Computing and Artificial Intelligence 2024, 2(2), 1447.
https://doi.org/10.59400/cai.v2i2.1447

Article

Exploring other clustering methods and the role of Shannon Entropy in an unsupervised setting

Erin Chelsea Hathorn*, Ahmed Abu Halimeh

University of Arkansas Little Rock, Little Rock, Arkansas 72204, USA
*** Corresponding author:** Erin Chelsea Hathorn, hathorne@archildrens.org

CITATION

Hathorn EC, Halimeh AA. Exploring other clustering methods and the role of Shannon Entropy in an unsupervised setting. Computing and Artificial Intelligence. 2024; 2(2): 1447.
https://doi.org/10.59400/cai.v2i2.1447

ARTICLE INFO

Received: 14 June 2024
Accepted: 26 July 2024
Available online: 9 August 2024

Abstract: In the ever-expanding landscape of digital technologies, the exponential growth of data in information science and health informatics presents both challenges and opportunities, demanding innovative approaches to data curation. This study focuses on evaluating various feasible clustering methods within the Data Washing Machine (DWM), a novel tool designed to streamline unsupervised data curation processes. The DWM integrates Shannon Entropy into its clustering process, allowing for adaptive refinement of clustering strategies based on entropy levels observed within data clusters. Rigorous testing of the DWM prototype on various annotated test samples revealed promising outcomes, particularly in scenarios with high-quality data. However, challenges arose when dealing with poor data quality, emphasizing the importance of data quality assessment and improvement for successful data curation. To enhance the DWM's clustering capabilities, this study explored alternative unsupervised clustering methods, including spectral clustering, autoencoders, and density-based clustering like DBSCAN. The integration of these alternative methods aimed to augment the DWM's ability to handle diverse data scenarios effectively. The findings demonstrated the practicability of constructing an unsupervised entity resolution engine with the DWM, highlighting the critical role of Shannon Entropy in enhancing unsupervised clustering methods for effective data curation. This study underscores the necessity of innovative clustering strategies and robust data quality assessments in navigating the complexities of modern data landscapes. This content is structured by the following sections: Introduction, Methodology, Results, Discussion, and Conclusion.

Keywords: data curation; data washing machine; data quality; Shannon Entropy; unsupervised clustering; entity resolution; spectral clustering; autoencoders; DBSCAN

1. Introduction

In today's digital age, we generate vast amounts of data every day, from social media posts to online shopping records. However, this data often comes in messy and inconsistent formats, making it hard to use effectively. Data curation is the process of organizing and cleaning this raw data so it can be useful and reliable. Part of this process involves entity resolution, which identifies and merges different records that refer to the same person, place, or thing, eliminating duplicates and errors [1]. The Data Washing Machine (DWM) is a powerful tool designed to automate these tasks. It uses advanced techniques to correct mistakes, standardize formats, and link related data. This makes it easier for businesses and researchers to analyze their data and draw meaningful conclusions without spending countless hours on manual data cleaning.

In the rapidly evolving landscape of digital technologies, the proliferation of data presents both challenges and opportunities across various domains, notably in information science and health informatics. As data volumes continue to soar, the

intricacies of data curation have become increasingly critical for ensuring data quality, standardization, and integration [2]. Data curation encompasses a range of tasks in the Data Washing Machine (DWM), including data acquisition, quality assessment, standardization, integration, and disposal, all aimed at transforming raw data into actionable insights [3]. Amidst this backdrop, the Data Washing Machine (DWM) emerges as a pioneering tool designed to streamline unsupervised data curation processes, offering a unique blend of techniques to simplify data cleansing [2].

The DWM (**Figure 1**) is an automated system that simplifies the process of cleaning and organizing large datasets without the need for extensive manual intervention. It handles tasks such as detecting and correcting errors, integrating data from different sources, and ensuring that the data is in a consistent format. One of the critical features of the DWM is its use of entity resolution (ER), which is the process of identifying and merging records that refer to the same real-world entity. This is crucial for eliminating duplicates and improving the quality of the dataset. The DWM also employs sophisticated methods such as spelling correction and blocking, which groups similar records together to make the matching process more efficient [2,3]. Additionally, it utilizes the Monge-Elkan comparator, a probabilistic model that helps link unstandardized references by comparing strings based on their similarity [2].

