Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 11, No. 3, 2024

Clustering Analysis of Gas Wells in Carbonate Reservoirs

Yu Fu, Qinghui He

Petroleum Engineering School, Southwest Petroleum University, Chengdu, China

Abstract: Capacity clustering analysis of carbonate gas reservoirs can identify groups of wells with capacity characteristics. Differentiated production management strategies can be developed for different groups to improve overall production efficiency. In order to overcome the limitations of traditional capacity categorization methods that rely on human adjustments with the inability to meet field demands in real time. This study is based on the production capacity historical data and geological data of each block in the target gas reservoir. By comparing multiple capacity clustering machine learning methods, a machine learning model for carbonate rock capacity clustering analysis based on capacity test well data was established, with a correlation coefficient of more than 0.9. The results of this study show that the predicted capacity classes can truly reflect the distribution of gas reservoir capacity. This study identifies potential high-risk gas wells for the gas reservoir, so that measures can be taken in advance to reduce the risk and improve the level of safe production.

Keywords: Capacity Clustering Analysis; Carbonate Gas Reservoirs; Data Processing; Prediction Model.

1. Introduction

With the continuous growth of energy demand and the increasing depletion of traditional energy resources, the development and utilization of unconventional energy resources have attracted more and more attention. Among them, carbonate gas reservoirs, as an important unconventional natural gas resource with huge reserves and wide distribution, have received extensive attention and research from the global energy industry. However, the development and evaluation of carbonate gas reservoirs face a series of challenges due to their complex reservoir characteristics and difficult development.

In the development process of carbonate reservoirs, capacity clustering analysis, as an important evaluation technique, plays a crucial role in the capacity evaluation of gas wells, the assessment of gas reservoir potential and the optimization of production strategies. Currently, traditional capacity analysis methods mainly include single well capacity evaluation and overall reservoir analysis, etc. Although they are widely used in the field of reservoir capacity evaluation to a certain extent, they usually lack the detailed analysis of the heterogeneity of gas wells, which makes it difficult to formulate accurate development strategies for different types of gas wells. Therefore, to address the limitations of traditional capacity analysis methods, this study focuses on the carbonate rocks in Block X of the Sichuan Basin, and establishes a capacity clustering analysis model for carbonate reservoirs based on capacity data by combining the capacity test data of multiple gas wells, historical data, and geologic data, and applying the cluster analysis method for data processing.

The establishment of this cluster analysis model provides a reliable support for the evaluation of gas well production capacity, effectively evaluates the potential of the gas reservoir, and provides a scientific basis for the formulation of corresponding production management measures for the gas reservoir. The results of this study will be able to more accurately analyze the capacity clustering of gas reservoirs, provide more reliable technical support for the development and management of carbonate reservoirs, and promote the scientific, efficient and sustainable development of gas reservoir development.

2. Optimization of Capacity Clustering Methods

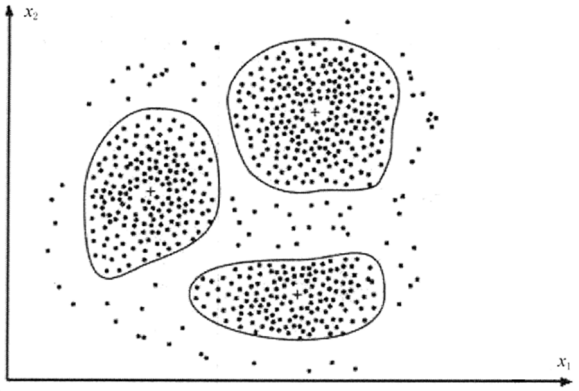

Figure 1. An example of clustering by distance between data objects

Clustering: Cluster analysis refers to the analytical process of grouping a collection of data objects into multiple classes composed of similar objects. According to a particular criterion (such as distance) to a data set divided into different classes or clusters, so that the similarity of data objects within the same cluster as large as possible, at the same time not in the same cluster of data objects as large as possible differences. After clustering the data of the same class are aggregated together as much as possible, and the data of different classes are separated as much as possible, which simply means that similar data are divided together, and the specific division does not care about the label of this class, and the goal is to aggregate similar data together, which is a kind of Unsupervised Learning (Unsupervised Learning) method, which is different from supervised learning (Supervised Learning), unlike Supervised Learning, the categorization or grouping information in the clusters that indicates the category of the data is absent. The similarity

between data is discriminated by defining a distance or similarity coefficient.[1]

Figure 1 shows an example of clustering by distance between data objects, where data objects with similar distances are classified into a cluster.

