International Journal of Engineering Research & Management Technology

ISSN: 2348-4039

Volume 9, Issue-6 November-December- 2022

Email: editor@ijermt.org www.ijermt.org

A STUDY ON K-MEANS CLUSTERING ALGORITHM AND ITS SIGNIFICANCE

Sunil Singh
Research scholar
Dept. Of computer application, OPJS University,
Churu, Rajasthan

Dr. Rajesh Pathak
Professor
OPJS University, Churu, Rajasthan

ABSTRACT

K-Means was developed by the Institute of Mathematics and its Applications (IMA). The information may be partitioned into k distinct clusters, each of which is determined by the level of similarity that exists between the groups. In order for the procedure to operate as intended, one is required to be aware of the value that is stored within the integer variable K. The K-mean method is the one that is used the most frequently for the purpose of clustering, and it has the ability to choose the appropriate cluster for new data based on the majority of the distance. This is because the K-mean method uses a weighted average of the distance between each pair of data points. The first k cluster centroids are chosen in a haphazard method by a random selection process. After that, the points are given to the centroids that are physically located in the closest proximity to them, and the centroids are recalculated to account for the newly formed group. The K-means method is advantageous in many ways, including the fact that it is simple to construct and explain, in addition to being efficient in terms of computing. A disadvantage of the technique is that it is difficult to estimate what the value of K is while employing this method, which is a drawback of the strategy. If the clusters take on a spherical form, then their efficiency will suffer. At the first stage, there are going to be two different groupings of objects that are present. They should proceed to the following stage, which is to determine the centroids of each set. The clusters that were responsible for creating the various dataset clusters are established when the data is reformed according to the centroid.

KEYWORDS:

K Means, Centroid, Clustering

INTRODUCTION

K-Means is an unsupervised learning approach that is often used in the process of categorizing data on the basis of the nearest neighbor. K-Means was developed by the Institute of Mathematics and its Applications (IMA). The information may be partitioned into k distinct clusters, each of which is determined by the level of similarity that exists between the groups. In order for the procedure to operate as intended, one is required to be aware of the value that is stored within the integer variable K.

Machine learning is a technique for learning a software model across a large database. Machine learning is a term that was coined by computer scientists. The data set that was employed in the process of training the software model is referred to as the "training data set." The databases that, when combined, produce the data set that will be used for training are referred to collectively as training samples, and they have been selected at random from the sample population. The training data set is used in its entirety in the development of the software model.

Test data set is the name given to the collection of information that is used to evaluate the quality of the software model. In order to evaluate the accuracy of the software model's prognostic capabilities, the test data set is utilized. The training samples are used to choose the test samples, but the test samples are not used in the selection process. In any given test data set, the accuracy of a software model would be measured by the proportion of test samples that are accurately predicted by the software model. In the event that the level of accuracy provided by the software model is enough, it will be possible to use the software model to classify previously unseen data records in the event that the value is unknown.

Another way that artificial intelligence helps doctors is by supplying them with the most recent clinical practice knowledge gleaned from journals, websites, and electronic books. Likewise, a system that utilizes artificial intelligence may lessen the likelihood of making diagnostic and treatment errors, both of which are inherently possible. As technology progresses, artificial intelligence will be able to provide historical health risk warnings and treatment outcome projections from huge amounts of population data by first filtering out the key information from those data using a real-time reference.

Infectious disorders are responsible for 15% of all cancers that are diagnosed in the world today. It is thought that anywhere between 5 and 10% of cases of cancer are due to genetic predispositions that were handed down from one generation to the next by an individual's parents. Cancer can be diagnosed by a comprehensive evaluation of a wide range of symptoms and screening practices. Following this, more testing may be carried out with the use of medical imaging, and the diagnosis may be verified with the assistance of a biopsy.

Mellitus Diabetes mellitus, which is more often known to simply as diabetes, is a metabolic illness that is defined primarily by aberrant insulin secretion and behavior. Diabetes is an abbreviated form of the medical term diabetes mellitus. The disease known as diabetes mellitus belongs to the group of conditions known as metabolic diseases. Hyperglycemia and problems with the metabolism of fatty acids, proteins, and carbohydrates are both signs of insulin insufficiency, which is the underlying condition that leads to elevated blood glucose levels.

K-MEANS CLUSTERING ALGORITHM AND ITS SIGNIFICANCE

The K-mean approach is the one that is used for clustering the most commonly, and it is able to select which new data should be placed in which cluster based on the majority of the distance. This makes it possible to cluster data more effectively.

The first k cluster centroids are chosen in a haphazard method by a random selection process. After that, the points are given to the centroids that are physically located in the closest proximity to them, and the centroids are recalculated to account for the newly formed group. Certain K-means are affected by centroids, which makes them more vulnerable to noise and outliers. This is because some of the K-means take into account centroids.

The K-means algorithm is advantageous in many ways, including the fact that it is simple to construct and explain, in addition to being efficient in terms of computing. A disadvantage of the technique is that it is difficult to estimate what the value of K is while employing this method, which is a drawback of the strategy. If the clusters take on a spherical form, then their efficiency will suffer.

