Identification and Classification of Big Data

Sasank M P(201CS153)
Sharique Nadim(201CS156)

Group 9

CS473 - **Big Data Analytics**

Overview

- A Structured Approach Towards Big Data Identification
- A Survey on Classifying Big Data with Label Noise
- A literature review on one-class classification and its potential applications in big data
- A Survey on IoT Big Data: Current Status, 13 V's Challenges, and Future Directions
- LSDStrategy: A Lightweight Software-Driven Strategy for Addressing Big Data Variety of Multimedia Streaming
- T-PCCE: Twitter Personality-based Communicative Communities Extraction System for Big Data

A Structured Approach Towards Big Data Identification

Outline:

- The structured approach presented in this study provides a straightforward method to identify and classify big data based on the Volume, Velocity, and Variety characteristics.
- It offers a practical and effective means to optimize resources, allowing for more efficient handling of big data workloads.
- The approach demonstrates clear benefits by saving up to 58% of main memory, reducing disk reads by 44%, and minimizing the use of computational resources.
- By categorizing big data into strong, moderate, or weak levels, it aids in making well-informed decisions on resource allocation and workload offloading.

A Structured Approach Towards Big Data Identification

Limitations:

- The approach primarily focuses on the processing aspect of big data, leaving out other important dimensions, such as Veracity and Value.
- It may not address the full spectrum of complexities associated with big data, as certain nuances and data-specific requirements could be missed.
- The proposed model relies on mathematical equations, which may not capture all real-world intricacies and may need adjustments for different contexts.
- Future work should aim to extend the model to encompass all aspects of big data and consider variations in specific applications and platforms.

A Survey on Classifying Big Data with Label Noise

Outline:

- This survey comprehensively covers the treatment of label noise within big data, focusing on the challenges posed by high-volume, high-variety, and high-velocity data.
- It outlines various techniques, including distributed kNN implementations, ensemble learners, and deep learning methods, tailored to mitigate the impact of label noise on machine learning algorithms.
- Results indicate the efficiency of distributed kNN solutions and the power of ensemble filters in reducing over-editing and improving classification performance.
- Deep learning approaches highlight the importance of pre-training on clean benchmark datasets, making them robust to noisy data in large-scale, diverse domains.
- Data streaming research emphasizes the effectiveness of combining local and global ensemble noise filters, particularly in addressing class-label noise with concept drift.

A Survey on Classifying Big Data with Label Noise

Limitations:

- Primarily focuses on label noise issues and solutions, leaving out other data quality aspects, such as data inaccuracies (veracity).
- Given the vast diversity in datasets and application contexts, the study may not encompass all possible big data scenarios and real-world complexities.
- While the techniques presented are effective, they may not cover every conceivable data challenge or all possible domains.
- The survey largely concentrates on supervised learning with label noise and big data, potentially overlooking unsupervised learning and other relevant machine learning domains.
- The applicability of certain techniques, such as deep learning methods, to structured data needs further exploration.
- Continuous research is needed to address these limitations, explore broader domains, and advance the state of the art in big data label noise treatment.

A literature review on one-class classification and its potential applications in big data

Outline:

- Definition: One-Class Classification (OCC) as a technique for detecting abnormal data points.
- Focus: Addressing issues related to imbalanced datasets commonly found in big data.
- Survey Scope: An overview of OCC literature published in the last decade, covering outlier detection, novelty detection, and deep learning in OCC.
- Application Areas: Biomedical data analysis, neuroimaging, industrial machine monitoring, and fraud detection.
- Notable Techniques: OCSVM, SVDD, Parzen Windows, LOF, Deep Learning, and Gaussian Process Regression.
- Key Takeaway: OCC's potential in mitigating class imbalance and addressing anomalies in various domains.

A literature review on one-class classification and its potential applications in big data

Limitations:

- Lack of specific studies on OCC in the context of big data.
- Absence of research addressing severe class imbalance, class rarity, noisy data, and feature engineering.
- Overfitting issues in SVDD with noisy or uncertain data.
- Limited studies on feature selection and transfer learning in OCC.

