Keywords A to Z Vol. 32

Keyword Clustering

A type of machine learning in which algorithms classify a huge amount of data according to the degree of similarity. Clustering can be used in the analysis of big data to find data features and automatically classify data into groups.

An efficient data grouping technique, essential for AI

Recently we have often heard about clusters in news about the spread of COVID-19. The word, of course, means a group of something, and it is the root of the word clustering. Early detection of a cluster where COVID-19 infection is spreading is one of the countermeasures for preventing the spread of infection.

A large amount of data is also called a cluster when processed in connection with AI or analysis of big data. The amount of data that can be collected has increased explosively due to the development of the Internet of Things (IoT). Big data is said to be a bonanza that can lead to innovation, but analyzing huge amounts of data is not easy. Therefore, it is expected that analysis will be done by AI. In recent years, various machine learning algorithms have been developed and their potential has greatly expanded.

Machine learning can be roughly divided into two types: supervised learning, which learns rules or patterns based on a set of data for which correct answers are known (training data) and builds a model; and unsupervised learning, which learns about data structure, features, or similarities from a set of data (not training data) and classifies it into groups. For example, supervised learning algorithms first learn the characteristics of cats and then extract pictures of cats from a large number of pictures. On the other hand, unsupervised learning algorithms extract the characteristics from an enormous number of pictures that were input and classify them into groups. Each piece of data may seem disjointed at a glance, but if the data is classified into groups appropriately, it becomes meaningful and can be used effectively.

Three methods for accurate data classification

There are three types of clustering methods.

The hierarchical clustering method builds a hierarchy of clusters by comparing each piece of data and combining clusters with high similarity or dividing clusters with low internal similarity. This process is repeated until all data is classified into a single cluster. With this method, data can be classified into the number of groups that is desirable for the purpose.

In the non-hierarchical clustering method, the number of clusters to be created is determined first, and the algorithm searches for optimal division without building a hierarchy of clusters. Since this method requires less calculation than hierarchical clustering, it is effective for analyzing big data.

The density-based clustering method defines an area where data density is high as a cluster and distant data items are treated as noise. Since dissimilar data is clearly identified, this method can improve the accuracy of clustering.

Among these methods, non-hierarchical clustering is the most widely used. In particular, “k-means clustering” can be considered to be the typical method. The k-means algorithm arranges clusters by repeatedly reassigning data points to clusters based on the distance between the data point and the center of a cluster.

AI can easily group students based on their grades

Here is an example of how the k-means method can be used for grouping students into English classes.

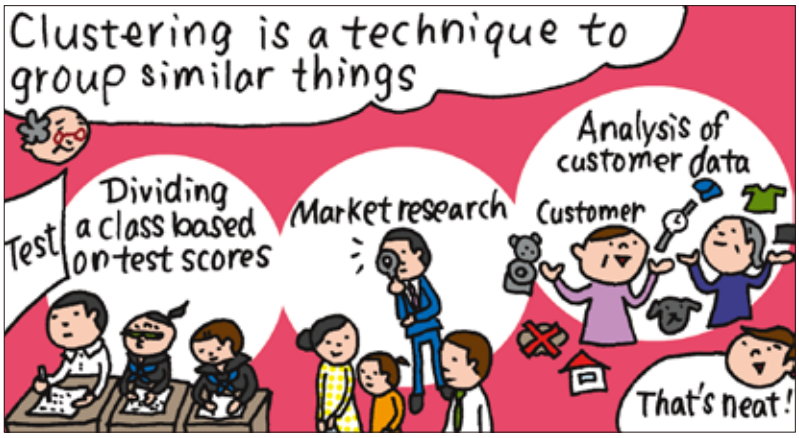
In a school, reading and listening tests were conducted to divide students into five classes according to their test scores. It would be possible to group the students based on their average scores, but some are good at reading only, some are good at listening only, and others are about the same at reading and listening.

Given this, the k-means method was used to classify the students. By plotting their scores on a graph with the vertical axis as reading and the horizontal axis as listening, clusters were identified. Five clusters were determined based on the students’ strengths and weaknesses, and students with similar academic abilities were successfully grouped.

If there are not many students, a teacher can group them manually, but if there are many students or if more subjects must be considered for making the classes, there are limits to manual clustering. If a huge amount of complex data must be handled, a clustering algorithm is effective.

In this example, the number of clusters was decided in advance, but in many cases that may not be possible. Depending on how many clusters are created, the results will change. If it is difficult to decide the number in advance, you can first calculate using a desired number of clusters, and then recalculate using a larger or smaller number, and compare the results to approach the optimum solution.

Clustering algorithms are effective for classifying and utilizing a huge amount of information that cannot be easily visualized. Research in the field of big data analysis using AI is progressing day by day.



©ad-manga.com

This article was published in April 2022.