Agricultural Fertilizer Application Behaviour and its Influencing Factors based on Data Mining and K-means algorithm

Jie Shang, Mingjuan Shang*
Northeast Forestry University, Harbin 150040, China
*Corresponding author(E-mail: 782468269@qq.com)

Abstract
The practice of agricultural development shows that fertilization, especially the application of chemical fertilizer, is the fastest and most effective measure to increase production. However, unreasonable fertilization often results in serious environmental problems. In this paper, the author analyse the agricultural fertilizer application behaviour and its influencing factors based on data mining and k-means algorithm. Influencing factors of fertilization behaviour mainly caused by household personal characteristics, operating characteristics and farmer's perception of environmental pollution. Through empirical analysis, the author calculated the index weight of the Farmers fertilizer influencing factors and put forward related suggestions.

Key words: Agricultural fertilizer, K-means algorithm, Farmers behaviour, influencing factors

1. INTRODUCTION

With the development of computer technology in various fields, a variety of formats (pictures, music, video and documents, etc.) of the massive data generated and stored in the database. Mining useful knowledge from massive data, and applying it to the production practice of human society, has become a key task in the field of information industry(Ali,2014). But too much data also brings a problem as data is rich, but lack of effective information, it means the cost of knowledge is too high. Change in demand driven technology. Distributed data mining provides a solution to this problem(Delgado, 2015). One of the key steps to realize the distributed data mining system is to transfer the traditional data mining algorithms to the distributed computing platform. After decades of development, a variety of algorithms have emerged in the field of distributed data mining, which have different needs and different tasks(Aurelio, 2013). These algorithms can be improved to adapt to the new distributed environment, which is an important step in the whole distributed data mining. In this paper, we analyze agricultural fertilizer application behaviour and its influencing factors based on data mining and k-means algorithm.

The protection of grain production is related to the national economy and the people's livelihood, especially in China, more and more people, in the limited land to meet the growing needs of the population, can only improve the grain yield(Gopal, 2014). The practice of agricultural development in the world shows that fertilization, especially the application of chemical fertilizer, is the fastest and most effective measure to increase production. Since the beginning of the last century to achieve the artificial synthesis of chemical fertilizers, crop yields per unit area of the world has been greatly improved(Ilyoo, 2013). According to the United Nations Food and Agriculture Organization, in developing countries, fertilization can increase the yield of grain crops 51%. The input of chemical fertilizer is the basis to ensure the grain yield, but the relevant data and a large number of studies have shown that the fertilization behavior of farmers in China is unreasonable. Combined with the literature and field research, we know that the main problems and difficulties of farmers' grain fertilization in our country are the large intensity of fertilization, the low utilization rate of chemical fertilizer, the low level of social service and the lack of scientific and technological extension(Joshua, 2015). At present, China's agricultural socialization service is still not standardized, various industries and departments. Besides, the agricultural scientific research departments, scientific research and technology promotion department because many factors can not be applied to the production as soon as possible(Milad, 2015). The chemical fertilizer in agricultural productivity continues to improve, increasing the amount of chemical fertilizer is due to some economic factors, but with the increase of fertilizer, has also brought environmental problems. In general, the higher the amount of fertilizer application, the more the amount of loss to the environment, the higher the degree of pollution of the ecological environment. The heavy use of chemical fertilizers in the farming industry has caused the deterioration of the quality of agricultural soil, water, rivers, lakes, bays and atmosphere, and has become the main cause of agricultural non-point source pollution in China.

