INT404: ARTIFICIAL INTELLIGENCE PROJECT

**A study of pattern recognition and classification of Iris flower based on Machine Learning from AI.**

Section: K18GR:

By :

Roll no :

1. J MAHENDER K -14

2. UDAY KIRAN MAHAPATRO - 01

3.MOHAK GAUTAM - 08

4.RANJIT RAJ - 18

**CONTENTS:**

|  |  |  |  |
| --- | --- | --- | --- |
| **1** | **Introduction** | |  |
|  | **1.1** | **Objectives** |  |
|  | **1.2** | **Collecting data set** |  |
|  | **1.3** | **Using K-means algorithm to achieve clustering** |  |
|  | **1.4** | **Evaluating result** |  |
| **2** | **Literature review** | |  |
|  | **2.1** | **Basic introduction to machine learning** |  |
|  |  | 2.1.1 Basic structure of machine learning system workflow |  |
|  | **2.2** | **Machine learning in pattern recognition** |  |
|  |  | 2.2.1 Machine learning algorithm in pattern recognition |  |
| **3** | **K-means clustering** | |  |
|  | **3.1** | **Introduction to clustering** |  |
|  |  | 3.1.1 K-means algorithm workflow |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | 3.1.2 The advantages and disadvantages of K-means |  |
|  |  | 3.1.3 Why choose K-means clustering algorithm |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **4** | **Implementation** | |  |
|  | **4.1** | **Python** |  |
|  | **4.2** | **SciKit-learn** |  |
|  | **4.3** | **numpy & scipy & matplotlib** |  |
|  | **4.4** | **Prepare Iris flower dataset** |  |
|  | **4.5** | **Machine learning system design** |  |
|  | **4.6** | **Using python to implement the program** |  |
| **5** | **Evaluating results** | |  |
| **6** | **Conclusion** | |  |

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.

1.2 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.

**This module is done by J MAHENDER K:**

**2．Literature review**

2.1 Basic introduction to machine learning

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.

2.1.1 Fundamental structure of machine learning system

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Environment |  | Learning | |  | Knowledge |  | execution | |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |



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.

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

2.2 Machine Learning in pattern recognition

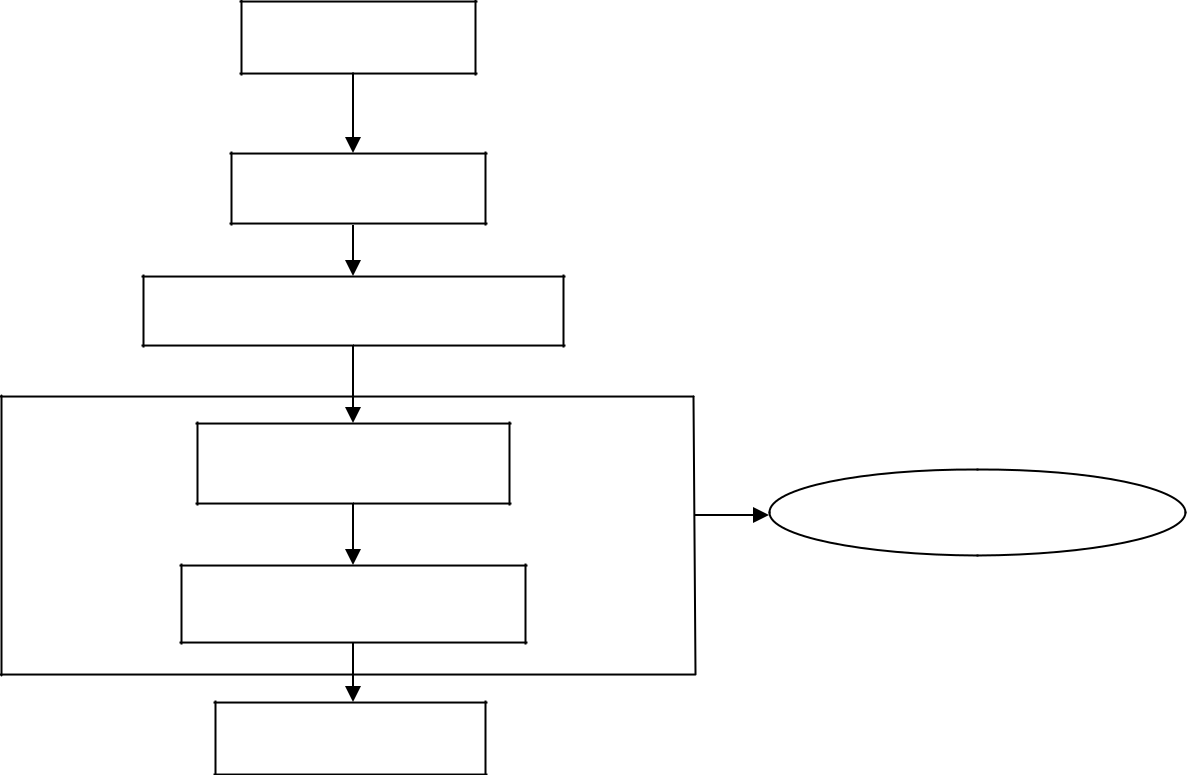
As mentioned above, the method of machine learning can also be used in pattern recognition. In fact, pattern recognition really needs machine learning to achieve its objective.

