CHAPTER 11

COMPUTER SCIENCE

Doctoral Theses

01. AGGARWAL (Geeta)
Combining Bi-clustering Solutions for Gene Expression Data.
Supervisor : Dr. Neelima Gupta
Th 23742

Contents

1. Introduction 2. Biological overview 3.Preliminaries and related work 4.BiETopti-Biclustering ensemble technique using optimization 5. BiETclassiBiclustering ensemble technique using classifiers 6.BiETmetaclus-Biclustering ensemble technique using metaclustering 7. Concluding Remarks. Bibliography.

02. BHARDWAJ (Manju)
Design and Comparative Analysis of Classification Ensembles.
Supervisor : Dr. Vasudha Bhatnagar and Prof. Debasis Dash
Th 23375

Abstract
(Verified)

This thesis covers two important facets of ensemble methodology – ensemble construction and ensemble pruning. For each of the two, an algorithm has been proposed and validated with experimental results. Both the algorithms use majority voting as the combiner function. Towards the end, a novel function is proposed in the thesis for comparison of pruned ensembles. A novel algorithm for SVM-based ensemble generation called Self-Chastising Ensemble (SCE) is presented. Inspired by Adaboost, the algorithm follows "learning-from-mistakes" paradigm and generates dependent classifiers. Each classifier attempts to rectify the mistakes made by the previous one. The algorithm accumulates the unbounded support vectors from the generated models and uses them as representatives of correctly classified instances. Undersampling of mistakes is done to ensure an equal mix of mistakes and unbounded support vectors. Novelty of the method lies in the way unbounded support vectors have been used along with mistakes to aid learning of difficult instances. Next, an accuracy oriented Ensemble pruning algorithm- Accu-Prune (AP) has been presented in the thesis. AP algorithm exploits the accuracy of individual base classifiers to implicitly achieve diversity in the pruned ensemble. It has been theoretically justified that the diversity of ensemble increases on addition of a pair of low accuracy classifiers. Given an accuracy ordered pool of classifiers, AP uses multipronged hill-climbing and reduced error pruning to discover the optimal ensemble. Lastly, we propose an objective function called Accrual function to quantify the difference in performance and size of two compared ensembles, to gauge their relative cost-effectiveness. The function is continuos, differentiable and can be parameterized to meet the user specific need for according priority to either size or performance. In our opinion, this is the most significant contribution of the thesis.

Contents

1. Introduction 2. Classification ensembles 3.Self chastising Ensembles 4. Learning from imbalanced data 5.Towards an optimally pruned classifier ensemble 6.Cost-

effectiveness of classification ensembles 7. Conclusion and future work. Result tables from selected publication. Bibliography.

03. GAUTAM(Anjali)
Matrix Factorisation Based Recommender System for Large-Scale Data.
Supervisor : Prof. Poonam Bedi
Th 23373

Abstract
(NotVerified)

Advent of e-commerce has contributed data in large volumes and variety i.e. large-scale data leading to the problem of information overload posing users with challenge of finding relevant items from this vast pool. Recommender systems (RSs) play a vital role in assisting users to overcome this challenge by helping users to find relevant items which are in accordance with the tastes and preferences of users. Generating recommendations from large-scale data i.e. news data press the recommendation technique to be able to generate timely and accurate recommendations which this research work aims at. This work proposes MapReduce Content-Based Recommendation (MRCBR) technique implemented on Apache Hadoop, which uses vector space model to extract document features and to create user profile followed by computing similarity between the two to generate recommendations. User tends to behave differently in different situations. Generating recommendations by taking into account situational information about the user i.e. user's context (such as time, location) will result in improved quality of recommendations. The work proposes a Content Boosted Context-Aware RS (CBCARS) by incorporating item content and user context resulting in sparse user-item-context rating tensor. Recommendations for users are generated using Tensor Factorization (TF). TF models the sparse user-item-context rating tensor by mapping users, items and context into a subspace of latent factors which explains observed ratings by recording hidden properties of users, items and context. To improve the computational efficiency of generating context-aware recommendations from large-scale data TF is implemented on scalable and distributed framework of Apache Spark. The proposed approach also takes into account both the explicit and the implicit user feedback. To generate quality recommendations, traditional CF technique exploits notion of similarity inducing monotony in recommendation list. Monotony in the list is avoided by introducing diversity in recommendation list by developing a cross-domain RS using matrix factorization

Contents

04. GOEL (Nidhi)
Quality estimation of Tomatoes Using Machine Vision: A Multi-Modal Image Retrieval Perspective.
Supervisor : Dr. Priti Sehgal
Th 23370

Abstract
(Verified)

The work proposed is an attempt to exploit machine vision using soft computing techniques in the realm of agriculture industry. Quality estimation techniques using machine vision that assess the quality of tomatoes in terms of ripeness stage and firmness class have been proposed. A technique is presented named as fuzzy rule based classification (FRBC-R) for estimating ripeness of preharvest tomatoes based on color. The proposed system considers natural lighting conditions and does not interfere in the usual growth of tomatoes. Prediction model for firmness estimation is built using regression analysis. Grounded on the prediction model, fuzzy rule based classification (FRBC-F) is proposed to classify tomatoes based on their predicted firmness into three categories soft, medium, and hard. Another contribution of this research is the proposal of techniques for image retrieval that allow multimodal retrieval (MMR) of images based on text and content whileunderstanding user's intention behind the search. While developing these techniques, variouschallenges in MMR have been addressed. Finally, the dissertation presents an application wherethe developed MMR and proposed machine vision techniques are mutually exploited to extract,automatically compute cognitive description, and retrieve tomato images based on their quality.*Outcomes of this research are distinctive as proposed system takes viability of digital camera todetermine the quality of pre-harvest tomatoes and whilst add flavor of multimodal image retrieval byautomatically assigning cognitive description to images.*

