SEMI-AUTOMATIC LABELING OF BIG DATA FOR AI TRAINING

YU-TING HWANG, CHIH-YU WEN

Department of Electrical Engineering,
National Chung Hsing University,
No.145, Hsing-Da Road, Taichung, Taiwan
{E-mail: huangxiaode22@gmail.com, cwen@dragon.nchu.edu.tw}

Abstract: Abstract: To tackle the issue of labeling for AI training, this paper firstly introduces a semi-automatic big data labeling method based on confusion area analysis (CAA), minimizing labeling effort and precisely identify classification errors caused by unsupervised clustering algorithms. Initially, the method utilizes unsupervised clustering algorithms to automatically classify big data, followed by identifying confusion areas based on classification results. The vast majority of data resides in non-confusion areas, where each cluster can be manually labeled with minimal effort. Only the small fraction of data in confusion areas requires individual manual labeling. Experimental results indicate that the manual labeling rate can be reduced to less than 1%, while the accuracy of the semi-automatically labeled data can reach 90.78%, surpassing previous approaches. Moreover, this work proposes a method for data parameter transformation to enhance the efficiency of unsupervised learning and further reduce manual labeling rates. Instead of using traditional Cartesian coordinates, this work develops the polar coordinate and fingerprint parameter transformations, based on scale-invariant feature transform (SIFT) algorithm, for handwritten digit recognition, where the experimental results demonstrate an improvement in unsupervised learning accuracy by 1-3%, with the overall accuracy of semi-automatically labeled data reaching 93.39%. Keywords: semi-automatic labeling, unsupervised learning, confusion area analysis, handwritten digit recognition.

1. Introduction

The outstanding performance of AI supervised learning frameworks has been widely applied in various fields. However, supervised learning requires a large amount of labeled data for training, which necessitates professional expertise and consumes significant time. Therefore, automatic labeling of big data has become a critical research topic. In previous studies, Hinkle et al. [1] introduced an interactive UI with visual synchronization, parameter display, and time-series visualization tools to assist in manually labeling big data. Desmond et al. [2] proposed predictive models to guide and assist manual labeling; these predictive models learn by continuously observing the manual labeling process and integrate UI tools to facilitate manual labeling. Although these methods utilize UI tools to accelerate the labeling process, manual effort is still required to label each data point individually. The manual labeling rate for these methods remains as high as 100%.

Moreover, Ferreira et al. [3] proposed a supervised method using Support Vector Machines (SVM) to learn from 10% of pre-labeled text data. The remaining 90% of the text dataset is then automatically labeled, achieving a semi-automatic labeling goal with an accuracy of up to 98%. However, this method requires at least 10% of the data to be manually labeled, and it lacks analysis and discussion on more ambiguous or confusing data. By combining multiple parameter transformations and various unsupervised methods, Vajda et al. proposed semi-automatic labeling methods [4–7], which first automatically cluster big data, and then categorize data based on multiple clustering results using a majority-vote strategy. This approach requires minimal manual labeling effort to achieve a semi-automatic labeling goal. For example, using the MNIST handwritten digit database, an accuracy of 89.13% was achieved.

This study, similar to the purpose of the works proposed by Vajda et al. in [4–7], aims to develop a semi-automatic annotation system based on unsupervised methods. However, the approaches adopted to achieve this objective are fundamentally different. In [4–7], a voting mechanism is employed to determine the correct labels for the data, which is not only time-consuming but also requires a larger quantity of annotated data, thereby increasing the system's overall workload. In contrast, the method proposed in this study focuses on identifying ambiguous regions within the data and performing precise annotations specifically in those areas, thereby improving annotation accuracy and system efficiency. This paper proposes a method focused on detecting confusing data, with the primary goal of improving the accuracy of automatic labeling. Initially, the method leverages unsupervised learning to extract a very small amount of data prone to classification errors, followed by manual labeling. Experimental results demonstrate that the proposed method achieves exceptionally high recognition accuracy (%) in semi-

automated big data labeling, while maintaining a very low manual labeling ratio (%).

The rest of the paper is organized as follows: Section 2 details the confusion area detection method, Section 3 presents experimental results and analyzes system performance, and Section 4 concludes the study.

