EAI Endorsed Transactions
on Energy Web Research Article EAI.EU

Application of Probability Statistics of Set Algebraic Systems Based on Data Mining in the Energy Field

Jianxi Yu

Henan Institute of Economics and Trade, Zhengzhou, Henan, 450046, China

Abstract

As the global demand for energy continues to grow and the rapid development of renewable energy sources, the energy sector faces complex data processing and analysis challenges. This paper discusses the probabilistic and statistical application of set algebraic system based on data mining in the field of energy, uses data mining technology to effectively integrate multidimensional data such as energy consumption, production and distribution, and uses set algebraic system to build data models. Then, probabilistic statistical methods are used to analyze the energy data to identify potential patterns and trends. Evaluate the economic and environmental impacts of different energy technologies through case studies. The research shows that the set algebra system based on data mining can effectively improve the ability to analyze energy data and help identify the key drivers of energy consumption. At the same time, probability statistical analysis can predict the effects of different energy policies after implementation, providing data support for decision-making. The utilization rate of renewable energy significantly reduces carbon emissions after adopting this method. Therefore, the set algebra system based on data mining combined with probability statistics provides an innovative solution for the energy field, which can better data analysis and decision support, and promote the efficient use of energy and sustainable development.

Keywords: data mining; Set algebraic system; Probability statistics; Energy field

Received on 08 July 2024, accepted on 30 August 2024, published on 03 October 2024

doi: 10.4108/ew.6549

*Corresponding author. Email: Yjxxjh1314@163.com

1. Introduction

In recent years, the world is faced with the increasingly severe energy crisis and environmental pollution, and the sustainable use of energy has become the core issue of economic development and social progress of all countries. With the continuous advancement of technology, especially the rapid development of information technology and data science, the amount of data generated in the field of energy has increased dramatically. These data cover all aspects of energy production, consumption, storage, transmission and management, providing valuable resources for in-depth understanding of the operating mechanism of the energy system and optimizing resource allocation. However, the traditional data analysis method is often powerless when dealing with complex, multi-dimensional and high-dimensional data, and it is difficult to dig out the potential rules and patterns. At the same time, the rapid development of data mining technology provides a new idea for data analysis in the field of energy. By applying data mining techniques, researchers can extract valuable information from massive amounts of data and identify key factors that affect the efficiency and sustainability of energy systems. In addition, set algebraic systems are widely used in modeling and optimization of complex systems because of their unique advantages in expressing and dealing with uncertainty. In the field of energy, combining probabilistic and statistical methods of data mining and set algebra can not only improve the analytical ability of various energy data, but also help decision makers to effectively predict the effect and impact of policy implementation. This integration not only provides a scientific basis for energy demand forecasting, supply chain management, smart grid optimization and other issues, but also provides a new solution for the development of renewable energy and the transformation of traditional energy.

Set algebra expands its application scope on the basis of Boolean algebra, accurately characterizing complex phenomena in the real world, especially in the field of data mining [1]. It not only solves difficult problems, but also ensures the reliability, certainty, and interpretability of results, promoting the development of decision support systems. Probability theory includes probability logic and statistical probability, the former combining the advantages of probability theory and logic to deal with uncertainty problems; The latter explains probability phenomena from a frequency perspective [2]. The set algebra system can accurately reflect and simulate probability problems in various fields, and has important academic and practical value [3]. Signals are an indispensable element in daily life, and information can be obtained through signal acquisition and processing to achieve various functions [4]. Traditional digital signal processing often has redundancy, and the compression sensing theory adopts non adaptive linear measurement to reduce the number of sampling points, achieve high probability and accurate signal reconstruction, and break through the limitations of Nyquist sampling theorem [5]. The combination of data mining technology with databases and artificial intelligence has become a cutting-edge research direction, applied in fields such as information management, query response, and decision support. This article proposes a set algebra system that combines compressed sensing and data mining, applied to probability and statistics problems, to improve stability and robustness, and to address issues such as outliers and missing values, demonstrating significant advantages [6].

2. Related work

In the context of growing global energy demand and increasing awareness of environmental protection, the energy sector is facing serious challenges. In order to cope with these challenges, more and more researchers begin to pay attention to the application of data mining and statistical analysis in the field of energy, especially the probabilistic statistical method based on set algebraic system. The application of data mining in the field of energy has been paid more and more attention. Energy management involves large-scale, high-dimensional and complex data, which is difficult to deal with effectively by traditional data analysis methods. It is mentioned in the literature that data mining technologies, such as cluster analysis, classification algorithm and association rule mining, are widely used in energy consumption pattern recognition, load forecasting, anomaly detection and so on. Literature uses clustering algorithm to analyze residential electricity consumption data, successfully identifies user behavior patterns, and puts forward targeted energy-saving suggestions. This research not only optimizes energy management, but also improves the efficiency of resource use. The set algebraic system provides a flexible way to model the relationship between data, which makes the analysis of uncertainty and fuzziness more efficient. Literature shows that set algebra can help to deal with the complexity problems caused by multi-source data fusion. In the study of smart grid management, the set algebra method is applied to establish the dynamic model of energy flow, which improves the response speed and reliability of the system. In probabilistic statistics, the literature applies statistical methods to analyze and forecast energy demand and supply. Traditional statistical models often assume that data are independent and equally distributed, but in reality, energy data may show complex temporal dependence and spatial correlation. For this reason, more and more researchers have begun to try methods such as hybrid models and Bayesian statistics. The literature uses Bayesian networks to analyze the impact of existing energy policies on future energy demand, and the results show that the interaction between different policies significantly affects the accuracy of the forecast. This approach provides a scientific basis for energy policy making and promotes more sustainable energy development strategies.