2. IMPROVED K-MEANS ALGORITHM AND ITS PARALLEL IMPLEMENTATION

2.1 K-means++ algorithm
Compared with the classical algorithm, the K-means++ algorithm is different from the initialization strategy. Algorithm from the data set of random selection of the first center, and the rest of the central point is through the probability to filter; the probability is proportional to the distance between the node and the initial center. The pseudo code of the algorithm is as follows:

Algorithm 1: K-means++
Input: k, X={x1,x2,x3…xn}
Output: C={c1,c2,c3…cn}
Step1: C−⇒0
Step2: Randomly selected from the collection node x as the first initial node C=C∪{x}
Step3: while (|C|<k) do
Step4: Traversal x∈X, select the node as a new initialization center. C=C∪{x}
Step5: end while, |C|=k.
Step6: The element in the set C is used as the initialization center node, Lloyd’s iteration according to the standard K-means algorithm, until convergence of the algorithm.

The algorithm makes use of such an intuitive fact: a good clustering algorithm, the cluster should be relatively dispersed. So in choosing a new cluster center should be as far as possible to choose from the other center of the node. Both in theory and in engineering practice, K-means++ has been proved to be far superior to the classic K-means. Although K-means++ algorithm has achieved great efficiency, but it also has shortcomings, limiting its application, that is, the inherent nature of the algorithm. When the input data set for the D dimension, the size as n, the number of clusters for K, the total running time of the algorithm is O (nkD), which is similar to the time spent on a single s’ Lloyd iteration. In spite of this, the algorithm can not be implemented in parallel. In the initialization, a node can is chosen to be the probability of the i-th center depends entirely on before I1 a center of the selection. The implementation of the original K-means++ algorithm, can only run on a single machine, scanning the local all the data to generate the initial center. When dealing with massive data, this limit is even more serious. First, with the growth of the data set, the number of categories will increase. For example, in a typical case, the algorithm has to scan all the data at once to complete the initialization action by integrating the millions of nodes into 100 or 1000 clusters. If the residual algorithm steps (such as Lloyd’s iterative process) can through parallel implementation, such as MapReduce to improve running speed, then initialize spent time will become performance bottlenecks in the clustering process. So in a lot of applications, the researchers used similar K-means++ other initialization algorithm, they can achieve efficient parallel. In addition, in the K-means++ algorithm, the Euclidean distance is still used to calculate the distance of each data node to the center point. In the calculation process, the different attributes of the node are treated specially. However, in practical applications, some of the attributes must have a greater impact on the clustering results, and some of the properties are weak.
2.2. Improved algorithm

Attribute weight is the influence degree of the clustering result. If the weight is large, the influence of the attribute to the clustering result is bigger. This is more in line with the practical application of the situation. In order to calculate the attribute weights, the concept of information entropy is introduced. The information entropy to represent objects contained in the amount of information, entropy is small, is included in the greater the amount of information, distribution to the weights of the attributes should also bigger. Conversely, the smaller the amount of information, the smaller the value should be. An input data set contains n number as m dimensional objects, the input data can be represented as follows.

\[
X = \begin{bmatrix}
    x_{11} & x_{12} & \cdots & x_{1m} \\
    x_{21} & x_{22} & \cdots & x_{2m} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n1} & x_{n2} & \cdots & x_{nm}
\end{bmatrix}
\]  

(1)

In order to calculate the attribute entropy and need with different dimension attribute values can be compared with the various properties of 0-1 standardization:

\[
b_{ij} = x_{ij} / \sum_{i=1}^{n} x_{ij}
\]

(2)

We can get the attribute weight matrix B

\[
B = \begin{bmatrix}
    b_{11} & b_{12} & \cdots & b_{1m} \\
    b_{21} & b_{22} & \cdots & b_{2m} \\
    \vdots & \vdots & \ddots & \vdots \\
    b_{n1} & b_{n2} & \cdots & b_{nm}
\end{bmatrix}
\]

(3)

The information entropy of the calculation of the j dimension attribute

\[
s_j = -p \sum_{i=1}^{n} b_{ij} \ln b_{ij}
\]

(4)

Calculate the difference value of the j dimension attribute
The difference value indicates the influence degree of the attribute to the clustering result, when the difference value is $C_j$ is smaller, then the $S_j$ is bigger, and the influence degree of the attribute to the clustering result is smaller. Calculate the weight of the $j$ dimension attribute:

$$w_j = \frac{C_j}{\sum_{j=1}^{m} C_j}$$

(6)

The weighted Euclidean distance formula of the object is expressed as:

$$d(x_i, x_k) = \sqrt{\sum_{j=1}^{m} w_j (x_{ij} - x_{kj})^2}$$

(7)

By multiplying the weights in the formula, we can describe the contribution of the attributes to the distance between the objects. The greater the attribute value, the greater the contribution, which is more in line with the actual application situation.

