Sort and Manage Huge Amounts of Unorganized Data Using Data Mining and Neural Networks

Amit Kumar[1]*, Dr. Faizanur Rahman[2]

[1] Research Scholar, Kalinga University

[2] Research Guide, Departmet of Computer Science, Kalinga University.

Abstract - The exponential growth of unstructured data presents a formidable challenge to organizations seeking valuable insights and actionable information. This paper explores innovative methods for sorting and controlling massive amounts of unstructured data through the integration of data mining techniques and neural networks. The first part of this study delves into the significance of unstructured data, outlining its sources, complexities, and potential value. It underscores the limitations of traditional data management approaches and highlights the need for advanced methodologies to extract meaningful patterns and knowledge. Data mining, a fundamental component of this research, is examined in depth. Various data mining techniques, such as clustering, classification, and association rule mining, are explored, with a focus on their adaptability to unstructured data. Additionally, we discuss the role of feature engineering and dimensionality reduction in enhancing the efficiency of data mining processes.

Keywords - Data Mining, Neural Networks, Sort, Manage, Unorganized

----------------------------------x----------------------------------

1. INTRODUCTION

The rise of databases in the modern world is indicative of the information explosion. Each year, databases accumulate more and more information. Their size is now often measured in terra bytes, rather than just giga bytes. Traditional manual techniques of data analysis, such as spreadsheets and ad hoc queries, are insufficient when faced with enormous amounts of data since they are not easily interpretable. These can only provide you with a broad overview of the data; they can't go into the details and hone down on the insights that matter. Furthermore, companies risk missing out on insights gleaned from data's latent associations if they fail to do so. That's why we need cutting-edge methods and software that can automatically and intelligently analyze massive amounts of data using previously unknown information. In 1995, during the First International Conference on Knowledge Discovery in Databases, the novel method was officially dubbed "Knowledge Discovery in Databases." Data mining (DM) is a step in the knowledge discovery process that uses a variety of algorithms to glean insights from massive datasets. Presently, DM is a relatively young field of study that has attracted a lot of interest and excitement from academics and is quickly gaining traction in the commercial sector.[1]

Relationships, especially complicated ones, are often buried or lost in a mountain of data, but DM algorithms can tease them out. It is difficult to separate the relevant pieces of information since the connections between them are not established. In contrast to how fields in a tuple are typically related in a database, the DM system can search for relations that are not immediately apparent or established. Experts in fields such as machine learning, pattern recognition, databases, statistics, A.I., NNs, knowledge acquisition, and data visualization are interested in DM. Methods, algorithms, and approaches from these many areas are often included in DM.

2. DATA MINING FOR UNSTRUCTURED DATA

A review of the literature on unstructured data suggests that data mining is a logical outgrowth of the development of computing. The evolution of the databases and data management business led to the improvement of various crucial features, including data collecting, database building, data administration, storage, and retrieval. Knowledge Discovery in Databases (KDD) is an interdisciplinary branch of computer science that deals with the process of finding patterns in large datasets, and data mining is the analytical phase of this process.

To mine more refined data sets from somewhat unstructured data is the goal of the field of unstructured data mining. It often involves gathering information from resources that aren't typically mined for such purposes.[2]

The majority of the time, textual information is processed and evaluated while dealing with unstructured data. Text mining may be broken down into two stages: (1) text refining, which converts raw text documents into an intermediate format; and (2) text mining itself. Extraction of essential information from a larger body of data or information in an intermediate form. Semi-structured forms include representations like conceptual graphs, whereas structured forms include relational data models.