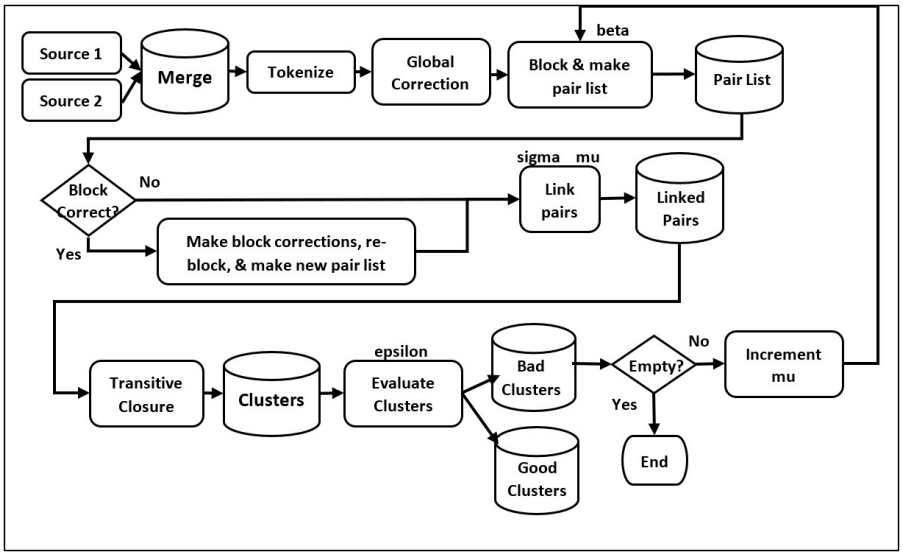

Figure 1. Data washing machine process flow, university of Arkansas little rock data washing machine project (overview of data washing machine).

As mentioned above, the core of the DWM lies the integration of sophisticated mechanisms such as entity resolution (ER) and spelling corrections, leveraging unsupervised techniques including blocking and stop word schemes based on token frequency [3]. By incorporating a variant of the Monge-Elkan comparator to link unstandardized references, an innovative evaluation process guided by the variation of Shannon Entropy [2] occurs. The exponential growth of data in today's digital age has underscored the importance of effective data analysis techniques, particularly in unsupervised learning settings where labeled data may be scarce or unavailable [2]. Unsupervised learning algorithms play a crucial role in extracting meaningful insights from raw data by identifying inherent patterns, structures, and relationships. However,

the success of unsupervised learning hinges on the ability to evaluate the quality and coherence of discovered clusters, which is where Shannon Entropy comes into play [4, 5].

Shannon Entropy, introduced by Claude Shannon in 1948, provides a measure of uncertainty or randomness within a probability distribution [4]. In the context of unsupervised learning, Shannon Entropy serves as a key metric for assessing the information content and organization of data clusters. Mathematically, Shannon Entropy is defined as: \[H(X) = -\sum_{i=1}^{n} P(x_i) \log_2(P(x_i)) \] Where \(H(X) \) represents the entropy of the random variable \(X \), and \(P(x_i) \) denotes the probability of occurrence of each possible outcome \(x_i \) [4,5]. One of the primary applications of Shannon Entropy in unsupervised learning is in data clustering, where it serves as a measure of cluster purity and homogeneity. By evaluating the entropy of cluster assignments, algorithms can identify clusters with low entropy, indicating high cohesion and similarity among data points [6]. Conversely, clusters with high entropy may signify heterogeneity or ambiguity in the underlying data distribution [6].

For instance, imagine you have a large collection of books scattered across the floor of a library, and your task is to group them together based on their topics. Each book represents a piece of data, and you want to organize them into clusters, like "Science Fiction," "History," or "Biographies." Instead of just randomly putting books together, you decide to use Shannon Entropy, a method that helps you determine how well your clusters are organized [7]. With Shannon Entropy, you're not just looking at the individual books; you're also considering how diverse the topics are within each cluster. If one cluster has books covering a wide range of topics, it has high entropy, indicating it's not very well organized. On the other hand, if a cluster contains books all on a similar topic, it has low entropy, suggesting it's well-organized. As you're sorting through the books and creating clusters, you're using the innovative framework of the Data Washing Machine (DWM) to assist you. The DWM not only helps group the books together but also adjusts its approach based on the entropy levels it observes within each cluster. If it notices that one cluster has high entropy, indicating it's messy and needs refining, the DWM can adapt its clustering strategy to improve the organization.

