Data Mining Project

Pengzhong Sun
School of Computer Science and Technology
Harbin Institute of Technology, China
18S103141@stu.hit.edu.cn

ABSTRACT

The report details my data mining experiments. I chose the cifar-10 database and categorized it for this database. I have used three methods: KNN (Nearest Neighbor Classifier), SVM (Support Vector Machine) and ANN (Artificial Neural Network).The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. In this article, I will show the differences and advantages of KNN, SVM, Ann.

KEYWORDS

KNN, SVM, CNN, CIFAR-10 dataset,Min-Max Normalization

ACM Reference Format:
Pengzhong Sun. 2018. Data Mining Project. In *Proceedings of ACM Woodstock conference (WOODSTOCK'97),* Pengzhong Sun (Ed.). ACM, New York, NY, USA, 3 pages. https://doi.org/10.475/123_4

1 PROBLEM

My research direction is biological images, so I chose the cifar-10 database for a relatively simple classification problem. This classification problem is very simple, but it is very suitable for practicing the knowledge learned in the data mining class, because I can use the nearest neighbor classifier, support vector machine, Bayesian belief network and artificial neural network, etc., and this implementation for me Future growth is also crucial. Through this experiment, I will complete the following:

(1) Learn to pre-process data.
(2) Complete the classification of nearest neighbor classifiers, support vector machines, and artificial neural networks.
(3) Generate sufficient knowledge of these three classifiers, to be able to distinguish their similarities and differences.
(4) Establish each model and gradually improve the accuracy.

2 FORMALIZATION

The CIFAR-10 dataset (Canadian Institute For Advanced Research) is a collection of images that are commonly used to train machine learning and computer vision algorithms. It is one of the most widely used datasets for machine learning research.

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. The 10 different classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. There are 6,000 images of each class. In **figure 1**, Here are the classes in the dataset, as well as 10 random images from each. The format of the training set is 50000*32*32*3, of which 50000 is the number of training samples, 32 is the width and height of the picture, and 3 is the RGB three channel. And There is an ASCII file that maps numeric labels in the range 0-9 to meaningful class names. It is merely a list of the 10 class names, one per row. The class name on row i corresponds to numeric label **i**.Therefore, this is a classification task.

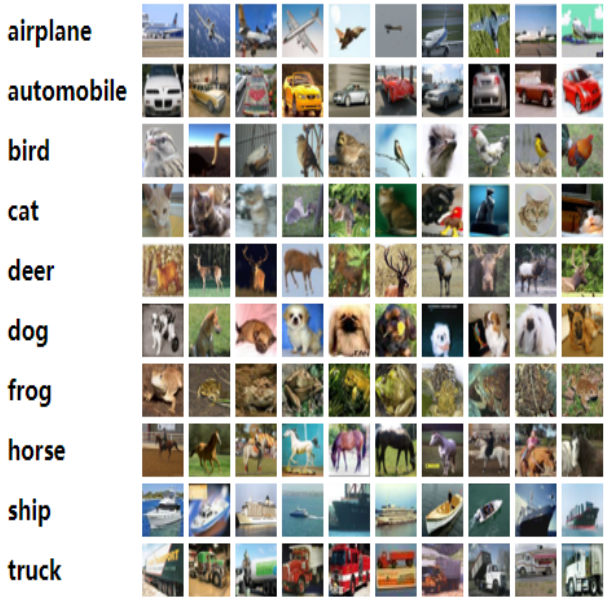

Figure 1: the classes in the dataset, as well as 10 random images from each.

3 ALGORITHMS

I have used a total of three methods: the nearest neighbor classifier, support vector machine, artificial neural network. These three classifiers are very different, and the effects are also good and bad. I will explain the three classifiers in detail.

3.1 Nearest Neighbor Classifier

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors,

WOODSTOCK'97, July 1997, El Paso, Texas USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

```
Got 49 / 500 correct => accuracy: 0.098000
```

Figure 2: This is the knn accuracy.

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 neighbor.k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

My kNN classifier consists of two stages:

- During training, the classifier takes the training data and simply remembers it.
- During testing, kNN classifies every test image by comparing to all training images and transfering the labels of the k most similar training examples.

I used Min-Max Normalization and divided each pixel value by 255. Because if you use the pixel value to calculate the distance directly, data overflow may occur.

I use method ***compute-distances*** to compute the distance matrix between all training and test examples. For example, if there are **Ntr** training examples and **Nte** test examples, this stage should result in a **Nte x Ntr** matrix where each element **(i,j)** is the distance between the i-th test and j-th train example.Then I used ***predict-labels*** for category prediction.

