Abnormal Data Mining Method in Environmental Monitoring Data of Animal Husbandry Farm

XU Xiao-hua[1], CHEN Chang-xi[2*]

{yg69030@163.com[1],xuxh69850@163.com[2*]}

(1.Tianjin Agricultural University,Tianjin 300384,China;

2.Tianjin Agricultural University,Tianjin 300384,China)

Abstract: In order to solve the problem of low accuracy of traditional anomaly data mining methods, this paper proposes an anomaly data mining method in the environmental monitoring data of livestock farms. Through collecting the environmental monitoring data of animal husbandry by sensors, after getting the environmental monitoring data, the environmental monitoring data is preprocessed, and the data after preprocessing is mined to complete the design of abnormal data mining method in the environmental monitoring data of animal husbandry. Compared with the traditional methods of outlier data mining, the experimental results show that the proposed outlier data mining method has higher mining accuracy.

Key words：Animal husbandry; Environmental monitoring; Abnormal data; Mining;

1 Introduction

With the adjustment of agricultural structure and the support of relevant national policies, agriculture, forestry, animal husbandry and fishery have developed rapidly, especially animal husbandry accounts for 1 / 3 of the overall output value. According to the data of the National Bureau of statistics, animal husbandry is maintained at an average value of about 30% every year, which not only increases the economic income of farmers, but also promotes the circulation of the national economy and increases the GDP [1-3]. Although large-scale breeding has gradually replaced the traditional extensive breeding mode, this trend still cannot change the extensive management mode in the breeding process. The ventilation, sewage, heating and other facilities are not complete, or even not, resulting in the poor growth environment of livestock, which is easy to cause animal disease, death, and poor quality of animal products [4-6].

With the concept of "healthy breeding", some domestic farms have improved their understanding of breeding management, strengthened supervision and ensured product quality. However, imperfect breeding management is still common, resulting in frequent occurrence of animal diseases. Traditional breeding, extensive breeding, and sky dependent breeding still account for a considerable proportion [7]. With the emergence of new animal husbandry production organization forms, the traditional family and single family farming forms decrease

EAI MOBIMEDIA 2020, August 27-28, Harbin, People's Republic of China

DOI 10.4108/eai.27-8-2020.2294506

year by year. It is of great practical significance to use high technology to carry out environmental monitoring on livestock farms for improving the quality of livestock products and reducing the occurrence of various animal diseases. By mining the abnormal data in the environmental monitoring data of the animal husbandry farm, we can find the abnormal environment in the animal husbandry farm, so as to deal with it in time. However, in the traditional animal husbandry environmental monitoring data, the method of abnormal data mining has the disadvantage of low mining accuracy, which needs further research. On this basis, a method of mining abnormal data in the environmental monitoring data of livestock and poultry farms is proposed.

2 Method of abnormal data mining in environmental monitoring data of livestock farm

2.1 Data collection

First of all, the environmental monitoring data of livestock farms are collected by sensors. The sensor is responsible for collecting data, and then transmitting it to the microcontroller, which then transmits it. Figure 1 shows the flow chart of the sensor.

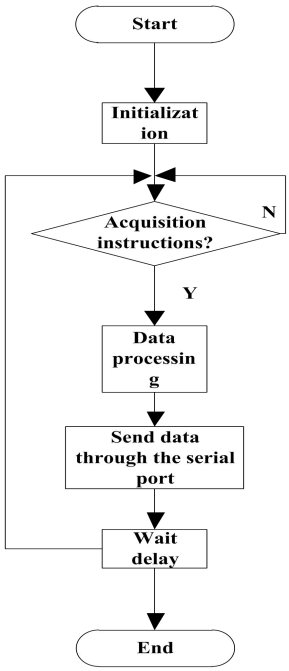

Figure 1 Processing flow of acquisition sensor

As can be seen from Figure 1, the upper computer can control the acquisition of data by sending instructions. The upper computer sends instructions, and MCU receives the collection

information instructions from the monitoring end, then packs the data and sends the data to the upper computer system. After obtaining the environmental monitoring data of animal husbandry farm, the pretreatment is carried out [8].

