International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 11 Issue 10, October-2022, Impact Factor: 7.957

Classification and Clustering using Intelligent Technique

Vrushali R. Parkhede[1], Prof. J.M. Patil[2]

[1]PG Student, Computer Engineering, Shri Sant Gajanan Maharaj College of Engg Shegaon, India
[2]Prof. CSE Dept., Shri Sant Gajanan Maharaj College of Engg Shegaon, India

ABSTRACT

Data is more readily available than ever before in the age of information, but processing it poses its own set of obstacles. As one of the most common uses for machine learning, data mining may be used to identify patterns in a dataset. The end outcome of this procedure is referred to as database knowledge discovery. They should be valuable and pertinent, but humans aren't usually good at finding these emerging patterns in large amounts of raw data. It's becoming increasingly common to use artificial intelligence (AI) to help with everything. In the current study, we employ K-means clustering and Support Vector Machines (SVM) to classify the DIGIT dataset (SVM). The dimensionality of feature vectors can be reduced using Principal Component Analysis (PCA). Cluster metrics are evaluated based on their purity. Prior knowledge of the true class is used to determine the validity of the cluster.

***Index Terms*—Artificial Intelligence, Cluster Metrics, Data Mining, Machine Learning.**

INTRODUCTION

A. Motivation

The ability of a human to solve and detect problems is unmatched, but this is not the case in the case of a computer. To function as a human, a variety of approaches and methodologies should be incorporated. After taking into account all of the accomplishments that have been made in this field, there is still a huge research gap that needs to be addressed. Consider the difference between online handwriting recognition and offline handwriting recognition. Because stroke information is gathered dynamically in online handwriting recognition of letters, an on-the-fly compilation of letters is conducted while writing in online handwriting recognition of letters. In contrast, while using offline recognition, the letters are not collected in real time. Because of the lack of information, online handwriting recognition is more accurate when compared to offline handwriting recognition. As a result, it is possible to conduct research in this field in order to improve offline handwriting recognition. The most difficult task in offline handwriting recognition is recognising the character of words in the handwriting. When recognising the characters of a word in offline handwriting, there are several ways that can be used.

B. Objectives

1. In order to recognise handwritten digits. (because handwritten digits' recognition includes a wide range of options for how the digits are written)
2. Utilize the Classification technique to categorise each object based on its properties.
3. Using clustering, collect similar data points and group them into clusters.

C. Approach

Classification of handwritten digits is done using SVM. In this work use of the DIGIT database for handwritten digits and SVM to categorise numbers 0 to 9. Digits Dataset is a collection of data that is a component of the sklearn library. A large number of datasets for practising machine learning techniques are included with Sklearn, and one of these datasets is digits. Digits has 64 numerical features (8x8 pixels) and a target variable with a ten-class classification system (0-9). The Digits dataset can be used for both classification and clustering purposes. This dataset contains 1797 images with a resolution of 8x8 pixels. Each image depicts a digit that has been written by hand. The transformation of an 8x8 figure into a feature vector with a length of 64 would be necessary before we could use it. Following that, the performance of K-means clustering is examined. The K-means method is reviewed and compared to other algorithms for testing the missing clusters for a total of ten clusters.

LITERATURE SURVEY

Jayadevan, Kolhe, Patil and Pal[1] discussed in their paper about automatic processing of handwritten bank cheque images. Nasrabadi[2] in his paper discussed pattern recognition using machine learning techniques. Plamondon and Srihari[3] in their paper comprehensively discussed various online and offline handwriting recognition techniques. Bozinovic and Srihari[4] showed us the methods of word recognition on offline cursive scripts.Marukawa, Koga, Shima and Fujisawa[5] in their paper discussed techniques of hand written Chinese character recognition using high speed matching algorithm. Smith, McNamara and Bradburn[6] applied various pattern classification techniques to character segmentation in their work. Lauer, Suen and Bloch[7] discussed about trainable feature extractor for handwritten digit recognition in their research. Lopes, Da Silva, Rodrigues and Filho[8] discussed in their paper about recognition of handwritten digits using the signature features and optimum path forest classifier. Oliveira, Sabourin, Bortolozzi and Suen[9] discussed a methodology for feature selection using multi objective genetic algorithms for handwritten digit recognition. Fanany[10] in their work discussed handwriting recognition on form document using convolutional neural network and support vector machines. Nju and Suen[11] developed a novel CNN-SVM classifier for recognizing handwritten digits.