And I used the cross-validation method to choose the optimal k value. Finally I got an accuracy of about 0.1.

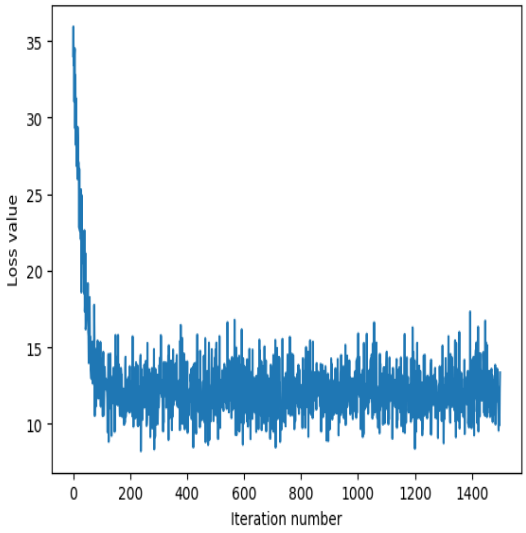

Figure 3: This is the svm loss.

```
training accuracy: 0.100162
validation accuracy: 0.092066
```

Figure 4: This is the svm accuracy.

3.2 Support Vector Machine

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

I Split the data into train, val, and test sets. In addition we will create a small development set as a subset of the training data.I can use this for development so our code runs faster.And then I used Min-Max Normalization and divided each pixel value by 255. I had vectorized and efficient expressions for the loss, the gradient and my gradient matches the numerical gradient. In the end, our accuracy rate is slightly better than the nearest neighbor method.

3.3 Artificial Neural Network

Artificial neural networks (ANN) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. The neural network itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as "cat" or "no cat" and using the results to identify cats in other images. They do this without any prior knowledge about cats, for example, that they have fur, tails, whiskers and cat-like faces. Instead, they automatically generate identifying characteristics from the learning material that they process.

An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it.

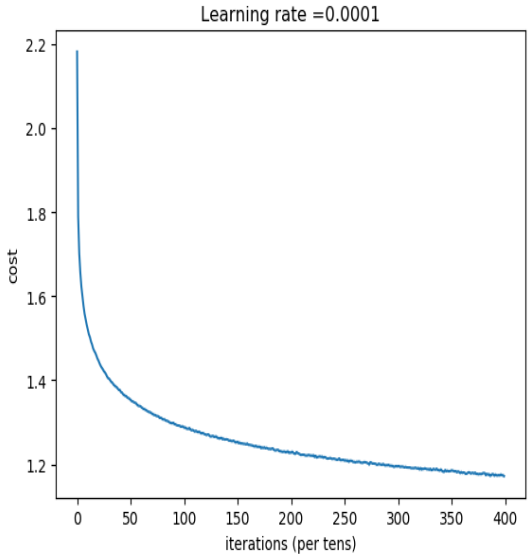

Figure 5: This is the Ann's cost.

```
Train Accuracy: 0.57398
Test Accuracy: 0.4268
Your algorithm predicts: y = 5
```

Figure 6: This is the Ann accuracy.

I used Min-Max Normalization and divided each pixel value by 255.And I used the TensorFlow framework to complete the ANN network. The network consists of the following components: ***LINEAR -> RELU -> LINEAR -> RELU -> LINEAR -> SOFTMAX***.

Finally, I got an accuracy of 0.57 on the training set and an accuracy of 0.42 on the test set, as shown in Figure 6. Finally, I selected a picture of the dog to identify it, and the recognition was successful. Figure 7 is the picture of the dog. The lowest algorithm predicts at the bottom of Figure 6 is the predicted category.The prediction is right.

4 EVALUATION

As mentioned above, the artificial neural network achieved the best accuracy, around 0.5. The effects of the nearest neighbor classifier and support vector machine are only about 0.1. Therefore, the nearest neighbor classifier and support vector machine are not particularly suitable for the classification of pictures. Because the nearest neighbor classifier only computes the "nearest" neighbor samples, the number of samples in a class is large, or such samples are not close to the target sample, or such samples are very close to the target sample. In any case, the quantity does not affect the results of the operation. The support vector machine can only solve the small sample machine learning problem and lack data sensitivity.

Figure 7: This is the dog to be predicted.

In summary, artificial neural networks are more suitable for image classification tasks among the three. And now the popular convolutional neural network can produce nearly 100

ACKNOWLEDGMENTS

I would like to thank Professor Zou Zhaonian for training in recent months. The teacher is conscientious and never misses classes. I am deeply touched. I will keep this thank you and move to complete the next academic year.