once you get the art of data analysis right, it is just a matter of practicing your newly-found skills well enough to become proficient.

Data science is interdisciplinary, incorporating elements of statistics, data mining, and predictive analysis, and focusing on processes and systems that extract knowledge and insights from data. It is also known as “analytics transformation” because the goal is to “transform” raw data into usable insights. It has also been called “industrial analytics” because the context is industrial rather than scientific – to analyze data for competitive or quality improvements that can be gained by having a better understanding of one’s customers, potential customers, service model, and almost any aspect of the organization that can be represented in bytes.

II. DM, Knowledge Discovery & Data Science

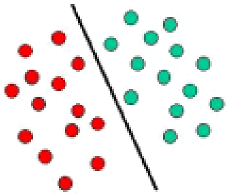
Data mining usually refers to automated pattern discovery and prediction from large volumes of data using ML techniques. Data mining can also be used to refer to online analytical processing (OLAP) or SQL queries that entails retrospectively searching a large database for a specific query. OLAP queries, also known as decision-support queries, are typically complex expensive queries that take a long time and touch large amounts of data. The process of extracting useful information or knowledge from the structured/ unstructured data and databases (relational and non-relational), using data mining and ML techniques, is called *knowledge discovery*, sometimes collectively called *KDD (knowledge discovery in databases)*. This knowledge can be in the form of brief and concise visual reports, a predicted value or a model of a larger data generating system. Data science is an inderdisciplinary field in which different KDD techniques and processes are studied. Next, we briefly describe the trend of non-relational databases to store unstructured data followed by an introduction to predictive analytics that helps in knowledge discovery from the huge volumes of structured/unstructured data.

III. Big Data Techniques

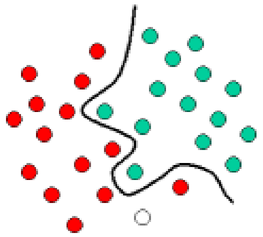
Modern datasets, or the *big data*, differ from traditional datasets in 3 V's: volume, velocity and variety. In today's age huge volumes of data is being generated at huge pace (or velocity) and the numerous sources of data give vast variety to it. All of this data, if harnessed intelligently, can truly realize the notion of the *information age*. Actionable information can be gathered from the data after performing intelligent processing and analytics on the available data. The techniques (specially related to machine learning) in order to gather, store, process and analyze this vast amount of data are the subject matter of this section. We also try to link this discussion, and different examples considered here to explain various concepts, to the humanitarian development. The aim of this section is to provide readers with a brief background and related work of the relevant techniques to help them understand their applications when discussed in the perspective of humanitarian development.

Support vector machine (SVM)

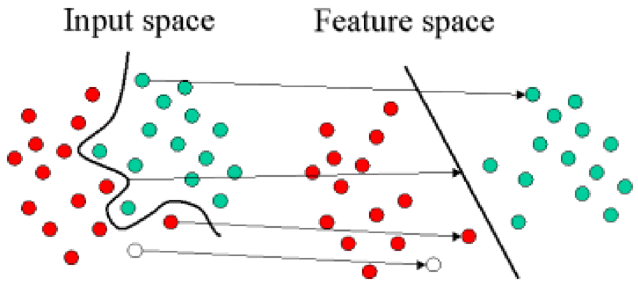
This algorithm learns to define a hyperplane to separate data into two classes. A hyperplane is the line that divides a group but is based on a property or attribute rather than location.This algorithm can help figure out an underlying separation mechanism between people who will buy a product and those who won't. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).



The above is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective groups (GREEN and RED in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). This situation is depicted in the illustration below. Compared to the previous schematic, it is clear that a full separation of the GREEN and RED objects would require a curve (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suited to handle such tasks.



The illustration below shows the basic idea behind Support Vector Machines. Here we see the original objects (left side of the schematic) mapped, i.e., rearranged, using a set of mathematical functions, known as kernels. The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) is linearly separable and, thus, instead of constructing the complex curve (left schematic), all we have to do is to find an optimal line that can separate the GREEN and the RED objects.



