Ciência e Natura, v. 37 Part 2 jun. 2015, p. 102–107

ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

ciência e natura

The analysis of agricultural data with regression data mining technique

Hooman Fetanat[1], Leila Mortazavifarr[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

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. 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. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. In this research, analyzed data with regression technique showed the effect of chlorophyll content on the number of flowers.

Keywords*: agriculture, data mining, regression*

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

1 Introduction

Data mining is the process of discovering previously unknown and potentially interesting patterns in large datasets (Piatetsky-Shapiro and Frawley 1991). 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 (Cunningham and Holmes 1999). 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. Technically, data mining simply is the process of finding correlation or patterns among dozen of field in large RDBMS (Ansar 2010). Data mining is a high-level application technique used to present and analyze data for decision-makers. There is an enormous wealth of information embedded in huge databases belonging to enterprises and this has spurred tremendous interest in areas of knowledge discovery and data mining. Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and analysis methods (Cunningham and Holmes 1999).

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. The analysis of these agricultural data base with different data mining methods may have some advantages in agriculture. This survey 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 Review of literature

A number of studies have been carried out on the application of data mining techniques for agricultural data sets. For example, Ibrahim (1999) on a sample data set applied six classification algorithms to 59 data sets and then six clustering algorithms were subsequently applied to the data generated. The results were studied and the patterns and properties of the clusters were formed to provide a basis for the research. The research provided a comparison of performance for the 6 classification algorithms set to the default parameter settings. It can be concluded that a large investigation is required which uses more data sets and data set characteristics.

Georg Rub (2008) showed that support vector regression can serve as a better reference model for yield prediction. It is computationally less demanding, at least as understandable as the hitherto multi-layer perception and, most importantly, produces better yield predictions. The results also show that model parameters which have been established on one data set can

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

be carried over to different (but similar with respect to the attributes) data sets. Data mining is widely applied to agricultural problems. For instance, the prediction of wine fermentation can be performed by using k-means approach (Pardalos et al. 2009) and a technique for classification based on the concept of biclustering (Mucherino and Urtubia 2010). Note that these works are different from the ones where a classification of different kinds of wine is performed. The data mining technique can also be used for estimating soil water parameters. Apples and other fruits are widely analyzed in agriculture before marketing. Apples running in conveyors can be checked by humans and the bad apples (the ones presenting defects) (Pardalos et al., 2009).

3.1 Regression technique

Regression is a data mining (machine learning) technique used to fit an equation to a dataset. 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 (chaple 2011).

Regression technique can be used in agriculture. Osman and et al. (1997) reported the results of a regression analysis of the relationship between energy use and agricultural productivity. The study was based on the analysis of the yearbook data for the period 1971-2003. The study investigated the impacts of energy use on productivity of Turkey's agriculture.

Mai and et al. (2006) showed regression technique can predict real demand water in Indian agriculture. A major task in agriculture production is water and fertilizer management. Deficit application, however, may limit the growth of crop.

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.

4 Methodologies

4.1 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.

4.2 Data collection

Data were collected from different ornamental plants which are in selected regions. 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.

4. 3 Data mining process

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

4. 3.1 Application of regression

Regression analysis states the relation between some independent variables and one dependant variable in order to predict the variation of dependant variables through independent variables. It also determines the portion of each independent variable in variation of depended variable (Mehralizade, 2007). This research showed the effect of chlorophyll content on the number of flowers.

5 Result

The regression equation for effect of chlorophyll content on the number of 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

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

equation 1,Which is predicted for duration of flowering by rising chlorophyll content ,Also it has confidence interval(months).

