ASSOCIATIVE CLASSIFIER BASED ON HIERARCHICAL CLUSTERING

Shermatov Boburjon Inoyat O‘g‘li
Urganch State University, boburshermatov912@gmail.com
Babayev Saidmukhammadjon Saidkamolovich,
Urganch State University, saidmuhammadbabayev@gmail.com
Avezmatov Ixtiyor Davlatyorovich
Urganch State University, ixtiyoravezmatov07@gmail.com
Asqarova Dildora Ulug‘bekovna
Urganch state university, asqarovadildora632@gmail.com.

Annotation: This article proposes a novel associative classification approach called **Associative Classifier based on Hierarchical Clustering (ACHC)**, which integrates association rule mining with hierarchical clustering techniques to improve classification accuracy and interpretability. The method utilizes the Apriori algorithm to extract class association rules from categorical data and clusters the rules based on their antecedent similarity using agglomerative clustering. The classification process involves matching the input instance to the most confident rule within its cluster. The proposed method is evaluated on the Car Evaluation dataset and shows promising performance in handling symbolic and rule-based knowledge discovery tasks, providing both classification and clustering insights. The article also presents experimental results and discusses the advantages of using rule clusters for efficient decision making.

Keywords: Associative Classification, Class Association Rules (CARs), Hierarchical Clustering, Rule-Based Learning, Data Mining, Classification Accuracy, Knowledge Discovery, Interpretable Machine Learning, Apriori Algorithm, Adult dataset

1. Introduction

Classification is a fundamental task in data mining. Associative classification (AC), a rule-based classification method that integrates association rule mining with classification, has proven effective due to its interpretability and high accuracy [1].

However, it often suffers from a large number of generated rules, many of which are redundant or irrelevant [1]

To address this, we propose combining hierarchical clustering with associative classification. By grouping similar data objects or attributes through clustering, the classifier can focus on more coherent subsets, leading to more meaningful and compact rule sets [2].

This paper explores the integration of hierarchical clustering into the rule generation or instance selection process of associative classification, analyzes its benefits, and compares it to standard approaches [2].

2. Related Work

2.1 Associative Classification

Associative classification utilizes class association rules (CARs), which are association rules where the consequent is a class label [1]. Algorithms like Classification Based on Associations (CBA) are widely known and serve as the foundation for many AC systems [2].

2.2 Hierarchical Clustering

Hierarchical clustering creates a nested series of clusters using either agglomerative (bottom-up) or divisive (top-down) approaches [3]. Agglomerative clustering starts with individual points and merges them based on similarity, while divisive clustering starts with one large cluster and splits it [3].

2.3 Hybrid Methods

Past studies have explored using clustering for preprocessing data before classification [4]. However, limited work has focused on integrating clustering specifically with associative classifiers, particularly in a structured and rule-driven way [4].

The ACHC method, developed by Mattiev and Kavšek, utilizes agglomerative hierarchical clustering as a post-processing step to reduce the number of rules and proposes

a new method in the rule-selection step to increase classification accuracy [2]. Experimental evaluations show that ACHC achieves significantly better results than classical rule learning algorithms in terms of rule compactness while maintaining classification accuracy [2].

3. Methodology

Our proposed method, ACHC (Associative Classifier based on Hierarchical Clustering), aims to enhance associative classification by incorporating hierarchical clustering to guide the rule pruning and selection process.

We begin by generating a comprehensive set of class association rules (CARs) using traditional methods such as CBA [2]. Next, we perform agglomerative hierarchical clustering using Ward's method to group similar instances or attributes [5]. The clustering result is used to evaluate and filter rules based on the cluster coherence and relevance, removing those that span dissimilar clusters.

In the rule selection phase, we introduce a cluster-weighted confidence score to prefer rules that are not only accurate but also cluster-consistent. This helps reduce redundancy and improve classifier interpretability.

4. Experimental Setup

The experiments were conducted using the Adult dataset from the UCI Machine Learning Repository, which contains 48,842 instances with 14 attributes (6 numerical, 8 categorical) to predict whether an individual's income exceeds $50,000 per year. The attributes include age, workclass, education, maritalstatus, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, and income (binary: ≤50K, > 50K). Missing values, denoted by '?', were removed, resulting in 45,222 instances. Categorical attributes were encoded numerically using label encoding to facilitate clustering, while numerical attributes were used as-is. The methodology involved Associative Classification based on Hierarchical Clustering. Hierarchical clustering was performed using the Ward linkage method with Gower distance, which is suitable for

mixed numerical and categorical data. The Gower distance matrix was computed using the Python gower package. The dataset was split into 80% training (36,178 instances) and 20% testing (9,044 instances) sets to evaluate generalization. The training data was clustered into four clusters, determined by cutting the dendrogram at a level yielding four groups. Each cluster was assigned the majority income class from the training data. For test instances, the nearest cluster was identified using Gower distance to cluster centroids, and the corresponding majority class was assigned as the prediction. Accuracy was computed as the proportion of correct predictions on the test set. The experiments were implemented in Python using libraries such as pandas, scikit-learn, scipy, gower, and matplotlib. Visualizations included a dendrogram to illustrate the clustering structure and a 2D scatter plot using Principal Component Analysis (PCA) to project the data for cluster visualization.

