GONZALES, Mark Edward M.
SY, Jacob Adrianne L.
UY, Lorene C.

Automatic Recommendation of Distance Metric for *k*-Means Clustering: A Meta-Learning Approach

Background of the Study:

- **Meta-Learning**
  - Meta-learning — derived from the concept of "learning to learn" in educational psychology — is a subfield of machine learning that explores the automatic recommendation of algorithms, as well as the improvement of their performance [1].
  - These recommendations for the selection or improvement of algorithms are determined using meta-data, which consist of meta-features (characteristics of the dataset, such as the number of instances and feature entropy) and performance evaluations on training datasets [2].
  - A problem with the current machine learning pipeline is that each stage involves a selection from a wide array of alternatives [3]; for example, in clustering, researchers are presented with several possible distance metrics. Following a theoretical approach in selecting the optimal option requires specialized knowledge that may not be available to all researchers; meanwhile, brute-forcing through all the options is computationally expensive and resource-intensive [4].
  - Meta-learning aims to address this problem by extracting general concepts when learning tasks and adaptively applying them to new tasks [5].

- **Clustering and Distance Metrics**
  - As clustering algorithms function based on the distance or similarity between one data point and another, the distance metric to be used in the algorithm implementation plays a critical role in its performance; the impact of the choice of the distance metric on the clustering quality has also been empirically demonstrated in several researches [5, 6, 7].
  - In the specific case of centroid-based clustering algorithms (e.g., *k*-means), the distance metric is used in determining the cluster assignment of a data point, the impact of which on the overall clustering quality has also been reported in studies [8, 9].
  - In practice, although the Euclidean (or L_2) distance is the most widely used distance metric [10], it is not suitable for every problem. Specifically, in view of the "curse of high dimensionality," the performance of L_k norms (e.g., Manhattan and Euclidean distances) is sensitive to the value of k as the number of dimensions increases, with the Euclidean distance becoming less preferable both theoretically and empirically when applied to high-dimensional datasets [11].
  - Although the optimal distance measure for a specific clustering task may be selected through theoretical or experimental methods, the former approach hinges on deep domain expertise and understanding of the geometry of the dataset while the latter demands a significant amount of time and resources [3].

- **Application of Meta-Learning to Clustering**
  - Recent researches have extended the notion of meta-learning from purely algorithm selection to the recommendation of choices in other stages of the machine learning pipeline. For instance, in clustering tasks, meta-learning has been applied to recommend not only the most suitable algorithm [12, 13] but also the number of clusters [14] and the cluster validity index [15],
  - However, there is currently a literature gap in the use of meta-learning to recommend distance measures. In 2020, Zhu, Li, Wang, Zheng, and Fu [3] worked on the first published study on a meta-learning model that performs better compared to defaulting to fixed or arbitrarily chosen metrics. An assessment of its methodology opens some room for improvement:
    - Their model used structural meta-features related to 1-itemsets and 2-itemsets, but it only considered a limited set of general[1] (number of instances, number of features, and data dimensionality) and distance-based meta-features (mean, variance, standard deviation, skewness, kurtosis, and percentage of normalized distances belonging to given intervals).
    - It only used a single information-theoretic meta-feature: mean normalized feature entropy.
    - The paper did not provide a comparative investigation or a justification for the machine learning algorithm employed in the construction of the meta-learning model: random forest for the single-metric recommendation and *k*-nearest neighbors for the ranking of metrics.
    - The model achieved an F-measure that is only between 0.3 and 0.4 for the single-metric recommendation and a Spearman's rank correlation coefficient of only 0.2 for the task of ranking the distance metrics based on suitability.
  - A systematic typology of meta-features can be found in [16]: general, statistical, information-theoretic, complexity, model-based, and landmarking. A similar list can also be found in [17].
    - Since this is a general-purpose typology (i.e., it is not tailored solely for clustering tasks), some categories, such as model-based and landmarking, are only for supervised tasks.
    - This typology is also not exhaustive. For instance, it does not include the distance-based meta-features employed in [3]. Conversely, the model in [3] does not consider complexity meta-features (e.g., the average number of PCA dimensions per point [17, 18]).
  - Therefore, it may be interesting to investigate whether selecting applicable meta-features from [3, 16, 17] and combining them can help improve the performance of meta-learning models for the automatic recommendation of distance metrics for *k*-means clustering.

