International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

OPTIMIZE THE STORAGE VOLUME USING DATA MINING TECHNIQUES

P. Nagababu[1], M. Prabhu Kalyan[2], E. Roshan Francis[3], N. Venkata Raju[4]

Assistant Professor, B. Tech Students
* Department of Information Technology
** ANDHRA LOYOLA INSTITUTE OF ENGINEERING & TECHNOLOGY **

***Abstract*-** Data mining is a modern area of science of extracting useful information from large data sets or databases. Applications of Data mining can be found in various areas. This paper introduces new contributions by optimization as a key technology in data mining. The methods suggested for solution of such important problems as where it deals with large data. The component of a larger data mining process and new data mining techniques can be built using entirely optimization-based method. Data Mining framework to compare and efficient optimization of storage location using sketch-based algorithm and sign-based algorithm.

***Keywords*-** Parking Slot, Deep Learning, Automated Parking, CNN, Mask R-CNN, YOLO, Image Processing

I. INTRODUCTION

Data Mining is the process of automatic discovery of useful information in large data repositories. Generally, Data Mining can be divided into two categories according to the objective of algorithms: 1) Classification Analysis 2) Association Analysis. Many Data Mining methods involve with mathematical programming techniques. Optimization can be a component of a larger Data Mining process and New Data Mining techniques can be built using entirely optimization-based method. These optimization-based Data Mining techniques are applied mainly in Classification Analysis, whereas very few algorithms in Association Analysis based as optimization. Data mining is an iterative process that typically involves the following phases: Problem definition: A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required. Data exploration: Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data.

II. AIM & OBJECTIVE

To Optimize the storage space using Data Mining Techniques and provide Minimal Space for the users.

We have been motivated observing the benefits of physically handicapped people like deaf and dumb. But if any normal human being or an automated system can understand their needs by observing their facial expression then it becomes a lot easier for them to make the fellow human or automated system understand their needs.

III. EXISTING SYSTEM

The overall architecture of PDP-Miner is The system, from bottom to top, consists of two components: Data Analytics Platform (including Task Management Layer and Physical Resource Layer) and Data Analysis Modules. Data Analytics Platform provides a fast, integrated, and user-friendly system for data mining in distributed environment, where all the data analysis tasks accomplished by Data Analysis Modules are configured as workflows and also automatically scheduled. Details of this module are provided in Section 2.1. Data Analysis Modules provide data-mining solutions and methodologies to identify important production factors, including controlling parameters and their underlying correlations, in order to optimize production process. These methods are incorporated into the platform as functions and modules towards specific analysis tasks. In PDP-Miner, there are 3 major analytic modules: data exploration, data analysis, and result management. In Section 2.2, more details are provided by presenting our data mining solutions customized for PDP production data.

IV. PROPOSED SYSTEM

The primary goal is to improve the yield rate of products by optimizing the manufacturing workflow. To this end, one important question is to identify the key parameters (features) in the workflow, which can significantly differentiate qualified products from defective ones. However, it is a non-trivial task to select a subset of features from the huge feature space. To tackle this problem, we initially experimented several widely used feature selection approaches. Specifically, we use Information Gain [11], mRMR [5] and ReliefF [24] to perform parameter selection. Figure 7 shows the top 10 selected features by these three algorithms on a sampled PDP dataset.

As observed in Figure 7, the three feature subsets share only one common feature ("Char 020101-008"). Such a phenomenon indicates the instability of feature selection methods, as it is difficult to identify the importance of a feature from a mixed view of feature subsets. In general, the selected are the most relevant to the labels and less redundant to each other based on certain criteria.

