Ciência e Natura, v. 37 Part 2 2015, p. 108–116
Revista do Centro de Ciências Naturais e Exatas – UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

The Application of Data Mining Techniques in Agricultural Science

Hooman Fetanat[1]*Leila Mortazavifar*[2] *Narsis Zarshenas*[3]

[1]*Iran,Shiraz, Faculty:Zand Department of ICT,Computer Faculty,Zand University*
[2]*Iran,Shiraz, Faculty:Zand Department of ICT,Computer Faculty,Zand University*
[3]*Iran, Shiraz, Faculty: Agricultur Faculty,Shiraz University*

Abstract

Information Technology has a positive impact on other disciplines. Using today's technology, precision agriculture and Information Technology are mixed together. Use of Information Technology in agriculture will lead to improvements in productivity. For this purpose, the raw data is transformed into useful information through data mining. This research determined whether data mining techniques can also be used to improve pattern recognition and analysis of large growth factors of ornamental plants experimental datasets. Furthermore, the research aimed to establish data mining techniques can be used to assist in the classification and regression methods by determining whether meaningful patterns exist various growth factors of ornamental plants characterized at various research sites across Kish Island. Different data mining techniques were used analyze a large data base of ornamental plants properties attributes. The data base has been collected from different plants of Kish Island in various areas in order to determine, classify and predict effective growth factors on blooming. In this research, analyzed data with regression technique showed the effect of chlorophyll content on the number of flowers. The analysis of these agricultural data base with different data mining methods may have some advantages in agriculture

I

Keywords: *information technology, data mining, cluster, agriculture, regression*

Recebido: dia/mês/ano Aceito: dia/mês/ano

1 Introduction

nformation Technology (IT) constitutes an important part of our life today. Data management has become one of the most important parts in the industry. Using techniques from data management to increase productivity, especially in agriculture is very important. Large amounts of data to increase efficiency, but using the correct techniques for managing and organizing this information is required. Hence data mining may help to convert this raw data into useful information for improving agricultural uses, because Data Mining represents a set of specific methods and algorithms aimed solely at extracting patterns from raw data [1]. Data mining is the process of discovering previously unknown and potentially interesting patterns in large datasets [1]. The 'mined' information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification [2]. Data mining, also termed as knowledge discovery, is the process of analyzing data from different perspective and summarizing it into valuable or non-trivial information. This information can be used for variety of purposes-research, cost cuts and revenue, future forecasting or prediction and so on. Data mining applications serve as a fraction of a number of analytical tools for analyzing data. The data can be analyzed from many different dimensions, categorized & summarized the relationships identified. Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and analysis methods [2]. The more common model functions in current data mining practice include [3]:

1. Classification: classifies a data item into one of several predefined categorical classes [4].

2. Regression: maps a data item to a real valued prediction variable [5,6,7].

3. Clustering: maps a data item into one of several clusters, where clusters are natural groupings of data items based on similarity metrics or probability density models [8,14,15,19].

4. Rule generation: extracts classification rules from the data [9].

5. Discovering association rules: describes association relationship among different attributes [10].

6. Summarization: provides a compact description for a subset of data [11].

7. Dependency modeling: describes significant dependencies among variables [12].

8. Sequence analysis: models sequential patterns, like time-series analysis. The goal is to model the states of the process generating the sequence or to extract and report deviation and trends over time [13].

This paper discusses data mining technique such as regression and clustering which is a process model for analyzing data and describes the support that SPSS provides for this model. SPSS-based analysis and application construction process is illustrated through a case study in the agricultural domain-ornamental plants. Cluster analysis or clustering was used is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. In this survey clustering technique divides growth factors into several independent categories. Also, regression technique which was used includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. This paper showed us the impressive factors on flowering of ornamental plants which lead to improve and maintain Kish Island landscape. A comparison of data mining techniques and statistical methods could produce a model for further understanding the data. The advantage of a better understanding of ornamental plants could improve elegance in landscape. The overall aim of the research was to introduce valid data mining techniques in agriculture and how these techniques have most efficient in determining patterns when compared to standard statistical analysis techniques.