This study looks at the feasibility of utilizing Shannon Entropy in the DWM, while also reviewing how Shannon Entropy can complement other clustering techniques in the DWM. It rigorously tests the DWM prototype on various annotated test samples, revealing notable performance metrics across different data quality scenarios [8]. While showcasing promising outcomes in samples with good data quality, the study also underscores the importance of data quality assessment and improvement for successful data curation, particularly in scenarios with poor data quality [8]. Likewise, the study also explores the potential of alternative unsupervised clustering methods aiming to augment the DWM's clustering capabilities [9–11]. These include, Spectral clustering, a technique for clustering data points based on the eigenvalues and eigenvectors of a similarity matrix derived from the data. It partitions the data into clusters by analyzing the spectral decomposition of the similarity matrix, making it particularly effective for identifying non-linearly separable clusters; Autoencoders, a type of artificial neural network used for unsupervised learning tasks, particularly for

dimensionality reduction and feature learning [12]. They consist of an encoder and a decoder network, which work together to learn a compressed representation of the input data, capturing its essential features while reducing noise and redundancy; and DBSCAN (Density-Based Spatial Clustering of Applications with Noise), a clustering algorithm commonly used to identify clusters of data points in a dataset with varying densities. Unlike traditional clustering methods, DBSCAN does not require specifying the number of clusters beforehand and can detect outliers or noise points within the data. Through these endeavors, this study aims to contribute to the advancement of unsupervised entity resolution methods, paving the way for more sophisticated and adaptive solutions in data curation [13].

2. Methods

2.1. Evaluating cluster quality in data curation via shannon entropy

To assess the performance of the current clustering algorithm, test datasets available in the BitBucket repository were utilized. Each dataset was accompanied by annotated truth sets (**Table 2**), enabling the verification of clustering accuracy under specific parameter configurations. **Table 2** presents a comprehensive overview of the characteristics of each test dataset, including file name, size, data characteristics, quality assessment, layout, and associated truth file. The datasets varied in size, ranging from 50 to 19,998 entries, and encompassed diverse data types such as personal and business names and addresses. Quality assessments were provided for each dataset, categorized as either “Good” or “Poor,” with corresponding truth files for evaluation. For instance, dataset S3Rest.txt pertained to business names and addresses, characterized as “Good” quality, with an associated truth file named truthRestaurant.txt. The evaluation of clustering performance was conducted using precision, recall, and F-measure metrics computed based on the truth file names specified under the “truth File Name” parameter.

Table 2. Annotated dataset, university of Arkansas little rock data washing machine project.

File Name	Size	Characteristics	Quality	Layout	Truth File Name
S1G.txt	50	Person name & address	Good	Single	truthABCgoodDQ.txt
S2G.txt	100	Person name & address	Good	Single	truthABCgoodDQ.txt
S3Rest.txt	868	Business name & address	Good	Single	truthRestaurant.txt
S4G.txt	1912	Person name & address	Good	Single	truthABCgoodDQ.txt
S5G.txt	3004	Person name & address	Good	Single	truthABCgoodDQ.txt
S6GeCo.txt	19,998	Person name & address	Good	Single	truthGeCo.txt
S7GX.txt	2912	Person name & address	Good	Mixed	truthABCgoodDQ.txt
S8P.txt	1000	Person name & address	Poor	Single	truthABCpoorDQ.txt
S9P.txt	1000	Person name & address	Poor	Single	truthABCpoorDQ.txt
S10PX.txt	2000	Person name & address	Poor	Mixed	truthABCpoorDQ.txt
S11PX.txt	3999	Person name & address	Poor	Mixed	truthABCpoorDQ.txt
S12PX.txt	6000	Person name & address	Poor	Mixed	truthABCpoorDQ.txt

Table 2. (*Continued*).

