

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

The undersigned hereby assigns to the Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the Doctor of Philosophy.

Date:

Place: **(Shivang Agarwal)**

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

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

Abstract	ii
List of Tables	viii
List of Figures	xi
1 Introduction	1
1.1 Learning Paradigms	1
1.1.1 Eager Learning	1
1.1.2 Lazy Learning	2
1.1.3 EaZy Learning	3
1.1.4 Incremental Learning	3
1.1.5 Ensemble Based Models	5
1.2 Spoof Fingerprint Detection	5
1.3 Automatic Hate Speech Detection on SMPs	6
1.4 Structure of the Thesis	8
2 Literature Survey	9
2.1 Ensemble Learning	9
2.2 Incremental Learning	9
2.3 Spoof Fingerprint Detection	11
2.4 Automatic Hate Speech Detection	13
3 EaZy Learning: An Adaptive Variant of Ensemble Learning	16
3.0.1 Learning Paradigms for Spoof Fingerprint Detection	17
3.1 EaZy Learning	19
3.2 Experimental Setup	22

3.2.1	Datasets	22
3.2.2	Features	23
3.2.3	Setup	23
3.3	Results and Discussion	25
3.3.1	Discussion	26
4	AILearn: An Adaptive Incremental Learning Model for Fingerprint Liveness Detection	28
4.1	AILearn: A Generic Model for Incremental Learning	31
4.2	AILearn for Spoof Fingerprint Detection	33
4.2.1	Feature Extraction	34
4.2.2	Ensemble Generation	36
4.3	Experimental Results and Discussion	36
4.3.1	Experimental Settings	36
4.3.2	Results	38
4.3.3	Feature-Level Comparison	40
4.3.4	Comparison with State-of-the-art	42
4.3.5	Discussion on Results	43
5	A-Stacking and A-Bagging	46
5.1	Introduction	46
5.1.1	Contributions	47
5.2	Stacking	48
5.2.1	A-Stacking	48
5.3	Bagging	49
5.3.1	A-Bagging	50
5.4	Results and Discussion	53
5.4.1	Experimental Setup	53
5.4.2	Datasets and Pre-processing	54
5.4.3	Results	54
5.4.4	Discussion on Results	61

6	Handcrafted V/S Deep Features	64
6.1	Introduction	64
6.2	Feature Representation	65
6.2.1	Handcrafted Features	66
6.2.2	Deep Features	66
6.3	Experimental Study	67
6.3.1	Feature Extraction	67
6.3.2	Dataset	67
6.3.3	Classifiers	67
6.3.4	Experimental Protocol	69
6.3.5	Results	70
6.3.6	Discussion	74
7	Combating Hate Speech using an Adaptive Ensemble Learning Model with a case study on COVID-19	76
7.1	Introduction	76
7.2	Importance of Automatic Hate Speech Detection	81
7.2.1	Hate Speech in the Times of COVID-19	81
7.2.2	Hate Speech Related to US Presidential Election	82
7.3	Proposed Model for Automatic Hate Speech Detection	82
7.4	Parallel Ensemble Learning Models	84
7.4.1	Parallelized Bagging	85
7.4.2	Parallelized A-Stacking	85
7.4.3	Parallelized Random-Subspace	86
7.5	Experimental Setup, Analysis and Discussion	87
7.5.1	Datasets	88
7.5.2	Experimental Setup	90
7.5.3	Results	91
7.5.4	Discussion	97
8	Conclusions and Future Work	101
A	List of Publications	104

List of Tables

3.1	Description of datasets.	23
3.2	Performance evaluation of EaZy learning on Category-1.	25
3.3	Performance evaluation of EaZy learning on Category-2.	26
4.1	Partitioning of the datasets in Phase I and Phase II for evaluation of the AIIearn algorithm.	37
4.2	Stability-Plasticity calculation on LivDet 2011 [1].	42
4.3	Stability-Plasticity calculation on LivDet 2013 [2]-LivDet 2015 [3] dataset.	43
4.4	Performance evaluation of AIIearn in comparison to the state-of-the-art [4,5] on LivDet2011 [1] datasets. In this table, FPR, NF and KF denotes false positive rate, new fake and known false, respectively.	44
5.1	Performance evaluation of A-Stacking on class-balanced datasets.	55
5.2	Performance evaluation of A-Bagging on class-balanced datasets.	56
5.3	Performance evaluation of A-Stacking on class imbalanced Biometrika datasets.	56
5.4	Performance evaluation of A-Stacking on class imbalanced DigitalPersona datasets.	57
5.5	Performance evaluation of A-Stacking on class imbalanced ItalData datasets.	58
5.6	Performance evaluation of A-Stacking on class imbalanced Sagem datasets.	59
5.7	Performance evaluation of A-Bagging on class imbalanced Biometrika datasets.	60
5.8	Performance evaluation of A-Bagging on class imbalanced DigitalPersona datasets.	60
5.9	Performance evaluation of A-Bagging on class imbalanced ItalData datasets.	61
5.10	Performance evaluation of A-Bagging on class imbalanced Sagem datasets.	62
6.1	Description of the LivDet datasets used in this study.	68

