A Decision Analysis Platform for Energy Big Data Based on Artificial Intelligence

Yuyan Wang*, Shujun Huang
Strategic Research Center of Oil and Gas, MNR, Beijing, China 100860

*Corresponding author: wangyy@sinooilgas.org.cn

Abstract. With the advent of the knowledge economy era, the application of energy big data has gradually emerged, and smart energy has begun to receive widespread attention. Decision analysis in the energy sector is fundamentally different in frequency, breadth, and complexity compared to the past. Traditional decision analysis techniques cannot meet the increasingly fierce market competition and corporate needs. Applying artificial intelligence technology into the energy field, realizing intelligent analysis and in-depth mining of energy big data, and providing better scientific decision-making have become a problem worthy of in-depth study in the new era. Through the analysis of the development of energy big data, this paper proposes the overall framework of an artificial intelligence-based energy big data analysis platform. We conducted analysis and evaluation from four aspects: data standardization, clustering analysis, classification analysis, and regression analysis. Through the discussion of data types and structures, we put forward a data simulation and evaluation plan, thus providing the energy resources provided in this article. The data decision analysis platform has more advantages in intelligent analysis capabilities.

1. Introduction
In recent years, technologies such as the Internet of Things (IoT), big data, and cloud computing have been continuously applied in power, oil, gas and other energy fields. The interaction, sharing and openness of data information has continued to accelerate, and the energy field has quietly entered the era of energy big data. As a basic strategic resource for enterprises, big data provides enterprises with comprehensive, accurate, and real-time business insights and decision-making guidance. Through the analysis and mining of the in-depth knowledge and value contained therein, it can provide decision support for various practical applications that are difficult to provide by other resources [1]. Big data is leading the arrival of data-intensive science, forming the fourth scientific paradigm after experimental science, theoretical science, and computational science [2], which is expected to promote the assumption-driven model of traditional science to data-intensive models based big data exploration.

With the complex environment and intensive competition of the energy sector, companies need to collect, store, and analyze massive amounts of data every day. Traditional decision-making models and ways of thinking are facing important challenges, and the intelligence and accuracy of decision-making analysis needs to be improved urgently. In the process of decision-making, the number of uncertain factors has increased sharply, the difficulty of decision analysis has increased, and hidden factors cannot be quantified [3]. Traditional decision-making methods are based on expert experience, small amount
of data analysis and individual case reasoning, which are far from meeting the individualized, diversified and complex decision-making needs of enterprises in the era of "energy big data".

In view of the above analysis, this article proposes to build a decision analysis platform for energy big data based on artificial intelligence technology, in order to improve the collection, storage, and analysis efficiency of energy data, and provide enterprises with scientific management and decision-making capabilities. Therefore, accelerating the research on artificial intelligence-based energy big data decision-making analysis platform not only has the significance of theoretical research, but also has practical application value.

2. Knowledge Preparation

2.1. K-means clustering algorithm

K-means algorithm is a partition-based clustering algorithm [4], and it is also one of the most important and widely used algorithms in unsupervised learning. Euclidean distance is generally used as an index to measure the similarity between objects. The greater the similarity, the smaller the distance. The core idea of K-means algorithm is to (1) randomly select k points from the data set as the initial cluster center \( C_i \) \( (1 \leq i \leq k) \), (2) classify them into the class with the smallest distance to the k centers, and (3) update the centroid of each clusters by averaging all samples. The sum of squared error (SSE) is used as the cost function, and the iteration is ended when the SSE converges. The SSE is calculated as follows:

\[
SSE = \sum_{i=1}^{k} \sum_{p \in C_i} ||p - m_i||^2
\]  

(1)

In the above formula, SSE measures the clustering quality, where \( k, m_i \) and \( p \) are the number of clusters, average vector of the ith cluster and the vector of the sample belonging to the ith cluster, respectively. However, the K-means algorithm has two shortcomings: it is sensitive to the initial value and easy to fall into the local optimal solution [5]. The current improvements to the k-means algorithm mainly focus on (1) the selection of the initial k value in the algorithm, (2) the initial clustering center point selection, (3) the detection and removal of outliers, and (4) the distance and similarity measurement.

2.2. Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a type of feedforward neural network (that includes convolution calculations and has a deep structure, which forms the basis of deep learning [6]. In the field of pattern classification, it effectively avoids complex data reconstruction and feature extraction processes, leading to its wide application.

