PSZ 19:16 (Pind. 1/07)

UNIVERSITI TEKNOLOGI MALAYSIA

DECLARATION OF THESIS / UNDERGRADUATE PROJECT PAPER AND COPYRIGHT

Author's full name : _____

Date of birth : _____

Title : _____

_____

_____

Academic Session: _____

I declare that this thesis is classified as :

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*

☐ **RESTRICTED** (Contains restricted information as specified by the organization where research was done)*

☐ **OPEN ACCESS** I agree that my thesis to be published as online open access (full text)

I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:

1. The thesis is the property of Universiti Teknologi Malaysia.
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by :

_____ _____

SIGNATURE **SIGNATURE OF SUPERVISOR**

_____ _____

(NEW IC NO. /PASSPORT NO.) **NAME OF SUPERVISOR**

Date : Date :

NOTES : * If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

"I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Master of Science (Information Assurance)"

Signature	:	
Name of Supervisor	:	Prof. Dr. *Azizah Bt. Abd. Manaf*
Date	:	21st JANUARY 2014.

USING ROUGH SET THEORY FOR CLASSIFICATION OF IMAGE SEGMENTATION DATA

DIALLO ABDOULAYE KINDY

A project submitted in partial fulfilment of the

requirements for the award of the degree of

Master of Science (Information Assurance)

Advanced Informatics School

Universiti Teknologi Malaysia

JANUARY 2014

DECLARATION

I declare that this thesis entitled “Using Rough Set Theory for Classification of Image Segmentation Data” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature:…….....

Name: Diallo Abdoulaye Kindy

Date: 21 January 2014

DEDICATION

To my parents, siblings and everyone who loves taking challenges in his/her life.

ACKNOWLEDGEMENT

All praise be to Allah, All Mighty, the Most Gracious, the Most Merciful. Blessings and peace be upon His beloved messenger Muhammad and also upon his family and his companions till the Day of Resurrection.

My deepest gratitude is extended to Prof. Azizah Abd Manaf, for all her assistance, patience, and invaluable support provided as my supervisor and advisor throughout this entire programme. May Allah reward you for all your kind help. Thanks a lot for being very helpful and a caring mentor.

In preparing this thesis, I was in contact with the main reference of Rough Set in Malaysia, Prof. Dr. Azuraliza Abu Bakar, Deputy Dean (Research And Innovation) Universiti Kebangsaan Malaysia, who co-supervised my work and guide me through this journey. I wish to express my sincere appreciation to her for all her guidance.

I would like also to express my appreciation to my fellow postgraduate students and friends from the International Islamic University Malaysia and National Taiwan University of Science and Technology.

Last but not least, my appreciation to all those who contributed either directly or indirectly to the success of this great attempt.

ABSTRACT

Knowledge Discovery in Database (KDD) can be defined as a technology or a process that helps to extract valuable information including hidden and unseen patterns, trends and relationships between variables from a large amount of data. The information learnt and the discovery made can help in applying the new found pattern in the training set to an unseen data, known as test set, that can guide and facilitate a crucial business decision making task. A large number of data mining techniques have been proposed in the literature for classification purpose. In this work, we are using the Rough Set Classifier (RSC) for mining image segmentation data set obtained from an online machine learning data repository. The RSC is a rule based data mining technique which generates rules from large databases and has great capabilities to deal with noise and uncertainty in data set. In order to find out the best accuracy method, we conducted around 10 experiments by varying the proportions between the training and test sets. The best method gave us an accuracy of 85.71%.

