Geospatial Data Mining Techniques: Knowledge Discovery in Agricultural

KEYWORDS: Data Mining, KDD process, OLAP and Data Warehouses, Geo graphic Data Mining, Agriculture

Shital Hitesh Bhojani

Assistant Professor, IT Center AAU - Anand, Gujarat - 381001

ABSTRACT *Agriculture data are highly expanded in provisions of nature, interdependencies and resources. The Agricultural yield is primarily depends on weather conditions, diseases and pests, planning of harvest operation, geographical and biological factors etc. For balanced and sustainable development of agriculture these resources and factors need to be calculated, monitored and examined so that proper policy implication could be drawn. Sometimes ago data mining techniques are not used in agriculture but currently data mining and knowledge management in agriculture assisting data classification, forecasting, predictive and preconception purposes. Some type of evaluation has already been attempted for the agriculture data and till it is continuing. The aim of this attempt to write a paper is to apply the computational characteristic to the needs of agriculture data, as they are uncertain and fundamentally seasonal so use of data mining techniques be helpful in some aspect of agriculture.*

Indian agriculture is known for its diversity which is mainly result of variation in resource and climate, to topography and historical, institutional and socio economic factors. The agriculture is extremely branch out in terms of its crops, climate, soil, resources like fisheries , water, livestock et . Moreover, production performance of agriculture sector has followed on uneven path and large gaps have development in productivity between different geographic locations across the country. In addition, the diversity among resources generates interactions among many different factors. These resources and factors need to be evaluated, monitored, and allocated optimally for balanced and sustainable development of the country.

Knowledge Management System in Agriculture

Computerizing techniques are used to facilitating extraction, storage, retrieval, integration, transformation, visualization, analysis, dissemination, and utilization of larger data.

Knowledge Management System is a platform which uses these techniques. Using this system we can identify and extract the valid and potentially useful data; also we can build the databases and data warehouse; Knowledge discovery from databases(KDD) and generate a new mechanism as per the organization requirements. Knowledge discovery from databases (KDD) is a response to the enormous volumes of data being collected and stored in operational and scientific databases. Continuing improvements in information technology (IT) and its widespread adoption for process monitoring and control in many domains is creating a wealth of new data. There are no common standards or rules that are applied in data collection. In order to use the data or information for further reference and in decision-making , data have to be integrated and aggregated properly. For this purpose it's challenging to Design data warehouses to integrate the collected information or data.

The KDD Process and Data Mining

KDD refers to the overall process of discovering useful knowledge from data. The KDD process usually consists of several steps, namely, data selection, data preprocessing, data enhancement, data reduction and projection, data mining, and pattern interpretation and reporting to make the decision of what qualifies as knowledge. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process. The primary goals of data mining are:

- **Prediction**
  - Involves using some variables or fields in the database to predict unknown or future values of other variables of interest.
- **Description.**
  - Focuses on finding human-interpretable patterns describing the data.

Geographic aspects of Data Mining and Knowledge Discovery

It's a special case of KDD. When we relate data mining to geographic, other features like location, dimension, distance etc. are come into existence. Geographic data mining includes:

- Spatial segmentation (clustering, classification)
- Spatial dependency (spatial association rules)
- Spatial trend detection
- Geographic characterization and generalization

Spatiotemporal objects and relationships tend to be more complex than the objects and relationships in non-geographic databases. Data objects in non-geographic databases can be meaningfully represented as points in information space. Size, shape, and boundary roperties of geographic objects often affect geographic processes, sometimes due to measurement artifacts. Geographic information has always been the central commodity of geographic research. Geographic information was difficult to capture, store,and integrate. Cost revolutions in geographic research have been fueled by a technological advancement for geographic data capture, referencing, and handling, including the map, accurate clocks, satellites, GPS, and GIS. The current explosion of digital geographic and geo-referenced data is the most dramatic shift in the information environment for geographic research in history.

Computational Needs of Agriculture with Data mining

As stated earlier, agriculture data are very uncertain because of many reasons. Nowadays many research scholars are working on uncertain data in databases. The problem with uncertain data is how to represent and query the data with uncertainty? Data mining is the process of extracting important and useful information from large sets of data. Common applications of data mining are for instance the search for interactions among the genes of a living being and the search for relationships among the Web pages available in the Internet.

Data Mining Techniques

Data mining techniques can be mainly divided in two groups:

- **classification**
  - Classification techniques are designed for classifying unknown samples using information provided by a set of classified samples.