2. The Proposed Methodology

The classification of large data sets requires professional expertise, which implies substantial manual effort. Therefore, the development of semi-automatic labeling techniques for large data is the focus of this paper. Within large data sets, the majority of data points in multidimensional space cluster together, forming distinct groups, as illustrated in Figure 1. These distinct clusters represent different characteristics and classifications of the data. However, a small portion of the data exhibits ambiguous spatial distributions, which we refer to as "confusing data (CD)". These data reside in the boundary regions between two clusters, known as "confusion areas (CA)". Such data often lead to misclassification during automatic categorization. Consequently, relying solely on unsupervised learning with limited parameters is insufficient to accurately classify these confusing data.

Therefore, this paper proposes a confusion area detection method. The method aims to identify confusion areas precisely and within a small scope to encompass confusing data as much as possible. Manual labeling is then applied to these confusing data to improve the accuracy of semi-automatic labeling.

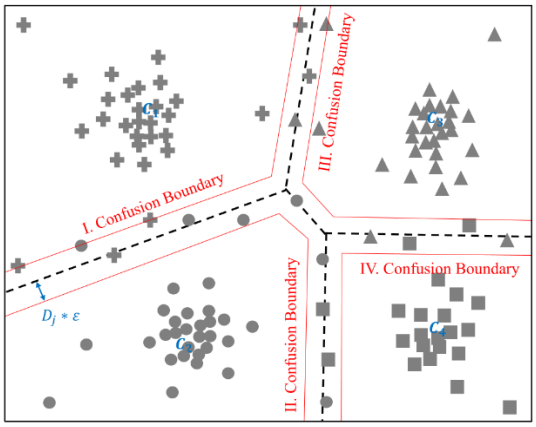

Figure 1. Illustration of confusing data in clustering: the cluster centroids and confusion boundaries.

2.1 The Confusing Data Detection (CDD) Algorithm

To classify these confusing data, the proposed confusing data detection (CDD) algorithm begins with unsupervised clustering algorithms to classify (M) data points ($X_i, 1 \le i \le M$) into N clusters (codewords). Next, the centroids of the N clusters ($C_j, 1 \le j \le N$) are calculated. Subsequently, for each data point X_i, the two closest centroids (e.g., C_{i1} and C_{i2}) are identified. Finally, based on distance ratios or boundary conditions, the confusing data are determined through Methods 1 and 2 as described in Algorithm 1. The detailed steps are presented below:

Algorithm 1. The methodology for the CDD algorithm	
Step	Action description
1	Use unsupervised algorithm (e.g., vector quantization (VQ) or self-organizing map (SOM) algorithm to classify unlabeled Data($X_i, 1 \le i \le M$) into N clusters.
2	Calculate the centroid C_j of j-th cluster. $C_j = \frac{1}{N_j}\sum_{k=1}^{N_j} X_k, 1 \le j \le N$, N_j is the number of data point X_i which is fall in j-th cluster.
3	The number n of confusing data is set to zero. n & i set to 0.
4	***Method 1:*** 1. For data point X_i, search the centroid C_{i1} and C_{i2}, which are the closest and second closest centroids to the data sample X_i. 2. Calculate the distance ratio $R_i = \frac{\|X_i - C_{j1}\|}{\|X_i - C_{j2}\|}$ 3. If $\|R_i - 1.0\| < \delta$, then the data point X_i is set as confusing data. (δ is a small value) ***Method 2:*** 1. For data point X_i, search the centroid C_i, which is the closest centroid to the data point X_i. (i.e. $C_i = C_{i1}$) 2. Calculate the maximum distance $D_j = \max(\|X_i - C_i\|)$ 3. If $D_{ij}/D_j > \varepsilon$, then the data point X_i is set as confusing data under the condition. (ε is a small value)
5	Next i, Repeat Step.4, until $i \ge N$

Finally, manual labeling is applied to the categories of N codewords and n confusing data points. Therefore, the manually labeling rate (MLR) is calculated as $(N + n)/M$. The confusion values δ and ε are numbers close to 0 and 1, respectively. The selection of δ and ε is determined by the percentage of the most confusing data points to be extracted. As for the accuracy analysis of semi-automatic labeling, it relies on manually labeled data for evaluation. In this study, we use the handwritten MNIST dataset for analysis, and detailed experimental data analysis is presented in the next section.

