One shot classification based on elements of Human Learning

Thushar A K
Department of Computer Science and Engineering
Jain University, Bengaluru, Karnataka, India

V.N. Manjunath Aradhya
Department of Computer Applications
JSS Science and Technology University, Mysuru, Karnataka, India

Abstract- The objective of one-shot learning is to mimic the way humans learn in order to make classification or prediction on a wide range of similar but novel problems. The core constraint of this type of task is that the algorithm should decide on the class of a test instance after seeing just one test example. Machine learning approaches including deep learning requires a large training set to perform at same levels of accuracy as humans do for novel tasks. We present a computational approach that models human learning abilities in digit recognition. Digits are modelled using a multinomial probability distribution of strokes and the generative model captures the process of digit formation in a way similar to how humans perceive the digits. We compare the model with traditional approaches of machine learning as well as our own discriminative Bayesian classifier for one shot classification of pen digits' data. The results obtained suggests that emulating the human process of digit generation can suggest ways to train a model with very few examples that is capable of generalising without extensive training.

Keywords – Human Concept learning, Compositionality, Causality, Generative model, One Shot Classification

I. INTRODUCTION

One of the remarkable features of human learning is generalisation from very few examples. Humans can understand new concepts by relating them to already known examples. Even after observing an image of an object for the first time, a child can recognise new objects of similar type. For example, after a child first sees a moving object such as a bicycle, the child parses it into a structure of familiar parts and learns the concept of a "moving object". On seeing a related object like a motor bike, the child identifies it as belonging to the same category of the already learned concept. This ability of human cognition is as a result of intelligence developed through learning and evolution. This suggests that approaching a recognition problem from the learning aspect can give insights into efficient generalization.

In this paper we focus on a generative model based on Compositionality and causality similar to the human learning approach. Many of the objects can be seen as being composed of parts and subparts. This provides a natural way to humans to learn new concepts. Causality provides the inductive priors or constraints during learning process. Handwritten characters present a very good example for learning based on these concepts. Characters can be considered to be built using parts and subparts. They contain a rich internal structure of pen strokes which provides a good apriori reason to explore part-based approach to representation learning. Psychological studies have shown that knowledge about how characters are produced from strokes influences basic perception including classification. Even though characters contain complex internal structure, they can be learned through tractable computational models unlike natural images. Handwritten digits have received major attention in machine learning. Even though traditional Machine Learning algorithms including Deep learning perform well on these datasets, these are far from human level competence. For example, the standard MNIST dataset provides thousands of examples of each class. In wide contrast, humans only need one example to learn a new character. In this work we have chosen to address the problem of one-shot classification using elements of human learning. The classification accuracy with single training example is computed using a generative model based on human learning elements. The model is compared against standard Naïve Bayes, kNN and multi-Layer Neural network based classifiers for one shot classification and the results are promising.

Some of the major challenges faced by Machine learning and deep learning are 1)Data gathering: Collecting sufficient relevant data for each category for machines to learn is laborious. 2)Data labeling: Often, labeling data requires experts or is impossible due to privacy, safety, or ethical issues. 3)Hardware constraints: Due to the large amount of data, as well as large parametric models, expensive hardware (GPUs and TPUs) is required to train them. One-shot learning is an approach to learn a new task using limited supervised data with the help of strong prior knowledge. One of the earliest works that resulted in high accuracy for the one-shot image classification problem dates back to the 2000s by Fei-Fei et al.[1]. They deploy a constellation model which uses the knowledge gained from previously learned categories in learning new categories. Prior information from previously learned categories is represented with a suitable prior probability density function on the parameters of their models. In recent years, researchers have made good progress tackling one-shot learning through different deep learning architectures and optimization algorithms, such as matching networks, model agnostic meta-learning, and memory-augmented neural networks. One-shot learning has a lot of applications in several industries—the medical [15, 16] and manufacturing industries in particular. In medicine, we can use one-shot learning when there is limited data available, for example, when working with rare diseases; whereas in manufacturing, we can reduce man-made errors such as edge case manufacturing defects.

