International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Review of Big Data Pre-processing

V. Maria Antoniate Martin*[1], Dr. K. David[2], N. Bala Sankar [3]

[1]Research Scholar, Department of Computer Science, Research and Development Centre, Bharathiar University, Coimbatore, Tamil Nadu, India

[2]Assistant Professor, Department of Computer Science, The Rajah's College, Pudukkottai, Tamil Nadu, India

[3]Student, Department of Information Technology, St. Joseph's College, Trichy, Tamil Nadu, India

ABSTRACT

The massive growth in the scale of data has been observed in recent years being a key factor of the Big Data scenario. Big Data can be defined as high volume, velocity and variety of data that require a new high-performance processing. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The presence of data pre-processing methods for data mining in big data is reviewed in this paper. The definition, characteristics, and categorization of data pre-processing approaches in big data are introduced. The connection between big data and data pre-processing throughout all families of methods and big data technologies are also examined, including a review of the state-of-the-art. In addition, research challenges are discussed, with focus on developments on different big data framework, such as Hadoop, Spark and Flink and the encouragement in devoting substantial research efforts in some families of data pre-processing methods and applications on new big data learning paradigms.

Keywords: Big Data, Pre-processing, Data Quality

I. INTRODUCTION

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.Datapreprocessing is used database-driven applications such as customer relationship management and rule-based applications[1].

Data Pre-processing

- ✓ Data cleaning
- ✓ Data integration and transformation
- ✓ Data reduction
- ✓ Discretization

II. LITERATURE REVIEW

Major Tasks in Data Pre-process

In this section, we look at the major steps involved in data pre-processing, namely, data cleaning, data integration, data reduction, and data transformation.

Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. If users believe the data are dirty, they are unlikely to trust the result so any data mining that has been applied. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output. Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust.

Instead, they may concentrate on avoiding over fitting the data to the function being modelled. Therefore, a useful pre-processing step is to run your data through some data cleaning routines. Section 3.2 discusses methods for data cleaning.

"The data set I have selected for analysis is HUGE, which is sure to slow down the mining process. Is there a way I can reduce the size of my data set without jeopardizing the data mining results?"

Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. Data reduction strategies include dimensionality reduction and numerosity reduction.

In **dimensionality reduction**, data encoding schemes are applied so as to obtain a reduced or "compressed" representation of the original data. Examples include data compression techniques (e.g., wavelet transforms and principal components analysis), attribute subset selection (e.g.,removing irrelevant attributes), and attribute construction (e.g., where a small set of more useful attributes is derived from the original set).

In numerosity reduction, the data are replaced by alternative, smaller representations using parametric models (e.g., regression or log-linear models) or nonparametric models (e.g., histograms, clusters, sampling, or **data aggregation**). Data reduction is the topic of Section.

Discretization and concept hierarchy generation are powerful tools for data mining in that they allow data mining at multiple abstraction levels. Normalization, data discretization, and concept hierarchy generation are forms of data transformation. You soon realize such data transformation operations are additional data pre-processing procedures that would contribute toward the success of the mining process. **Data integration** and data discretization are discussed in Sections[2].

A. Data pre-processing in new data mining fields

Many data pre-processing methods have been devised to work with supervised data, since the label provides useful information that facilitates data transformation. However, there are also pre-processing approaches for unsupervised problems. For instance, FS has attracted much attention lately for unsupervised problems[3].

B. Big Data and the MapReduce Programming Model

"Big data" is a term used to describe huge amounts of data so large and complex that cannot be processed by traditional tools and techniques in an easy way. Initially, Douglas Laney's Gartner analyst defined this concept as a three Vs model (Volume, Velocity and Variety), where "Volume" refers to the vast amounts of data that needs to be processed and analysed in order to obtain valuable information, "Velocity" states that the data must be processed in an acceptable response time, and finally, "Variety" means that the data can be presented in different formats. Later, additional Vs have been introduced to expand the description of the "big data" term, and some of these characteristics are Variability, Veracity, Volatility, Validity or Value[4].