Figure 1. The typical structure of CNN.
As shown in Figure 1, the overall architecture of a typical CNN follows the first half of stacked convolutional layers, several pooling layers in the middle to form a feature extractor, and finally a fully connected layer as a classifier to form an end-to-end network model [7]. The function of the convolutional layer is to use convolution operations to extract features. The more convolutional layers are, the stronger the expressive ability of features become. The pooling layer is based on the principle of local correlation for pooling sampling, thereby reducing the amount of data while retaining useful information. The structure of the fully connected layer is the same as the hidden layer structure of the fully connected neural network. Each neuron in the fully connected layer is connected to each neuron in the next layer. In order to make the feature displacement invariant, the activation function of the feature mapping structure generally chooses a nonlinear activation function, such as sigmoid and tanh functions.

2.3. Cloud and fog computing technology
With the widespread application of IoT in the energy field, the number of various IoT terminal devices has exploded, and the amount of raw data acquired is very large. Once the data transmission distance is large, it may cause I/O bottlenecks between the transmission terminal and the cloud server. This leads to a significant drop in transmission efficiency and even service interruption [8]. In view of the fact that a large number of terminal devices at the edge of the network require low-latency, location-aware, and mobile services, a service framework for fog computing is proposed, in order to extend the cloud computing network architecture to the edge of the network [9].

![Figure 2. IoT architecture based on cloud computing.](image)

As shown in Figure 2, we can expand a fog access point closer to the terminal device between the cloud and the IoT terminal device, in order to achieve data storage, data calculation, data analysis, seamless coverage, and air interface support. The key technologies of cloud and fog computing mainly include edge computing, edge storage, and cloud and fog collaboration. In addition, the key challenges lies in the decomposition of data tasks according to specific business scenarios, including the clarification of (1) what data is processed at the local node, (2) what data needs to be distributed to other fog access points for collaborative processing, and which data needs to be uploaded to the cloud computing node for processing.

3. The overall framework of the platform
According to the demand-oriented, application-driven design principle, the decision analysis platform realizes full coverage of energy companies, upstream and downstream enterprises of energy companies, government departments, scientific research institutions, and energy users. In order to realize the
panoramic intelligent analysis of energy big data, the platform needs to adhere to a unified standard, pay attention to practical application, and support the management of the whole life cycle of energy big data.

**Figure 3.** Overall architecture diagram of the decision analysis platform.

As shown in Figure 3, the overall framework of the decision analysis platform proposed in this paper mainly includes the source layer, the exchange layer, the storage layer, the support layer, the application layer, and the user layer. The source layer mainly includes structured data, semi-structured data, unstructured data, and IoT Internet raw data. Among them, the raw data of the IoT uses the intelligent edge processing of the fog node to complete the aggregation and collection. The exchange layer cleans and converts data from different sources, and completes the aggregation of various data. At the same time, relying on the data management platform and data scheduling platform to ensure the reliability and accuracy of data exchange and improve data quality. The storage layer completes the optimized storage of energy big data. According to different application types and scenarios, it is mainly divided into: basic big data area, theme data area, real-time data area, historical data area, streaming data area, application data area, test analysis data area, customized product data area, and public data release area. The supporting layer mainly includes various model algorithms and computing capabilities. At the same time, it provides various business services including data visualization services, data retrieval services, data integration services, and data analysis and prediction services to provide support for the upper application layer. The application layer mainly provides various information services for internal users and external users. For internal users, it provides internal business applications and business sand table deductions, helping users conduct business analysis for different scenarios and topics. On the other hand, it provides data product customization and data product customization and public data release for external users. The user layer is mainly for internal users, external users and IT personnel.

### 4. Platform implementation and evaluation analysis

In this section, we will discuss the implementation and the evaluation of the machine learning algorithms applied in the platform, which is shown in Figure 4.
4.1. Data Pre-processing
After the data collected at the source layer, we preprocess the data at the exchange layer, the main purpose of which is to convert the data into a structured format for downstream storage, analysis and visualization. We divide the data into two categories: continuous data and discrete data. For the former, we can directly perform z-score correction on it, that is, let each feature of these data obey the standard normal distribution, so that it is comparable in downstream analysis, that is, for feature i, we have $z_{score_i} = \frac{x_i - \mu_i}{\sigma_i}$. For the binary feature in the latter, we can directly set it to 1 or 0; and for multiclass feature, we can use one-hot encoding to split it into multiple binary classification features as one-hot encoding features.

4.2. Data Simulation
We use the python software package sklean [10] for data simulation and evaluation analysis. We first use the sklearn. datasets. Make_ classification function to perform data simulation, and the output result is a batch of high-dimensional data with a specific dimension, and we can set the internal distribution a priori by setting the number of clusters, severing as the ground truth for downstream benchmarks.