METHODOLOGY

Following the digitization process, a grayscale image is created. Then, using Otsu's method, the thresholding technique is done to create binary images, which are then resized so that all of the images have the same height and width. Finally, the image is enlarged such that all of the images have the same height and width. Support vector machine is utilized for recognition and classification of digits. Unsupervised approach for clustering is used by K-means clustering.

A. Recognizing hand written digits

The whole system is developed in Python. In order to recognise images of handwritten digits ranging from 0-9 scikit-learn library is used.

B. Digit Dataset

The digits dataset contains images of digits that are 8x8 pixels in size. The grayscale values for each image are stored in an 8x8 array of grayscale values in the images property of the dataset. These arrays will be used to visualise the first four photos in this section. The target attribute of the dataset holds the digit that each image represents, and this information is incorporated in the titles of the nine plots shown as following.

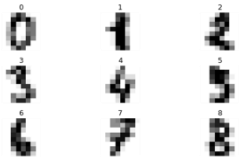

Fig. 1 shows the handwritten digits to be plotted.

C. Classification

If we want to use a classifier on this data, we must first flatten the images, which means converting each 2-D array of grayscale values from shape (8, 8) into shape (8, 8). (64,). Therefore, the full dataset will have the shape (n samples, n features), where n samples is the number of photos and n features is the total amount of pixels in each image.

Afterwards, we can divide the data into train and test subsets, and then we can fit a support vector classifier on the train samples. The fitted classifier can then be used to predict the value of the digit for the samples in the test subset using the data from the training set. The whole system is developed in Python. In order to recognise images of handwritten digits ranging from 0-9 scikit-learn library is used.

D. K-Means algorithm for clustering

K-Means Clustering is an unsupervised learning approach that is used to address clustering problems in machine learning or data science. This is also known by name as K-Means Clustering.

K-Means Clustering is an Unsupervised Learning technique that divides an unlabelled dataset into various clusters based on their similarities. K defines the number of pre-defined clusters that must be created in the process. In addition, it provides an easy technique to find the categories of groups in the unlabelled dataset on its own, without the need for any additional training or preparation.

In this algorithm, each cluster is associated with a centroid, and the algorithm is based on centroid-based data. The primary goal of this approach is to reduce the sum of distances between data points and their respective clusters to the smallest possible value. A clustering method takes an unlabeled dataset as input, separates the dataset into k-number of clusters, and repeats the procedure until the best clusters are not found. In this algorithm, the value of k must be predetermined.

As a result, each cluster contains datapoints that share some characteristics and is isolated from the others.

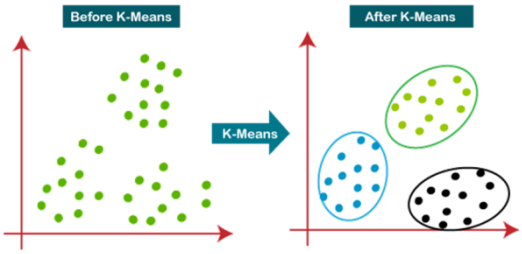

Fig. 2 shows clustering of datapoints before and after applying K-Means algorithm.