C. Classification with Imbalanced Datasets

Many real-world problems usually have a distribution of classes where one or more classes are represented by a large number of examples in contrast to the negligible number of examples of other classes. This is known as the problem of classification with imbalanced data and it is present in different domains such as finances, bioinformatics or medical applications. In these problems, the main concern is the correct identification of the underrepresented classes since they are the focus of interest. The traditional classification algorithms are often unable to address imbalanced datasets as they are built under the assumption of obtaining a greater generalization ability. For this reason, these algorithms try to get general rules that cover most of the examples, benefiting the most represented classes

and trying as noise the underrepresented classes[5][6].

D. Literature survey

In authors present a HACE (Heterogeneous, Autonomous, Complex and Evolving) theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. They analyze the challenging issues in the data-driven model and also in the Big Data revolution.

In authors present the KEOPS data mining methodology centered on domain knowledge integration. In this paper, the authors focuses first on the pre-processing steps of business understanding and data understanding in order to build an ontology driven information system (ODIS). Then they show how the knowledge base is used for the post-processing step of model interpretation[7].

E. Imbalanced data

Classification problems are typically formed by a small set of classes. Some of them come with a tiny percentage of instances compared with the other classes. These highly imbalanced problems are more noteworthy in Big Data environments where millions of instances are present. Some contributions to this topic have been implemented on Hadoop Map Reduce:

- The first approach in dealing with imbalanced large-scale datasets was proposed by Park et al. In this work, a simple over-sampling technique was employed using Apache Hadoop and Hive on traffic data with a 14 % of positive instances.
- Hu et al. proposed an enhanced version of Synthetic Minority Over-sampling Technique (SMOTE) algorithm on MapReduce. This method focused on replicating those minority cases that only belong to the boundary region to solve the problem of original SMOTE, which omits the distribution of the original data while yields new samples[8].

F. Instance reduction

Instance selection is a type of preprocessing technique, which aims at reducing the number of samples to be considered in the learning phase. In spite of its promising results with small and medium datasets, this task is normally undermined when coping with large-scale datasets (from tens of thousands of instances onwards)[9].

III. DIMENSIONALITY REDUCTION

When data sets become large in the number of predictor variables or the number of instances, data mining algorithms face the *curse of dimensionality* problem . It is a serious problem as it will impede the operation of most data mining algorithms as the computational cost rise. This section will underline the most influential dimensionality reduction algorithms according to the division established into Feature Selection (FS) and space transformation based methods[10].

A. Feature selection

Feature selection (FS) is “the process of identifying and removing as much irrelevant and redundant information as possible” . The goal is to obtain a subset of features from the original problem that still appropriately describe it. This subset is commonly used to train a learner, with added benefits reported in the specialized literature FS can remove irrelevant and redundant features which may induce accidental correlations in learning algorithms, diminishing their generalization abilities. The use of FS is also known to decrease the risk of over-fitting in the algorithms used later. FS will also reduce the search space determined by the features, thus making the learning process faster and less memory consuming.

The use FS can also help in task not directly related to the data-mining algorithm applied to the data. FS can be used in the data collection stage, saving cost in time, sampling, sensing and personnel used to gather the data. Models and visualizations made from data with fewer features will be easier to understand and to interpret [11][12][13].

B. Space transformations

FS is not the only way to cope with the curse of dimensionality by reducing the number of dimensions. Instead of selecting the most promising features, space transformation techniques generate a whole new set of features by combining the original ones. Such a combination can be made obeying different criteria. . The first approaches were based on linear methods, as factor analysis and PCA .

More recent techniques try to exploit nonlinear relations among the variables. Some of the most important, in both relevance and usage, space transformation procedures are LLE, ISOMAP and derivatives. They focus on transforming the original set of variables into a smaller number of projections, sometimes taking into account the geometrical properties of clusters of instances or patches of the underlying manifolds [14].

