Data Analytics with the IRIS Data set – finding the correlations between the products

G.N.V Vibhav Reddy[1],K Naga Sai Pravallika [2], M Deekshitha[3], E Shiva Ganesh[4], Y Praveen[5]

[1,2,3,4,5]Department of Computer Science and Engineering

[1,2,3,4,5] Sree Dattha Institute of Engineering and Science, Sheriguda, Telangana

ABSTRACT:

The term "correlation" refers to a mutual relationship or association between quantities. In almost any business, it is useful to express one quantity in terms of its relationship with others. Given two products X and Y, I have to find their correlation, i.e., their linear dependence/independence. Both products have equal dimension. The result should be a floating point number from [-1.0 .. 1.0]. The Iris flower data is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper "The use of multiple measurements in taxonomic problems" as an example of linear discriminant analysis. It is sometimes called Anderson’s Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula, “all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus.”The aim of the project is to find the special differences between the flower species and calculate their correlation terms with each other.

Index Terms: Correlation, Linear Dependence, Iris Flower Data Set, Multivariate Data, Linear Discriminant Analysis, Morphologic Variation, Species Comparison.

1.INTRODUCTION

It is estimated that 2.5 quintillion bytes of data are being produced daily; this has brought the world into the era of big data [1]. The growth of this data is continuing to increase exponentially [2]. Recent work has shown that big data has attracted unprecedented attention from both academics and industries. Massive amounts of data are being collected by many organizations on an ongoing basis. These datasets are being collected from various sources, including but not limited to the World Wide Web (WWW), social networks and sensor networks [3]. The discovery of knowledge from unstructured data accumulated from the WWW remains a difficult task because the content is suitable for human consumption rather than for machines [4]. Experimental evidence has shown that if big datasets are exploited and managed properly, it can give rise to critical intelligence that can motivate informed decisions and wider vision. The challenge in the big data era is to discover knowledge from big data using new techniques that were not imagined in the past [5]. The traditional methods such as econometric, statistical and mathematical models were mainly applied for data analytics in the 1970s [6]. However, the traditional methods are only effective in solving linear or near-linear problems, and some complex nonlinear time varying problems with limitations. The limitations of the traditional methods have prompted increasing attention in computational intelligence algorithms such as artificial neural network (ANN) due to their ability to solve complex real world problems better than the more traditional methodologies [7]. The ANN and other intelligent methodologies such as genetic algorithms (GA), rough sets, support vector machines (SVM), neurocomputing, fuzzy decision trees, fuzzy logic, etc. in hybrid, ensemble or single form can be exploited to perform data analytics [8]. Most of the algorithms pointed out effectively work on sizable datasets. The emergence of big data has posed a serious challenge to the research community in terms of how best to efficiently analyze these voluminous datasets, which are generally not in a central location. [5] has shown that the flow of data in a network has changed due to the emergence of big data. In general, data in a network move from one server to another. To realise the complete potential of big data, efficient and effective algorithms that can analyse this data are required; without these, the potential of big data cannot be explored [9]. Recently, researchers have made attempts to apply ANN within the context of big data analytics. Some progress has been made in this field using ANN, despite features of big data such as high volume, velocity, variety and diversity. In spite of the significance of the applications of ANN in big data analytics in exploring the potential of big data, we have not found a review that presents progress, challenges and future research direction on the application of ANN in big data analytics. This is to the best of the authors knowledge. However, [10] presented a survey that is limited to deep learning on big data analytics. This paper aims to review the attempts made by the research community in exploring the potential of big data using ANN. This is to provide researchers with the state-ofthe-art progress and the challenges of applying ANN within big data analytics, and to point out the potential for future research. This work also aims to identify for researchers the areas of ANN that have not been explored or which have received little attention from the research community. This review is intended for researchers to use as a standard in further exploration of the ANN that have not been explored in big data analytics. Unlike [10] the present study survey ANN applications to big data analytics in general without limitation to specific ANN architecture or model. The contributions of this work are summarised as follows: a. We present a taxonomy of the ANN according to their architecture within the big data analytics. b. We present a concise and precise summary of recent progress in the application of ANN to big data analytics. c. We show that ANN is the primary computational intelligence algorithm that has facilitated the exploration and exploitation of big data's potential. d. We demonstrate a trend in publications within big data analytics of the use of ANN, which is expected to grow rapidly in the very near future. e. We present several powerful computational intelligence algorithms that received little attention in big data analytics to researchers for easy identification. f. We highlight the challenges involved in previous attempts to explore big data using ANN and future research opportunities. The rest of this paper is organised as follows: Section II presents the taxonomy and basic theories of the ANN used in big data analytics. Section III gives a description, benchmarks and examples of big data, including real world examples. Section IV presents a review of the studies that have attempted the application of ANN within big data analytics. Section V presents a discussion which includes the strengths and limitations of the previous approaches.

