UNIVERSITI PUTRA MALAYSIA

APPLICATION OF OPTIMIZATION METHODS FOR SOLVING CLUSTERING AND CLASSIFICATION PROBLEMS

PARVANEH SHABANZADEH

IPM 2011 3

APPLICATION OF OPTIMIZATION METHODS FOR SOLVING CLUSTERING AND CLASSIFICATION PROBLEMS

By

PARVANEH SHABANZADEH

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

March 2011

Abstract of thesis presented to the Senate of University Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy.

APPLICATION OF OPTIMIZATION METHODS FOR SOLVING CLUSTERING AND CLASSIFICATION PROBLEMS

By

PARVANEH SHABANZADEH

March 2011

Chairman: Professor. Malik Hj. Abu Hassan, PhD

Faculty: Institute for Mathematical Research

Cluster and classification analysis are very interesting data mining topics that can be applied in many fields. Clustering includes the identification of subsets of the data that are similar. Intuitively, samples within a valid cluster are more similar to each other than they are to a sample belonging to a different cluster. Samples in the same cluster have the same label. The aim of data classification is to set up rules for the classification of some observations that the classes of data are supposed to be known. Here, there is a collection of classes with labels and the problem is to label a new observation or data point belonging to one or more classes of data. The focus of this thesis is on solving clustering and classification problems. Specifically, we will focus on new optimization methods for solving clustering and classification problems. First we briefly give some data analysis background. Then a review of different methods currently available that can be used to solve clustering and classification problems is also given.

Clustering problem is discussed as a problem of non-smooth, non-convex optimization and a new method for solving this optimization problem is developed. This optimization problem has a number of characteristics that make it challenging: it has many local minimum, the optimization variables can be either continuous or categorical, and there are no exact analytical derivatives. In this study we show how to apply a particular class of optimization methods known as pattern search methods to address these challenges. This method does not explicitly use derivatives, and is particularly appropriate when functions are non-smooth. Also a new algorithm for finding the initial point is proposed. We have established that our proposed method can produce excellent results compared to those previously known methods. Results of computational experiments on real data sets present the robustness and advantage of the new method. Next the problem of data classification is studied as a problem of global, non-smooth and non-convex optimization; this approach consists of describing clusters for the given training sets. The data vectors are assigned to the closest cluster and correspondingly to the set, which contains this cluster and an algorithm based on a derivative-free method is applied to the solution of this problem. The proposed method has been tested on real-world datasets. Results of numerical experiments have been presented which demonstrate the effectiveness of the proposed algorithm.

Abstrak tesis untuk dibentangkan kepada Senat Universiti Putra Malaysia bagi memenuhi syarat ijazah Doktor Falsafah.

PELAKSANAAN KAEDAHPENGOPTIMUMAN UNTUK MENYELESAIKAN MASALAH CLUSTERING DAN KLASIFIKASI

Oleh

PARVANEH SHABANZADEH

Mac 2011

Pengerusi: Profesor. Malik Hj. Abu Hassan, PhD

Institiut: Institiut Penyelidikan Matematik

Analisis berkelompok dan pengelasan adalah topik data lombong yang menarik yang boleh digunakan dalam banyak bidang. Berkelompok termasuk pencaman bagi subset untuk data yang serupa. Sampel dalam suatu kelompok yang sah adalah lebih serupa antara satu dengan lain daripada sample kepunyaan suatu kelompok berbeza. Sampel dalam kelompok yang sama mempunyai sama label. Tujuan pengelasan data adalah menyediakan petua untuk pengelasan bagi beberapa pengamatan yang kelas data diandaikan dah tahu. Di sini, ada suatu pungutan kelas dengan label dan masalah adalah melabelkan suatu pengamatan baharu atau titik data dipunyai oleh satu atau lebih kelas data. Tumpuan tesis ini adalah menyelesaikan masalah berkelompok dan penpelasan.Khususnya, kita akan tumpukan ke atas kaedah pengoptimuman baharu untuk menyelesaikan masalah berkelompok dan pengelasan. Mula mula kita berikan beberapa latarbelakang data analisis. Kemudian kita berikan suatu sorotan kaedah

