Caliphate Journal of Science & Technology (CaJoST)

ISSN: 2705-313X (**PRINT**); 2705-3121 (**ONLINE**) **Research Article**

Open Access Journal available at: https://cajost.com.ng/index.php/files and https://www.ajol.info/index.php/cajost/index

DOI: https://dx.doi.org/10.4314/cajost.v6i1.13

Article Info

Received: 20th December 2023

Revised: 29th March 2024

Accepted: 30th March 2024

Theoretical Utility of Data Value Metric and Genetic Algorithms for Variable Clustering in an Unsupervised Learning Environment

Okpako A. Ejaita[1] and Ojie D. Voke[2]

[1]Department of ICT, Faculty of Computing, University of Delta, Agbor.
[2]Department of Software Engineering, Faculty of Computing, University of Delta, Agbor.

***Corresponding author's email:**

ejaita.okpako@unidel.edu.ng

Cite this: *CaJoST,* 2024, 1, 103-108

Cluster analysis is regarded as one of the most important unsupervised learning tasks, with its natural application in dividing data into meaningful groups, also known as clusters, based on the information in the data by describing the objects in terms of their relationships and capturing the data's natural structure. Many traditional performance evaluation metrics for clustering algorithms abound in the literature, treating various attributes or variables equally when measuring similarity; however, different attributes or variables may contribute differently due to the amount of information they contain, which can vary greatly. Data Value Metric (DVM) is an information theoretic measure based on the concept of mutual information that has been shown to be a good metric for validating data quality and utility in a big data ecosystem and in traditional data. Because it uses a forward selection search strategy, Data Value Metric (DVM) suffers from local minima and loss of diversity in the population; however, hybridizing it with Genetic Algorithm will overcome the problem of local minima because there will be a blend of evolutionary search to ensure a balance between exploration and exploitation of the search space. This paper proposed a hybrid model of the Genetic Algorithm and the Data Value Metric (DVM) as an information-theoretic metric for quantifying the quality and utility of variable clustering selection that can be applied to traditional data.

Keywords: Genetic Algorithm, Convolution Neural Network (CNN), IoTs, Unmet Potential Data value

1. Introduction

Cluster analysis is one of the most important unsupervised learning tasks because of its natural application in dividing data into meaningful groups, also known as clusters, based on the information in the data by describing the objects of their relationships and capturing the natural structure of the data. Clustering is the grouping of specific objects based on their characteristics and similarities so that objects in the same group, known as a cluster, are similar to each other. Cluster analysis has long been used in many different fields, including psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. There is no doubt that real-world datasets contain a large number of features; however, many of these features may be irrelevant or redundant, and as such do not contribute meaningfully to the predictive power of the learning model and may even undermine it. Data cleaning, data integration, data transformation, and data reduction are the steps in data pre-processing (feature subset selection). A few dataset attributes may be redundant because their information is contained in other attributes. More factors can influence the computation time for diagnosis accuracy. Some of the data in the dataset may not be useful for diagnosis and can thus be removed before learning. The goal of feature selection is to find the fewest number of attributes so that the resulting probability distribution of the data classes is as close to the original distribution as possible [1]

The feature set is reduced based on feature relevance and redundancy in relation to the goal. The goal of feature selection is to maximize relevance while decreasing redundancy. It usually entails locating a feature subset that contains only relevant features. The two main clustering criteria are quality and stability. Quality ensures that the final clusters have strong cohesion and discrimination, whereas stability measures the solution's ability to be stable even when some space features are unknown.

Clustering algorithms' performance is highly dependent on cluster configuration, such as the number of clusters and the features or data characteristics. Good, relevant data and artificial intelligence (Machine Learning) are complementary abilities that work synergistically to improve decision support and other allied benefits depending on their areas of application. Redundancies and irrelevant data are potential problems in data that can harm any machine learning algorithm, particularly the clustering algorithm.

Although the Genetic Algorithm and its variants are evolutionary algorithms, they have been used in the literature to cluster data and for variable clustering. However, as Rubiyah (2012) points out, the probability of obtaining an optimal feature subset for classification is high when GA is used for feature selection with appropriate fitness functions and possible considerations.

