Application of an Improved Apriori Algorithm in Intelligence Greenhouse System

Liu Xiao-guo, Chen Yu

Jilin Agricultural University, Changchun, 130118 ,China

Keywords: Greenhouse, Association rule, Apriori, Sensor.

Abstract: To solve the problem that the accure data can't be pushed by the failure of local sensor in intelligence greenhouse system, it was presented that the Apriori algorithm which was based on association rule applied in the prediction of sensor fault data. Forcasting the greenhouse environment temperature is provided as an example in this paper, firstly, the classic Apriori algorithm is modified. Then it was used in the prediction of fault sensor data. The experimental results show that the improved Apriori algorithem could quickly find the association rule between the parameters in Greenhouse, thus estimated the range of the parameters of the fault sensor and the method could be proved to be feasible.

Introduction

With the technology of the Internet of things in agriculture application, and the concept of precision agriculture is proposed. By using the modern technology to agriculture, greenhouse using a variety of sensors, data acquisition can be timely accurate and effective transmission, is directly related to greenhouse crops is always in its optimal growth environment, thereby affecting the production of greenhouse crops. In this paper, combined with the actual engineering application background, firstly, the traditional association rules analysis Apriori algorithm is improved, and then the fault of the temperature sensor is analyzed.

Optimization and improvement of the traditional Apriori algorithm

A. The existing problems of the traditional Apriori algorithm

The most typical application of the traditional Apriori algorithm is the shopping cart analysis. The generation of the association rules is based on the relationship between the objects in the transaction database. However, we need to excavate the air humidity, light intensity, soil temperature, soil moisture in our intelligent greenhouse system, CO2 concentration and light intensity and air temperature according to association rules ask, that the form such as association rules A ^ B ^ C ^ D→E, and Apfiofi algorithm using discrete interval unlike the traditional of the same attribute parameter connection will produce a large number of redundant association rules, for example, A1, B1, C1 and Al, B1, B1, C2 connected to produce A1, C1, C2, C1 and C2 are assuming discrete interval illumination the different parameters, namely C1 and C2 are mutually exclusive the relationship cannot exist at the same time in the same association rules, so this kind of connection is meaningless, and will consume a large amount of system resources. So in order to remove redundant association rules to improve mining efficiency, we need to optimize the traditional Apriori algorithm to improve the efficiency of mining.

B. Improvement of traditional Apriori algorithm

Since the Apriori algorithm was put forward, in order to overcome its own shortcomings, improve the performance of the algorithm. Many scholars have conducted a lot of research, and put forward a variety of optimization and improved algorithm. Can be divided into the following four ideas: (1) method based on hash (hash): the main idea of the algorithm is frequently k-1 item sets are produced by k-itemsets by hash function mapping to different buckets (address mapping, and add the corresponding bucket count. (2) sampling method based on: the core idea is selected

random sample s in the transaction database, then s generated frequent itemsets, the essence is sacrificing accuracy exchange for raising the efficiency, especially for the calculation of the intensive frequent itemsets. (3) reduce the number of transactions: according to the transaction does not contain the frequent K item set does not contain the k+1 item set this conclusion, do not contain the candidate item set of transaction marks to be deleted, so as to reduce the number of scanning data. (4) partition based approach: the main idea is to block the transaction database according to the corresponding logic, and then consider each block to generate frequent item sets, and finally merge the frequent itemsets generated by each block. This paper is based on the division of the idea of Apriori algorithm optimization, and made a corresponding correction. First of all, we to set attribute dictionary sorting according to partition the transaction database, we in a two candidate as an example, such as Figure 1 the candidate set of itemsets are sorted lexicographically after delimit set of molecular subset.

