IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 04, Issue 03 (March. 2014), ||V6|| PP 33-35

A framework for Data Analysis using Data Mining

Harshank Godara, Pushpendra Kumar Petriya
[1](*M.Tech Scholar, Department of Computer Science and Engg, Lovely Professional University (LPU)*)
[2](*Assistant Professor, Department of Computer Science and Engg; Lovely Professional University (LPU)*)

Abstract: - The importance of data mining is increasing and studies have been done in many domains to solve tons of problems using various data mining techniques. But the data mining have not much applied in fields like poultry. The poultry data is huge and needs valuable and knowledgeable information from the large data produced every year. The data mining can be applied to it to get the knowledge and useful predictions. The analysis for the characteristics of poultry data is provided along with the proposed framework which consists of poultry data selection, poultry data pre-processing, poultry data mining and knowledge extraction. The list of tools which can be used for the analysis purpose is described and the expected results of how the tool produces are also shown. The framework can provide the methodical steps for scholars who are interested in the related researches of data mining and poultry farming. The framework can also be used as reference to other fields like agriculture.

Keywords: - *Data Mining, Classification, KDD, Poultry Data, Weka*

I. INTRODUCTION

Data Mining has been used widely in these recent years due to high availability of huge data. Data Mining can be applied to various kinds of data and aims at discovering interesting patterns from large amount of data. It creates these patterns by using models which takes input and produces one or more output like classification, prediction. The model is used for understanding phenomena from the data, analysis and prediction. For building model data must be there and this data is the data from the past. The data mining uses past data and produce the output which helps in taking better decisions in future. In this paper, we have purposed an analysis model for poultry data which is huge and needs to be mined for better usage.

II. DATA MINING & KDD IN REAL WORLD

Data Mining and KDD are used in contrast with each other. KDD provides the path-way to lead to the output whereas Data Mining provides the methods & techniques which helps in getting the desired result. Data mining application areas include financial data analysis, biological analysis, telecommunication industry, retail industry, medical data analysis & analysis in other scientific applications.[1]

Financial data analysis is required in loan payment prediction of customers, classification of customers for target marketing, detection of money laundering. Retail industry requires data mining for analysis of sales, customers, products, time, region, analysis of effective sales campaign, customer retention and product recommendation. Telecommunication industry requires mining of data for analyzing telecommunication data, fraudulent pattern analysis. Biological analysis includes discovery of various patterns in biological data like DNA, nucleotide/protein sequence, etc.

III. EXISTING WORK

Xiaojian long & Yuchun wu [2] provided a data mining model using decision tree algorithm for educational reform. R. Robu and C. Hora [3] discussed about the data mining technique on medical data and in more detail describe the classification technique on the medical data. Feixiang Huang, Shengyong Wang, and Chien-Chung Chan [4] predict hypertension from patient medical records by applying data mining process. Yuxiang Shao , Qing Chen , Weiming Yin [5] aims at improving teaching management, enhancing testing quality and distributing teaching resources with the help of data mining which is done with the improvement in ID3 algorithm. The dataset used is the student's employment rate. My Chau Tu, Dongil Shin, DongKyoo Shin [6] introduces Medical data mining with the use of decision tree C4.5 algorithm to identify the heart disease of a patient and compares the effectiveness among them. The data considered under study is from patients with coronary artery disease. B. M. Patil, Durga Toshniwal, R. C. Joshi. [7] presents an analysis on prediction of survivability of the burn patients. The algorithm C4.5 is used and is implemented in weka. Jianliang Meng & Yanyan Yang [8] discussed about C4.5 decision tree algorithm and uses the actual electric power marketing data as the basis for data mining & presents how this algorithm mines out marketing knowledge of behind the hidden electric power data so as to provide help for electric power enterprises operation and decision making.

IV. PROPOSED WORK

This section introduces a general framework which can be applicable to various datasets from various fields like medical, telecommunication and education. This framework considers the steps of KDD process and uses data mining technique to analyze the data.

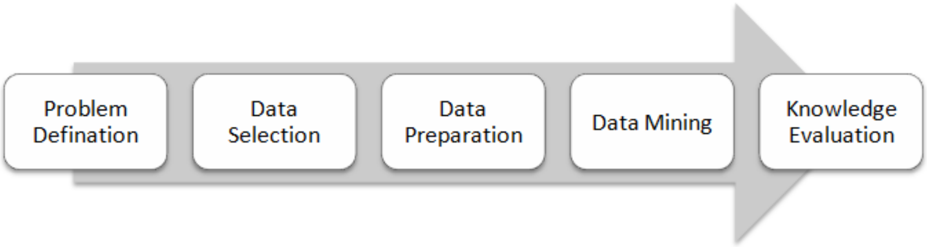

Fig.1. Framework inspired from KDD Process

4.2 Steps of Model

Step-1: Problem Definition: The data which needs to be analyzed must be identified in a particular field so that a particular requirement for analyzing is known and therefore valuable information is obtained. In poultry data the prediction of diseases is calculated and the result is analyzed and based on which it preventive measures can be taken.