The Data Value Metric introduced by [13] is a promising metric for feature selection; however, because it uses a sequential forward selection search strategy and is wrapper based, selecting the best features using Data Value Metric will suffer from local minima and redundancies in the result feature subset. The feature space is an important component in ensuring data quality, and doing so in a linear manner is time consuming and adds complexity. According to [4] one of the primary characteristics of the Genetic Algorithm is that it investigates the interdependencies between bits in a string, removing redundancies in the features and selecting the best features that will contribute to the proper clustering process. In light of this, Genetic Algorithms, which have been shown to avoid local minima, are proposed as a search strategy and clustering algorithm, with the Data Value Metric serving as the fitness function.

2. Related Literature

Dimensionality reduction techniques have been used to reduce data from large dimensions to smaller dimensions [8]. To reduce dimension, the most popular dimensionality reduction techniques combine features. Feature selection is one such dimensionality reduction method that selects features from a feature set without modifying them. Three (3) types of feature reduction methods have been identified:

i. Filter method: This method involves ranking features using appropriate criteria, with the highest-ranked features being chosen for application [1] The goal is to eliminate lower-ranked features. The most important aspect of this method is determining a feature's rank or relevance. There are several ranking methods.[1]

a. Correlation criteria: Detects linear dependencies between features. It is calculated using the Pearson correlation coefficient.

b. Mutual information: This metric is used to assess the interdependence of features. A value of 0 indicates that two features are unrelated.

This method has the advantage of being simple to compute and not relying on learning algorithms. The disadvantage is that the features chosen may not be guaranteed to be non-redundant.[1]

ii. Wrapper method: This method relies on classification to identify a feature subset. Exhaustive search methods may produce the best results, but they can be computationally expensive for large datasets. As a result, two types of wrapper methods are available [1]

a. Sequential Search Algorithms: - These algorithms add or remove features until they achieve the desired optimization function. The Sequential Forward Search algorithm begins with an empty set and adds features as they become available. Sequential Backward Search algorithms begin with the entire feature set and gradually eliminate those that fail to meet the performance criteria. It may result in local minima.

b. Heuristic Search Algorithms: - Genetic algorithms can be used to select features, with each chromosome representing the inclusion or exclusion of a set of features.
Although this is a convenient method for selecting features, the main disadvantage is that the entire model must be built and evaluated for each feature subset taken into account.

iii. Embedded methods: Attempts to compensate for the shortcomings of the filter and wrapper methods. It entails algorithms with built-in feature selection methods. This combines the steps of feature selection and performance evaluation into a single step [1]

2.1 Data Value Metric and Related Works

Several studies have proposed metrics for assessing or quantifying a given dataset's information gain. The Value of Information (VOI) analysis is one such metric. It is a decision-theoretical statistical framework that represents the expected increased inference accuracy or loss reduction based on prospective information [12]. The basic three types of VoI methods are as follows: (1) inferential and modeling cases for linear objective functions with simplified parameter distribution constraints, which limit their broad practical applicability; (2) inferential and modeling cases for nonlinear objective functions; and (3) inferential and modeling cases for [8] (2) methods for estimating the expected value of partial perfect information (EVPPI) that involve partitioning the parameter space into smaller subsets and assuming constant and optimal inference over local neighbourhoods within subsets [4];

For a specific parameter, the EVPPI is the expected inferential gain, or loss reduction, when is perfectly estimated. This is because the perfect is unknown in advance; this loss expectation reduction is applied to the entire parameter space:

EVPPI()=E(L(d,))+(E|(L(d,)))+(E|(L(d,))) … (1)

Where d represents a decision, inference, or action, d represents the optimal inference obtained when is known, is the model parameter vector, E represents the expectation, and L(d,) represents the likelihood function. It should be noted that VoI techniques are best suited for specific types of problems, such as evidence synthesis in decision theory.