2 Materials and Methods

2.1 Clustering Technique

Clustering is one method of data mining, were used in this research. When many data is available, it must develop algorithms, which can extract meaningful information from these data. Finding useful information in huge

Recebido: dia/mês/ano Aceito: dia/mês/ano

amounts of data has become known as the field of data mining. Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes called clusters [14,15].

Many studies have used clustering methods. Clustering techniques are used widely in various fields of science. Clustering is often followed by a stage in which a decision tree or rule set is inferred that allocates each instance to the cluster in which it belongs. Then, the clustering operation is just one step on the way to a structural description [19].

In the agricultural science, data mining clustering techniques are used in Optimizing pesticide use by data mining [16] ,explaining pesticide abuse by data mining [17] ,detecting weeds in precision agriculture [18] and The impact of data mining in the flowering [20].

Clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups. Clustering naturally requires different techniques to the classification and association learning methods we have considered so far. Many clustering methods are available and each of them may give a different grouping of a dataset. In [19,21] discussed about a few of them.

2.2 Regression

Regression Analysis is a statistical tool that uses the relation between two or more quantitative variables so that one variable (dependent variable) can be predicted from the other(s) (independent variables). But no matter how strong the statistical relations are between the variables, no cause-and-effect patterns are necessarily implied by the regression model. Many regression analyzes are available including simple linear, multiple linear, curvilinear and multiple curvilinear regression models [24].

Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation [22].

This paper discusses a process model for analyzing data and describes the support that SPSS provides for this model. SPSS-based analysis and application construction process is illustrated through a case study in the agricultural domain-ornamental plants

3 Methodology

3.1 Location of study

The study was conducted in Kish Island which is located in the Persian Gulf on the mainland Iran. The island in the geographic coordinates 53 degrees 53 minutes to 54 degrees 3 minutes east of the Greenwich meridian and 26 degrees 34 minutes north latitude, respectively.

Island weather is hot and humid, and the average annual temperature is 27 degrees Celsius. Four different regions of Kish Island were selected to do the experiment. They were Sanae, Sadaf, Pavion and Saffein

3.2 Data collection

Data were collected from different ornamental plants which are in selected regions in 2013-2014. The ornamental plants included *Nerium oleander*, *Tecoma stans*, *Thevetia nerrifolia* and *Delonix regia*. The data were related to growth factors such as number of flowers, florets, leaves, duration of flowering, leaf area, chlorophyll and proline contents, fresh and dry weight of leaves.

3.3 Data mining process

The data mining process was conducted with regression and clustering techniques on dataset.

3.4 Application of clustering

The clusters found by different algorithms vary significantly in their properties, and understanding these cluster models is a key to understanding the differences between the various algorithms. division of data into groups of similar objects is clustering And each group, called cluster, Objects within each cluster are similar but they are different with other Clusters.

Recebido: dia/mês/ano Aceito: dia/mês/ano

Clustering naturally requires different techniques to the classification and association learning methods.

The classic clustering technique is called *k-means.* First, you specify in advance how many clusters are being sought: this is the parameter *k*. Then *k* points are chosen at random as cluster centers. All instances are assigned to their closest cluster center according to the ordinary Euclidean distance metric. Next the centroid, or mean, of the instances in each cluster is calculated this is the "means" part.

These centroids are taken to be new center values for their respective clusters. Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain the same forever. This clustering method is simple and effective [19].

We used *K*-means clustering. This research considered growth factors included diameter of flowers, length of flowers, diameter and length of stems, chlorophyll and proline content, leaf area, dry and fresh weight of leaves in order to cluster them.

3.5 Application of regression

Regression analysis states the relation between some independent variables and one dependent variable in order to predict the variation of dependent variables through independent variables.

It also determines the portion of each independent variable in variation of depended variable. This research showed the effect of chlorophyll content on the number of flowers.

4 Result

The measured growth factors were classified into four clusters. In the first cluster, diameter of flowers, length of flowers and chlorophyll content were related to each other. However, they were independent to the other groups. Diameter and length of stem were in the second cluster. They were not related to the other groups. There were leaf areas, fresh and dry weight of leaves in cluster 3.