Problem Statement:

- Despite the known impact of distance metrics on the performance of clustering algorithms, there exists a literature gap in applying meta-learning to recommend distance metrics. While this idea was explored in [3], this study only considered a limited set of features and algorithms in the model construction, which may have curtailed its performance.
- The proposed research seeks to augment this work by combining selected meta-features from [3, 16, 17] and considering additional machine learning algorithms in creating the meta-learning model.

[1] In [3], they are classified as statistical features. However, in [16], they are termed general (or simple) features. For consistency of terminology in this proposal, the number of instances, number of features, and data dimensionality are classified as general features.

Objective:

- Implement a machine learning-based meta-learning model that utilizes (1) general, (2) statistical, (3) information-theoretic, (4) distance-based, (5) structural, and (6) complexity meta-features to automatically recommend a distance metric for *k*-means clustering
- Evaluate and compare the clustering quality using the meta-learning model's recommended distance metric against the clustering quality using fixed and randomly chosen distance measures

Significance of the Study:

- By exploring a meta-learning approach, this study provides machine learning researchers and practitioners an alternative paradigm to the selection of distance metrics for clustering tasks, which is less resource-intensive than a brute-force trial of possible distance measures and possibly more effective than simply defaulting to fixed or arbitrarily chosen measures.
- The meta-learning approach in this study can be incorporated or contextualized to domain-specific tasks that leverage clustering, such as grouping search results in information retrieval, detecting atmospheric trends in meteorology, finding patterns related to disease occurrence in epidemiology, analyzing genomes in biology, and identifying customer segments in business [19].
- With the current interest in adaptive approaches and auto-machine learning or AutoML [15, 20, 21], the results of this study can be used as a basis for comparison in evaluating the performance of subsequent automatic recommendation systems.

Methodology:

- **Data Collection**
  - This project will use the collection of datasets published by Pimentel [22] in OpenML [23].
    - This collection contains 217 datasets from across different task domains, including but not limited to medicine, biology, meteorology, physics, robotics, and engineering. It also includes well-known datasets used in clustering tasks, such as the Iris, Wine, and Zoo datasets.
    - These datasets have ground-truth assignments; this is advantageous for the project since both external and internal validation indices can be used to evaluate the clustering quality.
    - The suitability of this collection of datasets for meta-learning studies related to clustering is supported by its use in several studies [3, 12, 14, 24].
    - A preliminary description of this collection of datasets in view of the number of attributes and instances is given in [14]:

	Attributes	Instances
Min	2	100
Max	168	1000000
Mean	19.9636	4799.8545
Mode	5	100
Standard Deviation	23.9922	67249.5303
Skewness	2.5983	14.7309
Kurtosis	8.8355	215.0027

- Further exploratory data analysis (EDA) will be done during the course of the project.
- **As of writing, this collection of datasets has already been downloaded by the researchers.**

- **Labeling**
  - As of writing, the researchers have sent an email to the authors of [3], requesting a copy of the datasets labeled with the optimal distance metrics.
  - If no reply is received from the authors of [3], the researchers will label the datasets with the optimal distance metric based on the adjusted rand index (ARI) and the optimal distance metric based on the Davies-Bouldin index (DBI) when *k*-means clustering is performed.
    - Following [12, 14], the predictive attributes will be normalized to the range [0, 1].
    - ARI is an external cluster validation index (i.e., it requires ground-truth assignments) while DBI is an internal index that measures the separation between clusters.
    - Following [12], the value of *k* (number of clusters) in *k*-means clustering will be based on the ground-truth assignments.
    - Following [3], nine (9) distance metrics will be considered:
      - Manhattan distance
      - Euclidean distance
      - Chebyshev distance
      - Standardized Euclidean distance
      - Canberra distance
      - Mahalanobis distance
      - Cosine distance
      - Adjusted cosine distance
      - Pearson correlation distance