Furthermore, their computational complexity is high, necessitating nested Monte Carlo procedures. Another relevant study divides the differences (errors) between theoretical (population) parameters and their sample-driven estimates (statistics) into three independent components in a novel way. If and denote a theoretical characteristic of interest (e.g., population mean) and its sample-based parameter estimate (e.g., sample arithmetic average), then the error can be canonically decomposed as follows:

(error) =A(Data Quality)+B(Data Quantity)+ C(Inference Problem Complexity...) .... (2)

Another metric that quantifies the intrinsic classification limits is the Bayes error rate. The Bayes error rate in classification problems represents the smallest classification error achieved by any classifier [16]. The Bayes error rate is determined solely by the class distributions and characterizes the minimum achievable error of any classifier.

2.2 Data Value Metric for feature selection by Noshad et al, 2021 [13]

The definition of DVM for supervised problems can be extended to unsupervised clustering models. We don't have explicit outcomes to evaluate model performance in unsupervised problems.

```
Input: Input dataset, X = {X1, ... XN } Labels, Y = {Y1, ..., YN }
Desired number of output features, r
F: =φ,R: = {1, ..., r}
for each I ∈ R do
 f ←j∈R−F (DVM{F} − DVM{F ∪Xj})
Add f into F
Output: F
```

2.3 Limitation of the Existing System

Noshad et al., (2021) proposed using Data Value Metric for feature selection in both supervised and unsupervised machine tasks. These methods used a sequential forward selection search strategy, which is prone to local minima and has a negative impact on the machine algorithm's performance when outputted features are used. On the one hand, there is a need to avoid the local minima that is inherent in the Data Value Metric algorithm for feature selection, while on the other hand, there is a need to address the problem of some redundant features inherent in Mutual information, as the features chosen may not be guaranteed to be non- redundant [1]. The Genetic algorithm is well-suited to dealing with the dual problem. This paper proposes

1. The use of Genetic Algorithm and Data Value Metric to address the problem of local minima, as Genetic Algorithm is used to avoid it.

2. The use of a Genetic Algorithm to remove feature redundancy, which has been shown to be common with mutual information in the Data Value Metric.

3. Conceptual Framework of the Proposed System

Figure 1 shows the proposed Genetic Algorithm based Data Value Metric feature selection.

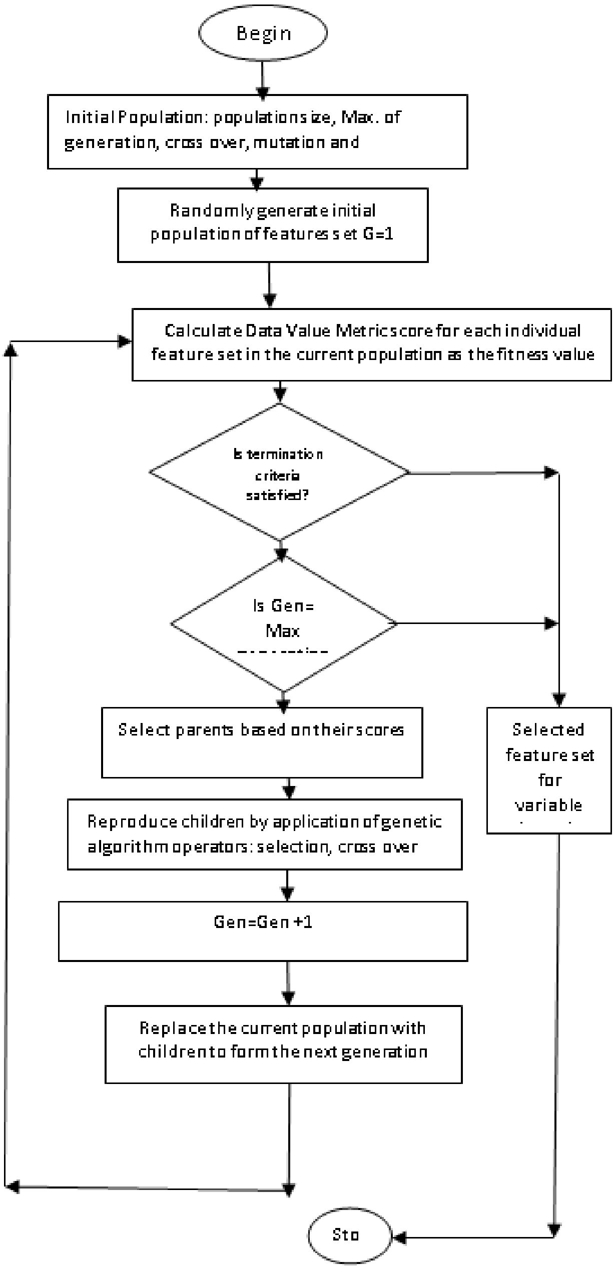
Begin
Initial Population: populationsize, Max. of generation, cross over, mutation and
Randomly generate initial population of features set G=1
Calculate Data Value Metric score for each individual feature set in the current population as the fitness value
Is termination criteria satisfied?
Is Gen= Max
Select parents based on their scores
Reproduce children by application of genetic algorithm operators: selection, cross over
Gen=Gen +1
Replace the current population with children to form the next generation
Selected feature set for variable
Sto