They were related to each other, but they were independent compared to other clusters. As shown in Table 1 in cluster 4, there was only proline content which was not related to other clusters.

Table 1: Clustering technique

	Diamete of flower	Length of flower	Chlorophyl content	Diameter of stem	Length of stem	Leaf area	Fresh weight of leaf	Dry weight of leaf	Proline content
Cluster1	✓	✓	✓						
Cluste2				✓	✓				
Cluster3						✓	✓	✓	
Cluster4									✓

There is Classification of Flowering Plants. A factor measured with the letters H-Z is shown.

H: Diameter of Flower, I: length of Flower, K: diameter of stem, L: length of stem, P: leaf area ,R: fresh weight of leaf, S: dry weight of leaf, V: Chlorophyll , Z: Proline

Area1:sanae ,Season 1:cold, species 1:oleander, X1 : number of flowers ,X7: number of flowers in season 1,X13: number of flowers in season 1,area 1, X19: number of flowers in season 1,area1 ,species1

4.1 Classification of Flowering Plants

Cluster Analysis of Variables: H, I, K, L, P, R, S, V, Z. The resulat is shown in figure 1.

Cluster 1
H I V
Cluster 2
K L
Cluster 3
P R S
Cluster 4
Z

Recebido: dia/mês/ano Aceito: dia/mês/ano

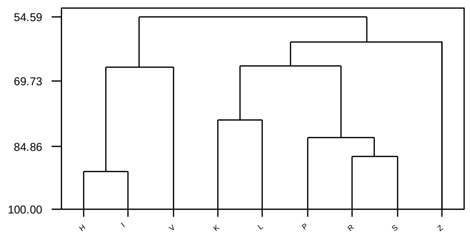

Figure1: Cluster Analysis of Variables

4.1.2 Classification of Flowering Plants in season 1

Analysis of Variables: H1; I1; K1; L1; P1; R1; S1; V1; Z1 .The resulat is shown in figure 2.

Cluster 1

H1 I1 V1

Cluster 2

K1 L1 P1 R1 S1

Cluster 3

Z1

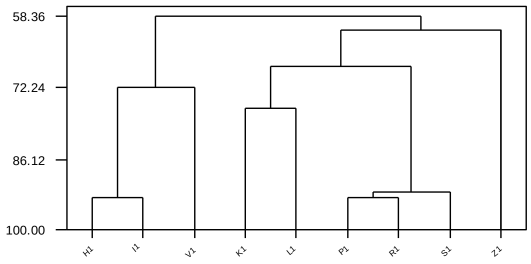

Figure2: Classification of Flowering Plants in season 1

4.1.3 Classification of Flowering Plants in season 1, area 1

Cluster Analysis of Variables: H11; I11; K11; L11; P11; R11; S11; V11; Z11. The resulat is shown in figure 3.

Cluster 1:

H11 I11 K11 L11 P11 R11 S11 V11

Cluster 2:

Z11

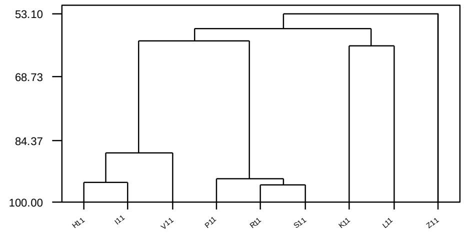

Figure3: Classification of Flowering Plants in season 1, area 1

The regression equation for effect of chlorophyll on the number of flowering plants flowers is:

X1=0.147118 + 0.173162V (eq.1)

X1 is number of flowers and V is chlorophyll content. There is prediction interval (months) for equation 1,Which is predicted for duration of flowering by rising chlorophyll content ,Also it has confidence interval(months). The result is shown In Figure 4.