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.

2.2.1 Machine learning algorithm in pattern recognition.



Data storage

Preprocssing

spanning characteristics

Feature selection

Machine Learning

Pattern classification

Post processing

Figure 1: pattern recognition system.

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.

**This module done by Uday Kiran:**

**3.1K-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.

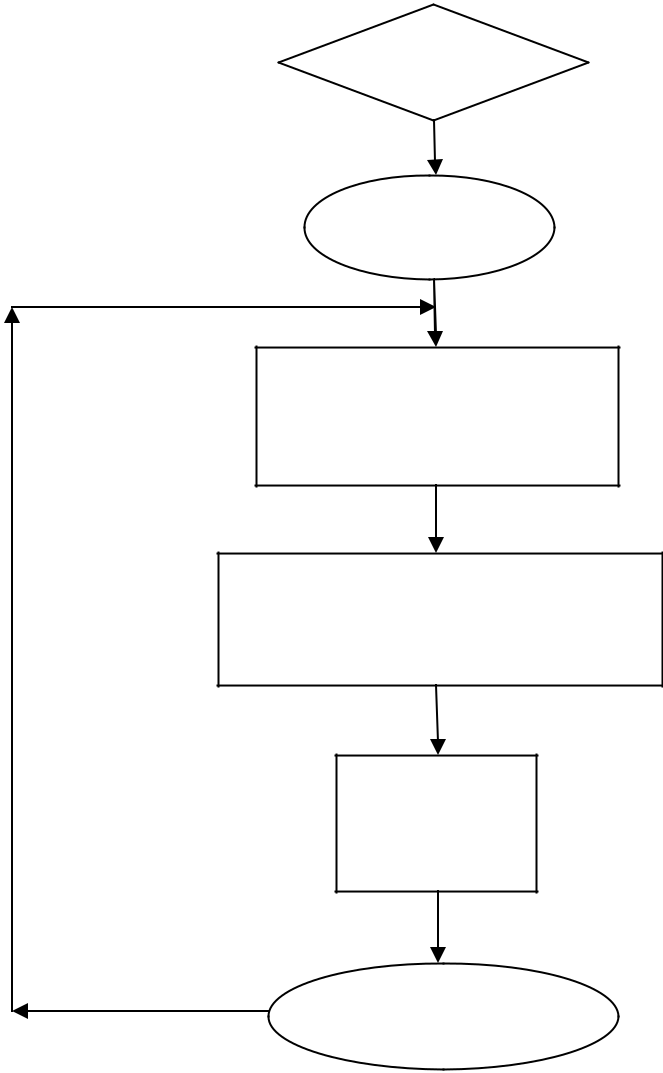
3.1 Introduction to clustering

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.

3.1.1 K-means algorithm workflow



start

K clusters

Determine the cluster

centers

Determine distance of each

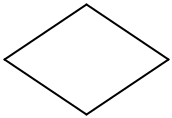
object to cluster centers

Grouping

objects

Yes other objects?

No 



end

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.

