Bachelor's Thesis (UAS)

Degree Program: Information Technology

Specialization: Internet Technology

2013

Yu Yang

A study of pattern recognition of Iris flower based on Machine Learning

TURUN AMMATTIKORKEAKOULU
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR'S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Degree Program: Information Technology | Specialization: Internet Technology

2013 | 43

Instructor: Patric Granholm

Yu Yang

A study of pattern recognition of Iris flower based on Machine Learning

As we all know from the nature, most of creatures have the ability to recognize the objects in order to identify food or danger. Human beings can also recognize the types and application of objects. An interesting phenomenon could be that machines could recognize objects just like us someday in the future. This thesis mainly focuses on machine learning in pattern recognition applications.

Machine learning is the core of Artificial Intelligence (AI) and pattern recognition is also an important branch of AI. In this thesis, the conception of machine learning and machine learning algorithms are introduced. Moreover, a typical and simple machine learning algorithm called K-means is introduced. A case study about Iris classification is introduced to show how the K-means works in pattern recognition.

The aim of the case study is to design and implement a system of pattern recognition for the Iris flower based on Machine Learning. This project shows the workflow of pattern recognition and how to use machine learning approach to achieve this goal. The data set was collected from an open source website of machine learning. The programming language used in this project was Python.

Keywords:

Pattern Recognition, Machine Learning, K-means algorithm, Python,dataset,SciKit-learn

CONTENTS

List of Figures

1. Introduction

Machine learning, as a powerful approach to achieve Artificial Intelligence, has been widely used in pattern recognition, a very basic skill for humans but a challenge for machines. Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

1.1 Background

Since the computer was invented, it has begun to affect our daily life. It improves the quality of our lives, it makes our life more convenient and more efficient. A fascinating idea is to let a computer think and learn as a human. Basically, machine learning is to let a computer develop learning skills by itself with given knowledge. Pattern recognition can be treated like computer being able to recognize different species of objects. Therefore, machine learning has close connection with pattern recognition.

In this project, the object is the Iris flower. The data set of Iris contains three different classes: Setosa, Versicolour, and Virginica. The designed recognition system will distinguish these three different classes of Iris.

1.2 Objectives

After the project has been settled, the computer should have the ability to aggregate three different classifications of Iris flower to three categories. The whole workflow of machine learning should work smoothly. The users do not need to tell the computer which class the Iris belongs to, the computer can recognize them all by itself.

The final purpose of this project is to let everyone who read this thesis have a basic understanding of machine learning. Even through someone never touched this field, they can realize that the machine learning algorithm will

become more popular and useful in the future. Moreover, the case study of Iris recognition will show how to implement machine learning by using Scikit-learn software.

1.3 Collecting data set

The data set contains three classes of 50 instances each, where each class refers to a type of iris plant. Each class is linearly separable from the other two classes. The attribute information will include sepal length, sepal width, and petal length and petal width. All of them have the same unit, *cm*.

1.4 Using K-means algorithm to achieve clustering

K-means algorithm was used for clustering Iris classes in this project. There are many different kinds of machine learning algorithms applied in different fields. Choosing a proper algorithm is essential for each machine learning project. For pattern recognition, K-means is a classic clustering algorithm. In this project, K-means algorithm can be implemented with the Python programming language.

1.5 Evaluating result

Evaluation will be the final part of this project. For each scientific project, the final result should be tested and evaluated if that is acceptable. The result will be automatically shown in the end of the program execution. For every machine learning algorithm, exceptions will always exist. In order to find the best result, result analyzing is necessary.

2. Literature review

2.1 Basic introduction to machine learning

Learning is a very important feature of Artificial Intelligence. Many scientists tried to explain and give a proper definition for learning. However, learning is not that easy to cover with few simple sentences. Many computer scientists, sociologists, logicians and other scientists discussed about this for a long time. Some scientists think learning is an adaptive skill so that the system can perform the similar task better in the next time(Simon 1987). Others claim that learning is a process of collecting knowledge(Feigenbaum 1977). Even though there is no proper definition for learning skill, we still need to give a definition for machine learning. In general, machine learning aims to find out how the computer algorithms can be improved automatically through experience(Mitchell 1997).

Machine learning has an important position in the field of Artificial Intelligence. At the beginning of development of Artificial Intelligence(AI), the AI system does not have a thorough learning ability so the whole system is not perfect. For instance, a computer cannot do self-adjustment when it faces problems. Moreover, the computer cannot automatically collect and discover new knowledge. The inference of the program needs more induction than deduction. Therefore, computer only can figure out already existing truths. It does not have the ability to discover a new logical theory, rules and so on.

2.11 Fundamental structure of machine learning system

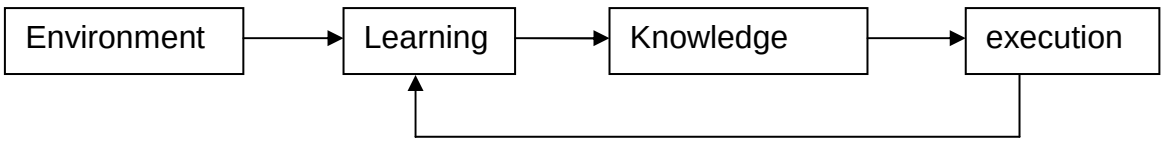

Figure 1. Learning system structure

Figure 1 shows the basic work structure of machine learning. The structure of machine learning system consists of four main parts: Environment, Learning, Knowledge base and Execute.

The environment represents a combination of information from external information source. That would include any information from persons or references materials and so on. It is the learning source for the whole machine learning system. The environment is responsible for transferring data to the system. The quality of the data is very important. In the reality, the data can be complex so it will be difficult for computer to process. In addition, the data can be incomplete, therefore the illation from the learning system is unauthentic.

Learning is the procedure of transferring the information from the environment to knowledge. The environment will give the computer external information, and then the computer will go through all the information by using analysis, comprehensive induction and analogy to process this information to knowledge. At last, all the knowledge would be imported to the knowledge base.

The knowledge base can be treated as the brain of the whole machine learning system. Different kinds of form and content of knowledge can have different influence on the designing of a machine learning system. Knowledge representation modes are eigenvector, First-order logic statements, production rule, and semantic system. Every mode has its own advantages and disadvantages. Therefore, when users want to design a machine leaning system, a good knowledge representation mode is very important for the whole system.

A proper knowledge representation mode should satisfy four basic requirements:

1. Strong expression
2. Easy theorization
3. Easy to modify the knowledge base
4. Easy to expand the knowledge represenation

Moreover, a machine learning system cannot create new knowledge from nothing. It always needs original knowledge to understand the information from environment. Then the computer can use this information to learn new

knowledge step by step. In conclusion, learning process in the whole system is a process of expansion and perfection of the knowledge base.

Execution is the core of the whole machine learning system. Each part of the system aims to make a progress for the execution part. On the other hand, execution also has a connection to each part, especially the learning process. The purpose of a learning process is to make the execution perfect. At last, the complexity, feedback and transparency of execution also has an influence on the learning process.

Complexity

The complexity of knowledge is different depending on the different learning tasks. Some tasks are quite easy, so the system does not need too much information. If the tasks are quite difficult, the system will need more information to learn.

Feedback

After the execution, the execution system can evaluate the leaning task, and then give feedback information to the learning process. The learning process will try to decide whether to collect information from environment to modify or improve the knowledge in knowledge base or not based on the feedback.

Transparency

From the result of execution part, users can easily see the structure of the knowledge base and give the evaluation for it.

2.1.2 The applications of Machine Learning.

Machine learning as a very likely approach to achieve human-computer integration and can be applied in many computer fields. Machine learning is not a typical method as it contains many different computer algorithms. Different

algorithms aim to solve different machine learning tasks. At last, all the algorithms can help the computer to act more like a human.

Machine learning is already applied in many fields, for instance, pattern recognition, Artificial Intelligence, computer vision, data mining, text categorization and so on. Machine learning gives a new way to develop the intelligence of the machines. It also becomes an easier way to help people to analyse data from huge data sets.

