A FRAMEWORK OF ENSEMBLE ARTIFICIAL NEURAL NETWORKS MODEL FOR CLASSIFICATION OF LARGE DATASETS

MUMTAZIMAH MOHAMAD

DOCTOR OF PHILOSOPHY
UNIVERSITI MALAYSIA TERENGGANU

2014

A framework of ensemble artificial neural networks model for classification of large datasets / Mumtazimah Mohamad.

A FRAMEWORK OF ENSEMBLE ARTIFICIAL NEURAL NETWORKS MODEL FOR CLASSIFICATION OF LARGE DATASETS

MUMTAZIMAH MOHAMAD

Thesis Submitted in Fulfillment of the Requirement for the Degree of Doctor of Philosophy in the School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu
Malaysia

October 2014

DEDICATION

To my loving and supporting husband, Razman Abd Aziz,
my loving children, Adib Rifqi, Deena Nur Mahirah and Adam Rifae
my father, Hj. Mohamad Omar,
&
in memory of my mother, Hjh. Misbah Musa and my mother in law, Hjh. Zalipah Long

Abstract of thesis presented to the Senate of Universiti Malaysia Terengganu in fulfillment of the requirement for the Degree of Doctor of Philosophy

A FRAMEWORK OF ENSEMBLE ARTIFICIAL NEURAL NETWORKS MODEL FOR CLASSIFICATION OF LARGE DATASETS

MUMTAZIMAH MOHAMAD

July 2014

Main Supervisor : **Professor Md. Yazid Mohd Saman, Ph.D.**

Co-Supervisor : **Associate Professor Muhammad Suzuri Hitam, Ph.D.**

School : **Informatics and Applied Mathematics**

Large datasets related tasks are a great challenge for applications that involve massive amounts of data with high dimensional spaces. hence have gained considerable attention in academic research. Due to its enormous size, the datasets poses a more complex problem especially in classification tasks where the datasets have different characteristics. Currently, most machine learning algorithms are able to deal with small to medium sizes of datasets with the use of single memory computer. However, classification of large datasets with the respective algorithms is impractical since it could reduce computer performance and is size-limited for a single processor with one memory. The literature shows that instead of complexity of the procedure, the use of techniques such as feature selection may affect variation, position and scale of the data. Consequently, a classifier has been designed with the aim to reduce the training time in classification task while at the same time preserving the accuracy of the classification results. Based on related literature, it has been proven that parallelization of Artificial Neural Networks (ANN) improves the generalization capability of each classifier of a processor. Therefore, in order to

enhance the performance and capability of ANN, an ensemble framework is proposed to model ANN for large dataset classification tasks.

An original structure model of Multi Layer Perceptron (MLP) was proposed to represent a modelling scheme for parallelization of large datasets and ensemble models. The proposed scheme consists of sequential, parallel and ensemble ANN models that have been evaluated based on performance and complexity. The proposed ensemble framework was constructed based on the characteristics of each model scheme which are inclusive of data pre-processing, reordering technique with partitioning schema, ANN training with stochastic pattern feed as well as output selection ang aggregating. The parallelization part utilizes Message Passing Interface (MPI) as a communication tool in passing data among processors in parallel ANN and ensemble ANN modelling scheme. The relevant models were designed to use composite performance measures such as accuracy, sum square error, training execution as well as speed up and efficiency factors. The performance measurements were designed to ensure only quality models are selected and aggregated as the modelling basis for the ensemble models.

Each model schemes shows the individual characteristics and the degree of ANN capability in classifying large datasets. The parallel ANN model shows good scalability, relevant speedup factor and efficiency. The experimental study shows that the proposed ensemble ANN models are able to improve the accuracy, reduce the training time as well as reduce errors in the classification test cases significantly. An enhanced reordering technique for multiple ANN diversity in ensemble ANN shows 11% improvement of corrected classification and an error reduction of 73%, demonstrated by seven parallelized classifiers. In addition, the reordering technique

has improved the performance of ensemble ANN for large datasets using both selection and combination methods. Hence, both methods contribute better results for different kinds of ensemble outputs, specific network diversity models and for other datasets.

Overall performance of the ensemble ANN model framework shows significant solution for any potential ensemble sizes and methods. ANN implementation for large datasets with ensemble parallel processors has almost no significant difference with the sequential ANN solution in terms of complexity. It was proven that such problem with high number of inputs with multiple classes could be solved with time complexity of $O(n^k)$ for some k which is a type of polynomial. This is in line with the importance of good performance that was achieved with the use of a combiner. The modelling scheme of large datasets in this thesis could be used as a guide for future development of learning techniques involving large datasets. The main contribution of this thesis lies in the utilization of the entire data source as well as proposing an integrated approach to deal with large datasets in classification tasks. In conclusion, the integrated strategies and measures proposed in this thesis could improve the classification performance of large datasets in terms of accuracy as well as speeding up the training.