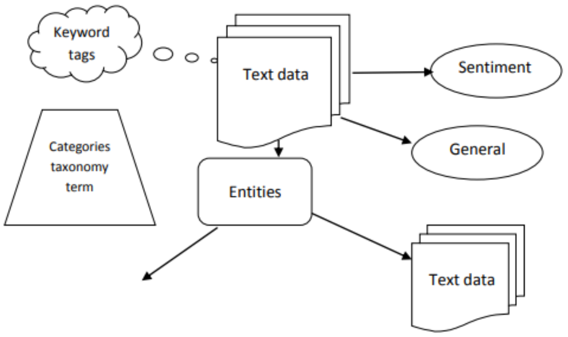

Figure 1.1: Prospective elements for mining in unstructured data sets

One major issue is the exponential growth in the quantity of unstructured data being collected. Due to its growing importance, the study of unstructured data has given rise to a variety of approaches for dealing with the challenges it presents.[2]

Present day there is a plethora of textual business and professional resources available. Text mining is used to analyze and make sense of unstructured, raw data that may be inconsistent. However, it does permit dealing with incomplete data, ambiguity, uncertainty, and fuzziness.

3. DATA MINING

Data mining refers to the act of sifting through massive amounts of data in search of useful patterns and correlations that may be used to inform analytical approaches to resolving business challenges. Companies may use data mining tools and processes to better foresee future trends and make educated business choices.

In data science, data mining is a core field that uses sophisticated analytical methods to extract actionable insights from data. Data mining is a part of the knowledge discovery in databases (KDD) process, which is a technique in data science for collecting, cleaning, and analyzing data. Although data mining and KDD may be used interchangeably in certain contexts, the two are often seen as separate methodologies. [3]

Historically, analyzing data has been a purely manual task. Before, analysts would utilize statistical methods to compile findings and write reports. This method used to work, but as data volumes and dimensions have grown, it has become inadequate. Manual analysis, even if feasible, cannot keep up with the rate at which data is increasing. No human being can process millions of records, each with hundreds of fields, and make sense of the data. Consequently, scholars and professionals concerned with the issue of automating data analysis have been slowly but gradually developing the area under the banners of KDD and DM. Inference from data is at the very core of the statistical science of statistics. Statistical methods are crucial for verifying hypotheses and discovering new insights from data via exploratory research. The performance of algorithms borrowed from the fields of statistics, pattern recognition, and artificial intelligence is severely hampered by the fact that the full dataset must fit in computer memory. Most modern DM algorithms have long since overcome this shortcoming. Instead, they are scalable and adaptable enough to deal with training data of any size.

The goal of any data analysis should be to learn something novel and practical from previously stored information. However, there are several reasons why data stored in databases might be unreliable. Errors in encoding or measurements, or even previously unknown factors, might be at blame. Because of this, it is difficult to conclude using statistical methods when the data is noisy. It is not only possible to extract information from clean data using data analysis that integrates NN, fuzzy logic (FL), or GA, but the such analysis also performs well with noisy data.[4]

3.1 Types of Data

Problems with analysis, capture, length, search, sharing, storage, transfer, visualization, and privacy breaches are among the most prominent throughout data analysis.

There are essentially three distinct categories of information.

i. Structured Data

ii. Semi-Structured Data and

iii. Unstructured Data

i. Structured Data

Schema-on-write refers to the practise of defining and formatting data according to a predetermined structure before it is stored. This is what we mean when we talk about structured data. The relational database is a prime illustration of structured data since it stores information in clearly delineated fields, such as customer names and addresses, that can be accessed using a database query language like Structured Query Language (SQL).

As a result of the normalization process and other settings, structured data is often inherent in relational database management systems (RDBMS).[5]

Some examples of structured data storage are as follows:

- Data warehouse
- RDBMS
- ERP

The data types used to determine the structure of the information specifying whether or not the data's underlying structures may be joined.

ii. Semi-structured data

Data that is not adequately arranged in a specialized repository like a database is called semi-structured data. However, unlike raw data, it does include associated information, such as metadata, which makes it more receptive to processing. Semi-structured describes the data, which in this case may comprise a large amount of linked online data.