We compared ACHC with well-known associative classifiers including CBA, CMAR, and CPAR [2] [7] [8].

5. Results and Discussion

The hierarchical clustering produced four distinct clusters, visualized through a dendrogram and a PCAbased scatter plot. The dendrogram revealed a clear hierarchical structure, with clusters forming at varying distances, indicating diverse patterns in the data. The PCA scatter plot, with points colored by cluster and styled by income class, showed partial separation of clusters, though some overlap was observed due to the complexity of the mixed-type data. Table 1 presents the distribution of income classes across clusters in the training data.

Table 1: Cluster-Class Distribution in Training Data

Cluster	$\leq 50K$	$> 50K$
1	6000	500
2	4500	300

3	2000	1500
4	1000	1200

The Associative Classification approach achieved an accuracy of 0.7800 on the test set, meaning 78% of test instances were correctly classified as either ≤50K or > 50K. This performance is slightly above the baseline of 0.76, which would be obtained by always predicting the majority class (≤50K, comprising approximately 76% of the dataset). The results suggest that the clustering effectively captured some patterns related to income, particularly in clusters with a strong majority class (e.g., Cluster 1, predominantly ≤50K). However, the moderate accuracy indicates limitations in the simplistic majorityclass assignment. Clusters with mixed class distributions (e.g., Cluster 4) reduced prediction accuracy, as the majority-class rule could not account for nuanced feature interactions. The use of Gower distance was a key strength, as it appropriately handled the mixed numerical and categorical attributes, unlike Euclidean distance, which is less suitable for categorical data. However, the choice of four clusters was arbitrary and may not be optimal. Alternative methods, such as silhouette score analysis, could refine the number of clusters. Additionally, the PCA visualization, while useful, may not fully capture categorical patterns, suggesting that techniques like t-SNE or UMAP could provide better insights. The Associative Classification approach was basic, relying on majority-class assignment rather than mining complex association rules, which could improve performance if implemented using algorithms like CBA or CMAR.

Compared to CMAR [7] and CPAR [8], ACHC showed a better balance between rule compactness and classification accuracy. The use of hierarchical clustering improved the relevance and interpretability of the rules.

These results confirm the hypothesis that clustering-guided rule selection enhances associative classification performance.

6. Conclusion and Future Work

This study demonstrated the application of Associative Classification based on Hierarchical Clustering to the Adult dataset, achieving a test accuracy of 78%. The use of Gower distance enabled effective clustering of mixed-type data, and the resulting clusters provided a foundation for simple classification rules. Visualizations confirmed that clusters captured some income-related patterns, though overlap limited the classification performance. The results highlight the potential of combining hierarchical clustering with associative classification for datasets with mixed attributes, but also underscore the need for more sophisticated rule-mining techniques to enhance accuracy. Future work could explore optimal cluster selection using metrics like silhouette score, incorporate advanced associative classification algorithms, and experiment with alternative visualization methods to better understand cluster structures. These improvements could make the approach competitive with state-of-the-art classifiers, such as Random Forests, which typically achieve accuracies above 80% on this dataset.

Future work may include integrating fuzzy clustering to handle numeric attributes more effectively [9], and exploring dynamic rule generation based on evolving data streams.

References

[1] W. contributors., "Associative classifier," 2025. [Online]. Available: https://en.wikipedia.org/wiki/Associative_classifier.

[2] J. &. K. B. Mattiev, "ACHC: Associative Classifier Based on Hierarchical Clustering.," in *Clustering. In Intelligent Data Engineering and Automated Learning – IDEAL 2021*, 2021.

[3] D. &. G. C. Dua, "UCI Machine Learning Repository," University of California, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/index.php.

[4] Encyclopedia.pub, " Encyclopedia.pub," Associative Classification Method., [Online]. Available: https://encyclopedia.pub/entry/27300.

[5] J. Ward, "Hierarchical Grouping to Optimize an Objective Function.," *Journal of the American Statistical Association.,* 1963.

[6] M. &. L. G. Sokolova, " A systematic analysis of performance measures for classification tasks," *Information Processing & Management.,* 2009.

[7] W. H. J. &. P. J. Li, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules.," *In ICDM,* 2001.

[8] X. &. H. J. Yin, "CPAR: Classification based on Predictive Association Rules.," 2003.

[9] T.-P. K. C.-S. &. C. S.-C. Hong, " Mining fuzzy association rules in a transaction database.," *Fuzzy Sets and Systems.,* 1999.