RESULT

A. Result of classification using digits dataset
Accuracy on the test set using SVM found 97.03%
Accuracy using KNN score: 96.11%
Logistic Regression score: 93.33%

Fig. 3 shows output of the system

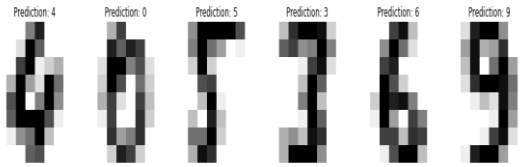

Fig. 4 shows image of prediction

From the above image of prediction, it is found that, all the six digits predicted are found correct. In the first column predicted digit images id of 4 and its accurately predicted. In the same way further 0, 5, 3 6 and 9 are also predicted correctly.

Table 1 Classification summary

Class	Precision	Recall	F1-score	Support
0	1.00	0.98	0.99	53
1	0.96	1.00	0.98	53
2	1.00	0.98	0.99	53
3	0.96	0.89	0.92	53
4	0.98	0.95	0.96	57
5	0.95	0.98	0.96	56
6	0.98	0.98	0.98	54
7	1.00	1.00	1.00	54
8	0.91	0.98	0.94	52
9	0.96	0.96	0.96	55
Accuracy			0.97	540
Macro avg	0.97	0.97	0.97	540
Weighted avg	0.97	0.97	0.97	540

From the classification report it is found that the digits 3 and 8 are having f1-score less than 92 and 94 respectively. For digit 0, 2 and 7 precisions are of 1 means 100% precision.

B. Result of clustering

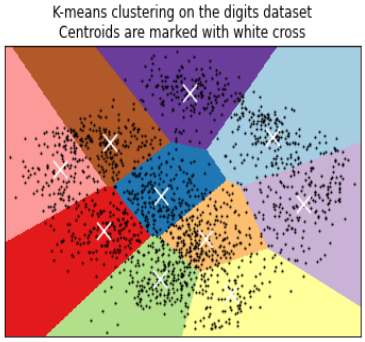

Fig. 5 K-means clustering on digits dataset

Table 2 Result of clustering
#digits:10; #samples:1797; #features 64

Init	time	Inertia	Homogeneity score	Completeness score	V measure	Adjusted rand index	Adjusted mutual information	Silhouette coefficient
K-means++	0.140s	69485	0.613	0.660	0.636	0.482	0.632	0.171
Random	0.083s	69952	0.545	0.616	0.578	0.415	0.574	0.145
PCA based	0.029s	72686	0.636	0.658	0.647	0.521	0.643	0.155

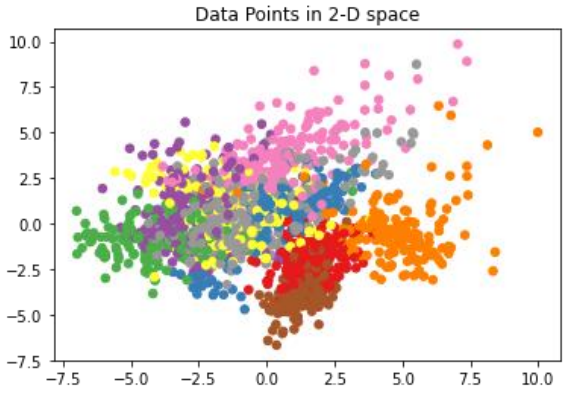

Fig. 6 Plotting of the PCA reduced data in 2-D space

DISCUSSION

We have used the digit dataset for both training and testing purposes. The accuracy on the test set using SVM was found to be 97.03 percent. We also compared the SVM, KNN, and Linear Regression and found that the SVM performed better for pattern classification with 1797 samples, and we anticipate that this score will increase slightly as the number of dataset samples increases. Also, we tested some of the predictions and naked that they were correct for handwritten digit recognition.

There is a learning curve. The cross-validated training and test scores for varied training set sizes are determined by this function. A cross-validation generator divides a dataset into training and test data k times, each time splitting the dataset in half. It will be necessary to train the estimator with different-sized subsets of the training set, and a score for each training subset size and the test set will be computed. Later, the scores will be summed over all k runs for each training subset size, with the highest score being selected. From the first graph of training it is found that Training and cross validation accuracy are nearly closed to each other with the classification accuracy of 97%. Thus we can say SVM found better in classification and recognition of handwritten digits.

