MACHINE LEARNING ALGORITHMS: SELECT STUDIES IN CLASSIFICATION, ARM AND THEIR APPLICATIONS

By

PALLATH PAUL VARGHESE

Department of Mathematics

Submitted

in fulfillment of the requirements of the degree

of

Doctor of Philosophy

to the

Indian Institute of Technology, Delhi

New Delhi-110016, India

May, 2008

Machine learning

Certificate

This is to certify that the thesis entitled **Machine Learning Algorithms: Select Studies in Classification, ARM and Their Applications**, which is being submitted by **Pallath Paul Varghese** for the award of the degree of **Doctor of Philosophy in Mathematics** to the **Indian Institute of Technology, Delhi**, is a bonafide research work done under my guidance and supervision.

The thesis has reached the standard fulfilling the requirements of the regulations relating to the degree. The results obtained in the thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

B. chandra

Dr (Mrs.) B.Chandra
Professor
Department of Mathematics
Indian Institute of Technology Delhi
Hauz Khas, New Delhi -110016
India

May, 2008

ACKNOWLEDGEMENTS

"GRATITUDE is the hardest of emotion, and often does not find adequate words to convey that entire, one feels." I express my deepest heartfelt gratitude to my supervisor Prof. B.Chandra for her extraordinary guidance and the amount of time spent in guiding me. Through her patient guidance, constant encouragement and extraordinary guidance, she has brought me closer to the reality. Her enthusiasm, inspiration and in-depth knowledge have provided a good basis for the present thesis. Prof. Chandra has always motivated me to set higher benchmarks and inspired me to look for solutions to problems rather than focusing on problems. She has always been prompt to respond to all my doubts and has always given me opportunity to work in new research areas in Machine Learning and their applications. She has always gone out of the way to help me in many ways. I am falling short of words to express even a fraction of my feelings for the tireless and incredible efforts that Prof. Chandra has put behind me.

I wish to acknowledge the Indian Institute of Technology, Delhi for providing me with outstanding research facilities. I convey my gratitude to all the faculty members of the Department of Mathematics for their help and support during the course of my research work. I take this opportunity to extend my sincere thanks to my SRC members Prof. Wagish Shukla and Prof. M.P.Gupta for their time, advice and suggestions. I would also like to thank the Head of the Department, Prof. Anshul Kumar for all the help and support provided to me.

I would like to thank the management of Nirma University, for encouraging me to pursue my Research at IIT-Delhi. I would also like to express sincere thanks to Head of the Department of Computer Science, Nirma University for strongly recommending me for further studies.

I would like to specially thank Dr. Ravi Kothari, IBM Research Labs for sharing his expertise to enhance my research work. I would also like to thank, Director, SAG, DRDO for sharing his valuable suggestions in the area of unique applications of Machine Learning for Crypt Analysis.

I would like to thank all my friends Niraj Dwivedi, Col .Rajesh Gandhi, Manish Gupta, Ashok Wardhan, Dr. Abhay Gupta for all the enjoyable and pleasant moments during my stay at IIT –Delhi.

Fore-most in my personal mention, I am extremely thankful to my parents for their blessings and prayers because of which I have excelled in my career. I would want to thank my wife for all the sacrifices she has made and without her strong support; I would not have been able to reach to this point. I would like to thank my daughter for being very patient and understanding through my research.

Above all, this piece of work will never be accomplished without our God Almighty with His blessings and His power that work within me and also without the people behind me for inspiring, guiding and accompanying me through thick and thin.

Pallath Paul

ABSTRACT

Machine learning techniques for Data mining can be summarized into various categories like Classification, Clustering and Association rule mining. Main emphasis of this thesis pertains to the area of classification and association rule mining. It also discusses unique application of machine learning algorithms.

Decision tree algorithms form a very effective tool in classification. A decision tree algorithm partitions the given training data in a manner that the resultant partitions have reduced impurity. Node splitting measures play an important role in construction of decision trees for selecting the best splitting attribute and the corresponding split point value.

New node splitting measures have been proposed in this thesis for decision tree induction. One of them is based on the number of distinct classes on either side of the partition after split whereas the other captures the information about the number of sequences of class information in the partitions. Several important properties of the proposed node splitting measure like convexity and well behavedness have been proved for each of the split measures. Benchmark datasets have been used to show the effective performance of the proposed split measures. The results on various datasets show that the proposed measures provide better classification accuracy and are more compact compared to those build using the two of the widely known node splitting measures. The proposed measures have better performance after pruning also.

A novel approach to speed up the process of decision tree construction has also been presented in the thesis. Here the node splitting measure is evaluated only at class

boundaries for selecting the best splitting attribute. It has been shown that on an average only 50% of the split points need to be evaluated to construct the decision tree resulting in significant reduction in time taken to construct the decision tree.

Fuzzy Decision Tree algorithm to fuzzify the binary SLIQ decision tree algorithm based on Gini Index is proposed in this thesis to overcome the sharp boundary problems existing in crisp decision trees. A new membership function is designed to fuzzify the decision boundary at the split point. More than 74% reduction in size is observed in the decision trees built using the proposed fuzzy decision tree algorithm with increased classification accuracy.

An alternate approach to Association rule mining for instant rule generation is also proposed in the thesis. A weighted directed graph is constructed with edges between adjacent nodes only if the two itemsets have confidence greater than minimum confidence. It is shown that the time required for generating all the essential rules does not increase significantly with the increase in the number of rules generated as is the case with the existing graph based association rule mining algorithms.

Unique applications of Machine learning algorithms in the field of crypt analysis are discussed in the thesis. It is shown how Machine learning algorithms can be effectively employed in the area of crypt analysis for better understanding of crypto systems. The results carried out so far gives a very good basis to consider that Machine learning algorithms are an effective tool for analyzing Crypto Systems. This has been explored for the first time in the literature.

CONTENTS