**Transactions on Engineering and Computing Sciences - Vol. 13, No. 05**
**Publication Date: September 10, 2025**
**DOI**:10.14738/tmlai.1305.19355.
Kaur, S., Mohil, M., Sharma, A., Bhaniya, H., Singh, H., & Bhardwaj, M. (2025). Amalgamation of Unsupervised and Supervised Approaches for Data Labeling. *Transactions on Engineering and Computing Sciences*, 13(05). 54-67.

# Amalgamation of Unsupervised and Supervised Approaches for Data Labeling

**Sharanjit Kaur**
Department of Computer Science,
Acharya Narendra Dev College,
University of Delhi, Delhi, India

**Meenu Mohil**
Department of Physics,
Acharya Narendra Dev College,
University of Delhi, Delhi, India

**Ansh Sharma**
Department of Computer Science,
Acharya Narendra Dev College,
University of Delhi, Delhi, India

**Hardik Bhaniya**
Department of Computer Science,
Acharya Narendra Dev College,
University of Delhi, Delhi, India

**Harshita Singh**
Department of Computer Science,
Acharya Narendra Dev College,
University of Delhi, Delhi, India

**Manju Bhardwaj**
Department of Computer Science,
Maitreyi College, University of Delhi,
Delhi, India

## ABSTRACT

**Data abundance is inevitable when every human activity is revolving around Internet of Things (IoT). The data is extensive, but it lacks the labels needed by the machine learning models to identify patterns and characteristics for accurate prediction and automation. Data labeling is a very crucial and essential task for consuming this abundant data for applications like customer relationship management systems, recommendation systems and pattern recognition. We propose a novel approach called Amalgamation of Unsupervised and Supervised Approaches for Data Labeling (AUSL), which integrates clustering and classification using rule-based refinement. Given the unlabeled data, AUSL offers a robust and interpretable framework for uncovering meaningful data labels. Ensemble-based**

**clustering and AdaBoost SVM ameliorates the selection of important attributes for data labeling, which are further processed by association rule mining to extract underlying significant data characteristics from the reduced domain. Experiments are conducted on four data sets to prove the robustness of the proposed method. The comparative performance of AUSL with an existing method is promising, achieving finer labels with an average hit rate exceeding 90% and confidence levels above 80%. These results indicate the robustness, adaptability, and superior label refinement ability of the proposed method. In conclusion, AUSL provides a scalable, interpretable, and effective solution for structured data labelling, with strong potential for real-world deployment in various data-driven applications.**

**Keywords:** Ensemble clustering, Unlabeled data, Association rules, AdaBoost ensemble, Supervised, Unsupervised, SVM, Data labels.

## INTRODUCTION

Applications such as web searches, network communication, IoT devices, network security, e-commerce and recommender systems generate a large amount of unlabeled data. The large volume of generated data can be effectively utilized for the automation of the desired applications like image analysis, anomaly detection, real-time recommendations and customer assistance with the presence of data labels [1,2]. Supervised machine learning models used for these applications need data labels to learn underlying context. Also, these labels are not only useful to understand the model's decision but also assist the individual in enhanced data interpretation to further improve the quality of the model [1]. Earlier, domain experts were engaged in labeling the data manually by studying and understanding the context [2,3]. However, this time consuming and expensive manual procedure slows down the process of data labeling [4-7]. Recently, clustering, an unsupervised method has gained popularity for accelerating the laborious task of data labeling [8,9].

Clustering is used to generate groups of related data for identification of its dominant characteristics and attributes [10]. It is employed for capturing customer segmentation in e-commerce transactions [11], for topic modeling in web search data [12], topic and entity modeling in information retrieval systems [13,14]. Various researchers are working out strategies to improvise the clustering algorithms, and quite a few of them are focusing on the usability of identified clusters in terms of its characterization and data labeling [9].

Indeed, the usage of clustering simplifies the annotation work for domain experts but on a reduced data space. The manual review of each of the generated clusters is still a tedious task, and affects its smooth applicability in systems such as customer relationship management, decision support, recommender systems, pattern recognition and information retrieval systems [15].

Clustering reveals groups of instances which are mutually dissimilar in characteristics, with each group consisting of dominant attributes that bind together the underlying instances. Machine learning models are apt for solving these issues, but need to be tuned for automatic revelation of these attributes [16]. Supervised machine learning techniques like neural network

and support vector machines are quite effective for revealing important data attributes [9]. Discovery of cluster-wise characteristics is crucial for reporting labels. The same set of attributes may be important for different clusters but require additional analysis to capture their contrasting characteristics. Association rule mining, an unsupervised approach is suitable for revealing such data characteristics and is being used to identify similar interest patterns among user's web search, purchase patterns, e-commerce transactions etc. [18,19]. In the proposed research work, this process is called data labeling, which aims to reveal finer characteristics of each cluster aka group using association rule mining.

