Journal of Artificial Intelligence Practice (2016) 1: 30-35
Clausius Scientific Press, Canada

Research and Application of Machine Learning on Geographic Information System

Zhenjiang Dong[1,a,] Peng Yang[2,b,] Zhicheng Ma[3,c], Yongbiao Chen[4,d]

[1]Shanghai Jiao Tong University, Shanghai, China
[2]Gansu State Grid Information & Telecommunication Co., Ltd.629 East Xijin Road, Qilihe, Lanzhou, Gansu Province, China

[3] Gansu State Grid Information & Telecommunication Co., Ltd.629 East Xijin Road, Qilihe, Lanzhou, Gansu Province, China

[4]Shanghai Jiao Tong University, Shanghai, China
[a]dongzhenjiangvip@163.com, [b]yangpeng@gs.sgcc.com.cn, [c]mazc@gs.sgcc.com.cn,
[d]chenyongbiao0319@sjtu.edu.cn

Keywords: Machine Learning; NoSQL Database; Neural Network; Recommender System; Collaborative Filtering.

Abstract: In the big data era, an information system that is able to flexibly scale out, store mass data and quickly response to concurrent requests is particularly important. Despite the mature mining technologies on structured data, the utilization of unstructured data is still inadequate which results in the waste of data sources. Under this circumstance, this paper adopts machine-learning technologies to build a salable information system by analyzing Geographical landform data.

1. Introduction

The advent of "Internet Plus" concept provides us with an original thought to apply the Internet technologies in the research of DANXIA information system. "DANXIA landform" is a special landform, which appeared in a specific historic period, which might help to reveal the features of crustal evolution. Currently, the research work on DANXIA landform mainly includes the collection of geological data, geomorphology data, physical geographic environment data and the analysis of the data collected. When the volumes of data reach a certain scale, the management of data will undergo severe efficiency declines with the increase of file volumes [1-4]. Meanwhile, the overheads of maintenance retrieval will be tremendously augmented.

In this paper, we propose a new way to apply machine-learning technologies into the research to model the landform data, and establish an information system which is easy to store information in and easy to be extended.

2. Background Information

2.1 Device Virtualization.

A neural network model designed for solving visual pattern recognition is proposed and named as Neocognition. A typical neural network model is comprised of large number of neurons.

Neural network consists of three layers. The first one is to take in the input signals. The second layer is comprised of three neurons, each of them accepting all the signals from the first layer and outputting a signal. The third layer consists of one neuron, accepting all the signals from the second layer and outputting the final result. The outputs of neurons on the second layer are demonstrated as the formulas (1), (2), (3). The final result is showed in the formula (4).

$$\alpha_3^{(2)} = g\left(\Theta_{31}^{(1)}\chi_1 + \Theta_{32}^{(1)}\chi_2 + \Theta_{33}^{(1)}\chi_3\right) \tag{1}$$

$$\alpha_2^{(2)} = g\left(\Theta_{21}^{(1)}\chi_1 + \Theta_{22}^{(1)}\chi_2 + \Theta_{23}^{(1)}\chi_3\right) \tag{2}$$

$$\alpha_3^{(2)} = g\left(\Theta_{31}^{(1)}\chi_1 + \Theta_{32}^{(1)}\chi_2 + \Theta_{33}^{(1)}\chi_3\right) \tag{3}$$

$$\alpha_1^{(3)} = g\left(\Theta_{11}^{(2)}\chi_1 + \Theta_{12}^{(2)}\chi_2 + \Theta_{13}^{(2)}\chi_3\right) \tag{4}$$

$\alpha_i^{(L)}$——The output signal of the *i*-th neuron on the L-th level.

$\Theta_{\mathrm{ji}}^{(L)}$—— The output signal of the *i*-th neuron acts` as the corresponding influence factor of the input signal of *J-th* neuron on the *L+1-th* layer.

2.2 Cost Function.

Cost function is another crucial concept in neural network training[5], measuring the gap between the specific solution and the best solution towards a specific problem. Basically, the neural network training process is to enable the cost function to get the minimum result. The corresponding cost function of neural network is as formula (5).

$$\mathrm{J}(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[\begin{matrix} y^{(i)}\log\left(h_\Theta\left(\chi^{(i)}\right)\right) + \\ \left(1 - y^{(i)}\right)\log\left(1 - h_\Theta\left(\chi^{(i)}\right)\right)\end{matrix}\right] \tag{5}$$

$\mathrm{J}(\Theta)$——The cost function which takes all the parameters of neural network as variables

m ——The number of data；

$y^{(i)}$ ——The result of the *i-th* set of training data；

h_Θ ——The corresponding compound function of the input and output of neural network；

$\chi^{(i)}$ ——The input of the *i-th* set of training data has been constructed based on ExpressEther .

