BIG DATA PRE-PROCESSING SOLUTIONS FOR TELECOM & MANUFACTURING SPECIFIC USE CASES

AJAY KUMAR

BHARTI SCHOOL OF TELECOMMUNICATION TECHNOLOGY & MANAGEMENT

INDIAN INSTITUTE OF TECHNOLOGY DELHI

NOVEMBER 2017

BIG DATA PRE-PROCESSING SOLUTIONS FOR TELECOM & MANUFACTURING SPECIFIC USE CASES

by

AJAY KUMAR
(Bharti School of Telecommunication Technology & Management)

Submitted

In fulfilment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

NOVEMBER 2017

CERTIFICATE

This is to certify that the thesis entitled "Big Data Pre-processing Solutions for Telecom & Manufacturing Specific Use Cases" being submitted by Ajay Kumar to the Indian Institute of Technology Delhi for the award of the degree of Doctor of Philosophy (Ph.D.) is a bonafide record of original research work carried out by him. He has worked under our supervision and has fulfilled the requirements for the submission of the thesis, which has reached the requisite standard for Ph.D. degree of this institute. The results presented in this thesis have not been submitted, in part or full, to any other university or institute for the award of any degree or diploma.

(Dr. Roma M. Debnath)
Department of Applied Statistics
Indian Institute of Public Administration
New Delhi, India

(Prof. Ravi Shankar)
Department of Management Studies
Indian Institute of Technology Delhi
New Delhi, India

ACKNOWLEDGEMENTS

I am immensely grateful to my supervisors Prof. Ravi Shankar and Dr. Roma M. Debnath for their guidance, unwavering support and encouragement. This thesis could not have attained its present form, both in content and presentation, without their active interest, direction and guidance. Their personal care has been the source of great inspiration. They have devoted their valuable time and took personal care in motivating me wherever I was disheartened.

I would like to thank the members of the Student Research Committee (SRC) comprising Prof. Surendra S. Yadav (Chairman) and Prof. M. P. Gupta (Internal Expert) of the Department of Management Studies and Prof. Mukesh Khare of the Department of Civil Engineering for giving useful comments and valuable suggestions at various stages of this study.

I would also like to take this opportunity to express my concern and gratefulness to the entire faculty and staffs of this Institution for having contributed in one way or the other in successfully completing my research. My thanks are due to fellow research scholars in particular, Dhanya, Ashish, Monika, Rachita, Vijayta, Divya, Devendra and Arun Purohit who evinced interest in my study and extended support during the research work.

I specially thank my wife Shaily for her support, patience and loving participation in accomplishing this task. I am also thankful to all the well-wishers for their direct and indirect support in accomplishing the research work. I also express my gratitude to my mother who remain a continuous source of inspiration for me.

Ajay Kumar

ABSTRACT

With an increasing importance of big data analytics in every domain of today's digital age from algorithmic trading and product recommendations to politics, there is a tremendous amount of research work going on in the field of big data pre-processing, which is taking us rapidly toward the advent of big data platforms and tools. Big data is 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 pre-processing and analysis. Data can be called the new oil and like oil, it needs to go through the refinement process before getting the actual value. With the usage of big data analytics, the organisations can get much useful information from interconnected, varied and complex datasets. This helps in forming valuable insights related to the state of the business and customer behaviour.

With the emerging technologies and all associated devices, it is predicted that large amounts of data will be created in the next few years – in fact, as much as 90% of current data were created in the last couple of years. In an era of complexity of the growing data, volume and the advent of big data, data pre-processing solutions have a key role to play to reduce high-dimensionality in machine learning problems. Because the volume, velocity, variety and complexity of datasets is continuously increasing, machine learning techniques have become indispensable in order to extract useful information from large amounts of otherwise meaningless data. Several studies have addressed the need to investigate underlying causes of poor classification in machine learning in big dataset, but few studies have focussed on the problems faced by the data scientists in the form of negative factors as noise, missing values, inconsistent & superfluous data, huge sizes in examples and features to learn and extract knowledge. Dimensionality is another major issue pertaining to big datasets, since it has a

large numbers of attributes and there is a dire requirement to reduce the data dimensionality for the learning purpose. In order to eliminate noisy, redundant or irrelevant attributes that could contribute to the deterioration of classification performance, 'big data pre-processing solutions' can be used. On the other side, the traditional methods have limitations when it comes to deal with the high dimensional big dataset, which has millions of instances and in successful extraction of the result within a fixed time.