Note that the generated data labels depend on the identified clusters, which vary with the clustering algorithm employed [9]. Recently, it has been shown that ensemble clustering outperforms any clustering algorithm, as it makes use of consensus among the clustering schemes to understand the inherent similarities among data points in a better way [20,21]. This motivated us to adopt ensemble clustering for generating a robust clustering scheme. Similarly, AdaBoost ensemble is used to improve learning capability of the classifier for attribute identification [22]. Currently, the cluster-specific range of identified attributes are reported as labels [9,23]. However, there is a further need to refine the data characteristics to capture the pattern and dependency in the data. Motivated by this, the research question we sought to answer in this paper is: "How to leverage the ensemble based supervised and unsupervised machine learning approaches to refine data labels?". In this paper, we propose a two-phase approach called AUSL (Amalgamation of Unsupervised and Supervised Learning) for data labeling, which leverages the coherence of ensemble clustering, classification ensemble, and association rule mining. The contributions are as follows:

- Use of unsupervised ensemble clustering to obtain good quality clusters.
- Learn the important cluster-wise attributes from the generated clusters using supervised AdaBoost ensemble with reduced uncertainties.
- Unsupervised association rule mining to identify prominent and fine rule-based data labels.
- Novel validation of the generated data labels by comparing them with those computed using corresponding ground truth.
- Comparative performance with existing work [9] using a robust performance metric, Jaccard similarity.

*Organization of the paper*: Section 2 briefly review the existing related literature. Section 3 explains the methodology followed in the proposed approach AUSL for data labeling followed by details of evaluation method used for comparative analysis and experimental setting. Sections 4 and 5 covers the results and discussion respectively. The paper is concluded in Section 6.

## LITERATURE REVIEW

In this section, we focus on the research works which employ unsupervised approach for data labelling, which is directly related to our proposed research solution. We deliberately exclude the compute intensive semi-supervised deep learning approaches from the review. In this

section, we only focus on techniques/methods used for labeling of structured and unstructured data.

Data arranged in a particular structure is called as structured data and is commonly used in various commerce, medical and social science applications. Clustering is directly applied on the structured data to reveal similar groups, the groupings are further used for generating data labels. Lopes et al. have shown the impact of discretization method on data labels identified using clustering and classification [9]. The supervised artificial neural network is trained to learn the important label for each cluster identified by a clustering algorithm. As the labels reported are influenced by the number of clusters generated, Silva et al. have refined the said method by using an inference approach to compute optimal number of clusters and demonstrated its applicability on four datasets [23]. Esmaeili et al. have extended the clustering to ensure group fairness while labeling the data for target marketing [24]. Recently, clustering has also been employed to generate synthetic class labels using an autoencoder to address the challenge for highly imbalanced credit card fraud detection datasets, resulting in improved model performance [25].

For unstructured data such as text, images, and audio, clustering algorithms cannot be used directly. Data needs to be converted to the structured form before revealing segments by a clustering algorithm. Here, we briefly review the literature which blends unsupervised and supervised approaches for extracting data labels from the unstructured data. Beil and Theissler [8] have adopted an interactive cluster clean-label approach for effective labeling of the unlabeled MNIST data set and have utilized the strengths of humans and unsupervised machine learning methods to achieve the net data labels. Recently, a comprehensive review on document clustering by Cozzolino and Ferrao [26] highlights the importance of clustering for organizing huge amount of unstructured text data without engaging a domain-expert explicitly. Kaya M.F. has used association rule discovery, topic modeling and decision trees for the semi-automated pattern labeling of business communication data [27]. It is shown that meaningful descriptions may be revealed from the generated labels for enhanced interpretation of the results compared to patterns delivered in high dimensional space. However, reliability is still a challenge due to mapping of original data space to reduced space using principal component analysis. Peganova et. al. [14] used the hierarchical clustering for revealing the labels for scientific documents using the text content of abstracts, which they found useful for searching an article under a particular topic. In computer vision, clustering-based methods help generate pseudo-labels for object recognition and categorization, which are still being used because of their ability to capture intrinsic data similarities when prior labels are unavailable [28-31]. Zhang et al. have used clustering consensus to refine pseudo labels by eliminating noise and have reported improvement over existing state-of-the-art clustering-based unsupervised methods for image re-identification [28].

*All of the proposed approaches have their benefits and limitations and are quite effective in generating labels, but no attention is being paid to the validation and refinement using the available ground-truths. In this research article, we propose a hybrid approach AUSL to generate*

*fine labels by amalgamation of unsupervised (clustering and association rules) and supervised approaches for structured data, which are validated using existing class labels.*

## METHODS

The proposed methodology aims to automate the generation of cluster-specific labels for unlabeled data by collaborating the outcomes of ensemble clustering, ensemble classifier and association rule mining. In this section, we detail the proposed approach AUSL that follows a structured pipeline visually described in Fig. 1. The proposed method consists of two phases viz. Phase I: Ensemble Clustering and Discretization, and Phase II: Identifying Dominant Attributes using AdaBoost for reporting Rule-based Data Labels. Each of these two phases are detailed in the following subsections.

### Phase I: Ensemble Clustering and Discretization

This phase involves clustering of data and its discretization, followed by cluster-wise segregation of discretized data to be used in Phase II. All the steps followed in this phase are explained below.