A. ***Classification SVM***

1) Classification SVM Type 1

For this type of SVM, training involves the minimization of the error function:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

subject to the constraints:

$$y_i\left(w^T\phi(x_i)+b\right) \geq 1-\xi_i \text{ and } \xi_i \geq 0, i=1,\ldots,N$$

where C is the capacity constant, w is the vector of coefficients, b is a constant, and ξrepresents parameters for handling nonseparable data (inputs). The index i labels the N training cases. Note that $y \in \pm 1$represents the class labels and xi represents the independent variables. The kernel $\phi$is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

2) ***Classification SVM Type 2***

In contrast to Classification SVM Type 1, the Classification SVM Type 2 model minimizes the error function:

$$\frac{1}{2}w^T w - \nu\rho + \frac{1}{N}\sum_{i=1}^{N}\xi_i$$

subject to the constraints:

$$y_i\left(w^T\phi(x_i)+b\right) \geq \rho-\xi_i, \xi_i \geq 0, i=1,\ldots,N \text{ and } \rho \geq 0$$

In a regression SVM, you have to estimate the functional dependence of the dependent variable y on a set of independent variables x. It assumes, like other regression problems, that the relationship between the independent and dependent variables is given by a deterministic function f plus the addition of some additive noise:

Regression SVM

y = f(x) + noise

The task is then to find a functional form for f that can correctly predict new cases that the SVM has not been presented with before. This can be achieved by training the SVM model on a sample set, i.e., training set, a process that involves, like classification (see above), the sequential optimization of an error function. Depending on the definition of this error function, two types of SVM models can be recognized:

3) ***Regression SVM Type 1***

For this type of SVM the error function is:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i + C\sum_{i=1}^{N}\xi^{*}{}_i$$

which we minimize subject to:

$$\begin{aligned}
&w^T\phi(x_i) + b - y_i \leq \varepsilon + \xi_i^{*}\\
&y_i - w^T\phi(x_i) - b_i \leq \varepsilon + \xi_i\\
&\xi_i, \xi_i^{*} \geq 0, i = 1,\ldots, N
\end{aligned}$$

4) Regression SVM Type 2

For this SVM model, the error function is given by:

$$\frac{1}{2}w^T w - C\left(\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\left(\xi_i + \xi_i^{*}\right)\right)$$

which we minimize subject to:

$$\begin{aligned}
&\left(w^T\phi(x_i) + b\right) - y_i \leq \varepsilon + \xi_i\\
&y_i - \left(w^T\phi(x_i) + b_i\right) \leq \varepsilon + \xi^{*}{}_i\\
&\xi_i, \xi^{*}{}_i \geq 0, i = 1,\ldots, N, \varepsilon \geq 0
\end{aligned}$$

There are number of kernels that can be used in Support Vector Machines models. These include linear, polynomial, radial basis function (RBF) and sigmoid:

B. ***Kernel Functions***

$$K\left(\mathbf{X}_i, \mathbf{X}_j\right) = \left\{\begin{array}{ll}
\mathbf{X}_i \cdot \mathbf{X}_j & \text{Linear}\\
\left(\gamma\mathbf{X}_i \cdot \mathbf{X}_j + \mathrm{C}\right)^d & \text{Polynomial}\\
\exp\left(-\gamma\,|\,\mathbf{X}_i - \mathbf{X}_j\,|^2\right) & \text{RBF}\\
\tanh\left(\gamma\mathbf{X}_i \cdot \mathbf{X}_j + \mathrm{C}\right) & \text{Sigmoid}
\end{array}\right\}$$

where $K\left(\mathbf{X}_i, \mathbf{X}_j\right) = \phi\left(\mathbf{X}_i\right) \bullet \phi\left(\mathbf{X}_j\right)$

that is, the kernel function, represents a dot product of input data points mapped into the higher dimensional feature space by transformation ϕ

***Gamma* is an adjustable parameter of certain kernel functions.**

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

IV. Technical Challenges

In today's age it is quite likely that big data will gain substantial potential and importance in order to shift the paradigm of the conventional humanitarian development process in almost every walk of life. It is, however, not a panacea to all the problems in the modern day world. Just like any other innovation, the wide scale adoption of big data is hindered by many potential challenges. In this section we discuss some of these challenges from two perspectives: technical and ethical. Correspondingly, we describe open issues and future work, which is required to address these challenges.