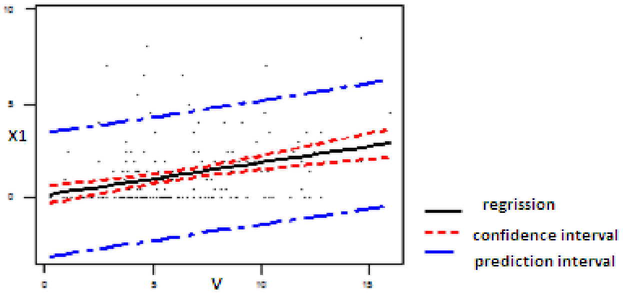

Figure4: Effects of chlorophyll content on number of flower

There is confidence interval (months) which is consist of a range of values (interval) that act as good estimates of the unknown population parameter (duration of flowering) .Also, prediction interval is an estimate of an interval in which future observations will fall, with a certain probability, given what has already been observed. As shown in Table 2 that rising chlorophyll content lead to increase number of flowering.

Table 2. Effects of chlorophyll content on number flower

observed chlorophyll	Number of flower	errors	confidence Interval	prediction Interval
8	1.53	0.13	2.26-1.80	0-4.79
9.5	1.79	0.17	1.45-2.13	0-5.05
10	1.87	0.18	1.51-2.24	0-5.14
18	3.26	0.45	2.36-4.16	0-6.63
20	3.61	0.52	2.57-4.64	0.2-7.01

Recebido: dia/mês/ano Aceito: dia/mês/ano

4.2 Effects of chlorophyll content on the number of flowering plants flowers in season 1

Chlorophyll content had an effect on the number of flowering plants flowering in the cold season. For example, as shown in figure 5 and table 3, with 20 milligrams per gram of fresh weight chlorophyll, obtain 3.79 number of flowers in winter with 0.747 errors, confidence interval 2.30-5.27 number of flowers and prediction Interval 0.80-7.50 number of flowers in the cold season, respectively.

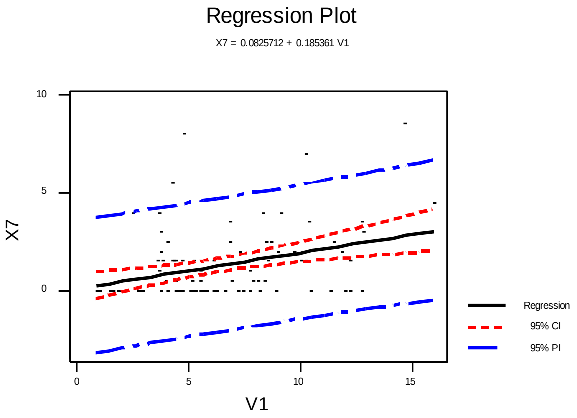

Figure5: Effects of chlorophyll content on the number of flowering plants flowers in season 1

Table3: Effects of chlorophyll content on the number of flowering plants flowers in season 1

observed chlorophyll	Number of flower	Errors	confidence Interval	prediction Interval
8	1.56	0.19	1.18-1.95	0-4.98
9.5	1.84	0.24	1.36-2.32	0-5.27
10	1.93	0.26	1.42-2.45	0-5.37
18	3.41	0.64	2.14-4.69	0.2-7.05
20	3.79	0.74	2.30-5.27	0.08-7.50

.

4.2.1 Effects of chlorophyll content on the number of flowering plants flowers in season 1 area1

The amount of chlorophyll on the number of flowers and flowering plants in winter in sanae area is affected For example, as shown in figure 6 and table 4, with 18 milligrams per gram of fresh weight chlorophyll, obtain 6.22 number of flowers in winter with 1.33 errors, confidence interval 3.44-8.96 number of flowers and prediction Interval 0.84-11.59 number of flowers in the cold season, respectively.