Clustering algorithms can automatically categorize unclassified data based on different distances between the data (e.g., Euclidean distance), which has higher accuracy and efficiency compared to manual classification. Currently commonly used clustering algorithms include BIRCH, DBSCAN, STING, SOM, FCM and K-means.[2]

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a comprehensive hierarchical clustering algorithm. It uses the concepts of Clustering Feature (CF) and Clustering Feature Tree (CF Tree) to generalize the clustering description. The Clustering Feature Tree summarizes the useful information for clustering and occupies much less space than a collection of metadata, which can be stored in memory, thus improving the speed and scalability of the algorithm for clustering on large data collections. The algorithm is able to perform clustering efficiently with one scan and can handle outliers effectively. Birch's algorithm is a distance-based hierarchical clustering that combines hierarchical coalescing and iterative relocation methods, first using bottom-up hierarchical algorithms, and then using iterative relocation to improve the results. Hierarchical coalescence uses a bottom-up strategy, where each object is first treated as a cluster of atoms, and then these clusters are merged to form larger clusters, reducing the number of clusters until all the objects are in a cluster or some end condition is satisfied. The main idea of the algorithm is to build a clustering feature tree initially stored in memory by scanning the database and then clustering the leaf nodes of the clustering feature tree.[3]

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a representative density-based clustering algorithm. DBSCAN is a density-based clustering algorithm that uses high-density connected regions to divide clusters. In this algorithm, clusters are characterized by the largest set of densely connected points and high-density regions. Unlike divisional and hierarchical clustering methods, it defines a cluster as the largest set of densely connected points, is able to classify regions with sufficiently high density as clusters, and can discover clusters of arbitrary shape in a noisy spatial database.[4]

STING (Statistical Information Grid) is a grid-based multi-resolution clustering technique that divides spatial regions into rectangular cells. For different levels of resolution, there are usually multiple levels of rectangular cells that form a hierarchy: each cell in the upper level is divided into multiple cells in the lower level. Statistical information about the properties of each grid cell (e.g., mean, maximum, and minimum values) is pre-computed and stored. These statistical variables can be easily used for the query processing described below. Statistical variables for higher-level cells can be easily computed from variables for lower-level cells. These statistical variables include: the attribute-independent variable count; the attribute-dependent variables m (mean), s (standard deviation), min (minimum), and max (maximum), and the type of distribution that the attribute values in the cell follow, e.g., normal, equilibrium, exponential, or none (if the distribution is unknown). When the data are loaded into the database, the variables count, m, s, min, and max are computed directly for the bottom cell. If the type of distribution is known in advance, the value of distribution can be specified by the user or obtained by hypothesis testing. After the hierarchy is established, the query processing time is O(g), where g is the number of bottom grid cells, which is usually much smaller than n. Since STING employs a multi-resolution approach to cluster analysis, the quality of STING clustering depends on the granularity of the bottom level of the grid structure. If the granularity is relatively fine, the cost of processing increases significantly; however, if the granularity of the lowest level of the grid structure is too coarse, it will reduce the quality of the cluster analysis. Moreover, STING does not take into account the relationship between a child cell and its neighboring cells when constructing a father cell. As a result, the shape of the resultant clusters is (isothetic), i.e., all the clustering boundaries are either horizontal or vertical, with no oblique dividing lines. Although this technique has fast processing speed, it may reduce the quality and accuracy of the clusters.[5]

SOM (Self-Organizing Maps), Self-Organizing Mapping Network, is a clustering algorithm based on neural networks. It is an artificial neural network developed by modeling the characteristics of the human brain for signal processing. After the model was proposed by Teuvo Kohonen, a professor at the University of Helsinki, Finland, in 1981, it has now become the most widely used self-organizing neural network method, in which the WTA (Winner Takes All) competition mechanism reflects the most fundamental features of self-organized learning. SOM networks can map any dimensional input pattern into a one- or two-dimensional graph at the output layer, and keep its topology unchanged; the network can make the weight vector space converge to the probability distribution of the input pattern through repeated learning of the input pattern, i.e., probability preservation. Neurons in the competitive layer of the network compete for the chance to respond to the input pattern, and each weight associated with the winning neuron is adjusted in a direction more favorable to its competition "i.e., with the winning neuron as the center of the circle, it exhibits excitatory lateral feedback to its nearest neighbors and inhibitory lateral feedback to its distant neighbors, with its near neighbors motivating each other and its distant neighbors inhibiting each other ". In general, a near neighbor is a neuron with a radius of about 50 μm to about 500 μm from the center of the neuron that sends the signal; a far neighbor is a neuron with a radius of about 200 μm to about 2 mm. Neurons farther away from the neurons than the distant neighbors show weak excitation, and since the curve of this interaction is similar to the hat worn by Mexicans, it is also called "Mexican hat". SOM algorithms have attracted a lot of attention for their properties such as topology preservation, probability distribution preservation, tutorless learning and visualization, etc. Various researches on the application of SOM algorithms have been conducted, and various studies on the application of SOM algorithms have been conducted. SOM algorithms have been widely used in speech recognition, image processing, classification and clustering, combinatorial optimization (e.g., the TSP problem), data analysis and prediction, and many other information processing fields. In short, the SOM algorithm has a wide range of applications, has a better development prospect, and is worthy of further research.[6]