A probabilistic generative model for handwritten characters was proposed by Lake et al[2]. The work addresses classification of characters from 20 different alphabets using single example from each category. The concepts of compositionality, causality and learning to learn are used to reduce the training examples. Concepts are represented as programs and their semantic structure is used to combine subparts and parts using a Bayesian criterion that best explain observed data. The approach reports an error rate as low as 3.5% in a one-shot classification task. Intelligence can be approached in two ways [3]. One is the statistical pattern recognition approach which treats prediction as primary usually in the context of a specific classification, regression or control task. The alternative approach treats models of the world as primary, where learning is the process of model building. The primary focus of Neural networks has been towards pattern recognition. Although humans and neural networks perform equally well on MNIST digit recognition [18, 19] and other large-scale image recognition tasks, there are two important differences in the learning process: people can learn to recognise a new character from a single character, thereby discriminating novel instances drawn by other people as well as similar looking non instances. Moreover, people can do much more than pattern recognition: they learn a concept, that is a model of the class that allows their acquired knowledge to be flexibly applied in new ways.

Generative models for handwritten digits perform equally well compared to discriminative approaches. Hinton et al[4] modelled each digit as a program which simulates the motor activity involved in its generation. The motor programs in addition to usage in classification can generate a large set of different images of the same class thus enlarging the training set available to other methods. A sequential generative model for one shot generalisation was proposed by Danilo[5]. Using the concepts of feedback and attention, the model reflects the principle of analysis by synthesis, in which analysis of information is continually integrated with constructed interpretation of it. Analysis is realised by attentional mechanisms that selectively process and route information from the observed data into the model. Interpretations of the data are obtained by sets of latent variables that are inferred sequentially to evaluate the probability of data.

One shot classification by employing deep learning with several modifications of Siamese and triplet architectures Koch[6] and Hoffer & Ailon[7]. A Siamese network consists of two identical computational subgraphs that share all parameters. The two branches produce embeddings for a pair of inputs and try to verify if the latter come from the same class by learning a distance measure. The triplet architecture is an extension to that but tries to learn embeddings between anchor, positive (same class) and negative (different class) triples of inputs instead. Siamese networks reported an accuracy of 92% for the one-shot classification on the omniglot data set. Triplet Networks have reported to outperform Siamese networks in problems ranging from deep ranking for image retrieval to face recognition. We have chosen to address the problem of classification of Pen Digits dataset[8] using single train digit as motivated by the human ability of learning concepts from few examples.We have adopted a bayesian approach in which handwritten characters are modelled as combinations of parts (strokes) and their transitions.

II. HUMAN CONCEPT LEARNING FOR ONE SHOT HAND WRITTEN DIGIT CLASSIFICATION

A Overview

Our approach is to apply a generative model based on compositionality and causality for the digit recognition problem. Given that digits are produced by pen strokes, a sequence of pixels through which the pen moves is a much more natural and efficient way to capture the essence of handwritten digits, as it relates more closely to the process that generated the data; each data point in such a sequence directly corresponds to a trajectory for movement of the pen. The parts of the digit and probability of occurrence are modelled using a Bayesian criterion. For test images, the model is used to predict the type of character.

For handwritten digits, there are many variations in writing the digit. If we take a probabilistic view of the problem, variations in the digit can be treated as a probability distribution over images. So, for each digit j = (0,1,2,3…9), we have a probability distribution $P_j(x)$ over the images x, which represents our uncertainty or imperfect knowledge about the variations among the images containing digit j. Thus $P_j(x)$ denotes the conditional probability of generating image x given that the digit is j:

$$P(x|y = j) = P_j(x)$$

Now if we are given an image x, to decide the class of digit, we apply the Bayes's rule to compute the posterior probability of each category having observed x:

$$P(y = j|x) = \frac{P(x|y = j).P(y=j)}{P(x)} = \frac{\pi_j P_j(x)}{\sum_{i=0}^{9} \pi_i P_i(x)}$$