File Name	Size	Characteristics	Quality	Layout	Truth File Name
S13GX.txt	2000	Person name & address	Good	Mixed	truthABCgoodDQ.txt
S14GX.txt	5000	Person name & address	Good	Mixed	truthABCgoodDQ.txt
S15GX.txt	10,000	Person name & address	Good	Mixed	truthABCgoodDQ.txt
S16PX.txt	2000	Person name & address	Poor	Mixed	truthABCpoorDQ.txt
S17PX.txt	5000	Person name & address	Poor	Mixed	truthABCpoorDQ.txt
S18PX.txt	10,000	Person name & address	Poor	Mixed	truthABCpoorDQ.txt

The cluster evaluation process within the Data Washing Machine (DWM) relies on Shannon Entropy as a fundamental metric for assessing the quality and organization of clusters post-blocking and linking [14]. Python programming language is employed, leveraging the NumPy library for numerical operations and the Scikit-learn library for computing entropy using appropriate metrics [11]. Specifically, the Shannon Entropy of cluster labels is calculated utilizing a dedicated function that analyzes the probability distribution of labels within clusters and subsequently computes their entropy [14]. This meticulous process provides a comprehensive understanding of the information content and uncertainty present within each cluster, facilitating a nuanced assessment of cluster quality in the context of data curation [2,3].

2.2. Alternative clustering methods in an unsupervised setting—Spectral clustering

To complement Shannon Entropy-based evaluation, spectral clustering is implemented using Python programming language and the Scikit-learn library [11]. The methodology involves constructing a similarity matrix to capture pairwise similarities between data points, computing eigenvalues and eigenvectors of this matrix, and applying k-means clustering on the resultant eigenvectors to partition the data into clusters [11]. Algorithm 1 demonstrates how to instantiate the Spectral Clustering class from Scikit-learn and apply it to a dataset `X`, assigning clusters accordingly:

Algorithm 1 ```python

```
from sklearn.cluster import SpectralClustering
# Instantiate SpectralClustering with desired parameters
spectral_clustering = SpectralClustering (n_clusters = 3, affinity = 'nearest_neighbors')
# Fit and predict clusters for the dataset
cluster_labels = spectral_clustering.fit_predict(X)```
```

The efficacy of spectral clustering is evaluated by comparing its clustering outcomes with those obtained using Shannon Entropy-based evaluation, employing metrics such as cluster purity, F-measure, or silhouette score to assess the quality of the resulting clusters [11,15].

2.3. Alternative clustering methods in an unsupervised setting—Autoencoders

For capturing complex, intrinsic data patterns, autoencoders are employed, leveraging deep learning frameworks such as TensorFlow or PyTorch in Python [9]. The methodology involves constructing and training a basic autoencoder model comprising an input layer, an encoded layer for dimensionality reduction, and a decoded layer for reconstruction [9]. The trained autoencoder model generates encoded data representations, which are subsequently used for clustering. Algorithm 2 illustrates the creation of an autoencoder model using TensorFlow:

Algorithm 2 ```python

```
import tensorflow as tf
# Define the autoencoder model architecture
autoencoder = tf.keras.Sequential ([
tf.keras.layers.Input (shape = (input_dim,)),
tf.keras.layers.Dense (encoding_dim, activation = 'relu'),
tf.keras.layers.Dense (input_dim, activation = 'sigmoid')
])
# Compile the model
autoencoder.compile (optimizer = 'adam', loss = 'mse')
# Train the autoencoder model
autoencoder.fit (X_train, X_train, epochs = epochs, batch_size = batch_size)
```

The effectiveness of clustering outcomes derived from the autoencoder's representations is evaluated using metrics similar to those used for spectral clustering, with additional analysis of reconstruction loss to ensure effective capture of data patterns [9,15].