ABSTRAK

Penemuan Pengetahuan dalam Pangkalan Data boleh ditakrifkan sebagai suatu teknologi atau proses yang membantu mengekstrak maklumat berguna termasuklah corak tersembunyi atau terlindung, arah aliran dan hubung kait antara pembolehubah dengan data dalam jumlah yang besar. Penemuan dan maklumat yang dipelajari dapat membantu dalam menerapkan corak data yang baru ditemui ke dalam set latihan. Set ujian pula dapat memberikan panduan dan memudahkan tugas membuat keputusan bisnes yang penting. Bagi tujuan pengklasifikasian, sejumlah besar teknik perlombongan data telah dicadangkan dalam penulisan ini. Dalam kajian ini, kami menggunakan Pengkelas Set Kasar (Rough Set Classifier - RSC) untuk melombong set data bagi segmentasi imej yang diperoleh dari repositori data pembelajaran mesin dalam talian. RSC adalah teknik perlombongan data berasaskan peraturan. Peraturan dijanakan dari pangkalan data yang besar dan mempunyai kebolehan untuk mengedalikan hingar dan ketidaktentuan dalam set data. Bagi mendapatkan teknik yang jitu, kami menjalankan 10 eksperimen dengan mempelbagaikan perkadaran diantara set latihan dan set ujian. Didapati kaedah terbaik menghasilkan kejituan 85.71%.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

RS	Rough Sets
DT	Data mining technique
KDD	Knowledge Discovery in Databases
CT	Computed Tomography
IE	Image Engineering

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

Data mining also known as Knowledge Discovery in Database (KDD) can be defined as a technology or a process that helps to extract valuable information including hidden and unseen patterns, trends and relationships between variables from a large amount of data. The information learnt and the discovery made can help in applying the new found patterns to unseen data that can guide and facilitate a crucial business decision making task. Data mining comprises several tasks such as classification, forecasting, prediction, clustering, association, deviation detection and so on. The definitive objective of data mining, in general, is to facilitate the prediction activity. Predictive knowledge discovery in database (KDD) is the most shared and common type that has the most direct impact in business applications [1]. The data mining process basically comprises of three stages: a) an initial exploration, b) a model design for pattern definition and c) an implementation of the proposed model in the new or unseen data for the prediction purpose.

Our data warehouses are rich with lots of hidden pieces of information that can be very handy for an intelligent decision making. Classification and prediction are two types of data analysis methods that can be used to mine models defining significant data classes or to predict future data tendencies or trends. This type of

analysis offers us a better understanding of the data at large. Classification predicts categorical (discrete, unordered) labels. However, prediction models continuous valued functions [46]. A medical researcher analysing a breast cancer data in order to predict which one of three specific treatments (“treatment A,” “treatment B,” or “treatment C”) a patient should receive, does a classification exercise. Hence he will construct a model or classifier for the prediction phase [46].

Data classification consists of two main tasks: (1) learning step: a classifier is built to describe the set of data classes or concepts of the objects in hand; (2) classification step: we apply what was learnt in step 1 in an unseen data sets and draw conclusion. The classification process is illustrated in the below figure:

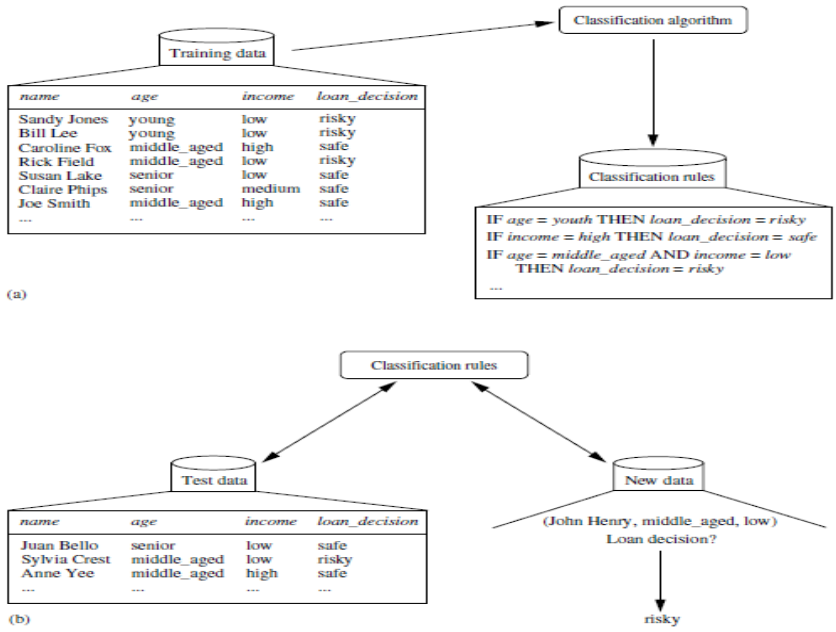

Figure 1.1 Data classification steps [46]

