K Nearest Neighbor Algorithm For Classification

Nearest Neighbor algorithms are non-parametric algorithms that use distance measure techniques for classification and regressions. This thesis uses the method of pruning to improve accuracy and efficiency of a nearest neighbor classifier and also states the different stages the pruning algorithm can be applied and shows the best stage for pruning which gives the maximum accuracy. The performance of the classifier is shown to be better than other improved nearest neighbor classifiers. A fast method of finding the optimal k in a k-nearest neighbor classifier is proposed in the thesis. A method of optimizing the distance measure using a second order training algorithm in a k-nearest neighbor algorithm is also proposed in this thesis which results to better accuracy than the traditional k-nearest neighbor classifier.

The Encyclopedia of GIS provides a comprehensive and authoritative guide, contributed by experts and peer-reviewed for accuracy, and alphabetically arranged for convenient access. The entries explain key software and processes used by geographers and computational scientists. Major overviews are provided for nearly 200 topics: Geoinformatics, Spatial Cognition, and Location-Based Services and more. Shorter entries define specific terms and concepts. The reference will be published as a print volume with abundant black and white art, and simultaneously as an XML online reference with hyperlinked citations, cross-references, four-color art, links to web-based maps, and other interactive features.

This textbook introduces fundamental concepts, major models, and popular applications of pattern recognition for a one-semester undergraduate course. To ensure student understanding, the text focuses on a relatively small number of core concepts with an abundance of illustrations and examples. Concepts are reinforced with hands-on exercises to nurture the student's skill in problem solving. New concepts and algorithms are framed by real-world context and established as part of the big picture introduced in an early chapter. A problem-solving strategy is employed in several chapters to equip students with an approach for new problems in pattern recognition. This text also points out common errors that a new player in pattern recognition may encounter, and fosters the ability for readers to find useful resources and independently solve a new pattern recognition task through various working examples. Students with an undergraduate understanding of mathematical analysis, linear algebra, and probability will be well prepared to master the concepts and mathematical analysis presented here.

In this carefully edited book some selected results of theoretical and applied research in the field of broadly perceived intelligent systems are presented. The problems vary from industrial to web and problem independent applications. All this is united under the slogan: "Intelligent systems conquer the world”. The book brings together innovation projects with analytical research, invention, retrieval and processing of knowledge and logical applications in technology. This book is aiming to a wide circle of readers and particularly to the young generation of IT/ICT experts who will build the next generations of intelligent systems.

This book provides an overview of the current state of the art in wireless networks around the globe, focusing on utilizing the latest artificial intelligence and soft computing techniques to provide design frameworks for wireless networks. These techniques play

a vital role in developing a more robust algorithm suitable for the dynamic and heterogeneous environment, making the network self-managed, self-operational, and self-configurational, and efficiently reducing uncertainties and imprecise information.
The field of data mining lies at the confluence of predictive analytics, statistical analysis, and business intelligence. Due to the ever-increasing complexity and size of data sets and the wide range of applications in computer science, business, and health care, the process of discovering knowledge in data is more relevant than ever before. This book provides the tools needed to thrive in today's big data world. The author demonstrates how to leverage a company's existing databases to increase profits and market share, and carefully explains the most current data science methods and techniques. The reader will "learn data mining by doing data mining". By adding chapters on data modelling preparation, imputation of missing data, and multivariate statistical analysis, Discovering Knowledge in Data, Second Edition remains the eminent reference on data mining. The second edition of a highly praised, successful reference on data mining, with thorough coverage of big data applications, predictive analytics, and statistical analysis. Includes new chapters on Multivariate Statistics, Preparing to Model the Data, and Imputation of Missing Data, and an Appendix on Data Summarization and Visualization Offers extensive coverage of the R statistical programming language Contains 280 end-of-chapter exercises Includes a companion website for university instructors who adopt the book
The past decade has seen greatly increased interaction between theoretical work in neuroscience, cognitive science and information processing, and experimental work requiring sophisticated computational modeling. The 152 contributions in NIPS 8 focus on a wide variety of algorithms and architectures for both supervised and unsupervised learning. They are divided into nine parts: Cognitive Science, Neuroscience, Theory, Algorithms and Architectures, Implementations, Speech and Signal Processing, Vision, Applications, and Control. Chapters describe how neuroscientists and cognitive scientists use computational models of neural systems to test hypotheses and generate predictions to guide their work. This work includes models of how networks in the owl brainstem could be trained for complex localization function, how cellular activity may underlie rat navigation, how cholinergic modulation may regulate cortical reorganization, and how damage to parietal cortex may result in neglect. Additional work concerns development of theoretical techniques important for understanding the dynamics of neural systems, including formation of cortical maps, analysis of recurrent networks, and analysis of self- supervised learning. Chapters also describe how engineers and computer scientists have approached problems of pattern recognition or speech recognition using computational architectures inspired by the interaction of populations of neurons within the brain. Examples are new neural network models that have been applied to classical problems, including handwritten character recognition and object recognition, and exciting new work that focuses on building electronic hardware modeled after neural systems. A Bradford Book
k-nearest neighbors (k-NN), which is known to be a simple and ef?cient approach, is a non-parametric supervised classi?er. It aims to determine the class label of an unknown sample by its k-nearest neighbors that are stored in a training set. The k-nearest neighbors are determined based on some distance functions. Although k-NN produces successful results, there have been some extensions for improving its precision. The