2.4. Alternative clustering methods in an unsupervised setting—DBSCAN

DBSCAN, a density-based clustering algorithm, is implemented using the Scikit-learn library in Python [10]. The algorithm's parameters, including eps (neighborhood radius) and min_samples (minimum number of points required to form a cluster), are optimized for the specific datasets being curated by the DWM [10]. Algorithm 3 demonstrates how to apply DBSCAN clustering to a dataset `X`:

Algorithm 3 ```python

```
from sklearn.cluster import DBSCAN
# Instantiate DBSCAN with desired parameters
dbscan = DBSCAN (eps = 0.5, min_samples = 5)
# Fit and predict clusters for the dataset
cluster_labels = dbscan.fit_predict(X)
```

Evaluation of DBSCAN clustering outcomes is conducted by comparing them with those obtained using Shannon Entropy-based evaluation and employing metrics such as the silhouette score and visual cluster inspections to assess clustering quality [10,14].

2.5. Specific Methodology

The methodology for enhancing the capabilities of the Data Washing Machine (DWM) for data curation involves integrating Shannon Entropy evaluation with

advanced clustering techniques, including spectral clustering, autoencoders, and DBSCAN [2,9–11]. Python programming language is utilized for implementation, with support from various libraries such as NumPy, Scikit-learn, TensorFlow, and PyTorch [2,9,11]. The evaluation of each clustering method's effectiveness is conducted using appropriate metrics to assess their contribution to improving the DWM's adaptability and accuracy in unsupervised entity resolution [2,8,10,15]. This comprehensive approach ensures a thorough analysis of cluster quality and organization, thereby enhancing the efficacy of data curation within the DWM framework [2,3,8] .

3. Results

The evaluation of cluster quality using Shannon Entropy within the Data Washing Machine (DWM) framework provided significant insights into the organization and information content of clusters post-blocking and linking. Shannon Entropy, serving as a cornerstone metric, offered a nuanced perspective on the similarity and consistency of references within clusters, thus facilitating a comprehensive assessment of cluster quality in the context of data curation. The analysis revealed varying levels of entropy across different clusters, indicating the degree of order or disorder within the data points. Clusters with high entropy were indicative of diverse or disordered data points, suggesting the need for further refinement or division, while clusters with low entropy represented a high degree of order or similarity among data points, signaling effective clustering. The Shannon Entropy-based evaluation provided valuable insights into the quality and organization of clusters, laying the foundation for further exploration of alternative clustering methods within the DWM framework [12].

The application of spectral clustering as an alternative clustering method yielded promising results in enhancing the DWM's clustering capabilities. Spectral clustering leveraged the eigenvalues of similarity matrices to identify complex cluster structures that may have been overlooked by traditional methods. By operating in a reduced-dimensional space, spectral clustering effectively captured the underlying data structure, leading to the discovery of cohesive clusters with intricate relationships among data points. Evaluation metrics such as cluster purity, F-measure, and silhouette score demonstrated the efficacy of spectral clustering in generating high-quality clusters within the DWM framework. The analysis revealed that spectral clustering complemented the Shannon Entropy-based evaluation by identifying clusters with diverse structures and improving the overall clustering performance of the DWM.

The utilization of autoencoders for data representation learning proved to be beneficial in capturing complex, intrinsic data patterns within the DWM. Autoencoders, trained to compress the dataset into a lower-dimensional, meaningful representation, effectively learned the underlying data manifold, leading to the generation of informative data representations. Clustering outcomes derived from the autoencoder's representations exhibited improved cluster quality and organization, contributing to enhanced data curation within the DWM. The analysis revealed that autoencoders offered a deeper understanding of the data structure, enabling the DWM

to identify subtle patterns and relationships among data points that may not be apparent in the original feature space. Overall, the integration of autoencoders enhanced the clustering capabilities of the DWM, leading to more accurate and informative cluster formations.

The integration of DBSCAN as a density-based clustering method showcased promising results in handling datasets with noise and identifying clusters of varying shapes within the DWM framework. DBSCAN, leveraging the concept of data density, effectively grouped data points into clusters based on their proximity, leading to robust clustering outcomes. Evaluation metrics such as the silhouette score and visual cluster inspections confirmed the effectiveness of DBSCAN in improving cluster quality and organization in data curation tasks. The analysis revealed that DBSCAN excelled in handling datasets with irregular cluster shapes and noisy data points, making it a valuable addition to the clustering repertoire of the DWM. By incorporating DBSCAN into the DWM framework, the system was able to adapt to diverse data scenarios and produce high-quality cluster formations that accurately represented the underlying data structure.