The K-means approach has a graphical representation which presents the process in graphical form. At the first stage, there are going to be two different groupings of objects that are present. They should proceed to the following stage, which is to determine the centroids of each set. The clusters that were responsible for creating the various dataset clusters are established when the data is reformed according to the centroid. This process will be repeated until the optimal clusters have been identified, at which point we will go to step.

Because it enables a deeper interpretation of medical data as well as the prediction of diseases, data mining plays an essential part in the area of healthcare and is essential to solving real-world medical issues. This is why it is so important. At present point in time, the methods of DM are applied by researchers in the diagnosis of a broad variety of illnesses, such as genetic algorithms and K-NN, all exhibited differing degrees of accuracy when they were put to use.

The clinical evaluation of diseases demands a lot of measurements to be done and elements to be studied since these diseases have a high mortality rate and a very wide range of possible manifestations. Because of the enormous amount of medical data is centered on the management of these conditions is of the utmost importance.

In addition, the biomedical industry has made use of a number of methodologies that are founded on artificial intelligence for the purpose of carrying out clinical research on a wide range of different diseases and conditions. As a consequence of this, the primary objective of this research is going to be to investigate the application of significant machine learning algorithms in the field of medicine, such as ANN, K-NN, DT, and SVM, for the purpose of diagnosing and treating the types of diseases that are being discussed in this study.

Its rapid rise is being driven by the fact that MLTs are becoming an increasingly valuable component of the healthcare industry. In this analysis, we will be focusing on diseases and conditions such as cancer, diabetes, heart disease, and hepatitis, all of which are considered to be widespread problems all over the world by the World Health Organization.

Cancer may begin in one part of the body and then spread to other parts of the body. Some of the observable signs and indications that indicate to the probability of cancer are changes in bowel movement, tumors, irregular bleeding, acute coughing, and unexplained weight loss. Other symptoms and indicators include cancer itself. Cancer is a disease that has more than one hundred subtypes, which can affect people. Oncologists and other educated medical experts classify different forms of cancer into a variety of subtypes based on the location of the main tumor.

Cancer may be broken down into its basic four subgroups, which are referred to as carcinomas, sarcomas, leukemias, and lymphomas respectively. Cancer of the colorectal colon is the sixth most prevalent form of the disease. Breast cancer, colorectal cancer, lung cancer, and cervical cancer are the types of cancer that are diagnosed in women the most frequently. The most prevalent kind of cancer is breast cancer.

In 2015, more than 90,5 million people all around the world were given a diagnosis of cancer. Almost 14.1 million new instances are recorded each year, and this figure does not take into account the incidence of skin cancers other than melanoma. Over 8.8 million individuals ended up passing away as a direct consequence of this, making about 15.7% of the total number of fatalities. Tobacco usage is responsible for over 22 percent of all cases of cancer-related deaths that occur in the world. There is a connection between being overweight, not getting enough exercise, having a poor diet, and consuming too much alcohol. This connection accounts for around 10% of the problem.

DISCUSSION

Diabetes, which affects more than 200 million people all over the globe, is frequently considered to be one of the most well-known endocrine ailments. This is because diabetes affects more than 200 million people all over the world. In the years to come, it is projected that there will be a considerable increase in the number of persons whose medical conditions will be classified as diabetes. There are three unique varieties of diabetes that may be identified from one another: type 1 diabetes, type 2 diabetes, and gestational diabetes.

In a broad sense, type 1 diabetes is the most common form of diabetes. Polyuria, polydipsia, and significant weight loss are some of the most common symptoms of diabetes. Diabetes is characterized by a range of symptoms, some of the most prominent of which are polyuria and polydipsia. Both a person's fasting plasma glucose (FPG) and postprandial plasma glucose (PPG) readings, which measure the amount of glucose in the blood after they have eaten, are taken into consideration when determining whether or not a person has diabetes.

The terms "cardiovascular disease" and "heart disease" are commonly used interchangeably with the word "cardiovascular disease." A condition of the coronary arteries, a stroke, and a problem of the peripheral arteries are all related with atherosclerosis. This might be due to factors such as smoking, polygenic disorders including diabetes mellitus, hypertension, high cholesterol levels in the blood, a lack of exercise, obesity, a bad diet, or excessive use of alcohol.

Cardiovascular diseases are the leading cause of mortality across the world. Both coronary artery disease and stroke are responsible for seventy-five percent of fatalities caused by cardiovascular disease in females and eighty percent of deaths caused by cardiovascular disease in males. It is thought that around 90 percent of cardiovascular diseases may be prevented.

Reducing risk factors for cardiovascular disease can be accomplished by adhering to a balanced diet and exercise routine, limiting exposure to secondhand smoke, and cutting back on the amount of alcohol consumed. Early detection and treatment, including counseling and, if necessary, medical intervention, are required for people who are already suffering from cardiovascular disease or who are at a high risk of developing cardiovascular disease due to risk factors such as hypertension, diabetes mellitus, and hyper lipidemia.

