Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 5, May-2016

Comparison of various data cleaning methos in mining

Kanu Patel[1], Priyank Bhojak[2], Vatsal Shah[3], Vikram Agrawal[4]

[1,2,3,4]*Assistant Professor, IT Department name, BVM Engineering College, Vallabh Vidhyanagar*

***Abstract* — Data mining is very wide area for research. Data quality is a main issue in quality information management. In this paper we focus on data cleaning methods, Data cleaning is one of the important aspect of data mining. Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining has various techniques that are suitable for data cleaning. In this paper we discuss three major data mining methods, In this paper we have to study various data cleaning techniques to use before implementation.**

Keywords- Data mining, Data cleaning, Functional dependency, Association rule, Binning method, smoothing.

I. INTRODUCTION

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Data cleaning is a process used to determine inaccurate, incomplete or unreasonable **data** and then improve the quality through correcting of detected errors and omissions. Generally **data cleaning** reduces errors and improves the **data** quality.

Major Tasks in Data Preprocessing ,,

Data cleaning: Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration: Integration of multiple databases, data cubes, or files ,,

Data transformation: Normalization and aggregation ,,

Data reduction: Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization: Part of data reduction but with particular importance, especially for numerical data

Why preprocessing?

1. Real world data are generally
   - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
   - Noisy: containing errors or outliers
   - Inconsistent: containing discrepancies in codes or names
2. Tasks in data preprocessing
   - Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
   - Data integration: using multiple databases, data cubes, or files.
   - Data transformation: normalization and aggregation.
   - Data reduction: reducing the volume but producing the same or similar analytical results.
   - Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

II. DATA CLEANING

Data pre-processing is an often neglected but important step in the data mining process. The phrase "Garbage In, Garbage Out" is particularly applicable to data mining and machine learning. Data gathering methods are often loosely controlled, resulting in out-of- range values (e.g., Income: -100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