3. INFLUENCING FACTORS OF FERTILIZATION BEHAVIOR

3.1. Theoretical basis of fertilization behavior

Based on the theory of "profit maximization" and "risk aversion" as the basic theoretical basis for the study of fertilization behavior. On the basis of this theory, it is assumed that he is willing to reduce the amount of chemical fertilizer applied when the excessive use of chemical fertilizers to farmers or their families is not maximized. The farmer is the main investor of the agricultural production factor, is the main operator of grain production and fertilizer application. Therefore, the most direct factor affecting the level of fertilizer application. First, find the influencing factors in the interpretation of chemical fertilizers by farmers from the family characteristics, in terms of basic household resources under certain external environment, and external factors of the farmers have a certain impact on fertilizer consumption.

![Figure 4. Risk aversion theory](image)

The farmer is "rational economic man", in our existing rural production system, farmers are relatively independent producers, its management and decision-making behavior has great autonomy, economic behavior of farmers is the largest family in pursuit of the production and operation of the income maximization and family investment profit as the goal. However, due to the incomplete of the rural market environment and the asymmetry of the information, the farmer's behavior is relatively maximized under the restriction of the specific social and economic environment. Because of China's rural social security system and mechanism of agricultural insurance are not perfect, and the agricultural production scale is small, the ability to resist risks is weak, there is a greater risk in agricultural production, therefore, farmers will affect the rational decision making on fertilization of grain production in the face of agricultural production and life's risk attitude. Although some choice is optimal under certain conditions, but if the minimum risk due to the existence of the expected return is less than acceptable, farmers will give up this decision to determine the optimal conditions, but the choice of a less risky decision.

3.2 Influencing factors of fertilization behavior

1) **The influence of the basic characteristics of the household on fertilization**: As the decision maker of the purchase and application of chemical fertilizer, the basic characteristics of the household will
affect the purchase and use of the behavior. The basic characteristics of the household include the education level of the head of household, the length of farming and the attitude towards risk. The farmers consider not only the utility maximization in the consumption of chemical fertilizer, and the fertilizer consumption behavior is influenced by the past habits and past income and consumption levels, education level headed fertilization therefore headed farming experience and age differences in personal characteristics may affect farmers determines the quality of the labor force to a great extent. The higher the education level of the head of household, the ability to accept agricultural knowledge is stronger, they will be more inclined to scientific fertilization and farming, so that the amount of fertilizer fertilizer showed significant difference in different levels of education among farmers.

2) The influence of farmers’ operating characteristics on fertilization: Now the family income is more and more diversified, simply rely on agriculture income of farmers is relatively small, the income structure of farmers farmers response resources allocation proportion, to a certain extent at the same time, also from the side reaction of the farmers to grow grain consumption in order to meet the purpose of rations, or by selling food access income. Generally, with the increase of farmers’ income, which is used for agricultural production expenditure will also increase in income, under certain conditions, if the household farming income accounted for the larger proportion of their income on agricultural production is more dependent on high, so fertilizer amount will be higher, and vice versa. In addition, family labor, refers to the age population in age between households, due to the application of fertilizers is an important part of agricultural production and labor, so labor supply situation may affect the fertilizer application level to a certain extent.

