International Journal Of Computer Science And Applications Vol. 1, No. 1, June 2008
ISSN 0974-1003

Data Mining Tool for Agriculture Applications

Jyothi Patil
Information Science Dept.
PDA College of Engg.
Gulbarga
Jyothip_pda2003@ yahoo.com

Dr. Goverdhan
Computer Science Dept.
JNT University
Hyderabad
goverdhan_cse@ yahoo.co.in

M.Narasimha Murthy
Computer Science Dept.
Indian Institute of Science
Bangalore
mnm@csa.iisc.ernet.in

ABSTRACT

This paper focuses on clustering in data mining. Clustering is a division of data into groups of similar objects. k_means clustering algorithm is used as a tool to cluster the weather data to develop an accurate and efficient weather prediction software for agriculture applications. The weather data contains date periodically observed with parameters temperature (min, max), rainfall, humidity (min, max). The Automatic weather prediction software is designed with user interaction to provide a period of observations such as weekly, fortnight, monthly, half yearly, yearly and so on, to make decision for required period for a particular region from which data is collected. k_means clustering algorithm clusters weather data on selected parameter.

Categories and Subject Descriptors

H 3.3 **[Information Storage and Retrieval]** : Information Search and Retrieval – clustering, search process, query formulation.

General Terms

Algorithm, Experimentation.

Keywords

k-means clustering, classification, prediction.

1 INTRODUCTION

We are living in a world full of data, every day, people encounter a large amount of information and store or represent it as data, for furher analysis and management. One of the vital means in dealing with these data is to classify or group them into set of categories or clusters. Actually as one of the most primitive activities of human beings, classification plays an important and indispensable role in the long history of human development.

In order to learn a new object or understand a new phenomenon, people always try to seek the features that can describe it, and further compare it with other known objects or phenomenon, based on the similarity or dissimilarity, generalized as proximity, according to some certain standards or rules. Basically classification systems[3][12] are either supervised or unsupervised. In supervised learning the class label of each training object is provided. In unsupervised learning (or clustering) the class label of each training object is not known, and the number or set of classes to be learned is not known in advance. Prediction is also a type of classification. Prediction predicts future state based on past and current data.

Clustering is a descriptive task seeks to identify homogenous groups of objects based on the values of their attributes (dimensions) [8][9]. Clustering techniques have been studied extensively in statistics[1], pattern recognition[3][7], and machine learning [2][10].

Clustering techniques can be broadly classified into two categories [8][9]. Partional and hierarchical. Given a set of objects and a clustering criterion [11], partitional clustering obtains a partition of the objects into clusters such that the objects in a cluster are more similar to each other than to objects in different clusters. A hierarchical clustering is nested sequence of partitions. An agglomerative, hierarchical clustering starts by placing each objects in its own clusters and then merges these atomic clusters into larger and larger clusters until all objects are in a single cluster. Divisive, hierarchical clustering reverses the process by starting with all objects in cluster and subdividing into smaller pieces [8].

Desiderate from the data mining perspective

Emerging data mining applications place the following special requirements on clustering techniques, motivating the need for developing new algorithms.

Effective treatment of high dimensionality: An object (data record) typically has dozens of attribute and the domain for each attribute can be large. Many dimensions or combination of dimensions can have noise or values that are uniformly distributed. Therefore distance functions that use all the dimensions of the data may be ineffective.

Interpretability of results : Data mining applications typically require cluster descriptions that can be easily assimilated by an end user as insight and explanations are of critical importance [6]. It is particularly important to have simple representations because most visualization techniques do not work well in high dimensional spaces.

Scalability and usability: The clustering technique should be fast and scale with the number of dimensions and the size of input. It should be insensitive to the order in which the data records are presented.

Current clustering techniques do not address all these points adequately although considerable work has been done in addressing each point separately.

