International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Research Article

A Comparative Study between Noisy Data and Outlier Data in Data Mining

L. Sunitha[a*],M. Bal Raju[a] and B.Sunil Srinivas[a]

[a]Department of computer Science and Engineering ,Vidya Vikas Institute of Technology, Hyderabad, India

Accepted 27 May 2013, Available online 1June 2013, **Vol.3, No.2 (June 2013)**

Abstract

In data mining pre processing means preparing data .It is the one of the important and compulsory task. Before applying the data mining techniques like association, classification or clustering noisy and outliers should be removed. In this paper we are trying to find similarities and differences between noisy data and outliers .Actually most of the data mining users are thing that these two are same but lot of differences are there.

Key Terms: *Noisy, Outlier, Pre Processing*

1. Introduction

Data preparation or pre processing is an important issue for both data warehouse and dada mining.

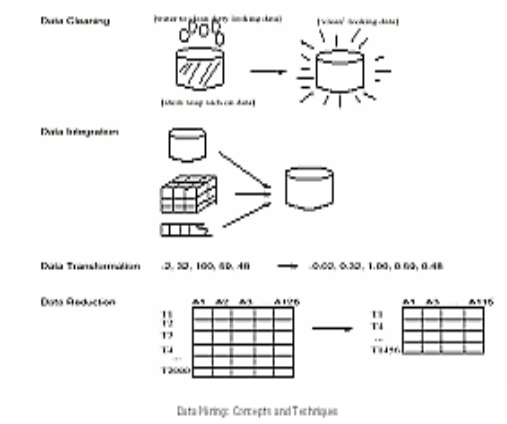

Fig 1. Data preprocessing forms

Cleaning and integration are two forms used data warehouse and selection and transformation these are used before applying data mining algorithm Noisy data is meaningless data. The term has often been used as a synonym for corrupt data. However, its meaning has expanded to include any data that cannot be understood and interpreted correctly by machines, such as unstructured text. Any data that has been received, stored, or changed in such a manner that it cannot be read or used by the program that originally created it can be described as noisy.

1.1 Problems with Noisy Data

1. Noisy data unnecessarily increases the amount of storage space required and can also adversely affect the results of any data mining analysis. Statistical analysis can use information gleaned from historical data to weed out noisy data and facilitate data mining.
2. Noisy data can be caused by hardware failures, programming errors and gibberish input from speech or optical character recognition programs.
3. Spelling errors, industry abbreviations and slang can also delay in machine reading.

Examples for Noisy

By 'noisy data' we mean errors scattered in the data. The term noisy means corrupted data
For example, due to inaccurate recording, data corruption. Some noise will be very common:
Data has incorrect type: (string in numeric attribute)
Data does not match enumeration (maybe in yes/no field)
Data is very dissimilar to all other entries (10 in an attr otherwise 0..1)
Some noisy will be rare typing 0.52 at data entry instead of 0.25.

Some possible solutions for Noisy

- Manual inspection and removal
- Use clustering on the data to find instances or attributes that lie outside the main body (outliers) and remove them
- Use regression to determine function, and then remove those that lie far from the predicted value
- Ignore all values that occur below a certain frequency threshold
- Apply smoothing function over known-to-be-noisy data.

*Corresponding author: L.Sunitha

- If noise is removed, can apply missing value techniques on it. If it is not removed, it may adversely affect the accuracy of the model.

1.2 Inconsistence

Some values may not be recorded in different ways. For example 'coke', 'coca cola', 'coca-cola', 'Coca Cola' etc .In this case the data should be normalised to a single form. Can be treated as a special case of noise. Some values may be recorded inaccurately on purpose! Email address: r.d.nospam.sanderson@...Spike in early census data for births on 11/11/1911. Had to put in some value, so defaulted to 1s everywhere.

Data Mining Just because the base data includes an attribute doesn't make it worth giving to the data mining task. For example, denormalise a typical commercial database and you might have:

ProductId, ProductName, ProductPrice, SupplierId, SupplierAddress. SupplierAddress is dependant on SupplierId so they will always appear together. A 100% confident, 100% support association rule is not very interesting!