3) Soil quality: It is also a very important factor that affects the amount of fertilization of farmers, whether the soil is rich in organic matter, water quality and so on, which has a significant impact on the demand of chemical fertilizer in the production process. If the soil quality is low, the land provided by the relatively small amount of chemical fertilizers, that is, the amount of fertilizer needed to invest more. Taking the land will affect the poor farmers in order to ensure the grain yield, grain yield, will increase the use of chemical fertilizer. This research area is the major grain producing areas in Henan Xinxiang County, according to the actual situation of the soil testing station to provide information and combining the investigation of soil quality in Xinxiang is in clay, and sand. The mixed soil. In the process of the investigation, the main way for the farmers to use fertilizer in less sand is to apply fertilizer and less fertilization. Because once in the sand with a lot of fertilizer, fertilizer and water infiltration into the ground, so that farmers in the barren land without large amounts of fertilizer, so the soil quality of chemical fertilizer is not significant.

4) Farmers’ awareness of environmental pollution: According to the rational choice and utility theory, if farmers on the current crop of excess fertilizer have correct cognition, farmers will be “utility maximization” principle as the theoretical basis of fertilizer, fertilizer if farmers think that home grown grain has been excessive and on food production yield began to decline, then influence of chemical fertilizer, they should have a higher reduced crop fertilizer quantity of grain will. In addition, if farmers recognize the excessive application of fertilizer in the process of food production may cause some damage to the environment, and excessive fertilization on their non utility maximization, they are also willing to reasonable and scientific application of fertilizers in order to achieve the purpose of environmental protection. On the contrary, if the farmers for the current utility situation of chemical fertilizer and grain production for lack of understanding or excessive fertilization would lead to the water pollution, air pollution, affecting the quality of agricultural products and human health also do not understand, then they will reduce the amount of chemical fertilizer deficiency.

4. EMPIRICAL ANALYSIS

4.1 Index system

In the course of the investigation found that the higher the level of education of the head of household, the ability to accept the relevant knowledge of agriculture is stronger, it will be more inclined to fertilization and scientific farming, but also know how to apply fertilizer. The following 5 years engaged in agricultural farmers, their education level is higher, and the high degree of industry farmers, tend to large one-time fertilization, from the agricultural time in more than 10 years’ old, risk averse, in the grain growing process will excessive application of chemical fertilizers to ensure food production. Farmers are limited rational economic people, they make decisions by the information constraints. Farmers to buy fertilizers, if a better understanding of the agricultural knowledge such as fertilizers on brand, quality and nutrient content, they will be reasonable collocation of fertilizer varieties, reasonable fertilization, rather than simply relying on the excessive application of nitrogen fertilizer to maintain and improve grain yield. Here refers to the training of agricultural technology extension departments of the relevant knowledge of farmers training. If farmers have participated in agricultural technology training, then it is possible for farmers to understand some of the technical aspects of information,
such as fertilization technology information, new varieties of information, so farmers will be more reasonable fertilization. According to the survey, the number of trained farmers was less than that of the farmers who had not been trained. This is because the current fertilizer is mainly based on chemical fertilizer, especially nitrogen fertilizer, in order to maintain food production, farmers in the grain production process a lot of fertilizer. If farmers know the current situation, as a rational economic man, it will be reasonable with the nutrient content, thus more reasonable, scientific fertilization, rather than relying on a large number of fertilizer to ensure food production.

Table 1. Index system

<table>
<thead>
<tr>
<th>first-level indicators</th>
<th>second-level indicators</th>
<th>third-level indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farmers fertilizer rate influencing factors A</td>
<td>Household characteristics B1</td>
<td>From agricultural time C1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Household risk attitude C2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Education level C3</td>
</tr>
<tr>
<td>cognition of chemical fertilizer B2</td>
<td>Agricultural technology training C4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Understanding of fertilizer brands C5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fertilization method C6</td>
</tr>
<tr>
<td>Farmers' awareness of pollution B3</td>
<td>Whether use of organic manure C7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Environmental pollution awareness C8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fertilizer utilization rate C9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Information channel C10</td>
</tr>
</tbody>
</table>

Analytic hierarchy process to determine the decision maker's experience to quantify, a decision-making problem is described as a hierarchical model, and then quantitative analysis and calculation. With respect to a CGI elements, we use 1 ~ 7 scaling the expert assignment method each element of B interlayers in pairwise comparison and judgment to determine the intermediate layer B element of a CGI elements relative importance, construct pairwise comparison judgement matrix A. Judgment matrix A (relative importance of each factor relative to the scientific and technological innovation and entrepreneurial team competence), see table 2.