1. *Data Preprocessing*: Correlation is used to identify the redundant attributes, which are removed before data clustering. Also, the remaining attributes are normalized to avoid bias of the machine learning algorithm towards domain attributes with larger values. It may be noted that this step is not explicitly mentioned in Fig. 1, but every data set is preprocessed before the clustering step. After preprocessing the data set, let $d$ be the number of remaining attributes identified as $a_1, a_2 .. a_d$ each with the corresponding domain range as $R_j$=[$l_j$,$u_j$], where $l_j$ and $u_j$ denote the lower and upper bounds for the attribute $a_j$ respectively.
2. *Ensemble clustering*: Three distinct clustering algorithms viz. k-means, spectral clustering, and agglomerative clustering are applied to the unlabeled data to fetch three clustering schemes. Afterwards, a co-association matrix is built using three clustering schemes to obtain a comprehensive and well-rounded representation of the entire data [26] required for ensemble clustering. This data matrix is used by k-means clustering algorithm to generate the desired number of $k$ clusters $\boldsymbol{C} = \{C_1, C_2, .., C_k\}$. The final cluster labels serve as reliable groupings of data points, which are later utilized for generating data labels.
3. *Discretization*: Data discretization is recommended for building accurate classifiers and is a pre-requisite for association rule mining used in Phase II for revealing informative labels. In order to improve data interpretability, the numerical continuous values are mapped to categorical values (distinct for each attribute) for simplifying the mapping process in phase II. Following the recommendation of [6,9], the *equal frequency discretization* method is used to map the domain range of numeric attributes to discrete categories. Let $\Delta(a_j)$ be the categorized values of the domain range $R_j$.
4. *Fabrication*: During this step, discretized data is mapped to its respective cluster labels that are generated by ensemble clustering. Cluster-wise discretized data is used by the supervised learner in Phase II for identification of cluster-wise dominant attributes, which are referred as *principal* attributes.

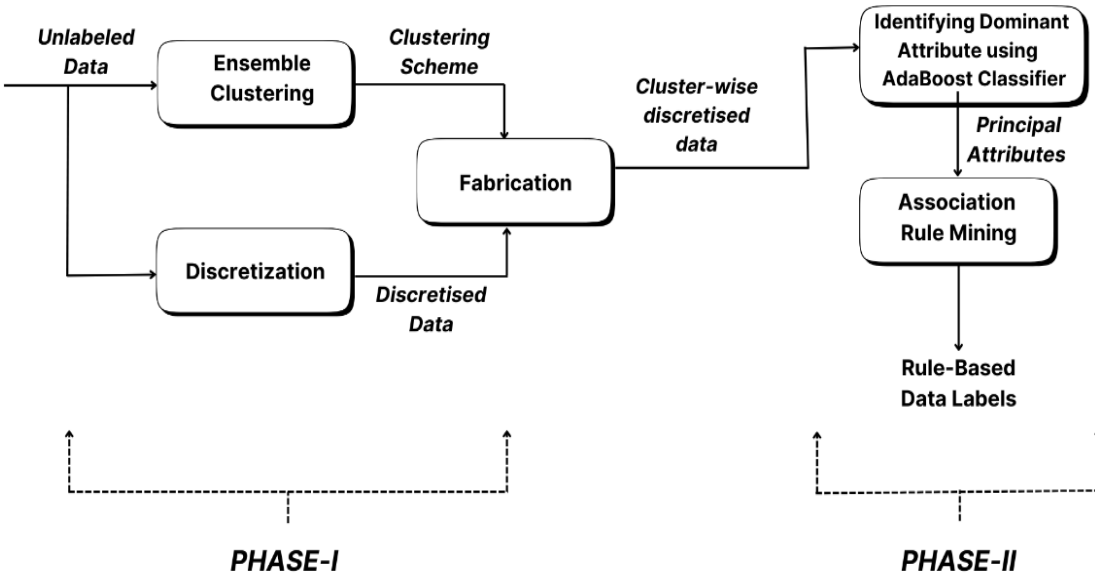


**Figure 1: Two-Phase Framework followed in AUSL for Cluster-wise Rule-Based Data Label Generation.**

## Phase II: Identifying Dominant Attributes using AdaBoost Classifier for Rule-based Data Labels

This phase uses discretized data and their corresponding cluster labels to identify principal attributes using AdaBoost ensemble. Subsequently, discrete categories of cluster-specific dominant (principal) attributes are processed using association rule mining to reveal important and informative characteristics as labels. The following steps are applied in phase II for generating rule-based data labels (see Fig. 1).

1. *Identifying Dominant Attributes*: In order to identify important attributes based on their predictive significance, we use Adaptive Boosting (AdaBoost) with Support Vector Machines (SVM) as base classifiers. Discretized records of each cluster are processed separately for identifying influential attributes from the original $d$ attributes. Each attribute is treated as a target variable, and the remaining attributes serve as predictors. A k-fold stratified cross-validation technique is used to record mean classification accuracy, referred as hit rate (H), for the selected target attribute. The process is repeated $d$ times and hit rate is computed for each attribute. Higher the hit rate of an attribute, higher is the predictive strength of attribute in classifying instances within the cluster. A dynamic statistical method is used to select important attributes with strong predictive performance from the $d$ attributes on the basis of computed hit rates. Note that the selected attributes termed as *principal attributes* may vary across the clusters and are crucial for data labeling. Let $L_i = \{q_1, q_2, .., q_{p_i}\}$ be the identified $p_i$ principal attributes for each cluster $C_i$.
2. *Association Rule Mining*: It constitutes a critical component of the proposed methodology for extracting informative and explanatory data labels from the reduced domain consisting of the principal attributes. For each identified cluster $C_i$ , association rules are generated using the famous Apriori rule mining algorithm that identifies frequent item

