Address Clustering Optimization Through Advanced ML and DL Techniques

Aditya Mehta(1226588222), Andi Lian(1219538359), Manav Kamlesh Parmar(1229897281), Soumya Gupta(1229081362), Vanshaj Gupta(1228730141), Xinyuan Wang(1230936943)

Keywords: Clustering, Machine Learning, Deep Learning, Address Standardization

1. Abstract

The "Address Clustering Optimization through Advanced ML and DL Techniques" project focuses on efficiently clustering residential addresses to enhance e-commerce delivery systems, particularly in countries like India where address formats lack standardization. By applying both traditional and deep learning methods, the project aims to identify similar addresses and those nearby, facilitating optimized delivery routes, reduced delivery times, and improved geocoding accuracy. This approach addresses logistical challenges by refining address clustering, potentially lowering delivery costs.

2. Problem Definition

The project is dedicated to overcoming the challenges of address clustering in densely populated areas like India, where non-standardized and incomplete address data often impedes efficient delivery and logistical operations. By leveraging advanced Machine Learning (ML) and Deep Learning (DL) techniques, it aims to systematically organize addresses into cohesive clusters based on proximity. This initiative is crucial for enhancing delivery route optimization, reducing transit times, and improving address verification processes, ultimately facilitating more effective and cost-efficient logistics solutions in complex urban environments.

3. Data Set

For our project, we will utilize the DeepParse Address Data, sourced from GRAAL-Research's GitHub repository. This dataset comprises 26,000 instances, each with four main features. To facilitate our analysis, we plan to divide this data into training and testing subsets using an 80:20 split, although this ratio may be adjusted based on further research insights. The preprocessing steps will include feature transformation and cleaning, crucial for preparing the dataset for the clustering algorithms we intend to implement. This comprehensive dataset will serve as the foundation for our exploration of advanced machine learning and deep learning techniques in address clustering optimization.

Dataset Source and Information: https://github.com/GRAAL-Research/deepparse-address-data

4. State-of-Art Methods & Algorithms

Some potential machine learning and deep learning methods and algorithms that we will explore for address clustering include:

- K-Means clustering: Divides data into k clusters by minimizing intra-cluster distances and maximizing inter-cluster distances.

- DBSCAN: Identifies clusters based on density, capable of discovering clusters of arbitrary shape and distinguishing noise.
- Self-organizing maps: Though not a traditional clustering method, SOMs can be used to visualize and cluster high-dimensional data in a lower-dimensional space, preserving topological properties.
- Hierarchical clustering: Builds a tree of clusters by continuously merging or splitting data points, revealing data structure at multiple scales.
- Neural networks: While primarily used for prediction, neural networks can be applied to clustering by learning representations of data that can then be clustered by traditional methods.
- Convolutional neural networks: Specialized in processing structured grid data such as images, using convolution operations to capture spatial hierarchies.
- Recurrent neural networks (LSTM): Designed for sequential data, capable of learning long-term dependencies across time steps.
- Transformer models (BERT): transformers can generate embeddings for text data that could be clustered to identify patterns or group similar texts.

5. Research Plan

5.1. Introduction and Reference Collecting

To establish a foundation for the project by collecting information we need and reviewing existing research related to clustering, machine learning, and deep learning techniques. Identify the difficulties of doing clustering in current status for non-standardized address formats.

5.2. Problem Definition Establish

Find the problem statement based on the information gained from research and reviews. Focus on how to use Machine Learning and Deep Learning to optimize address clustering.

5.3. Dataset Preparation

Prepare the clean, structured address dataset for analysis and make it suitable for Machine Learning and Deep Learning models.

5.4. Methodology Development

By going through experiments, develop and refine the Machine Learning Model and Deep Learning Model for clustering addresses based on similarity and accuracy. Develop a fully functional UI that is able to run the selected models in the best way.