2.2 The description of Machine Learning forms

A learning method is a complicated topic which has many different kinds of forms. Everyone has different methods to study, so does the machine. We can categorize various machine learning systems by different conditions. In general, we can separate learning problems in two main categories: supervised learning and unsupervised learning.

2.2.1 Supervised learning

Supervised learning is a commonly used machine learning algorithm which appears in many different fields of computer science. In the supervised learning method, the computer can establish a learning model based on the training data set. According to this learning model, a computer can use the algorithm to predict or analyze new information. By using special algorithms, a computer can find the best result and reduce the error rate all by itself. Supervised learning is mainly used for two different patterns: classification and regression.

In supervised learning, when a developer gives the computer some samples, each sample is always attached with some classification information. The computer will analyze these samples to get learning experiences so that the error rate would be reduced when a classifier does recognitions for each patterns.

Each classifier has a different machine learning algorithm. For instance, a neutral network algorithm and a decision tree learning algorithm suit to two different classifiers. They have their own advantages and disadvantages so that they can accomplish different learning objectives.

2.2.2 Unsupervised learning

Unsupervised learning is also used for classification of original data.

The classifier in the unsupervised learning method aims to find the classification information for unlabeled samples. The objective of unsupervised learning is to let the computer learn it by itself. We do not teach the computer how to do it. The computer is supposed to do analyzing from the given samples.

In unsupervised learning, the computer is not able to find the best result to take and also the computer does not know if the result is correct or not. When the computer receives the original data, it can find the potential regulation within the information automatically and then the computer will adopt this regulation to the new case. That makes the difference between supervised learning and unsupervised learning.

In some cases, this method is more powerful than supervised learning. That is because there is no need to do the classification for samples in advance. Sometimes, our classification method may not be the best one. On the other hand, a computer may find out the best method after it learns it from samples again and again.

2.3 Machine Learning in pattern recognition

As mentioned above, the method of machine learning can also be used in pattern recognition. In fact, pattern recognition really needs machine learning to achieve its objective.

Both supervised learning and unsupervised learning are useful for pattern recognition, for example, in this thesis, K-means clustering algorithm in unsupervised learning. The K-means clustering algorithm is always used for image segmentation. The image segmentation is so important for image pattern recognition. Because of the technology of image segmentation, it is easier to do the image analyzing so that it will achieve much better results for image pattern recognition.

Moreover, the technology of machine learning has been used in almost every field in pattern recognition. For example, image pattern recognition, voice recognition, fingerprint recognition, character recognition and so on. They all need machine learning algorithms to select features from the objects and to do the analyzing.

2.3.1 Basic introduction to pattern recognition

Pattern Recognition is a fundamental human intelligence. In our daily life, we always do ‘pattern recognition’, for instance, we recognize faces and images. Basically, pattern recognition refers to analyzing information and identifying for any kind of forms of visual or phenomenon information. Pattern recognition can describe, recognize, classify and explain the objects or the visual information.

As machine learning, pattern recognition, can be treated as two different classification methods: supervised classification and unsupervised classification. They are quite similar to supervised learning and unsupervised learning. As supervised classification needs a teacher that gives the category of samples, the unsupervised classification is doing it the other way around.

Pattern recognition is related to statistics, psychology, linguistics, computer science, biology and so on. It plays an important role in Artificial Intelligence and image processing.

2.3.2 Machine learning algorithm in pattern recognition.

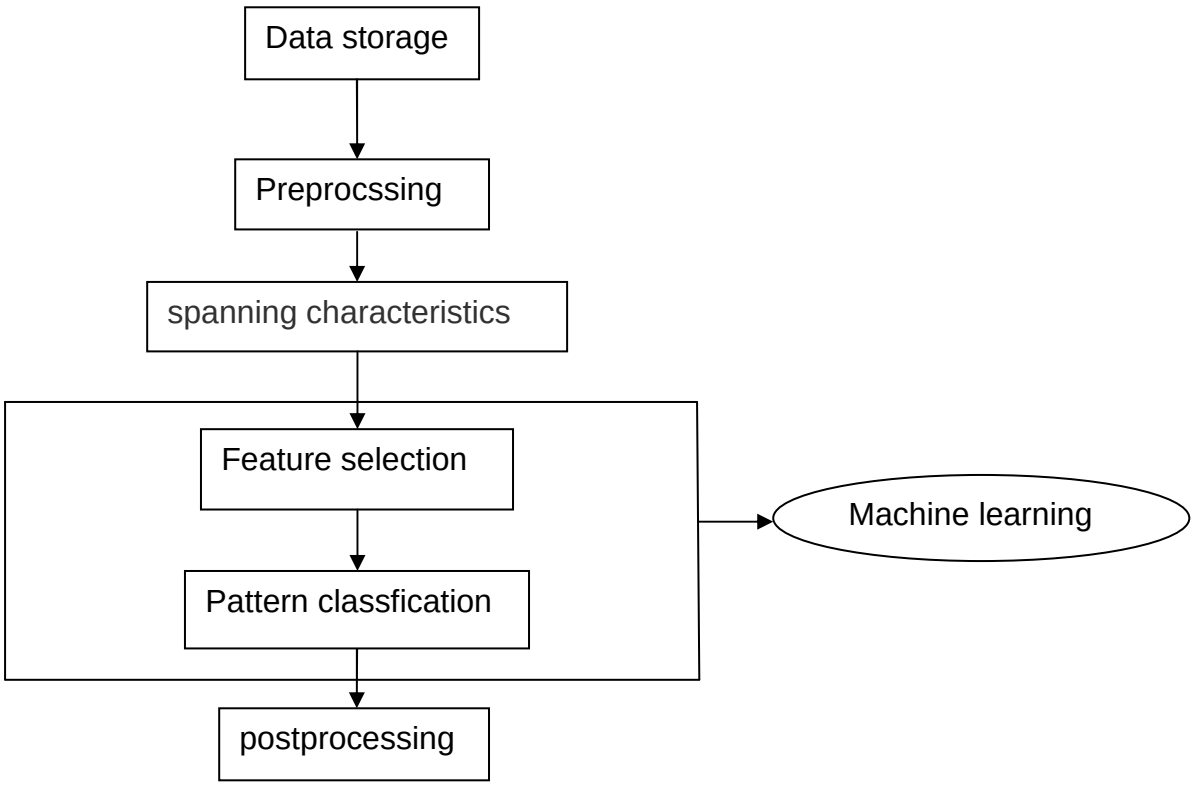

Figure 2. Pattern recognition framework

From Figure 2, we can see that feature selection and pattern classification are the main parts of the whole pattern recognition system.

The core of machine learning is mainly about searching. For different types of patterns, machine learning needs a suitable method to find the proper feature from all information. In order to achieve this, many scientists create many kinds of machine learning algorithms. These algorithms are made for feature selection and pattern classification. For instance, the genetic algorithm, the neural network algorithm, SVM, the K-nearest neighbor algorithm, all support different types of learning objectives. The process of feature selection is so important that it can have a great effect on the result of pattern recognition. Sometimes, the property of objects is so different. If the selection algorithm is not chosen right, the result of the pattern classification will be different or bad. Bad algorithms could cause plenty of information redundancy. Some useful data are not used. On the contrary, some unuseful data may be used for feature

selection. In this case, in the processing of pattern classification, the computer will classify objects in an inappropriate way with many errors. The end, the result would not be acceptable.

3. K-means clustering

As mentioned earlier in this thesis, machine learning consists of many kinds of learning algorithms for different learning methods. In this thesis, the classification information is assumed to be unlabeled. In this case, the best choice in unsupervised learning is the K-means clustering algorithm.

3.1 Introduction to clustering

The K-means clustering algorithm is one of the most popular clustering algorithms in the world. Clustering aims to classify data from the whole data space. The difference between each data object in the same class is similar. However, the difference between each data objects in different classes is large. Clustering belongs to the unsupervised learning method and it can automatically sort data sets.

Basically, the result of clustering algorithm is to find the same classification of different data in the whole data sets. For example, the data set contains monkey, lion, banana, apple, four different data units. After clustering, these four data will be divided into two main sections. One section includes monkey and lion representing the class of animals. The other section includes apple and banana, this section representing the class of fruits.