sets iteratively while pruning the search space using the anti-monotone property [33]. For data labels, two-item rules $X \rightarrow Y$ are generated such that $X \in \Delta(L_i)$. The consequent part ($Y$) of the rules may contain discrete value of any other attribute which may not be part of $\Delta(L_i)$. Rules, signifying dependency between categories, are captured at the user-specific support, confidence and lift thresholds. Let $G_i$ be the set of such identified categories of size $s_i = |G_i|$ such that $s_i \leq p_i < d$. These identified categories are then mapped back to original data space to generate cluster specific characteristics represented by data label $t_i$ for each cluster $C_i$ as shown below.

$$t_i = \cup_{j \in s_i} \; ]\alpha_j, \beta_j] \subset R_j \quad (1)$$

Here $\alpha_j$ and $\beta_j$ denote the lower and upper limit of the identified categories with $R_j$ as the range of $j^{th}$ attribute in original data set. It is important to note here that user-specified thresholds may need to be lowered for some clusters in case no data labels are reported because of underlying data sparsity and dependency.

Complete methodology followed in the proposed approach is concisely described in Algorithm 1.

## Evaluation Methods

This section lists the data sets used for the evaluation of the proposed method AUSL, followed by the description of validation metrics employed for performance evaluation.

### *Data Sets and Experimental Settings:*

Table 1 lists the four data sets used in this study. The data sets are downloaded from UCI repository [34], and their details are presented in Table 1. To assess the effectiveness of AUSL, its performance on these data sets is compared with that for an approach proposed in [9].

**Algorithm 1: AUSL Method**

**Input:** Data Set $D$ with #Instances $N$, #Clustering schemes $M$, #Clusters $k$, Hit rate $H$, User-specified support $S$, Confidence $F$ and Lift $T$

**Output:** Data labels for $k$ clusters

**Phase I: Ensemble Clustering and Discretization**

a) Pre-process the data set D to remove redundant (correlated) attributes and normalize the data. Let $d$ be the number of remaining attributes $a_1, a_2 \ldots a_d$ with corresponding domain range $R_j = [l_j, u_j]$ where $l_j$ and $u_j$ represent the lower and upper limit of the attribute $a_j$, $j \in [1, d]$.

b) Generate $M$ clustering schemes and create a co-association matrix showing frequency of co-occurrence of a pair of points in $M$ schemes. Apply k-means algorithm on the matrix to get an ensemble clustering scheme $\boldsymbol{C} = \{C_1, C_2, \ldots, C_k\}$.

c) Apply *equal frequency discretization* method to get discrete categories $\Delta(a_j)$ of each attribute $a_j$.

d) Map each attribute value in a cluster to corresponding discrete category (identified in step (c)).

**Phase II: Identifying Dominant Attributes using AdaBoost Classifier for Rule-based Data Labels**

For each cluster $C_i \in C$ do

i. Identify important attributes using AdaBoost with SVM as base classifier with hit rate >= $H$ and refer them as $L_i = \{q_1, q_2, .., q_{p_i}\}$ with $p_i$ as number of principal attributes selected for $C_i$.
ii. Apply Apriori algorithm to generate two-item sets rules $X \rightarrow Y$ $s.t$ $X \in \Delta(L_i)$.
iii. Filter out rules with user-specified support (S), confidence (F) and lift (T).
iv. Generate data labels using $X$, $Y$ of the filtered rules by mapping them to corresponding original attribute values.

**Table 1: Details of Data sets Used in the study.**

| S.No | Data set | # Attributes | # Records | #Clusters | URL |
|---|---|---|---|---|---|
| 1 | Wheat Seeds | 8 | 210 | 3 | https://archive.ics.uci.edu/dataset/236/seeds |
| 2 | Iris | 5 | 150 | 3 | https://archive.ics.uci.edu/dataset/53/iris |
| 3 | Wine | 14 | 178 | 3 | https://archive.ics.uci.edu/dataset/109/wine |
| 4 | Algerian Forest Fire | 14 | 244 | 2 | https://archive.ics.uci.edu/dataset/547/algerian+forest+fires+dataset |

All experiments in this study were carried out on a computer system with an Intel Core i3 processor (13th generation) 1.60 GHz, 8 GB RAM, and a 512 GB SSD. The development environment was meticulously configured with Python 3.12.8, complemented by the essential libraries—NumPy, Pandas, Scikit-learn, and Matplotlib—that facilitated robust data analysis and visualization.