Here πi, i= 0,1,2…9 denote the prior probability of observing particular class of digit. Now that we have the posterior distribution, image x is classified as the digit that maximises the posterior probability.

$$h(x) = arg\,max_j P(y = j|x)$$

B Compositionality

Humans easily identify parts and subparts in objects and concepts. This also applies for Handwritten characters and digits. Digits appear to humans as composed of parts and subparts. A part is a long sequence of pen movement without any break. Parts can be separated by brief pauses of pen called subparts. For instance, digit 9 can be considered to be composed of 2 parts, one curve and a line. All handwritten digits can be generated from a library of basic parts and identifying the parts in the digit gives a useful representation for classification.

1.1 Causality

The parts and subparts follow a certain relation during handwriting process. This relation is automatically decoded by the human cognition while recognising characters. Relation between parts is a form of domain knowledge which can be provided to the learning model while recognizing images. This is captured at an abstract level by the generative model for digits. Causal relation between strokes while humans generate digit 2 is illustrated in Figure 1.

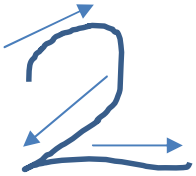

Figure 1. Causal Structure of handwritten digits

1.2 Learning to Learn

Each digit can be considered to be generated by an underlying probability distribution based on parameters specific to the type of the digit. Further, we can consider that there is a top-level generative model which generates each of the sub models based on global parameters of the recognition process. Handwritten digits share similar shape elements and causal structure. Having learned a structure for one type of digit enables a model to use its parameters while recognising another type. For instance, the initial shape element for both the digits 2 and 3 is a curve. On observing this curve humans get a first-hand information that the digit can be 2 or 3 since the underlying generative process is same for all the digits. This way hierarchical priors can be learned and shared across concepts.

II. PROPOSED APPROACH

A Experimental Setup

One shot classification is carried out using the pen digit data set which is one of the widely used hand written digit dataset from the UCI Machine Learning Repository.
Compositionality & causality are applied for recognising a test digit after training on a single example from the training set. We randomly sampled 10 digits from the training set, one from each class of digit as Training set and 3000 images from the test set is used for validation. The basic approach is summarized in Figure 2.

