Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4020723, 10 pages
https://doi.org/10.1155/2020/4020723

Research Article

An Improved HotSpot Algorithm and Its Application to Sandstorm Data in Inner Mongolia

Ren Qing-dao-er-ji, Rui Pang, and Yue Chang

School of Information Engineering, Inner Mongolia University of Technology, Hohhot 010051, China

Correspondence should be addressed to Ren Qing-dao-er-ji; renqingln@imut.edu.cn

Received 27 September 2019; Revised 15 December 2019; Accepted 3 March 2020; Published 10 April 2020

Academic Editor: Sergio Ortobelli

HotSpot is an algorithm that can directly mine association rules from real data. Aiming at the problem that the support threshold in the algorithm cannot be set accurately according to the actual scale of the dataset and needs to be set artificially according to experience, this paper proposes a dynamic optimization algorithm with minimum support threshold setting: S_HotSpot algorithm. The algorithm combines simulated annealing algorithm with HotSpot algorithm and uses the global search ability of simulated annealing algorithm to dynamically optimize the minimum support in the solution space. Finally, the Inner Mongolia sandstorm dataset is used for experiment while the wine quality dataset is used for verification, and the association rules screening indicators are set for the mining results. The results show that S_HotSpot algorithm can not only dynamically optimize the selection of support but also improve the quality of association rules as it is mining reasonable number of rules.

1. Introduction

Association rule mining is one of the important research contents in the field of data mining [1], which is widely used in finance, internet, medicine, and other fields. General association rule algorithms can only process discrete data, such as classical Apriori algorithm [2]. If real-type data is to be processed, it needs to be discretized, and in practical applications, most of the data in reality is real. The HotSpot algorithm selected in this paper can directly mine association rules and dynamically acquire the range of real number intervals without discretization of real data, thus avoiding the influence of subjective factors in discretization, and the processing speed is extremely fast. For example, a dataset of 300,000 records with 45 discrete-valued fields can be processed in 10 seconds.

But like Apriori algorithm, HotSpot algorithm has one obvious shortcoming: support selection needs to be set artificially based on experience and cannot be set accurately according to the actual scale of the problem [3]. If the support threshold is set too low, a cumbersome and complicated tree structure may be generated; if the support threshold is set too high, some associated intervals existing in the rare target attribute values may be ignored. Therefore, in the process of support selection, multiple comparison experiments are needed to determine the optimal support based on the mining results.

In order to overcome the sensitivity of HotSpot algorithm in support threshold settings and improve the quality of association rule mining, this paper combines intelligent optimization technology with association rule-mining technology. With the help of the good global search ability of simulated annealing algorithm, support can be set by machine learning, so as to solve the problem that unreasonable threshold setting affects the effect of association rules mining.

2. Related Works

There have been plenty of important research studies in the data mining field. Li uses the HotSpot algorithm to perform handover event association rule mining [4]; Wang researched through HotSpot algorithm the associate degree between customer and evaluation index of express train

service quality [5]. Wang presented an improved *K*-means algorithm by combining the agglomerative hierarchical clustering algorithm to select the initial cluster centers for HotSpot discovery in Internet public opinions [6]. Mallik et al. propose a weighted rule-mining technique (say, RANWAR or rank-based weighted association rule mining) to rank the rules using two novel rule-interestingness measures, viz., rank-based weighted condensed support (WCS) and weighted condensed confidence (WCC) measures to bypass the problem of association rule-mining algorithms evolving huge number of rules [7]. Sitanggang and Fatayati used the SPADE algorithm to generate sequential patterns on hotspot datasets in Sumatra Island, Indonesia, in 2014 and 2015 and then used association rule mining to obtain association between the locations of hotspot sequences and weather conditions [8]. Liu proposed a model by using the traditional vector space model (VSM), *K*-means algorithm, and SVM classifier, for Internet public opinion hotspot detection and analysis [9]. Di Martino et al. showed that the extended fuzzy C-means (EFCM) algorithm works better than the classical FCM algorithm when detecting hotspots [10]. Nisa et al. developed clustering on hotspot dataset to monitor the pattern of forest fires [11]. Agrawal and Choudhary used lung cancer data from SEER dataset with 13 predictor attributes for association rule-mining analysis [12]. Zhang et al. made some improvements to the basic *K*-means algorithm according to the characteristics of hotspot discovery [13].

While association rule mining is a core problem of data mining, the efficiency of mining algorithms is influenced by the data size. When there are very long patterns present in the data, it is often impractical to generate the entire set of frequent itemsets or closed itemsets. The set of maximal frequent itemsets (MFI) contains all frequent sets [14–17]. The MFI is the smallest possible representation of the data that can still be used to generate the frequent itemsets which is way bigger than the number of MFI. Once the frequent itemset is generated, the support information can be easily recomputed from the transactional database. One big issue of the Maximal Frequent Itemset Algorithm (MAFIA) is that the MFI loses the support information of the subset, that is, the support of its subset cannot be determined according to the support of the MFI. In our method, the S_HotSpot algorithm generates an association rule tree of which every node stores a support of corresponding frequent itemsets.