*1) **Technical challenges***

There are various technical challenges involved in implementing BD4D. As an example, with the daily production of vast amounts of data, are the processing and storing capabilities scaling proportionally? Below we present and describe a few of the technical challenges:

- Data Supply Chain: With all the benefits, policy analysis by utilizing big data is a precarious task. Many potential challenges and perils entail this process . Privacy is a major, and largely debated, concern in gathering data from users. During the data gathering process (the *big-data supply chain*) the context and semantics of the data can be altered resulting in faulty and sometimes controversial policies. Present day data sources are also prone to temporal and spatial restrains, due to disparity in worldwide technology proliferation, resulting in a statistical bias, which in turn can result in inefficient policies.

- Technology Usage: The context, specially in the online data collected about students, is very crucial to consider in LA. A problem that arises, while tracking the data-trail left by students online, is that every individual has a different attitude towards the usage of technology. The social network and sentiment analysis should be performed with care so that the students who use the Internet less, or differently, as compared to other students should not be penalized in the data analysis.

- Spatial Problem: Many users update their status with the information related to a crisis sitting, all together, at a different geographical site. This is a challenge in pin pointing an actual place of crisis for which the information was provided at the first place. So, the data gathered from the actual ground based surveys and aerial imagery should be corroborated with these for the effective actions to fight a crisis situation.

- Vulnerability of Connectivity: Although, scientists are working on trust management systems for the verification of the information gathered for an appropriate action: Fraudulent information and entities can still infiltrate the information network. This information can then be treated like normal data and has the potential to diffuse and infect other connected entities of the information network. This vulnerability is primarily caused by the connected nature of information producing and consuming entities, this *vulnerability of connectivity* and cacasding errors/failures are discussed by Barabasi in his book.

- Interoperability: Big data analytics often include collecting and then merging unstructured data of varying data types. As an example call detail records from cell phone companies, satellite imagery data and face-to-face survey data have to be corroborated together for the better and less-biased analysis. Merging and harmonizing this data for analysis is a challenging task. For effective data analytics a system is needed that could make data streams of potentially different formats homogenous.

- Fragmentation: The challenge of fragmentation is one of the major impediment to large-scale deployment of big data analytics. As an example, a patient might be seeing different specialists for, seemingly, different medical reasons. These specialists, then further, can prescribe different types of clinical tests resulting in different kinds of results. If, however, some protocol or a system is developed to integrate these fragments together and run analysis on them collectively then a clear and big picture of a patient's current health can be extracted. If the issue of fragmentation is resolved then this can not only speed up the diagnosis process but also has the ability to provide personalized treatment most suitable for the patient under consideration.

- Technology Scaling: In recent times, the technologies of cloud computing and software-defined networking (SDN) have proved very useful for efficiently implementing big data solutions: going forward, more work is needed to ensure that the computing and networking facilities scale to the ever-increasing scale of data.

*2) **Ethical***

Besides all the technology related challenges presented above it is imperative to consider the ethical dimensions of utilization BD4D. Throughout the paper, we have tried to outline, besides all the benefits, the potential challenges and harms incurred by the deployment of big data for development purpose. We saw that privacy is one of the major issues in almost every field where big data analytics are applied. Besides privacy, the challenge of fragmentation is one of the major impediment to large-scale deployment of big data analytics. Besides these well-known issues, there are a few subtle challenges as well: most of which fall into the realm of ethics and abuse of technology. Here we list a few of the challenges faced in the perspective of ethics when dealing with BD4D.

- Privacy: This concern tops the list. As an example, with large amounts of data being collected about individuals, it is of utmost importance that such information should not be abused for any sort of personal or financial gains.

- Digital Divide: This divide is simply the nonuniform diffusion of technological advancement and expertise through out the world. The result of this divide harm nations that lack the infrastructure, economic affordability and *data-savvy* faculty. The digital divide, the well-known issue of privacy, and the control and monopoly of entities exploiting the data are among the important challenges that hinder the wide scale deployment of big data techniques for development.

- Open Data: There are also many possible issues with open data. For greater transparency, it is desirable that government/development data is openly accessible. However, it is also important to think about who has the right to access, use, link, and repurpose open data (and how much flexibility is desirable, keeping in view various misuse and privacy issues). With the rising use of big data in humanitarian and development aid, governance efforts should focus on ensuring that sensitive information (such as the location of humanitarian actors and internally displaced persons (IDPs)) does not become open, since such data may maliciously be exploited by malevolent actors.

Finally, the evolution of data science, in itself is a challenge. This is because the field requires expertise and collaboration of people from various fields and disciplines. Interdisciplinary efforts should be encouraged and financially incentivized so that big data can be analyzed with the right perspectives and ethics in place.

