Advanced Journal of
AI and Robotics
Book Chapter

DATA MINING

Published in April 2020 ISSN: 2737-4440

https://ajrjournal.com/

Author

Dr. M. Arathi

Associate professor

JNTUH, Hyderbad, India

LARIX INTERNATIONAL
Your Global Research Partner

Advanced Journal of AI and Robotics
Available online through
https://ajrjournal.com/

Book Chapter
ISSN: 2737-4440

Vol. 01 Issue 01, 2020

DATA MINING

Dr. M. Arathi
Associate Professor in JNTUH, Hyderabad, Andhra Pradesh, India

Received on: 05-04-2020; Revised and Accepted on: 28-04-2020

INTRODUCTION

Most of the transactions have been computerized in day to day life, which in turn produces lot of digital data. For example point-of-sale, internet shopping & browsing, credit cards, information on credit cards, purchase patterns, payment history, sites visited etc. One trip by one person generates info on destination, airline preferences, seat selection, hotel, rental car, name, address, restaurant choices and so on. Data size is so huge that it cannot be processed or even inspected manually. Automated data collection tools and mature database technology lead to tremendous amounts of digital data stored in databases, data warehouses and other information repositories. Vast quantities of data are collected and stored out of fear that important information will be missed.

The Data volume grows so fast that old data is never analyzed. Only a small portion of data collected is analyzed.

The database systems do not support queries like "Who is likely to buy product X", "List all reports of problems similar to this one", "Flag all fraudulent transactions". But these may be the most important questions!

Why mine data? There is often information 'hidden' in the data that is not readily evident. More often, data mining yields unexpected nuggets of information that open the company's eyes to new markets, new ways of reaching customers and new ways of doing business". Human analysts may take a very long time to discover useful information.

What Is Data Mining? Data mining (knowledge discovery in databases) is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. The alternate names for Data mining are Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Data Mining can be done on following types of data:

- **Database-oriented data sets and applications**
  - Relational database, data warehouse, transactional database
- **Advanced data sets and advanced applications**
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web
- In data mining, following patterns can be mined:
  - Concept/class Descriptions
  - Mining frequent patterns, Associations & correlation
  - Classification & Prediction
  - Cluster Analysis

***Corresponding author:**
M. Arathi
Associate Professor in JNTUH, Hyderabad
Andhra Pradesh, India.
Email: arathi.jntu@gmail.com
DOI: https://doi.org/10.5281/zenodo.7110924

- Outlier Analysis
- Evolution Analysis

In concept/class description, summary of category of tuples will be obtained. In association rule mining, frequent patterns will be identified and interesting association rules be generated be considering correlation among objects. In classification & prediction, a model will be generated to classify or predict a value for given attribute. In clustering, the data tuples are grouped into clusters distance/similarity measure between data points. It is called unsupervised learning as opposed to classification which is supervised learning. In outlier analysis, the data objects that do not comply with general behavior or model of data are identified. Evolution analysis describes & models regularities or trends for objects whose behavior changes over time.

Data pre-processing

Data in the real world is dirty. That is, it is incomplete, noisy and inconsistent. In data pre-processing, we try to remove such type of data. There are various reasons for the data to be incomplete such as attributes of interest are not available, data not recorded because of misunderstanding or malfunctions, data may have been recorded and later deleted, etc. The various reasons for data to be noisy or inconsistent are faulty instruments for data collection, human or computer errors, errors in data transmission, technology limitations (e.g., sensor data come at a faster rate than they can be processed) etc.

The data pre-processing includes some of the following tasks:

- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization

In data cleaning, missing and noisy data are identified and removed.

To handle missing data, following strategies are used:

- Ignore the tuple
- Fill in the missing value manually
- Use a global constant to fill in the missing value: e.g., "unknown". This leads to a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class to fill in the missing value
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

There are various methods to handle Noisy Data. Some of those are:

- Binning method
- Clustering: the outliers are detected through clustering and removed.
- Combined computer and human inspection :computer detects suspicious values, which are then verified by humans
- Regression: smooth by fitting the data into regression functions.

Data Integration

Redundant data occur often when integration of multiple databases is done.

The redundant attributes are detected by correlation analysis (also called Pearson's product moment coefficient) and Chi square test.

Correlation Analysis (Numerical Data)

The Correlation coefficient for two numeric attributes A and B is computed as follows:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

where N is number of tuples, ai & bi are the values of A & B in tuple i respectively, and means of A and B, σA and σB are the respective standard deviation of A and B, Σ(aibi) is the sum of the AB cross-product.

The coefficient value will be between -1 and +1 (inclusive). If $r_{A,B} > 0$; A and B are positively correlated. If $r_{A,B} = 0$; A and B are independent; and if $r_{A,B} < 0$; A and B are negatively correlated.