A clustering algorithm groups all the same kind of data into one single class. The computer will recognize the specific features of all data so that it can separate data to the proper classes.

3.2 K-means algorithm

The K-means algorithm is based on the distance from each data to the initial cluster centers. The distance is the evaluation standard for the similarity of the data. This means that if the distance between two objects is small, then the similar level is high.

In the K-means algorithm, an initial set of k as the clustering point is chosen. The computer will find all the data which is close to the initial set of k. After that,

by using the method of iteration, the computer will update the value of k to get the new cluster for the rest of data. Then, the computer will retrieve the best result after running it again and again.

The formula of K-means algorithm is as following:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad (1)$$

All the results will be related to the initial set of k, therefore the random value of k is very important to the whole algorithm system.

Suppose we have a data set $\{X_1, X_2, \ldots\ldots X_n\}$ consisting of N observations of a random variable x.

1) We choose a number of k cluster centroids from N observations as the initial clustering centroid, $X_1(1), X_2(1), \ldots\ldots X_K(1)$, here (1) means the number of times of iteration.

2) Based on the means of each observation, we calculate the distance between each observation to the initial clustering point. According to the rule of minimum distance, we distribute every sample to the one of the k cluster centroids.

$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2. \quad (2)$$

3) We calculate the vector value of each cluster centroid. $X_j(k+1)$, j = 1,2,…K. Then we calculate and take the value of sample mean vector as the new cluster centroid.

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}. \quad (3)$$

4) We loop step 2 and 3 until every cluster does not change anymore.

3.2.1 K-means algorithm workflow

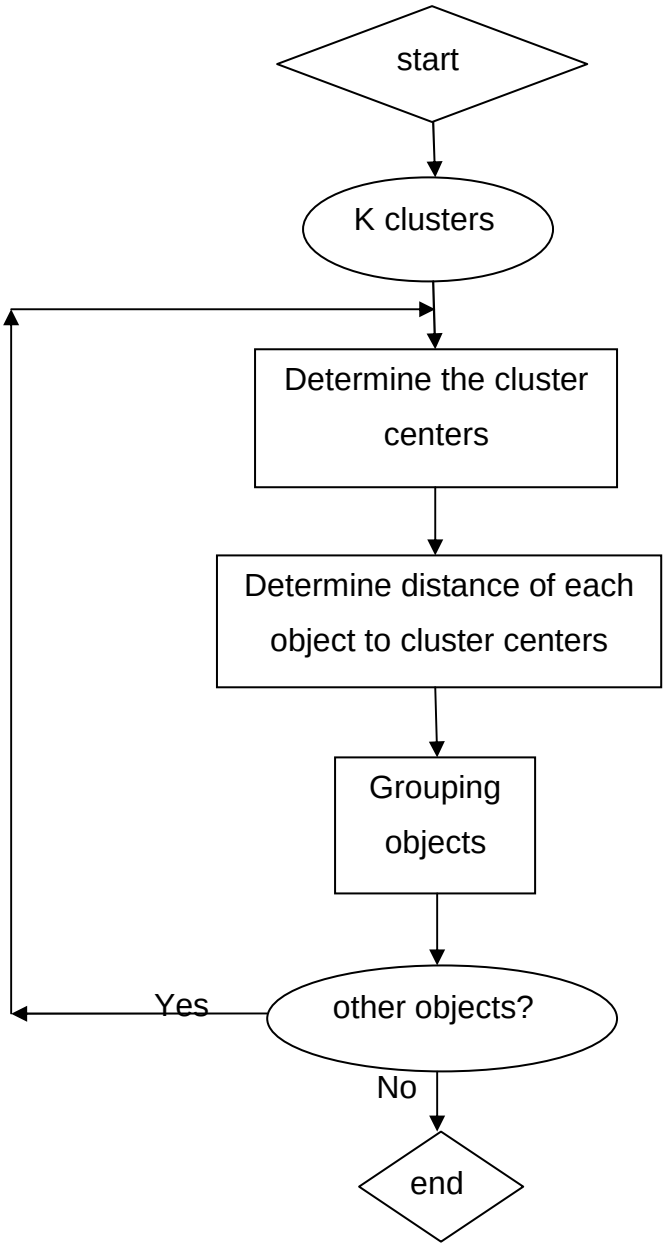

Figure 3. K-means algorithm workflow

Figure 3 shows the workflow of K-means algorithm. In the very first beginning, the system will choose a number of k clusters from a number of N observations. In the next, for the rest of the objects, the system will distributes these objects to the closest clusters based on the mini distance between objects to the cluster center. Moreover, it will calculate the means of all objects in the same cluster to get the new cluster center. These two steps are repeated until the formula (3)

convergences. In general, the equation (3) is based on mean square deviation theory.

The following tables show a sample of workflow of K-means. The dataset contains 30 samples and the number of clusters is 3.

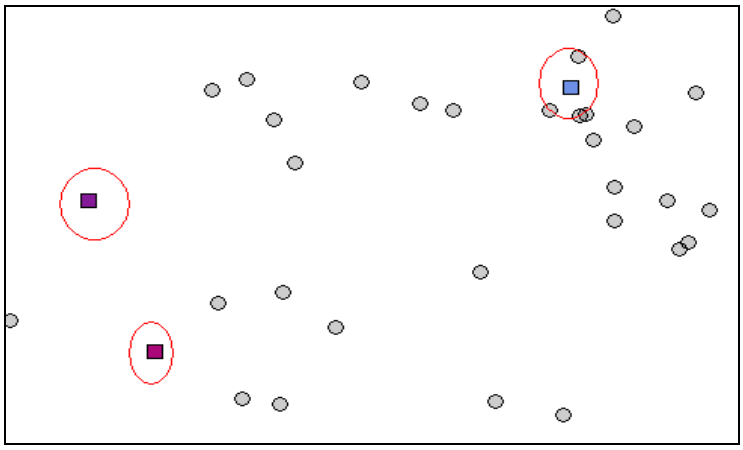

Figure 4. K-means clustering step 1

Now the system generates three cluster points with randomly. There are three different colours: purple(Top left), blue(Top right), pink(Bottom left). These three colours stand for three different clusters.

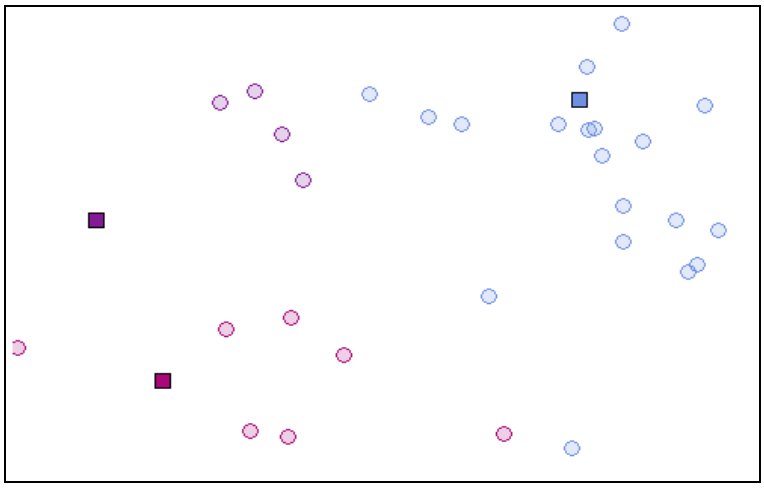

Figure 5. K-means clustering step 2

With the initial the point of k, then the system should calculate the distance of each object to the cluster centers. The new blank box indicates the new cluster center.