Methods that combine data mining with probability statistics are also being used in the field of renewable energy. For example, many studies have explored the prediction of renewable energy sources such as wind and solar energy, using time series data mining technology to improve the accuracy of power generation prediction. Data mining method was used to analyze the relationship between historical meteorological data and power generation data, from which weather factors were found to have a significant impact on renewable energy power generation, and then a prediction model based on machine learning was proposed, which effectively improved the scheduling capability of renewable energy. Therefore, the probabilistic statistics of set algebraic system based on data mining is gradually applied in the field of energy, and relevant research results provide important support for energy management and policy making. However, current research still faces several challenges, including data quality and availability, model complexity, and the integration of different data sources.

The literature points out that set algebra systems have wide applications in various fields such as electronic circuits, navigation, and aviation [7]. These fields are often accompanied by time delay and noise interference, and considering time delay or random factors can form corresponding set algebraic systems. Due to the difficulty in obtaining theoretical solutions for most systems, the study of numerical methods is particularly crucial. Therefore, the study of stability and convergence of numerical methods has become an important topic in numerical analysis. The literature indicates that based on the data obtained in the previous stage, this stage is dedicated to deep processing it to construct a dataset suitable for data mining [8]. This process involves two core steps, and through various measures, we have laid a solid foundation for subsequent data mining work. The literature points out that traditional information acquisition models often adopt a sampling followed by compression approach [9]. In this process, only the relevant numbers with larger absolute values are written, so that coefficients with zero are equal or similar. This approach actually treats a lot of valuable information as useless data, leading to the loss of information resources. This not only increases the burden of storage space, but also extends the time for collecting samples, thereby restricting the development of the information processing field. In order to overcome this limitation, we urgently need to explore new information processing methods to more effectively utilize information and promote rapid progress in the field of information processing. The literature indicates that reconstruction algorithms play a crucial role in sensing technology [10]. Not being able to carefully select the appropriate reconstruction algorithm may result in us being unable to accurately reconstruct the signal within the allowable error range. In algorithm design, two crucial factors are the precise reconstruction probability of compressed signals and the ease of calculation. The key to measuring algorithm performance is the precise reconstruction probability [11]. However, when using compressed sensing technology to reconstruct the original signal, the computational complexity is too high, resulting in a lengthy reconstruction process, which is not accepted in practical applications [12]. Therefore, the current research focus is on how to improve the efficiency and accuracy of signal reconstruction while reducing computational complexity. The literature points out that the NP hard problem cannot be quickly solved through polynomial time complexity algorithms due to its inherent difficulty [13]. When we face specific challenges, the calculation of determining the remaining probabilities is often more complex than decision-making problems. However, in an environment like Maple, if the number of values can make the algorithm very clear, the processing of probability reconstruction and table verification tasks can usually be carried out with linear time efficiency. This highlights the significant advantages brought about by decomposing the probability space into classes. Literature data mining is a complex process that involves multiple stages and requires repeated processing [14]. Each stage requires the collaboration of experts, data analysts, and other professionals. During the process, it is often necessary to make cyclic adjustments until satisfactory results are achieved. Its core is to mine knowledge patterns from raw data, with the goal of foresight and description. Predicting future values and describing them reveals interpretable data patterns [15].

3. Compression sensing technology

3.1. Basic Theory

It is essential to conduct in-depth analysis of signal characteristics and establish appropriate mathematical models before processing signals. In nature, analog signals often need to be converted into discrete digital signals through sampling processes. Therefore, we often model one-dimensional signals as vectors and two-dimensional signals as matrices. When constructing a mathematical model for compressed sensing, in order to simplify processing, we assume that the signal is a sequence of discrete values obtained based on the Nyquist sampling frequency, which can be represented by a one-dimensional vector. In this way, for signal $x(x\epsilon R^n)$, the compression sensing theory describes its measurement process

In most cases, signals in nature often do not have sparsity, so we need to transform signals into sparse signals through domain transformation, namely:

$x = \Psi\alpha(1)$

The measurement of signals can be done using the formula:

$y = \Phi\Psi\alpha = \Theta\alpha(2)$

The compressed sensing technique identifies the non-zero element in the vector α by the m measurements, thus restoring the original signal. The whole process of signal processing by compressed sensing is shown in Figure 1.