2.2 Data preprocessing

Data preprocessing includes data cleaning, data integration, data conversion, data discretization, data reduction and data reduction. The data preprocessing process is shown in Figure 2.

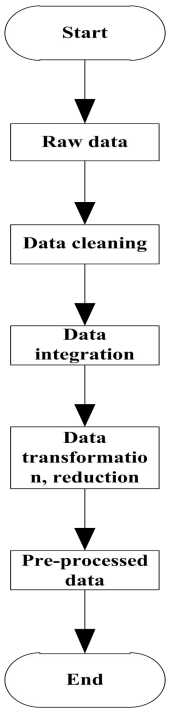

Figure 2 Flow chart of data processing

In the integration stage of data cleaning and data, first, simple data filtering and cleaning are carried out, errors and duplicate data are removed, and then relatively "clean" data is obtained, and then incomplete data is merged or completed, then data is standardized and normalized, and the final data values are all between [0,1]. So it provides a good foundation for the abnormal data mining of environmental monitoring data of livestock and poultry farms. [9].

The data reduction algorithm can well deal with attribute reduction under the condition of dynamic increase of environmental monitoring data in animal husbandry farm. The reduction effect is obvious, and the time is very short. The processing structure is shown in Figure 3.

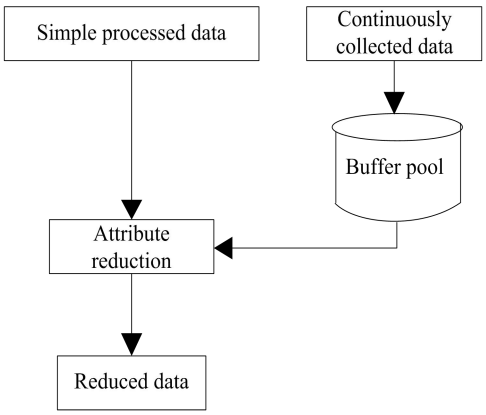

Figure 3 Data reduction processing structure

In Figure 3, the environmental monitoring data is reduced. The specific implementation process is as follows:

According to the rough set theory, mutual information equality can be used as the termination condition to find the relative reduction of knowledge. Based on this, a new calculation formula of attribute importance is proposed in combination with mutual information theory. This formula does not calculate the importance of data unrelated to attribute reduction, reduces the calculation time of attribute importance, and thus reduces the time of attribute reduction. The formula is described as follows:

$$\begin{aligned} &sig(a,B;C) = H(B \cup \{a\}) - H(B) \\ &= -\sum_{i=1}^{m_1} p(X_i)\log 2^{p(X_i)} + \sum_{i=1}^{m_2} p(Y_i)\log 2^{p(Y_i)} \end{aligned} \quad (1)$$

In formula (1), $p(X_i)$ represents the probability distribution of each object in $B \cup \{a\}$, $p(Y_i)$ represents the probability distribution of each object in B (where $p(X_i) = \frac{card(X_i)}{card(U)}$); $a = \left\{ b \middle| \left(b \in \{C - B\} \right) \wedge \left(\{b\} \cup B \in c_{ij} \right) \right\}$ and c_{ij} represent the elements in the difference matrix. The attribute importance is calculated by formula (1). The specific steps of the algorithm are as follows:

Input: livestock farm information $S = (U, C, V, f)$ and new object set (new livestock farm environmental monitoring data) $\{x_{s1}, x_{s2}, \cdots, x_{st}\}$; output: a attribute reduction B,

calculate attribute importance, if the conditions are met, the algorithm ends.

For the processing of missing value, Lagrange interpolation method is used: take out the five data before and after the missing value respectively. If the data before and after the missing value is empty or nonexistent, directly round off the data. According to the ten data, form a group, and use Lagrange difference formula to interpolate the data. The Lagrangian difference formula is shown in formula (2):

$$\begin{cases} L_n(x) = \sum_{i=0}^{n} l_i(x) y_i \\ l_i(x) = \prod_{j=0, j\neq 1}^{n} \dfrac{x - x_j}{x_i - x_j} \end{cases} \quad (2)$$

Where, x represents the x data, y_i represents the sample value, and $l_i(x)$ represents the Lagrangian polynomial.