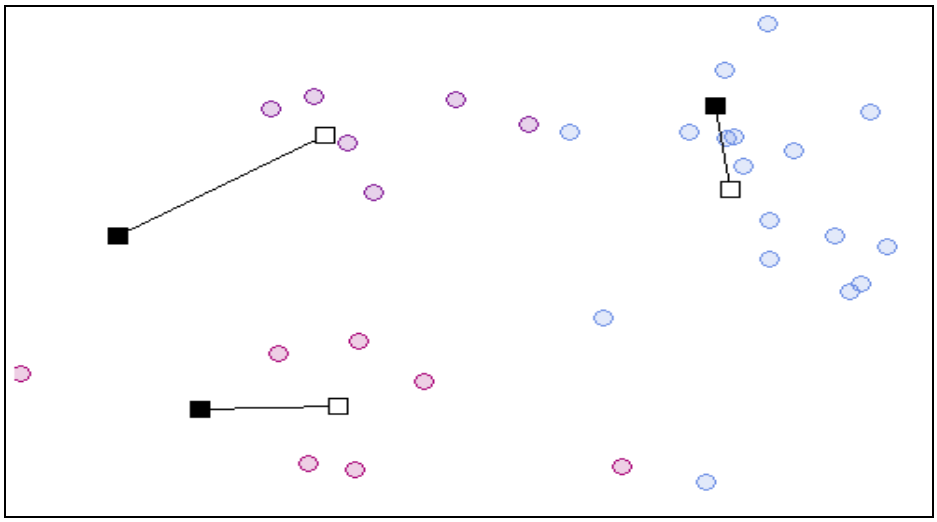

Figure 6. K-means clustering step 3

If there are still some objects missing, then the system will continue to find the new centroid for each cluster until all the samples are grouped. The system will loop equation (2) and (3) until the k cluster centroids will not move any longer. Therefore, In Figure 7, the k cluster centroids move to a new place and the calculation is continued.

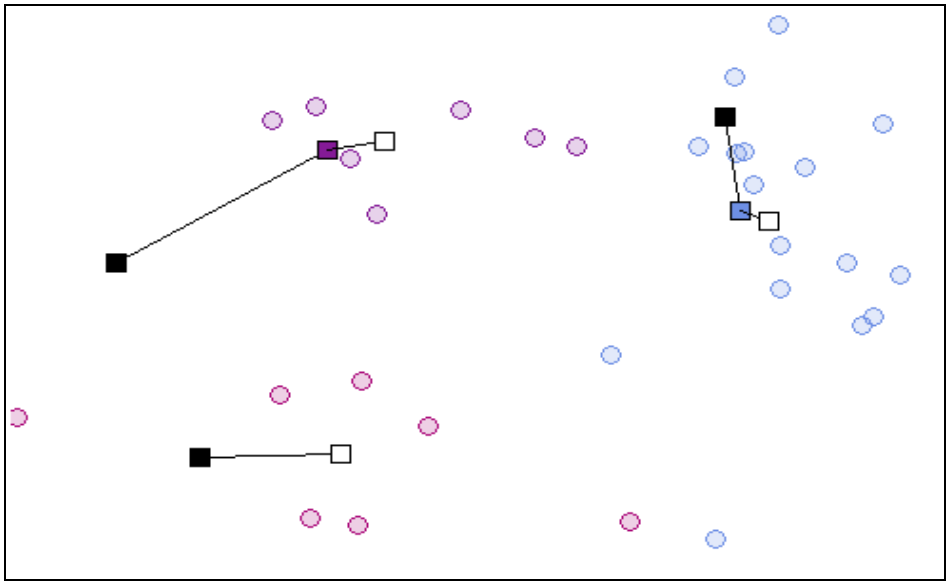

Figure 7. K-means clustering step 4

The next table is the final result. The principle of K-means algorithm is to make all samples in one cluster to be closer to each other, but the distance of each clusters should be larger.

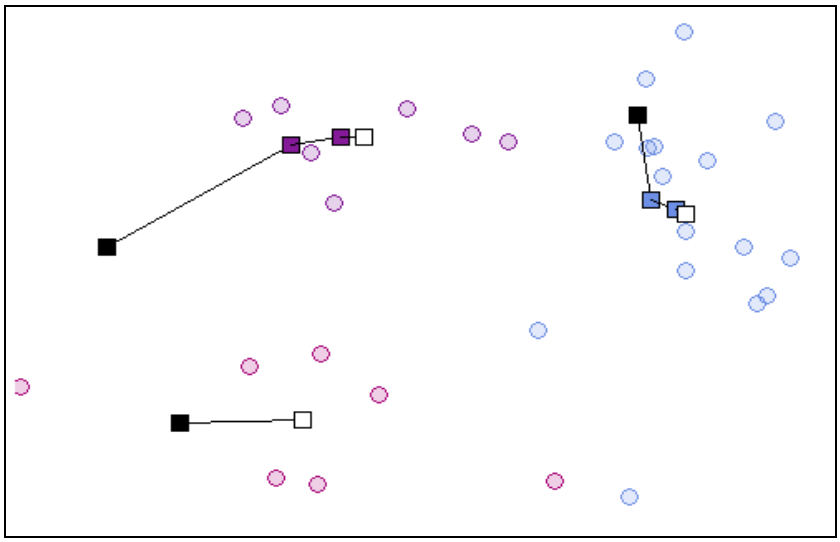

Figure 8. K-means clustering step 5

3.2.2 The advantages and disadvantages of K-means

Every machine learning algorithm has advantages and disadvantages. Here are the advantages and disadvantages of K-means.

Advantages of K-means:

- If the number of variables is large, K-means computes faster than other clustering algorithms.
- K-means can make clusters tighter if the centroid can be found properly.

Disadvantages of K-means:

- The value of k is too difficult to choose. Sometimes, the amount of types of dataset is unknown.
- Different initialization number affects output of cluster results.

3.2.3 Why choosing the K-means clustering algorithm

The K-means clustering algorithm belongs to unsupervised learning. If the classification information is not given, we do not know what kind of types of the object exist in the data set. But we know how many classes exist in the dataset. If we know the number of classes in the dataset, then we know the number of clusters. In this case, we should choose unsupervised learning to find out which sample belongs to which cluster.

The K-means algorithm is the simplest clustering algorithm. For this case, the data set contains 50 samples for each type of Iris flower. That means each type has the same amount of samples so that the centroid would be easier to calculate. The K-means algorithm is good enough for this case.

4. Implementation

4.1 Python

Python is a programming language created by Guido van Rossum in 1989. Python is an interpreted, object-oriented, dynamic data type of high-level programming languages.(Python Software Foundation 2013). The programming language style is simple, clear and it also contains powerful different kinds of classes. Moreover, Python can easily combine other programming languages, such as C or C++.

As a successful programming language, it has its own advantages:

Simple&easy to learn: The concept of this programming language is as simple as it can be. That makes it easy for everyone to learn and use. It is easy to understand the syntax.

Open source: Python is completely free as it is an open source software. Several of open source scientific computing storage has the API for Python. Users can easy to install Python on their own computer and use the standard and extend library.

Scalability: Programmers can write their code in C or C++ and run them in Python.

4.2 SciKit-learn

Scikit–learn is an open source machine learning library for the Python programming language. It features various classification, regression, and clustering algorithms and is designed to interoperate with the Python numerical libraries NumPy and SciPy (Pedregosa et al*.* 2011*).* SciKit-learn contains the K-means algorithm based on Python and it helps to figure out how to implement this algorithm in programming.

4.3 Numpy, Scipy and Matplotlib

In Python, there is no data type called array. In order to implement the data type of array with python, numpy and scipy are the essential libraries for analyzing and calculating data. They are all open source libraries. Numpy is mainly used for the matrix calculation. scipy is developed based on numpy and it is mainly used for scientific research.

By using them in Python programming, they can be used with two simple commands:

>>> *import numpy*

>>> *import scipy*

Then Python will call the methods from numpy and scipy.

Mathplotlib is a famous library for plotting in Python. It provides a series of API and it is suitable for making interactive mapping. In this case, we need to use it to find the best result visually.

4.4 Preparing the Iris flower data set

The data set of Iris flower can be found in UCI Machine Learning Repositor (Bache & Lichman 2013). In this thesis, the famous Fisher's Iris data set will be used.

The data set of Iris flower can be also found in the Scikit-learn library. In site-packages, there is a folder named sklearn. In this folder, there is a datasets subfolder to contain many kinds of data sets for machine learning study.