Images are meaningful medium to convey information. The need to understand them and extract information from their content brought in picture image segmentation which fulfills above requirements. Image segmentation, first step of most image analysis procedures, is the process of partitioning an image into multiple segments in order to turn it's representation into meaningful data: easy to understand and analyse as well, without human assistance. Quite a number of intelligent techniques (genetic algorithms, fuzzy sets, artificial neural networks, image processing, pattern recognition, edge-detection, thresholding, clustering, classifiers, region-based, knowledge-based analysis, etc) have been previously applied in this area of study. However, the concern of results accuracy is still remaining.

To tackle the existing issues in image classification and data mining, it become necessary to adopt a new approach which classify segmented images.

1.2 Project Background

Many mathematical theories and a wide range of algorithms have been proposed along the maturity of data mining decision systems in order to improve their impact and help facilitate practitioners' tasks. New algorithms are still being proposed however there is still a lack of solid and reliable machine learning systems. This concern is more noticed when it comes to the imaging field [54]. Managing, massaging and analyzing information data sets and utilizing the data in hand for a sound prediction in the similar issues (problems) in the future has been a hot and challenging area of research for a long time. Information is the most precious asset that we all must protect and use it for learning purpose that could help prevent upcoming/unseen matters.

Information can be analyzed through different means. Its classification is a very important part of business decision making tasks that every board of management and

decision makers rely on. Many decision making exercises are occurrences of classification problem. Hence, in general, most of those tasks can be formulated into a proper classification problem. It could be one of the following: prediction, forecasting, diagnosis, pattern recognition and matching problems. Classification of information can be done either by statistical method or data mining method. Therefore, all these tasks rely on information and depend heavily on uncertain and ambiguous data. These characteristics (uncertainty, rough data, imprecision) are the main basics of the proposed intelligent technique.

Due to above raised concerns; it's high time to bring in picture a new intelligence technique which should stand apart. It should not require any external parameters instead takes full advantage of the only given data. The technique should have the ability to tell whether the working/sample data at hand is complete or not based on the data itself. Moreover, if the data is not sufficient enough, it should suggest that more information about the objects is needed. On the other hand, if the data is complete, it should be in position to determine whether there are any redundancies and find the minimum data required for the classification purpose.

1.3 Problem Statement

There is no doubt that an efficient decision system is fundamental for accurate classification. We are all prone to mistakes and this have an impact in our output too. Many researches have proposed and implemented powerful methods for knowledge discovery in database (KDD) for different types of data sets such as medical data (tumor, cancer, etc.) and engineering data as well. However, a review of a wide range of the existing research works as highlighted in the literature review shows a lack of contribution pertaining to classification of images. This is mainly observed when considering the rough set classifier as a tool for implementation. There are still concerns around proper finding of best methods that can help easily identify the most contributing attributes that defines an image and assist in its

classification. Due to this matter designing decision systems based on a different method, compared to the widely known techniques, that solely classify segmented images is highly needed [54].

1.4 Project Aim

This project generally is meant to mine image knowledge for the purpose of classification. It focuses, mainly, on segmented images related to outdoor scenery. For the above purpose, we will be exploring the capabilities of an intelligent data mining technique known for its strength in dealing with uncertainty and vague data. Lastly, it aims to facilitate the task of mining segmented images which could be applied if successful in medical image knowledge discovery and other related decision support systems.

1.5 Project Objectives

The objectives for this project are threefold:

I. To identify and investigate existing intelligent techniques for image classification.

II. To propose and implement a different method with the hope to improve the classification accuracy.

III. To test and evaluate the proposed intelligent method.

1.6 Project Questions

The project questions are as the following:

i. To what extent we can rely on intelligent techniques for image classification?
ii. How helpful are existing intelligent techniques in dealing with image classification?
iii. What makes our proposed technique stand apart from other approaches in handling image classification?

