Analysis on Data cleaning and Transformation techniques

N.Marudachalam,
Reaseach scholar in Sathyabama University P.G & Research Department, Tamilnadu, India.
kiranmarudachalam@gmail.com

M.Ramakrishnan,
Professor and Chairperson, School of Information Technology,
Madurai Kamaraj University.Tamilnadu, India
ramakrishod@gmail.com

Abstract: In this modern world many companies produce raw data with missing values, In attributes and tuples we also find more errors. This type of data transformed with missing information stored are in repository, the data warehouse. Here large volume of data available for more number of years. Raw data from different sources are stored in the repository. The resources are streamlined and distributed to many companies like trading/business and IT vendors. The management uses it for decision making or decision support system for their company development. Before usage all management should clean the data in the dataware house, like correcting errors, removing duplicate data and avoiding unambiguous un wanted elements and etc.,

Keywords-Attributes,tuples, repository, uambiguous

I. Introduction

Data mining consist of several stages. In these stages data cleaning is a prilimary stage, after refinement the main result or main goal is achieved. This result is found hidden from external sources under the bottom layer. In which the the the knowledge is been discovered . This process is termed as knowledge discovery.

Through Requirements we are able to gain knowledge in the world. Here demand and requirements play key role in the market. Market basket Analysis solves the problem

Importance of Data Cleaning:

- Error free Repository storage
- Avoids Redundant data
- Unambiguous data are removed
- Customer satisfaction and user friendly
- Saves time and storage
- Error once cleared will be used for many time and any time
- Content is preserved
- Helps for decision making and decision support system
- Concentrates on quality assurance and data productivity.
- Provides Reliable data

II. Methods of Data cleaning

A.Hybrid Approach to Data Cleaning Method:

Here hybrid refers combination of HADCLEAN, PNR and its own features.

B.PNRS Algorithm/Modified/ PNRS/Original PNRS/Improved PNRS

The PNRS algorithm, coined by C. Varol et al. corrects the phonetic and typographical errors of current raw data

with standard dictionaries. It mainly employs two algorithms which are explained below by

- Inserting a blank space
- Swapping two letters
- Updating/ a letter

If it is a correct word then word is stored temporarily otherwise wil be ignored.

Phonetic Algorithm - Phonetic Algorithm checks the words through the sounds of the alphabets. Here we spell the word according to the suggestion given by the computer. If it matches with the already existing syllable it will say whether right or wrong. It takes the suggestion if it is right oitherwise word near by will be taken.

DOI: 10.21817/ijet/2017/v9i5/170905320

C. Transitive Closure Algorithm

Transitive Closure algorithm for data cleaning has been proposed by W.N. Li, et al [11]. This algorithm preprocesses the data to categorize millions and billions of records into groups of related data.The ETL tool (Extraction. Tansformation & Loading) using following algorithm processes the individual groups for data cleaning which involves Identifying and removal of redundancies - This is especially valuable when we are migrating data from different source systems where we might store same data in different formats Filling the blank cells.

Establishment of "group" relationship between different records leading to faster querying .In this technique the records are matched on the basis of matching of the keys (keys are selected attributes of the data). Each key is matched one after the other, so as to obtain related group of records. These groups can further be analyzed and corrected. Blanks can be filled, and redundancies can be removed. Next section explains the proposed data cleaning algorithm along with a case study.

Modified version of Transitive Closure algorithm

Transitive Closure algorithm matches two or more records into one group when one of the key (attribute) matches between the two records. But, sometimes by matching only one key to group records would result in mistakes as we cannot rely only on matching one key. Also this runs completely in semi automatic mode, where the records are grouped together and then leaves room for manual intervention to study these groups and then declare that duplication or correction of data in the records.

D. Hybrid Approach

A hybrid algorithm called HADCLEAN is being proposed in this paper that includes usage of modified versions of PNRS and Transitive Closure algorithms.