neutrosophic set (NS) de?nes three memberships namely T, I and F. T, I, and F shows the truth membership degree, the false membership degree, and the indeterminacy membership degree, respectively. In this paper, the NS memberships are adopted to improve the classi?cation performance of the k-NN classi?er.

Space support in databases poses new challenges in every part of a database management system & the capability of spatial support in the physical layer is considered very important. This has led to the design of spatial access methods to enable the effective & efficient management of spatial objects. R-trees have a simplicity of structure & together with their resemblance to the B-tree, allow developers to incorporate them easily into existing database management systems for the support of spatial query processing. This book provides an extensive survey of the R-tree evolution, studying the applicability of the structure & its variations to efficient query processing, accurate proposed cost models, & implementation issues like concurrency control and parallelism. Written for database researchers, designers & programmers as well as graduate students, this comprehensive monograph will be a welcome addition to the field.

This book constitutes the refereed proceedings of the Second EAI International Conference on Advanced Hybrid Information Processing, ADHIP 2018, held in Yiyang, China, in October 2018. The 71 papers presented were selected from 228 submissions and focus on hybrid big data processing. Since information processing has acted as an important research domain in science and technology today, it is the right time to develop deeper and wider use of hybrid information processing, especially information processing for big data. There are more remaining issues waiting for solving, such as classification and systemization of big data, objective tracking and behavior understanding in big multimedia data, encoding and compression of big data.

This book features high-quality research papers presented at the International Conference on Advanced Computing and Intelligent Engineering (ICACIE 2017). It includes sections describing technical advances in the fields of advanced computing and intelligent engineering, which are based on the presented articles. Intended for postgraduate students and researchers working in the discipline of computer science and engineering, the proceedings also appeal to researchers in the domain of electronics as it covers hardware technologies and future communication technologies.

Publisher Description

This book constitutes the refereed proceedings of the 5th International Conference on Intelligent Data Engineering and Automated Learning, IDEAL 2004, held in Exeter, UK, in August 2004. The 124 revised full papers presented were carefully reviewed and selected from 272 submissions. The papers are organized in topical sections on bioinformatics, data mining and knowledge engineering, learning algorithms and systems, financial engineering, and agent technologies.

Managing Information Technology Resources in Organizations in the Next Millennium contains more than 200 unique perspectives on numerous timely issues of managing information technology in organizations around the world. This book, featuring the latest research and applied IT practices, is a valuable source in support of teaching and research agendas.

Machine learning is the study of algorithms that automatically improve their performance with experience. That can provide signi cant competitive

advantages to many organizations by exploiting the potential of large data volume. Intelligently analyzed data is a valuable resource. At the heart of performance is classi cation accuracy in this speci ed task. Mostly a crucial problem in machine learning is identifying a representative set of features from which to construct a classi cation model for a particular task. The classi cation of data is based on the set of data feature used. The feature selection can provide optimizing performance by using Genetic Algorithm and strongly effect in classi cation. In making to get improved classi cation accuracy, author has taken advantage with using hybrid of machine learning methods rather than use of only machine learning approach. Author proposed Information based distance metric to overwhelm one of the weak points of k nearest neighbor classi er. Moreover it provide the comparison of the result of Information based distance metric and Euclidean distance metric on both majority voting and similarity score summing."