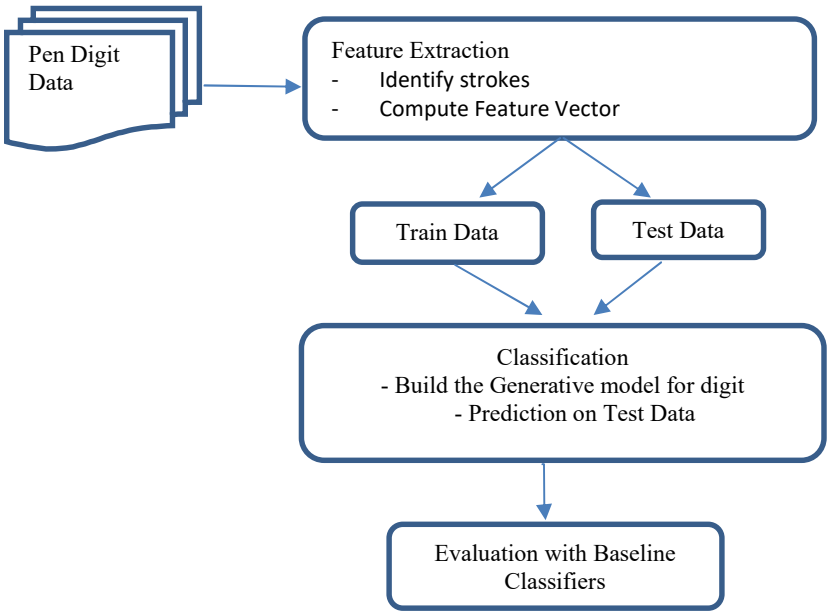

Figure 2. Classification Model for Pen Digits Data

B Generative Model based on Multinomial Stroke Distribution

Pen Digits Dataset [8] contains 10,992 samples of handwritten digits. Each sample consists of 16 features, which represent the sequence of 8 coordinate points sampled along the path of the pen's trajectory when drawing the digits, and each sample is labelled with the corresponding digit in the range of 0 to 9. The value of each feature is an integer in the range of 0 to 100. Therefore, each of the 8 sample points is a bounded in a 100x100 co-ordinate plane. Sample digits from the data set is shown in Figure 3 (a). Each digit of the Pen digits data set is composed of 7 sequential lines or strokes. Each of these strokes fall into one of the 8 categories as depicted in Figure 3(b)

Any image from the dataset can be seen as a combination of seven sequential strokes, each of which is chosen from one of the eight types. In our model, the strokes are considered independent of each other. Hence each pen digit with

seven strokes is equivalent to rolling an eight-sided dice 7 times and hence the probability distribution of obtaining a particular outcome, i.e., image, can be modelled by a multinomial distribution [9].

For a multinomial distribution with n trials and k possible outcomes, e1, e2 …. ek each having count of n1, n2, …. n_k , the probability distribution of the outcome is

$$P = [\, n! \,/\, (\, n1! * n2! * \ldots nk! \,) \,] * p_1^{n_1} \times p_2^{n_2} \times \ldots \times p_k^{n_k}$$

where $n = n1 + n2 + \ldots + nk$ is the number of trials and p1, p2, ...pk are the category probabilities.

For the pen digit data, n=7, number of strokes and k=8, number of categories of strokes. Hence n1,n2…n8 are the number of occurrences of each category of stroke in a digit and p1,p2,…p8 are the corresponding probabilities. Feature vector of a digit is represented as an eight-element array of the category probabilities. This is illustrated for a sample digit from the pen digits data set in Figure 3 (c).

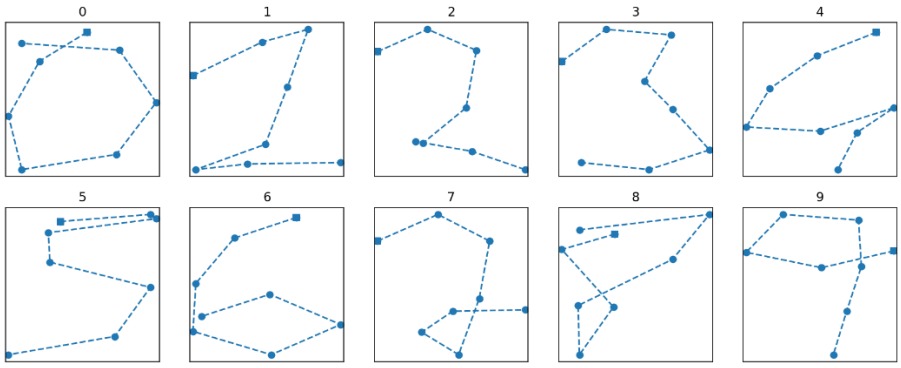

Figure 3(a). Sample digits from Pen Digits Dataset. Square markers indicate starting point

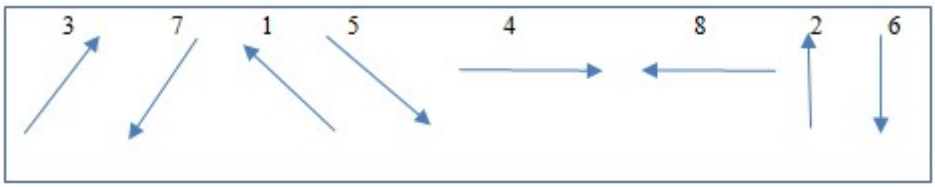

Figure 3(b) Eight Stroke Categories of each of the Pen Digit with numeric codes

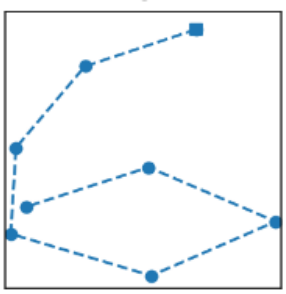

Figure 3(c). Feature Vector for digit 6 from the Pen Digit Data Set

Stroke Sequence: 7 7 7 5 3 1 7
Feature Vector:

Stroke Category	1	2	3	4	5	6	7	8

Probability	1/7	0	1/7	0	1/7	0	4/7	0

C Prediction based on the Generative Model

Prediction of the test digit is achieved by a goodness of fit test [10] of the multinomial probabilities of the test digit against model for the train digit for each of the digit types 0,1,2,…9.

Let H_m denote the null hypothesis that test digit belongs to type m. ϵ {0,1,2,..9}
H_m: $p_{m1}=x_{m1}$, $p_{m2}=x_{m2}$, ……..$p_{mk}=x_{mk}$.

We test which of the hypotheses fits the observed multinomial sample to the maximum and assign that model to the sample. Likelihood Ratio Test(LRT) which is the ratio of the probability of the observed result under the null hypothesis is used for comparing the models.

$$\text{LRT} = 2 \times \sum_{i=1}^{k} m_i \ln \frac{np_i}{m_i}$$

where m_i is the number of objects in category i and p_i is the hypothesized probability of that category.

LRT is computed for each test digit against each of the 10 multinomial models and the model with maximum LRT is chosen as the best match.

$$\text{Digit Class(j)} = \arg\max_{m \text{ in } 0,1,..9} \{\text{LRT}_m\}$$

D Discriminative Model for Prediction

As an alternative approach, we model the likelihood of observing a stroke pattern for a given digit. Using Baye's rule,

$$P(Digit\ Type{=}x \mid Stroke\ data){=}P(Stroke\ data \mid Digit\ Type{=}x) \times P(Digit\ Type{=}x)$$

Two digits are considered to be similar if their first stroke is similar and also has similar stroke transitions. Hence the following holds as per the multiplication rule

$$P(Stroke\ Data \mid Digit\ Type = x) = P(First\ Stroke \mid Digit\ Type = x) \times P(Stroke\ Transitions \mid Digit\ Type = x)$$

where $P(First\ Stroke \mid Digit\ Type = x)$ is the probability of the first stroke in a test digit if it matches with the first stroke of the Training digit x and $P(Stroke\ Transition \mid Digit\ Type = x)$ is the Probability of obtaining the same stroke transitions in Test and Train Data. This is calculated as the elementwise multiplication of the Causal Transition Probability Matrix of the Train and Test data.

A feature vector for an image in this approach is an 8 by 8 matrix of stroke transitions in the sequence. The matrix showing the transition of steps in 8 directions is denoted as *Causal matrix* since it depicts the causal relation between strokes. The step sequence and the causal matrix obtained for the sample digit 6 is shown in Fig 3. Finally, the matrix is normalised by dividing each element by the sum of all entries to convert each entry to a probability of transition.

For classification, we also use the First Stroke Probability of a digit which is the probability that the digit starts with a stroke in one particular direction as observed from the Train data. For calculating this, a first step probability distribution is calculated from the distribution of strokes in the Train data. From this, first step probability with respect to each of the eight steps for each Train digit is estimated. For the Test data, this is taken as the probability of

the first stroke that matches with the Train data. If the first stroke of Test digit does not match with the first stroke of any of the Train digit, a nominal constant first stroke probability is assigned.