6.2	Performance evaluation of hand-crafted and deep features in combination with different classifiers under Category-1.	71
6.3	Performance evaluation of hand-crafted and deep features in combination with different classifiers under Category-2. The experiments are performed by considering different sensors for training and testing and viceversa. The average of both experiments is reported.	72
6.4	Performance evaluation of hand-crafted and deep features in combination with different classifiers under Category-3.	73
7.1	Comparison with state-of-the-art.	80
7.2	Description of Datasets.	89
7.3	Performance evaluation of various models under within-dataset environment. (a) Waseem & Hovy, (b) SemEval 2019. The first row for each dataset represents the Micro average and the second represents the Macro average.	92
7.4	Performance evaluation of various models on COVID-19 datasets under within-dataset environment. (a) <i>Covid – Hate_{HL}</i> , (b) <i>Covid – Hate_{ML}</i> . The first row for each dataset represents the Micro average, and the second represents the Macro average.	93
7.5	Performance evaluation of the various models on US presidential election dataset. The first row for each method represents the Micro average and the second represents the Macro average.	94
7.6	Performance evaluation of the models under cross-dataset environment. (a) Train: Waseem & Hovy, Test: SemEval 2019 (b) Train: SemEval2019, Test: Waseem & Hovy. The first row for each dataset represents the Micro average and the second represents the Macro average.	95
7.7	Performance evaluation of the proposed model while considering the user-distribution on Waseem and Hovy dataset. We have restricted the number of tweets per user to 250. The first row for each method represents the Micro average, and the second represents the Macro average.	96
7.8	Performance evaluation of serial and parallel versions of ensemble classifiers under within-dataset environment. The top row for each classifier denotes the weighted average and the bottom one represents the macro average. . .	97

7.9	Performance evaluation of serial and parallel versions of ensemble classifiers under cross-dataset environment. The top row for each classifier denotes the weighted average and the bottom one represents the macro average. . .	98
7.10	Performance evaluation of serial and parallel versions of ensemble classifiers while controlling the user-bias. The top row for each classifier denotes the weighted average and the bottom one represents the macro average. . . .	98

List of Figures

1.1	Schematic Representation of Eager Learning.	2
1.2	Schematic Representation of Lazy Learning.	3
1.3	Conceptual model of EaZy learning.	4
3.1	Visual comparison between live and spoofs created using various spoof materials.	18
3.2	Conceptual Model of Adaptive Ensemble Learning [6].	20
4.1	Schema of the proposed AILearn incremental learning algorithm for Spoof Fingerprint Detection.	35
4.2	Comparison of the performance of AILearn when used with different features shown on Y axis. Percentage Gain on NF and percentage Loss on KF while learning in second phase are shown on X axis.	41
5.1	Conceptual model of A-Stacking.	49
5.2	Conceptual model of A-Bagging.	50
6.1	Accuracy comparison of various handcrafted and deep features under three environments. The accuracy is averaged across various classifiers used in the study.	75
7.1	Schematic diagram of the proposed model.	83
7.2	Schematic representation of the proposed parallelization of ensemble models.	85
7.3	Calculating the percentage drop in F1-score (micro) of various models when tested under cross-dataset environment in comparison with within-dataset environment. The significant drop in performance justifies the need of cross-dataset generalization. rival-1= [7], rival-2= [8]	99

7.4 Performance comparison of various models on Waseem and Hovy dataset under two environments: dataset with no restrictions on the number of tweets and dataset with a cap of 250 tweets per user. We show the F1 score-micro values for this comparison. Proposed model observes no drop. . 99