COMPUTING + MATHEMATICAL SCIENCES

AUT UNIVERSITY

Masters of Computer and Information Sciences

AUT UNIVERSITY COMPUTING + MATHEMATICAL SCIENCES

Survey of Three Data Mining Applications

Prepared for: Russel Pears
Prepared by: Mao Chuan Li
Student ID: 14854389
Submit Date: 2015/03/27
Paper Name: Data Mining & Machine Learning
Paper Number: COMP809

1 Introduction

Data Mining is a relatively new subdiscipline under computer science, which emerged in the 1990s, combines the power of mature statistical techniques and algorithms, artifical intelligence, and advanced database management systems to extract the hidden valuable knowledge from the muddling unmeaningful raw data to help people make better decisions.

Data Mining is data oriented, no matter what industry the data comes from. So it has been applied in a wide range of industries such as CRM, Banking, Healthcare, Insurance. More than 30 industries were voted in a public poll in 2014 (Kdnuggets.com, 2014).

This report selected 3 of the important industries: Education, Healthcare, and Weather forecast, and studied 3 applications of Data Mining techniques on each of them by understanding their backgrounds, analyzing the techniques used, and the benefits.

2 Discussion

2.1 Case A – DM Application in Education

2.1.1 Background

Far before the introduction of term Data Mining , universities have been established and served for more than hundreds of years. Each university has accumulated enormous students information, such as their names, gendar, religions, nationalities. Most important are their study degrees, marks for each subject and their graduation status.

By 2009, there were few studies with the help of Data Mining techniques to analyze the students' data and extract useful knowledge from them. The paper *Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining* was one of those papers to discover the hidden knowledge in students' history data.

The study was conducted by Siraj and Abdoulha(2009) from University Utara Malaysia, on the 8 years of data (1998 - 2006) collected in Sebha University in Libya, a universy established in 1983.

Sebha University was a public university in Libya (since there is no reliable information available in public about this university, the existence of it could not be determined at this time), which awarded both becheloar's and master's degrees in 9 faculties, including Science, Arts, Medicine, Dentisity, Law, Agronomy, Engineering, Accouting and Sports. All students were distributed in 6 branch campuses.

2.1.2 Data Description

The data was the enrollment dataset collected between 1998 and 2006 for all undergraduate's students, which contained 8510 records originally. Figure 1 shows a sample of the original student data (Siraj & Abdoulha, 2009).

Figure 1: Sample of Student Data

There were 38 attributes in the dataset, in which 8 attributes are numerical and others are categorical. The paper did not give any detail about how the raw data was cleaned in preprocessing phase, but just gave a result of 6830 records. Apparently, nearly 20% of noisy data had been cleaned out.

2.1.3 Tools and Algorithm Description

There were no specific tools mentioned in the paper, only a few Data Mining techniques are used for data analysis.

Descriptive Approaches

Firstly, cross tabulation analysis method was used to analyze the relationship among the 38 attributes.

Secondly, Kohonen network was used to determine the similarities among the attributes. Three groups of data were identified as Cluster 0, 1 and 2. The cluster attribute was also used for classification of each record, which was used to experiment the prediction of a student's classification.

Thirdly, correlation coefficient method was used to check which attributes had the closest relationship with the Cluster classification. Two significant attributes are identified. "Faculty" attribute had the strongest relation with Cluster, and "Nationality" had a medium relation with Cluster, which were further confirmed by the following predictive algorithms.

Predictive Approaches

With all data classified by the cluster(Cluster 0, 1 & 2), the records are partitioned into 3 groups: training (70%), validation (15%) and testing (15%). Neural Network, Logistic Regression and Decision Tree three algorithms are performed on the data.

Both regression and decision tree algorithms recognized that the 2 attributes "Faculty" and "Nationality" are importantly correlated to the class attribute Cluster found in previous steps. All three algorithms showed high accuracy of prediction, whilst Neural Network performed the best, which attained 99.98% accurate rate.

2.1.4 Discuss outcomes and benefits

The first half descriptive methods reaped a significant amount of valuable knowledge about the university's student gender distribution, population distribution among the different campuses and faculties. This provided the registrar office in Sebha University an overview of the current students status, and evidence for futher enrollment improvement.

The second half predictive methods showed a high accuracy of predicting the cluster of a

student, especially the Neural Network model, which can be used to predict the student status and help Sebha University to take proactive actions to help students as soon as possible.

2.1.5 Reference

Siraj, F., & Abdoulha, M.A. (2009). Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining. *Modelling & Simulation, 2009. AMS '09. Third Asia International Conference,*vol(no), 413-418. doi:10.1109/AMS.2009.117

2.2 Case B – DM Application in Healthcare

2.2.1 Background

When we are trying to solve a serious problem with the assistance of Data Mining techniques, the data might not always be ready there for us. We have to collect the data from scratch.

HIV is one of the world's deadly disease which could not be easily cured by present medical level. Researchers around the world are trying hard to fix this medical mystery by all means. The data about the HIV patients' were rarely collected. Two researchers from Guru Gobind Singh Indraprastha University(GGSIPU) developed a brand new system named ART(antiretroviral therapy) system, deployed it to all ART Clinics of All India Institute of Medical Sciences (AIIMS) in New Delhi.

All kinds of medical data was collected, in which most importantly there are 1054 records were for HIV patients. The paper was a research based on this data.