Figure 1: Proposed Genetic Algorithm based Data Value Metric feature selection

3.1 Data Value Metric (DVM)

Noshad et al., (2021), proposed a new information theoretical measure that quantifies the useful information content of large heterogeneous and traditional datasets. Data analytical value (utility) and model complexity are used by the DVM. It can be used to determine whether appending, expanding, or augmenting a dataset will benefit specific application domains. DVM quantifies the information boost or degradation associated with increasing the data size or the richness of its features, depending on the data analytic, inferential, or forecasting techniques used to interrogate the data. DVM is a combination of fidelity and regularization terms. The fidelity measures the utility of the sample data in the context of the inferential task.

The computational complexity of the corresponding inferential method is represented by the regularization term. Inspired by the concept of information bottleneck in deep learning, the fidelity term depends on the performance of the corresponding supervised or unsupervised model. DVM captures effectively the balance between analytical-value and algorithmic-complexity. Changes to the DVM highlight the tradeoffs between algorithmic complexity and data analytical value in terms of sample size and dataset feature richness. DVM values can be used to optimize the relative utility of various supervised or unsupervised algorithms by determining the size and characteristics of the data.

3.2 Genetic Algorithm (GA)

A GA is a heuristic search algorithm that is based on natural selection and genetics. To evolve a solution to a problem, the idea is to mimic biological processes such as survival of the fittest. GA is a method of evolving chromosome populations to new populations by combining selection with operations such as crossover and mutation [12]. Each chromosome contains genes. Selection operators select the fittest individuals from the population, whereas crossover and mutation mimic biological processes that introduce diversity into the population. Crossover and mutation are exploration processes, whereas selection is an exploitation process. Evolutionary algorithms are best suited for problems with a large search space, or a large number of possible solutions.

Other problems necessitate the creation of new solutions at each stage in order to investigate new options, or they involve complex solutions that cannot be processed by hand [12]. GAs, like the evolutionary process, relies on the fittest organisms/solutions to survive. The fitness of an organism/solution is determined by the problem at hand, and it is a factor that evolves over time.

3.3 Discussion of the work

This work presents a hybrid model of Data Value Metric and Genetic Algorithm for variable clustering in an unsupervised machine learning environment. The feature spaces of the data are represented as the initial population with values. The feature space is randomly selected to form the first generation which is sent into the evaluation metric where the Data Value Metric of that feature set is calculated. The Data Value Metric is used to quantify the amount of information present in the data. The Genetic Algorithm is applied to the features space of the data until a suitable high Data Value Metric is met.

Those feature sets that produced the highest Data Value Metric are then used in machine learning efforts in the hope of producing better classification performance. This is necessary because if only the data value Metric algorithm is used through the sequential forward selection search strategy, information or data may suffer from data redundancy, noise, and local minima.