- **Meta-Feature Extraction**
  - Combining the list of applicable meta-features in [3, 16, 17], the meta-features enumerated below will be collected from each dataset after undergoing normalization. Since clustering is unsupervised, the meta-features will be limited to only those that apply to unsupervised tasks:

    - **General Meta-Features**
      - Ratio between the number of attributes and instances
      - Ratio between the number of categorical and numeric features
      - Ratio between the number of instances and attributes
      - Total number of attributes
      - Number of binary attributes
      - Number of categorical attributes
      - Number of instances (rows)
      - Number of numeric features

    - **Statistical Meta-Features**
      - Canonical correlations
      - Correlation
      - Covariance
      - Eigenvalues
      - Geometric mean
      - Number of distinct highly correlated pairs of attributes
      - Number of attributes with at least one outlier value
      - Harmonic mean

- Interquartile range
- Kurtosis
- Median absolute deviation
- Maximum
- Mean
- Median
- Minimum
- Range of each attribute
- Standard deviation
- Skewness
- Sparsity metric
- Trimmed mean
- Variance

- **Information-Theoretic Meta-Features**
  - Concentration coefficient of each pair of distinct attributes
  - Shannon's entropy for each predictive attribute

- **Distance-Based Meta-Features**
  - Mean of the Euclidean distance values between all pairs of data points in the dataset (referred to as "distance values" in this listing)
  - Standard deviation of the distance values
  - Variance of the distance values
  - Skewness of the distance values
  - Kurtosis of the distance values
  - Percentage of normalized distance values in the interval [0, 0.1], …, (0.9, 1.0]
  - Percentage of standardized distance values in the interval [0, 1), … [3, ∞)

- **Structural Meta-Features**
  - [minimum, $\frac{1}{8}$ quantile, $\frac{2}{8}$ quantile, …, maximum] of frequencies of one-itemsets
  - [minimum, $\frac{1}{8}$ quantile, $\frac{2}{8}$ quantile, …, maximum] of frequencies of two-itemsets

- **Complexity Meta-Features**
  - Average number of features per dimension
  - Average number of PCA dimensions per point
  - Ratio of the PCA dimension to the original dimension

- **Model Construction**
  - The collection of datasets will be subjected to a 70%-30% train-test split.
  - The meta-features will be fed to meta-learning models. The tentative list of machine learning algorithms to be considered in building the meta-learning models is as follows:
    - Support vector machine
    - Logistic regression
    - Naive Bayes
    - *k*-nearest neighbors
    - Perceptron
    - Random forest
    - Stochastic gradient descent
    - XGBoost
  - Ten-fold cross-validation and hyperparameter tuning will be performed to optimize F1 scores.

- **Performance Evaluation**
  - Since the meta-learning approach in this study frames the distance metric recommendation as a single-label classification task, their F1 scores will be computed to evaluate the performance of the meta-learning models.
  - Ablation experiments will also be done to examine the performance of the models upon removal of meta-feature categories.
  - *k*-means clustering will be performed using the distance metric recommended by the best-performing model (based on F1). Following [3], its mean recommendation accuracy will be computed. The recommendation accuracy for a dataset d is given by $\frac{E_{recommended}(d)-E_{worst}(d)}{E_{best}(d)-E_{worst}(d)}$.
    - $E_{best}(d)$ and $E_{worst}(d)$ are the results of evaluating the clustering quality if the distance measures that perform the best and worst on d are used.
    - $E_{recommended}(d)$ is the result of evaluating the clustering quality if the distance measure recommended by the model is used.
    - The clustering quality will be evaluated using two indices: the adjusted rand index (external cluster validation index) and the Davies-Bouldin index (internal cluster validation index).
  - The recommendation accuracy of the best-performing model will be compared against the recommendation accuracies using fixed and randomly chosen distance measures. Hypothesis testing will be done to check if there is a significant difference between their performances.