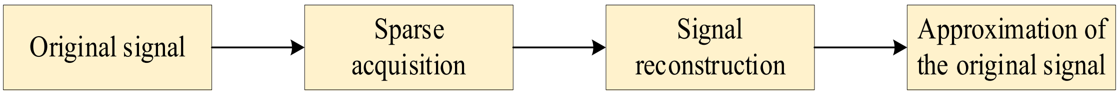

Figure 1. Compression sensing signal processing flow

Compression sensing theory mainly consists of three components: sparse representation, measurement matrix, and reconstruction algorithm.

3.2. Algorithm design

Facing the problem of wireless link, this paper proposes a new compressed sensing data collection algorithm based on sparse block observation matrix (CS-SBM). In this algorithm, the sparse block observation matrix operates within the cluster, and the member nodes determine whether they participate in the current round of data collection according to the non-zero states of their stored sparse observation vector elements. If the packet loss occurs in the data transmission and the data received by the cluster head is incomplete, the MC (Matrix Completion) theory can be used to recover the lost data.

However, the observation vector Y= ΦX obtained by directly using the observation vector Φ does not have the characteristics of low rank or approximate low rank, and there is no way to directly recover the observation value Y' of the loss with MC theory. To this end, this paper proposes a sparse block observation (SBM) matrix to observe the original data vector X. The matrix is characterized by only one non-zero element 1 per row, which reduces the energy loss of each round of data transmission, because only one node of data is sent per round. At the same time, the observation matrix still maintains the approximate low rank feature, making the data processing more efficient and accurate.

The matrix form in which the data vector is represented as m×n is:

$$X_{mn} = \begin{pmatrix} x_{11} & \cdots & & \cdots & \cdots & x_{in} \\ \vdots & x_{ij} & & \cdots & x_{i(j+q_k-1)} & \vdots \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ \vdots & x_{(i+p_k-1)} & \cdots & x_{(i+p_k-1)(j+q_k-1)} & & \vdots \\ x_{m1} & \cdots & & \cdots & \cdots & x_{mn} \end{pmatrix}_{m\times n} \quad (3)$$

Where m×n=N_1, its block matrix is:

$$X'_{p_k\times q_k} = \begin{pmatrix} x_{ij} & \cdots & x_{i(j+q_k-1)} \\ \vdots & \ddots & \vdots \\ x_{(i+p_k-1)1} & \cdots & x_{(i+p_k-1)(j+q_k-1)} \end{pmatrix}_{p_k\times q_k} \quad (4)$$

X exhibits the characteristic of approximate low rank, and its corresponding block matrix X' also possesses this property. In order to maintain the corresponding properties of the observation matrix, we constructed a sparse block observation matrix so that X' can be presented in the matrix form of the observation vector Y. In this way, we can effectively restore matrix X' using the MC principle. Therefore, the SBM matrix is defined as:

$$\Phi_E(h,g) = \begin{cases} 1, g = \Omega_h \\ 0, \text{others} \end{cases} (5)$$

3.3. Experimental Results

In order to comprehensively demonstrate the excellent functions of the algorithm proposed in this paper, the signal reconstruction probability was used as an evaluation indicator to compare the OMP (Orthogonal Matching Pursuit) algorithm with the algorithm proposed in this paper in detail. This comparison was made under different observation conditions of M. By comparison, we can more intuitively see the advantages of our algorithm in signal reconstruction. The specific results are shown in Figure 2.

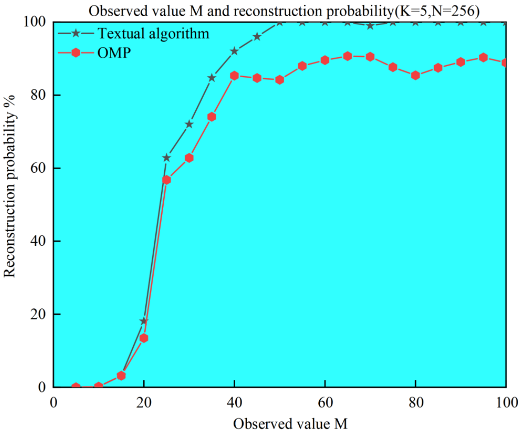

Figure 2. Compared the reconstruction probabilities of two algorithms under different observation values

Observing Figure 2, we can observe that when the observation value M is small, the reconstruction probabilities of the two algorithms are almost the same. However, as the observed value M gradually increases, the reconstruction probability of the algorithm proposed in this paper gradually surpasses that of the OMP algorithm, and the reconstruction probability of the algorithm proposed in this paper shows relatively stable performance. In addition, this comparison helps to comprehensively evaluate the performance characteristics of the two algorithms. The specific results are shown in Figure 3.