3. HotSpot Association Rule Algorithm-Related Definitions

Definition 1 ($\min S$). The minimum number of segments is used to determine whether there is enough data to segment real number intervals:

$$\min S = [\min \text{Support} * \text{card}(X) + 0.5], \tag{1}$$

where $\min \text{Support}$ means the minimum support and $\text{card}(X)$ means the total number of samples in the dataset.

Definition 2 ($\text{Max}\, C$). The maximum confidence level is used to determine the optimal segmentation point and reflects the reliability of the rules:

$$\text{Max}\, C = \begin{cases} \dfrac{\text{targetLeft}_i}{\text{LeftCount}_i}, & \text{targetLeft}_i \geq \min S, \text{targetRight}_i < \min S, \\ \text{Max}\left(\dfrac{\text{targetLeft}_i}{\text{LeftCount}_i}, \dfrac{\text{targetRight}_i}{\text{RightCount}_i}\right), & \text{targetLeft}_i \geq \min S, \text{targetRight}_i \geq \min S, \\ \dfrac{\text{targetRight}_i}{\text{RightCount}_i}, & \text{targetLeft}_i < \min S, \text{targetRight}_i \geq \min S, \\ \text{Max}\, C, & \text{targetLeft}_i < \min S, \text{targetRight}_i < \min S, \end{cases} \tag{2}$$

where targetLeft and targetRight, respectively, represent the number of transactions that meet the target value on both sides of the traversal point, and LeftCount and RightCount, respectively, represent the number of transactions on both sides of the traversal point. The targetLeft and LeftCount can be add-self obtained by judging the value of the target attribute by each traversal. targetRight and RightCount can be obtained by subself after each traversal judgment. ThetargetLeft, LeftCount, targetRight, and RightCount always satisfy formulas (3) and (4) during the traversal:

$$\text{LeftCount} + \text{rightCount} = \text{card}(X), \tag{3}$$

$$\text{targetLeft} + \text{targetRight} = \text{card}(Y), \tag{4}$$

where $\text{card}(Y)$ means that the dataset meets the number of target attribute values.

Definition 3 ($\max \text{Branches}$). The maximum number of branches limits the number of branches of the tree structure during the construction of the child node.

Definition 4 (targetValue). The target attribute support:

$$\text{targetValue} = \frac{\text{card}(Y)}{\text{card}(X)}. \quad (5)$$

Definition 5 ($\Delta \min \text{Imp}$). The minimum improvement, let G be a K itemset and add an itemset to G to make it G'. Namely, G' means $K + 1$ itemset. At this point, the improvement of $(G \longrightarrow Y)$ is improved by ΔImp, if $\Delta\text{Imp} \geq \Delta \min \text{Imp}$, and the itemset is added to the tree branch. This indicates that the addition of this set contributes to the relevance between G and Y:

$$\Delta\text{Imp} = \frac{\text{Max}\, C - \text{targetValue}}{\text{targetValue}}. \quad (6)$$

4. HotSpot Algorithm and Its Shortcomings

4.1. The Basic Idea of HotSpot Algorithm. HotSpot algorithm is an association rule-mining algorithm based on the tree structure [4], and it can mine discrete datasets or directly mine real data. Namely, the continuous data is divided by finding the associated interval, which is convenient for people to understand and analyze the relationship between the target attribute column and the associated attribute [18].

HotSpot algorithm finds the association interval that satisfies the minimum support, and the mining result is presented in a tree structure [19, 20]. When the given association attribute is continuous data, the specific steps of the algorithm are as follows: (1) statistical support of target attribute column, judging whether it is greater than the minimum support and taking the target attribute value as the root node. (2) Call the spanning tree algorithm, traverse the related attribute columns in turn, and arrange the associated attribute values of each column from small to large. Calculate the confidence of each value by formula (2) and record the new $\text{Max}\, C$ on the premise of increasing the confidence; otherwise, calculate the $\text{Max}\, C$ of the next value, and finally determine the best segmentation point by $\text{Max}\, C$. (3) Calculate the improvement degree of the segmentation point. If $\Delta\text{Imp} \geq \Delta \min \text{Imp}$, it is regarded as the child node joining candidate queue, and the segmentation point is recorded to output the HotSpot association rule tree. (4) Decide how to recursively call the spanning tree algorithm according to the queue size until no new child nodes are generated. (5) Output the association rule tree.

4.2. Problems in HotSpot Algorithm. The HotSpot algorithm for the mining of association rules depends on the artificially based minimum support and cannot be accurately set according to the actual scale of the dataset. This may lead to the following problems in the execution of the algorithm:

(1) If support is set too low, a cumbersome and complicated tree structure may occur, which makes the association rules too much and is difficult to extract. In addition, it is easy to result in fewer samples satisfying the range of association attributes, which makes association rules meaningless.