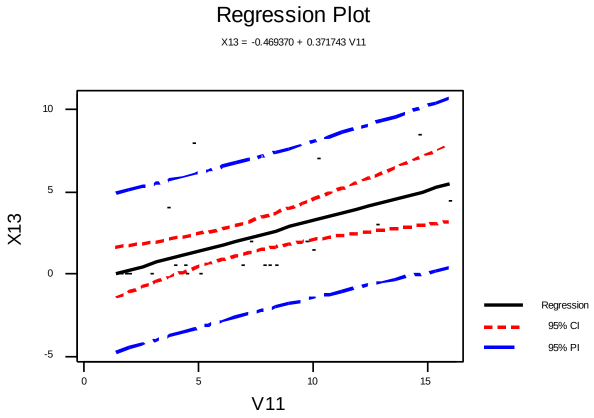

Figure 6: season 1 area1

Recebido: dia/mês/ano Aceito: dia/mês/ano

Table4: season 1 area1

observed chlorophyll	Number of flower	errors	confidence Interval	prediction Interval
8	2.50	0.47	1.51-3.49	0-7.21
9.5	3.06	0.55	1.91-4.21	0-7.80
10	3.24	0.58	2.02-4.46	0-8.00
18	6.22	1.33	3.44-8.99	2.84-11.59
20	6.96	1.54	3.75-10.17	1.35-12.57

.

4.2.2 Effects of chlorophyll content on the number of flowering plants flowers in season 1 area1 Species 1

The amount of chlorophyll affected in the number of flowering plants flowers in winter and oleander plants in sanae area. For example, as shown in figure 7 and table 5,with 8 milligrams per gram of fresh weight chlorophyll, obtain 3.62 number of flowers in winter with 1.55 errors, confidence interval 0-7.93 number of flowers and prediction Interval 0-11.68 number of flowers in the cold season.

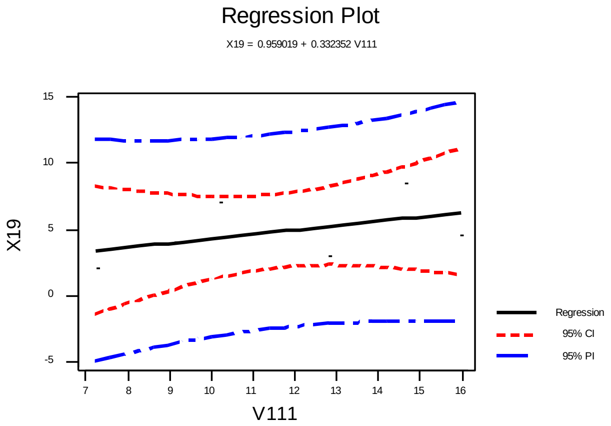

Figure 7: season 1 area1 Species 1

Table5: season 1 area1 Species 1

observed chlorophyll	Number of flower	errors	confidence Interval	prediction Interval
8	3.62	1.55	0-7.93	0-11.68
9.5	4.12	1.22	0.72-7.51	0-11.72
10	14.28	1.14	1.13-7.44	0-11.79
18	6.94	2.29	0.58-13.30	0-16.26
20	7.61	2.89	0-15.63	0-18.13

5. Discussion and conclusion

The collection of information and data has increased with the advent of new computing technology, but establishing patterns within this data has become more difficult and requires new approaches and tools if it is to be undertaken. The advent of this problem has provided an opportunity from which data analysis has started to take over from current methods. Furthermore, this technology has reduced the time taken to undertake data analysis and increased automation of the process [25]. Since, a great deal of data was collected in this research some effective outcome gained. As shown in Table 1 growth factors were in four clusters. These clusters varied independently, for instance, proline variation which was in cluster 4 didn't affect any other growth factors. Also Table 2 described Figure 4, increasing chlorophyll content causes number of flowering to extend. This result can be related to the amount of carbohydrate which made by chlorophyll. [23]

Recebido: dia/mês/ano Aceito: dia/mês/ano

also used regression to estimate leaf number currently held on the plant and degree of leaf sheding occurred was carried out in two Cassava (Manihot esculenta) morphotypes (Philippine and Nagra) at Mymensingh (24°75′N 90°50′E). [23] Four linear regression Models were developed for estimating leaf number from length of main stem and primary branch. For the future study we consider the effects of chlorophyll content on the number of flowering plants flowers in warm season.

References

Fayadd, U., Piatesky–Shapiro, G., Smyth, P. (1996).Data Mining to Knowledge Discovery in Databases, pp.50-67.

Cunningham, S. J., Holmes, G. (1999). Developing innovative applications in agriculture using data mining. In the Proceedings of the Southeast Asia Regional Computer Confederation Conference, pp.25-29.