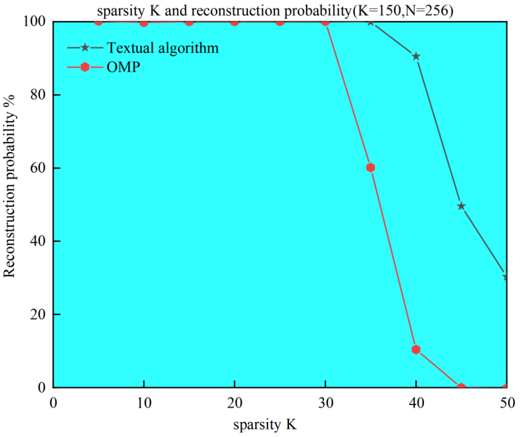

Figure 3. Comparison of reconstruction probabilities between two algorithms under different sparsity levels

It can be clearly seen from the graph that as the sparsity K increases, the reconstruction probabilities of both algorithms show a decreasing trend. However, as the sparsity K gradually increases, the reconstruction probability of our algorithm is always higher than that of the OMP algorithm. Specifically, when the sparsity K is 45, the reconstruction probability of our algorithm reaches 50%, while the reconstruction probability of the OMP algorithm is 0. Furthermore, it is worth noting that in the case of sparsity K not exceeding 30, the reconstruction probabilities of both algorithms have reached their maximum value, which is 100%.

Figure 4 illustrates the relationship between the number of observations M and the normalized mean absolute error (NMAE) when p=0.4.

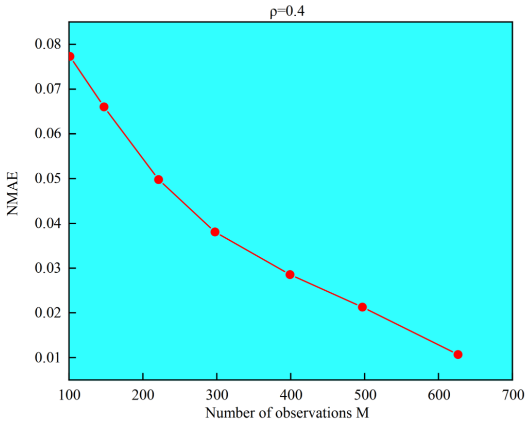

Figure 4. The relationship between observation frequency and reconstruction accuracy

Even when facing challenges, the algorithm proposed in this paper can still effectively improve the reconstruction accuracy by increasing the number of observations, thereby significantly reducing the adverse impact of packet loss on data reconstruction accuracy.

4. Data mining algorithms

4.1. Algorithm design

Given the differences in source and semantic information among modal data, we need to design specific feature extraction methods for each modality to accurately capture its unique representation. Next, we will take the integration scenario of three sparse modalities as an example to elaborate in detail. Imagine a spatiotemporal prediction task that includes data from three modalities: $x \epsilon R^{R \times C \times Tx}$, $Y \epsilon R^{R \times C \times TY}$, and $z \epsilon R^{R \times C \times TZ}$. For these three modalities, we have designed corresponding feature extractors to extract their key information. The data distribution obtained after processing by these feature extractors is defined as:

$$p_{\theta_X}(X^e) = \int_X q(X^e|X, \theta_X)p(X)dX \, ; p_{\theta_Y}(Y^e) = \int_Y q(Y^e|Y, \theta_Y)p(Y)dY \, ; p_{\theta_Z}(Z^e) = \int_Z q(Z^e|Z, \theta_Z)p(Z)dZ \quad (6)$$

Where $q(X^e|X, \theta_X)$ represents that the feature extractor maps input data X to the hidden space Xe function, θ_X is the core parameter, and the hidden space $P_{\theta X}(X^e)$ reveals the feature distribution characteristics. For information X, Y, and Z, the respective feature extractors are used to obtain feature representations such as X^e, Y^e, and Z^e. These extractors are based on deep neural networks. Compared with other modes, the corresponding representations proposed are all used in the same channel dimension.

Although single-modal representation can be directly applied to prediction tasks in nature, it may be accompanied by semantic bias and semantic insufficiency. Especially sparse mode, because of the limited amount of data, its representation often lacks sufficient semantics. Therefore, it is necessary to make up the semantic deficiency through the interaction of features between modes. In this paper, we aim to explore the interaction between modes by integrating high-dimensional representations of all single modes. Inspired by the idea of graph fusion, we design a multi-layer feature interaction module (MFI), which is similar to the structure of hierarchical neural networks, and can successively simulate the interaction process between single, two-mode and three-mode modes. The MFI module is composed of single-mode, two-mode and three-mode dynamic learning layers. It treats each interaction as a vertex and sets the weight of the edge according to the similarity between the interactive vertices and the importance of the vertices.