The inflammatory condition of the liver is referred to as hepatitis. It's possible that the condition may go away on its own, but it also has the potential to cause cirrhosis, liver cancer, and fibrosis (scarring). Although a viral infection is the most common cause of hepatitis, there are other possible triggers that might result in the condition.

These conditions include autoimmune hepatitis as well as hepatitis that is caused by medicines, narcotics, chemicals, or drinking alcohol.

Autoimmune hepatitis is essentially a medical disorder that develops as a result of the body's production of antibodies that are directed against the tissue of the liver. There are five primary types of hepatitis viruses, and they are designated by the letters A, B, C, D, and E, respectively. Every kind of hepatitis that may be passed on through the transmission of viruses is caused by a different virus.

Hepatitis A has always been an acute condition that only lasts for a brief length of time, but hepatitis B, C, and D have a greater likelihood of becoming chronic if they are not treated. Hepatitis E is an acute form of the disease; nonetheless, it poses a particularly grave threat to women who are pregnant. Rapid onset of signs and symptoms is characteristic of acute hepatitis.

Pain in the abdomen, a lack of appetite, lethargy, an unexpected loss of weight, signs of fever, black urine, and other symptoms are among the most typical of these signs and manifestations. Since chronic hepatitis develops over time and steadily worsens, its signs and symptoms may be too mild to notice.

In the year 2006, an information system was put into place with the objective of providing assistance to medical professionals throughout the process of making clinical decisions. The provision of the necessary information to those working in the healthcare industry was the primary purpose of the decision support system that was being discussed. The goal of providing these medical experts with this information was to assist them in their attempts to provide patients with a more effective therapy.

Specifically, the Support Vector Machine (SVM) and the Genetic Algorithm (GA) were the two distinct methods of machine learning that were utilized in a research study that utilized a combination of the aforementioned. Over the course of this inquiry, the investigators made use of the research software packages WEKA and LIBSVM. These accuracies were generated using the SVM and GA technique.

Thyroid illnesses were identified with the use of a machine learning method referred to as K-Nearest Neighbor (K-NN), which was employed in an experiment. A research study was conducted out in 2013 that made use of medical data for the goal of discovering widespread diseases by applying a range of DM techniques 44. This study was carried out in order to locate prevalent ailments.

The investigation of diagnostic microscopy (DM) techniques that have been used in the past to diagnose diseases including lung cancer, breast cancer, and cardiovascular abnormalities was the major purpose of this research effort. The genetic algorithm and fuzzy logic were the two machine learning approaches that were utilized in the process of developing a model for the diagnosis of diabetic Mellitus 45. This model was built via the usage of a blend of these two approaches. This model was brought into existence with the assistance of the software application MATLAB.

CONCLUSION

Machine learning is becoming an increasingly significant tool for modeling human processes in a range of sectors, including medicine, as a result of the volume of data that is available. Because the quality of the dataset has a significant impact on the precision of the machine learning algorithm, it is necessary to first collect and then have qualified medical professionals verify the correctness of the data. This dataset is used to train classifiers in order to improve their accuracy and make better predictions. The ability to forecast illnesses is becoming easier thanks to machine learning.

REFERENCES

1. Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 179, 104992.
2. Devarajan, M., & Ravi, L. (2018). Intelligent cyber-physical system for an efficient detection of Parkinson disease using fog computing. Multimedia Tools and Applications, 1-25.
3. Gokulnath, C. B., & Shantharajah, S. P. (2019). An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Computing, 22(6), 14777-14787.
4. Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2016). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE access, 4, 9145-9154
5. Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2016). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE access, 4, 9145-9154
6. Kirk, R. A., & Kirk, D. A. (2017, July). Introducing a Decision Making Framework to Help Users Detect, Evaluate, Assess, and Recommend (DEAR) Action Within Complex Sociotechnical Environments. In International Conference on Human Interface and the Management of Information (pp. 223-239). Springer, Cham. 35.
7. Masood, A., Sheng, B., Li, P., Hou, X., Wei, X., Qin, J., & Feng, D. (2018). Computerassisted decision support system in pulmonary cancer detection and stage classification on CT images. Journal of biomedical informatics, 79, 117-128.
8. Nilashi, M., Ibrahim, O., Samad, S., Ahmadi, H., Shahmoradi, L., & Akbari, E. (2019). An analytical method for measuring the Parkinson's disease progression: A case on a Parkinson's telemonitoring dataset. Measurement, 136, 545-557.
9. Yadav, S. S., & Jadhav, S. M. (2020). Detection of common risk factors for diagnosis of cardiac arrhythmia using machine learning algorithm. Expert Systems with Applications, 163, 113807
10. Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., & Lyu, C. (2019). A reliable method for colorectal cancer prediction

based on feature selection and support vector machine. Medical & biological engineering & computing, 57(4), 901-912. 30.