The attributes were normalized (z-score normalization) using StandardScaler() function from scikit-learn library. For all data sets, features were discretized using KBinsDiscretizer() function for equal-frequency binning, where the values were divided into quantile-based bins resulting in different number of bins for different data sets. Cluster ensemble was generated using three distinct clustering algorithms viz. While clustering, k clusters are generated where k is set as the number of classes as ground truth with the data set. In case of unlabeled data, k may be computed using method given in [21]. For the classification component, an AdaBoost ensemble was generated using Support Vector Machine (SVM) with linear kernel as the base classifier, and the number of boosting rounds was fixed at 10. For each cluster, attributes exhibiting hit rates surpassing the median for that cluster were selected as the principal attributes unique to it. This range was carefully chosen to ensure that only the most informative attributes were retained for refined data labels, unless explicitly specified otherwise in the experimental context. The proposed approach AUSL has three main components viz. ensemble clustering, ensemble classifier and statistical method for attribute selection. To assess the contribution of each individual component to the overall quality of data labeling, we also designed an experimental study where only one component is modified at a time. Keeping rest of the components fixed, we vary Ensemble Clustering in AUSL-1, Adaboost Ensemble with

Linear SVM as base classifier in AUSL-2 and median-based attribute selection method in AUSL-3, as followed in the compared method ALCD [9].

**Validation Metrics:**

To ensure robustness and significance of the extracted rules, we employ established interestingness measures [35] for filtering rules, which are briefly explained below.

- *Support* of a rule $X \rightarrow Y$ is defined as:

$$S(X \rightarrow Y) = \frac{\left| t_g|_{X,Y \subseteq t_g,\ t_g \in C_i} \right|}{|C_i|}$$

  with $t_g$ as $g^{th}$ instance in the cluster $C_i$ with its total records $|C_i|$. Considering Y as null, S(X) can be computed using above formula. We use minimum support threshold as 0.5 for all experiments.

- *Confidence* of a rule $X \rightarrow Y$ quantifies the reliability of the inference from $X$ to $Y$ and is defined as:

$$F(X \rightarrow Y) = \frac{S(X \rightarrow Y)}{S(X)}$$

  We have used 0.8 as minimum confidence threshold in the experiments.

- *Lift* metric of a rule $X \rightarrow Y$ captures the positive correlation between antecedent ($X$) and consequent ($Y$) of the rule, and is calculated as:

$$T(X \rightarrow Y) = \frac{F(X \rightarrow Y)}{S(Y)}$$

All rules with T ≥ 1 are filtered out for data labeling. As the data labeling is constructed using identified cluster-specific rules (see (1)), we use Jaccard similarity score for their quality assessment [36]. The Jaccard similarity score is defined as $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, which quantifies the similarity between two sets $A$ and $B$ by measuring the size of their commonality relative to all. We compute $J(A,B)$ by comparing the cluster specific rules with the corresponding class specific rules to validate the robustness of the proposed method. Note that corresponding class of a cluster is identified using dominant class labels within it. Let $J_i$ be the Jaccard similarity score between cluster $C_i$ and the corresponding dominant class $Cl_i$. The average Jaccard similarity score for k clusters is computed as $\sum_{i=1}^{k} \frac{J_i}{k}$.

In all experiments, number of clusters (k) is set as the of number of classes in the data set to have a fair matching of generated labels. We have used S ≥ 0.5, F ≥ 0.8 and T ≥ 1 in all the experiments unless mentioned explicitly.

## RESULTS

This section describes the results reported by the proposed method AUSL on the four UCI data sets listed in Table 1. Delineated data labels are generated using high confidence association rules. We validate the effectiveness of the delivered data labels using Jaccard similarity score by assessing the alignment between cluster-derived and class-specific knowledge. We report

and discuss the results of different experiments conducted to affirm the performance of the proposed method.

In the initial experiment, we examine the effect of confidence threshold F on the stability and consistency of cluster-specific labels using two data sets. We compute the Jaccard similarity score for confidence thresholds of 60%, 70%, 80%, and 90%. As illustrated in Figure 2, the cluster-wise comparison reveals that the Jaccard similarity remains stable up to an 80% threshold across both data sets. Beyond this point, increasing the threshold impacts the similarity score.