(2) If support is set too high, some association intervals existing in the rare target attribute values may be ignored, and it is difficult to find targeted association rules, which makes the conversion of data into knowledge inefficient.

5. Improvement of HotSpot Algorithm

5.1. Simulated Annealing Algorithm. The idea of simulated annealing algorithm was proposed by Metropolis et al. [21] in 1953, which is a global optimal algorithm [22]. At present, it has been widely used in production scheduling, artificial intelligence, and other fields. The simulated annealing algorithm starts to cool down from a higher initial temperature and combines the probability jump feature in the annealing process to randomly search for the global optimal solution in the solution space. The specific ideas are as follows:

(1) Determine the range of the model solution space, randomly generate an initial solution X_0 in the solution space, and calculate the corresponding objective function value $E(X_0)$

(2) Set the initial temperature $T = T_0$

(3) Perturbation according to the current solution X_i, generates a new solution X_j, and calculating the corresponding objective function value $E(X_j)$ to obtain $\Delta E = E(X_j) - E(X_i)$

(4) Determine whether to accept the new solution according to the Metropolis criterion for the incremental value ΔE

(5) In the homogeneous algorithm, steps (3) and (4) are repeated L_K times at temperature T_K; in the non-homogeneous algorithm, this step is ignored

(6) Annealing according to the temperature update function, namely, let T be equal to the next value T_K in the annealing schedule

(7) Repeat steps (3)–(6) until the termination condition are met

5.2. HotSpot Algorithm Improvement Ideas. The S_HotSpot algorithm proposed in this paper combines association rule-mining technology with intelligent optimization technology and overcomes the shortcomings of the setting of the support threshold in the HotSpot algorithm that needs to be set artificially many times according to experience. Finally, the support threshold is set by machine learning and combined with the mining background so that the quality of the mining association rules is improved and the quantity is reasonable, and the mining effect of the association rules is optimized. The key to the improvement of HotSpot algorithm lies in the setting of energy function E, state generation function, and temperature update function. The idea of setting these functions is emphasized below.

5.2.1. Energy Function E. There are two main factors influencing the results of HotSpot algorithm mining: (1) confidence: Conf; (2) cover: the number of samples

satisfying the range of association attributes in each association rule. Therefore, this paper uses the average confidence of all association rules to measure the importance of mining results while constructing the energy function, which is called the importance factor Conf_aver; the sample coverage is measured by the ratio of the number of samples in all association rules that satisfy the scope of the associated attribute (the number of association rules $^*\mathrm{Card}(x)$), called the sample coverage factor Cover_rate.

In view of this, this paper constructs an energy function model by Euclidean distance formula. As shown in Figure 1, where O is the ideal point, namely, (Conf_aver, Cover_rate) = (100%, 100%), A is the value of Conf_aver and Cover_rate under current support and improvement. The energy function is the distance from point A to point O, see formula (7).

From formula 7, n means the number of association rules and count means the number of samples in each association rule that satisfy the range of association attributes:

$$E = \sqrt{\left(1 - \frac{\sum_{i=1}^{n} \mathrm{Max}\, C}{n}\right)^2 + \left(1 - \frac{\sum_{i=1}^{n} \mathrm{count}}{n \times \mathrm{card}(X)}\right)^2}. \quad (7)$$

5.2.2. State Generation Function. The essential function of the state generation function is to make the candidate values of the support as far as possible throughout the solution space 0 to 1 so that the selection of the support is reasonable and global. Therefore, this paper sets a small enough initial support degree as θ, he base of support, and $(1-\theta)/L_n$ as the step size and then accumulates the support in turn:

$$\begin{cases} L_n = \log_{\alpha}^{T_f/T_0}, \\ X_{\mathrm{new}} = \theta + \mathrm{iter} \times \dfrac{1-\theta}{L_n}, \end{cases} \quad (8)$$

where θ is a constant less than 1%; iter means the number of current iterations; L_n means the total number of iterations; T_0 means the initial temperature; T_f means the termination temperature; and α means the iteration coefficient. After many comparative experiments, this paper makes $T_0 = 10^{20}$ and $T_f = 10^{-30}$.

5.2.3. Temperature Update Function. Because of the law of temperature change in the process of exponential annealing, the annealing rate is only related to the value of α, so it is a common annealing strategy. A better exponential anneal function is shown in formula (9), where α is an empirical value, which is a constant slightly less than 1, generally in the range of [0.65, 1). After many comparative experiments, this paper makes $\alpha = 0.9$:

$$T_k = \alpha \times T_{k-1}, \quad K \geq 1, 0 < \alpha < 1. \quad (9)$$

5.3. Improved Implementation of HotSpot Algorithm. The flow chart of S_HotSpot algorithm for optimizing support threshold using simulated annealing algorithm is shown in Figure 2. The specific calculation steps are as follows:

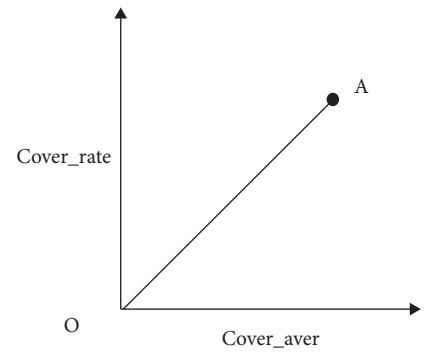

Figure 1: Energy function construction model.

