To

My family

CERTIFICATE

It is certified that the work contained in the thesis titled **Effective Learning Models on Pattern Mining Applications** by **Shivang Agarwal** has been carried out under my supervision and that this work has not been submitted elsewhere for a degree. It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy and SOTA for the award of Ph.D. Degree.

Supervisor

Dr. Ravindranath Chowdary C.

Deptt. of Computer Science & Engg.

IIT(BHU)

Varanasi - 221005

DECLARATION

I, **Shivang Agarwal**, certify that the work embodied in this thesis is my own bona fide work and carried out by me under the supervision of **Dr Ravindranath Chowdary C.** from January-2017 to April-2021, at the Department of Computer Science Engineering, Indian Institute of Technology (BHU), Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.

Date:

Place: **(Shivang Agarwal)**

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

Dr. Ravindranath Chowdary C.

IIT(BHU), Varanasi

Signature of Head of Department/Coordinator of School

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: **Effective Learning Models on Pattern Mining Applications**
Name of Student: **Shivang Agarwal**

Copyright Transfer

Acknowledgments

Though, only my name appears on the cover of this dissertation, so many great people have contributed to its production. I owe my gratitude to all those people who have made this thesis possible and because of whom my post graduate experience has been one that I will cherish forever.

I present my sincere gratitude to my thesis supervisor Dr Ravindranath Chowdary C, for his continuous guidance during the course of my PhD degree. I am thankful to him for teaching me the value of discipline and consistency.

I take this opportunity to thank Dr R. S. Singh, Dr Lavanya Selvaganesh of Department of Computer Science and Engineering and Department of Mathematics, IIT (BHU), respectively and Dr Ajita Rattani of Wichita State University, USA, for their valuable inputs to this dissertation.

I am thankful to my family for always providing me with emotional support. Special mention to my niece Anumeha for always being available for a video call whenever writing was difficult. I thank my wife, Jyoti, for being a patient listener. Being a PhD scholar herself, she understands me better than anyone else.

I thank my labmates, Mr Chintoo Kumar, Mrs Deepika Shukla and Mr Paras Tiwari, who provided stimulating discussions and happy distractions for resting my mind outside of my research. My appreciation also goes to my batchmates and friends, Dr Tribikram Pradhan, Mr Sushant Pandey, Dr Ashish Gupta and Mr Anshul Sharma, for the cherished time spent together in Varanasi.

Date: ________________ Shivang Agarwal

Abstract

This dissertation investigates various learning paradigms' behaviour on two pattern mining applications: spoof fingerprint detection and automatic hate speech detection on social media platforms. It argues that learning paradigms must consider properties inherently present in the data while deciding the number of hypotheses to be used for classification. These data properties are vital in applications that require finding a specific pattern in a massive amount of data. In our study, spoof fingerprint detection is regarded as an open-set classification task, and the generalization abilities of hate speech detectors are explored rigorously. Therefore, the emphasis is on the performance under cross-sensor, cross-material and cross-dataset environments.

The dissertation's central claim is that pattern mining applications require the learning model to be adaptive to the properties intrinsic to the dataset. Therefore, we propose a novel learning model, EaZy learning which is midway between eager and lazy learning. EaZy learning overcomes the high storage requirements and low prediction efficiency while maintaining good local approximations. The proposed model can be regarded as a variant of ensemble learning that considers the properties of data and moves adaptively towards the eager or lazy nature of the underlying problem. EaZy learning differs from ensemble learning in the way it generates the ensemble and how it integrates the outputs of the ensemble members. One of the critical ensemble learning requirements is to have a pool of diverse base classifiers. It achieves this by performing clustering on the training set and training the base classifiers on each cluster. In that way, the model delivers diversity, which results in different generalization capabilities of base classifiers in the ensemble. EaZy learning is a plug-in solution capable of working with various base classifiers on any application.