Ashok Kumar, D., Kannathasan, N. A.(2011). Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining, 8(3): 20-32

Mitra. S., Mitra. P., Pal, S.K. (2001). Evolutionary modular Design of rough knowledge-based network using fuzzy attributes Neurocomputing. 3(6): 45-66

Xu, K., Wang, Z., Leung, K. S(1998). Using a new type of non Linear integral for multi-regression: an application of evolutionary Algorithms in data mining. In Proceedings of IEEE International Conference on Systems, Man, and Cybernetics. San Diego, CA, USA, pp. 2326-2331.

Brenning, A. (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation.Natural Hazards and Earth System, 5(6),853–862

Brenning, A., Itzerott, S. (2006). Comparing classifiers for crop identification based on multitemporal landsat tm/etm data. In: Proceedings of the 2nd workshop of the EARSeL Special Interest Group Remote Sensing of Land Use and Land Cover. Bonn, Germany, pp. 64–71

Russell, S. Lodwick, W. (1999). Fuzzy clustering in data mining for telco database marketing campaigns. In: Proceedings of NAFIPS 99. New York, USA, pp. 720-726

Zhang, Y. Q., Fraser, M. D. (2000). Gagliano R A, Kandel A. Granular neural networks for numerical-linguistic data fusion and knowledge discovery. IEEE Transactions on Neural Networks,11: 658-667

Au, W. H., Chan, K. C. C. (1998). An effective algorithm for discovering fuzzy rules in relational databases. In: Proceedings of IEEE International Conference on Fuzzy Systems, pp. 1314-1319

Kacprzyk, J., Zadrozny, S. (1998). Data mining via linguistic summaries of data: an interactive approach, pp. 668-671

Bosc, P., Pivert, O., Ughetto, L. (1999) Database mining for the discovery of extended functional dependencies, pp.580-584.

Lee, R. S., Liu, J. N. K. (2000). Tropical cyclone identification and tracking system using integrated neural oscillatory leastic graph matching and hybrid RBF network track mining techniques. IEEE Transactions on Neural Networks,11: 680-689

Lee, R. C. T. (1981). Cluster analysis and its applications. In Advances in Information Systems Science, pp. 580-584

Everitt, B. S. (1993). Cluster Analysis. third edition, pp.28-36

Abdullah, A., Brobst, S., Pervaiz, I., Umar, M., Nisar, A.(2004). Learning Dynamics of Pesticide Abuse through Data Mining. Australasian Workshop on Data Mining and Web Intelligenc, pp.56-71

Abdullah, A., Hussain, A. (2006). Data Mining a New Pilot Agriculture Extension Data Warehouse, 38 (3): 229–249

Tellaeche, A., BurgosArtizzu, X. P., Pajares, G., Ribeiro, A. (2008). A vision-based hybrid classifier for weeds detection in precision agriculture through the Bayesian and fuzzy K-means paradigms.Adv Soft Comp, 44:72-79

Recebido: dia/mês/ano Aceito: dia/mês/ano

Witten, I. H., Frank, E. (2005). Data mining practical machine learning tools and technique, pp.23-31

Oteros, J., García-Mozo, H., Hervás-Martínez, C., Galán, C. (2013). Year clustering analysis for modelling olive flowering phenology. Int J Biometeorol, 57 (4): 545-547

Arockiam, L., Baskar, S., Jeyasimman, L. (2011). Overview of Clustering Techniques in Agriculture Data Mining. Agric, 6(5): 222-225

Sahu, H., Shrma, S., Gondhalakar, S. (2001). A brief overview on data mining survey, (1)3: 149-151

Fakir, M. S. A., Mostafa, M. G., Karim, M. R., Prodhan, A. K. M. A. (2011). Prediction of leaf number by linear regression models in cassava, 9(1): 49-54

Jackson, J. (2002). Data mining: a conceptual overview, communications of the association for information systems, 8: 267-296

Armstrong, L., Diepeveen, D. (2007). The application of data mining techniques to characterize agricultural soil profiles: The sixth Australian data mining conference. p p.81-96

Recebido: dia/mês/ano Aceito: dia/mês/ano