The primary objective of the study was to identify and develop the best frameworks or set of tools to pre-process and reduce the size of the input data from telecom and manufacturing specific case studies. The study has identified the variables using fuzzy AHP that are affecting the subscriber's preferences in telecom sector for improving the operational efficiency and the variables that are accelerating the digital manufacturing for improving the manufacturing & operational efficiency. The research has developed and compared the four different big data pre-processing solutions for manufacturing & telecom datasets on four different fuzzy benchmark predictive models. The proposed frameworks have improved the run time, accuracy, reduced the misclassification rate, enhanced the model interpretability, improved the classifier performance and economized the storage problems by reducing the number of training data samples on both telecom & manufacturing datasets.

As evident from the existing literature review, the problem addressed in this study has not been discussed adequately either in the academic literature or in the practicing world. The study aimed to focusing the importance of pre-processing in the era of big data, where storage and processing of large data is as simple as processing the small structured data. In this research, two industry verticals viz. telecom (service sector) and manufacturing sector have been examined, both the sectors are using big data; industry-specific challenges exist in these industries and how big data enables to solve these challenges have been examined.

This research tried to capture the essence of big data in telecom and manufacturing sectors. The research has focussed on the questions “How much can companies in the telecommunications & manufacturing industry benefit from big data?”, “What is the relationship between big data pre-processing and machine learning to improve the classification performance in terms of scalability, efficiency and accuracy of high dimensional datasets?”, “What are the potential big data use cases in telecom and manufacturing industry and which of the available machine learning algorithms support best the functionality and usability in telecom & manufacturing industry data pre-processing tasks (including small and big data pre-processing)?” and “How can an organization implement the big data analytics to get more value out of the available data to optimize the business intelligence processes”. These are the critical questions and every company is exploring the means to increase the revenue and profits. Why to pre-process the data if other big data tools like Hadoop etc. support handling the large datasets effectively? If required, what kind of pre-processing are being discussed in the study? How different it is from the pre-processing that is being followed in a regular KDD process? What kind of tools work well in such a scenario and how it is done effectively on such a huge volume of data? These are some of the questions that the study has attempted to answer.

The research developed the four-big data driven frameworks for solving the data imbalance, feature selection, instance selection and space transformation big data pre-processing challenges effectively and efficiently after removing the noisy and boundary instances from training dataset.

So, the proposed frameworks in the research enhance the performance of support vector machine (SVM), artificial neural network (ANN) and induction algorithms by adding fuzzy membership and using advanced imbalanced, space transformation, instance selection &

feature selection solving methods to predict the performance decay on telecom and manufacturing datasets in big data environment. The experimental results were validated using different big datasets and the experiments show that the methods used in the proposed frameworks outperforms in optimality, efficiency and other statistical metrics.

The study facilitates the implementation of fuzzy machine learning algorithms to analyse the big data sets with the help of proposed data pre-processing solution. It also provides a comprehensive review of state-of-the-art fuzzy machine-learning literatures, including theoretical, empirical and experimental studies pertaining to the various needs and recommendations. Big data pre-processing is an inevitable solution for achieving better performance from machine learning techniques on large datasets. However, research societies emphasizes on advanced machine learning algorithm development and performance optimization. The crucial step of data pre-processing seems to be regarded by the same significance. This has been the motivation to conduct an extensive study in this area.

