Notification Number: 540/2023

Date of Award: 23/06/2023

Name of the Scholar	**:**	Aamish Izhar
Name of the Supervisor	**:**	Prof. S.M.K. Quadri (Sup) & Prof. S.A.M. Rizvi (Co-Sup)
Name of the Department/Centre	**:**	Department of Computer Science
Topic of Research	**:**	Big Data-Enabled Framework for Smart Cities
Keywords:	**:**	Big Data, Smart Cities, Traffic Congestion, Unlabeled Datasets, Data Optimization

Findings

According to projections, urban regions will be home to 68% of the world's population by 2050. The rapid urbanization and increasing opportunities in the cities, has prompted a large-scale emigration of people from rural areas to cities in quest of better opportunities, living circumstances and sources of income. So, managing such urban surge is very critical for any nation, and requires planning for the needed infrastructure and facilities. To achieve this, a city is needed to made sustainable which is possible only if it is smart. Many developed and developing nations across the world have shown interest in the "smart city" concept and has undertaken many initiatives to make their cities smart.

A smart city uses information and communications technology (ICT), new age technologies, and innovation developments to address urban concerns, such as enhancing livability, fostering economic development, creating a sustainable, safe environment, and supporting effective urban management practices. The ultimate goal of a smart city is to better utilize public resources, increase the caliber of services offered to inhabitants, and reduce operating costs related to public administration, even though there is no formal definition that is widely accepted.

ICT and the Internet of Things (IoT) have paved the way for the generation of huge amount of data by connecting citizens, digital devices and various sensors for sensing and reporting purposes. Big data provides the possibility for the city to gain insightful knowledge from a sizable amount of data gathered from diverse sources. Smart cities effectively store, analyze, and mine big data systems to produce information to improve various smart city services. It also assists decision-makers in planning for any growth in smart city areas, resources, or services.

Big data can serve many application area of a smart city including healthcare, transportation systems, water management, energy management, waste management and other community services. These services can be made available to the citizens anywhere anytime and at reasonable cost which enhances their quality of living. The full potential of big data can be leveraged through proper tools and techniques for effective and efficient data analysis. Such tools and technologies include: Hadoop, Spark, Hive, Hbase, data mining, machine learning,

big data analytics etc. among others. The application of big data to smart cities domain has brought many benefits as well as challenges. While on one side it helps in decision making and improving smart cities related services, on the other side it poses certain challenges due to its inherent nature.

This thesis aims to mitigate the challenges present in the smart cities domain. One of the main issues for cities around the world is mobility, which has an impact on social, economic, and environmental aspects of big cities. The prediction of road traffic congestion has an important role in managing the transport system in smart cities. Informed travel decisions can be made with the help of such predictions by improving and providing better traveler information services.

Moreover, precise traffic predictions can help improve road safety by preventing accidents and help to improve transportation costs and decrease air pollution. Though, there are various methods to tackle this problem, most of them suffer from improper label generation. Therefore, motivated by such shortcomings, we propose an intuitive and logical solution by proposing two methods for label generation which highlight and consider some crucial features to generate the labels for an unlabeled dataset in the context of traffic dataset. We also present detail descriptions related to label generation process, countering the class imbalance problem, choice of classifiers for model training and evaluation. The results indicate that labels generated based on our approach prove to be effective in properly discriminating congestion and non-congestion scenarios.

Further, this thesis presents a detail analysis of our proposed approach for traffic congestion prediction under big data as traffic data is generated from a number of sources and is of big data nature. Big data approach using massive traffic dataset having $\approx$ 13.5 million traffic instances has been utilized for accurate prediction of traffic congestion. To deal with data imbalance problem in generated labels, random undersampling, SMOTE and a combination of both, has been employed to negate any bias inherent to two individual sampling techniques. Extensive experiments using various well-known classifiers, have been performed for label-generation evaluation. The results show that our proposed approaches can effectively discriminate congestion and non-congestion scenarios under big traffic data.

Finally, this thesis presents a data optimization approach to tackle the size of data. Vast amount of data are being generated through various sources in a smart city. These data are gathered and processed to gain insight. This is called data-driven approach. However, recently the focus is shifted to data-centric approach, where the emphasis is on data quality rather than size of dataset. It entails investing more time in methodically modifying and enhancing datasets to improve machine learning applications' accuracy. The real challenge here lies in the fact that data should be reduced without compromising the performance. Through our proposed approach, we have reduced the size of dataset (row-wise data reduction) considerably (around 50%) without affecting the performance. All the steps of our proposed approach has been described and evaluation results performed on real-world traffic datasets prove our approach to be effective.