Table 2. Judgment matrix A

<table>
<thead>
<tr>
<th></th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1/4</td>
<td>1/3</td>
</tr>
<tr>
<td>B1</td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B2</td>
<td></td>
<td></td>
<td>1/2</td>
</tr>
<tr>
<td>B3</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

3.2. Index weight
Calculate the geometric mean of all the elements in the matrix A.

\[
\bar{w}_i = n \prod_{j=1}^{n} a_{ij}
\]

We can obtain:

\[
\bar{w}_i = \left(\bar{w}_1, \bar{w}_2, \cdots, \bar{w}_n\right)^T
\]

(12)

To be normalized, that is to calculate:

\[
W_i = \frac{W_i}{\sum_{j=1}^{n} W_i}
\]

(13)

Obtain:

\[
w = \left(w_1, w_2, Kw_n\right)^T
\]

(14)

This is the relative weight of each element of Bi to B:

\[
w = \left(w_1, w_2, w_3\right)^T = \left(0.123, 0.557, 0.320\right)^T
\]

(15)

Calculation of the maximum eigenvalue of the \(\lambda\) matrix A Max
\[ \lambda_{\text{max}} = \sum_{i=1}^{n} \left( \frac{A W_i}{n w_i} \right) \]

Among them \((A W)_i\) is the first \(i\) element of vector \(A W:\)

\[
AW = \begin{bmatrix}
1 & 1/4 & 1/4 \\
4 & 1 & 2 \\
3 & 1/2 & 1
\end{bmatrix}
\]

\[
\lambda_{\text{max}} = \sum_{i=1}^{n} \left( \frac{A W_i}{n w_i} \right) = 3.01832
\]

Judgment matrix to calculate the judgement matrix, the largest eigenvalue corresponds to the characteristic to (weight), and then calculate the consistency checking and ratios, if \(C \cdot R < 0.1\) that matrix consistency test requirements, if not satisfied, continue to modified until can meet only.

The consistency check of the calculation results is \(C\cdot I\), the formula and the calculation result is:

\[
C \cdot I = (\lambda_{\text{max}} - n)/(n - 1) = 0.009162
\]

Identify the corresponding \(R.I = 0.58\) by Table 3.

<table>
<thead>
<tr>
<th>(n)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>R.I</td>
<td>0</td>
<td>0</td>
<td>0.58</td>
<td>0.90</td>
<td>1.12</td>
<td>1.24</td>
<td>1.32</td>
</tr>
</tbody>
</table>

The random consistency ratio, the formula and the result are calculated:

\[
C \cdot R = C\cdot I / (R.I) = 0.009162 / 0.58 = 0.01579 < 0.1
\]

According to the random consistency ratio \(C \cdot R < 0.1\), indicating that the consistency of the judgment matrix \(A\) can be accepted, the middle layer of CGI a weight value \(w = (0.123, 0.557, 0.320)\), \(t\) is acceptable.

Similarly, compared with the middle layer \(B\) elements, we use \(1 \sim 7\) scaling the expert assignment method of the elements of the underlying \(C\) for pairwise comparison and judgment to determine the relative importance of underlying \(C\) elements of \(B\) elements of middle layer, we construct pairwise comparison judgment matrix \(B_1, B_2, B_3, B_4\) can also be obtained the weight of all elements in the third layer, and the judgment matrix consistency test, see table 4 and table 5 and table 6.