In the first layer, the single-modal dynamic learning layer, we introduce the single-modal vertices of three modes, whose information vectors are represented as Ψ_X, Ψ_Y, and Ψ_Z, respectively, which correspond to their respective high-dimensional representations X^e, Y^e, and Z^e. In this layer, we treat each vertex with a modal attention network to highlight the importance of different modes in the interaction process. By calculating the weighted average of the information of all single-mode vertices, we obtain the final single-mode dynamic representation, where the weight of each vertex is determined according to its contribution during the interaction, and the weight of each single-mode vertex is defined as:

$$A_\phi = MAN(\Psi_\phi; \theta_{MAN}) \quad (7)$$

In modal attention networks, θ_{MAN} as a parameter. And these weights, the final single modal vector is:

$$\Delta_U = \frac{1}{3} \Sigma_\phi A_\phi \cdot \Psi_\phi \quad (8)$$

The second layer is a bimodal dynamic learning layer. In this layer, we adopted a multi-layer neural fusion network (MLP) to fuse every two unimodal vertices, thereby generating corresponding bimodal vertices:

$$\Psi_{\phi_1 \phi_2} = MLP(\Psi_{\phi_1} \oplus \Psi_{\phi_2}; \theta_{MLP}) \quad (9)$$

Regarding the weight setting of the edges connecting the first and second layers, we first use inner product to estimate the similarity of every two unimodal information vectors in the first layer. Based on an assumption: if two information vectors are similar, there is relatively little complementary information between them, and this information has been fully explored in the first layer. Therefore, we infer that the higher the similarity between two information vectors, the lower the interaction importance between their corresponding bimodals. Here, we define the similarity between two information vectors as:

$$\mathrm{Sim}(\phi_1,\phi_2) = \widetilde{\Psi}_{\phi_1}^{T}\widetilde{\Psi}_{\phi_2} \quad (10)$$

Based on similarity estimation, we further define the weight of the second layer vertex $\Phi_1\Phi_2$, the formula is as follows:

$$\widehat{A}_{\phi_1,\phi_2} = \frac{A_{\phi_1}+A_{\phi_2}}{\mathrm{Sim}(\phi_1,\phi_2)+0.5} \quad (11)$$

We construct a bimodal dynamic network, which fuses bimodal vertices into three modal vertices to form a three modal dynamic learning layer. The MLP network here does not share parameters with the second layer. At the same time, the specific bimodal vertices are fused with the unimodal vertices that are not involved in their formation, resulting in an additional three modal vertices. Therefore, there are six three modal vertices in the learning layer. We use a unified weight calculation method and the same importance measurement in the second layer to summarize the weight information of each three modal vertex and get the final three modal information. Importance measurement method. Then we add the weighted information from each three modal vertex to obtain the final three modal information Δ_T.

4.2. Experimental Analysis and Results

Firstly, we constructed an undirected network with 200 nodes and two equally sized nodes using a random module model. Among them, half of the data that matches the community is randomly allocated to obtain the adjacency matrix and metadata information of the network data.

Subsequently, we combined the essential information of adjacency matrix and metadata, applied the data mining model proposed in this paper, and estimated the model parameters to ultimately understand the results of community partitioning. Statistical inference algorithms are very stable and not easily affected by data interference. Even if the data is incomplete or has issues, it can still effectively identify the communities in the data. Moreover, this algorithm is very powerful, even if half of the metadata of nodes in the network is random, it will not affect its effectiveness in conducting community testing. It is worth noting that although there is randomly assigned metadata in this network, it cannot have a significant impact on the final community detection performance. Figure 5 shows the approximate process of the proposed data mining model in artificial data.

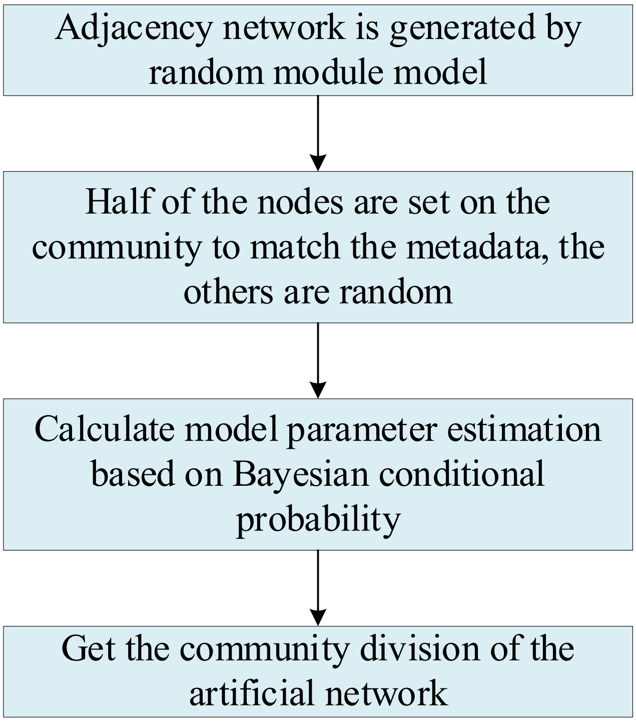
Adjacency network is generated by random module model
Half of the nodes are set on the community to match the metadata, the others are random
Calculate model parameter estimation based on Bayesian conditional probability
Get the community division of the artificial network

Figure 5. The application process of the model in artificial data in this article

The data mining model constructed in this article is applied to artificial networks, and the resulting data first lists the number of each node, followed by providing metadata information of the nodes. In addition, the results also show the probability values of nodes being assigned to the first and second communities. It is worth noting that the community detection results in this article are not simply classified, but rather use a probability form to reflect the possibility of nodes belonging to different communities. This processing method also has potential application value for the study of overlapping communities and can be one of the future research directions. Table 1 shows the performance indicators of community detection results obtained from the model in this article.