Published by Research Publications, Chikhli, India

1.1 k-means clustering

k-means clustering [4] is a method commonly used to automatically partition a data set into k groups, it proceeds by selecting k initial cluster centers and then iteratively refining them as follows

Each instance d_i assigned to its closest cluster center

Each cluster center C_j, is updated to be the mean of its constituent instances.

The algorithm converges when there is no further change in assignment of instances to clusters. The time completely of k-means is O(tkn) where t is the number of iterations.

1.2 Classifiers

1.2.1 NEAREST NEIGHBOR CLASSIFIER (NNC)

In this classifier, items are iteratively merged into the existing clusters that are closest. A threshold t is used to determine if items will be added to existing clusters or if a new cluster is created.

The complexity of the NNC [4] actually depends on the number of items for each loop, each item must be compared to each item already in a cluster. This will be n in the worst case. Thus the time complexity of NNC is $O(n^2)$ If there are n patterns $X = \{X_1, X_2 \ldots\ldots X_n\}$.Each of dimension d and each pattern is associated with a class c, and if we have a test pattern P then if $d(P, X_k) = \min \{d(P, X_i)\}$ where i=1…n. Assign pattern P to the class associated with X_k.

The distance between X and P is taken as the Euclidean distance

$$d(X,P) = \sqrt{(X[1] - P[1])2 + (X[2] - P[2])2}$$

1.2.2 K-NEAREST NEIGHBOR CLASSIFIER (KNNC)

In this classifier [4] instead of finding just one nearest neighbor as in the NNC, k neighbors are found. The majority class of these k nearest neighbor is the class assigned to the new pattern. The value chosen for k is crucial and with the right value of k, the classification accuracy will be better than the nearest neighbor classifier.

2 SIGNIFICANCE OF WEATHER DATA CLUSTERING

India is a predominantly agriculture based country with more than two thirds of its population living in rural areas where agriculture is the main occupation. In the field of agriculture weather plays an important role from sowing period, following the intermediate stages like germination, vegetation, reproduction and finally harvesting. The role of weather parameters is with respect to various departments like entomology, pathology, agronomy, soil science, genetics and plant breeding horticulture.

In the branch of entomology the weather parameters are used to examine the insect population that is insect population verses weather population. They work out the emergence of pest population in a particular crop based on the respective weather parameters like minimum temperature, maximum temperature, rainfall, minimum relative humidity and maximum relative humidity.

In the branch of pathology, disease verses different weather parameters are worked out. That is different weather parameters are co-related with the incidence of diseases there by giving the conclusion (prediction model), which states how weather parameters favours the disease development. Ex, high rainfall high relative humidity with moderate temperature favours discease development during rainy season.

In the branch of agronomy, the usage of weather parameters is between different kinds of crops like food crops, oil seeds, pulses, commercial crops, etc where growth can be annual.

In soil science weather parameters are used to examine the nature of soil that is whether the soil is acidic, sodic or alkaline. Ex, during heavy rainfalls soil may be acidic or in arid region like Raichur, soil is mostly sodic.

In horticulture department the weather parameters are used for growing of vegetables, fruits, plantations, flowers and ornamentals.

In genetics and plant breeding weather parameters are used to test pollen viability, increase in yield, increase in quality, pest resistance etc.

3 APPLICATION OF CLUSTERING TECHNIQUE FOR WEATHER DATA

3.1 Weather data

The weather data used in this paper is obtained from university of Agricultural sciences , Raichur. The data contains six dimensions namely date(MM/DD/YYYY), maximum temperature ,minimum temperature, rainfall, maximum relative humidity, minimum relative humidity. The university of agricultural sciences records data on days basis.

