UNIVERSITI PUTRA MALAYSIA

EXTENDED SPATIAL DECISION TREE ALGORITHM FOR CLASSIFYING HOTSPOT OCCURRENCE

IMAS SUKAESIH SITANGGANG

FSKTM 2013 6

EXTENDED SPATIAL DECISION TREE ALGORITHM FOR CLASSIFYING HOTSPOT OCCURRENCE

By

IMAS SUKAESIH SITANGGANG

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

February 2013

DEDICATIONS

This dissertation is dedicated to the memory of my beloved mother who passed away in 1997. Mom, you have the strongest and the biggest influence in my life. Thank you Mom for always being there for me.

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

EXTENDED SPATIAL DECISION TREE ALGORITHM FOR CLASSIFYING HOTSPOT OCCURRENCE

By

IMAS SUKAESIH SITANGGANG

February 2013

Chair: Razali Yaakob, PhD

Faculty: Computer Science and Information Technology

Forest fire in Riau Province Indonesia is a yearly disaster especially in dry season. It caused many negative effects in various aspects of life for people in Indonesia and neighboring countries including Singapore and Malaysia. In order to minimize the negative effects because of forest fires, classifying hotspots (active fires) occurrence is essential as an activity in fires prevention. The existing methods to classify hotspots occurrence including the logistic regression and the decision tree algorithms do not include spatial objects in the forest fires dataset because these methods are designed for non-spatial dataset. On the other hand, supporting factors for hotspots occurrence are mostly represented in spatial objects. Therefore spatial objects should be included in forest fires datasets for classifying hotspots occurrence in order to obtain the classifiers with high accuracy.

This work proposes a new spatial decision tree algorithm namely the extended spatial ID3 decision tree algorithm to classify hotspots occurrence from a forest fires dataset that contains point, line and polygon features. The method is an extension of the existing spatial decision tree algorithm which works on polygon features only. The proposed algorithm uses spatial information gain to choose the best splitting layer from a set of explanatory layers. The new formula for spatial information gain is proposed using spatial measures for point, line, and polygon features.

The extended spatial ID3 algorithm has been applied to the real forest fires dataset consisting of ten explanatory layers (river, road, city center, land cover, source of income, precipitation in mm/day, screen temperature in K, 10m wind speed in m/s, peatland type, and peatland depth) and a target layer. The target layer consists of true alarm data (hotspots 2008) and false alarm data. The result is a spatial decision tree with 134 leaves with the accuracy 71.12%. After pruning, the spatial decision tree becomes smaller with 122 leaves and its accuracy is 71.66%.

For comparison, classifiers for hotspots occurrence were also developed using the non-spatial methods namely the ID3 algorithm and the C4.5 algorithm as well as the logistic regression. The accuracy of decision tree generated by the ID3 and C4.5 algorithm is 49.02% and 65.24%, respectively. Meanwhile, the accuracy of the logistic regression model is 68.63%. Empirical results using the real spatial forest fires dataset demonstrate that the extended spatial ID3 algorithm has better performance in term of accuracy compared to the non-spatial methods.

The spatial decision tree has been tested using the new dataset on forest fires containing hotspots 2010. The experimental results show that the accuracy of the tree without pruning is 60.06%. Meanwhile, the accuracy of the tree with pruning is 61.89%. The pruned trees do not able to classify about 4.24% objects in the new dataset. These unclassified objects mostly take place in non-peatland areas in which source of income of people living in these areas are forestry and agriculture. Moreover, most of unclassified objects are located in plantation and dryland forest.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

ALGORITMA PEPOHON KEPUTUSAN RUANG DIPERLUAS UNTUK MENGKLASIFIKASIKAN KEJADIAN HOTSPOT

Oleh

IMAS SUKAESIH SITANGGANG

Februari 2013

Pengerusi: Razali Yaakob, PhD

Fakulti: Sains Komputer dan Teknologi Maklumat

Kebakaran hutan di Provinsi Riau Indonesia adalah bencana tahunan terutama di musim kemarau. Bencana ini menyebabkan banyak kesan negatif dalam pelbagai aspek kehidupan bagi penduduk di Indonesia dan negara-negara jiran termasuk Singapura dan Malaysia. Dalam usaha untuk mengurangkan kesan-kesan negatif ini, maka mengklasifikasikan kejadian hotspot adalah penting untuk dibangunkan sebagai satu aktiviti dalam pencegahan kebakaran. Kaedah yang sedia ada untuk mengklasifikasikan kejadian hotspot termasuk regresi logistik dan algoritma pepohon keputusan tidak melibatkan objek ruang dalam dataset kebakaran hutan kerana kaedah-kaedah ini direka untuk dataset bukan ruang. Sebaliknya, faktor sokongan untuk terjadinya hotspot kebanyakannya dinyatakan dalam objek ruang. Oleh kerana itu, objek ruang perlu dimasukkan dalam dataset kebakaran hutan untuk mengklasifikasikan kejadian hotspot dalam usaha untuk mendapatkan pengelas dengan ketepatan yang tinggi.