Table 1. Regarding the community detection results of the model in this article, community indicators

Classification of clubs	Club 1	Club 2
Accuracy rate	0.9028	0.7718
Recall rate	0.5889	0.8817
F1 metric	0.7489	0.8231

From these data, it can be seen that the algorithm has high accuracy in identifying nodes belonging to the community. These results fully demonstrate that the data mining model proposed in this article demonstrates excellent performance in computer synthesized data, effectively extracting structural information from artificial networks. This article observes the changes in runtime of the OMP algorithm and our algorithm, as shown in Figure 6.

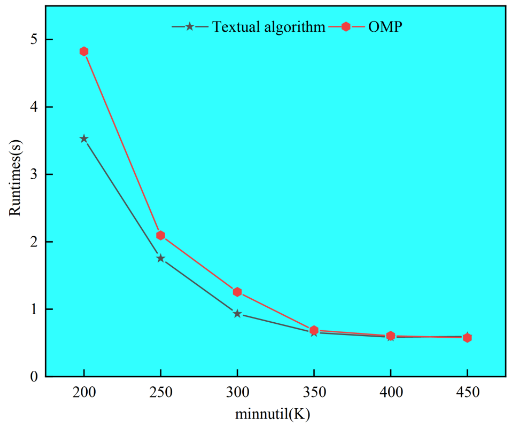

Figure 6. Run time

From Figure 6, it can be observed that in all datasets, the running time consumed by the algorithm in this paper is less than that of the OMP algorithm, and the startup time of the algorithm in this paper does not show a significant change with the variation of minutil, indicating that the algorithm is less sensitive to minutil.

Observations have shown that our algorithm outperforms the OMP algorithm in handling dense datasets. Because the algorithm in this article adopts an extended branch pruning strategy, which prunes the search space by using the utility upper limit of the extended branch itemset, abandoning the inefficient construction of linked lists, making the algorithm more efficient in data processing. Therefore, the algorithm in this article runs faster on dense datasets. As shown in Figure 7, a comparison was made between the algorithm proposed in this paper and the OMP algorithm in terms of memory wear.

As shown in Figure 7, the memory loss of our algorithm and OMP algorithm was compared.

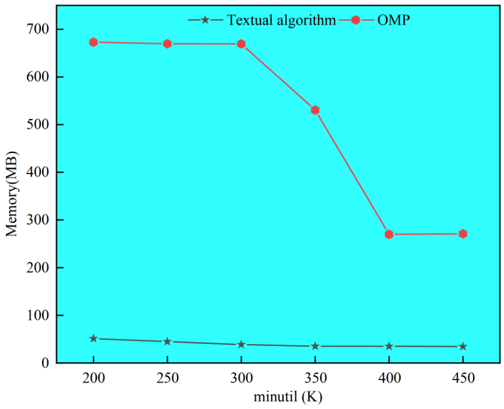

Figure 7. Memory consumption

Figure 7 shows the comparison between our algorithm and OMP algorithm in terms of memory consumption. On all datasets, our algorithm exhibits lower memory consumption. This is mainly due to a series of optimization measures adopted by the algorithm. Firstly, by using linked list overlay technology to reduce duplicate data in the linked list, the memory usage during the initial utility linked list construction is reduced. Secondly, utilizing a list buffering structure to efficiently utilize memory resources and optimize the storage of utility linked list information. In addition, the priority filtering strategy reduces the construction of inefficient linked lists, thereby reducing the number of utility linked lists and memory consumption. Finally, the extended branch pruning strategy prunes the search space based on the actual utility upper limit, effectively reducing memory usage. These innovative designs together make our algorithm superior to the OMP algorithm in terms of memory consumption.

5. Probability and Statistical Observations of Set Algebraic Systems

5.1. Preparatory knowledge

In order to discuss the relevant content of non smoothness analysis, we first need to clarify some basic terms and nature.

We define the Euclidean projection from x to Ω. If Ω is a bounded closed set, then any x∈R^n, set Π(x;Ω) are all non empty. Using Euclidean projection to define a normal cone on a finite dimensional space:

$$N(\bar{x};\Omega) \triangleq \limsup_{x\to\bar{x}}[\mathrm{cone}(x-\prod(x,\Omega))](12)$$

The formation of the subdifferential ∂Φ(x) depends on the specific normal cone and the image characteristics of the function Φ. Here, we set Φ as a finite real-valued function and introduce the corresponding cone concept. The original intention of this introduction comes from the in-depth study of the optimal control problem with terminal geometric constraints. In deriving the necessary conditions for this kind of problem, we find that the problem can be transformed into a more manageable form of free terminal control when the limit treatment is taken.