The data set can be found in Appendix 1.

In the species of this table, 0 represents setosa, 1 represents versicolor, 2 represents virginica.

In the process of preparing a training data set and a testing data set, the greatest problem is how to find the most appropriate way to divide the data set into training data set and testing data set. In some cases, by using sampling theory and estimation theory, we can separate the whole data set into training data set and testing data set. However, sometimes, the method would be changed. The attributes and the property of the data set would be different in various machine learning objects. Thus, in this kind of situation, in order to achieve a better result of machine learning, the data set will be separated according to the property of attributes of the data set.

The K-means algorithm and unsupervised learning does not use a training data set to compute the training sample. Therefore, there is no need to separate the dataset into a training data set and a testing data set. It can simply use this dataset to get the result of clustering.

4.5 Machine learning system design

In general, the principles of machine learning system design should follow two basic requirements :

- the model selection and creation and
- the learning algorithm selection and design.

In addition, different models can have different learning systems. On the other hand, the objective function is also different in different learning models. The objective function can help the machine to establish a learning system. Moreover, the accuracy and complexity of different algorithms would be the most important factor of the learning system. If the chosen algorithm is not very adaptive to the learning system, then the efficiency and result of the learning system would be reduced. The selection of training data set can have an influence on learning performance and feature selection.

4.6 Using Python to implement the program

For good implementation and good compatibility, Python version 2.7 will be in use. The Integrated Development Environment in this case will be PyScripter.

By using the Scikit-learn software package, there is no need to write a program to implement the K-means algorithm. After the installation has been finished, the K-means algorithm source code can be found in sklearn library. The source code of K-means clustering of Iris recognition can be found in the official website of Scikit-learn.

First of all, we need to import the library of numpy, dataset of Iris, K-means and Axes3D into the program. These are needed for this program. Numpy can helps to implement the K-means algorithm, the Iris dataset is the main data to be analyzed, Axes3d can make 3D outputs of this program, and the image will be more visual.

```
>>> import numpy as np
>>> import pylab as pl
>>> from mpl_toolkits.mplot3d import Axes3D
>>> from sklearn.cluster import KMeans
>>> from sklearn import datasets
```

Then, the program loads the Iris dataset and sets the centroid value and the number of clusters. In this program, the number of k clusters will be chosen as three and eight. In order to make a comparison, the third one will be the number of clusters 3, but with a bad initialization on the classification process. The initialiazation number has changed to 1. The default number is 10. Therefore the times the algorithm executes with different centroid seeds is reduced. This shows what happens to the result if the whole system has a bad initialization.

```
>>> np.random.seed(5)
>>> centers = [[1, 1], [-1, -1], [1, -1]]
>>> iris = datasets.load_iris()
>>> X = iris.data
>>> y = iris.target
>>> estimators = {'k_means_iris_3': KMeans(n_clusters=3),
...               'k_means_iris_8': KMeans(n_clusters=8),
...              'k_means_iris_bad_init': KMeans(n_clusters=3, n_init=1,
...                                              init='random')}
...
>>>
```

The result is shown as a table with three feature vectors. The feature vectors consists of petal width, sepal length and petal length. The output table will be three-dimensional.

```
>>> fignum = 1
>>> for name, est in estimators.iteritems():
...     fig = pl.figure(fignum, figsize=(4, 3))
...     pl.clf()
...     ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)
...     pl.cla()
...     est.fit(X)
...     labels = est.labels_
...     ax.scatter(X[:, 3], X[:, 0], X[:, 2], c=labels.astype(np.float))
...     ax.w_xaxis.set_ticklabels([])
...     ax.w_yaxis.set_ticklabels([])
...     ax.w_zaxis.set_ticklabels([])
...     ax.set_xlabel('Petal width')
...     ax.set_ylabel('Sepal length')
...     ax.set_zlabel('Petal length')
...     fignum = fignum + 1
...

KMeans(copy_x=True, init='k-means++', k=None, max_iter=300, n_clusters=8,
    n_init=10, n_jobs=1, precompute_distances=True, random_state=None,
    tol=0.0001, verbose=0)
<mpl_toolkits.mplot3d.art3d.Patch3DCollection object at 0x000000000F149E48>
[]
[]
[]
<matplotlib.text.Text object at 0x0000000016541CC0>
<matplotlib.text.Text object at 0x000000000F479E80>
<matplotlib.text.Text object at 0x000000000F470320>
KMeans(copy_x=True, init='k-means++', k=None, max_iter=300, n_clusters=3,
    n_init=10, n_jobs=1, precompute_distances=True, random_state=None,
    tol=0.0001, verbose=0)
<mpl_toolkits.mplot3d.art3d.Patch3DCollection object at 0x000000000F46CD68>
[]
[]
[]
<matplotlib.text.Text object at 0x000000000F45C358>
<matplotlib.text.Text object at 0x000000000F4745F8>
<matplotlib.text.Text object at 0x000000000F4488D0>
KMeans(copy_x=True, init='random', k=None, max_iter=300, n_clusters=3,
    n_init=1, n_jobs=1, precompute_distances=True, random_state=None,
    tol=0.0001, verbose=0)
<mpl_toolkits.mplot3d.art3d.Patch3DCollection object at 0x000000000F8BACC0>
[]
[]
[]
<matplotlib.text.Text object at 0x000000000F140898>
<matplotlib.text.Text object at 0x000000000F13D278>
<matplotlib.text.Text object at 0x000000000F88EBE0>
```

Then the program will show the standard plot of K-means clustering of Iris flower in supervised learning technique. The standard result of clustering is labeled with three species.

```
>>> fig = pl.figure(fignum, figsize=(4, 3))
>>> pl.clf()
>>> ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)
>>> pl.cla()
>>> for name, label in [('Setosa', 0),
...                     ('Versicolour', 1),
...                     ('Virginica', 2)]:
...     ax.text3D(X[y == label, 3].mean(),
...               X[y == label, 0].mean() + 1.5,
...               X[y == label, 2].mean(), name,
...               horizontalalignment='center',
...               bbox=dict(alpha=.5, edgecolor='w', facecolor='w'))
...
<mpl_toolkits.mplot3d.art3d.Text3D object at 0x0000000016A7D4A8>
<mpl_toolkits.mplot3d.art3d.Text3D object at 0x0000000016A7D438>
<mpl_toolkits.mplot3d.art3d.Text3D object at 0x0000000016A7D588>
>>>
```

The next step is to reorder the labels with the matched colors for the cluster results. After that all of the figures will be shown on the screen.

```
>>> y = np.choose(y, [1, 2, 0]).astype(np.float)
>>> ax.scatter(X[:, 3], X[:, 0], X[:, 2], c=y)
<mpl_toolkits.mplot3d.art3d.Patch3DCollection object at 0x000000000EE22240>
>>> ax.w_xaxis.set_ticklabels([])
[]
>>> ax.w_yaxis.set_ticklabels([])
[]
>>> ax.w_zaxis.set_ticklabels([])
[]
>>> ax.set_xlabel('Petal width')
<matplotlib.text.Text object at 0x000000000F8E4BE0>
>>> ax.set_ylabel('Sepal length')
<matplotlib.text.Text object at 0x000000000F8EED30>
>>> ax.set_zlabel('Petal length')
<matplotlib.text.Text object at 0x000000000F8F0CC0>
>>> pl.show()
>>>
```

5. Evaluating results

The result is shown in four images for the clustering results. Figure 9 will be the result with eight clusters. Figure 10 shows the result with three clusters.