berbeza yang sediada masa kini yang boleh digunakan untuk menyelesaikan masalah berkelompok dan pengelasan. Masalah berkelompok dibincangkan sebagai suatu masalah penngoptimuman tak licin, tak cembung dan sutu kaedah baharu bagi menyelesaikan masalah pengoptimuman ini di bangunkan. Masalah pengoptimuman ini mempunyai suatu bilangan ciri yang membuatkannya mencabar: ia ada banyak minimum setempat, pembolehubah pengoptimuman boleh jadi samada selanjar atau berkategori, dan tiada terbitan beranalitik. Dalam pengajian ini kita tunjukkan bagaimana untuk menggunakan suatu kelas khusus bagi kaedah pengoptimuman yang dikenali sebagai kaedah gelintaran corak untuk mengusulkan cabaran ini. Kaedah ini tidak menggunakan terbitan secara tak tersirat dan adalah sesuai apabila fungsi adalah tak licin. Juga suatu algoritma baharu untuk mencari titik awal dicadangkan. Kita telah buktikan bahawa kaedah baharu yang dicadangkan boleh mengeluarkan keputusan cemerlang berbanding dengan kaedah terdahulu. Keputusan ujikaji pengkomputeran di atas set data nyata mengesahkan keteguhan dan kebaikan bagi kaedah baharu. Kemudian masalah pengelasan data dikaji sebagai masalah pengoptimuman sejagat, tak licin dan tak cembung; pendekatan ini mengandungi kelompok perihalan untuk set latihan yang diberi. Vektor data diumpukkan kepada kelompok terhampir dan berpadanan kepada set yang mengandungi kelompok ini dan algoritma berasaskan ke atas kaedah bebas terbitan digunakan kepada penyelesaian masalah ini. Kaedah yang dicadang telah diuji ke atas set data dunia nyata. Keputusan ujikaji berangka telah dikemukakan dan menunjukkan keberkesanan algoritma yang dicadangkan.

ACKNOWLEDGEMENTS

At first I want to thank Allah for all of things that he has given in my life and then I offer my sincerest gratitude to my chairman, Professor Dr. Malik Hj. Abu Hassan who has supported me throughout my thesis from the initial to the final level with his patience and knowledge whilst allowing me the room to work in my own way. I would like to express my deep and sincere gratitude to my co-supervisor, Dr. Leong Wah June. His wide knowledge and his logical way of thinking have been of great value for me. His encouraging, detailed and constructive comments have enabled me to develop an understanding of the subject. I am also grateful to Dr. Mansor Monsi for serving in the supervisory committee.

I offer my regards to the Head of Institute, academic and general staffs of the Institute for Mathematical Research, University Putra Malaysia, who supported me in any respect during the completion of the project.

I owe great thanks to my precious parents for all things that they gave me or taught me. Without their encouragements, understandings and supports I would never have made any success and also it would have been impossible for me to finish this work. I owe my loving thanks to my dear sister and darling brother, for their loving supports and personal guidance.

I am greatly indebted to my darling law parents for financial and spiritual supporting them without their supports I would never have finished my PhD.

I owe special thanks to my dear's husband Kamyar Shameli and my sweetheart daughter Armita for all their helping in during my study without them I would never able to finish my PhD.

APPROVAL SHEET 1

I certify that Examination Committee has met on date of viva voce to conduct the final examination of Parvaneh Shabanzadeh on her Degree of Doctor of Philosophy thesis entitled "Investigation of Clustering and Classification Problems via Optimization Method" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the student be awarded the Degree of Doctor of Philosophy.

Member of the Examination committee were as follows

FUDZIAH BINTI ISMAIL, PhD
Assoc. Prof
Faculty of Science
Universiti Putra Malaysia
(Chairman)

DATO' MOHAMED SULEIMAN, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

NORIHAN MD. ARIFFIN, PhD
Assoc. Prof
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

DIPTESH GHOSH, PhD
Professor
Indian Institute of Management
Vastrapur, Ahmedabad, Gujarat, India
(External Examiner)

HASANAH MOHD GHAZALI , PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of Supervisory committee were as follows:

Malik B Hj Abu Hassan, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Leong Wah June, PhD
Assoc. Prof
Faculty of Science
Universiti Putra Malaysia
(Member)

Mansor B Monsi, PhD
Science Lecturer
Faculty of Science
Universiti Putra Malaysia
(Member)

HASANAH MOHD GHAZALI , PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I hereby declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and it is not concurrently, submitted for any other degree at University Putra Malaysia or at any institutions.

PARVANEH SHABANZADEH

Date: 29 March 2011

TABLE OF CONTENTS