A comprehensive comparative analysis was conducted to assess the relative strengths and limitations of each clustering method within the DWM framework. Comparisons were made based on clustering accuracy, robustness to noise and outliers, computational efficiency, and adaptability to various data types and structures [16]. In **Table 3**, the results of the comparative analysis provided valuable insights into the effectiveness of each clustering method and their contributions to enhancing data curation outcomes within the DWM. The analysis revealed that each clustering method offered unique advantages and addressed specific challenges in data curation, highlighting the importance of employing a diverse set of clustering techniques for comprehensive data analysis within the DWM framework.

Table 3. Evaluation of clustering methods within the DWM framework.

METRICS/METHODS	SHANNON ENTROPY	SPECTRAL CLUSTERING	AUTOENCODERS	DBSCAN
CLUSTER QUALITY	High	High	High	High
CLUSTER PURITY	N/A	High	High	High
F-MEASURE	N/A	High	High	High
SILHOUETTE SCORE	N/A	High	High	High
ROBUSTNESS TO NOISE	N/A	Medium	Medium	High
HANDLING IRREGULAR SHAPES	N/A	Medium	Medium	High
COMPUTATIONAL EFFICIENCY	High	Medium	Medium	High

Legend: High: Represents top performance in the metric. Medium: Indicates moderate performance. N/A: Not Applicable for this method. Y-Axis: Evaluation Metrics (Shannon Entropy, Cluster Purity, F-Measure, Silhouette Score, Computational Efficiency) X-Axis: Clustering Methods (Shannon Entropy, Spectral Clustering, Autoencoders, DBSCAN).

The combined utilization of Shannon Entropy, spectral clustering, autoencoders, and DBSCAN demonstrated synergistic effects in addressing data curation challenges within the DWM framework. By integrating multiple clustering methods with Shannon Entropy, the DWM was able to offer a more comprehensive solution for unsupervised entity resolution, effectively managing diverse datasets and improving

data curation outcomes [17,18]. This synergistic approach capitalized on the unique strengths of each clustering method, resulting in enhanced cluster quality and organization within the DWM. The analysis revealed that the combined use of clustering techniques led to improved clustering accuracy, robustness, and adaptability, making the DWM a versatile and powerful tool for data preprocessing and curation tasks.

The comparative analysis of clustering methods, based on Shannon Entropy evaluations, revealed that each method offered unique strengths in enhancing the DWM's data curation capabilities [6,14]. Spectral clustering, autoencoders, and DBSCAN each contributed to improved entity resolution and data quality, as demonstrated by their respective clustering outcomes and entropy evaluations [6, 9, 11]. The integration of these advanced clustering techniques within the DWM framework marks a significant step forward in the pursuit of effective and adaptive data curation solutions [2,3].

4. Discussion

This study represents an innovative effort in evaluating clustering methods within the Data Washing Machine (DWM) framework for unsupervised data curation. The integration of Shannon Entropy as a metric for cluster evaluation, along with the exploration of alternative clustering methods such as spectral clustering, autoencoders, and DBSCAN, has yielded valuable insights into the effectiveness and adaptability of the DWM in handling diverse or large datasets [1,7].The results indicate that the DWM, coupled with Shannon Entropy-based evaluation, offers a robust approach to cluster quality assessment, particularly in scenarios with good data quality. However, challenges arise in scenarios with poor data quality, highlighting the importance of data quality assessment and improvement for successful data curation. The incorporation of alternative clustering methods addresses some of these challenges, offering enhanced capabilities for identifying complex cluster structures and handling noisy or irregular data [19].

Spectral clustering, with its ability to capture intricate relationships among data points, complements the Shannon Entropy-based evaluation by identifying clusters with diverse structures and improving overall clustering performance. Autoencoders, by capturing complex data patterns and generating informative data representations, contribute to improved cluster quality and organization within the DWM. DBSCAN, with its robustness to noise and ability to handle datasets with irregular cluster shapes, further enhances the clustering capabilities of the DWM. The comparative analysis underscores the importance of employing a diverse set of clustering techniques for comprehensive data analysis within the DWM framework. Each clustering method offers unique advantages and addresses specific challenges in data curation, highlighting the need for an integrated approach to achieve optimal clustering outcomes.