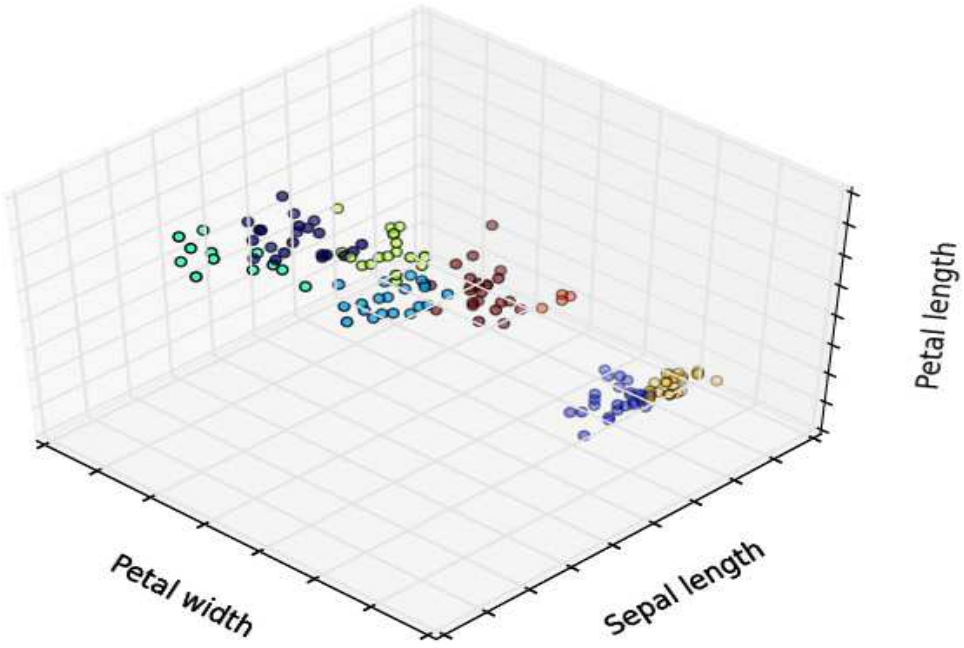

Figure 9. Clustering of Iris dataset with eight clusters

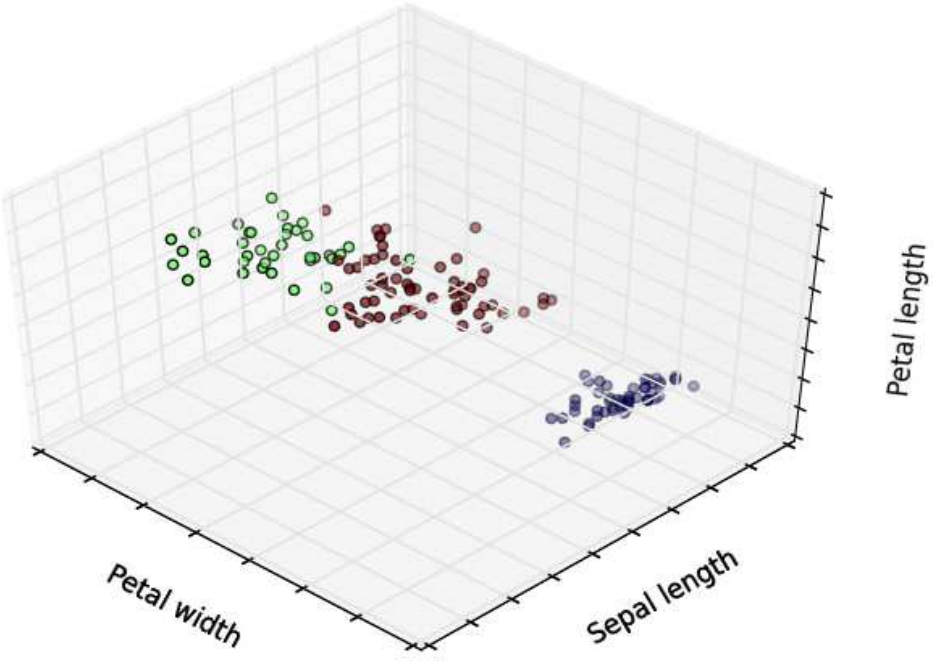

Figure 10. Clustering of Iris dataset with three clusters

As seen in Figure 9 and 10, the whole dataset is separated into eight clusters in Figure 9 and three clusters are shown in Figure 10 with different colors. In Figure 9, most of the samples stick together, it is really hard to distinguish them very clearly. The differences between each sample is small. In this case, the cluster result is not acceptable. On the other hand, in Figure 10, it can be easily seen that the cluster result is much better than in Figure 9. Even though there are still some overlapping parts between green and purple, but it quite clear to see the difference between these three clusters. This case shows the importance of choosing the number of clusters for K-means algorithm. Sometimes for the real datasets, it is difficult to know how many data sets should be used. Therefore, it is quite hard to choose the number of clusters. One method is to use the ISODATA algorithm, through the merging and division of clusters to obtain a reasonable number of k.

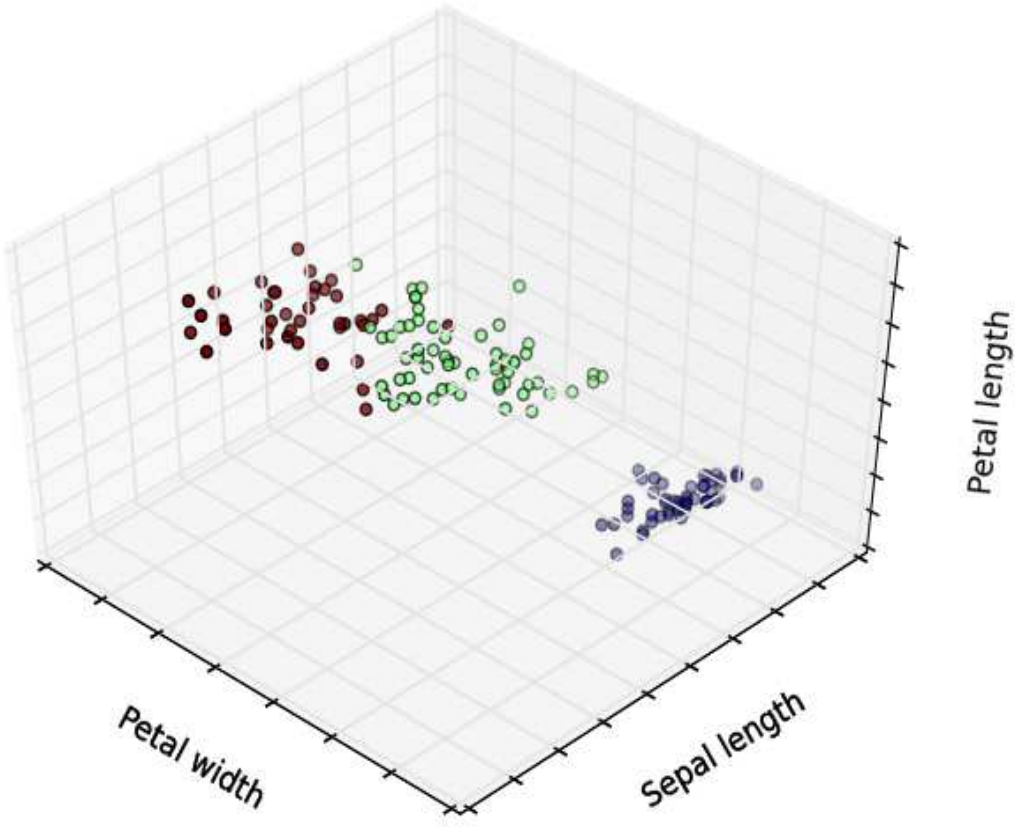

Figure 11. Clustering of Iris dataset with bad initialization

Figure 11 , shows the cluster result with three clusters but bad initialization. We can see that some of the samples change their class compare to the Figure 10. With a random initialization number, the system will obtain different cluster results. Therefore, a random initialization number is very important for a good cluster result. However, we do not know what could be a good initialization number. In this case, in some machine learning systems, the scientists will choose GA(Genetic Algorithm) to have the initialization point.

Figure 12 below illustrates a standard result of K-means clustering of Iris recognition. The term “ground truth” refers to the classification of training datasets in supervised learning. The number of clusters are three and with a good initialization point. This is the best classification of all shown here. The whole dataset has been separated properly and each dataset has good differences. In Figure 10, it shows the stardard result of classification in unsupervised learning. Compare to this figure, Figure 10 still has some small differences but it still works very well. Almost every data belongs to the right place.