Information gain(top10)	mRMR(top10)	ReliefF(top10)
Char_110101-004	Char_020101-016	Char_110101-009
Char_110101-003	Char_020101-004	Char_110101-008
Char_110101-005	Char_020101-008	Char_100102-079
Char_110101-002	Char_020101-009	Char_100101-199
Char_110101-006	Char_020101-010	Char_100101-208
Char_110101-001	Char_020101-007	Char_100101-212
Char_020101-008	Char_020101-006	Char_100102-013
Char_020101-017	Char_020101-003	Char_100101-213
Char_100101-168	Char_020101-014	Char_100102-081
Char_020101-013	Char_020101-013	Char_020101-008

V. STUDY OF THE SYSTEM

Data Analytics Platform:

Traditional data-mining tools or existing products [10, 21, 19, 18, 23, 30] have three major limitations when applied to specific industrial sectors or production process analysis: 1) They support neither large-scale data analysis nor handy algorithm plug-in; 2) They require advanced programming skills when configuring and integrating algorithms for complex data mining tasks; and 3) They do not support large scale of analysis tasks running simultaneously in heterogeneous environments.

Easy operation for task configuration:

Users, especially non-data-analyst, can easily configure a complex data mining task by assembling existing algorithms into a workflow. Configuration can be done through a graphic interface. Execution details including task scheduling and resource management are transparent to users. Flexible supports for various programs The existing data mining tools, such as data preprocessing libraries, can be utilized in this platform. There is no restriction on programming languages for those programs exist or to be implemented, since our data analytic platform is capable of distributing the tasks to proper runtime environments

Data Analysis Modules :

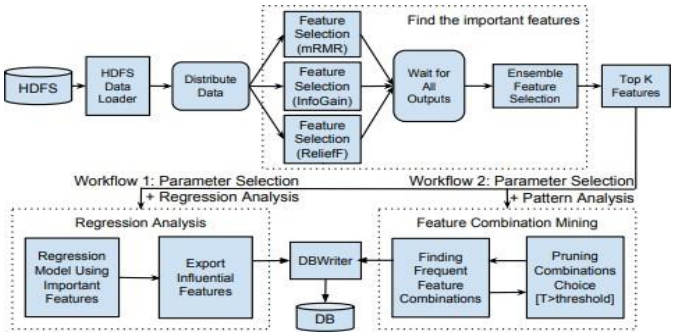
Data Exploration:

The Comparison Analysis and Data Cube are capable of assisting data analysts to explore PDP operation data efficiently and effectively.

Comparison Analysis :

Comparison Analysis, shown in Figure 6(a), provides a set of tools to help data analysts quickly identify parameters whose values are statistically different between two datasets according to several statistical indicators. Comparison Analysis is able to extract the top-k most significant parameters based on predefined indicators or customized ranking criteria. It also supports comparison on the same set of parameters over two different datasets to identify the top-k most representative parameters of two specified datasets.

Discriminative Analysis: Discriminative analysis (SeeFigure 6(e)) is an alternative approach to identify the feature values that have strong indication to the target labels (panel grade). By grouping and leveraging the features of individual panels, this approach is able to find the most discriminative rules (a set of



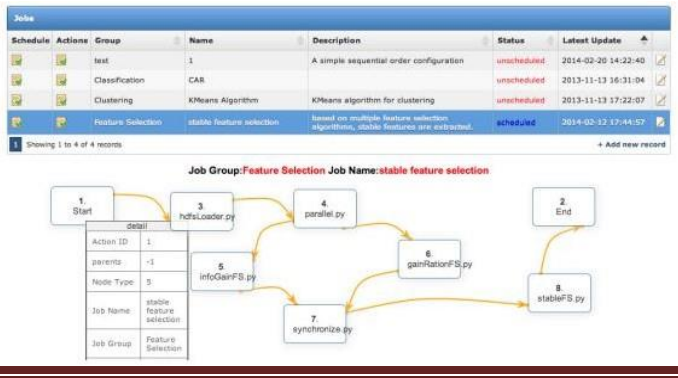
features with the values) to the target labels according to the data.

only need to explicitly create tasks dependencies before the workflow executing automatically byour platform.