2.2.2 Data Description

The data collected by ART systems included 1054 records. After cleanning and preprocessing, there were 672 valid records left for data mining. After feature selection, there were 9 significant attributes selected:

1) @attribute 'Age' numeric

2) @attribute 'Sex' {1, 2, 3}

3) @attribute 'Marital Status' {1, 2}

4) @attribute 'Route of transmission' {1, 2, 3, 4, 5, 6, 7, 8, 9}

5) @attribute 'Body Weight' numeric

6) @attribute 'HAART Regime' {1, 2}

7) @attribute 'TLC' numeric

8) @attribute 'DLC' numeric

9) @attribute 'Hemoglobin' numeric

2.2.3 Tools and Algorithm Description

In the preprocessing phrase, a few regular cleaning techniques were applied to the raw data, including: filling the missing values, identifying the outliers, smooth out the noisy data. For feature selection, no algorithms were used.

For mining the data, Microsoft Business Intelligence SQL Server 2008 was used to store these data, and help analyze them. Two techniques were used:

1. Microsoft Decision Trees was used to classfied the age of patients into 4 groups, in which different routes of transmission were associated. Other attributes such as gender and literacy level were also grouped to show the percentage of patients having the HIV disease.

2. Microsoft Association Rules was used to genereate an association dependency network. But it did not accurately express the relationship between the classifiers.

2.2.4 Discuss outcomes and benefits

First of all, the most important outcome of the reseach is the collection of the first hand

data about the HIV patients in India, which provided a baseline for HIV disease analysis.

Secondly, the decision tree algorithm applied to the data has showed an overview of the disease occurrence in different age, gender and literacy groups.

There are 3 important findings in the research:

- For those who are over 46 years old, blood transfusion is the main route of transmission of HIV disease; for whose under 45, heterosexual or mother-to-child is the main route

- In the sample 500 male are suffering the HIV, almost 3 times of female(172)

- From the education point of view, illiterate or less educated people are the main population that are suffering the deadly disease.

With these findings Indian government could effectively develop new polies and programs to tackle this dealy disease and eliminate the effect of HIV in India. For other countries, this might be a good reference and may apply the same method to study the HIV population.

2.2.5 Reference

Gosain, A., & Kumar, A. (2009). Analysis of health care data using different data mining techniques *Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009. International Conference vol(no)*. 1-6. doi:10.1109/IAMA.2009.5228051

2.3 Case C – DM Application in Weather Forcast

2.3.1 Background

Anyone could make a rainlfall prediction solely according to his or her feelings. But a scentific weather forcast is not as easy as that, which needs a huge amount of weather data and computing time on some high performance supercomputers. Even so, the forecast result accuracy is not close to 100%.

Unlike the traditional weather forecast models, like Weather Research and Forecasting (WRF) model and Global Forecast System (GFS), which requires intensive computing power to simulate the atmospheric changes and calculate to predict the weather, Data Mining models just need intensive historic data and moderate computing resources.

Two researchers from Veermata Jijabai Technological Institute (VJTI) obtained the weather data from Indian Meteorological Department (IMD) Pune, and applied the Data Mining techniques to analyze the data and predict the rainfall probobilities, which achieved a high rate above 90%.

2.3.2 Data Description

The two researchers carefully selected 6 months weather data provided by India Meteorological Department (IMD) Pune. There were 36 climate features in the raw dataset, 7 of them were selected as the key features relating to rainfall:

1. @attribute 'Temp' numeric

2. @attribute 'Station Level Pressure' numeric

3. @attribute 'Mean Sea Level Pressure' numeric

4. @attribute 'Relative Humidity' numeric

5. @attribute 'Vapor Pressure' numeric

6. @attribute 'Wind Speed' numeric

7. @attribute 'Rainfall' numeric

Totally there were 14560 records used for train the model, and 1542 records used for test the model.

2.3.3 Tools and Algorithm Description

As always, the weather raw data need cleaning and preprocessing. Although the paper did not give any detail about that, it gave a diagram to show how the data was processed,

from which we can tell that Normalization and Transformation techniques were used.

No tools were mentioned in this paper, but a pseudo code for how to calcuate the probability for a given event was given based on the Naive Bayes theroem.

P(H | E) =

$$\frac{P(Temp|H)*P(MSST|H)*P(WindSpeed|H)*P(Humidity|H)*P(VP|H)*P(SLP|H)*P(H)}{P(E)}$$

If P(Yes | E) > P(No | E), then the event is classfied as 'Yes', otherwise classfied as "No".

2.3.4 Discuss outcomes and benefits

The application of Data Mining techniques on the historic weather data to predict rainfall probability was proved to be almost as accurate as the present computing intensive models like WRF. Without the need of powerful supercomputers and clusters to calculate and predict the weather, the application of the establed model based on Bayesian algorithm could dramatically save the electric energy and decrease the investment of supercomputers for weather forecast.

Although there is no other algorithm used in the paper to compare the performance of prediction, the accuracy of the Naive Bayes model is satisfactory for weather forecast. The loose dependency between the weather data attributes may contribute more or less for the result.

2.3.5 Reference

Nikam, V.B., & Meshram, B.B. (2013). Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach. *Computational Intelligence, Modelling and Simulation (CIMSim) 2013 Fifth International Conference on vol(y),* 132-136. doi:10.1109/CIMSim.2013.29

3 Conclusion

As the above 3 case studies showed, Data Mining subject could apply to different kinds of

industries like Healthcare, Weather forecast, and even in Education. Wherever there is data, there could be hidden gold awaiting us to mine to help us better understand the world around us and help accurately predict what will happen in the further.

With the rapid development of information systems, big dataset are becoming more and more prevalent in each field. How to effectively manage these data and utilize them is a big chanllenge to those data owners. Data Mining techniques have been proved to be beneficial for all kinds of information areas with its mature techniques and products in market. The application of Data Mining shall play a more important role in those fields in the future.

4 References

KDnuggets.com. (2014). *Industries where you applied Analytics, Data Mining, Data Science in 2014. Retrieved from* http://www.kdnuggets.com/polls/2014/ industries-applied-analytics-data-mining-data-science.html