Kerja ini mencadangkan algoritma pepohon keputusan ruang yang baru iaitu algoritma ID3 pepohon keputusan ruang diperluas untuk membina pengelas untuk mengklasifikasikan kejadian hotspot dari dataset kebakaran hutan yang mengandungi ciri-ciri titik, garisan dan poligon. Kaedah ini adalah lanjutan daripada algoritma pepohon keputusan ruang yang sedia ada yang berguna pada ciri-ciri poligon sahaja. Algoritma yang dicadangkan menggunakan *information gain* ruang untuk memilih lapisan pemisahan terbaik daripada satu set lapisan penjelas. Formula baru untuk *information gain* ruang adalah dicadangkan menggunakan pengukuran ruang untuk ciri-ciri titik, garisan dan poligon.

Algoritma ID3 ruang diperluas telah diaplikasikan untuk dataset kebakaran hutan sebenar yang terdiri daripada sepuluh lapisan penjelas (sungai, jalan, pusat bandar, penutup tanah, sumber pendapatan, pemendakan dalam mm/hari, suhu skrin dalam K, 10m kelajuan angin dalam m/s, jenis tanah gambut, dan kedalaman tanah gambut) dan lapisan sasaran. Lapisan sasaran terdiri daripada data penggera benar (hotspot 2008) dan data penggera palsu. Hasil kajian ialah pepohon keputusan ruang dengan 134 daun dengan ketepatan 71,12%. Selepas pemangkasan, pepohon keputusan ruang menjadi lebih kecil dengan 122 daun dengan ketepatan 71.66%.

Sebagai perbandingan, pengelas untuk mengklasifikasikan kejadian hotspot telah dibangunkan dengan menggunakan kaedah bukan ruang iaitu algoritma ID3 dan C4.5 serta regresi logistik. Ketepatan pepohon keputusan yang dihasilkan oleh algoritma ID3 dan C4.5 masing-masing adalah 49.02% dan 65.24%. Sementara itu, ketepatan model regresi logistik adalah 68.63%. Hasil empirik menggunakan dataset kebakaran hutan sebenar menunjukkan bahawa algoritma ID3 ruang diperluas mempunyai prestasi yang lebih baik dari segi ketepatan berbanding dengan kaedah bukan ruang.

Pepohon keputusan ruang telah diuji menggunakan dataset kebakaran hutan baru yang mengandungi hotspot tahun 2010. Hasil eksperimen menunjukkan bahawa ketepatan pepohon tanpa pemangkasan adalah 60.06%. Sementara itu, ketepatan

pepohon dengan pemangkasan adalah 61.89%. Pepohon yang dipangkas tidak dapat mengklasifikasikan kira-kira 4.24% objek dalam dataset baru. Objek-objek yang tidak dapat dikelaskan kebanyakannya adalah dari tempat di kawasan bukan tanah gambut di mana sumber pendapatan penduduk yang tinggal di kawasan tersebut adalah dari segi perhutanan dan pertanian. Selain itu, kebanyakan objek yang tidak dapat diklasifikasikan terletak di kawasan penanaman dan hutan tanah kering.

ACKNOWLEDGEMENTS

First of all I would like to thank my ALLAH Almighty Who gave me the courage, health, and energy to accomplish this dissertation in due time and without Whose help this study would have not been possible to complete within the time limits.

I would like to express my sincere gratitude to my supervisor Dr. Razali Yaakob and my co-supervisors Associate Professor Dr. Norwati Mustapha and Associate Professor Dr. Ahmad Ainuddin B Nuruddin, for their support, guidance, time, motivation and encouragement throughout duration of this study. Their guidance helped me in conducting this research and writing this dissertation.

My sincere thanks also goes to Indonesia Directorate General of Higher Education (IDGHE), Ministry of National Education, Indonesia for supporting the PhD Scholarship. I also want to express my gratitude to Southeast Asian Regional Center for Graduate Study and Research in Agriculture (SEARCA) for providing the SEARCA PhD Research Scholarship to partially support the research. I am also thankful to my institution Department of Computer Science, Faculty of Natural Science and Mathematics, Bogor Agricultural University that has directly or indirectly contributed to the completion of this research.

I am grateful to the authority of the Meteorological Climatological and Geophysical Agency (BMKG) Indonesia, National Coordinating Agency for Survey and Mapping (BAKOSURTANAL) Indonesia, Wetland International, BPS-Statistics Indonesia, FIRMS MODIS Fire/Hotspot, NASA/University of Maryland, and U.S. Geological Survey for providing data for this research.

Acknowledgement would be incomplete without extending my gratitude to my entire family and friends, especially to my father and my sisters for moral support and prays for my health and successful completion of this dissertation within time limits.

I certify that a Thesis Examination Committee has met on 21st February 2013 to conduct the final examination of **Imas Sukaesih Sitanggang** on her thesis entitled "**Extended Spatial ID3 Decision Tree Model for Classifying Hotspots Occurrence**" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the **Doctor of Philosophy**.

Members of the Thesis Examination Committee were as follows:

Abu Bakar bin Md Sultan, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairperson)

Ali bin Mamat, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Md Nasir bin Sulaiman, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Ajith Abraham
Professor
Faculty of Electrical Engineering and Computer Science
Vsb-Technical University of Ostrava
Czech Republic
(External Examiner)

SEOW HENG FONG, PhD
Professor and Deputy 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 the Supervisory Committee were as follows:

Razali Yaakob, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairperson)

Norwati Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Ahmad Ainuddin bin Nuruddin, PhD
Associate Professor
Faculty of Forestry
Universiti Putra Malaysia
(Member)

BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I 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 is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

IMAS SUKAESIH SITANGGANG

Date: 21 February 2013

TABLE OF CONTENTS

Page

CHAPTER