2.2 Feature Parameter Transformations

Furthermore, for the CDD algorithm, this paper adopts unsupervised clustering methods such as VQ and SOM. It also

employs multiple feature parameter transformations, including Cartesian coordinate data, polar coordinate transformation, and fingerprint transformation for cross-referencing. Note that the Cartesian coordinate data and polar coordinate transformation respectively present the original image data type and the conversion of the Cartesian coordinate system to polar coordinates.

For the fingerprint transformation, the original data is subjected to circular segmentation. Assuming the angle is divided into A segments and the radius is divided into S segments, the resulting fingerprint data dimensions are A*S. This method applies the concept of scale-invariant feature transform (SIFT) algorithm, where each dataset is segmented and computed based on different kernels to create fingerprint data. Since the fingerprint data dimensions obtained from each kernel are identical, this method is highly suitable for clustering using VQ or SOM. The following are kernels applied for image feature transformation:

- Kernel 1: After dividing the angles and radius evenly, the average value within each region is calculated as the fingerprint parameter.
- Kernel 2: After evenly segmenting the angles and radius, the average value and standard deviation of all points within each region are computed as the fingerprint parameters.
- Kernel 3: The original image radius is mapped to an exponential function space to achieve exponential distribution segmentation larger blocks near the center and finer ones towards the edges. The average value within each region is then computed as the fingerprint parameters.
- Kernel 4: Following uniform division of angles and radius, the average value within each region is calculated as the fingerprint parameter. The process is repeated with the image rotated ±5 degrees, generating additional fingerprint parameters.

Detailed experimental data and performance analyses are presented in Section 3.

3. Performance Evaluation

Section 3 conducts experimental verification of the proposed semi-automatic labeling algorithm based on CA analysis. The dataset used for verification is the MNIST handwritten digit dataset, consisting of 60,000 training samples and 10,000 testing samples. Each data point has a manually labeled ground truth, making it highly suitable for performance evaluation of the accuracy of semi-automatic labeling algorithms. The experiments are divided into two parts. The first part evaluates the performance of unsupervised classification algorithms, while the second part focuses on the performance evaluation of the CDD algorithm, which is the core emphasis of this paper.

3.1 Performance with Feature Transformations

In the first part, the unsupervised classification algorithms (i.e., VQ and SOM) are applied. Three types of feature parameters, including Cartesian coordinate data, polar coordinate data, and fingerprint-transformed data, are tested for cross-verification. Table 1 presents the accuracy of automatic classification. From Table 1, it is evident that unsupervised clustering on the MNIST dataset achieves an automatic classification accuracy of up to 91.53%. Among the algorithms, SOM slightly outperforms VQ, and in terms of feature transformations, fingerprint transformation and polar coordinate transformation slightly outperform Cartesian coordinates. This part of the results, compared to the accuracy rate of 89.13% reported in [4], demonstrates an improvement of 2.4%, reaching 91.53%. This indicates a slight enhancement in the accuracy of automatic classification achieved by the proposed method in this study.

As shown in Table 1, there is no significant difference in classification accuracy between the Kernel-based method and the Cartesian-based method. However, a clear difference can be observed in terms of data size. When the data size decreases, meaning the number of data points is reduced, the classification time is shortened accordingly, which enhances classification efficiency. Furthermore, Table 2 presents the corresponding data sizes for each Kernel and Cartesian method, further illustrating the differences in data volume between these approaches.

TABLE 1. Correction Rate with different unsupervised methods and feature transformations.

Feature \ Alg.	VQ	SOM
Cartesian Coordinate	90.22%	91.41%
Fingerprint (Kernel 1)	90.91%	91.46%
Fingerprint (Kernel 2)	90.36%	91.53%
Fingerprint (Kernel 3)	90.33%	90.82%
Fingerprint (Kernel 4)	90.65%	91.36%

TABLE 2. Comparison of Data Sizes for Each Kernel and Cartesian Method

Feature	Data Size (Number of Data Points)
Cartesian Coordinate	784
Fingerprint (Kernel 1)	512
Fingerprint (Kernel 2)	512
Fingerprint (Kernel 3)	512
Fingerprint (Kernel 4)	512

3.2 Performance of the CDD Algorithm

The second part of the experimental verification analyzes the performance of the CDD Algorithm, which is used to identify mislabeled data resulting from automatic clustering and classification of large datasets. Initially, the relevant

parameters of the algorithm are defined in Table 3.

