Data Preprocessing in Real-time Education Management System

Kristine Hambardzumyan
NPUA
Yerevan, Armenia
e-mail: hambardzumyan.k@polytechnic.am

***Abstract*— The aim of this abstract is to present widely available educational data which provides the opportunity to prepare helpful data to be able to analyze. Throughout the educational process big data is collected which necessitates the need to analyze which, as a result, is likely to improve the educational quality. Data mining is becoming even more important in an educational context. Data mining basically depend on the quality of data. Raw data usually susceptible to missing values, noisy data, incomplete data, inconsistent data and outlier data. So, it is important for these data to be processed before being mined. Preprocessing data is an essential step to enhance data efficiency. Data preprocessing is one of the most data mining steps which deals with data preparation and transformation of the dataset and seeks at the same time to make knowledge discovery more efficient. Preprocessing includes several techniques like cleaning, integration, transformation and reduction. This study shows a detailed description of data preprocessing techniques which are used for data miming.**

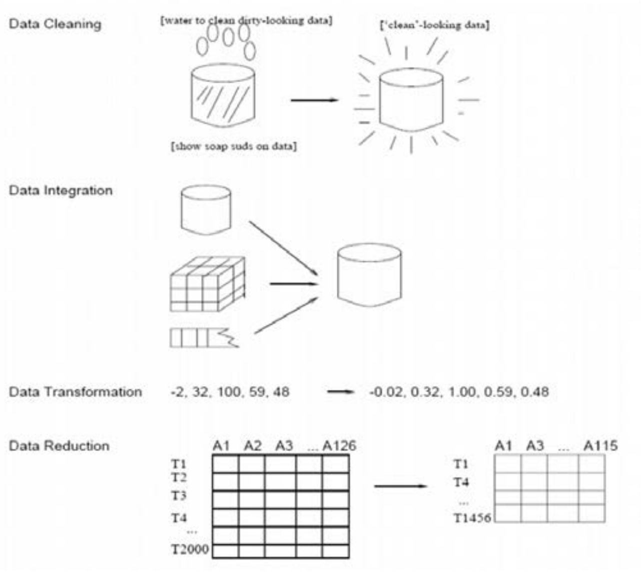
***Keywords*— Data mining, data preprocessing, data set, Naïve Bayes, dataset, Apriori algorithm.**

I. Introduction

Knowledge Discovery in Databases (KDD) is a process of extraction valuable information from huge data sources. Data mining is a step of KDD which is performs analysis and models for huge dataset using classification, clustering, association rules and many other techniques. The raw data are highly vulnerable to missing, noise, outliers and inconsistent because of their huge size, multiple resources, and their gathering methods. The poor-quality data will effect or data mining results [1].

II. Preprocessing techniques

Data preprocessing is one of the most data mining tasks which includes preparation and transformation of data into a suitable form to mining procedure. Data preprocessing aims to reduce the data size, find the relations between data, normalize data, remove outliers, and extract features for data. It includes several techniques like data cleaning, integration, transformation, and reduction Fg.1.



Fg.1: Preprocessing forms

III. Data cleaning

Row data may have incomplete records, noise values, outliers and inconsistent data. Data cleaning is a first step in data preprocessing techniques which is used to find the missing values, smooth noise data, recognize outliers and correct inconsistent. These dirty data will effects on miming procedure and led to unreliable and poor output. Therefore, it is important for some data-cleaning routines to be used [5].

Missing values: If there are records with unrecorded values for its records then these values may be filled using the following ways.

Ignore the tuple: This chois is selected when the value of class label is not existing. This method is not effective, but it is used when the tuple has several attributes with empty values.

Fill the missing value manually: This approach in general requires human effort and consuming. It cannot be used with the large size of dataset.

Use a global constant to fill the missing value: This method works by replacing missing values of attribute by a particular constant which is similar for all records for example using “Unknown” as a label. This method have problems because when the missing values are replaced by a specific term

“Unknown” as an example, the mining programs may believe that they form an important concept, since they a common value.

Use the attribute mean to fill the missing value: This method works by replacing the missing value for a particular attribute by the average value for that attribute.

Use the attribute mean for all samples belonging to the same class as the given tuple: For example, if we classify users depending on credit risk, the missing value can be replaced by the average value of income for the users which belong to the similar credit risk class for a given tuple.

Use the most probable value to fill the missing value: This approach is used with techniques like inference-based regression using a decision tree induction or Bayesian formalism.

Naïve Bayes: A naïve (or simple) Bayesian classifier based on Bayes’ theorem is a probabilistic statistical classifier [4], which the term “naïve” indicates conditional independence among features or attributes. Its major advantage is its rapidity of use because it is the simplest algorithm among classification algorithms. Hence, it can readily handle a data set with many attributes.