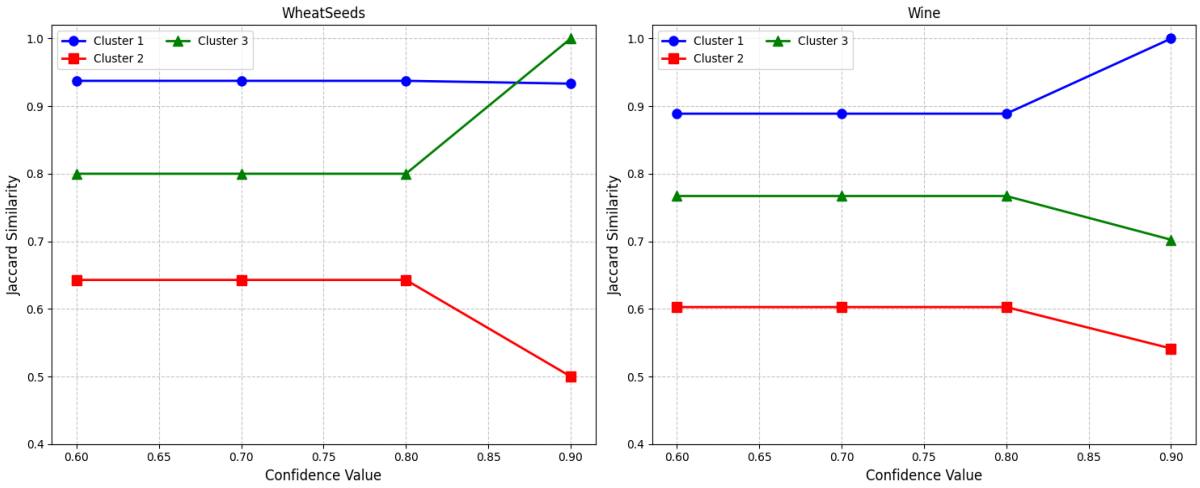


**Figure 2: Impact of varying confidence value on the Jaccard score used for capturing similarity between cluster-specific rules and class-specific rules.**

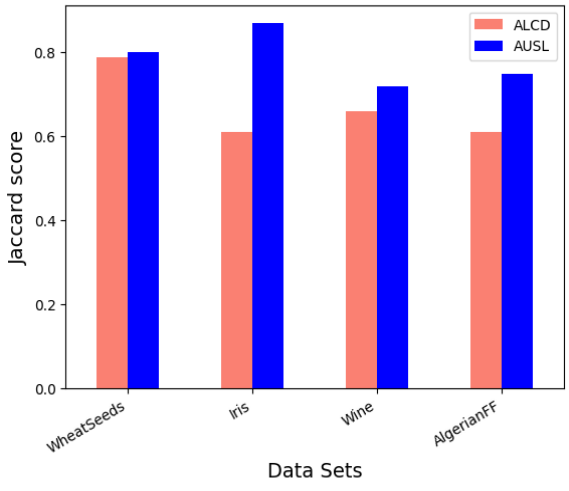


**Figure 3: Comparative performance of AUSL and ALCD on four data sets: AlgerianFF- Algerian Forest Fire.**

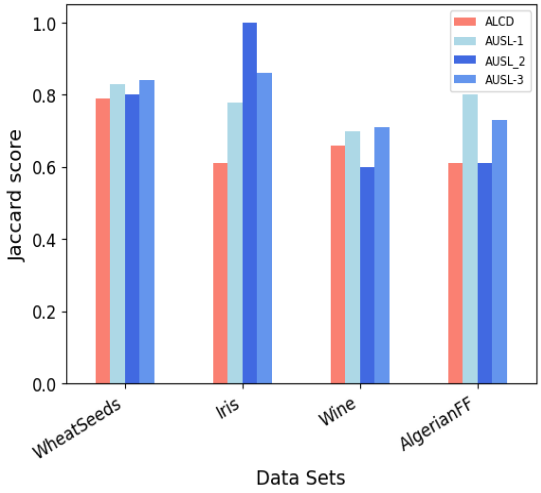


**Figure 4: Component-wise performance of AUSL and ALCD on four data sets.**

Consequently, we proceed with F≥0.8 in subsequent analyses. Next, we validate the enhanced data labeling performance of the proposed method AUSL by comparing its results across four data sets with those obtained using the approach introduced in [9], which we refer to as ALCD in the discussion. Figure 3 shows the average Jaccard similarity scores computed for the four datasets. The consistently higher similarity scores achieved by AUSL across all data sets is indicative of its superior ability to capture finer data labels, demonstrating its marked improvement over the existing ALCD approach.

We also check the effectiveness of individual component of the proposed method which we refered as AUSL-1, AUSL-2 and AUSL-3 (See Sec. 3.3.1 for details). As shown in Figure 4, the results confirm the positive impact of these modifications, demonstrating their positive influence on the approach. Notably, performance is comparable or improved across all cases, with the exception of AUSL-2 on the Wine data set as well as AlgerianFF. The deteriorated result in this case may be attributed to overfitting of the Adaboost SVM ensemble and require fine parameter tuning.

**Table 2: Comparing data labels reported by AUSL and ALCD using two data sets.**

| **Data Set** | | **AUSL** | | **ALCD** | |
|---|---|---|---|---|---|
| Wine | C# | Attribute | Range | Attribute | Range |
| | 1 | Alcohol<br>Color intensity | 11.03~12.52<br>1.28~3.74 | Alcohol<br>Color intensity<br>Magnesium | 11.03~12.52<br>1.28~3.74<br>70~90 |
| | 2 | Flavanoids<br>OD280/OD315 | 0.34~1.50<br>1.27~2.30 | Flavanoids<br>OD280/OD315<br>Hue | 0.34~1.50<br>1.27~2.30<br>0.48~0.87 |
| | 3 | Total phenols<br>Proline | 2.61~3.88<br>835~1680 | Total phenols<br>Proline<br>Flavanoids | 2.61~3.88<br>835~1680<br>2.65~5.08 |
| WheatSeeds | 1 | Area<br>Kernel Length<br>Kernel Width | 16.17~21.18<br>5.83~6.67<br>5.65~6.55 | Area<br>Perimeter | 16.17~21.18<br>15.18~17.25 |
| | 2 | Area<br>Kernel Width | 10.59~12.80<br>2.63~3.04 | Area<br>Kernel Length | 13.67~15.18<br>4.90~5.36 |
| | 3 | Area<br>Perimeter | 12.80~16.17<br>13.67~15.18 | Area<br>Perimeter | 12.80~16.17<br>13.67~15.18 |