TABLE 3. Relevant parameters of the CDD algorithm

Parameters	*definition*
M	The number of large dataset entries. (MNIST dataset size, M = 60,000)
N	The number of clusters (Codewords) after unsupervised automatic clustering. The MNIST dataset contains 10 categories, so N must be greater than or equal to 10. At this stage, these N clusters are manually labeled into categories.
E	The number of classification errors after unsupervised automatic clustering and labeling of large datasets. (Based on Table 1, $E \approx M * 10\%$)
D	The maximum distance from the center point to the boundary within a cluster after automatic clustering.
R	The ratio of the distance of each data point to its two closest cluster centroids.
K1 (k1%)	Using the Algorithm 1 with Method 1, the ratio of the distance of each data point to its two closest cluster centroids is calculated to identify and delineate the confusion area data, R ≥ k1%.
K2 (k2%)	Using the Algorithm 1 with Method 2, a total of K2 data points with boundary range distance (k2%) were identified as confusing data, D ≥ k2%.
C (c%)	The number of mislabeled data points among the K data points identified by the Algorithm 1, where $C = K \cap E$ and $c\% = \frac{C}{K}$.
P (p%)	The number and proportion of manually labeled data points, where $P = N + K$ and $p\% = \frac{P}{M}$.
Q (q%)	The total number and proportion of correctly labeled classifications, where $Q = M - (E - C)$ and $p\% = \frac{Q}{M}$.

In the semi-automatic labeling and classification results under the VQ with Cartesian model, the dataset consists of a total of 60,000 data points (M = 60,000). The unsupervised VQ clustering method groups the data into 256 clusters (N = 256). After clustering, these 256 clusters are manually labeled into categories from 0 to 9. Subsequently, the 60,000 data points are quantized and automatically labeled based on these 256 clusters. Among them, 6,000 data points (10%) had discrepancies between VQ automatic classification and their original classifications, resulting in (E = 6,000) automatic classification errors.

Based on the aforementioned unsupervised clustering results, the Algorithm 1 is used to identify and analyze the top 1% to 10% of the most confusing data points. Tables 4 through 7 present various metrics, including the hit rate (c%) of confusing data detection, the ratio of manually labeled data (P, p%), and the overall accuracy of labeling and classification (Q, q%).

TABLE 4. VQ with Method 1

K1(k%)	874 (0.99)	2289 (0.98)	4218 (0.97)	6192 (0.96)	8194 (0.95)
C (c%)	196 (22.4%)	501 (21.8%)	862 (20.4%)	1271 (20.5%)	1656 (20.2%)
P (p%)	1130 (1.89%)	2545 (4.24%)	4474 (7.46%)	6448 (10.7%)	8450 (14.0%)
Q (q%)	90.87%	91.38%	91.98%	92.66%	93.31%

TABLE 5. VQ with Method 2

K2(k%)	297 (1%)	472 (5%)	809 (10%)	2155 (20%)	5629 (30%)
C (c%)	141 (47.4%)	216 (45.7%)	349 (43.1%)	790 (32.9%)	1705 (30.2%)
P (p%)	553 (0.92%)	728 (1.21%)	1065 (1.77%)	2411 (4.02%)	5885 (9.81%)
Q (q%)	90.78%	90.90%	91.13%	91.86%	93.39%

TABLE 6. SOM with Method 1

K1(k1%)	3919 (0.99)	7630 (0.98)	10994 (0.97)	14211 (0.96)	17090 (0.95)
C (c%)	681 (17.3%)	1286 (16.8%)	1770 (16.1%)	2192 (15.4%)	2569 (15.0%)
P (p%)	5519 (9.20%)	9230 (15.3%)	12594 (20.9%)	15811 (26.3%)	18690 (31.1%)
Q (q%)	92.36%	93.36%	94.17%	94.88%	95.50%

TABLE 7. SOM with Method 2

K2(k%)	1717 (1%)	3147 (5%)	6496 (10%)	20016 (20%)	38512 (30%)
C (c%)	488 (28.4%)	876 (27.8%)	1561 (24.0%)	3468 (17.3%)	4903 (12.7%)
P (p%)	3317 (5.53%)	4747 (7.91%)	8096 (13.4%)	21616 (36.0%)	40112 (66.8%)
Q (q%)	91.85%	92.50%	93.64%	96.82%	99.21%

4. Discussion

The results of this study indicate that data located at the boundaries of clusters are more prone to classification errors, leading to higher confusion rates. Compared to data within the cluster, boundary data exhibit vague feature distributions with smaller differences from neighboring clusters, making it more challenging for classifiers to distinguish them accurately.