(1) First, a solution X_{old} is generated in the solution space (0, 1) of the support; let min Support = X_{old}

(2) Let the initial temperature be T_0, the termination temperature be T_f, and the temperature update function be T_k

(3) Taking X_{old} as the support value threshold, then using it to mine association intervals through the HotSpot algorithm for the task-related dataset, and calculating the energy value E_{old}

(4) Generating a new solution X_{new} according to the state generation function

(5) If $X_{\mathrm{new}} >$ targetValue, the algorithm ends; otherwise, step (6)

(6) Taking X_{new} as the new support value threshold, then using it to mine association intervals through the HotSpot algorithm for the task-related dataset, and calculating the energy value E_{new}

(7) Calculating the value difference of energy function ($\Delta E = E_{\mathrm{new}} - E_{\mathrm{old}}$) and judging whether to accept the new solution according to the Metropolis criterion

(a) If $\Delta E < 0$, then min Support = X_{new}

(b) If $\Delta E > 0$, the probability $p = \exp(-\Delta E/T_k)$ is calculated; if $P >$ Random (0, 1), min Support = X_{new}; otherwise, min Support = X_{old}

(8) Determining whether T_k has reached T_f; if T_f has been reached, terminate the algorithm and output the association rule tree; otherwise, repeat steps (4)–(6)

It can be seen that S_HotSpot algorithm makes the support value continuously change in the solution space according to the set state generation function and combines the difference of the energy function E with the Metropolis criterion to determine whether to accept the new support threshold. In this process, the support threshold is set. It can find by itself through machine learning, and there is no need to manually select the support threshold several times. The time complexity of the algorithm is $O(nh2^h)$, where h is the number of attribute indexes and n is the number of samples.

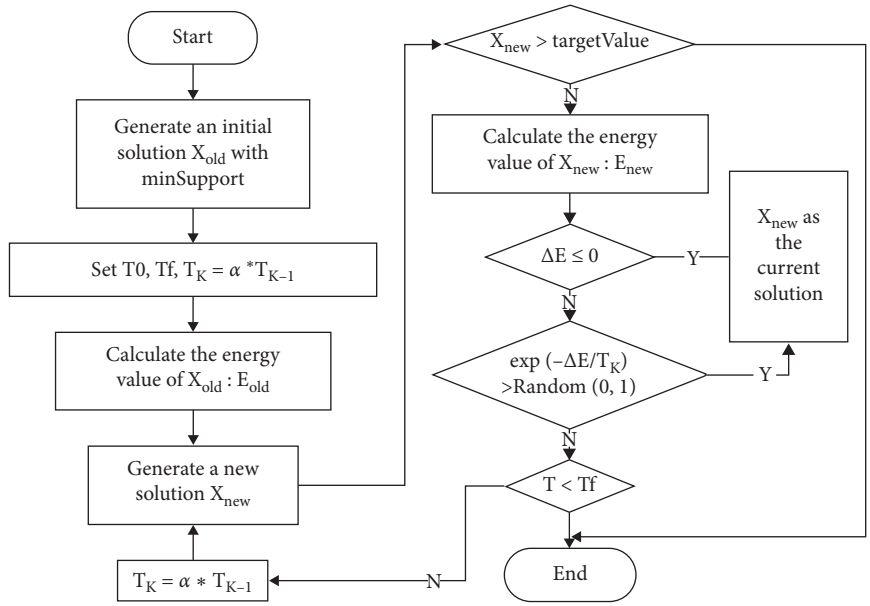

Figure 2: S_HotSpot algorithm flow chart.

6. Experiment

6.1. Experimental Setup. The experimental hardware configuration in this paper includes Inter(R) Core(TM) i7-6700 CPU 3.40 GHZ processor; 4 GB memory; and AMD Radeon R5 340 graphics card. Software configuration includes Windows 8.1, MATLAB 2016a, and SPSS20.

The experimental dataset is derived from "China's strong sandstorm sequence and its supporting dataset" and "China's surface climate data daily value dataset" (to obtain the experiment source code, please visit the following link: https://github.com/yaya0804/S_HotSpot-Algorithm.git). After data preprocessing such as data integration, data dimensionality reduction, and screening of Inner Mongolia sites [23], the association rule-mining objects are constructed in the form shown in Figure 3.

From Figure 3, the sandstorm level includes four intensities: I1: Heavy Sandstorm, I2: Strong Sandstorm, I3: Sandstorm, I4: Blowing Sand. In the mining process, the sandstorm grade is defined as the target attribute, the maximum number of branches is 2, the minimum improvement degree is 0.01, and the month and meteorological attributes (units in turn are 0.1 mm, 0.1 mm, 1%, 0.1 h, 0.1/C, and 0.1 m/s.) are related attributes. BS and AS represent the number of association rules before and after filtering according to the index of association rules, respectively.