2.LITERATURE SURVEY

2.1.Big data processing: Big challenges and opportunities

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of ``Big Data.'' While the promise of Big Data is real -- for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 -- there is currently a wide gap between its potential and its realization. Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modeling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the result

2.2. Affective neural networks and cognitive learning systems for big data analysis

Cognitive computing is a computational environment which is comprised of (1) a high-performance computing infrastructure powered by special processors such as multicore CPUs, GPUs, TPUs, and neuromorphic chips; (2) a software development environment with intrinsic support for parallel and distributed computing, and powered by the underlying computing infrastructure; (3) software libraries and machine learning algorithms for extracting information and knowledge from unstructured data sources; (4) a data analytics environment whose processes and algorithms mimic human cognitive processes; and (5) query languages and APIs for accessing the services of the cognitive computing environment. We have defined cognitive computing in terms of its functions, since it is not easy to define it precisely and completely by other methods. Cognitive analytics draws upon the cognitive computing environment to generate actionable insights by analyzing diverse heterogeneous data sources using cognitive models that the human brain employs.

2.3. A generalized Intelligent-agent-based fuzzy group forecasting model for oil price prediction

In this study, a generalized Intelligent-agent-based fuzzy group forecasting model is proposed for oil price prediction. In the proposed model, some single Intelligent-agent-based predictors with much disagreement are first created for crude oil price prediction. Then these single prediction results produced by these single intelligent predictors are fuzzified into some fuzzy prediction representations. Particularly, some methods of fuzzification are extended into a consolidated framework to make the later computation generalization. Subsequently, these fuzzified prediction representations are integrated into a fuzzy consensus, i.e., aggregated fuzzy prediction. Finally, the aggregated fuzzy prediction is defuzzified into a crisp value as the final prediction results. For verification and testing purposes, two typical oil price series are used to conduct the experiments.

2.4. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey

Financial distress prediction is of great importance to all stakeholders in order to enable better decision-making in evaluating firms. In recent years, the rate of bankruptcy has risen and it is becoming harder to estimate as companies become more complex and the asymmetric information between banks and firms increases. Although a great variety of techniques have been applied along the years, no comprehensive method incorporating an holistic perspective had hitherto been considered. Recently, SVM+ a technique proposed by Vapnik [17] provides a formal way to incorporate privileged information onto the learning models improving generalization. By exploiting additional information to improve traditional inductive learning we propose a prediction model where data is naturally separated into several groups according to the size of the firm. Experimental results in the setting of a heterogeneous data set of French companies demonstrated that the proposed model showed superior performance in terms of prediction accuracy in bankruptcy prediction and misclassification cost.

3.EXISTING SYSTEM

Data analysis is major challenge in the field of Information Technology as their will be continues growth in the data and new challenges will be keep on raising day by day. So in order to analysis the correlation between the terms of similar related data is again a huge task. As the values ranges differ with in no time. The system uses basic static methods in analyzing and finding the relation between the data.

Disadvantages of Existing System:

- No proper methods for sorting the data.
- Heavy algorithms to be used to in analysing the data.
- Critical programming techniques to be used.
- No logical packages for filtering the text.
- Libraries are not efficient in analysis user data is never considered in the enhancement of the product.

4.PROPOSED SYSTEM

We are using Python programming to analyse the data files. The Python program has many in built methods and libraries to work Programming is easy.

Advantages:

- Python has a huge packages which are more efficient to work on files and any kind of unstructured data.
- User friendly programming language.
- Has a Python Library.
- It has package as inbuilt libraries are used at most to analysis.

5.IMPLEMENTATION

MODULES:

1. Data Collection
2. Data Pre-Processing
3. Feature Extration
4. Evaluation Model

1.DATA COLLECTION

Data used in this paper is a set of product reviews collected from IRIS dataset records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already knowthe target answer. Data for which you already know the target answer is called labelled data.

2. DATA PRE-PROCESSING

Organize your selected data by formatting, cleaning and sampling from it.

Three common data pre-processing steps are: Formatting:

The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file. Cleaning: Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be removed from the data entirely. Sampling: There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

3.FEATURE EXTRATION

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Random forest. These algorithms are very popular in text classification tasks.

4.EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid over fitting, both methods use a test set (not seen by the model) toevaluate model performance. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions

5.VISUALIZATION

Data visualization is the display of information in a graphic or tabular format. Successful visualization requires that the data (information) be converted into a visual format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported. The goal of visualization is the interpretation of the visualized information by a person and the formation of a mental model of the information. In everyday life, visual techniques such as graphs and tables are often the preferred approach used to explain the weather, the economy, and the results of political elections. Likewise, while algorithmic or mathematical approaches are often emphasized in most technical disciplines—data mining included— visual techniques can play a key role in data analysis. In fact, sometimes the use of visualization techniques in data mining is referred to as visual data mining.