The naïve Bayesian classifier, or simple Bayesian classifier, works as follows [4]:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x_1 , x_2 , …, x_n), depicting n measurements made on the tuple from n attributes, respectively, A1, A2, …, An.

2. Suppose that there are m classes, C1, C2, …, Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if

$$\boldsymbol{PC_i|X) > P(C_j|X) for\ 1 \le j \le m; j \neq} \mathbf{i}$$

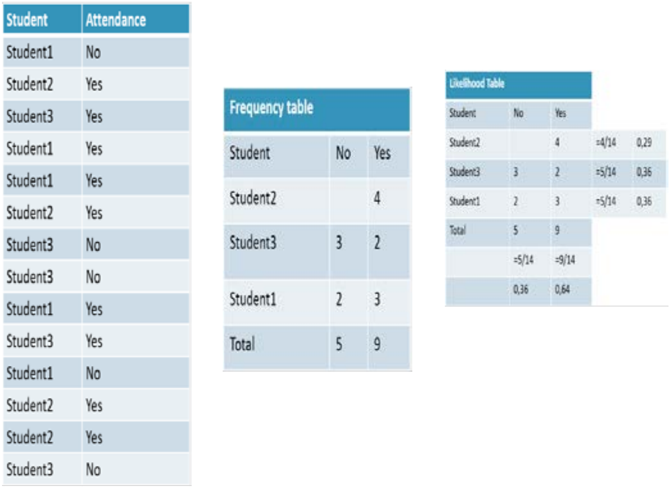
Thus, we maximize $\boldsymbol{P(C_i|X)}$. The class Ci for which $\boldsymbol{P(C_i|X)}$ is maximized is called the maximum posteriori hypothesis. By Bayes’ theorem,

$$\boldsymbol{P(C_i|X) = P(X|C_i)/P(X)}$$

3. As P(X) is constant for all classes, only $\boldsymbol{P(X|C_i)P(C_i)}$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) = … = P(Cm), and we would therefore maximize $\boldsymbol{P(X|C_i)}$**.**

Otherwise, we maximize $\boldsymbol{P(X|C_i)P(C_i)}$. Note that the class prior probabilities may be estimated by $\boldsymbol{P(C_i) = |C_i\ D|/|D|}$, where $\boldsymbol{|C_i, D_j|}$ is the number of training tuples of class Ci in D.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $\boldsymbol{P(X|C_i)}$. To reduce computation in evaluating $\boldsymbol{P(X|C_i)}$, the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes).



Student	Attendance
Student1	No
Student2	Yes
Student3	Yes
Student1	Yes
Student1	Yes
Student2	Yes
Student3	No
Student3	No
Student1	Yes
Student3	Yes
Student1	No
Student2	Yes
Student2	Yes
Student3	No

Frequency table		
Student	No	Yes
Student2		4
Student3	3	2
Student1	2	3
Total	5	9

Likelihood Table				
Student	No	Yes		
Student2		4	=4/14	0,29
Student3	3	2	=5/14	0,36
Student1	2	3	=5/14	0,36
Total	5	9		
	=5/14	=9/14		
	0,36	0,64		

Fg.2: Naïve Bayes example

P (Yes|Student1) = P(Student1|Yes) * P(Yes) / P (Student1)
P (Student1 |Yes) = 3/9 = 0.33
P (Student1) = 5/14 = 0.36
P(Yes)= 9/14 = 0.64
P (Yes | Sudent1) = 0.33 * 0.64 / 0.36 = 0.60

Noise data: One of the most problems which effects on mining process is noise. Noise is a random error or variance in a measured variable. Noise data means that there is an error in data or outliers which deviates from the normal. It can be corrected using the following methods:

Binning: Binning ways swish organized information esteem by direction the "sector", or qualities around it. The organized esteems unit of measurement circulated into varied "containers", or canisters. Since binning techniques counsel, the realm of qualities, they perform neighborhood smoothing.

Clustering: Outliers might be identified by grouping, where comparable esteems are sorted out into gatherings. Naturally, values which fall outside of the arrangement of bunches might be considered anomalies.

Regression: Data are going to be smoothed by fitting the information to a capability, as associate degree example, with relapse. Direct relapse includes finding the alone line to suit two factors; with the goal that one variable is going to be accustomed anticipate the other.

IV. Data Integration

This technique works by combining data from multi and various resources intone consistent data store, like in data warehouse. These resources cart have multi database, files or data cubes. In data integration There are a number of issues for consideration, like Schema integration, object matching and redundancy which are an important aspect. Each attribute like “annual revenue” is considered as redundant if it “derived” from another attribute or set of attributes. In consistence in attribute or dimension is another form of redundancies. Correlation analysis can be used to detect some redundancies. The correlation between two variables can measure how the attributes can imply one the other strongly. The correlation between (X, Y) attributes can be evaluated by finding the correlation coefficients.