Step-2: Data Collection: The data set for the study must be collected from the desired and accurate source. The data can be collected through personal interaction or interaction through telephone, schedules & questionnaires, surveys, interviews, etc.

Step-3: Data Preparation: Initially the raw data will get collected for the study. The data needs to be preprocessed which can be done in weka so that it can be made ready for the study and therefore the data quality is also improved. The data quality refers in terms of accuracy, completeness, consistency, timeliness, believability and interpretability. Many tools are available for data preprocessing like excel, RapidMiner, Weka, etc.

Step-4: Data Mining: This step provides the method & the tool for building the mining the data. It provides the model which does the actual work and leads to the desired output. All the models are built in 4 steps:

i. The first step is identifying a set of instances with a known behavior.
ii. The second step is data preparation – it includes data pre-processing.
iii. The third step is training the model by using some algorithm & methods and 80-20 split is generally used.
iv. The fourth step is testing the model where the accuracy of the model is checked.

Step-5: Knowledge Extraction: From a large data, information will be extracted which will be beneficial for the future purpose as well.

4.2.1 Methods of Data Mining

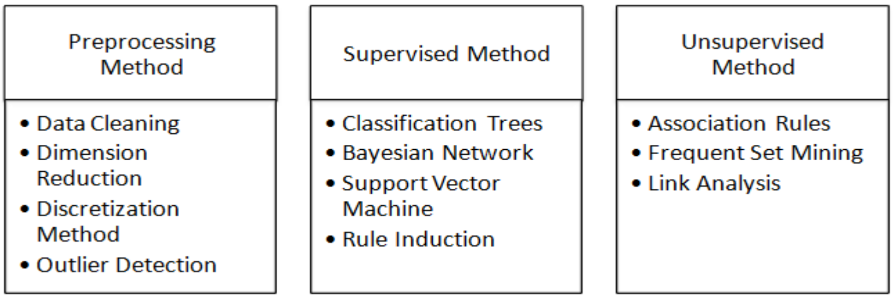

Fig.2. Methods in data mining

4.2.2 Tools used for mining data

Free Open-Source Tools	***Commercial tools***
Knime	*SPSS*
RapidMiner	*SAS*
Weka	*Oracle Data Mining*
Orange, R	*Microsoft Analysis Services*

4.2.3 Working of Model

1. Data will be divided into training set and test set.

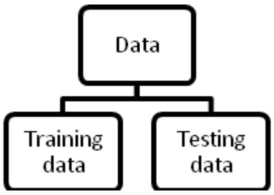

Fig.3. Training and Test Data

2. How these data sets will be applied.

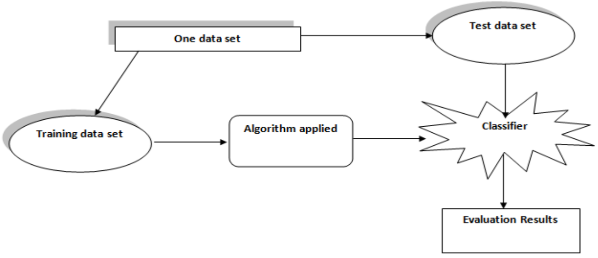

Fig.4. Schematic View of the Data Set Usage

V. CONCLUSION

In this paper, a framework for analyzing data is proposed which will generate refined results and will be useful for further analysis.

VI. ACKNOWLEDGEMENT

The paper making needed co-operation of lots of people so we thank all those who had helped with their kind support.

REFERENCES

[1] Jiawei Han & Micheline Kamber, Data Mining: Concepts and Techniques(Morgan Kaufmann Publication, San Francisco, 2012)

[2] Xiaojian Long & Yuchun Wu, Application Of Decision Tree In Student Achievement Evaluation, International Conference on Computer Science and Electronics Engineering (ICCSEE), 2012, 243 – 247

[3] R. ROBU and C. HORA, Medical data mining with extended WEKA, IEEE 16th International Conference on Intelligent Engineering Systems (INES), Lisbon, 2012, 347 – 350

[4] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, Predicting Disease By Using Data Mining Based on Healthcare Information System. IEEE International Conference on Granular Computing (GrC), Hangzhou, 2012, 191 - 194

[5] Yuxiang Shao , Qing Chen , Weiming Yin, The Application of Improved Decision Tree Algorithm in Data Mining of Employment Rate: Evidence from China. First International Workshop on Database Technology and Applications, Wuhan, Hubei, 2009, 202 – 205

[6] My Chau Tu, Dongil Shin, DongKyoo Shin, A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, Chengdu, 2009, 183 – 187

[7] B. M. Patil, Durga Toshniwal, R. C. Joshi., Predicting Burn Patient Survivability Using Decision Tree In WEKA Environment. IEEE International Conference on Advanced Computing. Patiala, 2009

[8] Jianliang Meng & Yanyan Yang , The Application of Improved Decision Tree Algorithm in the Electric Power Marketing. World Automation Congress (WAC), Puerto Vallarta, Mexico, 2009, 1 – 4