SAS® GLOBAL FORUM 2020 #SASGF

Paper 4891-2020

How SAS® Viya® Provides a Way to Deliver Analytics to Everyone: Image Classification Examples

Pavel Rogatch, Analytium BY

ABSTRACT

We are living in an era when machine learning (ML) helps to solve complex tasks for enterprises and improve productivity. ML techniques such as image recognition are continually improving, but they are of no use if an enterprise cannot easily implement them. The scarcity of data science experts and the high costs associated with hiring them can be a significant barrier, especially for small and midsize enterprises. In this presentation, we argue that modern tools like SAS® Viya® provide an opportunity even for small enterprises to build complex models with high accuracy. For this purpose, we use MNIST and Fashion-MNIST data sets that have been used to test image classification methods. We show that using SAS Viya with default settings and without coding, you can get a model that misclassifies about 3.5−6.0% and 11.1−11.9% of images for the first and second data sets, respectively. Simple tuning with a graphical user interface enables you to achieve misclassification rates as low as 2.3−2.7% and 9.8−10.7%. In many cases, such rates are satisfactory for enterprises, and they can benefit from machine learning techniques even if for any reason they lack data scientists. SAS Viya makes the process of building complex models simple and time-efficient and opens opportunities for all kinds of people within enterprises to perform data analysis. This presentation is intended for a technical audience with a basic knowledge of SAS® as well as for non-experts in data science who want to start using ML.

INTRODUCTION

Humans have the ability to collect and analyze information. This ability is a great asset on its own, but it can be boosted by using invented mathematical algorithms and computational power of modern devices, especially as volumes of collected data are rapidly increasing. Data analysis helps to solve problems and make decisions based on the evidence.

There are many benefits of analytics for enterprises. It enables enterprises to discover trends and patterns, make predictions, determine best actions, etc. which results in increased effectiveness. Here are just a few examples of how enterprises can benefit from analytics:

- get sales forecasts,
- determine marketing strategies based on the customer behavioral patterns,
- detect fraudulent financial activities,
- perform market segmentation,
- recommend the right products to customers,
- recognize the potential loss of customers (churn analysis),
- process images and recognize what is depicted on them,
- etc.

One of the methods of data analysis is machine learning, its algorithms are continually improving and becoming more complex for better performance. However, what matters for many enterprises is not how sophisticated algorithms are but rather how easy they can be used to solve real business problems.

TYPICAL PROBLEMS WITH IMPLEMENTING ANALYTICS AND MACHINE LEARNING AND SOLUTIONS

Typical problems with implementing analytics and machine learning in enterprises include:

- The scarcity of data scientists.
- High costs that are associated with hiring data scientists. Many enterprises cannot afford it.
- Much time is required to create a model, so just a few models are deployed.
- A small percentage of employees have access to.

Possible solutions:

- Augmented analytics. The world of data science is shifting towards augmented analytics – automation of analytical tasks. First, it reduces the time spent by data scientists on models and frees them for more complex tasks. Consequently, it speeds up model deployment and helps to achieve greater results. Second, augmented analytics opens doors to analytics for everyone.
- Training of business users. With little training, many business users who work with different data will be able to fulfill the role of so-called "citizen data scientists" who do not need to be experts in data science or programming, but who can build machine learning models using appropriate software. In fact, some predict that soon citizen data scientists will be producing more advanced analysis than professional data scientists (Sallam, R., Howson, C. and Idoine, C., 2018).

So, there is hope for those who want to use machine learning but lack resources. For example, SAS® Viya® delivers capabilities that enable different users within organizations to easily access, explore, transform, analyze, and govern their data. Self-service data preparation toolset in SAS® Viya® is interactive and does not require coding skills. With it, users have access to all the data, the power to manipulate the data, and no need to ask for help. The advanced analytics capabilities of SAS® Viya® allow people who are not data scientists to get value from machine learning as illustrated by the examples that follow.

MODEL BUILDING IS MADE SIMPLE IN SAS® VIYA® AS ILLUSTRATED BY IMAGE CLASSIFICATION EXAMPLES

We use MNIST and Fashion-MNIST data sets that have been used to test image classification methods.

EXAMPLE 1: MNIST DATA SET

Example 1 uses MNIST data set – a large data set of handwritten digits (60000 images). Sample images from this dataset are shown in Figure 1.

The goal was to train an algorithm that would be used to recognize handwritten digits. This is a classic image recognition problem.

We have tried to model a situation when a person without deep data science knowledge is solving this problem using the advanced analytics capabilities of SAS® Viya®. The data set was split into training (70%), validation (20%) and test (10%) data sets.

