International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-3, Issue-4, April 2017 Pages 85-87

Classification of Handwritten Digits using Machine Learning Techniques

Prashasti Gupta, Navni Bhatia

***Abstract*— The MNIST dataset (Mixed National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems. [1][2] The database is also widely used for training and testing in the field of machine learning. The MNIST database contains 60,000 training images and 10,000 testing images. [3]**
In this paper, we aim to apply classification techniques to predict labels for records in the MNIST dataset using machine learning. In total, there are 10 labels ranging from 0-9. Classification will be done using Random Forest Classification Algorithm. We also aim to implement Principle Component Analysis to reduce the dimensionality of the data while retaining its variance. To this data, we aim to apply K Nearest Neighbors Classification Algorithm

***Index Terms*— MNIST dataset, Classification techniques, Random Forest Classification Algorithm, Principal Component Analysis, K Nearest Neightbors Classification Algorithm.**

I. INTRODUCTION

Each image in the dataset is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the intensity of lightness or darkness of that pixel, with higher numbers indicating darker. This pixel-value is an integer between 0 and 255, inclusive. The training data set has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image.

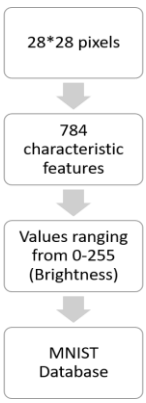

Figure 1: MNIST dataset

A. *Knowledge Discovery*

It can sequentially be explained as:

- Data cleaning: to remove redundant data
- Data integration: to combine multiple data sources
- Data selection: to select analysis relevant data
- Data transformation: to transform or consolidate data into forms appropriate for mining by performing summary or aggregation operations
- Data mining: to extract data patterns
- Pattern evaluation: to identify the patterns representing knowledge
- Knowledge presentation: to visually represent the mined knowledge

B. *Object Classification*

A classifier is an algorithm that takes a set of features that characterize objects and uses them to determine the class of each object. The classic example in astronomy is distinguishing stars from galaxies. For each object, one measures a number of properties (speed, size, compactness, boundary box etc.); the classifier then uses these properties to determine whether each object is a star or a galaxy. There are of two types of classification supervised and unsupervised. [4] In supervised classification, meaning that a human expert both has determined into what classes an object may be categorized and has provided a set of sample objects with known classes. This set of known objects is called the training set because it is used by the classification programs to learn how to classify objects. In unsupervised classification methods in which the induction engine works directly from the data, and there is neither training sets nor pre-determined classes. There are four steps to develop classifiers:

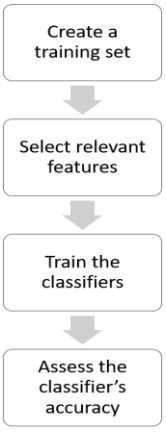

Figure 2: Classification Process

Prashasti Gupta, Information Technology, Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India.
Navni Bhatia, Information Technology, Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India.

C. Machine Learning

Machine learning, is a branch of artificial intelligence, which is concerned with the construction and study of systems that can learn from data. There are three approaches in machine learning model i.e. supervised learning, semi-supervised learning and unsupervised learning. For example, a machine learning system could be trained on images to learn to distinguish between pedestrian and non-pedestrian images. After learning, it can then be used to classify new images into pedestrian and non-pedestrian folders. The machine learning algorithms used in present work are:

- Random Forest Classifier
- Principal Component Analysis
- K Nearest Neighbors Classifier

II. ALGORITHMS

A. Random Forest Classification

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. [5]

B. Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (or sometimes, principal modes of variation). The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations. [6]

This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

C. K Nearest Neighbors Classification

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.[7] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbors.

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

III. METHODOLOGY

The working can be explained as:

Step 1: Data Pre-Processing
The datasets were read in the IPython notebook in .csv format and stored for processing, after importing all relevant modules.

Step 2: Data Splitting
A function was defined to study the nature of each classifier. It trains on a fraction of the data corresponding to a predefined split ratio, and then evaluates on the rest of the data points.

Step 3: Model Building: Random Forest Classifier
A random forest classifier object was instantiated to evaluate the performance as a function of the number of estimators. For each estimator, classification was carried out 10 times and the mean and standard deviation of each observed score was noted.

Step 4: Model Assessment: Random Forest Classifier
The resultant accuracy score obtained for each of the estimator was finally plotted on a logarithmic scale.

Step 5: PCA Application
Principal Component Analysis was applied to the data to reduce the dimensionality. The data was visualised and it was observed that PCA separated the feature space into visible clusters already for 2 components. Intuitively, the number of components was increased.

Step 6: Model Building: K-Nearest Neighbors Classifier
A k-nearest neighbors classifier object was instantiated to perform classification on the PCA output. The classifier was evaluated as a function of its performance on the number of PCA components.

Step 7: Model Assessment: K-Nearest Neighbors Classifier
The resultant accuracy score obtained for each of the number of PCA components was finally plotted on a logarithmic scale.

IV. EXPERIMENTAL RESULTS

A. Random Forest Classification

The accuracy of the classifier as a function of the number of estimators was as shown below.