Contents

05. GUPTA(Shikha nee shikha Agarwal)
Community Detection in Social Networks: A Quantum – inspired evolutionary approach.
Supervisor : Prof. Naveen Kumar
Th 23369

Abstract
(Verified)

A social network may be abstracted as a graph whose nodes denote the individuals and the edges denote the relationships between pairs of nodes. In this thesis, we examine the problem of detecting disjoint communities in a social network as an optimization problem with the objective of maximizing the network modularity, a measure of the goodness of the communities formed. We propose three community detection algorithms, employing variants of well-known quantum-inspired evolutionary algorithm (QIEA). Like any other evolutionary algorithm, a QIEA is also characterized by the representation of the individual, the evaluation function, and the population dynamics. However, individual bits called qubits, are in a superposition of states. As chromosomes evolve individually, the quantum-inspired evolutionary algorithms (QIEAs) are intrinsically suitable for parallelization. Our first algorithm is a hierarchical bi-partitioning algorithm. The algorithm begins with local search to capture the local variations in the density of the social network, followed by a search for communities on the reduced network structure. We next propose a numeric variant of the standard quantum-inspired evolutionary algorithm. In a numeric observation QIEA, the classical chromosome is represented as an array of numeric values and each element of the quantum chromosome is a superposition of k qubits. We have

also proposed parallel implementations of these algorithms on CUDA-enabled GPUs, employing a single-population fine-grained approach that is suited for massively parallel computers. In this approach, each element of a chromosome is assigned to a separate thread. The parallel implementations achieve significant speedup, and due to their highly parallel nature, an increase in the number of multiprocessors and GPU devices may lead to a further speedup. Finally, we propose an algorithm for multidimensional networks that allows heterogeneous links between the nodes. The proposed algorithms are able to detect high-quality communities and work well for both weighted and unweighted networks.

Contents

06. SHARMA(Meera)
Developing Prediction Model to Assist Software Developers and Support Managers.
Supervisor : Dr. V. B. Singh
Th 23371

Abstract
(NotVerified)

Software repositories, such as bug repositories, source control repositories, archived communications, deployment logs and code repositories contain a rich and detailed information about the evolution of a software project. These repositories help software developers/triagers in the bug fixing process and software managers in the maintenance and evolution of the software products. Reporting a bug requires several attributes to be filled at the time of bug submission in the form of bug report.Bug report data is useful in various bug attributes prediction, namely bug priority, severity, CC list, fix time and assignee. Cross project validation is an important concern in empirical software engineering where we train classifier on one project and test it for prediction on other projects in the absence of historical data for training. A clear understanding of bug attributes, their interdependence and their contribution in predicting the other attributes will help in improving the quality of software.It is the need of the hour to discover the association of bug fix time and assignee with other bug attributes.Open source software is evolved through an active participation of the users in terms of reporting of bugs, request for new features and feature improvements. The code changes done in source code files due to these issues fixing increase the complexity of code which is used to predict the possible code changes in the software over a long run (potential complexity of code changes).A software is upgraded with the inclusion of new features and improvements in existing features. Some of the issues remain unresolved in the current release. These unresolved/leftover issues are added to the initial issue content of the next release and get fixed in subsequent releases. For the release time problem, the source code changes and left over issues of the previous release can be considered

Contents

models 8. Release time planning in multi-release software 9. Conclusion. Appendix. References. List of publications

07. THAKRAL(Sonika)
Approximation Algorithms For Data Placement Problem.
Supervisor : Prof. Neeliam Gupta
Th 23372

Abstract
(NotVerified)

Data placement problems deal with cost effective placement of data on servers in order to serve a given set of clients. The cost function may include different parameters, such as the cost of placing data on servers or the total sum of distances" between the clients and the servers they are assigned to. We address two types of data placement problems under capacity constraints - the Replica Placement problem and the typed Data Placement problem. The two problems differ in the notion of capacity. Whereas in the Replica Placement problem capacity defines the number of clients that can be served by one server, in the typed Data Placement problem capacity indicates the maximum number of services that any server may offer. We study variants of these problems that are NP hard and present LP rounding based constant factor approximation algorithms for them. We study the following variants of the replica placement problem: • On tree graphs with unit-length edges; we present a polynomial time O(1) approximation algorithm for this variant. We extend our techniques for graphs having bounded tree-width and present an O(t) approximation algorithm where t is the tree-width of the graph. • On graphs with arbitrary edge lengths, having bounded degree and bounded tree-width (called BDBT graphs); we present polynomial time O(d + t) approximation algorithm for this variant where d and t respectively denote the degree and tree-width of the graph. • On a generalization of BDBT graphs wherein BDBT graphs are connected in a tree-like manner; we call such graphs Trees of Bounded Degree Bounded Tree-width graphs (TBDBT graphs). We present O(d+t) approximation algorithm for this variant where d and t respectively denote the degree and tree-width of any component BDBT graph. For the typed Data Placement problem, we study the variant having two different service types and no facility opening costs; we present a polynomial time 4- approximation algorithm for this variant.

Contents