4. Conclusion

We proposed variable clustering using the Genetic Algorithm and the Data Value Metric algorithm in this paper. Although there are efforts to collect data, not all data is of equal quality or value. Effective data preparation, which includes feature selection, is a panacea for a high-performing machine learning algorithm such as clustering. It is hoped that when the proposed system is implemented, it will result in a better clustering experience, a comparison of Genetic Algorithm based Data Value Metric (GA-DVM) variable clustering and other variable clustering methods to validate its computational effectiveness, and an investigation into the relationship between the proposed algorithm on performance evaluation metrics and data size to assay its performance in the big data ecosystem.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We acknowledge the support from the University of Delta, Agbor, Delta State in terms of moral support and other academic support in ensuring we have a conducive environment for research.

References

[1] Chandrashekar, G., & Sahin, F. (2014).A survey on feature selection methods.*Computers & Electrical Engineering,40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[2] Gómez F. and Quesada A.(2017). Genetic algorithms for feature selection in Data Analytics.[Online]. Available: https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection. [Accessed 05 12 2021]. https://doi.org/10.1049/cp.2010.0556

[3] Hunter, A. (2000). Feature selection using probabilistic neural networks. *Neural Computing & Applications,9*, 124–132. https://doi.org/10.1007/s005210070023

[4] Huang, H., Wu, Y., Chan, Y., & Lin, C. (2010). Study on image feature selection: A genetic algorithm approach," in *IET International Conference on Frontier Computing. Theory, Technologies and Applications*, Taichung.

[5] James, A.N, Hsein W, Herbert C, Mathew A., (2019) Machine Learning Applcations of Artificial Inteligence to Imaging and Diagnosis, Biophsy Rev, 11(1) 111-118

[6] Khuri, S. (2017).Genetic Algorithms. Helsinki, 2017

[7] Kushwaha, P., &Welekar, R. (2016). Feature Selection for Image Retrieval based on Genetic Algorithm. *International Journal of Interactive Multimedia and Artificial Intelligence,4*(16), 16–21. https://doi.org/10.9781/ijimai.2016.423

[8] Leskovec, J., Rajaraman, A., & David, J. (2014).Dimensionality Reduction.In Mining of Massive Datasets (pp. 415–447).Cambridge University Press.https://doi.org/10.1017/CBO9781139924801.013

[9] Liang, L., Peng, J., & Yang, B. (2012). Image Feature Selection Based on Genetic Algorithm. In *International Conference on Information Engineering and Applications*, Chongqing

[10] Lin, C. H., Chen, H. Y., & Wu, Y. S. (2014). Study of image retrieval and classification based on adaptive features using genetic algorithm feature selection," *Expert Systems with Applications,* vol. 41, no. 15, pp. 6611-6621.

[11] Lu J., Zhao T. and Zhang Y., (2008). Feature selection based-on genetic algorithm for image annotation. *Knowledge-Based Systems,21*(8), 887–891.

[12] Mitchell, M. (1998).An Introduction to Genetic Algorithms.MIT Press.https://doi.org/10.7551/mitpress/3927.001.0001

[13] Noshad M, Jerome C, Yuming S, Alfred H &I vo D. Dino(2021).A data value metric for quantifying information content and utility.*Journal of Big Data .Vol.8:82* https://doi.org/10.1186/s40537-021-00446-6

[14] Sahiner, B., Chan, H. P., Wei, D., Petrick, N., Helvie, M. A., Adler, D. D., &Goodsitt, M. M. (1996). Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue. *Medical Physics,23*(10), 1671–1684. https://doi.org/10.1118/1.597829

[15] Sivasankar, E., & Rajesh, R. S. (2012). Design and development of efficient feature Selection and classification techniques for Clinical decision support system. Shodhganga, Tirunalveli.

16] Woznica, A., Nguyen, P., & Kalousis, A. (2012). Model mining for robust feature selection., KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM New York, NY, USA, PP 913-921, 2012.

[17] Yu, L., & Liu, H. (2004).Efficient Feature Selection via Analysis of Relevance and Redundancy.*Journal of Machine Learning Research,5*, 1205–1224.

[18] Yusof, R., Khairuddin, U., & Khalid, M. (2012).A New Mutation Operation for Faster Convergence in Genetic Algorithm Feature Selection.*IJICIC,8*, 7363–7378.