4.3. Unsupervised clustering and evaluation
We use the sklearn. cluster. KMeans function to perform unsupervised clustering on the standardized data. For each sample, we will finally get a class label. We use sklearn. Metrics. adjusted_ rand_ score to evaluate the predicted label and its real label. The larger the value, the closer the clustering result is to the real label. The calculation formula of adjusted rand score is as following, where $n_{ij}$ is the number of samples predicted to be label j whose real label is i, $a_i$ is the number of samples whose real label belongs to i, $b_j$ is the number of samples whose predicted label belongs to j.

$$ARI = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left( \sum_{i} \binom{a_i}{2} \right) \left( \sum_{j} \binom{b_j}{2} \right) / \binom{|U|}{2}}{\frac{1}{2} \left( \sum_{i} \binom{a_i}{2} \right) \left( \sum_{j} \binom{b_j}{2} \right) / \binom{|U|}{2} - \frac{1}{2} \sum_{i,j} \binom{n_{ij}}{2} - \left( \sum_{i} \binom{a_i}{2} \right) \left( \sum_{j} \binom{b_j}{2} \right) / \binom{|U|}{2} }$$

\hspace{1cm} (2)

4.4. Supervised classification and evaluation
Next, we use the pytorch software package to build a convolutional neural network, which can effectively extract valuable features and eliminate the correlation between features, so as to perform supervised classification of data to predict the discrete labels. We use cross-validation to evaluate the performance of CNN with the function sklearn. Model_ selection. Cross validate to randomly divide the data set for several times. In each time of split, a part of the division is used as the training set, and the remaining part is used as the test set without labels. Then we put the test set into the trained CNN to output a set of predicted labels. We finally use the sklearn. metrics. Classification_ report function to evaluate the consistency of the output label with its true label, including evaluating the precision and recall for each cluster, along with the overall accuracy for the entire dataset.
4.5. Supervised regression and evaluation
Finally, we use the above-mentioned CNN and change the structure of the final output layer to perform supervised regression on the data to be able to predict continuous labels. We also use cross-validation strategy to evaluate the performance as mentioned above. We finally evaluate the effect of the model by calculating Root Mean Square Deviation (RMSD) as following, where N, x_i and x̄_i are the number of samples, true value of the sample and the predicted value of the sample, respectively.

\[ RMSD = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x}_i)^2}{N}} \]  

(3)

5. Conclusion and discussion
Intelligent energy big data analysis is of great significance to all aspects of the energy field. Based on artificial intelligence technology, this paper proposes the overall architecture of the decision analysis platform for energy big data, which realizes the in-depth development and intelligent application of energy big data. Compared with traditional techniques, the intelligence, accuracy and authority of energy big data analysis have been greatly improved.

The energy field is a huge and complex system, and intelligent decision-making analysis is a complex systematic project. With the different complexity of decision-making issues, there are many challenges in the future, and there is still a lot of work to be further studied for the improvement of the decision analysis platform.

Reference
[1] Yang S L, Zhou K L. Management issues in Big Data: The resource-based view of Big Data [J]. Journal of Management sciences in China, 2015, 18 (5): 1-8.
[2] Tolle K M, Tansley Stewart, Hey Tony. The Fourth Paradigm: Data-Intensive Scientific Discovery. Proceedings of the IEEE, 2011, 99 (8): 1334-1337.
[3] Liu D R, Ke C K. Knowledge support for problem-solving in a production process: A hybrid of knowledge discovery and case-based reasoning [J]. Expert Systems with Applications, 2007, 33(1):147-161.
[4] A R C D A, B B M A. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering [J]. Pattern Recognition, 2012, 45(3):1061-1075.
[5] Xu Y, Shaw S L, Zhao Z, et al. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach [J]. Transportation, 2015, 42(4):625-646.
[6] Hao X, Zhang G, Ma S. Deep Learning [J]. International Journal of Semantic Computing, 2016, 10(03):417-439.
[7] Lecun Y, Bottou L. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.
[8] Chang H , Hari A , Mukherjee S , et al. Bringing the cloud to the edge[C]// IEEE Infocom -ieee Conference on Computer Communications Workshops. IEEE, 2014:346 - 351.
[9] Luan T H, GAO L, Li Z, et al. Fog Computing: Focusing on Mobile Users at the Edge [J]. Computer Science, 2015.
[10] Buitinck et al. API design for machine learning software: experiences from the scikit-learn project, arxiv, 2013.