1.7 Project Scope

The scope for this study focuses on:

i. Classifying outdoor images using intelligent techniques: we will only be using Rough Sets method instead.
ii. We will be only using segmented images extracted from the Image Segmentation Data Set under the UCI Machine Learning publicly available repository [38]. Images types are: RGB.
iii. The tool we will be using is the Rough Set Classifier namely Rosetta.

1.8 Significance of this study

The significance of this work can be outlined as following:

i. Contribution to the data mining research area by exploring the strength of our proposed classification method.
ii. To ease finding the best model within our proposed technique for image classification with less manual work and computation.
iii. Facilitating the task of data mining of images by providing more reliable and fast approaches for image classification.
iv. Taking advantage of the strength of the Rough Set theory in terms of ambiguity and imprecision.

1.9 Organization of the Report

This report consists of 5 chapters. Chapter one describes the introduction, background of the study, research objectives and questions, the scope of the study and its primary objectives. The second chapter reviews available and related literature on data mining methods, the rough set classifier usage for knowledge discovery in database, and image segmentation techniques. Chapter three describes the research methodology. Chapter four deals with the implementation and highlights the findings as well. The last chapter which is chapter five concludes this work and suggests future improvements.

REFERENCES

[1] Lahiri, R. (2006). Comparison of data mining and statistical techniques for classification model (Doctoral dissertation, Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Master of Science in The Department of Information Systems & Decision Sciences by Rochana Lahiri BE, Jadavpur University, India).

[2] Pawlak, Z., "Rough Sets," *Int. Journal of Man-Machine Studies*, *21*, pp. 127-134, 1984

[3] Pawlak Z (1982) Rough sets. Inf J Comput Inf Sci 11(5):341–356

[4] He, F., Wang, X., & Liu, B. (2010, August). Attack detection by rough set theory in recommendation system. In Granular Computing (GrC), 2010 IEEE International Conference on (pp. 692-695). IEEE.

[5] Senthilkumaran, N., & Rajesh, R. (2009). A study on rough set theory for medical image segmentation. International Journal of Recent Trends in Engineering, 2(2), 236-238.

[6] Prasantha, H. S., Shashidhara, H. L., Murthy, K. N. B., & Madhavi, L. G. (2010). Medical image segmentation. International Journal on Computer Science and Engineering, 2(4), 1209-1218.

[7] Schalkoff R.J., Pattern Recognition: Statistical, Structural and Neural Approaches, John Wiley and Sons, 1992.

[8] Lei H., Cheng S.i, Ao M. and Wu Y. "Application of an Improved Genetic Algorithm in Image Segmentation". 2008 International Conference on Computer Science and Software Engineering.

[9]] Milan S., Vaclav H. and Roger B. “Image Processing, Analysis, and Machine Vision”, 2nd edition, Thomas Learning, pp.123-190 (1998)

[10] M. Jirari, Fuzzy Medical Image Segmentation "Presentation-I for Pattern Recognition Class"

[11] Hassanien, A. E., Abraham, A., Peters, J. F., & Schaefer, G. (2009). Rough Sets in Medical Informatics Applications. In *Applications of Soft Computing* (pp. 23-30). Springer Berlin Heidelberg.

[12] Pawlak Z. Rough Sets. Theoretical Aspects of Reasoning About Data. Kluwer, The Netherlands, 1991

[13] Polkowski L. Rough Sets. Mathematical Foundations. Physica-Verlag, Heidelberg, 2003.

[14] Gonzalez RC, Woods RE. Digital image processing. 2nd ed. 2004. Pearson Education

[15] Pal NR, Pal SH. A review on image segmentation techniques. Pattern Recog. 1993;26:1277–94.

[16] Engr. V. C. Chijindu et.al., Medical Image Segmentation Methodologies – A Classified Overview, African Journal of Computing & ICT. ISSN 2006-1781, Vol 5. No. 5, Sept 2012.

[17] Sharma, N., & Aggarwal, L. M. (2010). Automated medical image segmentation techniques. Journal of medical physics/Association of Medical Physicists of India, 35(1), 3.