After that, the data is normalized. If the data values of different dimensions are different or distributed unevenly, the training time will be longer and the data processing effect will be affected. Normalization processing in order to solve this problem, it will be used in the abnormal data mining of environmental monitoring data of animal husbandry farm, and the environmental monitoring data will be normalized using formula (3):

$$NL = \frac{L - \min(x)}{\max(x) - min(x)} \quad (3)$$

In formula (3), L represents the original value before data normalization, $\max(x)$ represents the maximum value of all values under the dimension before data normalization, $min(x)$ represents the minimum value of all values under the dimension before data normalization, and NL represents the normalized characteristic value. Through the above process, realize the pretreatment of the environmental monitoring data of the livestock farm [10].

2.3 Data mining

Data mining is done by clustering and fast computing algorithm. In clustering algorithm, outliers are usually regarded as leaf nodes in clustering feature tree, and the density of leaf nodes is very low, usually set a fixed threshold. If the number of sample data objects of one leaf node is less than the set threshold or the number of samples in this leaf node is the least compared with other leaf nodes, then it can be considered that it is very possible that the leaf

node is an outlier, and then the leaf node is added to the outlier data set. If the set threshold changes, so that the sample object that has been included in the abnormal data set does not meet the condition of being an abnormal point, it should be re classified into the clustering feature tree. First, cluster the original data set, and calculate the size of the local anomaly factor, and it is also related to the density of its surrounding neighbors.

The density based lof algorithm mainly involves the calculation of $k-$ distance, $k-$ distance neighborhood, reachable distance and reachable density of data objects. The $k-$ distance of data object p can determine which data objects in the data set can constitute p 's neighbors. To a certain extent, it measures the density of the area around p . if the $k-$ distance of p is larger, it indicates that the area around p is sparse; on the contrary, if the $k-$ distance of p is smaller, it indicates that the area around p is dense. Figure 4 vividly explains the $k-$ distance of p and the neighbor object of p .

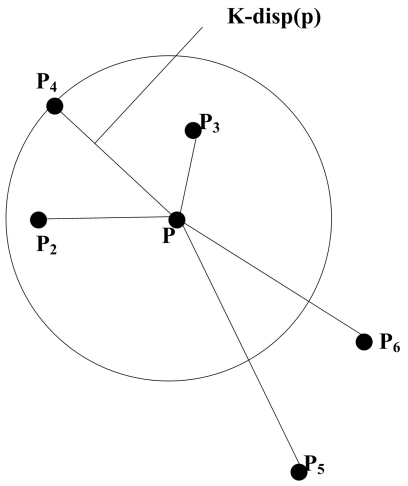

Figure 4 Neighborhood of $k-$ distance and $k-$ distance of p

In Figure 4, a total of six data objects are set, which are $p\left(p_1\right)$, p_2, p_3, p_4, p_5 and p_6 respectively. At the same time, the distance from each data object to p is marked. Suppose n represents the total number of objects in the sample data set, there are $k-1$

objects in the circle, and the remaining $n-k-1$ objects are distributed outside the circle, then the distance between the object p_4 and the object p_1, which is just on the circumference boundary, is the $k-$ distance between the object p_1 and the object p_1. The data object located on and within the circle boundary is defined as the neighbor object of object p_1, for example, in Figure 4 (p_2, p_3, p_4), it is the neighbor object of object p. According to figure 4, the following definitions are given:

Defines the k distance of data object p. For any positive integer k in the data set, the k distance of object p is defined as the Euclidean distance between p and an object o, where object o belongs to data set D and the following conditions need to be met simultaneously: there are at least k data objects $o' \in D \setminus \{p\}$ and $d(p,o') \le d(p,o)$; there are at most $k-1$ data objects $o' \in D \setminus \{p\}$ and $(p,o') < d(p,o)$. When calculating the $k-$ distance of object p, you can first calculate the distance of all objects reaching p respectively, and then select k minimum distances. The maximum distance of k objects already selected is defined as the $k-$ distance of p.