This book constitutes the refereed proceedings of the First International Conference on Human.Society§Internet, held in Seoul, Korea, in July 2001. The 32 revised full papers presented together with 4 invited papers were carefully reviewed and selected from a total of 85 submissions. The papers are organized in topical sections on digital economy, electronic commerce, digital divide, Internet status and new applications, virtual enterprises, cyber education, digital governance, medical computing, mobile computing, and human computing.

The k-nearest neighbor (k-NN) pattern classifier is a simple yet effective learner. However, it has a few drawbacks, one of which is the large model size. There are a number of algorithms that are able to condense the model size of the k-NN classifier at the expense of accuracy. Boosting is therefore desirable for increasing the accuracy of these condensed models. Unfortunately, there does not exist a boosting algorithm that works well with k-NN directly. We present a direct boosting algorithm for the k-NN classifier that creates an ensemble of models with locally modified distance weighting. An empirical study conducted on 10 standard databases from the UCI repository shows that this new Boosted k-NN algorithm has increased generalization accuracy in the majority of the datasets and never performs worse than standard k-NN.

K-Nearest-Neighbors (KNN) search is a fundamental problem in many application domains such as database and data mining, information retrieval, machine learning, pattern recognition and plagiarism detection. Locality sensitive hash (LSH) is so far the most practical approximate KNN search algorithm for high dimensional data. Algorithms such as Multi-Probe LSH and LSH-Forest improve upon the basic LSH algorithm by varying hash bucket size dynamically at query time, so these two algorithms can answer different KNN queries adaptively. However, these two algorithms need a data access post-processing step after candidates' collection in order to get the final answer to the KNN query. In this thesis, Multi-Probe LSH with data access post-processing (Multi-Probe LSH with DAPP) algorithm and LSH-Forest with data access post-processing (LSH-Forest with DAPP) algorithm are improved by replacing the costly data

access post-processing (DAPP) step with a much faster histogram-based post-processing (HBPP). Two HBPP algorithms: LSH-Forest with HBPP and Multi-Probe LSH with HBPP are presented in this thesis, both of them achieve the three goals for KNN search in large scale high dimensional data set: high search quality, high time efficiency, high space efficiency. None of the previous KNN algorithms can achieve all three goals. More specifically, it is shown that HBPP algorithms can always achieve high search quality (as good as LSH-Forest with DAPP and Multi-Probe LSH with DAPP) with much less time cost (one to several orders of magnitude speedup) and same memory usage. It is also shown that with almost same time cost and memory usage, HBPP algorithms can always achieve better search quality than LSH-Forest with random pick (LSH-Forest with RP) and Multi-Probe LSH with random pick (Multi-Probe LSH with RP). Moreover, to achieve a very high search quality, Multi-Probe with HBPP is always a better choice than LSH-Forest with HBPP, regardless of the distribution, size and dimension number of the data set.

With the increasing popularization of the Internet, together with the rapid development of 3D scanning technologies and modeling tools, 3D model databases have become more and more common in fields such as biology, chemistry, archaeology and geography. People can distribute their own 3D works over the Internet, search and download 3D model data, and also carry out electronic trade over the Internet. However, some serious issues are related to this as follows: (1) How to efficiently transmit and store huge 3D model data with limited bandwidth and storage capacity; (2) How to prevent 3D works from being pirated and tampered with; (3) How to search for the desired 3D models in huge multimedia databases. This book is devoted to partially solving the above issues.

Compression is useful because it helps reduce the consumption of expensive resources, such as hard disk space and transmission bandwidth. On the downside, compressed data must be decompressed to be used, and this extra processing may be detrimental to some applications. 3D polygonal mesh (with geometry, color, normal vector and texture coordinate information), as a common surface representation, is now heavily used in various multimedia applications such as computer games, animations and simulation applications. To maintain a convincing level of realism, many applications require highly detailed mesh models. However, such complex models demand broad network bandwidth and much storage capacity to transmit and store. To address these problems, 3D mesh compression is essential for reducing the size of 3D model representation.