The term "semi-structured data" is used to describe information that is often unstructured but has metadata attached to it that indicates specific features. Because of the information included in the metadata, the data may be organized, searched, and analyzed with more efficiency than would be possible with just unstructured data. You might think of semi-structured data as a hybrid between traditional structured data and the chaos of unstructured data.

In contrast to a database with CRM tables, a tab-delimited file holding client information is an example of semi-structured data. Semi-structured data, on the other hand, has more structure than unstructured data; the tab-delimited file is more particular than a list of comments from a customer's Instagram account.[6]

iii. Unstructured data

Data that does not follow a predetermined data model and lacks a clear, machine-readable structure is considered unstructured. Since unstructured data is not arranged predictably and does not adhere to a standard data model, it cannot be stored in a traditional relational database.

Unstructured data is any data that lacks a defined structure. No particular order or method may be ascribed to the unstructured material. It's information that doesn't fit neatly into any preexisting data model and has no clear organizational scheme that would allow a machine to process it efficiently.

4. CONCEPT OF FILTERING AND MANAGING UNSTRUCTURED DATA

The phrase "information filtering" is used to refer to several different methods of getting relevant data to the right people. This phrase is being used increasingly in both popular and technical writing to describe software like electronic mail, multimedia distributed systems, and electronic governmental documents. Filtering and its associated operations, like retrieval, routing, classification, and extraction, are frequently confused with one another. Only by drawing that difference can the unique research challenges connected with filtering be recognized and addressed.[7]

Managing information that lacks a standard format requires what is known as "unstructured data management." This entails everything from acquiring the data to determining how to best store, organize, and analyze it.

Excel, Google Sheets, and relational databases are all useful for managing structured data, but dealing with unstructured data necessitates the use of more sophisticated software as well as elaborate rules and methods.

Every day, businesses deal with massive volumes of unstructured data. This data comes from a wide variety of places, including market research, customer comments, app reviews, social media posts, and more. All of this data includes useful information that businesses can use to make better choices, develop data-driven strategies, enhance existing operations, save costs, and gain a competitive edge.

4.1 Challenges of Managing Unstructured Data

Businesses typically confront a set of problems that can be holding them back from handling unstructured data:[8]

- **Data quality**

When dealing with unstructured data, it is common to practice first cleaning the data. Poor quality data may be the consequence of having data that is duplicated, out of date, untrustworthy, erroneous, or includes outliers, all of which can distort outcomes in unstructured data analysis. Getting the most out of your data requires cleaning and preparing large datasets, which can be a significant challenge for enterprises.

- **Siloed data:**

Data is collected independently by each group and kept in their unique forms and databases. Information, however, has to be kept in a centralized location where it can be accessed by anybody and where its retrieval is rapid and easy. Businesses will need to invest time and money in data routing before implementing new processes, but they may not have the capacity to do so.

- **Data Growth & Costs:**

In addition to the rising expenses associated with data management, storing your growing amount of unstructured data will be a must. You may reduce the amount of space data takes up by compressing it (and eliminating duplicates), which will help you save money and maximize the efficiency of your data management.[9]

5. DATA MINING APPLICATION

Some significant and helpful uses for DM have emerged as a result of recent technology advances. Market segmentation, real estate pricing, client acquisition and retention, cross-selling, credit card fraud detection, risk management, and attrition analysis are just a few instances where DM has been effective.

Market segmentation is done so that a product's marketing and sales efforts may concentrate on the most promising leads. If done well, this may ensure that the sales and marketing budget is used to its fullest potential. Customer segments may be defined using DM predictions on customer behavior. Multiple methods of data segmentation are available to DM. The data segment is represented as a leaf in the decision tree DM method. In the KNN, the winning neuron is used to define the data subset.