Then we are the subset of itemsets and subset 2 item set connection to generate (A1, B1, C1); (A1, B1, C2}; (A1, B1, C3}; (A1, B2, C1); (A1, B2, C2); (A1, B2, c3}; {A1, B3, c1}; {A1, B3, and {Al CI}, B3, C1} a total of nine candidate set. And if the use of the traditional Apriori algorithm will generate (A1, B1, C1}; {AI, B1, C2}; (A1, B1, C3); (A1, B2, C1}; (A1, B2, C2}; (A1, B2 and C3); (A1, B3 and C1); {A1, B3, C1}; {A1, B3, C1}; (A1, C1, C2); (A1, Cl, and c3} (A1, C2, C3) a total of 15 candidate set, which contains six redundant rules.

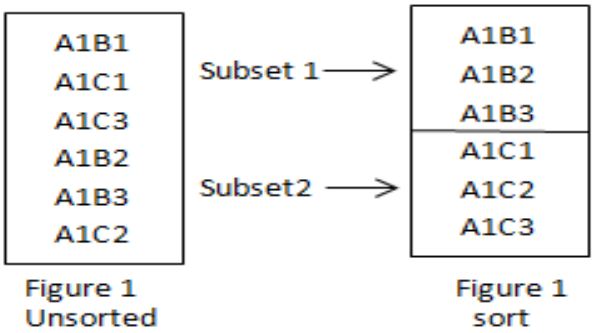

Figure 1 the candidate set of itemsets are sorted lexicographically after delimit set of molecular subset

C. algorithm efficiency analysis

In order to further verify the superiority of the algorithm used in this paper, we will compare the performance of the proposed optimization algorithm with the traditional Apriori algorithm. Experimental verification hardware environment for Inter (R) core (MT) CPU 4.00GH i3-2630M, memory for the PC 2GB. The software environment is win7 operating system, the database system is oraclel10, the development language is R language.

We compare and analyze the two algorithms from the following three aspects: A. in the same degree of support of the two algorithms in a number of different database transactions run the time consumed. B. in support of the same database in the same number of transactions in the case of the same number of cases, the number of frequent sets generated. C. the number of transactions in the database is equal to the degree of support for different cases, the two algorithms run the time consumed.

Figure 2 is respectively with the two algorithms of containing 10000, 20000, 30000, 40000, 50000 data acquisition sensor data sets of association rules mining consumption time, the red line represents the improved algorithm IMApriori, the blue line represents the traditional Apriori algorithm, minimum support degree is 10%. From Figure 3 we can see that with the expansion of the scale of the data set, the improved IMApriori algorithm is significantly less computing time than the traditional Apriori algorithm. This is because of the increasing number of redundant association rules generated due to the increase of data set data, the improved IMApriori algorithm can effectively filter out this part of association rules to speed up the operation.

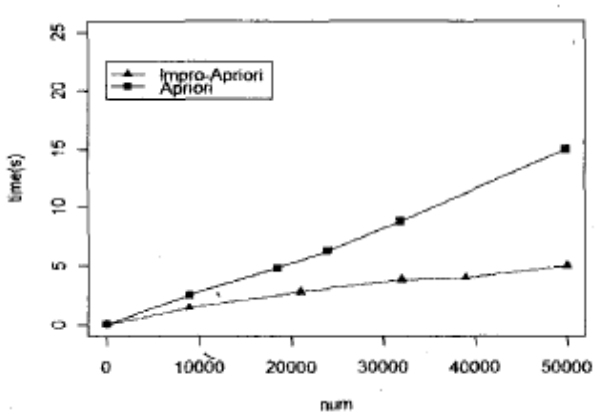

Figure 2 Two algorithms run in different data sets of consumption time contrast

Table3. The number of frequent sets generated by the two algorithms

Frequent collection	IMApriori	Apriori
1-item	39	29
2-item	301	406
3-item	2340	3654
4-item	3407	8701
5-item	4407	10291

On the table was IMApriori algorithm and Apriori algorithm in the database transaction set to 10000 generated frequent set number of contrast, by Table 1 we can see frequent sets the number of attributes for three the following time, IMApriori algorithm and Aprioi algorithm to generate frequent set number decreased but not significantly, but when the number of attributes more than 3, frequent IMApriori generated by the algorithm set number is much smaller than the size of Aprioi algorithm, this can greatly reduce the removal of redundant association consumed when asked, making the efficiency of the algorithm is more high.