Defines the $k-$ distance neighborhood of data object p. The definition of $k-$ distance of given object p in data set is the collection of all data objects within its $k-$ distance neighborhood whose distance from object p does not exceed $k-$ distance, namely:

$$Nk-dis\tan ce(p)(p)=\{q\in D\setminus\{p\}|d(p,q)\le k-dis\tan ce(p)\} \quad (4)$$

Defines the reachable distance of data object p. For a given natural number k, object p

is defined relative to object o as:

$$reach-distk(p,o)=\max\{k-dis\tan ce(o),d(p,o)\} \quad (5)$$

Defines the local reachable density of data object p. For the given positive number k, the reciprocal of the average reachable distance between the object p and its neighborhood relative to its distance $k-$ is defined as the local reachable density, and the calculation method is as follows:

$$lrd_k(p)=\frac{1}{\frac{\sum_{q\in N_k(p)} reach-dis_k(p,q)}{|N_k(p)|}} \quad (6)$$

In formula (6), $lrd_k(p)$ represents the local reachable density, and $\frac{\sum_{q\in N_k(p)} reach-dis_k(p,q)}{|N_k(p)|}$ represents the average reachable distance between the object p and its neighborhood relative to its $k-$ distance. Define the local outlier of data object p. For a given positive number k in the data set, the calculation method of the local outlier factor $LOF_k(p)$ of the object p is as follows:

$$LOF_k(p)=\frac{\sum_{q\in N_k(p)}\frac{lrd_k(q)}{lrd_k(p)}}{N_k(p)} \quad (7)$$

The local outlier indicates the abnormal degree of an object. If the larger the local outlier is, the greater the abnormal degree of the object is. Otherwise, the smaller. Because the local exception factor of each object of lof algorithm is related to the environment of the object, in the dynamic incremental database environment, new data objects often affect the local exception factor of some objects in the original data set, so it is necessary to call lof algorithm to recalculate the local exception factor for all data objects. However, the new data object will

only affect the local abnormal factors of some data objects, but not all data objects. Therefore, it is improved. If the data in the database is updated, only the affected data objects need to be recalculated, which greatly improves the speed of abnormal data mining.

For a given sample data set D, for a known object q in D, whenever an object p is added, the $k-$ distance neighborhood of q changes due to the addition of object q, which also changes the reachable density and local outlier of q, and then a series of changes of q will cause a series of changes. Here we define the affected object set: In a given sample object D, due to the addition of new object p, the local outlier of data in set D changes, then the set of data is called the affected object set.

When adding object p to data set D, it is necessary to recalculate $k-$ distance neighborhood and local outlier factors for the affected object set, and update the neighborhood in the following three cases. If q is the affected object and there are $d(q,p)=k-dis(q)$, then p can be directly added to the neighborhood of q. If q is the affected object, and there are $d(q,p)<k-dis(q)$, and at most $k-2$ data o make $d(q,o)<k-dis(q)$ exist at the same time, then p can also be directly added to the neighborhood of g. If q is the affected object and $d(q,p)<k-dis(q)$, $k-1$ data are met at the same time. Make $d(q,o)<k-dis(q)$, then p can be added to the neighborhood of object q, and the object farthest from q can be deleted. According to the above definition, the abnormal data in the environmental monitoring data of animal husbandry farm are clustered.

Cluster the data in the original data set: first, cluster the original data set, because according to the definition of the abnormal data, the abnormal data is some data that accounts for a small part of the data set, so after clustering the original data, the calculation of the normal data that accounts for a large part of the data can be eliminated, so the normal data cluster N and abnormal data U are formed by clustering , continue processing until the end of clustering, forming data clusters $N_1,N_{2,}\cdots,N_i$ and $U_1,U_2,\cdots,U_i$.