Table 1. Input (weather data)

Date	MXT	MNT	RF	RH1	RH2
1/1/1996	30.0	18.2	0.0	81	41
1/2/1996	30.4	18.2	0.0	77	38
1/3/1996	30.7	17.0	0.0	72	36
1/4/1996	30.1	18.0	0.0	90	36
1/5/1996	30.7	17.5	0.0	77	32
1/6/1996	31.5	17.5	0.0	79	39
1/7/1996	30.8	17.4	0.0	78	34

1/8/1996	31.9	17.0	0.0	83	38
1/9/1996	30.8	15.8	0.0	81	36
1/10/1996	30.9	18.4	0.0	84	43

3.2 Weather Prediction

Input (Weather Data) --→ Modified Input Data ------→ Feature Selection -→ Clustering using k-means on selected feature ------→ out put

The input data (weather data) Contains six dimensions namely date, max temperature, min temperature, rainfall, max relative humidity, min relative humidity.

The modified input data contains month(mm), day(dd) year (yyyy) average temperature, rainfall, average humidity as dimensions.

Table 2. Modified input data

MM	DD	YYYY	Avg. Temp.	Rainfall	Avg. Humidity
1	1	1996	24.1	0	61
1	2	1996	24.2	0	57.5
1	3	1996	23.85	0	54
1	4	1996	24.05	0	63
1	5	1996	24.1	0	54.5
1	6	1996	24.5	0	59
1	7	1996	24.1	0	56
1	8	1996	24.45	0	60.5
1	9	1996	23.3	0	58.5
1	10	1996	24.65	0	63.5

Any one feature i.e., avg temperature, rainfall or avg humidity is selected from input data and clustered using k-means. The selected feature data that is clustered is between starting and ending period specified by the user. The period can be weak, fortnight, month, year .

Output displays results of clustering from which we can predict temperature, humidity and rainfall in that region for a specified period from which data is collected.

4 EXPERIMENTAL RESULTS

```
Enter the starting period
Enter the Month: 01
Enter the day: 01
Enter the year: 1996
Enter the ending period
Enter the Month: 12
Enter the day: 31
Enter the year: 1996
Enter K value = 4
Enter clustering data choice,(4)Average Temperature(DegC) , (5)Rainfaill(mm) ,(6)Average Humidity(F) : 4
starting period Month Day year
          1          1       1996

ending period Month Day year
         12         31       1996

Total number of days between starting and ending period
  366

 iter   phase      num           sum
    1       1      366        14.843
    2       1       21       14.1646
    3       1       14       13.9664
    4       1       14       13.8635
    5       1        5         13.84
    6       1        3       13.8204
    7       2        1       13.8193
7 iterations, total sum of distances = 13.8193
```