FCM is fuzzy c-mean clustering algorithm, in many fuzzy clustering algorithms, FCM algorithm is the most widely used and more successful, it is through the optimization of the

objective function to obtain the affiliation of each sample point to all the class centers, to determine the class of the sample points to achieve the purpose of automatic classification of sample data. The idea is to maximize the similarity between objects classified into the same cluster and minimize the similarity between different clusters. Fuzzy C-mean algorithm is an improvement of the ordinary C-mean algorithm, which is hard for the division of data, while FCM is a flexible fuzzy division.[7]

K-means is a cluster analysis algorithm for iterative solution, the steps are, pre-dividing the data into K groups, then randomly selecting K objects as the initial cluster centers, then calculating the distance between each object and each seed cluster center, and assigning each object to the cluster center closest to it. The cluster centers and the objects assigned to them represent a cluster. For each sample assigned, the cluster centers of the clusters are recalculated based on the existing objects in the cluster. This process is repeated until some termination condition is met. The termination conditions can be that no (or a minimum number of) objects are reassigned to different clusters, no (or a minimum number of) cluster centers are changed again, and the sum of squared errors is locally minimized.

Based on the KNN algorithm we can see the difference between the data before and after processing, where the missing values have been automatically filled in.

Table 1. Analogy of different clustering methods

Clustering Method	Algorithm name	Merits	drawbacks
hierarchical clustering	BIRCH	Good interpretability	High time complexity
density clustering	DBSCAN	fast, insensitive to noise, and can find clusters of arbitrary shapes.	Parameters are difficult to control and have a large impact on the clustering effect
network clustering	STING	fast	Parameter sensitive, can't handle irregularly distributed data, dimensionality disaster
model clustering	SOM	in the form of a probability	Inefficient execution, especially when the number of distributions is large and the amount of data is small
fuzzy clustering	FCM	Works well for clustering data that satisfies a normal distribution	The performance of the algorithm depends on the initial clustering center
partition partitioning	K-means	Simple and efficient for large datasets, low time complexity and space complexity	Sensitive to noise and outliers

Comparing the advantages and disadvantages of different clustering algorithms, K-means algorithm is selected for clustering analysis in this study, which has the advantages of simplicity and efficiency, low time and space complexity, and the number of clusters can be formulated according to the needs of the clusters, with the highest analytical efficiency.

3. Clustering Results

K-means clustering is an unsupervised learning algorithm that divides a dataset into K clusters, and the process includes selecting the number of clusters K, randomly initializing K cluster centers, assigning data points to the nearest cluster based on their distances from the cluster centers, updating the cluster centers to be the mean value of the data points in the cluster and iterating the above steps repeatedly until the cluster centers are no longer changing significantly or until a preset number of iterations are reached to minimize the sum of squared distances of data points to their cluster centers. points to the sum of squares of the distances from the cluster centers to which they belong. In this study, the k-means clustering method was used to build a cluster analysis machine learning model based on historical yield data.

$$d\left(x_i,\mu_j\right)=\sqrt{\sum_{m=1}^{n}\left(x_{im}-\mu_{jm}\right)^2} \tag{1}$$

$$\mu_j=\frac{1}{\left|C_j\right|}\sum_{x_i\in C_j}x_i \tag{2}$$

$$J=\sum_{j=1}^{K}\sum_{x_i\in C_j}\left\|x_i-\mu_j\right\|^2 \tag{3}$$

Based on the Kmeans clustering algorithm, after taking the daily gas production, unimpeded flow rate, dynamic reserves, depth of the middle part of the producing layer, annual gas recovery, open well oil pressure, and water/gas ratio of each well as the final classification indexes for grading, the results of the classification results of the gas wells of the Gaomolengdeng Si reservoir obtained after the clustering analysis are as follows

According to the clustering results, the summarized three categories of wells have the following characteristics:

Class I wells: due to its high daily gas production, large unimpeded flow rate, high dynamic reserves and open well oil pressure, combined with the study of stable production influencing factors in the results IV, it has a strong ability to stabilize production, and it can be preferred as a medium- to long-term peaking and guaranteeing supply well.

