Solving the MNIST Dataset Using Python Scikit Learn Library

13514061 - Robert Sebastian Herlim [1]
School of Electrical Engineering and Informatics
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1] 13514061@std.stei.itb.ac.id

***Abstract*—Today data is the main component of an existing information system. It is easy to store and retrieve the data, but it would not be valuable if we just let them as it be. To make the data suitable to be consumed by the reader, it should be processed into different form, such as graph, tables, chart, and model. Data mining is a part of studies in data processing so that data itself can also be useful in later use. In this paper will briefly explain the definition, process, and problem types of data mining, and also describe the result of the author's experiment of mining the MNIST dataset using the Scikit-Learn library in Python programming language.**

Keywords— learning, data, python, scikit-learn, artificial intelligence.

I. Introduction

Nowadays, information technologies are rapidly evolving due to the increasing number of the technologies user. This factor has driven a new basic need for the human, which is the need to be connected anytime and anywhere. Everything seems like connected to anything without limits by the existence of the cheap and *ultra*-mobile internet connection provided by the Internet Service Provider (ISP).

So, today anything can be related to the internet. One of the most impactful product is the currently trending the *Internet of Things* (IoT). The key of IoT products are actually simple. We can simply make any stuff in our daily life usage to connect the Internet to make our life easier. We can control the uses of these appliances by using our fingertip (and also our *smartphones*) anytime and anywhere.

With the state of everything is connected to the Internet, it should be balanced with the capability of data processing. Today, data is the most valuable component in a system. By using the current user data, we can make our system adapting to the user preferences to increase the *user experience* (UX) gained from the user while using our application.

Storing the data is not a main problem these days. Main memory and secondary storage are very affordable in this year. The main problem of data processing is simply asking the question: "How should we process the user data so that the data can be a valuable information to the developer team?". This question leads to the data mining studies in many aspects.

II. Fundamental Concept

A. Data Mining

According to illinois.edu (p. 1), data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies, and significant structures, from large amounts of data stored in databases, data warehouses, or other information repositories. Due to the wide availability of huge amounts of data in electronic forms, and the imminent need for turning such data into useful information and knowledge for broad applications including market analysis, business management, and decision support, data mining has attracted a great deal of attention in information industry in recent years. [1]

B. Types of Problem in Data Mining

According to wideskills.com, there are several types of problem in Data Mining, which are described by the following figure.

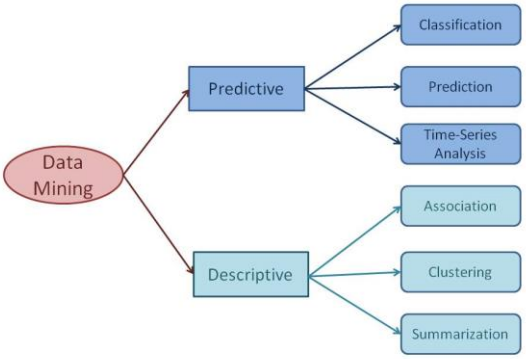

Figure 1 Types of Problem in Data Mining *(source: http://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks)*

1) Classification

In the classification problem, we create a model by using the inductive learning approach from the classified data (fact) provided. Using the built model, we have to classify some other unclassified data as accurately as possible.

2) Prediction

The prediction problem is pretty similar with the classification problem. The difference is we have to categorize some data into some determined class in the classification problem, but for the prediction problem we have to predict the possible values of the future data. Simply, the goal of the classification problem is classifying some data into discrete class, while in the other hand the goal of the prediction problem is classifying some data into more continuous label.

3) Time-series Analysis

In the time-series analysis problem, we are given a sequence of events where each event is associated with its preceding events. The goal of this type of problem is to analyze the sequence to extract useful patterns, trends, rules, and statistics.

4) Association

In the association problem, we have to find the association connection, and relationships between set of items.

5) Clustering

In the clustering problem, we have to group some data into some *clusters* (groupings) so that every item in the same cluster is similar to one another. The similarity between data can be determined using the attributes of the data itself.

6) Summarization

In the summarization problem, we have to generalize the data provided into a smaller set that gives aggregated information of the data using different abstraction levels and different angles. Simply, the goal of summarization problem is process the large uninformative data into smaller informative data for the reader using the aggregation functions such as sum, count, average, and etc. [2]

C. Data Mining Process

According to Fayyad & Patetsky-Shapiro & Smith (1996), the process of data mining is described by the following figure. [3]

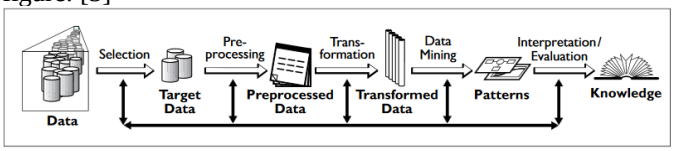

Figure 2 Process of Data Mining *(source: Process for Extracting Useful Knowledge from Volumes of Data, 1996)*

III. Problem Specification

For this paper writing purpose, the author tried to solve a common Data Mining problem "MNIST dataset" (https://www.kaggle.com/c/digit-recognizer). In this problem, we are given a training dataset consists of 42,000 instances of handwritten number images. Each instance in the training dataset is actually a set of 784 (28x28) pixels image. We are also given the correct label for each of the corresponding instance. The training dataset is given in the Comma-separated Values (CSV) format.

Using the model built from the training dataset, we are about to classify 28,000 unclassified instances from the test dataset.