Target Timeline:

Week	Target Task
0 **(As of writing)**	- The researchers have already downloaded the OpenML collection that contains the 217 datasets to be used in this proposed project. - The researchers have also emailed the authors of [3] to request a copy of the datasets labeled with the optimal distance metrics.
1	Labeling of datasets in the collection (If no response is received from the authors of [3]) - To expedite the process and promote reproducibility, the researchers will use `pyclustering` [25] for *k*-means clustering; this library has also been used in prior studies [26, 27]. - Except for the Mahalanobis, cosine, adjusted cosine, and Pearson correlation distances (to be implemented by the researchers), the distances listed in the methodology are built into this library.
2-3	Meta-feature extraction - To promote reproducibility, the researchers will use `pymfe` [17] for extracting meta-features; this library has been used in prior meta-learning studies [28, 29]. - Except for the distance-based and structural meta-features (to be implemented by the researchers), the meta-features listed in the methodology can be extracted using this library.

4-6	Model construction - To promote reproducibility, the researchers will use `scikit-learn` [30] for the machine learning algorithms.
7-8	Performance evaluation and comparison of the built models
9-10	Finishing of manuscript Two-week allowance for possible roadblocks in previous target tasks

References:

[1] C. Lemke, M. Budka, and B. Gabrys, "Metalearning: a survey of trends and technologies," *Artificial Intelligence Review*, vol. 44, no. 1, pp. 117-130, July 2013.

[2] A. Rivolli, L. P. F. Garcia, C. Soares, J. Vanschoren, and A. C. P. L. F. de Carvalho, "Meta-features for meta-learning," *Knowledge-Based Systems,* vol. 240, March 2022.

[3] X. Zhu, Y. Li, J. Wang, T. Zheng, and J. Fu, "Automatic recommendation of a distance measure for clustering algorithms," *ACM Transactions on Knowledge Discovery from Data,* vol. 15, no. 1, pp. 7-22, December 2020.

[4] C. So, "Exploring meta-learning: Parameterizing the learning-to-learn process for image classification," in Proc. International Conference on Artificial Intelligence and Communication (ICAIIC), 2021, pp. 199-202.

[5] E. P. Xing, A. Y. Ng, M. I. Jordan, S. Russell, "Distance metric learning, with application to clustering with side-information," in NIPS'02: Proc. 15th International Conference on Neural Information Processing Systems, 2002, pp. 521-528.

[6] J. L. Suarez, S. Garcia, and F. Herrera, "A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges," *Neurocomputing,* vol. 425, pp. 300-322, 2021.

[7] R. Quaddoura, H. Faris, I. Aljarah, J. Merelo, and P. Castillo, "Empirical evaluation of distance measures for nearest point with indexing ratio clustering algorithm," in Proc. 12th International Joint Conference on Computational Intelligence (IJCCI 2020), 2020, pp. 430-438.

[8] D. Usman and S. F. Sani, "Performance evaluation of similarity measures for K-means clustering algorithm," *Bayero Journal of Pure and Applied Science,* vol. 12, no. 2, 144-148, December 2019.

[9] M. K. Gupta and P. Chanda, "Effects of similarity/distance metrics on k-means algorithm with respect to its applications in IoT and multimedia: A review," *Multimedia Tools and Applications,* 2021.

[10] R. Shahid, S. Bertazzon, M. L. Knudtson, and W. A. Ghali, "Comparison of distance measures in spatial analytical modeling for health service planning," *BioMed Health Services Research,* vol. 9, no. 200, 2009.

[11] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in ICDT '01: Proc. 8th International Conference on Database Theory, 2001, pp. 420-434.