V. Data transformation

Data transformation includes transforming the data to forms suitable for mining process. It involves the following:

Smoothing: It removes noise from data. It includes techniques such as clustering, regression and binning.

Aggregation: It is the process of applying statistical metrics like means, median and variance which are necessary to summarize the data. The resulted aggregated data are used in data mining algorithms. For example, apply aggregation on the daily sales to compute monthly and annual sales. **Generalization:** It includes replacing the lower-level data (primitive) by higher level using hierarchical concepts. An example, street which is a type of categorical attributes may be replaced to city or country which is high level terms. Another example, age which is a type of numeric concepts can be mapped to senior, younger and youth which are high level concepts.

Normalization: This method works by a adjusting the data values into a specific range such as between 0-1 or -1-1. This method is useful for mining techniques like classification, artificial neural networks, and clustering algorithms. Using the normalization to scale the data attributes in tanning face for back propagation neural network algorithm can be used to speed the learning stage Minimum-maximum, z-score and decimal scaling are popular forms of normalization.

VI. Data reduction

These techniques can be used to reduce the representation of dataset in smaller volume with respect to maintain the integrity of the original dataset. Thus, a better data results can be obtained from applying mining techniques on that reduce data. The following subsection shows data reduction strategies [2]:

Data cube aggregation: This approach construct data cube by applying operations of aggregation on data without losing the necessary information for the data analysis. **Attribute subset selection**: It reduce the dataset size by removing redundant features or dimensions and irrelevant attributes.

Dimensionality reduction: Alsco known as (data compression) it uses the mechanisms of encoding to reduce the size of dataset. Reduction can be lossless and lossy based on the retrieved information from decoding. Wavelet Transforms and Principal Component Analysis (PCA) are two effected methods for lossy reduction (Larose, 2006). **Numerosity reduction**: It replaces or mapped data to an alternative or smaller representation of data. It consists of parametric and non-parametric models. The first model needs to store only the parametric of model without storing whole data. While in second model, it includes techniques such as sampling, histogram, and clustering.

Data discretization and concept hierarchy generation: These techniques can be used to replace the attributes data values by high level of conceptual or interval ranges. It is a type of numerosity reduction which is very useful for the generation of hierarchal automatically. One of the important tools of data mining is hierarchical and discretization which are perform mining in multi abstraction levels. Data discretization can be classified based on how it performed into supervised or unfollowing et al, it is top-down or bottom-up discretization It is consist of the following techniques:

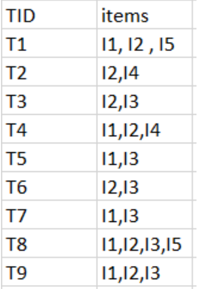
Binning: It is a splitting top to down technique which depend on the determined bins number. Binning methods which are used for data smoothing are also used in discretization and hierarchy generation. Equal-width or equal-frequency binning can be used to discretize the values of attribute by replacing bin value by mean or median as in smoothing. This process can recursively repeat to produce hierarchy concept. It is unsupervised technique because it doesn't use class label. It depends on the user specification for bin numbers.

Histogram: It is one of the unsupervised techniques that does not use class label. It distributes attributes values into ranges (buckets). The values are divided into equal ranges in equal width histogram while in equal frequency histogram, each part has the similar amount of data. The algorithm may be repeated reclusively to form multiple level hierarchies.

Entropy-based: It is one of the popular tools for discretization data. It is presented by Shannon during his study about information theory. It is top down and supervised technique. It uses class information to reduce the size of data. To discretize a numerical attribute X, it must choose the value of X with minimum entropy as a splitting point this step is repeated recursively to get hierarchical discretization [2].

Clustering: It one of a most method for data discretization. The algorithms of clustering can be used to discretize a numeric attributes X by dividing de values of attribute to groups or clusters. It produces high results of discretization. It can be classified into either top-down splitting or bottom-up merging strategy. In the first type, each cluster which forms anode can be further split into sub clusters, forming low level of hierarchy. While in the second type, clusters are produced by merging the neighbors' clusters to form high level concept [2].