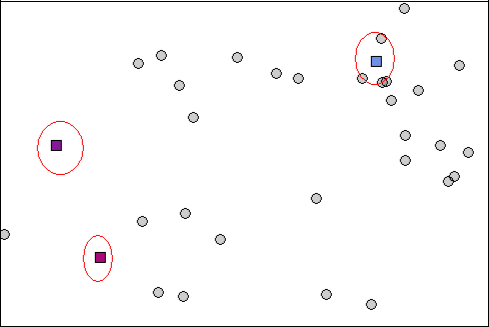


Figure 4. K-means clustering step 1

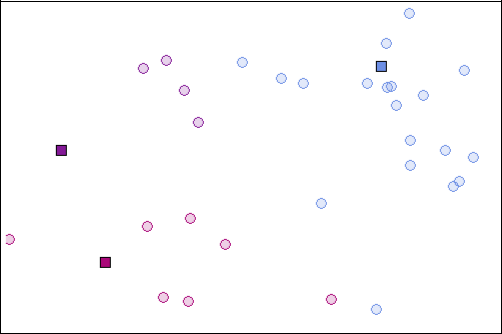
Now the system generates three cluster points with randomly. There are three different colours: purple(Top left), blue(Top right), pink(Bottom left). These three colours stand for three different clusters.

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.

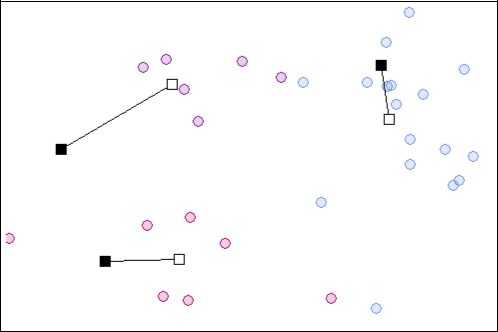


Figure 6. K-means clustering step 3

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.

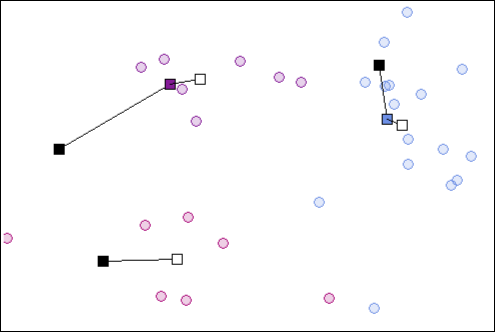


Figure 7. K-means clustering step 4

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.

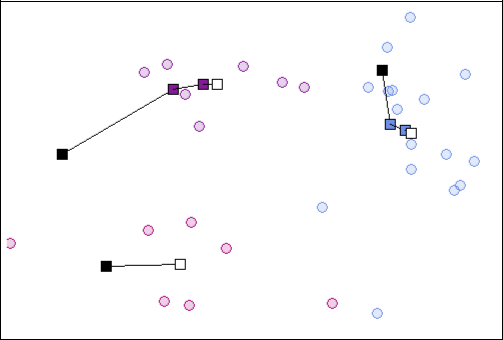


Figure 8. K-means clustering step 5

3.1.2 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.

3.1.3 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.

**This module done by Mohak Gautam:**

**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 simpleas 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 inPython

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.

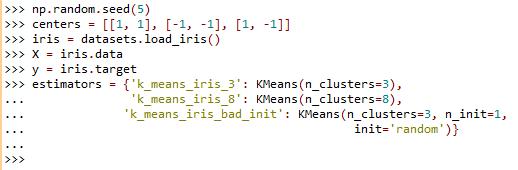
4.6 Using Python to implement the program

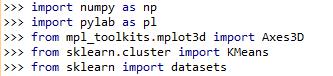
For good implementation and good compatibility, Python version 2.7 will be in use. The Integrated Development Environment in this case will be PyScripter.

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.

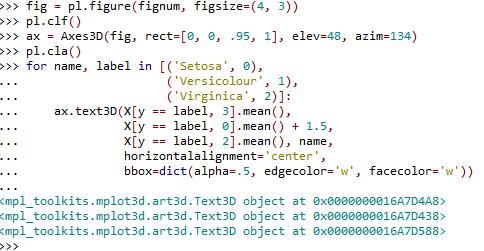
Then, the program loads the Iris dataset and sets the centroid value and the number of clusters. In this program, the number of k clusters will be chosen as three and eight. In order to make a comparison, the third one will be the number of clusters 3, but with a bad initialization on the classification 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.



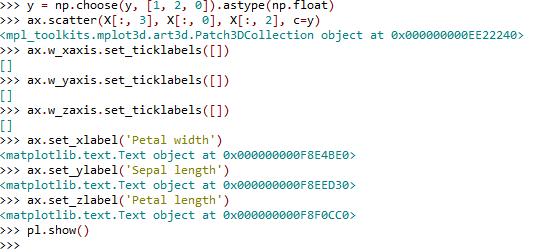
The result is shown as a table with three feature vectors. The feature vectors consists of petal width, sepal length and petal length. The output table will be three-dimensional.