[12] B. A. Pimentel and A. C. P. L. F. de Carvalho, “Statistical versus distance-based meta-features for clustering algorithm recommendation using meta-learning,” in Proc. 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-8.
[13] Y. Poulakis, C. Doulkeridis, and D. Kyriazis, “AutoClust: A framework for automated clustering based on cluster validity indices,” in Proc. 2020 IEEE International Conference on Data Mining (ICDM), 2020, pp. 1220-1225.
[14] B. A. Pimentel and A. C. P. L. F. de Carvalho, “A Meta-learning approach for recommending the number of clusters for clustering algorithms,” *Knowledge-Based Systems,* vol. 195, May 2020.
[15] S. Muravyov and A. Filchenkov, “Meta-learning system for automated clustering,” in Proc. International Workshop on Automatic Selection, Configuration and Composition of Machine Learning Algorithms co-located with the European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (AutoML@PKDD/ECML), 2017.
[16] J. Vanschoren, “Meta-Learning,” in *Automated Machine Learning*, F. Hutter, L. Kotthoff, and J. Vanschoren, Eds. Cham: Springer, 2019.
[17] E. Alcobaça, F. Siqueira, A. Rivolli, L. P. F. Garcia, J. T. Oliva, and A. C. P. L. F. de Carvalho, “MFE: Towards reproducible meta-feature extraction,” *Journal of Machine Learning Research,* vol. 21, no. 111, pp. 1-5, 2020.
[18] A. C. Lorena, L. P. F. Garcia, J. Lehmann, M. C. P. Souto, and T. K. Ho, “How complex is your classification problem? A survey on measuring classification complexity,” *ACM Computing Surveys,* vol. 52, no. 5, pp. 1-34, September 2020.
[19] P. N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining,* 2nd ed. New York: Pearson.
[20] M. Khodak, M. F. Balcan, and A. Talwalkar, “Adaptive gradient-based meta-learning methods,” in NIPS’19: Proc. 33rd International Conference on Neural Information Processing Systems, 2019, pp. 5917-5928.
[21] H. Rakotoarison, L. Milijaona, A. Rasoanaivo, M. Sebag, and M. Schoenauer, “Learning meta-features for AutoML,” presented at ICLR 2022 - International Conference on Learning Representations, Virtual, 2022.
[22] B. Pimentel, “Datasets,” *OpenML,* 2017. [Online]. Available: https://www.openml.org/s/88/data. [Accessed: Apr. 2, 2022]
[23] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, “OpenML: Networked science in machine learning,” *ACM SIGKDD Explorations Newsletter,* vol. 15, no. 2, pp. 49-60, June 2014.
[24] A. Jilling and M. Alvarez, “Optimizing recommendations for clustering algorithms using meta-learning,” in Proc. 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-10.
[25] A. V. Novikov, “PyClustering: Data mining library,” *The Journal of Open Source Software,* vol. 4, no. 36, p. 1230, April 2019.
[26] M. C. Romão, N. F. Castro, J. G. Milhano, R. Pedro, and T. Vale, “Use of a generalized energy mover’s distance in the search for rare phenomena at colliders,” *The European Physical Journal C,* vol. 81, no. 192, February 2021.
[27] M. Hahsler, M. Piekenbrock, and D. Doran, “dbscan: Fast Density-Based Clustering with R,” *Journal of Statistical Software,* vol. 91, no. 1, October 2019.

[28] M. Kotlar, M. Punt, Z. Radivojević, M. Cvetanović, and V. Milutinović, “Novel meta-features for automated machine learning model selection in anomaly detection,” *IEEE Access,* vol. 9, pp. 89675-89687, June 2021.
[29] J. P. Monteiro, D. Ramos, D. Carneiro, F. Duarte, J. M. Fernandes, and P. Novais, “Meta-learning and the new challenges of machine learning,” *International Journal of Intelligent Systems,* vol. 36, no. 11, pp. 6240-6272, November 2021.
[30] F. Pedregosa *et al.,* “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825-2830, 2011.