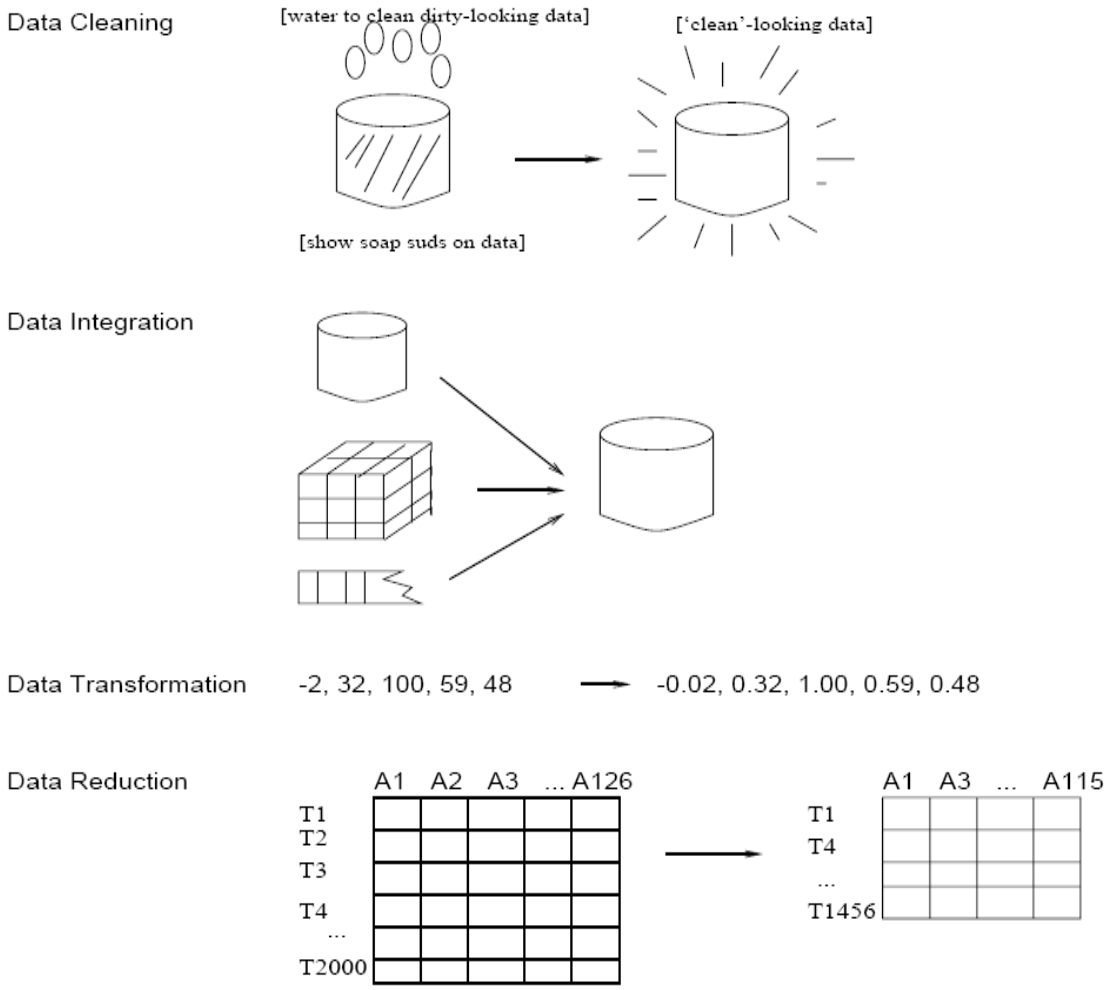

Fig. 1 Data Processing types

- Data auditing: The data is audited with the use of statistical and database methods to detect anomalies and contradictions: this eventually gives an indication of the characteristics of the anomalies and their locations. Several commercial software packages will let you specify constraints of various kinds (using a grammar that conforms to that of a standard programming language, e.g., JavaScript or Visual Basic) and then generate code that checks the data for violation of these constraints. This process is referred to below in the bullets "workflow specification" and "workflow execution." For users who lack access to high-end cleansing software, Microcomputer database packages such as Microsoft Access or File Maker Pro will also let you perform such checks, on a constraint-by-constraint basis, interactively with little or no programming required in many cases.
- Workflow specification: The detection and removal of anomalies is performed by a sequence of operations on the data known as the workflow. It is specified after the process of auditing the data and is crucial in achieving the end product of high-quality data. In order to achieve a proper workflow, the causes of the anomalies and errors in the data have to be closely considered.
- Workflow execution: In this stage, the workflow is executed after its specification is complete and its correctness is verified. The implementation of the workflow should be efficient, even on large sets of data, which inevitably poses a trade-off because the execution of a data-cleansing operation can be computationally expensive.
- Post-processing and controlling: After executing the cleansing workflow, the results are inspected to verify correctness. Data that could not be corrected during execution of the workflow is manually corrected, if possible. The result is a new cycle in the data-cleansing process where the data is audited again to allow the specification of an additional workflow to further cleanse the data by automatic processing.

Data cleaning types

1. Fill in missing values (attribute or class value):
    - Ignore the tuple: usually done when class label is missing.
    - Use the attribute mean (or majority nominal value) to fill in the missing value.

- Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
- Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.

2. Identify outliers and smooth out noisy data:
   - Binning
     - Sort the attribute values and partition them into bins (see "Unsupervised discretization" below);
     - Then smooth by bin means, bin median, or bin boundaries.
   - Clustering: group values in clusters and then detect and remove outliers (automatic or manual)
   - Regression: smooth by fitting the data into regression functions.
3. Correct inconsistent data: use domain knowledge or expert decision.

III. FIRST-ORDER HEADINGS

Missing Values:

Imagine that you need to analyze *AllElectronics* sales and customer data. You note that many tuples have no recorded value for several attributes, such as customer *income*.How can you go about filling in the missing values for this attribute? Let's look at the following methods:

1. Ignore the tuple: This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
2. Fill in the missing value manually: In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.
3. Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like *"Unknown"* or □¥. If missing values are replaced by, say, *"Unknown,"* then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of *"Unknown."* Hence, although this method is simple, it is not foolproof.
4. Use the attribute mean to fill in the missing value: For example, suppose that the average income of *AllElectronics* customers is $56,000. Use this value to replace the missing value for *income*.
5. Use the attribute mean for all samples belonging to the same class as the given tuple: For example, if classifying customers according to *credit risk*, replace the missing value with the average *income* value for customers in the same credit risk category as that of the given tuple.
6. Use the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for *income*. Decision trees, regression,

Value may be missing because it is unrecorded or because it is inapplicable, In medical data, value for Pregnant? Attribute for Jane or Anna is missing, while for Joe should be considered Not applicable. Some programs can infer missing values

Ignore / delete the instance: (not effective when the percentage of missing values per attribute varies considerably).

Fill in the missing value manually: expert based + infeasible?

Name	Age	Sex	Pregnant	..
Mary	25	F	N	
Jane	27	F	?	
Joe	30	M	-	
Anna	2	F	?	