Notably, this cone is non-convex, and the convex envelope of the subdifferential of a locally Lipschitz continuous function is just the Clarke generalized subdifferential of that function. If Φ_k is lower half constant near point x, here:

$$\hat{\partial}\phi(x^0) \triangleq \{x^{**} \in R^n \left| \liminf_{u \to x^0} \frac{\phi(u)-\phi(x)-\langle x^{**},u-x\rangle}{|u-x|} \geq 0\right\} (13)$$

Represents the Frechet subdifferential. Using the positive and negative symmetric structure method, we have:

$$\partial^+\phi(x^0) \triangleq -\partial(-\phi)(x^0) (14)$$

It should be noted that for a few Lipschitz functions, the basic subdifferential and the Frechet subdifferential are probably not the same.

5.2. Set algebraic system model

Variable delay differential algebraic systems are widely used in many fields of science and engineering, such as computer aided design, circuit analysis, mechanical systems, chemical reaction simulation and real-time simulation of automatic control systems, so it is of great theoretical significance and practical value to study such problems. However, such equations have both delay terms and algebraic constraints, which makes the analysis process extremely complicated. Therefore, this paper mainly discusses a class of variable delay differential algebraic systems with specific indexes. Let x '(t) be the inner product in the space CM and ‖·‖ be the norm derived from that inner product.

The upper bound is obtained for every x,u∈C^M, y∈C^N, and for example L1,L2,L3 are fixed and invariant numbers of appropriate size. In addition, another step assumes that f satisfies the following conditions:

$$\mathrm{Re}\langle x_1 - x_2, f(x_1 - u_1, y) - f(x_2 - u_2, y)\rangle = \alpha\|x_1 - x_2\|^2 + \beta\|u_1 - u_2\|^2 (15)$$

$$\|f(x,u,y_1) - f(x,u,y_2)\| \leq \gamma\|y_1 - y_2\|, x, u \in C^M, y_1, y_2 C^N (16)$$

5.3. Probability Statistics of Set Algebraic Systems

According to the finite nature of storage tapes, automata can be divided into two categories: finite band automata and infinite band automata. Among them, finite automata play a crucial role in modeling discrete input-output systems due to their limited control states and symbol sets. In practical applications, finite automata are further divided into Moore type finite state machines and Mealy type finite state machines based on whether input signals are utilized or not. A notable feature of the Moore type finite state machine is that its output signal is influenced by the current state, meaning that its output is actually a function of the current state. As a form of deterministic finite automaton, i.e. a 5-tuple:

$$A = (Q, \sum, \delta, h, q_0) (17)$$

Finite automata can be regarded as a device composed of an input band and a controller. The input strip is subdivided into multiple squares, each of which is used to store a specific symbol, all of which come from a limited set of symbols Σ. The controller has a limited number of potential states, which constitute the set Q. The controller is equipped with a read in head to capture symbols from the input band. It is worth noting that the concept of time here is discrete, and at system startup, the controller is initialized to state q_0. Its core function is to determine the next state based on the current state q and the symbol a captured from the input by the read in header, and to achieve the transition from state q to state q':

$$q' = f(q, a) (18)$$

And move the reading head one grid to the right. As shown in Figure 8.

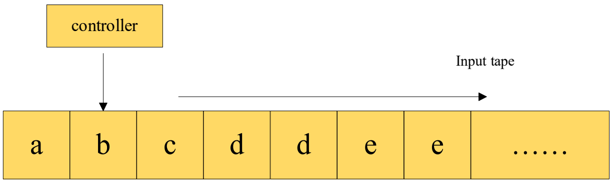

Figure 8. Finite automaton

The Moore machine identifies its condition by conveying letters, and it is not very attentive to accepting requests, so it is not used to accept or reject input, but to perform calculations. For special input sequences, Moore's machines can perform accurate calculations like finite deterministic automata, and their output results are sequentially accessed state identification series. When a Moore's machine has countless states, it can be represented by a list containing two aspects: one is a square table derived from natural number pairs, used to describe the transition function, i.e. m exists in (i, j); The other part is a list of state identifiers, arranged in the order of the states in the transition table, where the first state is the initial state. The status identifier has three values: 0 (false), 1 (true), and 2 (undetermined).

In computer algebra systems, a probability table is a complex data structure that has a unique function of extracting and deriving probability values that are not directly stored but can be calculated from existing data. This means that users only need to input some initial data, and the system can automatically calculate the remaining probability. Taking n independent original events as an example, users only need to input 2^n probabilities, and other probabilities can be automatically calculated as needed. If this feature is not available, users will need to manually input all 2^n probability values. This table is based on the principle of independence between events and allows for reconstruction of unspecified values. The atoms of a free Boolean algebra, as the index of the original event, can accurately define a discrete probability space, where the measurable element field is the Boolean algebra. A probability space can be decomposed into the product of multiple small probability spaces, which we call a class. The characteristics of each class are determined by the probability assignment method of its atoms, and elements with zero probability do not need to be specifically specified.

There are two strategies for dealing with unclear probabilities:

(1) Class level: The probability of missing is usually considered zero;

(2) Probability space level: The probability associated with different atomic events should remain in an unspecified state, and these probabilities can be reconstructed under the premise of inter class independence.