For K-means, we have investigated and compared the various initialization procedures, both in terms of runtime and the quality of the results. Because the ground truth is known in this case, we can use a variety of cluster quality metrics to determine how well the cluster labels correspond to the ground truth. And found the PCA based approach is better for clustering.

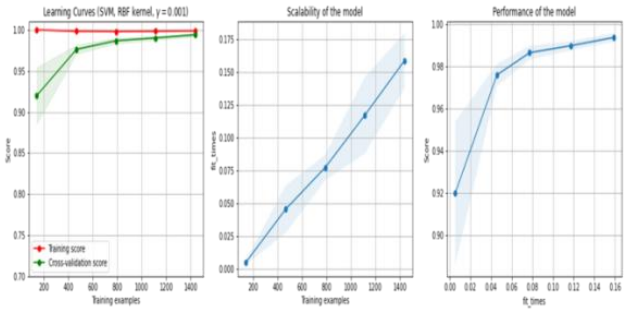

Fig. 7 Learning curve for the SVM classifier

CONCLUSION AND FUTURE SCOPE

The fundamental goal of this project is to develop an Automatic Handwritten Digit Recognition system for document images, as well as to apply classification and clustering techniques to the results of the system. The performance comparison of SVM, KNN, naive bayes, and Logistic Regression methods to determine the best feasible classification algorithm performance of SVM was shown to be superior to the performance of the other three classifiers used in the comparison.

Machine learning techniques like as k-means clustering are a wonderful approach to get started. It is technically reasonable approachable, and it provides a high-level overview of how machine learning and data mining techniques are applied in real-world situations.

REFERENCES

[1]. R. Jayadevan, S. R. Kolhe, P. M. Patil, and U. Pal, "Automatic processing of handwritten bank cheque images: a survey," *Int. J. Doc. Anal. Recognit. IJDAR*, vol. 15, no. 4, pp. 267–296, 2012.
[2]. N. M. Nasrabadi, "Pattern recognition and machine learning," *J. Electron. Imaging*, vol. 16, no. 4, p. 049901, 2007.
[3]. R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 63–84, 2000.
[4]. R. M. Bozinovic and S. N. Srihari, "Off-line cursive script word recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 1, pp. 68–83, 1989.
[5]. K. Marukawa, M. Koga, Y. Shima, and H. Fujisawa, "A High Speed Word Matching Algorithm for Handwritten Chinese Character Recognition.," in *MVA*, 1990, pp. 445–450.
[6]. R. W. Smith, J. F. McNamara, and D. S. Bradburn, "Pattern Classification Techniques Applied to Character Segmentation," *TRW Financ. Syst*. Berkeley CA, 1998.
[7]. F. Lauer, C. Y. Suen, and G. Bloch, "A trainable feature extractor for handwritten digit recognition," *Pattern Recognit.*, vol. 40, no. 6, pp. 1816–1824, 2007.
[8]. G. S. Lopes, D. C. da Silva, A. W. O. Rodrigues, and P. P. Reboucas Filho, "Recognition of handwritten digits using the signature features and Optimum-Path Forest Classifier," *IEEE Lat. Am. Trans.*, vol. 14, no. 5, pp. 2455–2460, 2016.
[9]. L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen, "A methodology for feature selection using multiobjective genetic algorithms for handwritten digit string recognition," Int. J. Pattern Recognit. Artif. Intell., vol. 17, no. 06, pp. 903–929, 2003.
[10]. M. I. Fanany, "Handwriting recognition on form document using convolutional neural network and support vector machines (CNN-SVM)," in *Information and Communication Technology (ICoIC7)*, 2017 5th International Conference on, 2017, pp. 1–6.
[11]. X. X. Niu and C. Y. Suen, "A novel hybrid CNN–SVM classifier for recognizing handwritten digits," *Pattern Recognit.*, vol. 45, no. 4, pp. 1318–1325, 2012.