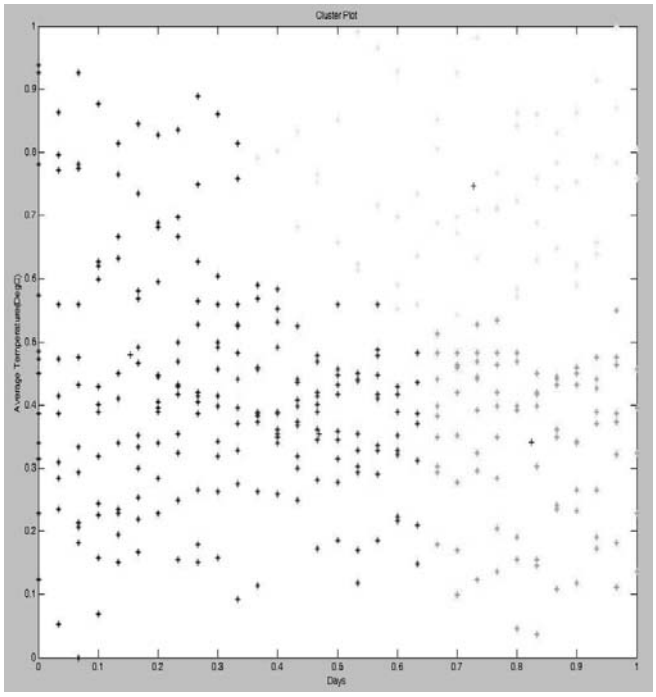

Fig. 1. One year observation for temperature

```
Enter the starting period
Enter the Month: 01
Enter the day: 01
Enter the year: 2003
Enter the ending period
Enter the Month: 12
Enter the day: 31
Enter the year: 2003
Enter K value = 4
Enter clustering data choice,(4)Average Temperature(DegC) , (5)Rainfaill(mm) ,(6)Average Humidity(F) : 5
starting period Month Day year
          1          1       2003

ending period Month Day year
         12         31       2003

Total number of days between starting and ending period
   365

  iter   phase       num           sum
     1       1       365       7.81084
     2       1        57       6.72087
     3       1        25       6.59199
     4       1         3       6.56305
     5       1         2       6.53916
     6       1         1        6.5062
     7       2         0        6.5062
7 iterations, total sum of distances = 6.5062
```

```
Enter the starting period
Enter the Month: 01
Enter the day: 01
Enter the year: 2000
Enter the ending period
Enter the Month: 12
Enter the day: 31
Enter the year: 2000
Enter K value = 4
Enter clustering data choice,(4)Average Temperature(DegC) , (5)Rainfaill(mm) ,(6)Average Humidity(F) : 6
starting period Month Day year
          1          1       2000

ending period Month Day year
         12         31       2000

Total number of days between starting and ending period
   366

  iter   phase       num           sum
     1       1       366       13.5609
     2       1        40        12.466
     3       1        14       12.2486
     4       1         7        12.199
     5       1         5       12.1794
     6       1         5       12.1264
     7       1         2       12.1109
     8       2         1       12.1089
8 iterations, total sum of distances = 12.1089
```