6.2. Screening and Analysis of Experimental Results

6.2.1. Association Rules Screening Indexes

(1) Objective evaluation index: after the rules that satisfy the minimum support in the HotSpot algorithm are output in the form of a tree structure, the user can form an initial rule set, but at this time some rules have little meaning. For example, when the support is 0.04 and the target attribute is I4, the tree structure is illustrated in Figure 4, where the components of each path represent an optimal segmentation point except the root node, the dotted line box indicates the maximum confidence corresponding to the best segmentation point, the denominator in the solid line box indicates the number of samples that satisfy the range of meteorological factors, and the numerator indicates the number of samples in which the target attribute value occurs on the basis of satisfying the range of meteorological factors

Figure 4 includes five association rules, although the confidence is high, it is not reliable. For example, one of the rules is when the average wind speed is <=3.5 m/s and the sunshine hours are >1.1 h, the probability of occurrence of I4 is 78.28%. However, from Figure 4, it can be found that there are only 396 samples which satisfy the average wind speed <=3.5 m/s and the sunshine duration >1.1 h meteorological factor range, accounting for only 5.1% of the total sample. The small denominator causes the occurrence of the meteorological factor range itself to be a small probability event, which is not representative. Such an association rule has little meaning. Therefore, the rules mined by the HotSpot algorithm are not as important as the higher confidence level, and it is also necessary to make the number of samples satisfying the range of meteorological factors not too low.

Therefore, this paper conducts the meteorological factor range coverage test and confidence test for the association rules of HotSpot mining so that the mining association rules are reliable. On this basis, the association rules are screened as follows:

Sandstorm grade	Month	Small evaporation	Accumulated precipitation at 20 – 20	Average relative humidity	Sunshine hours	Average temperature	Average wind speed
I4	3	39	0	50	49	–3	54
I4	4	58	0	33	88	43	36
I4	4	54	0	48	96	56	15

Figure 3: Dataset part data.

```
Total: 7734 instances
Target attribute: sandstorm grade
Target value: I4 [Number:3884 instances (50.2198%)]
minS: 309 instances (4% of total)
Sandstorm grade = I4 (50.2198% [3884/7734])
|   average wind speed <= 35 (75.9815%) [329/433])
|   |   Sunshine hours > 11 (78.2828%) [310/396])
|   average temperature > 189 (61.0312%) [509/834])
|   |   average wind speed <= 65 (65.3527%) [315/482])
|   |   Small evaporation <= 195 (63.2653%) [310/490])
```

Figure 4: HotSpot algorithm association rule-mining result.

(i) Meteorological factor range coverage test: keep the association rule of meteorological factor range coverage (Cover) > 0.3:

$$\text{Cover} = \frac{\text{count}}{\text{card}(X)}. \tag{10}$$

(ii) Confidence test: higher confidence indicates that the probability of occurrence of the target attribute is greater; this paper specifies Conf > 0.15.

(1) Simplicity: simplicity reflects the user's comprehensibility, which is reflected in the number of rules. Since the HotSpot algorithm has specified the target attribute column before mining, the number of consequential items of the rule is 1. Only the number of precursors of rules is considered, namely, meteorological factors related to meteorological factors. This paper stipulates that the front part is larger than three items for deletion, which is convenient for analysis.

6.2.2. Analysis of HotSpot Mining Results. From Table 1, it can be found that the HotSpot algorithm runs extremely fast, with an average running time of less than half a second. The total number of association rules after screening through subjective evaluation indicators and objective evaluation indicators: sum (AS) = 12. In addition, it can be found that when the target attribute is listed as I4 and the support is 0.3 and 0.5, the number of association rules before and after the selection of association rules remains unchanged, indicating that when the support set is 0.04, 0.1, and 0.3, and 0.5 and 0.3–0.5 is the optimal support for I4. Similarly, 0.3 is the optimal support for I3 and 0.1 is the optimal support for I2. When the target attribute is listed as I1 and the support is as low as 0.04, the number of association rules is still 0. This is because the heavy sandstorm recorded only 76, accounting for less than 1%.

Table 1: HotSpot mining results.

Target attribute column grade	Support	BS	AS	Runtime (s)
I4	0.04	5	0	0.36
	0.1	4	0	0.34
	0.3	2	2	0.37
	0.5	1	1	0.36
I3	0.04	24	1	0.35
	0.1	5	1	0.33
	0.3	3	3	0.31
	0.5	0	0	0.33
I2	0.04	12	0	0.35
	0.1	4	4	0.36
	0.3	0	0	0.33
I1	0.009	75	0	0.87
	0.04	0	0	0.31

6.2.3. Analysis of S_HotSpot Mining Results. In Table 2, in terms of the number of association rules, the total number of association rules after screening through subjective evaluation indicators and objective evaluation indicators is sum (AS) = 5. In addition, the optimal support interval indicates that, in this interval, the value of energy function is the same and the minimum, namely, the mining results are the same in this interval. It can be found that when the target attribute column is I4, the optimal support interval is about 0.5; when the target attribute column is I3, the optimal support interval is about 0.3; and when the target attribute column is I2, the optimal support interval is around 0.1. The results are basically consistent with those of Section 6.2.2, so S_HotSpot algorithm can dynamically optimize the support threshold.