HARDWARE & SOFTWARE REQUIREMENTS:

HARDWARE REQUIRMENTS:

- System : Intel i3 2.4 GHz.
- Hard Disk : 500 GB.



- RAM : 4GB

SOFTWARE REQUIRMENTS:

- Operating system : Windows 8.
- Coding Language : python

VI. SYSTEM ARCHITECTURE

The complete working procedure of the project.

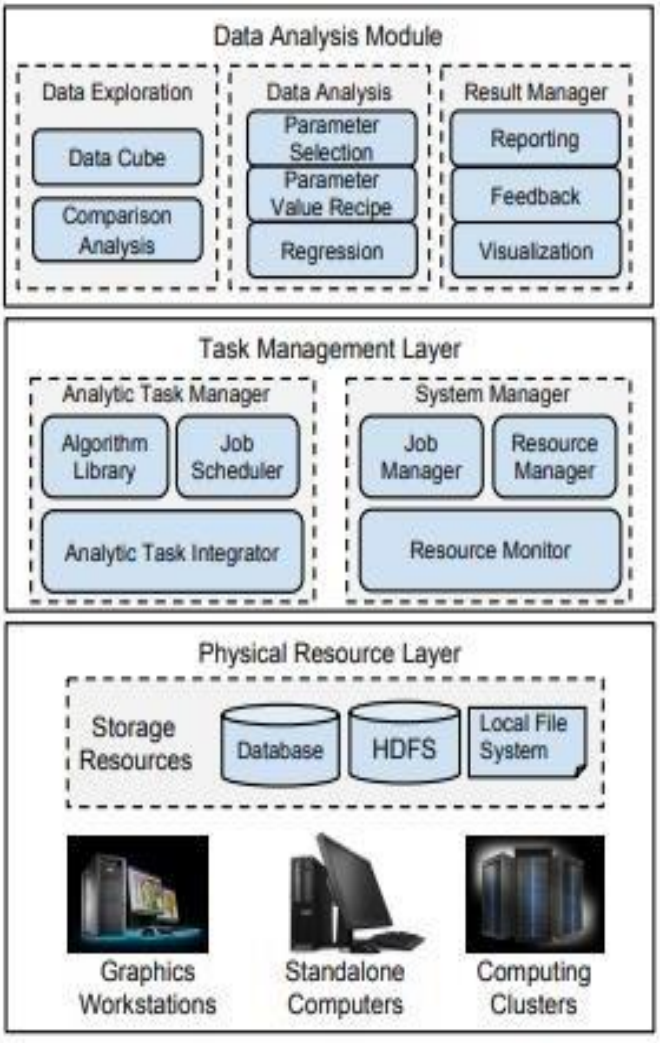


Figure 3: System Architecture.

VII. SYSTEM DESIGN

System design shows the overall design of system. In this section we discuss in detail the design aspects of the system.

METHODOLOGY INVOLVED IN THIS PROJECT

TensorFlow

TensorFlow is a free and open-source software library for dataflow and differentiableprogramming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks. It is used for both research and production at Google. TensorFlow was developed by the Google Brain team for internal Google use. It was released under the Apache 2.0 open-source license on November 9, 2015.

Numpy:

It is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensionalcontainer of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze.

Matplotlib

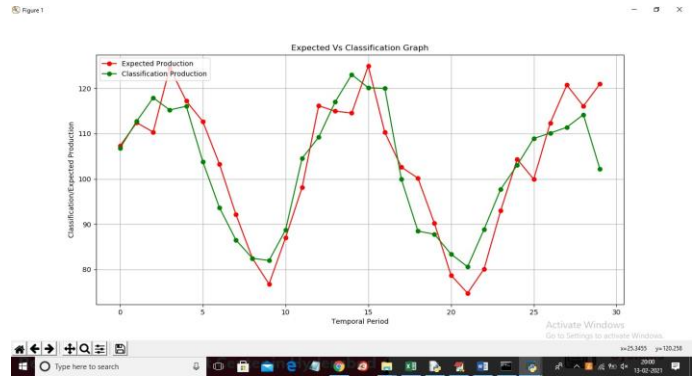
Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

