PLOS ONE

RESEARCH ARTICLE

Use of relevancy and complementary information for discriminatory gene selection from high-dimensional gene expression data

Md Nazmul Haque[1]*, Sadia Sharmin[2], Amin Ahsan Ali[3], Abu Ashfaqur Sajib[4]*, Mohammad Shoyaib[1]

1 Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh, **2** Department of Computer Science & Engineering, Islamic University of Technology, Dhaka, Bangladesh, **3** Department of Computer Science & Engineering, Independent University, Dhaka, Bangladesh, **4** Department of Genetic Engineering & Biotechnology, University of Dhaka, Dhaka, Bangladesh

* bsse0635@iit.du.ac.bd (MNH); abu.sajib@du.ac.bd (AAS)

Citation: Haque MN, Sharmin S, Ali AA, Sajib AA, Shoyaib M (2021) Use of relevancy and complementary information for discriminatory gene selection from high-dimensional gene expression data. PLoS ONE 16(10): e0230164. https://doi.org/10.1371/journal.pone.0230164

Editor: Y-h. Taguchi, Chuo University, JAPAN

Received: February 22, 2020

Accepted: September 21, 2021

Published: October 6, 2021

Data Availability Statement: GDS6063, GDS5306, GDS4431, GDS3341, GDS3610, GDS4824, GSE106291 datasets are available at the Gene Expression Omnibus (GEO) database at https://www.ncbi.nlm.nih.gov/geo/.

Funding: AAS received research grant from the innovation fund (2019-2020) of the ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh (http://ims.ictd.gov.bd/). MNH received research fellowship (56.00.0000.028.33.029.19-329) from ICT Division, Ministry of Posts, Telecommunications

Abstract

With the advent of high-throughput technologies, life sciences are generating a huge amount of varied biomolecular data. Global gene expression profiles provide a snapshot of all the genes that are transcribed in a cell or in a tissue under a particular condition. The high-dimensionality of such gene expression data (*i.e.*, very large number of features/genes analyzed with relatively much less number of samples) makes it difficult to identify the key genes (biomarkers) that are truly attributing to a particular phenotype or condition, (such as cancer), *de novo*. For identifying the key genes from gene expression data, among the existing literature, mutual information (MI) is one of the most successful criteria. However, the correction of MI for finite sample is not taken into account in this regard. It is also important to incorporate dynamic discretization of genes for more relevant gene selection, although this is not considered in the available methods. Besides, it is usually suggested in current studies to remove redundant genes which is particularly inappropriate for biological data, as a group of genes may connect to each other for downstreaming proteins. Thus, despite being redundant, it is needed to add the genes which provide additional useful information for the disease. Addressing these issues, we proposed Mutual information based Gene Selection method (*MGS*) for selecting informative genes. Moreover, to rank these selected genes, we extended *MGS* and propose two ranking methods on the selected genes, such as MGS_f—based on frequency and MGS_{rf}—based on Random Forest. The proposed method not only obtained better classification rates on gene expression datasets derived from different gene expression studies compared to recently reported methods but also detected the key genes relevant to pathways with a causal relationship to the disease, which indicate that it will also able to find the responsible genes for an unknown disease data.

Science of the Total Environment

Economic and Health Burden Associated with the Ambient Air Pollution in Bangladesh Over the Last Decade (2008-2019)

--Manuscript Draft--

Manuscript Number:	STOTEN-D-22-06433
Article Type:	Research Paper
Section/Category:	
Keywords:	ambient air pollution; long-term exposure; health burden; economic burden; cardiovascular and respiratory diseases
Corresponding Author:	Abdus Salam, PhD Dhaka University Dhaka, Dhaka BANGLADESH
First Author:	Abdus Salam, PhD
Order of Authors:	Abdus Salam, PhD
	Md Riad Sarkar Pavel
	Shahid Uz Zaman
	Khaled Shaifullah Joy
	Nazmul Haque
	Farah Jeba
	Juwel Rana
Abstract:	Exposure to high levels of air pollutants (PM2.5, PM10, CO, O3, NO2, and SO2) has serious negative health and economic consequences, notably in a heavily polluted country like Bangladesh. The Health and economic burden attributed to air pollutants in the eight administrative divisions (Barishal, Chattogram, Dhaka, Khulna, Rajshahi, Rangpur, and Sylhet) of Bangladesh have been estimated for the last decade (2008-2019). Reanalysis products from the Copernicus Atmosphere Monitoring Service (CAMS) were validated and utilized for this study. Exposure-response function based on the concentration, exposed population, incidence rate of health outcomes, threshold values were regarded to estimate the health burden. Value of statistical life (VSL) and cost of illness (COI) methods were utilized to quantify the economic burden on national Gross Domestic Products (GDP). Yearly premature all-cause mortality due to ambient air pollution was estimated to be 0.80 ± 0.23 million where the contribution of cardiovascular (CVD) and respiratory diseases was approximately 51.71% and 21.58%. PM2.5, PM10, CO, O3, NO2, and SO2 each accounted for approximately 59.29%, 9.20%, 4.61%, 14.71%, 4.01%, and 8.19% of all-cause premature mortality, respectively. Yearly ~12.59 million hospitalizations were attributed to CVD and respiratory diseases. The annual economic burden of these health outcomes was estimated to be 12.11 ± 7.45 billion dollars, which equaled 6.26 ± 1.64% of the national GDP during this period. Among the divisions, the majority (5.19 ± 1.43% of GDP) of the national economic burden was attributed to Dhaka. According to this study, minimizing ambient air pollution, particularly PM2.5, in Dhaka may save the government of Bangladesh a significant amount of GDP.
Suggested Reviewers:	Md Firoz Khan, PhD Professor, North South University firoz.khan@northsouth.edu Expert
	Talib Latif, PhD Professor, National University of Malaysia: Universiti Kebangsaan Malaysia talib@ukm.edu.my Expert
	Hiroshi Tanimato, PhD Professor, National Institute of Environmental Health Sciences