सार

एल्गोरिदमिक व्यापार और उत्पाद सिफारिशों से राजनीति तक आज के डिजिटल युग के हर डोमेन में बड़े डेटा विश्लेषिकी के बढ़ते महत्व के साथ, बड़े डेटा पूर्व प्रसंस्करण के क्षेत्र में बहुत अधिक शोध कार्य चल रहा है, जो हमें तेजी से बड़े डेटा प्लेटफार्मों और उपकरणों का आगमन बड़े डेटा को उच्च मात्रा, गति और विभिन्न प्रकार के डेटा के रूप में परिभाषित किया जाता है, जिनकी आवश्यकता होती है नए उच्च-प्रदर्शन प्रसंस्करण। बड़े डेटा को संबोधित करना एक चुनौतीपूर्ण और समय मांगने वाला कार्य है जिसके लिए एक बड़ी कम्प्यूटेशनल अवसंरचना की आवश्यकता होती है ताकि सफल डेटा पूर्व-प्रसंस्करण और विश्लेषण सुनिश्चित किया जा सके। डेटा को नए तेल और तेल की तरह कहा जा सकता है, वास्तविक मूल्य प्राप्त करने से पहले उसे परिशोधन प्रक्रिया के माध्यम से जाना चाहिए। बड़े डेटा विश्लेषिकी के उपयोग के साथ, संगठन इंटरकनेक्टेड, विविध और जटिल डेटासेट से बहुत उपयोगी जानकारी प्राप्त कर सकते हैं। यह व्यवसाय की स्थिति और ग्राहक व्यवहार से संबंधित मूल्यवान अंतर्दृष्टि बनाने में मदद करता है।

उभरती हुई प्रौद्योगिकियों और सभी संबंधित उपकरणों के साथ, यह अनुमान लगाया जाता है कि अगले कुछ सालों में बड़ी मात्रा में डेटा तैयार किया जाएगा - वास्तव में, पिछले कुछ वर्षों में वर्तमान आंकड़ों का जितना 90% बनाया गया था। बढ़ते आंकड़ों, मात्रा और बड़ी आंकड़ों के आगमन की जटिलता के युग में, मशीन सीखने की समस्याओं में उच्च आयामी कम करने के लिए डेटा प्री-प्रोसेसिंग समाधानों की महत्वपूर्ण भूमिका है। चूंकि डेटासेट की मात्रा, वेग, विविधता और जटिलता लगातार बढ़ती जा रही है, इसलिए मशीन सीखने की तकनीकों को अन्यथा अर्थहीन डेटा की बड़ी मात्रा से उपयोगी जानकारी निकालने के लिए अपरिहार्य बना दिया गया है। कई अध्ययनों ने बड़े डेटासेट में मशीन सीखने में गरीब वर्गीकरण के अंतर्निहित कारणों की जांच करने की आवश्यकता को संबोधित किया है, लेकिन कुछ अध्ययनों ने डेटा वैज्ञानिकों के सामने नकारात्मक समस्याओं के रूप में शोर, अनुपयोगी मूल्य, असंगत और अनावश्यक डेटा के रूप में समस्याओं पर ध्यान केंद्रित किया है, ज्ञान और ज्ञान प्राप्त करने के लिए उदाहरणों और विशेषताओं में विशाल आकार। आयाम एक बड़ी समस्या है जो बड़े डेटासेट से संबंधित है, क्योंकि इसमें बहुत अधिक गुण हैं और सीखने के उद्देश्य के लिए डेटा आयाम कम करने की आवश्यकता है। शोर, बेमानी या अप्रासंगिक गुणों को समाप्त करने के लिए, जो वर्गीकरण के प्रदर्शन की गिरावट में योगदान दे सकता है, 'बड़े डेटा प्री-प्रोसेसिंग समाधान' का उपयोग किया जा सकता है दूसरी तरफ, पारंपरिक तरीकों में उच्च आयामी बड़े डेटासेट से निपटने के लिए सीमाएं हैं, जिनमें लाखों उदाहरण हैं और निश्चित समय के भीतर परिणाम के सफल निष्कर्षण में।