Chart 4.These two kinds of algorithm in the minimum support degree is 5%, 10%, 15%, 20%, 25% mining association rules the consumption of time, the database transaction set number 10000. From Figure 3 we can see with the minimum support degree reduce the computation time of the algorithm IMApriori is obviously less than that of the computation time of the traditional Apriori algorithm is improved.

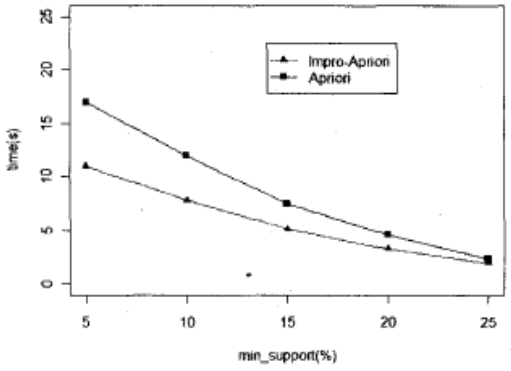

Figure 3 Two algorithms run in different data sets of consumption time contrast

Two algorithms run in different support degree of consumption time comparison table generated by the two algorithms frequent itemsets number table 1 for IMApriori algorithm and Apriori algorithm in the database transaction set to 10000 generated frequent set number of contrast, from table 1 we can see frequent sets the number of attributes for three following time, IMApriori algorithm and Aprioi algorithm to generate frequent set number decreased but is not obvious

enough, but when the number of attributes more than 3, frequent IMApriori generated by the algorithm set size is less than that of the algorithm Aprioi. This can reduce the removal of redundant association consumed when asked, making the algorithm efficiency more high. In summary, we believe that, in the context of the generation of association rules with the greenhouse temperature prediction, the improved IMApriori algorithm performance is better than the traditional Aprior algorithm performance. Application of improved IMApriori algorithm in greenhouse temperature prediction.

Application of improved IMApriori algorithm in greenhouse temperature prediction

A. greenhouse sensor acquisition data preprocessing

This paper adopts intelligent greenhouse in May 5, 2015 day of the data collected by sensor nodes as the original data set, shown in Table 2 as part of the original data. The METER property of the ID bar, SARH SSTP on behalf of environmental humidity, soil temperature, soil moisture SSRH, SACD on behalf of CO_2 concentration, SSUN light intensity, SATD environment temperature. table 2. we can see in this moment SATD 0:49:55 VALUE value is 0, that is to say failure fails to push accurate data at this time. The temperature sensor and the temperature as the important factors affecting the growth of crops, crops in the suitable temperature range if their growth, can the growth and development, when the higher or lower than the growth of crops will reduce the scope of activity, its growth will eventually lead to the products will be affected Decline in the quantity or quality. If this temperature sensor failure to push the temperature value is 0), in the greenhouse is self executing agencies will perform wrong operation, making the greenhouse temperature is no longer in the crop of optimal growth environment. This time we need based on association rules of the numerical range of sensor temperature make a reasonable estimate, we manually operated greenhouse actuator to provide a scientific basis.

Table2.The part of the original data

DATA_ID	COLLECTOR_ID	METER_ID	PARAM_ID	VALUE	COLLECT_TIME
67861	420402789	SACD	1	400.1	2016/5/615:57:52
67862	420402789	SSTP	1	15.1	2016/5/615:57:52
67863	420402789	SSRT	1	18.4	2016/5/615:57:52
67864	420402789	SARH	1	17.7	2016/5/615:57:52
67865	420402789	SSUN	1	16.4	2016/5/615:57:52
67866	420402789	SATD	1	83.6	2016/5/615:57:52
67867	420402789	SSRH	1	17.6	2016/5/615:57:52