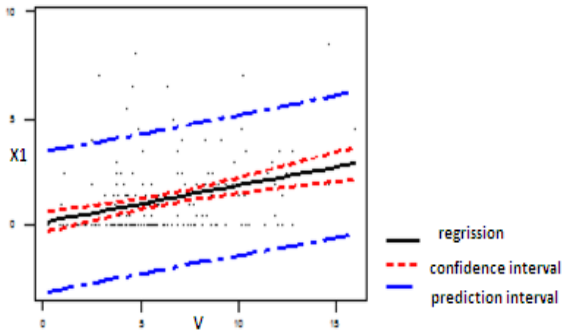

Fig 1: (effects of chlorophyll content on number of flower)

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 table 1 showed that rising chlorophyll content lead to increase number of flowering .

Table1: (effects of chlorophyll content on number flower)

observed chlorophyll	Number of flower	errors	confidence Interval	prediction Interval
10	1.87	0.18	1.51-2.24	0-5.14
9.5	1.79	0.17	1.45-2.13	0-5.05
8	1.53	0.13	2.26-1.80	0-4.79
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

Chlorophyll content had an effect on the number of flowering plants flowering in the cold season. For example, as shown in table 2, with 20 milligrams per gram of fresh weight chlorophyll, obtain 3.79 numbers 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.

Table2 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

The amount of chlorophyll on the number of flowers and flowering plants in winter in sanae area is affected For example, For example, as shown in table 3, 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.

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

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

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 table 4,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.

Table4 Effects of chlorophyll content on the number of flowering plants flowers in season 1 area1 Species1

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

6 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 (Armstrong et. al. 2004). Since, a great deal of data was collected in this research some effective outcome gained. As table 1 illustrated 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. This result can be related to the amount of carbohydrate which made by chlorophyll. Fakir et al. (2011) 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). Four linear regression Models were developed for estimating leaf number from length of main stem and primary branch.

References

Ansari, A., Ansari, S. (2010). the concept of data mining, its applications & Issues,pp.67-78.

Armstrong, L., Diepeveen, D. (2004). The application of data mining techniques to characterize agricultural soil profiles. Confpapers.

Arockiam, L., Baskar, S. (2011). Overview of clustering techniques in agriculture data mining. Agri. J 6: 222-225.

Ashkvand, rad. E. (2007). Mining the Ovarian Cancer Ascites Proteome for Potential Ovarian Cancer Biomarkers. Data mining journal 8(4): 661–669

Chapple, M. (2011). Regression, Accessed 2 Feb 2012,pp.98-106.

Cunningham, S., Holms, G. (1999). developing innovative application in agriculture using data mining, pp.6-15

Fakir, M., Mostafa, M. (2011). Prediction of leaf number by linear regression models in cassava. JBAU 2011 9(1): 49-54

Gregory, M., Piatetsky-Shapiro. (2007). Data Mining and Knowledge Discovery journal, pp.56-71.

Ibrahim, R. (1999). Data Mining of Machine Learning Performance Data. Unpublished Master of Applied Science (Information Technology), Publisher; RMIT University Press

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

Klise, K. McKenna. (2006). Water quality change detection Multivariate algorithms. proceeding of the SPIE (International Society for Optical Engineering), Defense and Security Symposium, April 18-20, Orlando. Florida, pp.211-222.

Leemans, v. Destain, M. (2004). A real-time grading method of apples based on features extracted from defects. J. Food Eng 61: 83-89

Mai, J. (2006), Data mining application journal of amercing no 2, pp:17-22.

Mucherino, A., Urtubia, A. (2010). Consistent Biclustering and Applications to Agriculture, IbaI Conference Proceedings, and Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop Data Mining in Agriculture (DMA10), pp. 105–113.

Osman K (1997) Arterial hypertension in Saudi Arabia," Annals of Saudi Medicine, Vol. 17, No. 2, pp. 170-174.

Pardalos, P., Papajorgji, P. (2009). A Data mining in agriculture. University of Florida, 303 Weil Hall, Gainesville, FL 32611-6595, USA.

Rub G (2008) Data Mining of Agricultural Yield Data: A Comparison of Regression Models, pp.17-28

Tellaeche, A., BurgosArtizzu, X., Pajares, G. (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