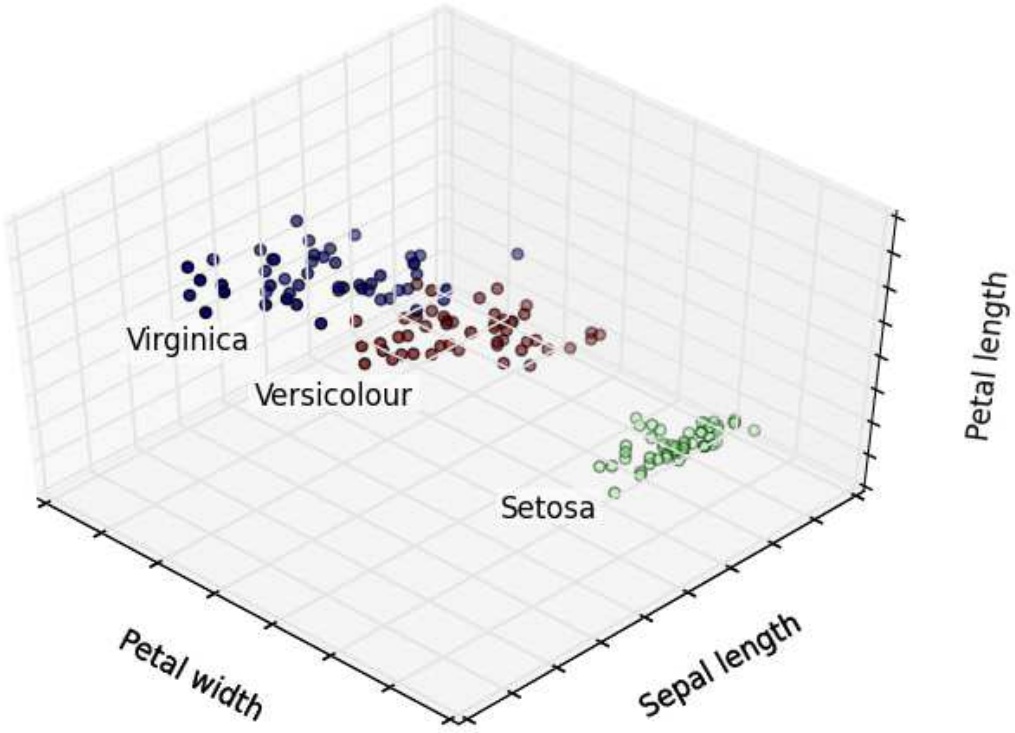

Figure 12. Clustering of Iris dataset in ground truth

These results show the effect that the number of k and the random initialization number have on the clustering result. It is also possible to see the advantages and disadvantages of the K-means clustering algorithm.

6. The future prospects

The Iris recognition case study above shows that the Machine Learning algorithm works well in this pattern recognition. The speed of computing is fast and the result is acceptable. However, the K-means clustering algorithm is just one of the clustering algorithm in unsupervised learning. There are more algorithms for different work objectives in different scientific fields.

As it is mentioned above, Machine Learning can be separated into supervised learning and unsupervised learning. However, sometimes, a whole dataset have both labeled data and unlabeled data. In order to process this kind of dataset, a new learning method called Semi-supervised(SSL) Learning has become a research hotspot. Because of this learning method, both machine learning and pattern recognition have a new research direction. It saves a lot of time and human resource to label those large amounts of unlabeled data. The SSL is also significant on improving learning performance of a computer.

Moreover, a learning system always consists of two parts, learning and environment. The environment gives knowledge to the computer and the computer will transfer this knowledge and store them and select useful information to implements different learning objectives. Therefore, different learning strategies can also be separated into rote learning, learning from instruction, learning by deduction, learning by analog, explanation-based learning and learning from induction. All of them have different algorithms to process different work objectives.

The implemented case in this thesis is only a simple example of machine learning and pattern recognition. Moreover, the K-means algorithm used in this thesis is a basic algorithm. However, if the data set has many feature dimensions and it is complicated, and if the learning objective is not that simple, the K-means algorithm can not be used.

Nowadays, GA (Genetic Algorithm), Artificial neural network and other machine learning algorithms have become more and more stable and useful. Many

scientists are working on improving the performance of machine learning algorithms.The K-means has also its own improved parts. The K-means can also be used along with other algorithms, such as ISODATA ,EM and K-means++. A better machine learning algorithm can obtain better results for pattern recognition. As the technology of pattern recognition develops, it requires more professional and more perfect machine learning algorithms. In this case, machine learning has a huge potential for growth.

In general, besides pattern recognition, machine learning can also be widely used in many fields of computer science and Artificial Intelligence. More and more Artificial Intelligence products are coming out on the market. Nowadays, people can use the Artificial Intelligence products every day. For example, people use Google search for seeking information which it is also based on the clustering algorithm of machine learning. All in all, machine learning definitely has a bright prospect.

7. Conclusion

With the rapid development of technology, AI has been applied in many fields. Machine learning is the most fundamental approach to achieve AI. This thesis describes the work principle of machine learning, two different learning forms of machine learning and an application of machine learning. In addition, a case study of Iris flower recognition to introduce the workflow of machine learning in pattern recognition is shown. In this case, the meaning of pattern recognition and how the machine learning works in pattern recognition has been described. The K-means algorithm, which is a very simple machine learning algorithm from the unsupervised learning method is used. The work also shows how to use SciKit-learn software to learn machine learning.

References

‘Clustering - K-means demo’, K-means-Ineractive demo, Available at: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html. Consulted 22 AUG 2013

Bache, K.& Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Bishop, C. 2006. Pattern Recognition and Machine Learning. New York: Springer, pp.424-428.

Fisher, R.A. 1936. UCI Machine Learning Repository: Iris Data Set. Available at: http://archive.ics.uci.edu/ml/datasets/Iris. Consulted 10 AUG 2013

Improved Outcomes Software.,2004. K-Means Clustering Overview, Available at: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm. Consulted 22 AUG 2013

Mitchell, T. 1997. Machine learning. McGraw Hill.

Mjolsness, E. & Decoste, D. 2001. Machine learning for science: state of the art and future prospects.Science, 293 (5537), pp. 2051--2055.

Pedregosa, F.& Varoquaux, G. 2.11., Scikit-learn: machine learning in Python — Scipy lecture notes, Available at: http://scipy-lectures.github.io/advanced/scikit-learn/. Consulted 22 AUG 2013

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Others*.,2011*. Scikit-learn: Machine Learning in Python., JMLR 12, pp. 2825-2830.

Python Software Foundation., 2013. General Python FAQ — Python v2.7.5 documentation. Available at: http://docs.python.org/2/faq/general.html. Consulted 20 AUG 2013

Qin, X. and Zheng, S. 2009. A new method for initialising the K-means clustering algorithm. Knowledge Acquisition and Modeling, 2009. KAM '09. Second International Symposium on. vol. 2, pp. 41--44.

Robin. 2010. Machine-Learning - Artificial Intelligence. Available at: http://intelligence.worldofcomputing.net/category/machine-learning. Consulted 22 AUG 2013

Theodoridis, S.& Koutroumbas, K. 2006. Pattern Recognition. 3rd Edition. Publisher: Academic Press.