Abstrak tesis yang dikemukakan kepada Senat Universiti Malaysia Terengganu sebagai memenuhi keperluan untuk Ijazah Doktor Falsafah

KERANGKA KERJA MODEL PENGHIMPUN RANGKAIAN NEURAL BUATAN BAGI PENGELASAN SET DATA BERSKALA BESAR

MUMTAZIMAH MOHAMAD

Julai 2014

Penyelia Utama : **Professor Md. Yazid Mohd Saman, Ph.D.**

Penyelia Utama : **Associate Professor Muhammad Suzuri Hitam, Ph.D.**

Pusat Pengajian : **Informatik dan Matematik Gunaan**

Kerja-kerja berkaitan set data berskala besar merupakan cabaran yang besar terhadap aplikasi-aplikasi yang melibatkan data yang banyak dengan ruang dimensi yang berbilang telah mendapat perhatian dalam penyelidikan akademik. Set data bersaiz besar boleh menyebabkan terjadinya masalah yang lebih kompleks terutamanya bila melibatkan kerja pengelasan di mana data tersebut mempunyai ciri-ciri yang tersendiri. Kebanyakan algoritma pembelajaran mesin kini dilihat mampu untuk mengendalikan set data berskala kecil hingga skala sederhana dengan menggunakan ingatan komputer tunggal. Walau bagaimanapun, pengelasan set data berskala besar mengikut algoritma-algoritma tersebut adalah tidak praktikal di mana ia boleh mengurangkan prestasi komputer dan juga ianya terhad kepada satu pemproses tunggal dengan satu ruang ingatan. Kajian literatur menunjukkan, selain dari kekompleksan prosedur, penggunaan teknik seperti pemilihan ciri juga boleh menjejaskan kepelbagaian, kedudukan dan skala sesuatu data. Oleh itu, suatu pengelas telah dibangunkan dengan tujuan untuk mengurangkan masa latihan kerja

pengelasan dalam masa yang sama memelihara ketepatan keputusan pengelasan. Berdasarkan kajian literatur yang berkaitan, ia telah dapat dibuktikan bahawa penyelarian Rangkaian Neural Buatan (RNB) meningkatkan kebolehan generalisasi setiap pengelas pada setiap pemproses. Dengan itu, untuk menambahbaik kebolehan generalisasi RNB, satu kerangka kerja penghimpun dicadangkan untuk memodelkan RNB bagi kerja-kerja pengelasan set data berskala besar.

Struktur model asas Lapisan Pelbagai Perceptron (LPP) telah dicadangkan sebagai skema pemodelan bagi penyelarian set data berskala-besar dan model penghimpun. Skema yang dicadang terdiri daripada model RNB berjujukan, RNB selari dan RNB penghimpun yang diukur dari sudut prestasi dan kekompleksan. Kerangka kerja penghimpun cadangan telah dibina berasaskan ciri-ciri setiap skema model iaitu termasuk pra-pemprosesan data, teknik penyusunan semula dengan skema pemecahan, latihan RNB dengan suapan corak *stochastic*, pemilihan output dan agregasi output. Penyelarian terlibat menggunapakai *Message Passing Interface* (MPI) sebagai alatan komunikasi untuk menghantar data kepada semua pemproses dalam skema RNB selari dan RNB penghimpun. Model yang relevan telah direkabentuk dengan menggunakan prestasi yang komposit seperti ketepatan, jumlah ralat kuasa dua, pelaksanaan latihan dan juga faktor kelajuan serta kecekapan. Pengukuran prestasi telah dibuat untuk memastikan hanya model-model berkualiti yang akan dipilih dan diagregasikan sebagai asas pemodelan bagi model penghimpun.

Setiap skema model menunjukkan ciri-ciri individu dan aras keupayaan RNB dalam mengelaskan set data berskala besar. Model RNB selari menunjukkan perubahan skala yang baik, faktor kelajuan dan kecekapan yang relevan. Hasil eksperimen

menunjukkan bahawa model penghimpun RNB cadangan dapat meningkatkan ketepatan, mengurangkan masa latihan dan mengurangkan kesilapan kes-kes ujian pengelasan secara signifikan. Teknik penyusunan semula ditambahbaik bagi kepelbagaian RNB dalam model penghimpun menunjukkan peningkatan 11% daripada ketepatan pengelasan dan pengurangan ralat sebanyak 73%, melalui tujuh pengelas yang diselarikan. Selain itu, penambahbaikan pada teknik penyusunan semula telah meningkatkan prestasi penghimpun RNB untuk set data berskala besar dengan menggunakan kedua-dua kaedah pemilihan dan gabungan. Oleh itu, kedua-duanya telah menyumbang keputusan yang lebih baik untuk setiap jenis output penghimpun yang berbeza, model kepelbagaian rangkaian yang terperinci dan untuk set data lain.

Prestasi keseluruhan kerangka kerja model penghimpun RNB menunjukkan ia merupakan penyelesaian yang signifikan dan berpotensi bagi mana-mana saiz dan kaedah penghimpun. Pelaksanaan RNB untuk set data berskala besar dengan pemproses selari penghimpun hampir tidak mempunyai perbezaan yang signifikan dengan penyelesaian RNB jujukan dalam aspek kekompleksan. Ia terbukti apabila sesuatu masalah yang mempunyai banyak bilangan input dengan pelbagai kelas dapat diselesaikan dengan kekompleksan masa $O(n^k)$ untuk pembolehubah k yang berjenis polinomial. Ini adalah selaras dengan kepentingan prestasi yang baik yang telah dicapai dengan penggunaan penghimpun. Skema pemodelan set data berskala besar dalam tesis ini perlu dijadikan panduan kepada pembangunan teknik pembelajaran akan datang yang melibatkan set data berskala besar. Sumbangan utama tesis ini adalah dalam menggunakan keseluruhan sumber data yang yang ada serta menggalakkan pendekatan bersepadu untuk menangani set data berskala besar dalam kerja-kerja pengelasan. Kesimpulannya, dengan pengintegrasian dan penilaian

strategi seperti yang dicadangkan di dalam tesis ini, prestasi pengelasan set data berskala besar boleh ditambahbaik dalam menentukan ketepatan serta meningkatkan prestasi kelajuan latihan.