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.



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.



**This module done by Ranjit Raj:**

**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.

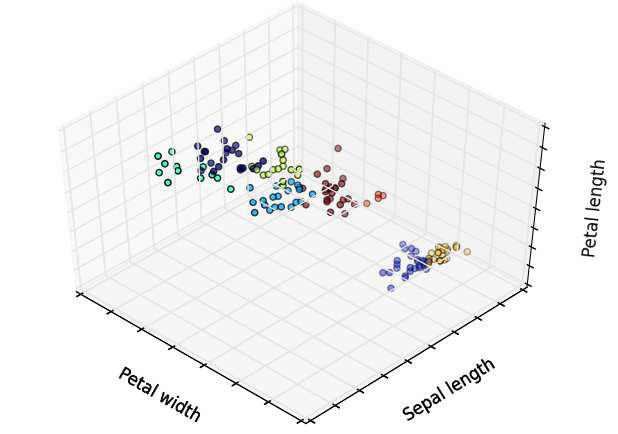


Figure 9. Clustering of Iris dataset with eight clusters

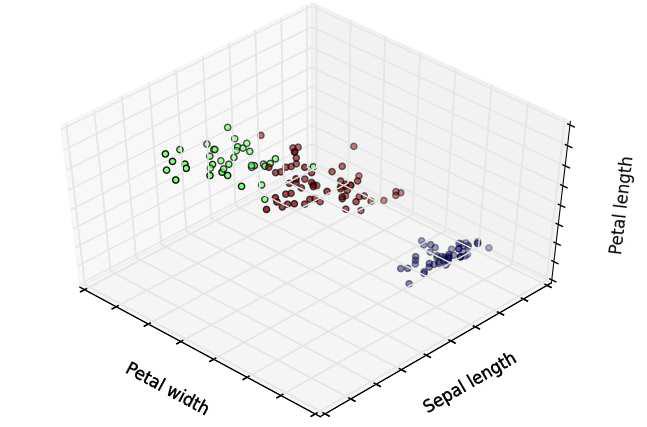


Figure 10. Clustering of Iris dataset with three clusters

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.

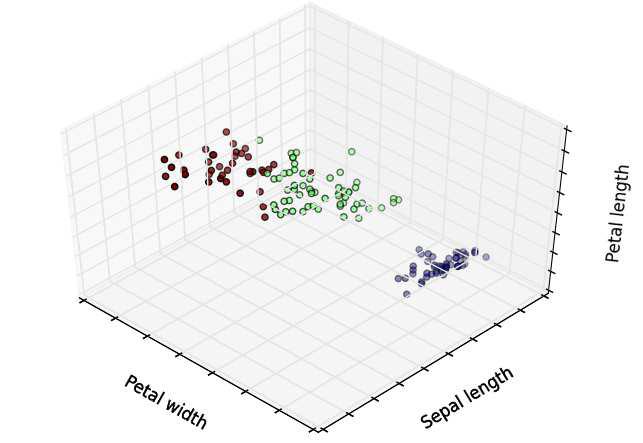


Figure 11. Clustering of Iris dataset with bad initialization

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.

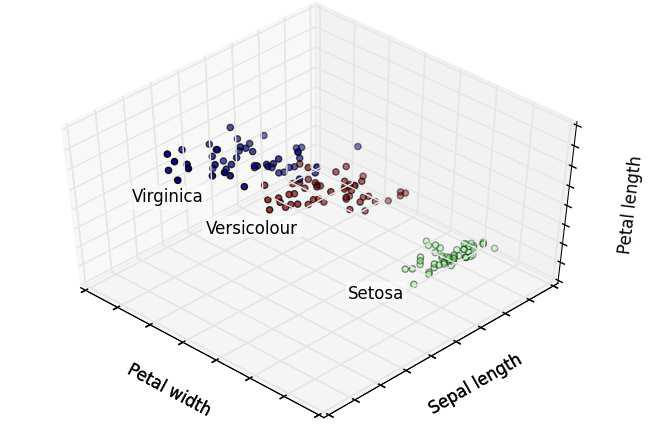


Figure 12. Clustering of Iris dataset in ground truth

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.

**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.