In detecting confusion areas, VQ outperforms SOM as the confusion areas identified by VQ contain a higher proportion of confusing data. According to the experimental results in Section 3, the confusion areas delineated by VQ encompass a larger proportion of confusing data, demonstrating its stronger capability in detecting cluster boundaries. This may be attributed to the vector quantization mechanism of VQ, which allows it to more precisely identify highly confusing regions

and enhance its ability to capture confusing data.

According to the experimental results in Section 3, VQ with Method2 demonstrates the best performance in detecting cluster boundaries, capturing up to 47.4% of confused data. Regardless of whether VQ or SOM is used, Method2 consistently achieves higher accuracy with fewer labeled samples, highlighting its efficiency in reducing annotation effort while maintaining classification performance.

5. Conclusion

This study explores the application of semi-automatic annotation technology based on confusion area analysis in image processing and verifies its effectiveness in improving classification accuracy and efficiency. The proposed algorithm effectively identifies confusion area in image classification, allowing manual annotation to focus on key data points, thereby reducing annotation costs while enhancing classification performance. The results show that combining confusion area analysis with unsupervised classification significantly reduces annotation time and labor costs while maintaining classification performance comparable to supervised methods. Additionally, selecting appropriate feature transformation methods and unsupervised classification techniques can further enhance model accuracy, making this approach more adaptable and practical. However, challenges remain, as the performance of unsupervised classification may be limited in more complex image datasets. Future research should focus on improving feature selection strategies and confusion area identification, as well as integrating deep learning techniques to enhance model adaptability.

In conclusion, this study proposes an effective combination of semi-automatic annotation and unsupervised classification, offering a promising solution for image classification, especially for large-scale datasets with high annotation costs. Future work will emphasize optimizing confusion area analysis, improving automatic annotation accuracy, and incorporating advanced machine learning and deep learning techniques to further expand the applicability of this approach.

References

[1] Lee B. Hinkle; Tristan Alexander Rech; Tyler Lynn; Gentry Atkinson; Vangelis Metsis, “Assisted learning visualizer (ALVI): a semi-automatic learning system for time-serial data”, IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), 2023.

[2] Michael Desmond, Evelyn Duesterwald, Kristina Brimijoin, Michelle Brachman, and Qian Pan, “Semi-Automated Data Labeling”, Journal of Machine Learning Research Vol.133, pp.156–169, 2021.

[3] Alessandro dos Santos Ferreira, Daniel Matte Freitas, Gercina Goncalves da Silva, Hemerson Pistori, and Marcelo Theophilo Folhes, “Unsupervised deep learning and semi-automatic data labeling in weed discrimination”, Computers and Electronics in Agriculture, pp.1~11, 165 (2019) 103963, 2019.

[4] Szilard Vajda; Akmal Junaidi; and Gernot A. Fink, “A Semi-supervised Ensemble Learning Approach for Character Labeling with Minimal Human Effort” 2011 International Conference on Document Analysis and Recognition 2011.

[5] Szilard Vajda, Daekeun You, Sameer K. Antani, George R. Thoma, “Label the many with a few: Semi-automatic medical image modality discovery in a large image collection”, IEEE Symposium Series on Computational Intelligence, 2014.

[6] Szilard Vajda, · Daekeun You · Sameer Antani · George Thoma, “Large image modality labeling initiative using semi-supervised and optimized clustering”, International Journal of Multimedia Information Retrieval June 2015.

[7] Szilárd Vajda, Yves Rangoni, and Hubert Cecotti, “Semi-automatic ground truth generation using unsupervised clustering and limited manual labeling: Application to handwritten character recognition”, Pattern Recognition Letters, pp.23~28, Vol.58, 2015.