IV. Method and Procedure

To solve the MNIST dataset problem, the author tried to learn using Python programming language by using the Scikit-Learn library. Scikit-Learn library is a common library consists of machine learning tools for data mining and data analysis.

To acquire the model, the author used the following steps:

1) Data selection

The first thing we should do before pre-processing the data is finding *noise*. Data noises are defined as inconsistency and incorrectness of data, including the missing-valued data. We must omit these data noises so that we can build an unbiased model based from the valid data.

2) Data transformation

In the MNIST dataset, each instance consists of 784 pixels. Each pixel represented using value between 0-255. So, if we calculate number of computation using 42,000 instances, the program must compute 42,000 * 784 = 32,928,000. It is indeed a huge number of computation.

The idea of data transformation is to simplify each instance so that we can obtain more generalized model from the data. In the other side, the more simplified instance also increase the performance of model-building.

For the MNIST dataset, the idea of data transformation we can perform is by reducing the size of each image by half. The original size of each image is 28x28 pixels. Here, we can reduce the size into 14x14 pixels by calculating each 4 (2x2) adjacent cells into a single cell. For each cell in the reduced image, we calculate the average of the values from the original picture. The example of this transformation is shown in the following figure.

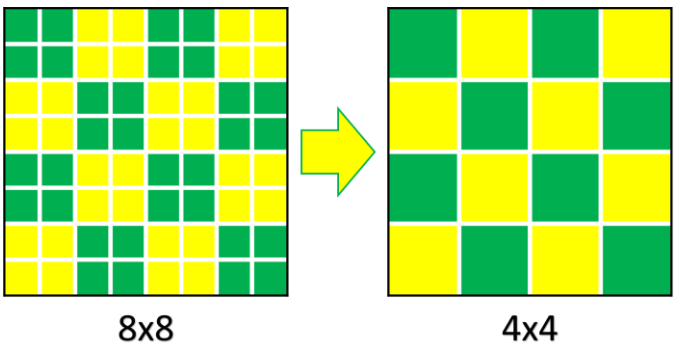

Figure 3 Image Transformation

3) Data pre-processing

After we have the 14x14 reduced image instances of the training dataset, we must process each instance so that each instance could be informative and meaningful for the model

we build. In this current state, each cell of the 14x14 images has the average value of the original 28x28 image.

In the pre-processing step for this dataset, the author make another transformation for each of the reduced images we have. The key of this transformation is to minimize the variation of each cell value so that every cell can only contain 1/0 value (on/off). The author did this transformation by picking a threshold value so that every cell with value larger than the threshold can only be written as a "*on*" value. Otherwise, it is written as "*off*" value.

4) Model building

The last step is start building the model. This step is requiring pretty much time from the whole data mining process. In this step, we should choose the most suitable algorithm and modeling technique is the best for the problem. In this time, the author picked the Random Forest classifier to classify the handwriting images. Some other variant of techniques such as Naïve Bayes, Neural Network, Decision Tree, etc. We also have to specify the best parameters such as *number of trees*, *minimum split*, etc.

To perform the model building step, we need to execute the following code:

```
clf = RandomForestClassifier(n_estimators=250)
clf.fit(X_train_valid, y_train_valid)
```

V. Results

After performing those steps, in this state we have a classifier ready to classify the unclassified test dataset. To classify the test dataset, we need to execute the following code:

```
clf_probs = clf.predict_proba(X_test)
```

The prediction must be written into CSV and submitted to Kaggle using the following format:

```
ImageId,Label
1,3
2,7
3,8
```

And from the technique we used just like the procedure described above, we will get the following submission result:

Figure 4 Submission Result

From the 28,000 unclassified instances, we have correctly classified 94.729% of the test dataset (about 26,500 instances correctly classified). This result briefly describes the performance measurement of our model which is pretty good. But in data mining studies, we shall not grateful for that result. We should strive for better accuracy result by tuning the model parameters, or trying different approach.

VI. Conclusion

In conclusion, the data mining process is indeed challenging and interesting because it is typically new for the author. There are so many different techniques we can try, but to choose the best approach for a problem requiring lots of experiences building data models. Learning about data mining also teaches us not to easily give up and grateful for the result that we have, but we have to strive more for better result.

In this era of information, the ability to process data into something informative is indispensable for everyone. Data is practically stored and used in every system as the main component of the system. The author highly recommends the reader to study more about data processing. For beginners and newcomers in data processing studies should try data processing using *spreadsheets* application such as Microsoft Excel.

Acknowledgment

First of all, the author would like to thank God for His grace and blessing so that the author could finish this writing well but also in time. The author also would like to thank the lecturers of IF3280 Socio-informatika dan Profesionalisme, that are Dr. Ir. Rinaldi Munir, MT., Dr. Eng. Ayu Purwarianti, ST., MT., and Dessi Puji Lestari ST, M. Eng., Ph.D for their guidance and patience while lecturing the course for the whole semester. The last but not least, the author would like to thank all of the author's friends that cannot be listed one by one for being a supportive friend for the author.

References

[1] http://scikit-learn.org/stable/documentation.html

[2] http://web.engr.illinois.edu/~hanj/pdf/ency99.pdf

[3] http://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks

[4] Fayyad, U..M., Piatetsky-Shapiro, G., Smyth, P., 1996, The KDD Process for Extracting Useful Knowledge from Volumes of Data

Declaration

I hereby certify that this paper is a copyright on my own, neither a copy nor a translation of any other paper, and not an act of plagiarism.

Bandung, May 5th 2017

Robert Sebastian Herlim