E. Robust Cleaning

It gives the safe cleaning without eliminate or damage the data

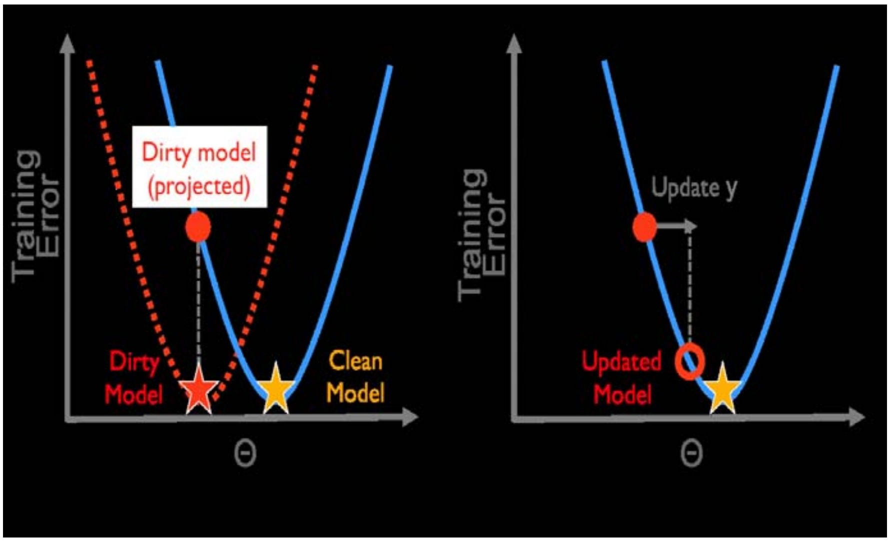
F. Predictive Modeling

The user provides a relation R and wishes to train a model using the data in R softare. This work focuses on a class of well analyzed predictive analytics problems; ones that can be expressed as the minimization of convex loss functions. Convex loss minimization problems are amenable to a variety of incremental optimization methodologies with provable guarantees (see Friedman, Hastie, and Tibshirani [15] for an introduction).

Examples include generalized linear models (including linear and logistic regression), support vector machines, and in fact, means and medians are also special cases. We assume that the user provides a featurizer F(_) that maps every record r 2 R to a feature vector x and label y. For labeled training examples f(xi; yi)gNi=1, the problem is to find a vector of *model parameters* _ by minimizing a loss function _ over all training examples:

G.Genetic Derivation Algorithm

The update algorithm intuitively follows from the convex geometry of the problem. Consider the problem in one dimension (i.e., the parameter _ is a scalar value), so then the goal is to find the minimum point (_) of a curve l(_). The consequence of dirty data is that the wrong loss function is optimized. Figure A illustrates the consequence of the optimization. The red dotted line shows the loss function on the dirty data. Optimizing the loss function finds _(d) that at the minimum point (red star). However, the true loss function (w.r.t to the clean data) is in blue, thus the optimal value on the dirty data is in fact a suboptimal point on clean curve (red circle).



increasing the batch size O(p1b). In the experiments, we use a batch size of 50 which converges fast but allows for frequent model updates. If a data cleaning technique requires a larger batch size than 50, i.e., data cleaning is fast enough that the iteration overhead is significant compared to cleaning 50 records, ActiveClean can apply the updates in smaller batches. For example, the batch size set by the user might be b = 1000, but the model updates after every 50 records are cleaned. We can disassociate the batching requirements of SGD and the batching requirements of the data cleaning technique.

H. Compared Algorithm

Here are the alternative methodologies evaluated in the experiments: **Robust Logistic Regression .** Feng et al. proposed a variant of logistic regression that is robust to outliers. We chose this algorithm because it is a robust nsion of the convex regularized loss model, leading to a better apples-toapples

comparison between the techniques.

Discarding Dirty Data. As a baseline, dirty data are discarded. **SampleClean (SC) .** SampleClean takes a sample of data, applies data cleaning, and then trains a model to completion on the sample.

Active Learning (AL) [18]. To fairly evaluate Active Learning, we first apply our gradient update to ensure correctness. Within each iteration, examples are prioritized by distance to the decision boundary (called Uncertainty Sampling in [31]). However, we do not include our optimizations such as detection and estimation.