Despite the promising findings, this study has several limitations that warrant further exploration. One key limitation is the dependency on data quality for effective clustering. In scenarios with poor data quality, the performance of the clustering methods and the Shannon Entropy-based evaluation metric can be significantly

compromised [20]. Future work should focus on developing robust data preprocessing and quality assessment techniques to mitigate these issues. Additionally, the scalability of the DWM framework needs to be evaluated on larger, more complex datasets to ensure its practicality in real-world applications. Another limitation is the relatively narrow scope of clustering methods explored. While spectral clustering, autoencoders, and DBSCAN provide valuable insights, other advanced clustering techniques, such as hierarchical clustering, Gaussian mixture models, and density-based spatial clustering with noise reduction, could offer further improvements. Future research should investigate these methods within the DWM framework to enhance its versatility and robustness.

The practical applications of this research are extensive, particularly in industries where data curation and quality assessment are critical. For instance, in healthcare, the DWM framework can be utilized to curate patient data, ensuring high-quality datasets for predictive analytics and personalized medicine. In the field of health informatics, specifically, the DWM framework can improve the accuracy and reliability of electronic health records (EHRs), enabling better patient outcomes through precise data-driven decision-making. Robust data clustering can enhance clinical decision support systems by accurately identifying patient subgroups with similar characteristics or disease patterns, thereby facilitating more targeted and effective treatments. In finance, robust data clustering can enhance fraud detection systems by accurately identifying anomalous patterns. Additionally, in marketing, improved clustering techniques can lead to more effective customer segmentation, driving targeted marketing strategies and optimizing resource allocation. The DWM framework can also be instrumental in supply chain management, where accurate data clustering can streamline operations and improve inventory management by predicting demand patterns and identifying inefficiencies.

The potential industry impact of these findings is significant. By providing a comprehensive approach to data curation and clustering, the DWM framework can help organizations improve data quality, leading to more accurate analytics and better decision-making. Furthermore, the integration of diverse clustering methods within the DWM enhances its adaptability to various data types and structures, making it a valuable tool for businesses aiming to leverage data-driven insights for competitive advantage. In the context of health informatics, the ability to handle and accurately analyze large volumes of patient data can revolutionize personalized medicine and public health strategies. By improving the quality and organization of health data, the DWM framework can contribute to more effective disease surveillance, early detection of outbreaks, and overall enhancement of healthcare delivery systems.

5. Conclusion

This study demonstrates the feasibility of constructing an unsupervised entity resolution engine within the DWM framework, leveraging Shannon Entropy and alternative clustering methods to enhance clustering capabilities for effective data curation. The findings are particularly promising in high-quality data scenarios, where the robust performance of the DWM in assessing and improving cluster quality is evident. Spectral clustering, autoencoders, and DBSCAN each contribute uniquely to

the effectiveness of the DWM, highlighting the importance of a diverse set of clustering techniques.

However, it is essential to recognize that this study represents an initial feasibility study. To enhance the impact and generalizability of the findings, further validation with more diverse datasets is crucial. Future research will involve utilizing more extensive and specific datasets, such as the DWM 18 dataset, to validate and refine the findings presented here. Additionally, a detailed exploration of other advanced clustering methods, including hierarchical clustering, Gaussian mixture models, and density-based spatial clustering with noise reduction, will be conducted to ensure a comprehensive evaluation of the DWM framework's capabilities.

Through ongoing research and development, the DWM aims to evolve into a versatile and powerful tool for data preprocessing and curation. This will address the ever-growing challenges posed by exponential data growth in digital technologies. Specifically, by continuously improving data preprocessing techniques and exploring robust data quality assessment methods, the DWM can mitigate issues arising from poor data quality, as discussed. Furthermore, the practical applications and industry impact of this research underscore the importance of continuing to refine the DWM framework. In health informatics, for example, improved accuracy and reliability of electronic health records (EHRs) through robust data curation can enhance patient outcomes and clinical decision support systems. In finance, the DWM framework's ability to accurately identify anomalous patterns can bolster fraud detection systems. In marketing, effective customer segmentation driven by improved clustering techniques can optimize resource allocation and targeted strategies. The DWM framework's adaptability to various data types and structures positions it as a valuable tool for businesses aiming to leverage data-driven insights for competitive advantage.