IV. CONCLUSION

At the present, the size, variety and velocity of data is huge and continues to increase every day. The use of Big Data frameworks to store, process, and analyse data has changed the context of the knowledge discovery from data, especially the processes of data mining and data pre-processing. In this paper, we presented a review on the rise of data pre-processing. Here presented an updated categorization of data pre-processing contributions under the big data framework. The review covered different families of data pre-processing techniques, such as feature selection, imperfect data, imbalanced learning and instance reduction as well as the maximum size supported and the frameworks in which they have been developed. The key issues in big data pre-processing were highlighted.

V. REFERENCES

[1] Aggarwal CC. Data Mining: The Textbook. Berlin, Germany: Springer; 2015

[2] .Pyle D. Data Preparation for Data Mining. San Francisco: Morgan Kaufmann Publishers Inc.; 1999.

[3] Li Z, Tang J. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. IEEE Trans Image Process. 2015; 24(12):5343–355.

[4] A. Fern´andez, S. R´ıo, V. L´opez, A. Bawakid, M. J. del Jesus, J. M. Ben´ıtez and F. Herrera, “Big Data with Cloud Computing: An Insight on the Computing Environment, MapReduce and Programming Frame- works,”WIREs Data Mining and Knowledge Discovery, vol. 4, no. 5, pp. 380–409, 2014.

[5] H. He and E. A. Garc´ıa, “Learning from imbalanced data,” IEEE Transac-tionsonKnowledgeandDataEngineering,vol.21,no .9,pp.1263–1284, 2009.

[6] V. L´opez, A. Fern´andez, S. Garc´ıa, V. Palade and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” Information Sciences, vol. 250, pp. 113–141, 2013.

[7] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, “Data Mining with Big Data”, (In Press) IEEE Transactions on Knowledge and Data Engineering, 2013.

[8] Park SH, Ha YG. Large imbalance data classification based on mapreduce for traffic accident prediction. In: 8th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS). Birmingham: 2014.

[9] Triguero I, Peralta D, Bacardit J, García S, Herrera F. MRPR: A mapreduce solution for prototype reduction in big data classification. Neurocomputing. 2015; 150 Part A: 331–45.

[10] Bellman RE. Adaptive Control Processes - A Guided Tour. Princeton, NJ: Princeton University Press; 1961.

[11] Hu F, Li H, Lou H, Dai J. A parallel oversampling algorithm based on NRSBoundary-SMOTE. J InfComput Sci. 2014; 11(13):4655–665

[12] .Hall MA. Correlation-based feature selection for machine learning.Waikato University, Department of Computer Science. 1999.

[13] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003; 3:1157–82.

[14] .Chandrashekar G, Sahin F. A survey on feature selection methods.ComputElectr Eng. 2014; 40(1):16–28.

[15] Kim JO, Mueller CW. Factor Analysis: Statistical Methods and Practical Issues (Quantitative Applications in the Social Sciences). New York: Sage Publications, Inc; 1978.

VI. AUTHOR DETAILS

V. Maria Antoniate Martin is a Research Scholar in Computer Science at Bharathiar University, Coimbatore, Tamil Nadu, India. He is also working as an Assistant Professor in Department of Information Technology at St. Joseph's College, Tiruchirappalli, Tamil Nadu, India. He received his Bachelor of Science degree in Computer Science from Bharathidasan University in 2003, He completed his Masters in Science in Computer Science from the same University in 2006. He also completed his Masters in Philosophy in Computer Science from the same University in 2011. He has seven years of teaching experience. He has published nine research articles in reputed International Journals. He is also the co-author of a publication in a National Conference of importance. His area of research is Data Mining.

Dr. K. David is an Assistant Professor in the Department of Computer Science at H.H. the Rajah's College, Pudukkottai, TamilNadu, 622001. He has over fifteen years of teaching experience. He has published scores of papers in peer reviewed journals of National and International repute and is currently guiding seven Ph.D., scholars. His research interests include, UML, OOAD, Knowledge Management, Web Services and Software Engineering.

N. Bala Sankar, is a student of M.Sc. Computer Science, St. Joseph's College, Trichy-620002. He received him Bachelor of Science degree in Computer Science from Bharathidasan University in 2016.