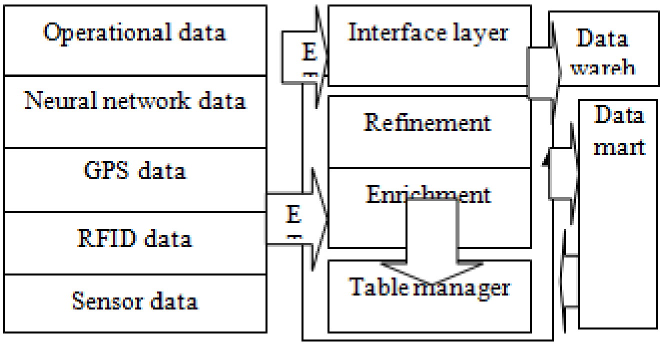
ActiveClean Oracle (AC+O): In ActiveClean Oracle, instead of an estimation and detection step, the true clean value is used to evaluate the theoretical ideal performance of Active- Clean.

L. Traditional Statistical Algorithm

This algorithm Tradtional statical Alogorithm deals with ineference, standard deviation, mean, median and mode. And improved cases of the above said.

II.Data Transformation:

The data are transformed or consolidated into forms suitable for data mining. This can involves smoothing,Aggregation,Generalization of data ,Normalization and attribute construction.



Operational Data:

Here, we will get traditional ETL(Extraction, Transform and loading)approach the most current information available from Internet of things(IoT). such as sensor data, RFID data(Radio frequency identification data, Neural network data, Big data Analytics and GPS(Global positioning system sources. These are reffered as operational data for data warehouse environment.

ETL(Extraction,Transformation.

Loading):

Operational data goes to interface layer after the ETLProcess .Here extraction is fetching the suitable data and convert it to most suitable form for working process for another operation .This operation known as transformatation the original data in this stage translated into abstract form and load to the data ware house for non-volatile storage.

Interface Layer :

Here there are three segment namely Refinement, Enrichment and table manager.

Refinement:

Refinement is the data under value Enggineering to remove unwanted things.

Enrichment:

Here the data polished(improved refinement) for data mining .Now we apply hybrid dimensional online analytical processing for report analysis to the final action(OLAP,MOLAP).

Table Manager:

It is indexed form of refinement and enrichment data .

III. Data Warehouse:

Data warehouse plays vital role in KDD (Knowledge Discovery in Data bases)process with the following properties.

i)Time dependent- Long period of time,ex:5-15 years data.

ii)Non-volatile-ex: WORM

iii)Subject oriented-Data design specifically for same pool.

iv)Integrated-Variety of information gathered.

Data Mining:

It is KDD process through the following properties. information requirement, data selection, data cleaning and transformation,Data enrichment ,coding and data mining techniques(clustering-shared density based clustering ,micro clustering and DBSCAN clustering) segmentation,Prediction finally gives decision support system or report to take further action.

Conclusion & Future Work

Preprocessing of the data mining in the data warehousing system we clean the data with WEKA tool. We transform the preprocessed data into another form like Normalization, Discretization, Generalization and smoothing techniques.

We have to plan to analyses data cleaning and transformation techniques in Artificial Intelligence

IV. REFERENCES

[1] Arindam Paul, Varuni Ganesan, Jagat Sesh Challa, Yashvardhan Sharma” A Novel Approach for Data Cleaning by Selecting the Optimal Data to Fill the Missing Values for Maintaining Reliable Data Warehouse” https://www.researchgate.net/publication/301543768 Mar-2012

[2] Sanjay Krishnan, Jiannan Wang, Eugene Wu y, Michael J. Franklin, Ken Goldberg” ActiveClean: Interactive Data Cleaning While Learning Convex Loss Models” Columbia University,Jan-2016

[3] Dollars for docs. http://projects.propublica.org/open-payments/.

[4] For big-data scientists, ’janitor work’ is key hurdle to insights. http://www.nytimes.com/2014/08/18/technology/for-big-datascientists-hurdle-to-insights-is-janitor-work.html.

[5] A pharma payment a day keeps docs’ finances okay.https://www.propublica.org/article/ a-pharma-payment-a-day-keeps-docs-finances ok.

[6] A. Alexandrov, R. Bergmann, S. Ewen, J. Freytag, F. Hueske, A. Heise, O. Kao, M. Leich, U. Leser, V. Markl, F. Naumann,