2.3 System Design

DANXIA landform information system is comprised of many subsystems. The first problem to solve when we are developing the system is landform data modeling. DANXIA landform data embraces the following features: large volume of data, few connections among data, low consistency requirements and low security requirements. For these reasons, NoSQL database is adopted for this system.

The second is recognizing the DANXIA scenic spot that the given picture belongs to. The system design adopts convolution neural network to categorize the pictures, thus recognizing different places according to the given pictures.

The last part is recommending appropriate scenic spots to users. For this purpose, an optimized collaborating filtering algorithm, enabling the algorithm to consider other influence factors except the scenic spot when breaking down the scenic spot factors.

2.4 Implementation

This part will focus on the implementation details of data persistence layer, and elaborate the implementation details of the automatic recognition and scenic spots recommendation functions. Guest OS configures the remote PCI devices using remote mounting strategy, transparent to user applications. We have also demonstrated the workflow of how the whole system works to show the feasibility of our design. Now, the implementation of the system prototype is working in progress. And some necessary experiments will be done to test the performance and practicality of our design in near feature.

2.5 Information System Architecture

Node.js is applied as the server. Node.js is a platform, which is created to enable fast and extendable web applications.

Express is the web applications structure applied in the Node.js. Its functions include: powerful request routing service, reliable Internet environment, flexible control over the status of web application by configuring environmental parameters.

AngularJS is a MVC frame in client side. It offers two-way data binding, automatically synchronizing data between View and Model. It also provides dependency injection function, cutting down the coupling degree of codes.

MongoDB is an open-source NoSQL Document-oriented database. In a Document-oriented database, all the data is stored as independent cells in the form of documents, making data storage more flexible.

2.6 View spot Image Recognition System

2.6.1 Data acquisition and Pre-processing

View Spot Image Recognition System captures all the view spot images from MongoDB database to a set of original images. All the pictures in the set are tagged with different DANXIA landforms it belongs to. Meanwhile, the original data will be scaled up or down to fit a specified resolution requirement. Then the RGB color pictures are greyed to scale down the dimensionality of the input data without threatening the information volume of input data.

2.6.2 Model architecture and relevant parameters

View spot Image Recognition System recognizes images automatically by training the Convolutional Neural Network which can categorize the scenic images. The figure below is the model architecture that will be utilized.

2.6.3 Optimized View Spot recommendation System

Owing to the fact that traditional collaborative filtering algorithm cannot represent users' fancy grade towards a specific scenic spot, an optimized collaborative filtering algorithm is proposed.

3. Evaluation

The experiments are conducted in two parts:

3.1 Automatic view spot recognition test

4000 pieces of data from the DANXIA landform database are selected for this experiment. 20

images from every piece of data will be chosen to comprise the original data set of 80000 images for this test. 48000 images are randomly chosen as the training data set from the original set. Then another 16000 pictures are selected as cross validation data set. The left 16000 images are presented as the test data set.

First, we would conduct pretreatments over all the images selected, converting them into formats that could be used by the system. For Hyper Parameter that cannot be trained, we will try different values and pick up the optimal parameter by utilizing the cost function in the cross validation data set. Once the Hyper Parameter is confirmed, convolutional neural network will be trained by using the training data set. Then the test data set will be utilized to test the neural network will have already been trained before. By analyzing the difference between the test results and real results, we can calculate the prediction error rate, which is showed in the following graphs .

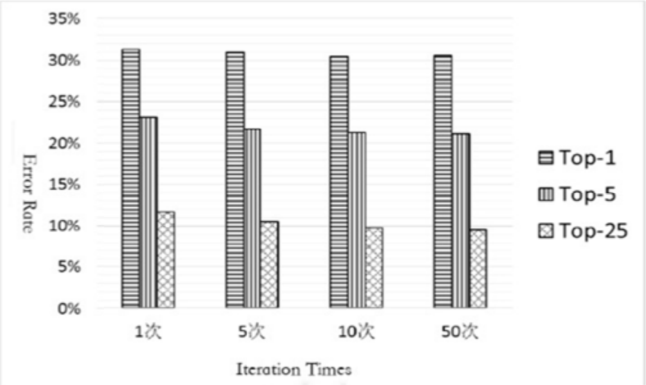

Fig.1 The error rate histogram of view spot image recognition system

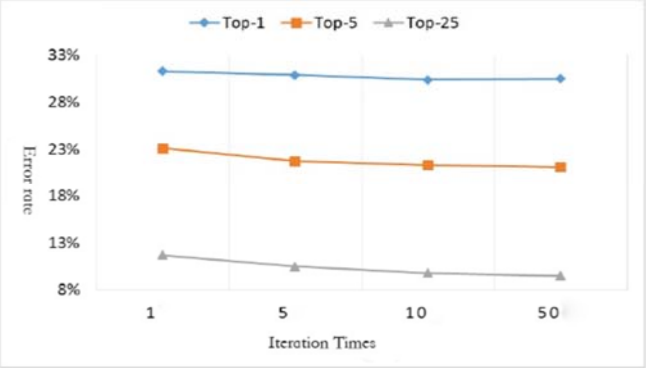

Fig.2 The error rate line chart of view spot image recognition system

Fig 2 is the corresponding line chart. Top-1 is the most possible spot it predicts. If it is not the real place, then it is wrong; Top-5 means it will provide 5 spots with highest possibility, it any one these 5 places conform to the real place. Then the prediction is correct. If not, it is wrong. Top-25 is similar to Yop-5, with the scale of possible places being 25.