0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	1	0	2	0
0	0	0	0	0	0	0	0

Stroke sequence: 7 7 7 5 3 1 7
First step: 7
Step Transitions: 7->7->7->5->3->1->7
No of transitions: 6

0	0	0	0	0	0	1/6	0
0	0	0	0	0	0	0	0
1/6	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1/6	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	1/6	0	2/6	0
0	0	0	0	0	0	0	0

Normalised Causal matrix obtained by dividing frequency of transitions by total number of Transitions

Figure 4. Causal Transition Matrix and Normalised Causal Matrix for Digit 6

For prediction, first Stroke and Transition Matrix are extracted for the Test image by traversing the stroke sequence. A likelihood score is assigned for the test image against each train image. i.e.,

$$\text{Score(j)} = \text{First Stroke Probability (j)} \times \text{Causal Probability of the Stroke Sequence(j)}$$

where j=0,1,2…9 represent the Train images. Test image is classified as the label of the Training image which has the maximum likelihood score.

. .

III. EXPERIMENT AND RESULT

The simulations were conducted using R version 3.5.1 on Intel i5 CPU with 1.6GHz speed and 8GB RAM. We chose R due to the rich repertoire of tools for scientific computing tasks, especially in implementing statistical methodologies

A. Comparison with Naïve Bayes

The first experiment was to test the accuracy with respect to Naïve Bayes model[11] where the probability distribution of each label over an image is considered as product of independent coin flips. Each pixel value is considered independent of all other pixels. It is observed that on 100 simulations of the classification experiment using 10 train digits and 3000 test digits, accuracy with Naïve Bayes is 9.01% and with our approach is 26.5%, which indicates an average improvement of 66% on using the Human concept learning model.

B. Comparison with kNN

Accuracy of the approach was also compared against the k Nearest Neighbour (kNN) classifier [12] using a single Train digit. Euclidean distance between pixel values was used as the distance metric for kNN. When the input data has higher dimensions, kNN suffers from the curse of dimensionality because all vectors were almost equidistant to the target vector. When the kNN was run with single Training digit and 3000 test digits, the accuracy obtained was 9.31%.

C. Comparison with Neural Network

A multilayer perceptron (MLP) is a class of feedforward artificial neural network[13, 17] with at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. It can distinguish data that is not linearly separable. To study the effect of training with single example, a 4-layer Neural network was trained with one example on the Test dataset with pixel-based input. We used an MLP with 2 hidden layers of 5 and 10 nodes and Resilient Backpropagation (RPROP) algorithm for Training[14] . It was found that accuracy obtained is 14.1%. When the MLP was trained with the feature vector of causal probability as used in our generative model, the accuracy obtained is 24.11%, which suggests that proposed approach can improve performance of neural classifiers with less training data. The generative model yields higher accuracy compared to the discriminative model suggesting that it models the real world process of digit generation more accurately. The results obtained using various approaches are summarised in Table 1.

Approach	Accuracy
Naïve Bayes [11]	9.01%
kNN [12]	9.31%
4 layer MLP [13]	14.1%
MLP with causal features	24.11%
Human Concept Learning – Discriminative Model	24.8%
Human Concept Learning – Generative Model	26.5%

Table 1: Comparison of accuracy on One-shot Classification

IV.CONCLUSION

The study suggests that the Human learning approach can give promising results in classification with lesser training examples. Deep Learning algorithms need hundreds to thousands of training examples to achieve the benchmark accuracies. In novel learning tasks such as scene understanding, language acquisition and speech recognition humans still outperform the best machine learning approaches. Here the AI component based on cognition can help in developing more plausible models. In other words, reverse engineering the human understanding to difficult computational problems can help to overcome the data gathering and labelling challenges faced by deep neural networks. The approach can be extended to various areas where compositionality, causality and learning to learn can be applied. In the domain of speech, a word can be split into syllables and each syllable can be further broken down to voice primitives such as phoneme. Learning can progress in a generative way using probabilistic priors representing each new structure. The methodology is also relevant in areas where few shot learning is unavoidable. For example, consider the problem of identity verification in financial instruments using methods such as recognition of handwriting / biometric features such as thumb impression or face. The areas which use these features for pattern recognition demands high level of security. In such applications, getting large number of training examples has practical constraints and recognition needs to be performed with only a few available training data points

REFERENCES

[1] Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4), 594-611

[2] Brendon M Lake, Rusian Salakhutdinov, Joshua B Tanenbaum. Human Level concept learning through probabilistic program induction, Science Magazine, Vol 350, Issue 6266, 1332-1338 (2015).