Fig 2. Missing data table

Noisy Data

"What is noise?" Noise is a random error or variance in a measured variable. Given a numerical attribute such as, say, *price*, how can we "smooth" out the data to remove the noise? Let's look at the following data smoothing techniques
Outliers – graphical identification

1. Binning: Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or *bins*. Because binning methods consult the neighborhood of values, they perform*local* smoothing. illustrates some binning techniques. In this example, the data for *price* are first sorted and then partitioned into *equal-frequency* bins of size 3 (i.e., each bin contains three values). In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the *bin boundaries*. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be *equal-width*, where the interval range of values in each bin is constant. Binning is also used as a discretization technique.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

2. Regression: Data can be smoothed by fitting the data to a function, such as with regression. *Linear regression* involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other. *Multiple linear regression* is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Use simple statistics and graph tools - Statistica

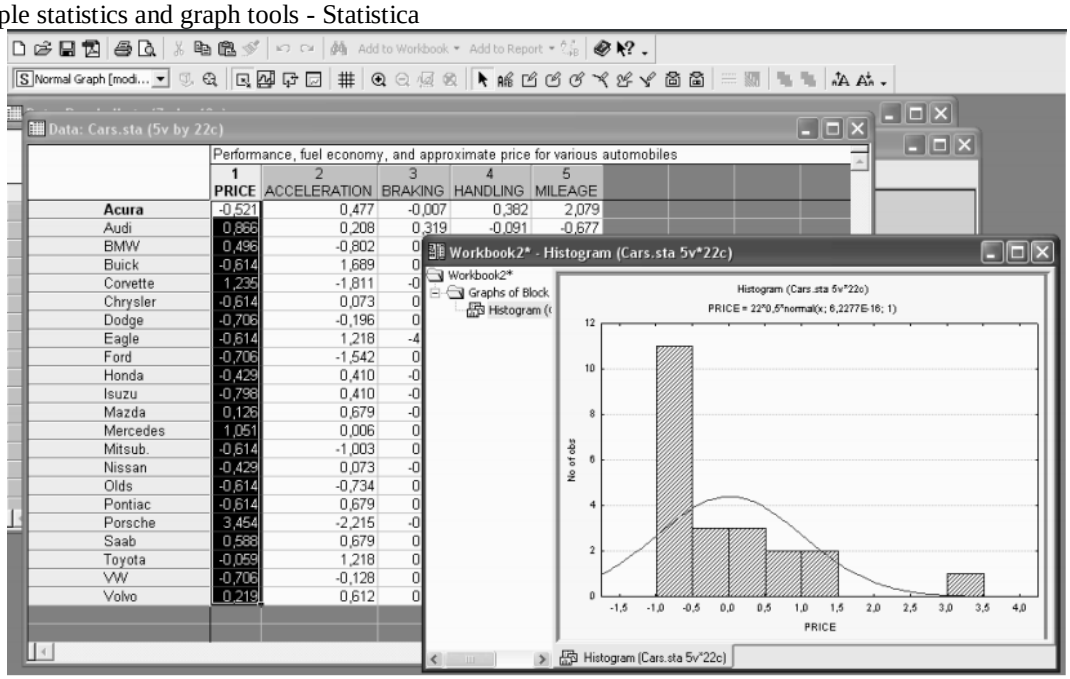

Fig 3 Use simple statistics and graph tools - Statistical

Regression – outliers

An example of „corn flakes" [Larose 08] – 2 points could be outliers for a linear regression → standardized residuals

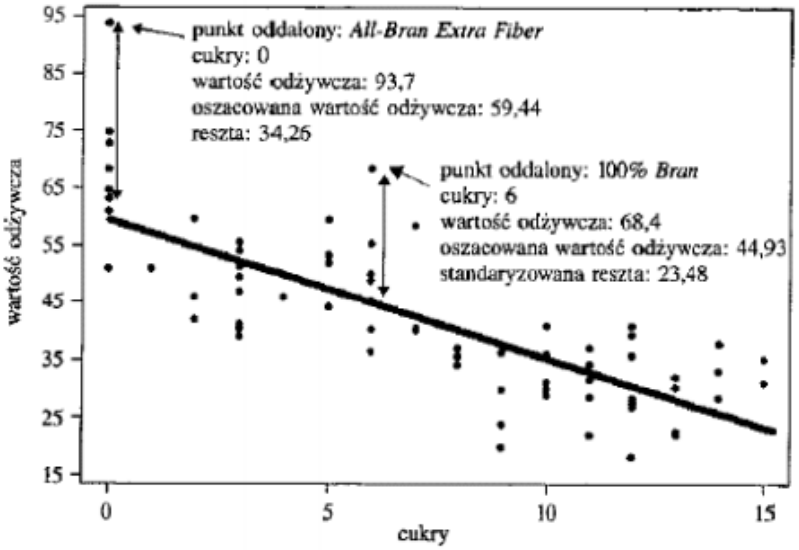

Fig 3. Regression data plotting[4]