From the above results, we can see that

1: The error rate will descend with the number of iteration times climbing under the same calculation method.

2: Under the same calculation method, the error rate will fall dramatically when the iterations times increase from 1 to 5.

3: The more predictions given, the lower error rate.

4: Top-1 embraces a very high error rate no matter how many times it has been iterated.

3.2 View spot recommendation test

This experiment will be conducted by simulating the real situations. 16 million pieces of data will be randomly generated conforming to a certain form. 1million pieces of them will be selected as training and cross validation data set. Those, which are not selected, will serve as test data.

Owing to the fact that every user can only give one score to a certain scenic spot according to the traditional collaborative filtering algorithm, special handling of training data set is required. Through averaging multiple scores of the same scenic spot graded by the same user in the training data set and cross validation data set, it produces the training data set and cross validation data set for the traditional Collaborative filtering algorithm. However, the training data set will not undergo special handlings, thus, making it possible to compare the results of traditional collaborative filtering algorithm with the corresponding optimized one.

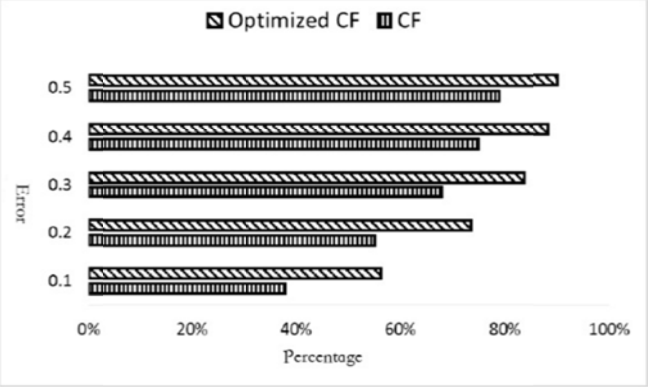

Fig.3 The relation of cost function on training data and cross validation data

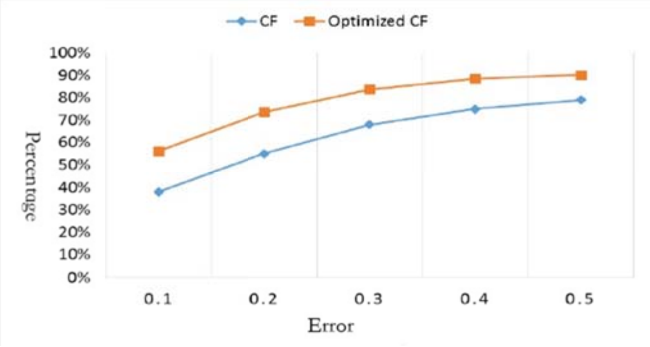

Fig 4 is the comparison graph between traditional and optimized collaborative filtering.

CF stands for the optimized collaborative filtering algorithm. From Fig 4, we can see that the performance of the optimized collaborative filtering algorithm is superior to the traditional one. From the line chart, it is clear that within 0.1, optimized algorithm enjoys a higher accuracy rate over 50% while the traditional one being under 40%. When the error scale keeps increasing, the accuracy rates of both keep rising with the climbing speed slowing down. Meanwhile, it can be seen that the gap between the two different algorithms reach its highest point within 0.1 and keeps shrinking with the augment of error scale. The accuracy of optimized algorithm makes it an ideal one for DANXIA view spot recommendation system.

References

[1] Kuota Chan. On the Subdivisions of the Red Beds of South-Eastern China[J]. Bulletin of the Geological Society of China, 1938, 18:315-316.

[2] DeCandia G, Hastorun D, Jampani M, et al. Dynamo: amazon's highly available key-value store[C]. ACM SIGOPS Operating Systems Review, 2007, 41(6): 205-220.

[3] Chang F, Dean J, Ghemawat S, et al. Bigtable: A distributed storage system for structured data[C]. ACM Transactions on Computer Systems (TOCS), 2008, 26(2): 4.

[4] Mitchell TM. The discipline of machine learning[M]. Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006.

[5] Werbos P. Beyond regression: New tools for prediction and analysis in the behavioral sciences[M]. 1974.

[6] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[C]. Biological cybernetics, 1980, 36(4): 193-202.