6.3. Performance Comparison. In terms of the quality of association rules, take the target attribute I3 as an example for comparison experiments. The mining result of S_HotSpot algorithm is shown in Figure 5.

From Figure 5, it can be found that the probability of I3 sandstorm is 33.74% when the average wind speed is ≥3.6 m/s. At this time, the optimal support interval is 31.72%–31.82%, and the meteorological factor range coverage rate is 94.06%. It can be seen from Section 6.2.2 that 0.3 is the optimal support for I3, so the support is set to 0.3. The mining result of the HotSpot algorithm is shown in Figure 6.

Figure 6 has three association rules. The meteorological factor coverage of the three rules is calculated according to equation (10): 87.04%, 93.05%, and 88.03%,

Table 2: S_HotSpot mining results.

Target attribute column grade	Optimal support/interval (%)	BS	AS	Runtime (s)
I4	48.08–48.36	1	1	39
I3	31.72–31.82	1	1	26.8
I2	15.45	3	3	8.55
I1	0.92	68	0	3.98

```
Optimal support interval: 31.7243% - 31.8157%
Total : 7734 instances
Target attribute: sandstorm grade
Target value : I3 [Number:2532 instances (32.7386%)]
minS : 2454 instances  (31.7243% of total)
Sandstorm grade = I3 (32.7386% [2532/7734])
|       average wind speed > 36 (33.7354% [2461/7295])
```

Figure 5: S_HotSpot mining results.

```
Total: 7734 instances
Target attribute: sandstorm grade
Target value: I3 [Number:2532 instances (32.7386%)]
minS: 2320 instances (30% of total)
Sandstorm grade = I3 (32.7386% [2532/7734])
|      average wind speed > 45 (34.492% [2322/6732])
|      Accumulated precipitation at 20-20 <= 5 (33.4167% [2405/7197])
|      |      average wind speed > 36 (34.4108% [2342/6806])
```

Figure 6: HotSpot mining results.

both of which are lower than the meteorological factor range coverage of rule 2 in Figure 5. Moreover, the S_HotSpot algorithm mining results consider the meteorological factor coverage and confidence so that the energy function is the lowest and the association rules are practical, so the S_HotSpot mining results are better than the HotSpot mining results.

6.4. Discussion. From the aspect of support selection, S_HotSpot algorithm can directly obtain the optimal support or interval. In terms of running time, S_HotSpot algorithm is slower than HotSpot algorithm. In terms of the number of association rules, the number of association rules mined by S_HotSpot algorithm is concise and easy to filter and analyze. In summary, the performance comparison is shown in Table 3.

In Table 3, although the S_HotSpot algorithm runs slowly, this runtime represents the time it takes to find the optimal support threshold, while HotSpot algorithm corresponds to a single runtime. Although it takes less time, it needs several comparative experiments to determine the optimal support according to the mining results. Generally speaking, it takes longer time than S_HotSpot algorithm.

In summary, the overall performance of the S_HotSpot algorithm is better than the HotSpot algorithm.

6.5. Comparative Study. The paired t-test is used to compare the values of means from two related samples. This paper made comparisons using SPSS between the HotSpot algorithm and the S_HotSpot algorithm before and after filtering according to the index of association rules, respectively, in order to determine whether the results provide sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis. The P value is set to 0.05. Figure 7 shows the test results of both BS and AS dataset.

It can be seen from the table-*paired samples statistics* in Figure 7 that the average number of association rules after filtering using the HotSpot algorithm is 2.60, while using S_HotSpot, the average number is 4.20. Table-*paired samples correlations* shows the correlation coefficient between the two algorithms is 0.293, the corresponding probability is 0.210 which is greater than the P value 0.05. It can be considered that the two paired algorithms have no correlation. Table-*paired samples test* shows the mean value of the paired differences between the two samples is −1.6, the t-value is −1.256, and the corresponding probability is 0.224 which is greater than 0.05; therefore, the null hypothesis cannot be rejected. On the contrary, the t-value of the algorithm matching test before association rule filtering is 0.028, smaller than 0.05. This shows that when there is no evaluation index screening, the difference between HotSpot algorithm and S_HotSpot algorithm is significant, indicating that merging the simulated annealing algorithm into Hotspot algorithm will influence the association rule mining. After filtering, the association rules mined by them are similar, indicating that the good association rules have been mined by both algorithms (Figure 8).