IV. BIG DATA FOR DEVELOPMENT AREAS

Big data projects for different development areas and tools for big data analytics using DM techniques in Data Science.

Project	***Type***	***Open Data***	***Description***
Digital Humanitarian Network	Online Service	✗	Network of IT to fight human crises.
Open Street Map	Non-profit	✗	Real-time, online and a map of natural crisis of a region.
Google Flu Trends	Non-profit	✓	Presently inactive but provides data about flu and dengue trends for different regions.
Data.gov (Health)	Governmental	✓	Open health data, tools and applications from the US Government.
Educational Data Mining	Non-profit	✗	Community dedicated for R&D based on learning data mining of several education data.
Data.gov (Education)	Governmental	✗	Open data, tools and applications related to education at all levels from the US Government.
Education Data (The World Bank)	Non-profit	✓	Open data from The World Bank’s projects and data study related to education.
Data.gov	Governmental	✓	Provides open data for different field such as health, education,

Project	***Type***	***Open Data***	***Description***
			agriculture etc.
Hadoop	Open Source Tool	NA	Open source tool for distributed computations on large amounts of data.
Elastic Map Reduce (Amazon)	Processing Tool	NA	Data processing service for large amounts of data.

Table I : Big Data Projects for Development Areas

V. CONCLUSIONS

In this paper, we have reviewed the literature focused on using big data techniques for human development (BD4D). Our aim in this paper is to highlight to a broad audience the immense potential of BD4D in a variety of settings including humanitarian emergencies (including disaster response and migrant crisis), agriculture, poverty alleviation, food production, healthcare and education. We have highlighted the various challenges and pitfalls associated with BD4D. We envision that in the future BD4D will play a big role in human development and global prosperity, but to succeed with BD4D, it is imperative that researchers are able to tackle and solve the challenges identified.

VI.REFERENCES

[1] The home of the U.S. Government’s open data (Education). http://www.data.gov/education. [Online; accessed 01-October-2017].

[2] Kwon et al., 2014: O. Kwon, N. Lee, B. ShinData quality management, data usage experience and acquisition intention of big data analytics International Journal of Information Management, 34 (3) (2014), pp. 387-394

[3] Barbier and Liu, 2011: G. Barbier, H. LiuData mining in social media C.C. Aggarwal (Ed.), Social network data analytics, Springer, United States (2011), pp. 327-352

[4] Alltuition. https://www.alltuition.com/. [Online; accessed 02-October-2017].

[5] W. Hu, N. Xie, L. Li, X. Zeng, S. MaybankA survey on visual content-based video indexing and retrieval IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 41 (6)(2011), pp. 797-819

[6] Beaver et al., 2010: Beaver, S. Kumar, H.C. Li, J. Sobel, P. Vajgel Fin ing a needle in haystack: Facebook's photo storage Proceedings of the nineth USENIX conference on operating systems design and implementation, USENIX Association, Berkeley, CA, USA (2010), pp. 1-8

[7] Open Data Kit. https://opendatakit.org/. [Online; accessed 04-October-2017].

[8] Cukier, 2010: Cukier K., The Economist, Data, data everywhere: A special report on managing information, 2010, February 25, Retrieved from http://www.economist.com/node/15557443.

[9] Diebold, 2012: F.X. DieboldA personal perspective on the origin(s) and development of “big data”: The phenomenon, the term, and the discipline (Scholarly Paper No. ID 2202843)

V. Sree Rekha is presently working as Lecturer, Department of Computer Science in Sri Durga Malleswara Siddhartha Mahila Kalasala. She worked as Assistant Professor in the Department of Computer Science from June 2008 to May 2017 and had an experience of 9 years in Montessori Mahila Kalasala. She participated actively in few National and International Seminars. Her publications are also published in both National & International Journals.

Dr. R. Padmavathy is presently working as Head, Department of Commerce in Montessori Mahila Kalasala Degree College. She Completed 2 Minor Research Projects sponsored by UGC. She participated actively in few National (36) and International (8) Seminars. Her publications are also published in both National & International Journals.

Dr. B.Syam Sundar Raju is presently working as Incharge Principal & Head, Department of History in Montessori Mahila Kalasala Degree College. He published book - “Vijayanagaram Zamindar in the Colonial Andhra”. He participated actively in few National (35) and International (9) Seminars. His publications are also published in both National & International Journals.