In the next experiments, we compare the data labels generated by AUSL with that of ALCD on two data sets viz. Wine and WheatSeed using H>=80% and F>=0.8. The experiments were recreated as reported in ALCD using 3 bins for discretisation of the attributes and keeping the number of clusters (k) same as number of classes in the data set. Table 2 shows the comparative data labels reported by ALCD and AUSL. The obtained clusters for both methods are mapped using majority of overlapped data points. In case of Wine data, both methods have reported two common attributes with same data range in all three clusters. However, ALCD has reported one additional attribute for each cluster that was dominant attribute. This attribute is removed in case of AUSL because it could not meet the confidence threshold of its rule patterns in the

corresponding cluster. Results on WheatSeed data also vindicate better labeling by AUSL due to detection of finer patterns, as compared to ALCD. Manual inspection of the data also confirms the obtained results. Final or similar data range, but with superior Jaccard score affirms (see Figs. 3 and 4) better capability of AUSL to deliver refined data labels as per underlying patterns and structural associations. Finer data labels with improved performance captured through higher Jaccard similarity score on the labeled data vindicate the effectiveness of the proposed approach. Therefore, applying the method to unlabeled data will similarly showcase its capacity to accurately identify labels that truly represent the underlying data patterns.

## DISCUSSION

The proposed AUSL framework combines unsupervised ensemble clustering with supervised AdaBoost SVM and unsupervised association rule mining to automate data labeling for structured data. AUSL reduces manual effort while utilizing the strengths of both types of machine learning approaches. Experimental results on four datasets show that AUSL achieves finer, more interpretable labels than an existing method and its variation, with average hit rates exceeding 90% and confidence levels above 80%, thus demonstrating robust and reliable label generation. Also, the validation of the generated labels with those computed using ground-truth class labels using Jaccard score affirms the robustness of the proposed approach. These findings imply that AUSL is highly effective for applications requiring large-scale, accurate data labeling, such as customer relationship management, target marketing, and recommender systems.

### Limitations

Although the approach has been proved to be highly effective, it incurs an increased computational complexity for the high dimensional data due to integration of multiple machine learning algorithms [37] and is sensitive to choices of parameter values in association rule mining, which still require manual intervention. Overall, AUSL provides an automatic, scalable and interpretable approach that enhances automated data labeling, while also highlighting opportunities for future enhancements in efficiency, scalability and adaptability to large unstructured data.

## CONCLUSION

This study provides an innovative method AUSL, which is an amalgamation of ensemble clustering, Adaboost SVM ensemble and association rule mining for labeling the structured data. Ensemble clustering is employed to get robust clustering scheme so that generated data labels are of good quality. Original numeric data are discretized using equal frequency discretization method to get discrete values imperative for classification and association rule mining in the proposed methodology. The instances of the obtained clusters are then mapped to discretized space and assessed by Adaboost SVM ensemble for revealing corresponding dominant attributes. Subsequently, association rule mining is applied for identifying prevalent characteristics of the dominant attributes and those with high confidence are reported as data labels by mapping to original data space. The finer data labels in terms of rules not only capture the dependency between data characteristics but also reflects the refined domain range of the corresponding attribute. Future work could focus on scaling the method to high-dimensional

data, adapting it for unstructured formats like text or images, and incorporating automated parameter tuning through optimization or weak supervision techniques.

## References


[1] Sun Y, Lank E, Terry M. Label-and-learn: Visualizing the likelihood of machine learning classifier's success during data labeling. In Proceedings of the 22nd International Conference on Intelligent User Interfaces 2017 Mar 7 (pp. 523-534).

[2] Chavez T, Zhao Z, Jiang R, Koepp W, McReynolds D, Zwart PH, Allan DB, Gann EH, Schwarz N, Ushizima D, Barnard ES. A machine-learning-driven data labeling pipeline for scientific analysis in MLExchange. Applied Crystallography. 2025 Jun 1;58(3).

[3] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP. Preparing medical imaging data for machine learning. Radiology. 2020 Apr;295(1):4-15.

[4] Zhang S, Jafari O, Nagarkar P. A survey on machine learning techniques for auto labeling of video, audio, and text data. arXiv preprint arXiv:2109.03784. 2021 Sep 8.

[5] Wang P, Vasconcelos N. Towards professional level crowd annotation of expert domain data. Proc IEEE CVPR. 2023;3166–3175.

[6] Fredriksson T, Mattos DI, Bosch J, Holmström Olsson H. Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In: Morisio M, Torchiano M, Jedlitschka A, editors. Product-Focused Software Process Improvement. Cham: Springer International Publishing; 2020. p. 202–216.

[7] Chegini M, Bernard J, Berger P, Sourin A, Andrews K, Schreck T. Interactive labeling of a multivariate dataset for supervised machine learning using linked visualisations, clustering, and active learning. Visual Informatics. 2019;3(1):9–17.