Apriori algorithm: Usage of Apriori algorithm for data reduction. Apriori is designed to operate on databases containing transactions. Each transaction is seen as a set of items (an itemset). Given a threshold {\display style C}C, the Apriori algorithm identifies the item sets which are subsets of at least {\display style C}C transactions in the database.[3] Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length {\display style k} k from item sets of length {\display style k-1} k-1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent {\display style k} k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

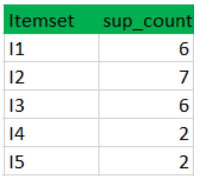


TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

Fg.3: Apriori algorithm example

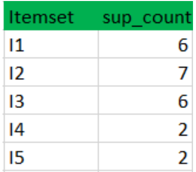
Step-1: K=1

(I) Create a table containing support count of each item present in dataset – Called C1(candidate set)



Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

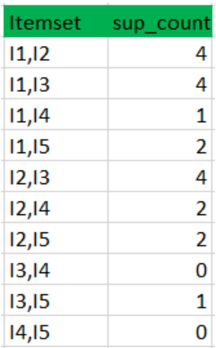
(II) compare candidate set item’s support count with minimum support count (here min_support=2 if support_count of candidate set items is less than min_support then remove those items). This gives us itemset L1.



Itemset	sup_count
I1	6
I2	7
I3	6
I4	2
I5	2

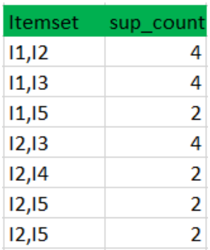
Step-2: K=2

- Generate candidate set C2 using L1 (this is called join step). Condition of joining Lk-1 and Lk-1 is that it should have (K-2) elements in common.
- Check all subsets of an itemset are frequent or not and if not frequent remove that itemset. (Example subset of {I1, I2} are {I1}, {I2} they are frequent. Check for each itemset)
- Now find support count of these itemset by searching in dataset.



Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

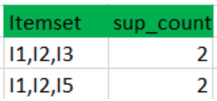
(II) compare candidate (C2) support count with minimum support count (here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L2.



Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

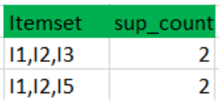
Step-3:

- Generate candidate set C3 using L2 (join step). Condition of joining Lk-1 and Lk-1 is that it should have (K-2) elements in common. So here, for L2, first element should match.
  So itemset generated by joining L2 is {I1, I2, I3} {I1, I2, I5} {I1, I3, i5} {I2, I3, I4} {I2, I4, I5} {I2, I3, I5}
- Check if all subsets of these itemset are frequent or not and if not, then remove that itemset. (Here subset of {I1, I2, I3} are {I1, I2}, {I2, I3}, {I1, I3} which are frequent. For {I2, I3, I4}, subset {I3, I4} is not frequent so remove it. Similarly check for every itemset)
- find support count of these remaining itemset by searching in dataset.



Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

(II) Compare candidate (C3) support count with minimum support count (here min_support=2 if support_count of candidate set item is less than min_support then remove those items) this gives us itemset L3.



Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

Step-4:

Generate candidate set C4 using L3 (join step). Condition of joining Lk-1 and Lk-1 (K=4) is that they should have (K-2) elements in common. So here, for L3, first 2 elements (items) should match.

Check all subsets of these itemset are frequent or not (Here itemset formed by joining L3 is {I1, I2, I3, I5} so its subset contains {I1, I3, I5}, which is not frequent). So, no itemset in C4

We stop here because no frequent itemset are found further.

Thus, we have discovered all the frequent item-sets.

VII. Conlusion

Data can be incomplete, inconsistent, and missing, data preprocessing s one of the important matters for data mining, data preprocessing includes data cleaning, data integration, data transformation and data reduction. Data cleaning method is used to remove the noisy data, completed on uncompleted data and remove unnecessary data. Data integration method is integrated to different source of data in one place. Data transformation method change forms of data and data reduction reduce the volume of database by schema integration. The presented Naïve Bayes classifier of data purification and the Apriori algorithm for data extraction are presented. Thus, the initial processing of data is of great importance for the preparation, analysis, and processing of the big data.

References

[1] А. А.Барсегян, М. С.Куприянов, В. В.Степаненко, И. И. Холод, *ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ: Data Mining, Visual Mining, Text Mining, OLAP* 2-е издание, Санкт-Петербург «БХВ-Петербург» 2007.

[2] Suad A. Alasadi and Wesam S. Bhaya, *Review of Data_Preprocessing Techniques in Data Mining*, Journal of Engineering and Applied Sciences, vol. 12, no. 16, pp. 4102-4107, 2017.

[3] Jiao Yabing, *Research of an Improved Apriori Algorithm in Data Mining Association Rules*, International Journal of Computer and Communication Engineering, vol. 2, no. 1, pp. 25-27, 2013.

[4] Evaristus Didik Madyatmadja, Mediana Aryuni, “Comparative Study Of Data Mining Model For Credit Card Application Scoring In Bank”, *Journal of Theoretical and Applied Information Technology*, Jakarta, vol.59, no. 2, pp. 269-274 2014.

[5] B. Mounika, V. Satyanarayana, “A Survey On Data Cleaning Techniques”, *International Journal of Engineering Computational Research and Technology*, Hyderabad, vol. 2, no. 1, pp. 501-510 2017.