Chi – square test (Categorical Data)

Suppose A has c distinct values, namely a1,a2,..ac and B has r distinct values namely b1,b2,…br. The data tuples described by A & B can be shown as contingency table, with the c values of A making up the columns and r values of B making up the rows.

	a1	a2	…..	ac
b1				
b2				
…..				
br				

X2 (chi-square) test :

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where,

$$e_{ij} = \frac{count(A = a_i) * count(B = b_j)}{N}$$

where N is no. of tuples, count(A=ai) is no. of tuples having value ai for A and

e_{ij} is expected frequency for i^{th} row and j^{th} column.

The X2 test the hypothesis that A & B are independent. The larger the X2 value, the more likely the variables are related.

Data Transformation

The various data transformation techniques are as follows:

1. Smoothing: remove noise from data

2. Aggregation: summarization, data cube construction

3. Generalization: concept hierarchy climbing

4. Normalization: scaled to fall within a small, specified range. Some of the normalization techniques are min-max normalization, z-score normalization, normalization by decimal scaling

5. Attribute/feature construction

6. New attributes constructed from the given ones

Min-max normalization

Suppose that minA & maxA are the min & max values of an attribute A. Mapping a value v of A to v' in the range [new_minA, new_maxA] :

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Z-score normalization:

The value v is normalized as follows:

$$v' = \frac{v - \mu_A}{\sigma_A}$$

where μ is mean of vector v and σ is standard deviation of v.

Normalization by decimal scaling

V is mapped to v' as:

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that Max(|v'|) < 1

Suppose values of A range from -83 to 67. The max. absolute value of A is 83.

To normalize each value is divided by 102. Hence, -83 and 67 are mapped to -0.83 and 0.67 respectively.

Data reduction

- The DW will be usually too huge to be analyzed.
- Complex data analysis & mining on such data may take a very long time.
- Data reduction techniques obtain a reduced representation of the data set, yet closely maintains the integrity of original data.

Strategies for data reduction are:

- Data cube aggregation
- Attribute subset selection
- Dimensionality reduction
- Numerosity reduction

- Discretization and concept hierarchy generation
- Attribute subset selection
- Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to mining task.
- Reduced data set can be obtained by removing irrelevant attributes.

Basic methods for attribute subset selection are:

- Step-wise forward selection
- Step-wise backward elimination

Initial attribute set:

–{A1, A2, A3, A4, A5, A6, A7, A8, A9, A10}

–{A1, A2, A3, A4, A5, A6, A8, A9, A10}

–{A1, A4, A5, A6}

Reduced attribute set:

–{A1, A4, A6}

- Combination of forward selection and backward elimination
- Decision tree induction

Dimensionality reduction

- In data compression, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data.
- Data compression techniques can be categorized as:

  – Lossy

  – lossless

Numerosity Reduction

• Reduce data volume by choosing alternative, smaller forms of data representation

• Parametric methods

– Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data

• Non-parametric methods

– histograms, clustering, sampling

• Regression Analysis

• Linear regression:

A random variable y can be modeled as a linear function of another random variable x,

$$Y = w X + b$$

–Here x & y are numerical attributes

–Two regression coefficients, w and b, represent slope and y-intercept respectively.

• Multiple regression:

$$Y = b0 + b1 X1 + b2 X2.$$

Histograms

- A histogram for an attribute, A, partitions the data of A into disjoint subsets or buckets.
- If each bucket represents only a single value, the buckets are called singleton buckets.
- Consider following data which represents prices of commonly sold items at a store:

1,1,5,5,5,5,5,8,8,10,10,10,10

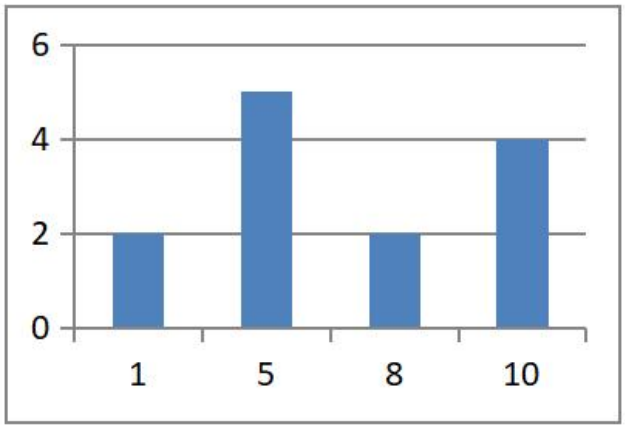
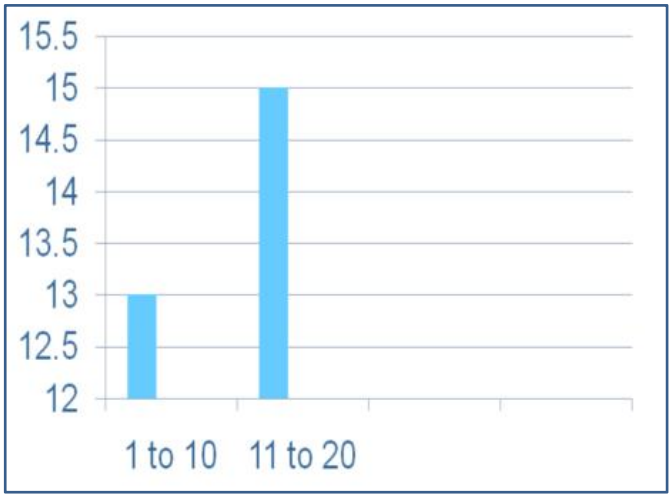