[18] Rastgarpour, M., & Shanbehzadeh, J. (2011). Application of AI Techniques in Medical Image Segmentation and Novel Categorization of Available Methods and Tools. Proceedings of the International MultiConference of Engineers and Computer Scientists 2011 Vol I, IMECS 2011, March 16-18, 2011, Hong Kong.

[19] Patil, D. D., & Deore, M. S. G. (2013). Medical Image Segmentation: A Review. International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 1, January 2013, pg. 22 - 27.

[20] Zhang, Y. J, An Overview of Image and Video Segmentation in the last 40 years, Proceedings of the 6th International Symposium on Signal Processing and Its Applications, pp. 144-151, 2001.

[21] Kaganami, H. G. and Beij, Z. "Region Based Detection versus Edge Detection", IEEE Transactions on Intelligent information hiding and multimedia signal processing, pp. 1217-1221, 2009.

[22] Sang, C. A. Digital Image Processing (UST 2007 Fall).

[23] Barbu T. A Pattern Recognition Approach to Image Segmentation, Proceedings of the Romanian Academy, Series A, Volume 4, Number 2, pp. 143-148, May-August 2003.

[24] Pawlak, Z. (2001). Rough sets and their applications. Computational Intelligence in Theory and Practice, 8, 73.

[25] Sommerville, I. (2007) Software Engineering (8th ed.) England. Pearson Education Limited.

[26] Guo, Z., Fujioka, S., Blancaflor, E.B., Miao, S., Gou, X. and Li, J. (2010) TCP1 modulates brassinosteroid biosynthesis by regulating the expression of the key biosynthetic gene DWARF4 in Arabidopsis thaliana. Plant Cell, 22, 1161–1173.

[27] Lundström, C. (1997). Segmentation of Medical Image Volumes. (Student paper). Linköpings universitetLinköpings universitet.

[28] Dass, R. Priyanka, and Devi, S. "Image Segmentation Techniques", IJECT Vol.3, Issue 1, Jan-March. 2012

[29] Seerha, G. K. and Kaur, R. "Review on Recent Image Segmentation Techniques" IJCSE 2013; 5(2):109-112. ICID: 1033584. IC™ Value: 3.00

[30] Extracted from http://www.isi.uu.nl/Research/Databases/ on 19/05/2013

[31] Hvidsten, T. R. (2010) "A tutorial-based guide to the ROSETTA system: A Rough Set Toolkit for Analysis of Data" retrieved from http://www.trhvidsten.com/docs/ROSETTATutorial.pdf.

[32] A. Lenarcik, and Z. Piasta, Rough classifiers in Rough Sets, Fuzzy Sets and Knowledge Discovery, W. Ziarko, Ed. London, Springer-Verlag, pp. 298-316, 1994.

[33] Bakar, A.A.; Othman, Z.; Ismail, R.; Abdullah, R.M., "Managing the air cargo offload problems using rough set theory," *Electrical Engineering and Informatics, 2009. ICEEI '09. International Conference on* , vol.01, no., pp.1,6, 5-7 Aug. 2009, doi: 10.1109/ICEEI.2009.5254828.

[34] Zdzisaw P. Chapter 1: Elementary Rough Set Granules: Toward a Rough Set Processor.

[35] A tutorial-based guide to the ROSETTA system: A Rough Set Toolkit for Analysis of Data by Torgeir R. Hvidsten.

[36] Keyun H. , Yuchang L. and Chunyi S. (1982) 'Feature Ranking in Rough Sets', Journal of Computer and Information Sciences, Vo.11, No.5, pp.341-356.

[37] Hassanien, A. E. & Ali, J. M. H, Rough Set Approach for Generation of Classification Rules of Breast Cancer Data, INFORMATICA, 2004, Vol. 15, No. 1, 23–38.

[38] UCI: UCI - Machine Learning Repository (2013). http://archive.ics.uci.edu/ml/. Accessed 01 Sept 2013.