Class II wells: with high daily gas production, large unimpeded flow rate, high dynamic reserve and open hole oil pressure, and strong stable production capacity, these wells can be preferred as short-term peak-peaking supply wells.

Class III wells: with low daily gas production, small non-resistance flow rate, low dynamic reserve and open well oil pressure, combined with the study of factors affecting production stabilization in Result 4, their production stabilization capacity is weak, and they can be used as auxiliary wells for peak shifting and supply maintenance, and appropriately adjusted downward to prolong the time of production stabilization after the end of peak shifting and supply maintenance.

Table 2. Capacity clustering results

WELL	Daily gas production (million cubic meters per day)	Unimpeded flow (million cubic meters per day)	Open well oil pressure (MPa)	Water-to-gas ratio (square/million cubic meters)	wells
WELL-1	13.27	76.09	17.66	0.16	I
WELL-2	37.50	123.83	30.88	0.16	I
WELL-3	36.86	160.78	27.81	0.16	I
WELL-4	30.21	127.20	34.54	0.16	I
WELL-5	35.11	133.95	35.31	0.16	I
WELL-6	13.70	94.53	19.13	0.16	I
WELL-7	25.46	73.91	22.29	0.11	I
WELL-8	36.98	89.77	31.43	0.16	I
WELL-9	24.93	135.89	30.54	0.16	I
WELL-10	13.19	30.41	33.01	0.16	III
WELL-11	14.49	39.05	33.14	0.16	III
WELL-12	9.37	13.96	11.11	0.16	III
WELL-13	19.90	48.67	26.23	0.16	III
WELL-14	14.29	43.65	12.89	0.16	III
WELL-15	19.72	44.01	24.18	0.16	III

4. Conclusion and Outlook

In this study, we applied the K-means clustering method to establish a cluster analysis model based on historical production data after comparing various clustering methods in machine learning. Through this method, we can effectively categorize and evaluate the production capacity of gas wells, thus providing a scientific basis for the development and management of gas reservoirs.

The results of the study show that the established cluster analysis model reaches 0.9 in terms of accuracy, which indicates that the method has high reliability and practicability in identifying and classifying the production capacity characteristics of gas wells. Through differentiated management of gas wells with different clusters, resource allocation can be optimized and production efficiency can be improved, thus promoting the scientific, efficient and sustainable development of gas reservoirs.

Looking ahead, further research can be carried out in the following aspects: first, more feature variables and data sources can be explored to improve the accuracy and robustness of the model; second, other machine learning algorithms, such as hierarchical clustering or DBSCAN, can be combined in order to compare and analyze the effects of different algorithms; finally, the research results can be applied to the development and management of other unconventional natural gas resources in order to validate the broad applicability and practical value of the method. Through continuous research and optimization, the K-means clustering method is expected to play a greater role in the development of unconventional energy resources and provide more reliable technical support for the sustainable development of the energy industry.

References

[1] Ricco AJ, Crooks RM, Osbourn GC. Surface Acoustic Wave Chemical Sensor Arrays: New Chemically Sensitive Interfaces Combined with Novel Cluster Analysis to Detect Volatile Organic Compounds and Mixtures[J], 1998.

[2] Villegas R,Salim A,Collins M,. Dietary patterns in middle-aged Irish men and women defined by cluster analysis[J], 2004.

[3] Camargo,Suzana J,Robertson,. Cluster Analysis of Typhoon Tracks. Part II: Large-Scale Circulation and ENSO. [J], 2007.

[4] Houser C,Masters G,Shearer P,. Shear and compressional velocity models of the mantle from cluster analysis of long-period waveforms[J], 2010.

[5] Zanchi AM. Artificial neural networks and cluster analysis in landslide susceptibility zonation[J].

[6] Mouridsen K,Christensen S,Gyldensted L,. Automatic selection of arterial input function using cluster analysis[J], 2010.

[7] Hu F,Hao Q,Bao K. A Survey on Software-Defined Network and OpenFlow: From Concept to Implementation[J], 2014.