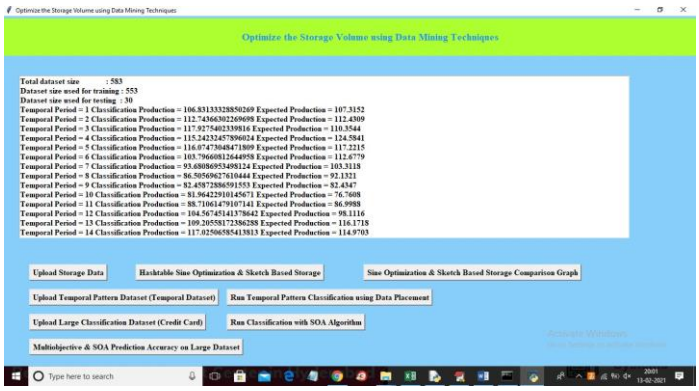
Matplotlib canbe used in Python scripts, the Python and IPython shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, seethe sample plots and thumbnail gallery. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

Scikit – learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.

THE RESULTS

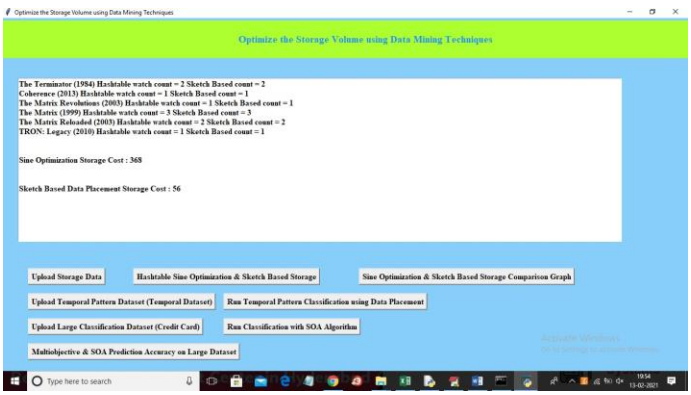


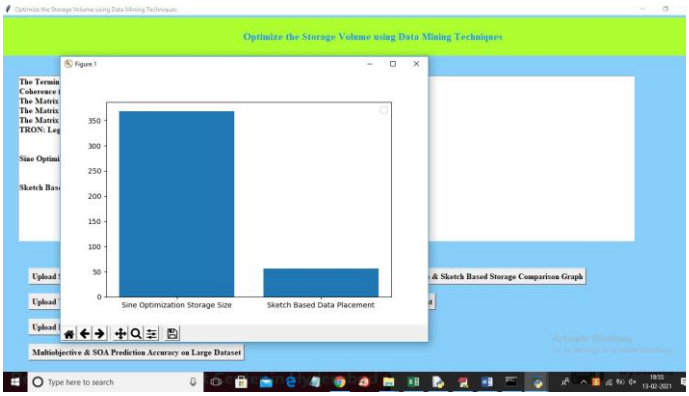


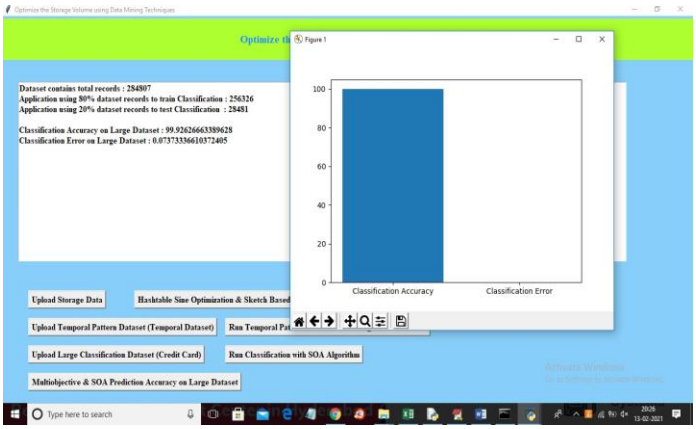
LITERATURE SURVEY

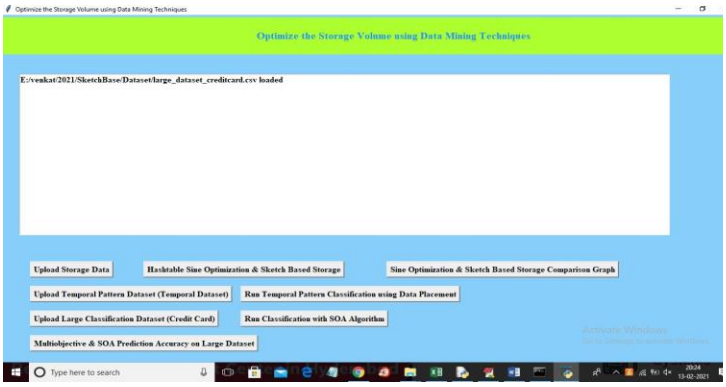
Some Optimization methods can be used in specific areas of data mining for classification and clustering which they find inter related many Data Mining tasks such as classification, clustering and model selection can be formulated as multi-criteria optimization problems. These problems later can be cast into multi-criteria optimization framework. It has introduced in few search optimization algorithm strategies. The two main categories are continuous optimization strategy and discrete optimization strategy, which was further divided into smaller search strategy. Optimization is a scientific discipline that deals with the detection of optimal solutions for a problem, among alternatives. The optimality of solutions is based on one or serval criteria that are usually a problem and user dependent. It has identified opportunities for interactions between DM and optimization in e- Customer Relation Management(eCRM). These opportunities are to develop methods for preprocessing data to optimize the performance of eCRM models and to develop optimization based for active learning with feature selection methods as well as to select for selecting best patterns or models generated by Data Mining. sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

FINAL OUTPUT









VIII. CONCLUSION