**Table 4. Calculation of weight in the second index system C1, C2, C3**

<table>
<thead>
<tr>
<th>(B_1)</th>
<th>(C_1)</th>
<th>(C_2)</th>
<th>(C_3)</th>
<th>(W_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1)</td>
<td>1</td>
<td>3</td>
<td>1/2</td>
<td>0.334</td>
</tr>
<tr>
<td>(C_2)</td>
<td>1/3</td>
<td>1</td>
<td>1/3</td>
<td>0.142</td>
</tr>
<tr>
<td>(C_3)</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0.525</td>
</tr>
</tbody>
</table>

**Table 5. Calculation of weight in the second index system C4, C5, C6**

<table>
<thead>
<tr>
<th>(B_2)</th>
<th>(C_4)</th>
<th>(C_5)</th>
<th>(C_6)</th>
<th>(W_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_4)</td>
<td>1</td>
<td>1/3</td>
<td>2</td>
<td>0.252</td>
</tr>
<tr>
<td>(C_5)</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0.589</td>
</tr>
<tr>
<td>(C_6)</td>
<td>1/2</td>
<td>1/3</td>
<td>1</td>
<td>0.159</td>
</tr>
</tbody>
</table>

**Table 6. Calculation of weight in the second index system C7, C8, C9, C10**

<table>
<thead>
<tr>
<th>(B_3)</th>
<th>(C_7)</th>
<th>(C_8)</th>
<th>(C_9)</th>
<th>(C_{10})</th>
<th>(W_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_7)</td>
<td>1</td>
<td>5</td>
<td>1/3</td>
<td>2</td>
<td>0.245</td>
</tr>
<tr>
<td>(C_8)</td>
<td>1/5</td>
<td>1</td>
<td>1/7</td>
<td>1/2</td>
<td>0.063</td>
</tr>
<tr>
<td>(C_9)</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>0.572</td>
</tr>
<tr>
<td>(C_{10})</td>
<td>1/2</td>
<td>2</td>
<td>1/5</td>
<td>1</td>
<td>0.119</td>
</tr>
</tbody>
</table>

The bottom element \(C\) for the synthetic weight goals of \(A\), each index weight see table 7.

**Table 7. Each index weight**

<table>
<thead>
<tr>
<th>first-level indicators</th>
<th>second-level indicators</th>
<th>third-level indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farmers fertilizer rate influencing factors (A)</td>
<td>Household characteristics (B_1)</td>
<td>(C_1) (0.334)</td>
</tr>
</tbody>
</table>
5. CONCLUSIONS

Based on the utility theory, this study investigated the influence factors of farmers' Fertilization behavior. According to the research findings, and combined with the results of the survey data, the farmers' cognition for the fertilizer utilization rate, nearly 70% of the farmers do not know the specific circumstances. The utilization rate of fertilizer to strengthen universal education knowledge to protect the ecological environment of farmers. In the process of developing the rural economy, there is a lack of education for the protection of the ecological environment. Many farmers blindly pursue the maximization of production and labor intensity, in the process of planting a large number of fertilizers, pesticides, agricultural environmental pollution. This one-sided pursuit of economic efficiency, but to make their own money, ignoring environmental benefits, regardless of the behavior of environmental pollution so that unsustainable agricultural sustainable development. Although the excessive use of chemical fertilizer especially the excessive application of nitrogen fertilizer, can improve grain yield to a certain extent, but long-term use will cause excessive application of special land compaction, low productivity, in order to maintain production, the amount of chemical fertilizer also increased, which will cause the vicious spiral. Therefore, must attach importance to the transformation from the traditional only micro fertilizer sales and crop yield to the new concept of the concept of social and environmental costs and direct economic benefits simultaneously, so that the farmers realize the importance of environmental effects, consider the problem from the long-term interests, not one-sided pursuit of immediate interests.

Acknowledgements

National Natural Science Foundation of China: Study on the application of chemical fertilizer and agricultural non-point source pollution control in large scale growing plants: A Study on the mechanism and policy simulation(Item number: 71573036); and NSFC:70973016/G312.

REFERENCES