[8] Beil D, Theissler A. Cluster-clean-label: An interactive machine learning approach for labeling high-dimensional data. In: Proceedings of the 13th International Symposium on Visual Information Communication and Interaction. ACM; 2020.

[9] Lopes LA, Machado VP, Rabêlo RAL, Fernandes RA, Lima BV. Automatic labeling of clusters of discrete and continuous data with supervised machine learning. Knowledge-Based Systems. 2016;106:231–241.

[10] Jain AK. Data clustering: 50 years beyond k-means. Pattern Recogn Lett. 2010;31(8):651–666.

[11] Wasilewski A. Customer segmentation in e-commerce: A context-aware quality model for comparing clustering algorithms. Journal of Internet Services and Applications. 2024 Jul 25;15(1):160-78.

[12] Tohalino JV, Amancio DR. Extractive multi-document summarization using multilayer networks. Physica A. 2018;503:526–539.

[13] Aziz NAM, Ali AA, Naguib SM, Fayed L. Clustering-based topic modeling for biomedical documents extractive text summarization. J Supercomput. 2025;81(1):171.

[14] Peganova I, Rebrova A, Nedumov Y. Labeling hierarchical clusters of scientific articles. In: 2019 Ivannikov memorial workshop. IEEE; 2019. p. 26–32.

[15] Cozzolino I, Ferraro MB. Document clustering. Wiley Interdiscip Rev Comput Stat. 2022;14(6):e1588.

[16] Vădineanu S, Pelt DM, Dzyubachyk O, Batenburg KJ. Reducing manual annotation costs for cell segmentation by upgrading low-quality annotations. Journal of Imaging. 2024 Jul 17;10(7):172.

[17] Fredriksson T, Bosch J, Olsson HH. Machine learning models for automatic labeling: A systematic literature review. In: 15th Int Conf on Softw Technologies ICSOFT. SciTePress; 2020. p. 552–561.

[18] Dol SM, Jawandhiya PM. Classification technique and its combination with clustering and association rule mining in educational data mining—a survey. Eng A. 2023;122:106071.

[19] Najafabadi MN, Mahrin NM, Chuprat S, Sarkan HM. Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data. Comput Human Behav. 2017;67:113–128.

[20] Strehl A, Ghosh J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res. 2002;3(Dec):583–617.

[21] Bhatnagar V, Ahuja S, Kaur S. Discriminant analysis-based cluster ensemble. Int J Data Min, Mod Manag. 2015;7(2):83–107.

[22] Mehmood Z, Asghar S. Customizing SVM as a base learner with adaboost ensemble to learn from multi-class problems: A hybrid approach adaboost-msvm. Knowl Based Syst. 2021;217:106845.

[23] Silva ALMLS, Neres FJOD, Mendes APSdS, Machado VP, Santana AM, Rabêlo RAL. Method for inferring the number of clusters based on a range of attribute values with subsequent automatic data labeling. Proc Comput Sci. 2023;222:561–570.

[24] Esmaeili SA, Duppala S, Dickerson JP, Brubach B. Fair labeled clustering. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2022 Aug 14 (pp. 327-335).

[25] Kennedy RK, Villanustre F, Khoshgoftaar TM, Salekshahrezaee Z. Synthesizing class labels for highly imbalanced credit card fraud detection data. Journal of Big Data. 2024 Mar 9;11(1):38.

[26] Cozzolino I, Ferraro MB. Document clustering. Wiley Interdiscip Rev Comput Stat. 2022;14(6):e1588.

[27] Kaya MF. Pattern labeling of business communication data. Group Decision and Negotiation. 2022 Dec;31(6):1203-34.

[28] Zhang X, Ge Y, Qiao Y, Li H. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2021 (pp. 3436-3445).

[29] Lin, Yutian, et al. "A bottom-up clustering approach to unsupervised person reidentification." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[30] Zhang H, Peng Y. Image clustering: An unsupervised approach to categorize visual data in social science research. Sociological Methods & Research. 2024 Aug;53(3):1534-87.

[31] Chelebian E, Avenel C, Ciompi F, Wählby C. DEPICTER: Deep representation clustering for histology annotation. Computers in Biology and Medicine. 2024 Mar 1;170:108026.

[32] Golalipour K, Akbari E, Hamidi SS, Lee M, Enayatifar R. From clustering to clustering ensemble selection: A review. Engineering Applications of Artificial Intelligence. 2021 Sep 1;104:104388.

[33] Kumbhare TA, Chobe SV. An overview of association rule mining algorithms. Int J Comput Sci Info Technol. 2014;5(1):927–930.

[34] Kelly M, Longjohn R, Nottingham K. The UCI Machine Learning Repository, https://archive.ics.uci.eduAsuncion A, last accessed: May 2025.

[35] Geng L, Hamilton HJ. Interestingness measures for data mining: A survey. ACM Comput Surv. 2006;38(3):9–es.

[36] Fletcher S, Islam Z, et al. Comparing sets of patterns with the Jaccard index. Australasian J Inf Syst. 2018;22.

[37] Wilson A, Anwar MR. The Future of Adaptive Machine Learning Algorithms in High-Dimensional Data Processing. International Transactions on Artificial Intelligence. 2024 Nov 22;3(1):97-107.