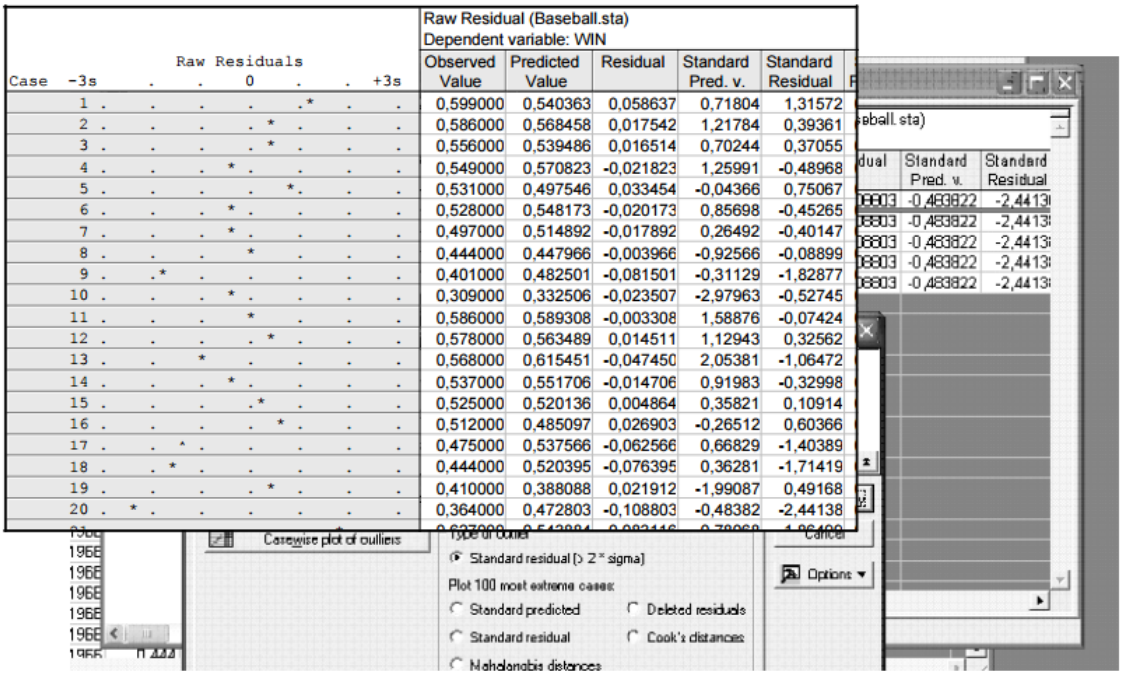

Fig 4 Data record

3. Clustering: Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers (Figure 2.12). Chapter 7 is dedicated to the topic of clustering and outlier analysis.

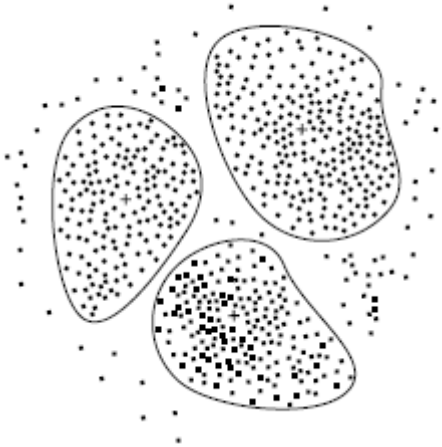

Fig 5 Clustering[6]

IV. CONCLUSION

Data cleaning is very necessary part of data mining. From the above study we can see that there are different types of problems in data cleaning .Data cleaning methods and approaches depend upon the type of data which we want to clean and according to that we apply particular methods. This paper also presents a comparison of data cleaning methods and determines the best one. Each method has its own specific features and depending upon the data we can use it to clean data.

REFERENCES

[1]. Lee, M.L.; Lu, H.; Ling, T.W.; Ko, Y.T.: Cleansing Data for Mining and Warehousing. Proc. 10th Intl. Conf. Database and Expert Systems Applications (DEXA), 1999.
[2]. Quass, D.: A Framework for Research in Data Cleaning. Unpublished Manuscript. Brigham Young Univ., 1999
[3]. Hernandez, M.A.; Stolfo, S.J.: Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data Mining and Knowledge Discovery 2(1):9-37, 1998.
[4]. http://www.ukessays.co.uk accessed on 7-4-2016 at 4:18pm
[5]. Müller Heiko & Christoph Freytag Johann , Problems, Methods, and Challenges in Comprehensive Data Cleansing ,Humboldt-Universität zu Berlin zu Berlin,10099 Berlin, Germany.
[6]. Han & Kembaer, Data mining & Data warehousing ,2nd Edition Book.