Clustering new data: when dealing with abnormal data mining in the dynamic

incremental data environment, the newly added data are usually normal data, only a small amount of abnormal data. The algorithm avoids the high complexity of re clustering when new data arrive. Make full use of the normal cluster N_i, calculate the center object O_i in each cluster N_i, then calculate the distance from each newly added object to the center object O_i, and record the cluster when the minimum distance is obtained. If the minimum distance calculated is less than the set radius value ε_i, the object can be directly added to the cluster, if the distance can not meet the conditions for adding any cluster , a new abnormal cluster is formed. Add all new data objects to a normal or abnormal cluster.

Calculate the local outlier factors for the data in the abnormal cluster: after the clustering, the initial abnormal cluster $U_1,U_2,\cdots,U_i$ and the newly added abnormal cluster $U_1',U_2',\cdots,U_i'$ are formed. Because the amount of data in these abnormal clusters is not very large, the data can be directly calculated as $k-$ distance neighborhood, local reachable density and local outlier factors.

In order to improve the speed of abnormal data mining in environmental monitoring data, the lof algorithm and DBSCAN algorithm are combined to improve the algorithm. First of all, DBSCAN algorithm is used to cluster the original data set. The dense data objects are subtracted from a large number of data sets, and the sparse data objects that are not clustered are retained to form abnormal clustering. On the basis of clustering, the data in the cluster is judged, and the abnormal data is added to the abnormal cluster. Finally, lof algorithm is used to calculate the local abnormal factors for the data objects in the abnormal cluster.

After clustering the new data objects, we can judge which data objects are dense and which are sparse, and add the sparse data objects to the new abnormal data set U'. Then, on the basis of the above improved algorithm clustering, we calculate the average of the cluster centroid and the distance from all data objects to the centroid of the cluster , and then calculate the distance between each data object in the cluster and the cluster centroid. If the distance is less than the average value, the data object is considered normal, otherwise, it is considered abnormal data object, and the abnormal data object is added to the new abnormal data set port. Because the main research object of lof algorithm is the data object with sparse distribution in the data set, so the data object with sparse distribution in the original data set and the data object with sparse distribution in the new data set need to use lof algorithm to calculate the

local abnormal factor.Because the sparse data objects in the original dataset (the data objects in U) have calculated the local exception factor with the lof algorithm in advance, when the new data objects arrive, the data objects in the new exception dataset U' and the exception dataset U are affected by the data objects in the new exception dataset U', resulting in the change of the local exception factor. Therefore, the data object in U can be recalculated with the local exception factor and sorted according to the size of the local exception factor, so the former data object is the real exception data object.

At this point, the design of abnormal data mining method in environmental monitoring data of livestock farm has been completed.

3 Experiment

In order to verify whether the anomaly data mining method in the monitoring data of animal husbandry farm has higher mining accuracy, the paper compares the proposed anomaly data mining method with the traditional anomaly data mining method

3.1 Experimental process

The experimental operation environment is shown in Table 1.

Table 1 Experimental environment parameters

Serial number	Items	Parameter
1	ComputerCPU	Pentium(R) Dual-Core T4300@2.10GHz
2	RAM	2.0GB
3	Hard disk	320G
4	Software operating environment	Windows 7,VC++6.0

The experiment was carried out in the above experimental environment. First of all, to ensure the smooth progress of the experiment, we use the shuttle data set, which contains 43500 data objects. Now take 3000 data objects as the initial data set, and take 10, 30, 100, 200, 500 as the number of newly added data objects. In the case of $Esp = 18.245, Minpts = 20$, when the new data objects arrive, the running time result of the algorithm used is shown in Table 2.

Table 2 Algorithm running time

Serial number	Number of new objects	Algorithm running time
1	10	0.174s

2	30	0.298s
3	100	0.659s
4	200	1.173s
5	500	2.845s

It can be seen from table 2 that with the increase of new data objects, the calculation time of the algorithm does not increase significantly. The results show that the algorithm is efficient and meets the calculation requirements.

On this basis, in view of the five groups of animal husbandry environmental monitoring data with different amount of abnormal data (10, 20, 30, 40, 50), the mining accuracy of this paper's abnormal data mining method and traditional abnormal data mining method is compared.