Volume : 3 | Issue : 1 | January 2013 | ISSN - 2249-555X

- **clustering techniques**
  - o clustering techniques can be used to split a set of unknown samples into clusters.

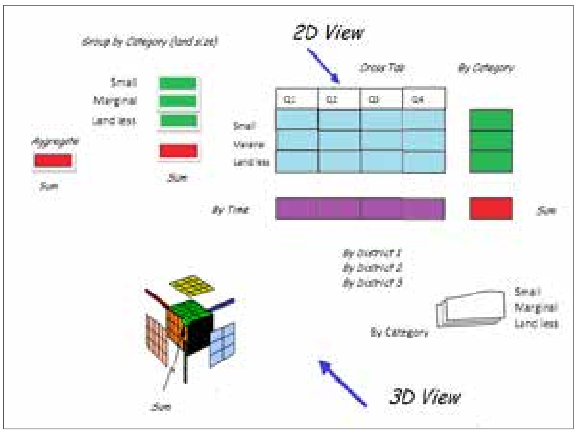
OLAP And Data Warehouses (DWs) -- its Application in Agriculture

The purpose of OLAP - Online Analytical Processing is to allow a multidimensional analysis of high-volume databases to conduct a special analysis of data. It provides an opportunity of viewing agriculture data from different perspective for analysis of Soil physical characteristics. Nowadays techniques like OLAP, Spatial OLAP, and multidimensional database applied to the management of national resources.

A DW is a repository that integrates data from one or more source databases. A DW usually exists to support strategic and scientific decision making based on integrated, shared information, although DWs are also used to save legacy data for liability and other purposes. Data and information are extracted from heterogeneous sources and new concepts are generated. A data warehouse is a repository of integrated information available for queries and analysis. So in brief we can describe that Data warehouse is a collection of databases that is used to store data for reporting and analysis, which can used in future for further research.

A data warehouse gives the option to analyze data from different sources under the same roof.

The following figure 1 depicts the OLAP cubes and DW sructure:



[Figure 1: Multi Dimensional Data Cubes for Data Warehousing]

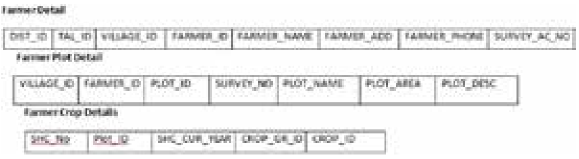
Gist of Agricultural data representation Using Data Mining Techniques

Data mining can be applied any type or kind of data/information with different approaches.

Data can be of different type like structured, unstructured, semi structured and may very each and every time. Currently we have the different kind of databases available like geographic database, relational database, multimedia database, time series database, data files or flat files, etc. Here describing some of the database which is used in mostly.

Relational Databases:

A relational database consists of set of tables, a table consists of set of attributes and attribute consist of set of values. A table is consisting of rows and columns, column represents the attributes and row represents the tuple i.e. record. A tuple is an object or relationship between objects and can be identified by set of values using primary or unique and foreign keys. In table 1 we present some relations like farmer details, farmer plot and crop details. These relations are part of Soil Health Card database, pretended for Gujarat government support program.



Farmer Detail

DIST_ID	TAL_ID	VILLAGE_ID	FARMER_ID	FARMER_NAME	FARMER_ADD	FARMER_PHONE	SURVEY_AC_NO

Farmer Plot Detail

VILLAGE_ID	FARMER_ID	PLOT_ID	SURVEY_NO	PLOT_NAME	PLOT_AREA	PLOT_DESC

Farmer Crop Details

SHC_NO	Plot_ID	SHC_CUR_YEAR	CROP_GR_ID	CROP_ID

[Table 1: Relational database Fragments for Agriculture Soil Health Card Farmer Detail]

Which kind of Data can be mined in Agriculture using relational database?