For example, event a, let its probability be:

$$P(a) = u, P(a') = 1 - u(19)$$

Note that probability is a numerical value that exists in symbolic form here. Based on the previous explanation, we may need to artificially assign probabilities to event a and its complement a', despite the existence of certain correlations between them. Although we can derive the probability of a', this derivation may bring some problems: (1) a large amount of calculations; (2) It is difficult to combine another more fundamental and easily implementable extension method, so we set it to zero for all unspecified probabilities.

So the system can infer the probability that the input is not yet clear:

$$P(b \wedge a) = P(b)P(a) = [P(b \wedge c) + P(b \wedge c')]P(a) = \frac{1}{4}u(20)$$

This article discusses in detail the calculation problem of probability statistics in set algebra systems. By implementing the conversion process from Moore's machine to probability table, we have verified the feasibility and enormous potential of the proposed method. Especially for large-scale set algebra problems, using this computational method can fully simplify the process, improve its efficiency, and demonstrate broad application prospects and important theoretical value.

6. Conclusion

With the continuous increase of global energy demand and the promotion of sustainable development goals, the application of data mining based set algebra system and probability statistics in the energy field is particularly important. Data mining has shown its powerful ability in the identification and analysis of energy consumption patterns, helping managers to better understand user behavior to formulate effective energy conservation strategies. At the same time, combined with the ensemble algebra system, it can process multi-dimensional and multi-source energy data, and enhance the real-time analysis ability of energy flow and equipment operating state. This is important for improving the reliability and efficiency of the energy system. The application of probability statistical method brings a new perspective for energy demand forecast. By using sophisticated statistical models, researchers are able to accurately assess uncertainty and analyze the impact of policy changes on energy demand, thereby providing a scientific basis for policy formulation. In addition, for the prediction and management of renewable energy, the combination of data mining technology and probabilistic models makes the scheduling of wind and solar power generation more accurate, and further promotes the utilization efficiency of renewable energy. Therefore, the probability statistics of set algebraic system based on data mining has shown a broad prospect in the field of energy, and has great practical value and research significance. As the technology develops and its application deepens, it is expected to provide strong support for building more efficient, reliable and sustainable energy systems.

References

[1] B. Lowe, S. Tarafder, Generalized algebra-valued models of set theory. The Review of Symbolic Logic 8(1) (2015) 192-205.

[2] P. Janotta, H. Hinrichsen, Generalized probability theories: what determines the structure of quantum theory?. Journal of Physics A: Mathematical and Theoretical 47(32) (2014) 323001.

[3] Y. Wang, A denotational mathematical theory of system science: system algebra for formal system modeling and manipulations. Journal of Advanced Mathematics and Applications 4(2) (2015) 132-157.

[4] S. L. Oh, Y. Hagiwara, U. Raghavendra, R. Yuvaraj, N. Arunkumar, M. Murugappan, U. R. Acharya, A deep learning approach for Parkinson’s disease diagnosis from EEG signals. Neural Computing and Applications 32 (2020) 10927-10933.

[5] Y. Sun, C. Cui, J. Lu, Q. Wang, Data compression and reconstruction of smart grid customers based on compressed sensing theory. International Journal of Electrical Power & Energy Systems 83 (2016) 21-25.

[6] A. M. Medina-Mardones, A computer algebra system for the study of commutativity up to coherent homotopies. Tbilisi Mathematical Journal 14(4) (2021) 147-157.

[7] C. Bright, I. Kotsireas, V. Ganesh, Applying computer algebra systems with SAT solvers to the Williamson conjecture. Journal of Symbolic Computation 100 (2020) 187-209.

[8] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, K. Crawford, Datasheets for datasets. Communications of the ACM 64(12) (2021) 86-92.

[9] S. Sun, J. Gong, A. Y. Zomaya, A. Wu, A distributed incremental information acquisition model for large-scale text data. Cluster Computing 22 (2019) 2383-2394.

[10] H. Kim, C. M. Park, M. Lee, et al., Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra-and inter-reader variability and inter-reconstruction algorithm variability. PloS one 11(10) (2016) e0164924.

[11] J. An, S. Cho, Variational autoencoder based anomaly detection using reconstruction probability. Special lecture on IE 2(1) (2015) 1-18.

[12] Y. Bao, H. Li, J. Ou, Emerging data technology in structural health monitoring: compressive sensing technology. Journal of Civil Structural Health Monitoring 4 (2014) 77-90.

[13] W. Li, Y. Ding, Y. Yang, R. S. Sherratt, J. H. Park, J. Wang, Parameterized algorithms of fundamental NP-hard problems: A survey. Human-centric Computing and Information Sciences 10 (2020) 1-24.

[14] V. Plotnikova, M. Dumas, F. Milani, Adaptations of data mining methodologies: a systematic literature review. PeerJ Computer Science 6 (2020) e267.

[15] A. Ragab, M. El Koujok, H. Ghezzaz, M. Amazouz, M. S. Ouali, S. Yacout, Deep understanding in industrial processes by complementing human expertise with interpretable patterns of machine learning. Expert Systems with Applications 122 (2019) 388-405.