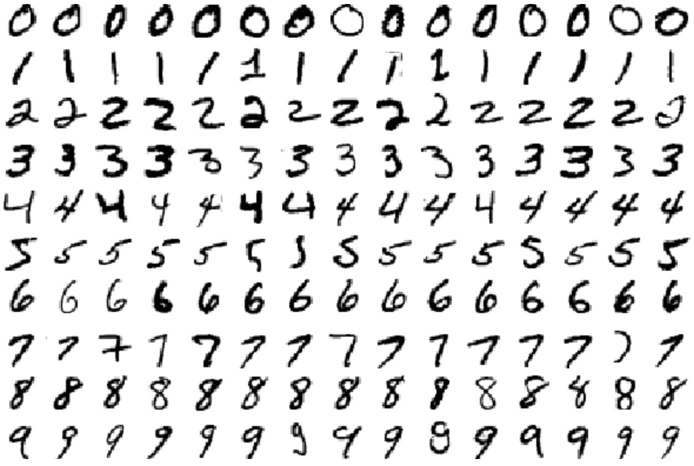

Figure 1. Sample images from MNIST data set

We have built several Neural Network and Gradient Boosting models using default parameters, Autotuning option and simple tuning using Graphical User Interface. "Simple tuning" in this case means the following: for the Neural Network model we manually set the number of neurons per hidden layer and Hidden layer activation function, for the Gradient Boosting model we manually set the number of trees and maximum depth.

Misclassification rates for the selected models are shown in Table 1.

Model	Default Parameters	Autotuning	Simple Tuning
Neural Network	0.053	0.048	0.023
Gradient Boosting	0.035	0.027	0.030

Table 1. Misclassification rates for the Example 1

These results are not too bad compared to the results received by data science experts. For example, the Wikipedia article on MNIST database (Wikipedia, the Free Encyclopedia, 2020) lists several ML methods used on this database with error rates ranging from 0.0017 to 0.076. Note that many of those models used data preprocessing.

EXAMPLE 2: FASHION-MNIST DATA SET

Example 2 uses Fashion-MNIST data set – a large data set of fashion grayscale images (60000 images) depicting things like T-shirt/top, Trouser/pants, Pullover shirt, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot. Sample images are shown in Figure 2.

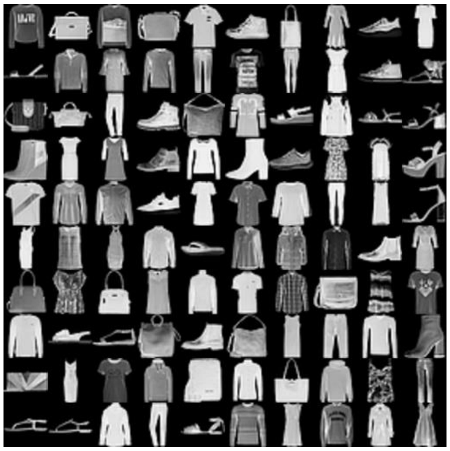

Figure 2. Sample images from Fashion-MNIST data set

The goal was to train an algorithm that would be used to recognize different items.

Again, we have tried to model a situation when a person without deep data science knowledge is solving this problem using the advanced analytics capabilities of SAS® Viya®. The data set set was split into training (70%), validation (20%) and test (10%) data sets.

We have built several Neural Network and Gradient Boosting models using default parameters, Autotuning option and simple tuning using Graphical User Interface "Simple tuning" in this case means the following: for the Neural Network model we manually set the number of neurons per hidden layer and Hidden layer activation function, for the Gradient Boosting model we manually set the number of trees and maximum depth.

Misclassification rates for selected models are shown in Table 2.

Model	Default Parameters	Autotuning	Simple Tuning
Neural Network	0.119	0.150	0.107
Gradient Boosting	0.111	0.106	0.098

Table 2. Misclassification rates for the Example 2

These are also not very bad results for this kind of task. Note that they are received without applying deep data science knowledge and without any data preprocessing.

CONCLUSION

About 35 years ago when spreadsheet programs were released and the number of personal computers increased, it has brought a fundamental change in how different business users gained control of their data and got access to the computing power and analytics that they needed for their work. But since then the demand for data analysis has grown tremendously and many organizations have outgrown Excel with some of its flaws.

Today augmented analytics may bring another fundamental change in the world of analytics making advanced analytics available for every business. Fancy phrases such as "Neural Networks", "Gradient Boosting" or "Machine Learning" should not scare business users anymore because they can easily run machine learning algorithms by themselves and get good results. Yes, data scientists can get better results using their knowledge, but organizations that do not have data scientists for any reason can still benefit from machine learning if they invest in educating their employees and use modern tools like SAS® Viya®. Analytics is becoming more available to everyone.

REFERENCES

"MNIST database." *Wikipedia, the Free Encyclopedia*. Wikipedia, the Free Encyclopedia. Accessed February 27, 2020. https://en.wikipedia.org/wiki/MNIST_database.

Sallam, R., Howson, C. and Idoine, C. "Augmented Analytics Is the Future of Data and Analytics." Gartner, Inc. Accessed February 27, 2020. https://emtemp.gcom.cloud/ngw/eventassets/common/research-notes/documents/gartner-research-augmented-analytics-2019.pdf.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Pavel Rogatch
Analytium BY
E-mail: pavel.rogatch@analytium.co.uk