6.6. Validation. The validation dataset is derived from "wine quality" [24]. After data preprocessing such as data integration and data dimensionality reduction, the association rules mining objects are constructed in the form shown in Figure 8.

In terms of the quality of association rules, take the target attribute as I6 as an example for validation experiments. Multiple experiments of different support have been conducted using HotSpot algorithm and it turns out that the best support is 0.3, while S_HotSpot algorithm shows that the optimal support is 0.27 which is pretty close to 0.3. The mining result of S_HotSpot algorithm is shown in Figure 9 and HotSpot in Figure 10.

As can be seen from the above, the S_HotSpot algorithm still maintains the abovementioned advantages over HotSpot algorithm using different datasets.

Table 3: Performance comparison between S_HotSpot and HotSpot.

Influence factor	HotSpot algorithm	S_HotSpot algorithm
Support selection	Artificial setting	Dynamic optimization
Running speed	Fast	Slow
Number of association rules	More, difficult to screen	Less, easy to screen
Association rule quality	Low	High

Paired samples statistics

	Mean	*N*	Std. deviation	Std. error mean
Pair 1 Hotspot_BS	2.600	20	3.25091	0.72693
S_Hotspot_BS	14.2000	20	20.53393	4.59153
Pair 2 Hotspot_AS	2.6000	20	3.25091	0.72693
S_Hotspot_AS	4.2000	20	5.72713	1.28062

Paired samples correlations

	N	Correlation	Sig.
Pair 1 Hotspot_BS & S_Hotspot_BS	20	−0.328	0.158
Pair 2 Hotspot_AS & S_Hotspot_AS	20	0.293	0.210

Paired samples test

	Paired differences					*t*	d*f*	Sig. (2−tailed)
	Mean	Std. deviation	Std. error mean	95% Confidence interval of the difference Lower	Upper			
Pair 1 Hotspot_BS − S_Hotspot_BS	−11.60000	21.81839	4.87874	−21.81132	−1.38868	−2.378	19	0.028
Pair 2 Hotspot_AS − S_Hotspot_AS	−1.60000	5.69764	1.27403	−4.26658	−1.06658	−1.256	19	0.224

Figure 7: Paired *t*-test for HotSpot and S_HotSpot algorithm.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Quality	Fixed acidity	Volatile acidity	Citric acid	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	pH	Sulphates	Alcohol
2	I6	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8
3	I6	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5
4	I6	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1
5	I6	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9
6	I6	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9
7	I6	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1

Figure 8: Wine-quality dataset part data.

```
Optimal support interval: 27.7027% - 27.7027%
Total: 4898 instances
Target attribute: quality
Target value: I6 [Number:2198 instances (44.8755%)]
minS: 1357 instances (27.7027% of total)
quality = I6 (44.8755% [2198/4898])
|   volatile acidity <= 0.27 (52.107% [1422/2729])
|   |   alcohol > 8.9 (53.3725% [1361/2550])
|   |   free sulfur dioxide > 13 (53.0739% [1364/2570])
|   alcohol > 9.8 (48.7354% [1503/3084])
```

Figure 9: S_HotSpot mining result using wine quality dataset.

```
Total: 4898 instances
Target attribute: quality
Target value: I6 [Number:2198 instances (44.8755%)]
minS: 1469 instances (30% of total)
quality = I6 (44.8755% [2198/4898])
|   volatile acidity <= 0.28 (51.0184% [1528/2995])
|   |   alcohol > 8.8 (52.0994% [1489/2858])
|   |   free sulfur dioxide > 12 (51.8792% [1477/2847])
|   alcohol > 9.8 (48.7354% [1503/3084])
```

Figure 10: HotSpot mining result using wine quality dataset.

7. Conclusion

In this paper, the simulated annealing algorithm is used to intelligently optimize the support threshold so that the quality of association rule-mining results is increased and the quantity is reasonable. The dataset is used as the mining background to conduct a comparative experiment and the optimal support mining results of HotSpot algorithm based on experience and S_HotSpot algorithm mining results are analyzed and compared from the running time of the algorithm, the selection of support, and the quality and quantity of association rules. By using HotSpot algorithm, one can see that the optimal support interval is 31.72%–31.82%, and the meteorological factor range coverage rate is 94.06%; S_HotSpot algorithm found three association rules, the meteorological factor coverage of the three rules is calculated according to equation (10): 87.04%, 93.05%, and 88.03%, both of which are lower than the meteorological factor range coverage of rule 2 in Figure 5.

The experimental results show that the S_HotSpot algorithm can automatically optimize the support; thus, high-quality association rules that users are interested in can be mined. Moreover, the S_HotSpot algorithm mining results consider the meteorological factor coverage and confidence so that the energy function is the lowest and the association rules are practical.

Although the overall performance of the improved S_HotSpot algorithm is better than the HotSpot algorithm, there is a bottleneck in the overall running time of the algorithm, which needs further improvement. In addition, the dataset used in the algorithm has many correlative attributes, and there is no data exploration and analysis for each sandstorm grade to explore the highly relevant meteorological attributes for each grade of sandstorm.