Model-based methods may be used to estimate property values. To estimate prices, regression is the method most often used. The challenge of regression analysis is in isolating the factors outside of the dependent variable that influences the cost. Later on, in the regression analysis, these variables will be utilized. Selecting variables for the regression may be done using DM methods. Similarly, real estate price modeling using the classification tree DM approach has shown promising results.[10]

In today's increasingly competitive marketplace, attracting and keeping customers is a top priority for businesses of all sizes. Getting new clients is a constant challenge for any business, but especially one that is just starting. Presently, however, keeping profitable consumers is the top priority for businesses. For one thing, keeping an old client costs less than finding a new one. Today, customer churn is the greatest challenge to service providers in the cellular telephone business. When paying customers to leave your service in favor of a rival, they are said to be "churners." Simply put, churn is the overall monthly loss of customers. Companies with a large number of customers who are in a position to quickly migrate to a rival for better facilities and services at a lower cost are more concerned about churn. By identifying at-risk clients and developing a knowledge of what criteria indicate high-risk customers, predictive approaches from DM may be applied for churn prevention to provide significant cost savings.

Introducing new goods or services to an organization's current clientele is known as cross-selling. When you cross-sell, you're aiming to boost earnings. DM allows for the cross-selling of new items to be performed with little risk to the company. Existing client data is modeled using DM approaches before cross-selling begins. In this framework, the desired action may motivate the consumer to buy a different product. The model is then used to rank consumers from those with the highest purchase likelihood to those with the lowest, with the highest-ranked subset being chosen as the target customers.[11]

6. DATA MINING METHODS

The process of data mining is searching for meaningful relationships among massive datasets. Insightful conclusions may be drawn from the data according to the methods brought about by this procedure. This process also creates fresh insights into the data we currently have at our disposal. Techniques like pattern recognition, classification, association, outlier analysis, cluster analysis, regression, and prediction are used. In the face of a potentially rapid shift in the data at hand, it is simple to identify patterns. We have gathered the data and divided it into parts for analysis. Data similarity is used to form clusters.

Many strategies exist for accomplishing DM objectives. These techniques may be categorized in several ways, including by the tasks they're best suited to and the kinds of settings in which they excel. This section discusses many of the most common DM techniques.

6.1 Classification

Data modeling (DM) is useful since it can classify data beforehand. Supervision is present. The following is the mathematical definition of the i_{th} class or class C_i:

$$C_i = \{o \in S | Cond_i(o)\}$$

where o is selected from dataset S based on whether or not o meets the class membership criteria C_i.

Methods for categorizing data may be seen in the identification of fraudulent credit card transactions and the approval of bank loans, for instance. we see a fictitious dataset with 16 examples in two dimensions. An individual who has received a loan from a certain financial institution is represented by each dot on the graph. People whose loans are in good standing with the bank have been separated from those whose payments have been missed. The bank may want to know the status of a loan application before they even consider them for a loan if prior examples are any indication. DM's categorization technique is suitable for this task. After reviewing the data, the classification technique develops a classifier (also known as a decision function). This classifier can predict what category a new instance will fall into. Various methods, such as regression, neural networks (NN), decision rules, etc., may be used to build the classifier.[12]