6.CONCLUSION

we have presents a survey on the application of ANN approaches within the context of big data analytics. In our review, state-of-the-art issues regarding the application of ANN in big data analytics is provided. The several attempts made by researchers in the application of ANN are discussed. The paper has discussed the challenges involving ANN in terms of handling big data, and future research opportunities are unveiled. The progress within big data analytics using ANN is described. Research on the applications of ANN in big data analytics is at an early stage and is expected to grow rapidly in the near future. We believe that expert researchers can use this review as a benchmark for future progress and development. In addition, the paper can be used as a starting point for novice researchers interested in big data analytics using ANN.

7.REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ‘‘Data mining with big data,’’ IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.

[2] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, ‘‘The rise of ‘big data’ on cloud computing,’’ Inf. Syst., vol. 47, pp. 98–115, Jan. 2015.

[3] Y. Liu, J. Yang, Y. Huang, L. Xu, S. Li, and M. Qi, ‘‘MapReduce based parallel neural networks in enabling large scale machine learning,’’ Comput. Intell. Neurosci., Aug. 2015, Art. no. 297672.

[4] A. Hussain, E. Cambria, B. W. Schuller, and N. Howard, ‘‘Affective neural networks and cognitive learning systems for big data analysis,’’ Neural Netw., vol. 58, pp. 1–3, Oct. 2014.

[5] J. Abawajy, ‘‘Comprehensive analysis of big data variety landscape,’’ Int. J. Parallel, Emergent Distrib. Syst., vol. 30, no. 1, pp. 5–14, 2015.

[6] M. A. Kaboudan, ‘‘Compumetric forecasting of crude oil prices,’’ in Proc. Congr. Evol. Comput., May 2001, pp. 283–287.

[7] L. Yu, S. Wang, and K. K. Lai, ‘‘A generalized Intelligent-agent-based fuzzy group forecasting model for oil price prediction,’’ in Proc. IEEE Int. Conf. Syst., Man Cybern., Oct. 2008, pp. 489–493.

[8] A. Verikas, Z. Kalsyte, M. Bacauskiene, and A. Gelzinis, ‘‘Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: A survey,’’ Soft Comput., vol. 14, no. 9, pp. 995–1010, 2010.

[9] D. Wan, Y. Xiao, P. Zhang, and H. Leung, ‘‘Hydrological big data prediction based on similarity search and improved BP neural network,’’ in Proc. IEEE Int. Congr. Big Data, Jun./Jul. 2015, pp. 343–350.

[10] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, ‘‘A survey on deep learning for big data,’’ Inf. Fusion, vol. 42, pp. 146–157, Jul. 2018.

[11] Y. Wang, B. Li, R. Luo, Y. Chen, N. Xu, and H. Yang, ‘‘Energy efficient neural networks for big data analytics,’’ in Proc. Conf. Design, Autom. Test Eur., Mar. 2014, pp. 1–2.

[12] X. Yu and O. Kaynak, ‘‘Sliding-mode control with soft computing: A survey,’’ IEEE Trans. Ind. Electron., vol. 56, no. 9, pp. 3275–3285, Sep. 2009.

[13] H. Jaeger, ‘‘Adaptive nonlinear system identification with echo state networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2003, pp. 609–616.

[14] S. Scardapane, D. Wang, and M. Panella, ‘‘A decentralized training algorithm for echo state networks in distributed big data applications,’’ Neural Netw., vol. 78, pp. 65–74, Jun. 2015.

[15] J. Herbert, ‘‘Echo state network,’’ Scholarpedia, vol. 2, no. 9, p. 2330, 2007.

[16] G. Hinton et al., ‘‘Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,’’ IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, Nov. 2012.

[17] Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521, no. 7553, p. 436, 2015.

[18] Y. LeCun et al., ‘‘Handwritten digit recognition with a back-propagation network,’’ in Proc. Adv. Neural Inf. Process. Syst., 1990, pp. 396–404.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ‘‘Gradient-based learning applied to document recognition,’’ Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[20] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep Learning, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[21] M. Klassen, Y. H. Pao, and V. Chen, ‘‘Characteristics of the functional link net: A higher order delta rule net,’’ in Proc. IEEE ICNN, Jun. 1988, pp. 507–513.

[22] J. C. Patra and R. N. Pal, ‘‘A functional link artificial neural network for adaptive channel equalization,’’ Signal Process., vol. 43, pp. 181–195, May 1995.

[23] Y. Yu, Y. Tian, N. Feng, and M. Lei, ‘‘Research on lifetime prediction method of tower crane based on back propagation

neural network,'' in Advances in Electronic Commerce, Web Application and Communication. Berlin, Germany: Springer, 2012, pp. 111–116.

[24] M. Lukoševičius and H. Jaeger, ''Reservoir computing approaches to recurrent neural network training,'' Comput. Sci. Rev., vol. 3, no. 3, pp. 127–149, 2009.

[25] C. Cortes and V. Vapnik, ''Support-vector networks,'' Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995