Economic and Health Burden Associated with the Ambient Air Pollution in Bangladesh Over the Last Decade (2008-2019)

Abdus Salam[1], Md Riad Sarkar Pavel[2], Shahid Uz Zaman[1], Khaled Shaifullah Joy[1], Nazmul Haque[3], Farah Jeba[1], and Juwel Rana[4]*

[1]Department of Chemistry, Faculty of Science, University of Dhaka. Dhaka- 1000

[2]Department of Chemistry, Iowa state university, Ames, IA-50011, USA

[3]Department of Computer Science and Engineering, Islamic University of Technology, Gazipur

[4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Canada

[*]Corresponding author's Email: asalam@gmail.com; asalam@du.ac.bd; Tel.: +8801817061160

Submitted to

Science of the Total Environment

March 20, 2022

A Study on Classifying Stack Overflow Questions based on Difficulty by Utilizing Contextual Features

Journal:	*IEEE Access*
Manuscript ID	Access-2022-16693
Manuscript Type:	Regular Manuscript
Date Submitted by the Author:	20-Jun-2022
Complete List of Authors:	RAIDA, MALIHA NOUSHIN; Islamic University of Technology SRISTY, ZANNATUN NAIM; Islamic University of Technology ULFAT, NAWSHIN; Islamic University of Technology MONISHA, SHEIKH MOONWARA ANJUM; Islamic University of Technology MOSTAFA, JUBAIR IBNA; Islamic University of Technology HAQUE, NAZMUL; Islamic University of Technology
Keywords: Please choose keywords carefully as they help us find the most suitable Editor to review:	Software engineering, Software maintenance, Text analysis, Feature extraction
Subject Category Please select at least two subject categories that best reflect the scope of your manuscript:	Computers and information processing, Computational and artificial intelligence
Additional Manuscript Keywords:	

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.
Digital Object Identifier 10.1109/ACCESS.2017.DOI

A Study on Classifying Stack Overflow Questions based on Difficulty by Utilizing Contextual Features

MALIHA NOUSHIN RAIDA [1, 2], ZANNATUN NAIM SRISTY [1, 2], NAWSHIN ULFAT [1, 2], SHEIKH MOONWARA ANJUM MONISHA [1, 2], MD. JUBAIR IBNA MOSTAFA [1, 2], MD. NAZMUL HAQUE [1, 2]
[1]Software Engineering Lab (SELab), Islamic University of Technology (IUT), Gazipur 1704, Bangladesh
[2]Department of Computer Science and Engineering, Islamic University of Technology (IUT), Gazipur 1704, Bangladesh
Corresponding author: Maliha Noushin Raida (e-mail: malihanoushin@iut-dhaka.edu).

ABSTRACT Technical question-answering sites like Stack Overflow are gaining enormous attention from the practitioners of specialized fields to exchange their programming knowledge. They ask questions on different topics, having various levels of difficulty and complexity. To answer such questions, all practitioners do not have the same level of expertise on those topics. However, the existing approach of Stack Overflow does not consider the difficulty and primarily filters out the questions based on topics only. For this reason, a large percentage of questions fail to attract the attention of appropriate users, resulting in questions having no answer or a significant delay in response time. To address these limitations, we incorporate three models, namely TF-IDF, LDA, and Doc2Vec, to extract semantic and context-dependent features that can measure the difficulty of questions. Each of these models is used with different classifiers along with other features to classify the questions based on difficulty. Extensive experiments on different datasets exhibit the effectiveness of our models, and the Doc2Vec outperforms the other models. We also discovered that the contextual features are correlated with question difficulty, and one subset of features outperforms others. The proposed approach can be beneficial for building an automatic tagger based on question difficulty.

INDEX TERMS Stack Overflow, Question Difficulty, Topic Modeling, Doc2Vec, TF-IDF

I. INTRODUCTION

Developers frequently use community Question and Answering (Q&A) sites like Stack Overflow (SO) to solve programming challenges. Every day, over 6,000 new questions are posted to SO, and approximately 10 million users [1] follow the site. The users, ranging from beginners to experts, participate in constructive exchanges of knowledge on this site, forming a dynamic programming community. Anyone can ask questions about various topics to fix their issues, and other users can respond or offer their thoughts on the same. To make this procedure more user-friendly, SO offers several filtering and preference choices such as Interesting[1], Bounties[2], Watched Tags[3], Ignore Tags[3] for suggesting appropriate ones. However, with qualitative and quantitative analysis on the live server [4], we found that it takes around 16 days to get an answer while the standard deviation varies up to 113 days. Besides, 30% of the total questions remain unanswered, which is a major concern because it hinders the proliferation of knowledge sharing in SO.

Many researchers have drawn this concern by addressing it from many angles, such as exploring the reasons behind unanswered questions, finding several factors behind becoming a question unanswered, and so on. For example, Wang et al. [2] conducted an empirical study on four Stack Exchange websites to find out the reasons for not getting the answers fast from the Q&A systems. However, they were more focused on the user profile, not the question. Besides, some reasons (e.g., frequent/non-frequent users) are hard to measure in practice, and how these reasons can be applied to predict the unanswered questions are not discussed. To solve these

[1]https://stackoverflow.com/?tab=interesting
[2]https://stackoverflow.com/?tab=bounties
[3]https://stackoverflow.help/en/articles/5611335-watch-or-ignore-tags
[4]https://data.stackexchange.com/stackoverflow/queries