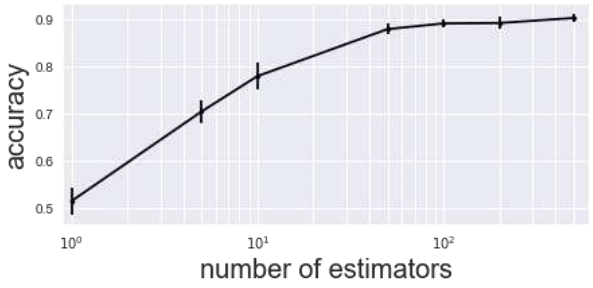

Figure 3: Random forest classification: Accuracy as a function of the number of estimators

B. *Principal Component Analysis*

It was observed that approximately 100 PCA Components were required to capture approximately 90% of the variance present in the data.

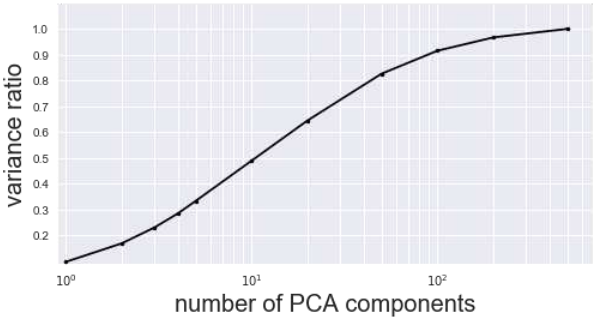

Figure 4: Principal component analysis and variance ratio

C. *K Nearest Neighbors Classification*

The accuracy appeared to saturate at approximately 90% for 20 or more PCA components. The accuracy appeared to drop for much larger numbers due to overfitting.

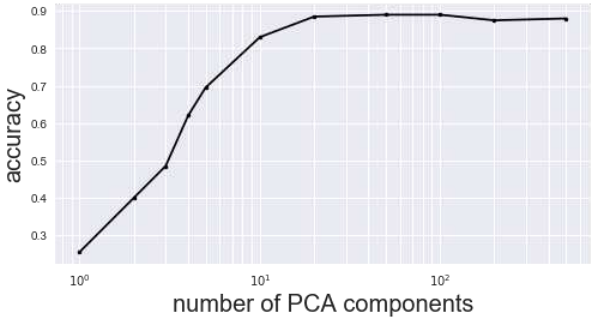

Figure 5: K nearest neighbors classification applied to PCA output

V. Conclusion

In this research, it is proved that PCA can be effectively used in reducing feature dimensionality in datasets. The scope, however, is not only limited to digits. The developed methodology can be made as a unified approach for classification of any character, provided a training set is available. Thus, it can also be used in other language scripts, as well as in handwritten character recognition. Developing commercial systems which can maintain high recognition rates regardless of the irregularities such as the quality of the input documents, and varying font styles is a challenging task for various scripts

Acknowledgment

This research paper was made possible with the help and support of our project guide. Firstly, we would like to thank **Er. Narinder Kaur** for her continuous support and encouragement and adding valuable advice towards the organization and the theme of the paper. We would also like to thank the IT Department of our college for its support. The end product of this research paper might have not been possible without their cooperation.

References

[1] "Support vector machines speed pattern recognition - Vision Systems Design". Vision Systems Design.
[2] Gangaputra, Sachin. "Handwritten digit database"
[3] Kussul, Ernst; Tatiana Baidyk (2004). "Improved method of handwritten digit recognition tested on MNIST database". Image and Vision Computing. 22 (12): 971–981.
[4] Zhang, Tianzhu, Si Liu, Changsheng Xu, and Hanqing Lu. "Mining semantic context information for intelligent video surveillance of traffic scenes." Industrial Informatics, IEEE Transactions on 9, no. 1 (2013): 149-160.
[5] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5
[6] https://en.wikipedia.org/wiki/Principal_component_analysis
[7] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbors nonparametric regression". The American Statistician. 46 (3): 175–185.

Prashasti Gupta was born in New Delhi on 20 September 1995. She is currently a 4th year student at Maharaja Agrasen Institute of Technology, New Delhi pursuing B.Tech. in Information Technology. She was a member of the winning team for the AICTE Smart India Hackathon 2017, a flagship event under the Prime Minister Mr. Narendra Modi's Digital India Movement. She is a gold medal recipient from Delhi Public School RK Puram. She has interned at Paytm and Limetray in the past. Her primary areas of interest include data analytics and machine learning.

Navni Bhatia, born on January 16th, 1996 is a final year B.Tech. student majoring in Information Technology at Maharaja Agrasen Institute Of Technology, Guru Gobind Singh Indraprastha University, New Delhi. The author has also been awarded the Times Scholar award and has been the recepient of various merit based scholarships at Delhi Public School, Faridabad and Maharaja Agrasen Institute Of Technology. She has interned with the MIS department of India's largest commercial enterprise, Indian Oil, and with Cashify, India's leading recommerce platform. She wishes to further apply herself in the data science field.