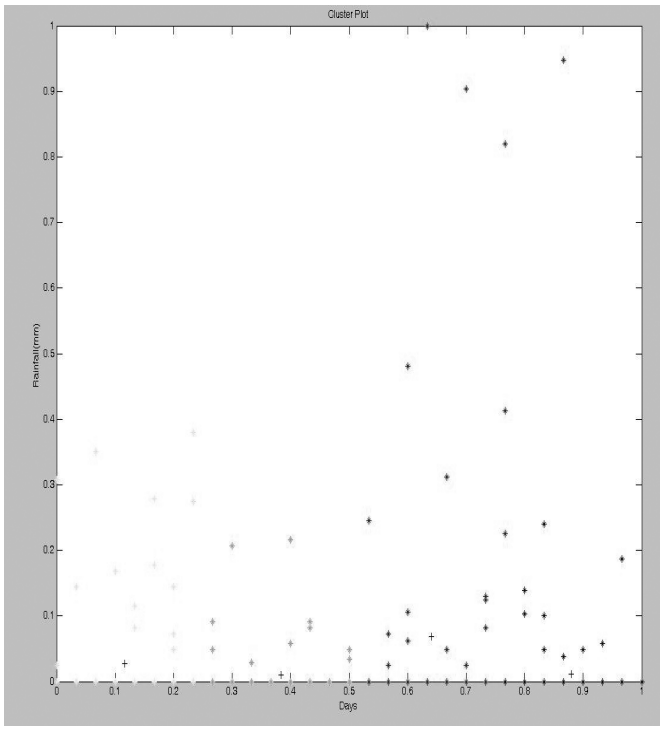

Fig. 2. One Year observation for rainfall

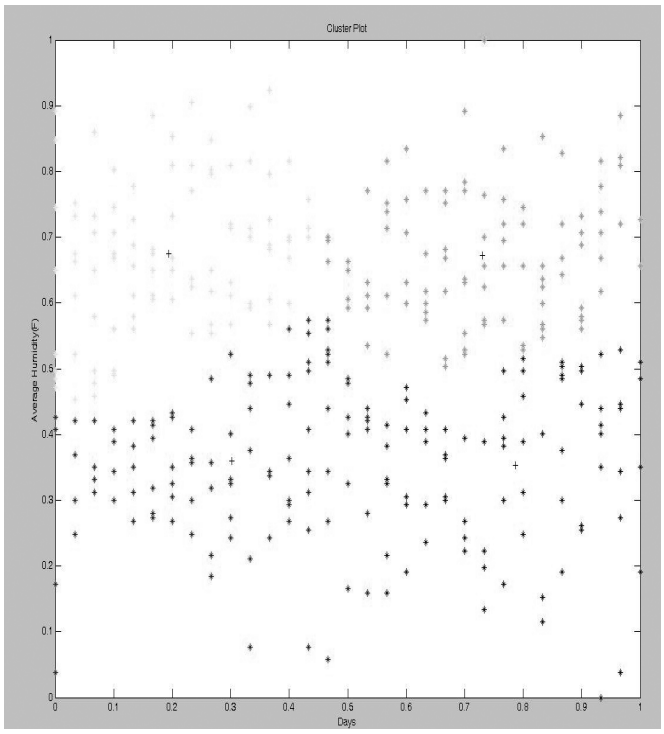

Fig. 3. One year observation for humidity

5 CONCLUSION

In this paper experimental results of k-means clustering algorithm on weather data show that it perform well by properly selecting the k value. Further work is being undertaken to use the clustered results to classify given object using NNC and kNNC classifiers and see whether k-means is suitable for clustering weather data.

6 REFERENCES

[1] P. Arabie and L.J. Hubert. An overview of combinatorial data analysis. In P.Arabie, L.Hubert, and G.D. Soete, editors, Clustering and classification, pages 5-63. World scientific pub., New Jersey, 1996.

[2] P.Cheeseman and J.Stutz. Bayesian classification (auto class): Theory and results. In U.M. Fayyad, G.Piatetsky-Shapiro, P.Smyth, and R. Uthuruswammy, editors, Advances in Knowledge Discovering and Data mining, Chapter 6, pages 153-180. AAAI/MIT press, 1996.

[3] R.O. Duda and P.E. Hart. Pattern classification and scene Analysis. John Wiley and Sons, 1973.

[4] S.Sridhar and Margaret H. Dunham. Data Mining Introductory and Advanced Topics. Pearson Education, 2006.

[5] Jiawei Han and Micheline Kamber. Data Mining Concepts and Techniques. Elsevier. 2000

[6] U.M.Fayyad, G. Piatetsky-Shapiro, P.Smyth, and R. Uthurusamy, editors. Avances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[7] K.Fukunaga. Introduction to Statistical pattern Recognition. Academic Press, 1990.

[8] A.K. Jain and R.C. Dubes. Algorithm for clustering Data. Prentice Hall, 1988.

[9] L.Kaufman and P.Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, 1990.

[10] R.S. Michalski and R.E. Stepp. Learning from Observation: Conceptual clustering. In R.S. Michaski, J.G. Carbonell, and T.M. Mitchell, editors, Machine Learning: An Artificial Intelligence Approach, Volume I, pages 331-363. Morgan Kaufmann, 1983.

[11] P.Sneath and R. Sokal. Numerical Taxonomy. Freeman, 1973.

[12] V.Cherkassky and F.Mulier, Learning From Data: Concepts, Theory, and Methods. New York : Wiley, 1998.

[13] R. Duda, P.Hart, and D.Stork, Pattern Classification, 2nd ed. New York: Wiley, 2001.

[14] Li, T., & Ma, S.(2004). IFD: Iterative Feature and Data Clustering. Proceedings of the 2004 SIAM International conference on Data Mining (SDM 2004).

[15] P.K.Agarwal and C.M.Procopiuc, “Exact and Approximation Algorithms for Clustering,” Proc. Ninth Ann. ACM-SIAM Symp. Discrete Algorithms, pp. 658-667, Jan. 1998.

[16] K.Alsabti, S.Ranka, and V.Singh, “An Efficient k-means Clustering Algorith”, Proc. First Workshop High Performance Data Mining, Mar. 1998.

[17] Indian Agriculture, Agriculture in India www.indianchild.com/india_agriculture.htm.

[18] Agriculture statistics at a glance 2004 http://agricoop.nic.in/statglance2004/AtGlance.pdf.