[3] Brendon M Lake, Tomer D Ullman, Joshua B Tanenbaum, Samuel J Greshman, Building machines that learn and think like People, Behavioral & Brain Sciences, 1-72(2017)

[4] G. E. Hinton and V. Nair. Inferring motor programs from images of handwritten digits. In Advances in Neural Information Processing Systems 19, (2006)

[5] Rezende, Danilo, et al.: One-Shot Generalization in Deep Generative Models. In Proceedings of the 33rd International Conference on Machine Learning, JMLR: W&CP volume 48, (2016).

[6] Koch, Gregory, Zemel, Richard,and Salakhutdinov, Ruslan: Siamese neural networks for one-shot image recognition. ICML deep learning workshop, vol. 2. 2015.
[7] Hoffer, Elad and Ailon, Nir. Deep metric learning using triplet network. CoRR, abs/1412.6622, 2014. URL http: //arxiv.org/abs/1412.6622.
[8] E. Alpaydin, Fevzi. Alimoglu(1998). UCI Machine Learning Repository
[9] [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[10] Evans, Morton; Hastings, Nicholas; Peacock, Brian (2000). *Statistical Distributions* (3rd ed.). New York: Wiley. pp. 134–136. ISBN 0-471-37124-6.
[11] Ostrovski, Vladimir (May 2017). "Testing equivalence of multinomial distributions". Statistics & Probability Letters. 124: 77–82
[12] Ng, Andrew & Jordan, Michael. (2002). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. Adv. Neural Inf. Process. Sys. 2.
[13] Cover TM, Hart PE (1967). Nearest neighbor pattern classification" IEEE Transactions on Information Theory. 13 (1): 21–27
[14] Hastie, Trevor & Tibshirani, Robert & Friedman, Jerome & Franklin, James. (2004). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Math. Intell.. 27. 83-85. 10.1007/BF02985802.
[15] Riedmiller M. and Braun H. (1993) A direct adaptive method for faster backpropagation learning: The RPROP algorithm. Proceedings of the IEEE International Conference on Neural Networks (ICNN), pages 586-591. San Francisco.
[16] Aradhya V.N.M., Mahmud M., Guru, D.S. et al, "One-shot Cluster-Based Approach for the Detection of COVID-19 from Chest X-ray Images", Cognitive Computation (2021). https://doi.org/10.1007/s12559-020-09774-w
[17] V. N. M. Aradhya, M. Mahmud, M. Chowdhury, D. S. Guru, M. S. Kaiser and S. Azad, "Learning Through One Shot: A Phase by Phase Approach for COVID-19 Chest X-ray Classification," *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2021, pp. 241-244, doi: 10.1109/IECBES48179.2021.9398761.
[18] Manjunath Aradhya, V. N and Hemantha Kumar. G, "Principal Component Analysis and Generalized Regression Neural Networks for Efficient Character Recognition," *2008 First International Conference on Emerging Trends in Engineering and Technology*, 2008, pp. 1170-1174, doi: 10.1109/ICETET.2008.214.
[19] V.N. Manjunath Aradhya, G Hemantha Kumar, S Noushath, "Unconstrained Handwritten Digit Recognition: Experimentation on MNIST Database, Advances In Pattern Recognition, pp. 140-143, 2006. https://doi.org/10.1142/9789812772381_0022
[20] V. N. M. Aradhya, G. H. Kumar and S. Noushath, "Robust Unconstrained Handwritten Digit Recognition using Radon Transform," *2007 International Conference on Signal Processing, Communications and Networking*, 2007, pp. 626-629, doi: 10.1109/ICSCN.2007.350685