3.2 Analysis of experimental results

Compare and analyze the mining accuracy of the proposed anomaly data mining method and the traditional anomaly data mining method. The comparison results are shown in Figure 5.

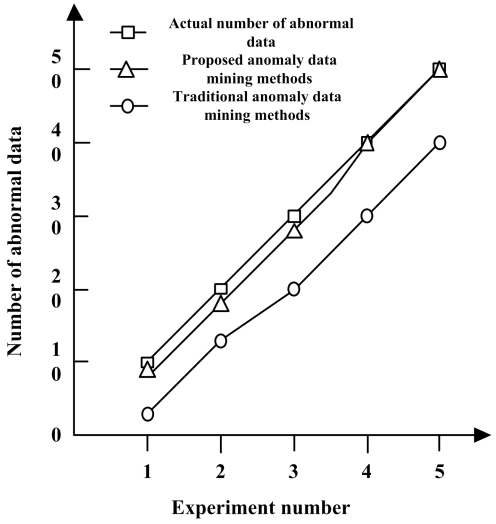

Figure 5 Comparison results of mining accuracy

As shown in Figure 5, the maximum difference between the number of abnormal data mined by traditional exception data mining method and the actual number of abnormal data is 10; the maximum difference between the number of abnormal data mined by the proposed exception data mining method is 2. Through comparison, it is found that the proposed anomaly data mining method has higher mining accuracy.

4 Concluding

In view of the low accuracy of the traditional methods of outlier data mining, a method of outlier data mining in the environmental monitoring data of animal husbandry is proposed. Through the comparative experiment, compared with the traditional abnormal data mining methods, the experimental results show that the proposed abnormal data mining method has a higher mining accuracy, hoping that it can provide a certain reference value for the research of abnormal data mining in animal husbandry environmental monitoring data.

5 Fund projects

Post Scientist of National Broiler Industry Technology System（CARS-41-G25）

The Science and Technology Planning Program of Tianjin Science and Technology Bureau, China (18ZXRHNC00080)

Information Post of Tianjin Dairy Industry Technology System Innovation Team(ITTCRS2020023)

Reference

[1] LI Ping. Anomaly Detection Method for Based on Quantum K-Means Clustering[J]. Computer Knowledge and Technology, 2018,14(06):198-199.

[2] Pan Zhian, Liu Qingjie, Wang Xiaoying, et al. Anomaly Detection Model of Satellite Data Based on Improved Clustering Algorithm[J]. Bulletin of Science and Technology, 2018,34(07):128-131.

[3] LIANG Zhihao, WU Jianghong, XIE Zili. Variable Frequency Room Air Conditioner Operation Pattern Recognition and Data Mining[J]. Journal of Mechanical Engineering, 2019,55(06):194-202.

[4] YANG Pinlin. Mining method for target feature data in color image database[J]. Journal of Shenyang University of Technology, 2018,40(1):60-64.

[5] LIU Xingzhong, HAN Xin, LI Rong, et al. Research on Product Family Planning Method Based on Data Mining[J]. Machine Design & Research, 2018,34(05):7-11.

[6] Wang Lijia. Detection of offensive information on Internet via data mining[J]. Journal of Hunan Industry Polytechnic, 2019,19(01):25-27,46.

[7] LIU Binghan, FU Zhongguang, WANG Yongzhi, et al. Big Data Mining Technology Based on Parallel Algorithm and Its Application in Power Plant Boiler Performance Optimization[J]. Journal Of Chinese Society Of Power Engineering, 2018,38(06):431-439.

[8] HE Shan, WANG Wenda, ZHANG Wei, et al. Evaluation Method for Operating State of Low-voltage Distribution Network Based on Data Mining[J]. Guangdong Electric Power, 2019,32(05):80-86.

[9] LI Yijin. Application of Data Mining Technology in Construction of Power Network Operation Monitoring Platform[J]. Telecom Power Technologies, 2019,36(04):283-284.

[10] HAN Jingling. Research on Interest Classification of Online Shopping Users Based on Data Mining[J]. Computer Simulation,2018,35(07):418-421.