अध्ययन का प्राथमिक उद्देश्य पूर्व प्रक्रिया के लिए सबसे अच्छा चौखटे या उपकरणों के सेट की पहचान करना और विकसित करना और टेलीकॉम से इनपुट डेटा के आकार को कम करना और विशिष्ट केस अध्ययनों का निर्माण करना था। अध्ययन ने फजी एएचपी का इस्तेमाल करते हुए वेरिएबल की पहचान की है जो परिचालन दक्षता में सुधार के लिए दूरसंचार क्षेत्र में ग्राहक की वरीयताओं को प्रभावित कर रहे हैं और वे चर जो विनिर्माण और परिचालन दक्षता में सुधार के लिए डिजिटल विनिर्माण में तेजी ला रहे हैं। अनुसंधान ने चार भिन्न फजी बेंचमार्क भविष्य कहने वाले मॉडल पर विनिर्माण और दूरसंचार डेटासेट के लिए चार अलग-अलग बड़े डेटा पूर्व प्रसंस्करण समाधान विकसित और तुलना किए हैं। प्रस्तावित ढांचे ने रन टाइम, सटीकता में सुधार किया, गलत वर्गीकरण दर को कम किया, मॉडल व्याख्यात्मकता को बढ़ाया, क्लासिफायरफ़ाइल प्रदर्शन में सुधार किया और दूरसंचार और विनिर्माण डेटासेट दोनों में प्रशिक्षण डेटा नमूनों की संख्या कम करके भंडारण समस्याओं को कम किया।

जैसा कि मौजूदा साहित्य की समीक्षा से स्पष्ट है, इस अध्ययन में संबोधित समस्या को शैक्षिक साहित्य या अभ्यास दुनिया में पर्याप्त रूप से भी चर्चा नहीं की गई है। इस अध्ययन का उद्देश्य बड़े आंकड़ों के युग में पूर्व-प्रसंस्करण के महत्व पर ध्यान केंद्रित करना है, जहां बड़े डेटा का भंडारण और प्रसंस्करण सरल रूप से छोटे संरचित डेटा प्रसंस्करण के रूप में सरल है। इस शोध में, दो उद्योग मंडल अर्थात् दूरसंचार (सेवा क्षेत्र) और विनिर्माण क्षेत्र की जांच की गई है, दोनों क्षेत्र बड़े डेटा का उपयोग कर रहे हैं; इन उद्योगों में उद्योग-विशिष्ट चुनौतियां मौजूद हैं और इन चुनौतियों का समाधान करने में बड़ी संख्या में कैसे जांच की गई है। इस शोध ने दूरसंचार और विनिर्माण क्षेत्रों में बड़े आंकड़ों का सार कब्जा करने की कोशिश की। इस शोध पर ध्यान केंद्रित किया गया है "बड़े डेटा से दूरसंचार और विनिर्माण उद्योग में कंपनियों को कितना लाभ मिल सकता है?", "स्केलेबिलिटी, दक्षता के मामले में वर्गीकरण के प्रदर्शन में सुधार के लिए बड़े डेटा पूर्व प्रसंस्करण और मशीन सीखने के बीच संबंध क्या है? और उच्च आयामी डेटासेट की सटीकता? "," दूरसंचार और विनिर्माण उद्योग में संभावित बड़े डेटा का उपयोग किस प्रकार होता है और उपलब्ध मशीन सीखना एल्गोरिदम का कौन सा टेलीकॉम और विनिर्माण उद्योग डेटा प्रसंस्करण कार्य (छोटे सहित और बड़े डेटा पूर्व प्रसंस्करण)? "और" कैसे एक संगठन बड़े डेटा विश्लेषिकी को कार्यान्वित कर सकता है ताकि व्यावसायिक डेटा प्रक्रियाओं को अनुकूलित करने के लिए उपलब्ध डेटा से अधिक मूल्य मिल सके "। ये महत्वपूर्ण सवाल हैं और हर कंपनी राजस्व और मुनाफा बढ़ाने के साधनों की खोज कर रही है। बड़े डेटासेट को प्रभावी ढंग से संभालने में सहायता करना आदि Hadoop जैसे अन्य बड़े डेटा टूल का डेटा पूर्व-प्रक्रिया क्यों करें? यदि आवश्यक हो, अध्ययन में किस प्रकार की पूर्व-प्रसंस्करण चर्चा की जा रही है? यह पूर्व प्रसंस्करण से कितना अलग होता है जिसे नियमित केडीडी प्रक्रिया में किया जा रहा है? इस तरह के परिदृश्य में किस तरह के औजार अच्छे तरीके से काम करते हैं और यह कितनी बड़ी मात्रा में डेटा पर प्रभावी ढंग से किया जाता है? ये कुछ ऐसे प्रश्न हैं जिनके अध्ययन ने उत्तर देने का प्रयास किया है।