Later, an incremental model is proposed, which accommodates new knowledge without having to retrain the model from scratch. Incremental learning enables the learner

to accommodate new knowledge without retraining the existing model. It is a challenging task that requires learning from new data and preserving the knowledge extracted from the previously accessed data. This challenge is known as the stability-plasticity dilemma. We propose AILearn, a generic model for incremental learning that overcomes the stability-plasticity dilemma by carefully integrating the base classifiers' ensemble on new data with the current ensemble without retraining the model from scratch using entire data. One of the significant challenges associated with spoof fingerprint detection is the performance drop on spoofs generated using new fabrication materials. Also, it is beneficial in automatic hate speech detection on social media, where the narratives change continuously over time. To the best of our knowledge, AILearn is the first attempt in incremental learning algorithms that adapts to data properties for generating a diverse ensemble of base classifiers.

Next, we propose A-Stacking and A-Bagging: the adaptive versions of ensemble learning approaches Stacking and Bagging, respectively. One of the main motives of ensemble learning is to generate an ensemble of multiple weakly correlated experts. The proposed models achieve this by producing a set of disjoint experts where each expert is trained on a different subset of the dataset. A-Bagging applies the same base learner to different subsets of data and combines their predictions using weighted majority voting. A-Stacking uses logistic regression as the meta-classifier, which resulted in better performance than the best individual base classifier. This justifies the extra effort of employing a meta-classifier.

Based on the analysis of the influence of various types of features on different classifiers, we conducted a comprehensive study on the impact of using handcrafted and deep features on presentation attack detection. We conduct a comprehensive study on the impact of handcrafted and deep features from fingerprint images on the classification error rate of the fingerprint liveness detection task. We use LBP, LPQ and BSIF as handcrafted features and VGG-19 and Residual CNN as deep feature extractors for this study. As the problem is targeted as an open-set classification task, the emphasis is on achieving better robustness and generalization capability. In our observation, handcrafted features outperformed their deep counterparts in two of the three cases under the within-dataset environment. In the cross-sensor environment, deep features obtained a better accuracy, and in the cross-dataset environment, handcrafted features brought a lower classification error rate.

Using a case study on hate speech propagation during the ongoing global pandemic, we show the usefulness of automatic hate speech detection and propose adaptive ensemble models to address it. Automatic hate speech detection on social media platforms is an essential task that has not been solved efficiently despite various researchers' multiple attempts. It is a challenging task that involves identifying hateful content from social media posts. Relying on manual inspection delays the process, and the hateful content may remain available online for a long time. The current state-of-the-art methods for tackling hate speech perform well when tested on the same dataset but fail miserably on cross-datasets. Therefore, we propose an ensemble learning-based adaptive model for automatic hate speech detection, improving the cross-dataset generalization. The proposed expert model for hate speech detection works towards overcoming the strong user bias present in the available annotated datasets. We conduct our experiments under various experimental setups and demonstrate the proposed model's efficacy on the latest issues such as COVID-19 and US presidential elections. In particular, the loss in performance observed under cross-dataset evaluation is the least among all the models. Also, while restricting the maximum number of tweets per user, we incur no drop in performance.

Later, hate speech detection performance is accelerated by parallelizing the models and achieving reasonable speedup and efficiency. To deal with large-scale data efficiently and accurately, we need a simple, scalable and robust framework. Therefore, we propose parallelization to the standard ensemble-based algorithms so that they can be used to speed up automatic hate speech detection on SMPs. We parallelize bagging, A-stacking and random sub-space algorithms and test both serial and 'parallel versions on the standard high-dimensional datasets for hate speech detection. We observe a significant speedup with high efficiency that claims that the proposed models are suitable for the considered application. We observed that the accuracy is not affected by parallelizing the algorithms compared with serial algorithms executing on a single machine.

The study is significant as it addresses the fundamental requirements of an ensemble model (i.e., diversity and accuracy) by generating disjoint base classifiers trained on subsets of the original training data. The dissertation concludes with a discussion on the proposed models' impact on the applications mentioned above under various test scenarios.

Contents

List of Tables

List of Figures