1.3 Redundant Values

In Data Mining putting the redundant values are harmful for Association rule mining and for other data mining tasks too. Can treat text as thousands of numeric attribute frequency from our inverted indexes? But not all of those terms are useful for determining (for example) if an email is spam. 'the' does not contribute to spam detection. Can treat text as thousands of numeric attributes: term/frequency from our inverted indexes. But not all of those terms are useful for determining (for example) if an email is spam. 'The' does not contribute to Spam detection. The number of attributes in the table will affect the time it takes the data mining process to run. It is often the case that we want to run it many times, so getting rid of Unnecessary attributes is important.

1.4 Number of Attributes

Data Mining Called dimensionality reduction we'll look at techniques for this, but some simplistic versions:

- Apply upper and lower thresholds of frequency
- Noise removal functions
- Remove redundant attributes
- Remove attributes below a threshold of contribution to classification

2. Outlier

The handling of anomalous or outlying observations in a data set is one of the most important The handling of anomalous or outlying observations in a data set is one of the most important tasks in data pre-processing tasks.

In data pre-processing outliers are extreme from normal Handling of outlier is important for three reasons.

- Outlying observations can have a considerable influence on the results of an analysis.
- Outliers are often measurement or recording errors, some of them can represent phenomena of interest, something significant from the viewpoint of the application domain.
- For many applications, exceptions identified can often lead to the discovery of unexpected knowledge.

3. Approaches to outlier management

There are two principal approaches to outlier management

(i) Outlier accommodation

Which is characterized by the development of a variety of statistical estimation or testing procedures which are robust against, or relatively unaffected by, outliers. In these procedures, the analysis of the main body of data is the key objective and outliers themselves are not of prime concern. This approach is difficult to be applied to those applications where explicit identification of anomalous observations is an important consideration, e.g. suspicious credit card transactions.

(ii) Retained or rejected

This approach is characterized by identifying outliers and deciding whether they should be retained or rejected. Many statistical techniques have been proposed to detect outliers and comprehensive texts on this topic are those by Hawkins and Barnet and Lewis. These approaches range from informal methods such as the ordering of multivariate data, the use of graphical and pictorial methods, and the application of simple test statistics, to some more formal approach in which a model for the data is provided, and tests of hypotheses that certain observations are outliers are set up against the alternative that they are part of the main body of data.

The identification of outliers has also received much attention from the computing community .However; there appear to be much less work on how to decide whether outliers should be retained or rejected. In statistical community, a commonly-adopted strategy when analyzing data is to carry out the analysis both including and excluding the suspicious values. If there is little difference in the results obtained then the outliers had minimal effect, but if excluding them does have an effect it may be better to find an alternative. This is where knowledge-based outlier analysis steps in. In order to successfully distinguish between noisy outlying data and noise free outliers, different kinds of information are normally needed. These should not only include various data characteristics and the context in which the outliers occur, but also relevant domain knowledge.

The procedure for analyzing outliers has been experimentally shown to be subjective, depending on the above mentioned factors. The analyst is normally given this task of judging which suspicious values are obviously

silly, impossible and which, while physically possible, should be viewed with caution. However in the context of data mining where a large number of cases are normally involved, the number of suspicious cases would be sizable too and manual analysis would become insufficient.

4. Conclusion

Most of the users of data mining can think that noisy data and outlier data are same both should be removed, actually here we are try to find the dissimilarities between noisy and outlier ,noisy is removed in pre processing where as outliers may or may not removed depending the data mining algorithm. Noisy data does not have any applications where as outliers may be observed in clustering technique as a by product .Outliers having different views depending on application and method, identified outliers are not leaved they should be analyzed, but noisy data is simply it is removed it does not have any applications.

References

http://iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf

n.wikipedia.org/wiki/Data_mining\

ww.itl.nist.gov/div898/handbook/eda/section3/eda35h.htDetection of Outliers,

http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

D. Hawkins (1988), Identification of outliers. *Chapman and Hall London*

S. Ramaswamy, R. Rastogi, and K. Shim (2000), Efficient algorithms fo mining outliers from large data sets, *ACM SIGMOD Record*, vol. 29,no. 2, pp. 427–438.

A. Ghosting, S. Parthasarathy, and M. Otey (2008), Fast Mining of Distance based Outliers in High-dimensional Datasets, *Data Mining and Knowledge Discovery*, vol. 16, no. 3, pp. 349–364.

F. Anguilla and F. Fassetti (2007), Very efficient mining of distance-based Outliers, *Proc. of the sixteenth ACM conference on Conference on Information and knowledge management*, pp. 791–800.