Future Research Directions:

- Exploration of OCC applications in domains with high class imbalance and class rarity.
- Enhanced solutions for handling noise in OCC, especially in big data scenarios.
- Investigate advanced feature engineering and selection techniques.
- Research on transfer learning's potential in OCC.

A Survey on IoT Big Data: Current Status, 13 V's Challenges, and Future Directions

Outline:

- IoT and Big Data: IoT leverages sensor-based data acquisition and cloud-based big data analysis for intelligent actions.
- Data-related issues are impeding IoT growth. This survey identifies challenges in IoT Big Data (IoTBD).

13 "V's" challenges are explored:

- Volume, Variety, Velocity, Veracity, Value
- Variability, Visualization, Validity, Vulnerability
- Volatility, Venue, Vocabulary, Vagueness
- Solutions and approaches for each challenge have been proposed in recent research.

A Survey on IoT Big Data: Current Status, 13 V's Challenges, and Future Directions

Limitations:

- Inefficient Data Acquisition: Current IoT data collection methods lack energy efficiency.
- Energy-efficient Acquisition: Focus on energy-efficient data collection methods for sustainable IoT devices.
- Scalable Analytics: Utilize ML and DL for scalable analysis, ensuring accuracy with large datasets.
- Optimized Network Processing: Efficient data allocation to reduce latency and bandwidth issues.
- Enhanced Data Governance: Develop lawful IoT data collection, usage, and sharing tools.

LSDStrategy: A Lightweight Software-Driven Strategy for Addressing Big Data Variety of Multimedia Streaming

Outline:

- Rapid generation of diverse data necessitates efficient solutions despite resource constraints.
- Traditional methods allocate cloud resources based on data characteristics, but big data streams are uncertain due to their randomness.
- LSDStrategy, a novel strategy, uses machine learning on multimedia streams to predict workload types.
- Multi-classifiers tested, including Decision Tree (DT), K-Nearest Neighbor (K-NN), and Random Forest (RF).
- Experiments showed DT consistently performed well for both artificial and real-world datasets.
- LSDStrategy agility and adaptivity tested with a synthetic stream.

LSDStrategy: A Lightweight Software-Driven Strategy for Addressing Big Data Variety of Multimedia Streaming

Results, Conclusion, and Future Work:

- LSDStrategy's purpose is to predict data types in multimedia streams.
- Extracted features to predict workload variety, balancing accuracy and processing time.
- Experiments showed promising results, with DT performing well.
- LSDStrategy can be deployed in a distributed environment to enhance performance.
- Future work includes the implementation of resource management based on LSDStrategy predictions.
- This approach aims to improve big data analysis and processing accuracy and efficiency.

Twitter Personality-based Communicative Communities Extraction System for Big Data

Outline:

- Objective: Identification of communicative communities on social media platforms, specifically Twitter, using users' personality traits as a key characteristic.
- System: T-PCCE (Twitter Personality-based Communicative Communities Extraction)
- Approach: Utilizes machine learning techniques for personality extraction and a modularity-based community detection algorithm extended with a post-processing step based on personality criteria.
- Methodology: Employs the cloud infrastructure and MapReduce Programming Environment.
- Results: Demonstrates the creation of more communicative communities by considering users' personality traits, improving information exchange within smaller communities.

Twitter Personality-based Communicative Communities Extraction System for Big Data

Flaws and Future Work:

- Limited Dataset: Experiments were conducted on relatively small datasets (696 and 1246 nodes).
- Single Personality Model: Utilizes the Big Five personality model, leaving scope for exploring other personality models.
- Focus on Twitter: Primarily designed for Twitter, with potential limitations in generalizing to other social media platforms.
- Time-based Analysis: Investigate the temporal evolution of communicative communities.
- Feature Influence: Study the influence of additional features on community ranking and communication within communities.

Anomaly Identification in Big Data using Apache Spark

- Setting up Apache Spark
- Data Ingestion
- Model Selection
- Training the Model
- Model Evaluation

Thank You!