[39] Karthigasoo. S, Cheah Y.N. and Manickam.S: Improving the Accuracy of Medical Decision Support via a Knowledge Discovery Pipeline using Ensemble Techniques. *Journal of Advancing Information and Management Studies, 2(1),* (Pub 2005).

[40] Nguyen H. S. and Skowron A.:Quantization of real-valued attributes. *In Proc. Second International Joint Conference on Information Sciences,* pp. 34–37 (Pub. 1995).

[41] Marzuki, Z and Ahmad, F. Data Mining Discretization Methods and Performances. *Proceedings of the International Conference on Electrical Engineering and Informatics Institut Teknologi Bandung, Indonesia* June 17-19, 2007.

[42] Liu, H. et al. Discretization: An Enabling Technique. Data Mining and Knowledge Discovery, 6,393-423.(Pub 2002).

[43] UCI: UCI - Machine Learning Repository (2013) http://archive.ics.uci.edu/ml/datasets/Image+Segmentation. Accessed 01 Sept 2013.

[44] Bakar, A. A., Othman, Z. A., Hamdan, A. R., Yusof, R., & Ismail, R. (2008). An Agent Based Rough Classifier for Data Mining. Intelligent Systems Design and Applications Eighth International Conference, 145-151. http://dx.doi.org/10.1109/ISDA.2008.29.

[45] Barnaghi, P. M., Sahzabi, V. A., & Bakar, A. A. (2012). A comparative study for various methods of classification. In *International Conference on Information and Computer Networks* (pp. 62-66).

[46] Han, J., Kamber, M., & Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann.

[47] Bakar, A. A., Othman, Z., Ismail, R., & Zakari, Z. (2009, August). Using rough set theory for mining the level of hearing loss diagnosis knowledge. In*Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on* (Vol. 1, pp. 7-11). IEEE.

[48] Rahman, S. A., Bakar, A. A., & Hussein, Z. A. M. (2009, August). Feature selection and classification of protein subfamilies using Rough Sets. In*Electrical Engineering and Informatics, 2009. ICEEI'09. International Conference on* (Vol. 1, pp. 32-35). IEEE.

[49] Kobashi, S. Kondo, K. and Hata, Y. Rough sets based medical image segmentation with connectedness. In 5th Int. Forum on Multimedia and Image Proceeding, pages 197-202, 2004.

[50] Chena, C-B. and Wang, L-Y Rough set-based clustering with refinement using shannon's entropy theory. *Computers and Mathematics with Applications*, 52(10–11):1563–1576, 2006.

[51] Mohabey, A. and Ray, A.K Fusion of rough set theoretic approximations and FCM for color image segmentation. In *IEEE Int. Conference on Systems, Man, and Cybernetics, volume 2, pages 1529–1534, 2000.*

[52] Peters, J.F. and Borkowski, M. K-means indiscernibility relation over pixels. In *Int. Conference on Rough Sets and Current Trends in Computing*, pages 580–585, 2004.

[53] S.Widz, K. Revett, and ´Sle¸zak D. Application of rough set based dynamic parameter optimization to mri segmentation. In 23rd Int. Conference of the North American Fuzzy Information Processing Society, pages 440–445, 2004.

[54]] Srimani P.K and Shanthi Mahes, "Knowledge Discovery in Image-Segmentation Data set Using Decision Tree Classifiers" International Journal of Current Research, Vol.4, Issue, 09, pp.135-140, september, 2012

[55] Schiezaro, M., & Pedrini, H. (2013). Data feature selection based on Artificial Bee Colony algorithm. EURASIP Journal on Image and Video Processing,2013(1), 1-8.

[56] H. R. Arkes, N. W. Dawson, T. Speroff, F. E. Harrel, Jr., C. Alzola, R. Phillips, N. Desbiens, R. K. Oye, W. Knaus, and A. F. Connors, Jr. The covariance decomposition of the probability score and its use in evaluating prognostic estimates. Medical Decision Making, 15:120–131, 1995.

[57] Bakar, A. A., (2013) Lecture Notes about Introduction to Mining Data using Rough Set Theory.

[58] Rosetta, Extracted from http://www.lcb.uu.se/tools/rosetta/ on October 6th, 2013.