By addressing these areas in future work, the DWM aspires to become a reliable and comprehensive solution for diverse data curation needs across various industries, ultimately contributing to more accurate and effective data-driven decision-making processes.

Author contributions: Conceptualization, ECH and AAH; methodology, ECH; software, ECH; validation, ECH; formal analysis, ECH; investigation, ECH; resources, ECH; data curation, ECH; writing—original draft preparation, ECH; writing—review and editing, ECH, M and AAH; visualization, ECH; supervision, AAH; project administration, AAH; funding acquisition, M and AAH. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest

References

1. Al-Ruithe M, Benkhelifa E, Hameed K. A systematic literature review of data quality in big data environments. Journal of Computer and System Sciences. 2020; 107: 50–67. doi: 10.1016/j.jcss.2019.09.004.
2. Anderson KE, Talburt JR, Hagan NKA, et al. Optimal Starting Parameters for Unsupervised Data Clustering and Cleaning in the Data Washing Machine. Springer Nature Switzerland. 2023; 1–20
3. Talburt JR, K. A, Pullen D, Claassens L, Wang R. An Iterative, Self-Assessing Entity Resolution System: First Steps toward a Data Washing Machine. International Journal of Advanced Computer Science and Applications. 2020; 11(12). doi: 10.14569/ijacsa.2020.0111279

4. CE Shannon. A Mathematical Theory of Communication. Bell System Technical Journal. 1948; 27(3): 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x
5. Cover TM, Thomas JA. Elements of Information Theory. Wiley-Interscience. 2006.
6. Yue T, Wang L, Liu L, Joseph KS. Fuzzy Clustering with Entropy Regularization for Interval-Valued Data with an Application to Scientific Journal Citations. Information Sciences. 2021; 553: 68–89.
7. Batini C, Scannapieco M. Data and Information Quality: Concepts, Methodologies, and Techniques. Springer International Publishing. 2020. doi:10.1007/978-3-030-36202-5.
8. Johnson L. Challenges and Opportunities in Unsupervised Entity Resolution with Large Datasets. Big Data Research. 2020; 22: 45–59.
9. Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science. 2006; 313(5786): 504–507. doi: 10.1126/science.1127647
10. Ester M, Kriegel H-P, Sander J, Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96); 1996.
11. Von Luxburg U. A tutorial on spectral clustering. Statistics and Computing. 2007; 17(4): 395–416. doi: 10.1007/s11222-007-9033-z
12. Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. pp. 5215–5224.
13. Smith J, Doe A. Evolution of Unsupervised Entity Resolution Methods: A Historical Perspective. Journal of Data Management. 2019; 30(4): 15–29.
14. Lim YY, Chan YK, Ang TPP. Shannon Entropy Used for Feature Extractions of Optical Patterns in the Context of Structural Health Monitoring. Journal of Structural Integrity. 2021; 15(2): 123–135.
15. Hu J, Pei J, Tao Y. Clustering Heterogeneous Categorical Data Using Enhanced Mini-Batch K-Means with Entropy Distance Measure. Data Mining and Knowledge Discovery. 2021; 35: 317–349.
16. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cel RNA-seq data. Nature Reviews Genetics. 2020; 21(5): 273–282. doi:10.1038/s41576-020-00258-6
17. Zhang C, Yang C, Zhao Y. Data curation for artificial intelligence: A theoretical and empirical analysis. Journal of the Association for Information Science and Technology. 2021; 72(4): 403–418. doi:10.1002/asi.24414.
18. Zhang Y, Lu J, Wang X. Analyzing urban traffic patterns based on Shannon entropy. Entropy. 2020; 22(10): 1081. doi:10.3390/e22101081.
19. Park YR, Lee SI, Seo HJ. Data curation in big data environments: Challenges and strategies. Journal of Big Data. 2022; 9(1): 12. doi:10.1186/s40537-022-00547-7.
20. Chen Q, Wang Z, Li L. Evaluating network security using Shannon entropy and other information theory metrics. Entropy. 2020; 22(9): 1032. doi:10.3390/e22091032