Teknomo,K.2006.K-Means Clustering: Numerical Example. Available at: http://people.revoledu.com/kardi/tutorial/kMean/NumericalExample.htm. Consulted 22 AUG 2013

APPENDIX

1. Fisher's Iris flower dataset

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	0
4.9	3	1.4	0.2	0
4.7	3.2	1.3	0.2	0
4.6	3.1	1.5	0.2	0
5	3.6	1.4	0.2	0
5.4	3.9	1.7	0.4	0
4.6	3.4	1.4	0.3	0
5	3.4	1.5	0.2	0
4.4	2.9	1.4	0.2	0
4.9	3.1	1.5	0.1	0
5.4	3.7	1.5	0.2	0
4.8	3.4	1.6	0.2	0
4.8	3	1.4	0.1	0
4.3	3	1.1	0.1	0
5.8	4	1.2	0.2	0
5.7	4.4	1.5	0.4	0
5.4	3.9	1.3	0.4	0
5.1	3.5	1.4	0.3	0
5.7	3.8	1.7	0.3	0
5.1	3.8	1.5	0.3	0
5.4	3.4	1.7	0.2	0
5.1	3.7	1.5	0.4	0
4.6	3.6	1	0.2	0
5.1	3.3	1.7	0.5	0
4.8	3.4	1.9	0.2	0
5	3	1.6	0.2	0
5	3.4	1.6	0.4	0
5.2	3.5	1.5	0.2	0
5.2	3.4	1.4	0.2	0
4.7	3.2	1.6	0.2	0
4.8	3.1	1.6	0.2	0
5.4	3.4	1.5	0.4	0
5.2	4.1	1.5	0.1	0
5.5	4.2	1.4	0.2	0
4.9	3.1	1.5	0.1	0
5	3.2	1.2	0.2	0
5.5	3.5	1.3	0.2	0
4.9	3.1	1.5	0.1	0
4.4	3	1.3	0.2	0
5.1	3.4	1.5	0.2	0

5	3.5	1.3	0.3	0
4.5	2.3	1.3	0.3	0
4.4	3.2	1.3	0.2	0
5	3.5	1.6	0.6	0
5.1	3.8	1.9	0.4	0
4.8	3	1.4	0.3	0
5.1	3.8	1.6	0.2	0
4.6	3.2	1.4	0.2	0
5.3	3.7	1.5	0.2	0
5	3.3	1.4	0.2	0
7	3.2	4.7	1.4	1
6.4	3.2	4.5	1.5	1
6.9	3.1	4.9	1.5	1
5.5	2.3	4	1.3	1
6.5	2.8	4.6	1.5	1
5.7	2.8	4.5	1.3	1
6.3	3.3	4.7	1.6	1
4.9	2.4	3.3	1	1
6.6	2.9	4.6	1.3	1
5.2	2.7	3.9	1.4	1
5	2	3.5	1	1
5.9	3	4.2	1.5	1
6	2.2	4	1	1
6.1	2.9	4.7	1.4	1
5.6	2.9	3.6	1.3	1
6.7	3.1	4.4	1.4	1
5.6	3	4.5	1.5	1
5.8	2.7	4.1	1	1
6.2	2.2	4.5	1.5	1
5.6	2.5	3.9	1.1	1
5.9	3.2	4.8	1.8	1
6.1	2.8	4	1.3	1
6.3	2.5	4.9	1.5	1
6.1	2.8	4.7	1.2	1
6.4	2.9	4.3	1.3	1
6.6	3	4.4	1.4	1
6.8	2.8	4.8	1.4	1
6.7	3	5	1.7	1
6	2.9	4.5	1.5	1
5.7	2.6	3.5	1	1
5.5	2.4	3.8	1.1	1
5.5	2.4	3.7	1	1
5.8	2.7	3.9	1.2	1
6	2.7	5.1	1.6	1
5.4	3	4.5	1.5	1
6	3.4	4.5	1.6	1
6.7	3.1	4.7	1.5	1

6.3	2.3	4.4	1.3	1
5.6	3	4.1	1.3	1
5.5	2.5	4	1.3	1
5.5	2.6	4.4	1.2	1
6.1	3	4.6	1.4	1
5.8	2.6	4	1.2	1
5	2.3	3.3	1	1
5.6	2.7	4.2	1.3	1
5.7	3	4.2	1.2	1
5.7	2.9	4.2	1.3	1
6.2	2.9	4.3	1.3	1
5.1	2.5	3	1.1	1
5.7	2.8	4.1	1.3	1
6.3	3.3	6	2.5	2
5.8	2.7	5.1	1.9	2
7.1	3	5.9	2.1	2
6.3	2.9	5.6	1.8	2
6.5	3	5.8	2.2	2
7.6	3	6.6	2.1	2
4.9	2.5	4.5	1.7	2
7.3	2.9	6.3	1.8	2
6.7	2.5	5.8	1.8	2
7.2	3.6	6.1	2.5	2
6.5	3.2	5.1	2	2
6.4	2.7	5.3	1.9	2
6.8	3	5.5	2.1	2
5.7	2.5	5	2	2
5.8	2.8	5.1	2.4	2
6.4	3.2	5.3	2.3	2
6.5	3	5.5	1.8	2
7.7	3.8	6.7	2.2	2
7.7	2.6	6.9	2.3	2
6	2.2	5	1.5	2
6.9	3.2	5.7	2.3	2
5.6	2.8	4.9	2	2
7.7	2.8	6.7	2	2
6.3	2.7	4.9	1.8	2
6.7	3.3	5.7	2.1	2
7.2	3.2	6	1.8	2
6.2	2.8	4.8	1.8	2
6.1	3	4.9	1.8	2
6.4	2.8	5.6	2.1	2
7.2	3	5.8	1.6	2
7.4	2.8	6.1	1.9	2
7.9	3.8	6.4	2	2
6.4	2.8	5.6	2.2	2
6.3	2.8	5.1	1.5	2

6.1	2.6	5.6	1.4	2
7.7	3	6.1	2.3	2
6.3	3.4	5.6	2.4	2
6.4	3.1	5.5	1.8	2
6	3	4.8	1.8	2
6.9	3.1	5.4	2.1	2
6.7	3.1	5.6	2.4	2
6.9	3.1	5.1	2.3	2
5.8	2.7	5.1	1.9	2
6.8	3.2	5.9	2.3	2
6.7	3.3	5.7	2.5	2
6.7	3	5.2	2.3	2
6.3	2.5	5	1.9	2
6.5	3	5.2	2	2
6.2	3.4	5.4	2.3	2
5.9	3	5.1	1.8	2
5.1	3.5	1.4	0.2	2

2. Source code

```
print(__doc__)



# Code source: Gael Varoqueux
# Modified for Documentation merge by Jaques Grobler
# License: BSD 3 clause

import numpy as np
import pylab as pl
from mpl_toolkits.mplot3d import Axes3D



from sklearn.cluster import KMeans
from sklearn import datasets

np.random.seed(5)
```

```
centers = [[1, 1], [-1, -1], [1, -1]]
iris = datasets.load_iris()
X = iris.data
y = iris.target

estimators = {'k_means_iris_3': KMeans(n_clusters=3),
              'k_means_iris_8': KMeans(n_clusters=8),
              'k_means_iris_bad_init': KMeans(n_clusters=3, n_init=1,
                                              init='random')}



fignum = 1
for name, est in estimators.iteritems():
    fig = pl.figure(fignum, figsize=(4, 3))
    pl.clf()
    ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)

    pl.cla()
    est.fit(X)
    labels = est.labels_

    ax.scatter(X[:, 3], X[:, 0], X[:, 2], c=labels.astype(np.float))

    ax.w_xaxis.set_ticklabels([])
    ax.w_yaxis.set_ticklabels([])
    ax.w_zaxis.set_ticklabels([])
    ax.set_xlabel('Petal width')
    ax.set_ylabel('Sepal length')
    ax.set_zlabel('Petal length')
    fignum = fignum + 1

# Plot the ground truth
```

```
fig = pl.figure(fignum, figsize=(4, 3))
pl.clf()
ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)

pl.cla()

for name, label in [('Setosa', 0),
                    ('Versicolour', 1),
                    ('Virginica', 2)]:
    ax.text3D(X[y == label, 3].mean(),
              X[y == label, 0].mean() + 1.5,
              X[y == label, 2].mean(), name,
              horizontalalignment='center',
              bbox=dict(alpha=.5, edgecolor='w', facecolor='w'))
# Reorder the labels to have colors matching the cluster results
y = np.choose(y, [1, 2, 0]).astype(np.float)
ax.scatter(X[:, 3], X[:, 0], X[:, 2], c=y)

ax.w_xaxis.set_ticklabels([])
ax.w_yaxis.set_ticklabels([])
ax.w_zaxis.set_ticklabels([])
ax.set_xlabel('Petal width')
ax.set_ylabel('Sepal length')
ax.set_zlabel('Petal length')
pl.show()
```