Data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. There are different king of query languages are available. But mostly, SQL (Structured Query Language) is used with relational database. Relational database uses DDL and DML commands to retrieve, insert, update and delete the data from database. Also it is providing many functions for calculating and displaying different kind of values. For instance, to select farmer plot and crop details we can use following SQL query:

```
select a.*,b.*,c.* from SHC_FARMER_MASTER a, SHC_FARMER_PLOT_MASTER b,SHC_MAIN_CROP c

where a.FARMER_ID=b.FARMER_ID and
        b.PLOT_ID=c.PLOT_ID
        and a.FARMER_ID='F00012706'
```

This will result into one record which holds the details for the particular farmer id i.e. 'F00012706' . If we want all farmer information then we have to just remove the fields from the query as below:

```
select a.*,b.*,c.* from SHC_FARMER_MASTER a, SHC_FARMER_PLOT_MASTER b,SHC_MAIN_CROP c

where a.FARMER_ID=b.FARMER_ID and
        b.PLOT_ID=c.PLOT_ID
```

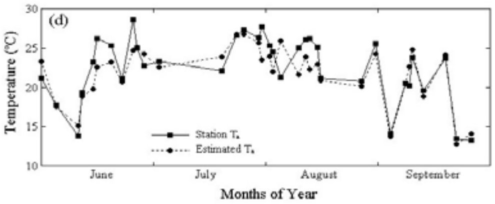
Transaction Databases

A database transaction is a logical unit of database operations which are executed as a whole to process user requests for retrieving, updating, deleting data from the database.

Since relational databases do not allow nested tables (i.e. a set as attribute value), transactions are usually stored in flat files or stored in two normalized transaction tables, one for the transactions and one for the transaction items. Normalization is the process where a database is designed in a way that removes redundancies, and increases the clarity in organizing data in a database. In easy English, it means take similar stuff out of a collection of data and place them into tables. Keep doing this for each new table recursively and you'll have a Normalized database. From this resultant database you should be able to recreate the data into its original state if there is a need to do so.

Time-Series Databases:

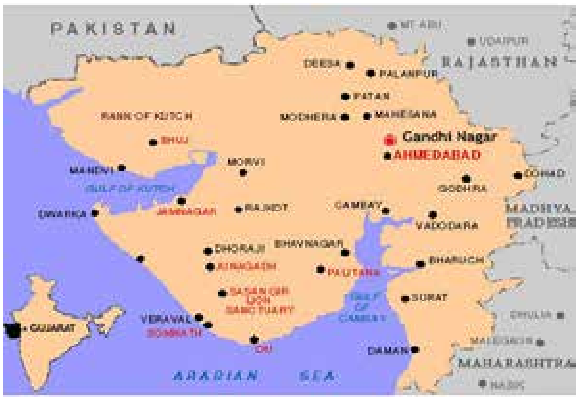
Time series data often arise when monitoring industrial processes or tracking corporate business metrics. "Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. " Data mining in such databases also includes the prediction of trends and movements of the variables in time. The following figure 2 shows some examples of time-series data.



[Figure 2: Time Series Database]

Geo Spatial Databases

These are the special kind of database which holds the geographic information like map, distance, location, longitude, latitude, etc. into the database. Such spatial databases features present new challenges to data mining algorithms. An ordinary database has types for strings, numbers and dates. A spatial database adds one or more types for representing geographic features. The basic geographic types are: GEOMETRY, POINT, LINESTRING, POLYGON, MULTIPOINT etc. The following figure 3 shows the visualization of Gujarat state.



[Figure 3: Visualization of Gujarat Spatial OLAP of Gujarat]

Conclusions

There are large amount of agricultural data and resources in which we can apply the data mining techniques. This is moderately a novel research field and it is expected to grow in the future. The area is open for this emerging and interesting research field. The multidisciplinary approach of integrating computer science with agriculture will generate new emission in forecasting/ managing agricultural information purpose. Using the data mining techniques we can face the challenges like:

a) Knowledge assessment
b) Knowledge processing
c) Knowledge Implementation

REFERENCE 1. “Data Mining: An effective tool for yield estimation in the agricultural sector”, by Raorane A.A., Kulkarni R.V, July – August 2012 (IJETTCS) | 2. "Data Mining of Geo-spatial Database For Agriculture Related Application" by Kiran Mai, C., Murali Krishna, I.V., A.Venugopal Reddy -- Map India 2006. New Delhi. | 3. “Principles of Knowledge discovery in Data bases”, by Osmar R.Zaiane -- department of computer science, University of Alberta. | 4. "Developing innovative applications in agriculture using data mining" by Cunningham S.J., G. Holmes. (2005). | 5. "The case for an agri data ware house: Enabling analytical exploration of integrated agricultural data". by Ahsan Abdullah, Stephen Brobst Brobst, M.Umer and M. Farooq Khan |