अनुसंधान डाटासेट से शोर और सीमा के उदाहरणों को हटाने के बाद डेटा असंतुलन, सुविधा चयन, उदाहरण चयन और स्थान परिवर्तन बड़े डेटा पूर्व-प्रसंस्करण चुनौतियों को प्रभावी ढंग से और कुशलतापूर्वक हल करने के लिए चार बड़े डेटा चालित ढांचे का विकास किया।

इसलिए, अनुसंधान में प्रस्तावित चौखटे फजी सदस्यता को जोड़कर समर्थन वाकर मशीन (एसवीएम), कृत्रिम तंत्रिका नेटवर्क (एएनएन) और प्रेरण एल्गोरिदम का प्रदर्शन बढ़ाते हैं और उन्नत असंतुलित, अंतरिक्ष परिवर्तन, उदाहरण चयन और सुविधा चयन के तरीके को सुलझाने के तरीके का अनुमान लगाते हैं। बड़े डेटा वातावरण में दूरसंचार और विनिर्माण डेटासेट पर प्रदर्शन क्षय प्रयोगात्मक परिणाम विभिन्न बड़े डेटासेट्स का उपयोग करके मान्य किए गए थे और प्रयोगों से पता चलता है कि प्रस्तावित ढांचे में उपयोग किए गए तरीके ऑप्टिमाइलिटी, दक्षता और अन्य सांख्यिकीय मैट्रिक्स में बेहतर प्रदर्शन करते हैं।

अध्ययन प्रस्तावित डेटा पूर्व प्रसंस्करण समाधान की मदद से बड़े डेटा सेट का विश्लेषण करने के लिए फजी मशीन सीखने एल्गोरिदम के कार्यान्वयन की सुविधा प्रदान करता है। यह विभिन्न आवश्यकताओं और सिफारिशों से संबंधित सैद्धांतिक, व्यावहारिक और प्रायोगिक अध्ययनों सहित, राज्य-के-अत्याधुनिक फजी मशीन-शिक्षण साहित्य की एक व्यापक समीक्षा भी प्रदान करता है। बड़े डेटा प्री-प्रोसेसिंग बड़े डेटासेट पर मशीन सीखने की तकनीक से बेहतर प्रदर्शन प्राप्त करने के लिए एक अनिवार्य समाधान है। हालांकि, अनुसंधान समाज उन्नत मशीन सीखने एल्गोरिथम विकास और प्रदर्शन अनुकूलन पर बल देता है। डेटा के पूर्व-प्रसंस्करण के महत्वपूर्ण चरण को उसी महत्व से माना जाता है। इस क्षेत्र में व्यापक अध्ययन करने के लिए यह प्रेरणा रही है।

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACF	Auto Correlation Function
AHP	Analytic Hierarchy Process
ANN	Artificial Neural Network
AUC	Area under Curve
CCA	Canonical Correlation Analysis
CDR	Call Details Records
CLV	Customer lifetime Value
EOQ	Economic Order Quantity
EPQ	Economic Production Quantity
FURIA	Fuzzy Unordered Rule Induction Algorithm
HDFS	Hadoop Distributed File Systems
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbours
LDA	Linear Discriminant Analysis
MAPE	Mean Absolute Percentage Error
MMM	Marketing Mix Modelling
MSE	Mean Square Error
PACF	Partial Auto Correlation Function
PCA	Principal Component Analysis
RBF	Radial basis Function
RNNR	Reverse Nearest Neighbours Reduction
ROC	Receiver Operating Characteristics
ROI	Return on Investment
SMOTE	Synthetic Minority Over-sampling Technique
SQL	Structured Query Language
SVM	Support Vector Machine