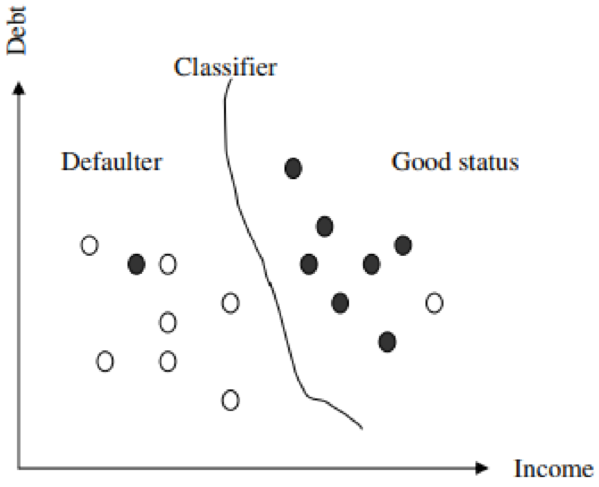

Figure 1.6: Example of a Classification

The most popular form of DM is classification. In the academic literature, a plethora of different approaches to categorization has been presented. The VSA ID3, AQ15, and CN2 documents are the foundational sources for classification theory and practice. The Version Space Algorithm (VSA) uses inductive learning to classify data based on two hypotheses—a broad hypothesis and a more narrowly focused hypothesis. The general hypothesis will always be supported by the presence of positive examples, even if those examples do not support the particular hypothesis. Thus, Negative cases will be in line with the narrow hypothesis but will contradict the broad one. ID3 is a decision-tree-based supervised learning system. The input is a subset of the training examples, and the decision tree is constructed from those subsets. Splitting the training data set along selected attributes results in constructing a tree for each subgroup, which is repeated until all of the nodes in a given subset are of the same class.

The inductive learning system in AQ15 creates categorization rules based on data patterns. To build a ranked set of rules, CN employs entropy as a search heuristic. The best rule from the set of rules that meet the user-defined criteria is used to categorize an instance.

5. CONCLUSION

The management and analysis of massive volumes of unstructured data are critical challenges in today's data-centric landscape. Unstructured data, stemming from sources such as text, images, audio, and more, has become increasingly abundant. Traditional data management techniques are ill-equipped to handle this deluge of information due to its inherent complexity and lack of predefined structure. Data mining techniques play a pivotal role in converting unstructured data into actionable insights. Methods like clustering, classification, and association rule mining are adaptable to various data types and can reveal hidden patterns and relationships within unstructured datasets.

REFERENCES

1. Etzion, O., Gilat, D., and Sharon, G. (2015), “Inference of Reactive Rules from Dependency Models,” LNCS, Springer-Verlag, Heidelberg, November 2003, vol. 2876, pp. 49-64.
2. Mues, C. and Vanthienen, J. (2016), “Using neural network rule extraction and decision tables for credit risk evaluation,” Management Science, vol. 49, no. 3, pp. 312-329.
3. Lyer B. and Swami A. (2019), “An Interval Classifier For Database Mining Applications,” International Conference on Very Large Databases(VLDB), pp. 560-573, Vancouver, Canada.
4. Rissanen, J, and Yu, B. (2017), “The Minimum Description Length Principle in Coding And Modeling,” IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2743-2760.
5. Imielinski, T. and Swami, A. (2017), “Database Mining: A Performance Perspective”, IEEE: Special issue on Learning and Discovery in KnowledgeBased Databases, pp. 914-925.
6. Wynne H. and Yiming M. (2018), “Integrating Classification and Association Rule Mining,” Proceedings of 4th International Conference on Knowledge Discovery and Data Mining, pp.80-86.
7. Ching-Kian, W. and Philip S, Y. (2016), “Scoring the Data Using Association Rules,” Applied Intelligence, vol. 18, no. 2, pp. 119-135.
8. Clark, P. and Niblett, T. (2018), “The CN2 Induction Algorithm,” Machine Learning, Vol. 3, pp. 262-283
9. T. Lu, N. Gu, (2016) "An algorithm for efficientprivacy-preserving item-based collaborative filtering", Future Generation ComputingSystem, 55, 311–320.
10. Adamcza, R. and Grąbczewski, K. (2018), “A New Methodology of Extraction, Optimization and Application of Crisp and Fuzzy Logical Rules”, IEEE Transactions on Neural Networks, vol. 12, pp. 277-306.
11. Goscinski A., Khalil I.,(2020) "PPFSCADA: Privacy preserving framework for SCADA data publishing",Future Generat. Comput. Syst., Vol. 37, pp. 496-511.
12. M. Haider, (2015) Beyond the hype: Big dataconcepts, methods, and analytics,

International Journal Information Management, Vol.35, No. 2, pp. 137–144.

Corresponding Author

Amit Kumar*

Research Scholar, Kalinga University