It has a practical data-driven solution that considers both system flexibility and algorithm customization is expected to fill the application gap between the manufacturer and data analysts. We firmly believe that, if properly being applied, the use of data analytics will become a dominating factor to underpin new waves of productivity growth and innovation, and to transform the way of manufacturing across industries in a fundamental manner.

REFERENCES

1. R Belz and P Mertens. Combining knowledge-based systems and simulation to solve rescheduling problems. Decision Support Systems, 17(2):141–157, 1996.

2. Injazz J Chen. Planning for erp systems: analysis and future trend. Business process management journal, 7(5):374–386, 2001.

3. Wei-Chou Chen, Shian-Shyong Tseng, and Ching-Yao Wang. A novel manufacturing defectdetection method using association rule mining techniques. Expert systems with applications, 29(4):807–815, 2005.

4 Chad A Davis, Fabian Gerick, Volker Hintermair, Caroline C Friedel, Katrin Fundel, Robert Kuffner, ¨ and Ralf Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. Bioinformatics, 22(19):2356–2363, 2006.

5. Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02):185–205,2005.

6. Gang Fang, Gaurav Pandey, Wen Wang, Manish Gupta, Michael Steinbach, and Vipin Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. TKDE, 24(2):279–294, 2012.

7. C Groger, Florian Niedermann, Holger Schwarz, and Bernhard Mitschang. Supporting manufacturing design by analytics, continuous collaborative process improvement enabled by the advanced manufacturing analytics platform. In CSCWD, pages 793–799. IEEE, 2012.

8 Christoph Gr¨oger, Florian Niedermann, and Bernhard Mitschang. Data mining-driven manufacturing process optimization. In Proceedings of the World Congress on Engineering, volume 3, pages 4–6, 2012.

9. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. ACM SIGKDD explorations newsletter, 11(1):10–18, 2009.

10. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. SIGKDD Explorations, 2009.