Fig. A histogram for price using singleton buckets

• Often, buckets represent continuous ranges for the given attribute.

• To further reduce the data , each bucket can represent a continuous range of values for the given attribute.

1,1,5,5,5,5,5,8,8,10,10,10,10,12,14,14,14,15,15,15, 15,15,15,18,18,20,20,20



Clustering

• Partition data set into clusters based on similarity, and store cluster representation

Sampling

Sampling: obtaining a small sample s to represent the whole data set N

Different methods:

– Simple random sample without replacement (SRSWOR) of size s: This is created by drawing s of N tuples from D(s<N), where the probability of drawing any tuple in D is 1/N, i.e.,all tuples are equally likely to be sampled.

– Simple random sample with replacement (SRSWR) of size s: This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced.

Cluster sample

• If the tuples are grouped into M mutually disjoint clusters, then select s clusters where s < M

• Stratified Sampling : Same as cluster sample except that instead of selecting clusters randomly, half of the tuples are selected from each cluster.

Discretizing Numeric Attributes

• We can turn a numeric attribute into a nominal/categorical one by using some sort of discretization.

• This involves dividing the range of possible values into subranges called buckets or bins.

• example: an age attribute could be divided into these bins:

child: 0-12

teen: 12-17

young: 18-35

middle: 36-59

senior: 60-

References:

1. R. Agrawal, T. Imielinski, and A. Swami (1993). "Mining associations between sets of items in massive databases," in Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data (pp. 207–216), New York: ACM Press.

2. M. J. A. Berry, and G. S. Linoff (1997). Data Mining Techniques. New York: Wiley.

3. M. J. A. Berry, and G. S. Linoff (2000). Mastering Data Mining. New York: Wiley.

4. L. Breiman, J. Friedman, R. Olshen, and C. Stone (1984). Classification and Regression Trees. Boca Raton, FL: Chapman & Hall/CRC (orig. published by Wadsworth).

5. J. Han, and M. Kamber (2001). Data Mining: Concepts and Techniques. San Diego, CA: Academic.

6. D. Hand, H. Mannila and P. Smyth (2001). Principles of Data Mining. Cambridge, MA: MIT Press.

7. T. Hastie, R. Tibshirani, and J. Friedman (2009). The Elements of Statistical Learning. 2nd ed. New York: Springer.

8. D. W. Hosmer, and S. Lemeshow (2000). Applied Logistic Regression, 2nd ed. New York: Wiley-Interscience.

How to cite this article:

M. Arathi, DATA MINING. **AJR 2020; 01 (01): 01-06**. **DOI:** https://doi.org/10.5281/zenodo.7110924

Conflict of interest: The authors have declared that no conflict of interest exists.
Source of support: Nil

ABOUT AUTHOR

Dr. M. Arathi is working as an Associate Professor in JNTUH, Hyderabad, Andhra Pradesh, India. It is more than 17 years since she has been with JNTUH. She received her B.E.(CSE) from MVSREC, Hyderabad, Andhra Pradesh, India, in 2001, M.Tech(CS), JNTUH, Hyderabad, Andhra Pradesh, India, in 2008 and Ph.D(CS) from JNTUH, Hyderabad, Andhra Pradesh, India in 2020. Her Major fields of study are Data Mining and Big data analytics. She is currently the coordinator for MCA program and Training and Placement officer at SIT, JNTUH, Hyderabad. Previously, she held the post of Officer-in-charge of Examinations for SIT, JNTUH, Hyderabad.

Dr. M. Arathi is an expert committee member for Institute for Innovations in Science and Technology. She has been a judge for many paper presentation contests in JNTUH. She has delivered many lectures on QTP testing tool and Data analytics in Refresher Courses held by JNTUH and other Universities. She has helped the University in finalizing syllabus for many subjects. She has published many papers in international/national journals and international/national conference. She has organized many workshops and conferences.