Data Availability

The dataset used to support the findings of this study have been deposited in the National Meteorological Information Center repository http://data.cma.cn/site/index.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (61966027) and Natural Science Foundation of Inner Mongolia (2018MS06021, 2016MS0605, and 2015MS0614).

References

[1] M. S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 866–883, 1996.

[2] Y. Cui and Z. Bao, "Survey of association rule mining," *Application Research of Computers*, vol. 33, no. 2, pp. 330–334, 2016.

[3] X. Wang, *Abnormal Objects Recognition in Video Based on Data Mining*, Taiyuan University of Technology, Taiyuan, China, 2013.

[4] J. Li, *Research Handover Method in GSM-R*, Zhejiang University, Hangzhou, China, 2011.

[5] Y. Wang, "Study on comprehensive evaluation of service quality of railway express freight trains," 2017.

[6] G. Wang, "Research on hotspot discovery in internet public opinions based on improved *K*-means," *Computational Intelligence and Neuroscience*, vol. 2013, Article ID 230946, 6 pages, 2013.

[7] S. Mallik, A. Mukhopadhyay, and U. Maulik, "RANWAR: rank-based weighted association rule mining from gene expression and methylation data," *IEEE Transactions on Nanobioscience*, vol. 14, no. 1, pp. 59–66, 2014.

[8] I. S. Sitanggang and E. Fatayati, "Mining sequence pattern on hotspot data to identify fire spot in peatland," *International Journal of Computing and Information Sciences*, vol. 12, no. 1, pp. 143–147, 2016.

[9] H. Liu, "Internet public opinion hotspot detection and analysis based on k means and SVM algorithm," in *Proceedings of the 2010 International Conference of Information Science and Management Engineering*, vol. 1, IEEE, Xi'an, China, pp. 257–261, August 2010.

[10] F. Di Martino, V. Loia, and S. Sessa, "Extended fuzzy *C*-means clustering algorithm for hotspot events in spatial analysis," *International Journal of Hybrid Intelligent Systems*, vol. 5, no. 1, pp. 31–44, 2008.

[11] K. K. Nisa, H. A. Andrianto, and R. Mardhiyyah, "Hotspot clustering using DBSCAN algorithm and shiny web framework," in *Proceedings of the 2014 International Conference on Advanced Computer Science and Information System*, IEEE, Jakarta, Indonesia, pp. 129–132, October 2014.

[12] A. Agrawal and A. Choudhary, "Identifying hotspots in lung cancer data using association rule mining," in *Proceedings of the IEEE 11th International Conference on Data Mining Workshops*, IEEE, Vancouver, Canada, pp. 995–1002, December 2011.

[13] H. Zhang, C. Liu, M. Zhang et al., "A hot spot clustering method based on improved kmeans algorithm," in *Proceedings of the 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, IEEE, Chengdu, China, pp. 32–35, December 2017.

[14] P. K. D. Sarma and A. K. Mahanta, "Reduction of number of association rules with inter itemset distance in transaction databases," *International Journal of Database Management Systems*, vol. 4, no. 5, p. 61, 2012.

[15] K. K. Sethi, R. Dharavath, and S. Nyakotey, "PPS: parallel pincer search for mining frequent itemsets based on spark," in *Advances in Intelligent Systems and Computing*, pp. 351–363, Springer, Cham, Switzerland, 2016.

[16] D. Burdick, M. Calimlim, J. Flannick et al., "MAFIA: a performance study of mining maximal frequent itemsets," 2003.

[17] K. Gouda and M. J. Zaki, "GenMax: an efficient algorithm for mining maximal frequent itemsets," *Data Mining and Knowledge Discovery*, vol. 11, no. 3, pp. 223–242, 2005.

[18] L. Zhang, Y. Tan, G. Xiao et al., *Matlab Data Analysis and Data Mining*, China Machine Press, Beijing, China, 2015.

[19] M. Hall, "HotSpot segmentation-profiling [EB/OL]," 2010, https://wiki.pentaho.com/display/DATAMI NING/HotSpot+Segmentation-Profiling.html,2010-08-24/2018-11-20.

[20] M. Hall, "HotSpot in weka [EB/OL]," 2018, http://weka.sourceforge.net/packageMetaData/HotSpot/index.html.2011-01-11/2018-11-22.

[21] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth et al., "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, 1953.

[22] X. Zhou, Y. Ma, and Y. Hu, "MixedGenetic algorithm and simulated annealing algorithm for solving job shop scheduling problem," *Journal of Chinese Computer Systems*, vol. 36, no. 2, pp. 370–374, 2015.

[23] J. Han, M. Kamber, J. Pei et al., *Data Mining: Concepts and Techniques*, China Machine Press, Beijing, China, 2012.

[24] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," in *Decision Support Systems*, vol. 47, pp. 547–553, no. 4, Elsevier, Amsterdam, Netherlands, 2009.