First, we need to pre process the data, according to the sensor's measurement range of the sensor data collected for the discretization of the classification, table 3 for the sensor performance parameters of the greenhouse. Environmental humidity (unit%):A1 representative10~15, L5 representative A2~20, A3 representative 25~20, A4 on behalf of 25~30, A5 on behalf of 30~35, A6 on behalf of 35-40, A7 on behalf of other (expressed by the numerical value of the same). On behalf of soil temperature (Unit degree Celsius):B15~10, B2 represent10 to 15, B3 represents15-20, B4 represents20~25,B5, represented the other. Soil moisture (%):C1represents18~19, C2 representative19 to 20, C3 represents 20~21, C4 represented from 21 to 22, C5 represents 22-23, C6 on behalf of 23 to 24, C7 represents the CO2 concentration (mg/L):D1 350 delegates and D2 represent 450~550, D3 on behalf of 550 to 650, D4 on behalf of 650 to 750, and D5 represented 750~850, D6 represents 850to 950, D7on behalf of other. Light intensity F unit KLUX):E1 on behalf of the weak light level (0~100), E2 on behalf of the light level (100~300), E3 on behalf of the intensity of light (300 or more). Ambient temperature (Unit degree Celsius):F1 representative of l0~15, F2 on behalf of 15~20, F3 on behalf of 20~25, F4 on behalf of 25 ~ 30, F5 on behalf of 30 ~35, F6 on behalf of 35-40, F7 on behalf of other.

Table 3. Sensor performance parameters

Sensor category	Installation location	measuring range	
Ambient humidity	middle part	0-100	0-15mA
soil emperature	15cm soil	-40-120	80μW
Soil humidity	15cm soil	0-100	80μW
CO_2 concentration	bottom	0-2000	4-50mA
illumination intensity	Central section	0-20	3.5vA
Environment emperature	Central section	0-50	0-15mA

Table 4. Part of data preprocessing

Serial number	the relative air humidity	soil emperature	Soil humidity	CO2 concentration	illumination intensity	air temperature	Time
1	A1	B3	C1	D1	E1	F1	2016/5/615:57:52
2	A2	B2	C2	D1	E1	F2	2016/5/615:57:52
3	A2	B4	C3	D1	E1	F2	2016/5/615:57:52
4	…	…	…	…	…	…	…

B. The generation and result analysis of association rules

Will improve IMApriori algorithm is applied to the pretreatment of data mining association rules, set the minimum support of 5%, minimum confidence is 50%, satisfy the minimum support and minimum confidence association rules.

Conclusion

In this paper, in the greenhouse environment temperature forecast the actual engineering background to improve the traditional Apriori algorithm. The experimental results show that the improved algorithm IMApriori enough good applied to smart greenhouse system, the generated association rules can help the relationship between greenhouse management and scientific analysis of the greenhouse environment parameters as the guidance of synthetic, the manual operation.

Reference

[1] Chen Guifen, Cao Liying, Ma Li. Application Status and Development Trend of Data Mining in Precision Agriculture. Journal of Jilin Agricultural University, 2008, 30 (4), 621-626.

[2] He Peng, Chu Yanhong. Research on Greenhouse Multi−parameter Control Algorithm Based on Data Mining. Journal of Agricultural Mechanization Research,2012 (10), 180-183.

[3] Malgaokar S, Surve S，Hirave T. Use of minging techniques to improve the effecitiveness of marketing of marketing andsales.Mumbai.India.IEEE.2013.1-5.

[4] Zhang Yaling, Chen Weimin, Zhang Peng, Hu Shunren, Huang Xiaowei, Zheng Wei. Overview of sensor fault diagnosis technology. Sensors and micro systems, 2009,28 (1), 4-6, 12.

[5] Zhang haifei ect, Application of multi-sensors information fusion in greenhouse humidity detection Transducer and Microsystem Technologies,2011,06,133-134