5.5. Implementation and Testing

To implement the chosen models on the prepared dataset and evaluate their performance. By going through refinement and development, make the model achieve optimal performance.

5.6. Results Analysis and Optimization

Analyze the results provided by selected models, identify the areas of improvement and update the models accordingly.

5.7. Conclusion and Future Work

Summarize the project findings and discuss the possible applications in the real world.

6. Evaluation Plan

We will evaluate the performance of our models by using metrics like accuracy, precision, recall, F1 score, and silhouette coefficient. We will also consider computational efficiency and ease of implementation. The best performing model will be selected based on its ability to cluster similar addresses together in an efficient manner accurately.

7. Project Timeline: Tasks, Descriptions, Deadlines

Task	Description	Deadline
Reference & Problems	**1. Collecting information and reference related to clustering, machine learning, and deep learning techniques.** **2. Analyze the difficulties we are facing, especially in countries that have non standardized address formats.**	**02/19/2024**
Challenge Definition	**1. Identify the specific logical challenges that need to be done by clustering.** **2. Get the corresponding solutions or ways for the given challenge.**	**03/26/2024**
Dataset Preparation	**1. Gather or download datasets we need.** **2. Perform data sorting and organizing.**	**03/11/2024**
Method Development	**1. Select ML and DL models.** **2. Evaluate models based on accuracy, efficiency, and their ability to handle datasets.** **3. Optimize models through tests.** **4. Develop functional UI for program.**	**03/25/2024**
Implementation & Test	**1. Ready for the Project Presentation on Apr 1, 2024.** **2. Apply the selected ML and DL models to cluster the address datasets.** **3. Conduct comparative analysis to determine the most effective techniques for address clustering.**	**04/01/2024**
Result Analysis	**1. Ready for the Group Demo Presentation on Apr 22, 2024.** **2. Identify the challenges and limitations encountered during the implementation stage.** **3. Optimize the models based on the feedback and**	**04/22/2024**

	analysis to improve performance.	
Conclusion	**1. Ready to turn in a comprehensive Final Project Report on May 1, 2024.** **2. Propose directions for future research including the possible applications of using ML and DL techniques in address clustering.**	**04/29/2024**

8. Division of Work

Team Member	Task
Aditya Mehta	Literature review, dataset preparation
Andi Lian	Report Writing and documentation
Manav Kamlesh Parmar	UI development
Soumya Gupta	Results analysis and optimization
Vanshaj Gupta	Presentation and project coordination
Xinyuan Wang	Clustering development and evaluation

9. References

1. Kumar, P. Santhosh, et al. "A Comprehensive Review on Deep Learning Algorithms and Its Applications." *The Institute of Electrical and Electronics Engineers, Inc. (IEEE) Conference Proceedings*, The Institute of Electrical and Electronics Engineers, Inc. (IEEE), 2021, https://doi.org/10.1109/ICESC51422.2021.9532767.
2. Zheng, Baokun, et al. "Identifying the Vulnerabilities of Bitcoin Anonymous Mechanism Based on Address Clustering." *Science China. Information Sciences*, vol. 63, no. 3, 2020, pp. 132101-, https://doi.org/10.1007/s11432-019-9900-9.
3. Jackson, Jo. *Machine Learning : Proceedings of the Ninth International Workshop (ML92)*. Edited by D. Sleeman et al., Morgan Kaufmann Publishers, 1992.
4. Hahsler, Michael, and Matthew Bolaos. "Clustering Data Streams Based on Shared Density between Micro-Clusters." *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 6, 2016, pp. 1449–61, https://doi.org/10.1109/TKDE.2016.2522412.
5. Küçük Matci, Dilek, and Uğur Avdan. "Address Standardization Using the Natural Language Process for Improving Geocoding Results." *Computers, Environment and Urban Systems*, vol. 70, 2018, pp. 1–8, https://doi.org/10.1016/j.compenvurbsys.2018.01.009.