The book provides foundations of machine learning and algorithms with a road map to deep learning, genesis of machine learning, installation of Python, supervised machine learning algorithms and implementations in Python or R, unsupervised machine learning algorithms in Python or R including natural language processing techniques and algorithms, Bayesian statistics, origins of deep learning, neural networks, and all the deep learning algorithms with some implementations in TensorFlow and architectures, installation of TensorFlow,

neural net implementations in TensorFlow, Amazon ecosystem for machine learning, swarm intelligence, machine learning algorithms, in-memory computing, genetic algorithms, real-world research projects with supercomputers, deep learning frameworks with Intel deep learning platform, Nvidia deep learning frameworks, IBM PowerAI deep learning frameworks, H2O AI deep learning framework, HPC with deep learning frameworks, GPUs and CPUs, memory architectures, history of supercomputing, infrastructure for supercomputing, installation of Hadoop on Linux operating system, design considerations, e-Therapeutics's big data project, infrastructure for in-memory data fabric Hadoop, healthcare and best practices for data strategies, R, architectures, NoSQL databases, HPC with parallel computing, MPI for data science and HPC, and JupyterLab for HPC.

This book constitutes the refereed proceedings of the Third International Conference on Advances in Visual Informatics, IVIC 2013, held in Selangor, Malaysia, in November 2013. The four keynotes and 69 papers presented were carefully reviewed and selected from various submissions. The papers focus on four tracks: computer visions and engineering; computer graphics and simulation; virtual and augmented reality; and visualization and social computing.

High-throughput sequencing and functional genomics technologies have given us a draft human genome sequence and have enabled large-scale genotyping and gene expression profiling of human populations. Databases containing large number of sequences, polymorphisms, and gene expression profiles of normal and diseased tissues in different clinical states are rapidly being generated for human and model organisms. Bioinformatics is thus rapidly growing in importance in the annotation of genomic sequences, in the understanding of the interplay between genes and proteins, in the analysis the genetic variability of species, etc. The 3rd APBC brings together researchers, professionals, and industrial practitioners for interaction and exchange of knowledge and ideas. The proceedings contains the latest results that address conceptual and practical issues of bioinformatics. Papers presented at APBC'05 and included in this proceedings volume span the following: Novel Applications in Bioinformatics, Computational Analysis of Biological Data, Data Mining & Statistical Modeling of Biological Data, Modeling and Simulation of Biological Processes, Visualization of Biological Processes and Data, Management, Migration, and Integration of Biological Databases, Access, Indexing, and Search in Biological Databases. Contents:A Better Gap Penalty for Pairwise-SVM (H N Chua & W-K Sung)A Graph Database with Visual Queries for Genomics (G Butler et al.)Consensus Fold Recognition by Predicted Model Quality (J Xu et al.)Toward Discovering Disease-Specific Gene Networks from Online Literature (Z Zhang et al.)Hybrid Registration for Two-Dimensional Gel Protein Images (X Wang & D D Feng)Exact Algorithms for Motif Search (S Rajasekaran et al.)Voting Algorithms for Discovering Long Motifs (F Y L Chin & H C M Leung)A Highly Scalable Algorithm for the Extraction of Cis-Regulatory Regions (A M Carvalho et al.)A Support Vector Machine Approach for Prediction of T Cell Epitopes (L Huang & Y Dai)Protein Informations Towards Integration of Data Grid and Computing Grid (H Nakamura)Computing the Assignment of Orthologous Genes via Genome Rearrangement (X Chen et al.)and other papers Readership: Computational biologists, bioinformaticists, computer scientists, biologists. Keywords:Bioinformatics;Computational Biology;Datamining;Biological Data;Modeling;Visualization;Database Management;Database Integration;Biological Database Indexing;Biological Database Search

Master's Thesis from the year 2012 in the subject Computer Science - Didactics, , course: COMPUTER SCIENCE & ENGINEERING, language: English, abstract: During the last years, semi-supervised learning has emerged as an exciting new direction in machine learning research. It is closely related to profound issues of how to do inference from data, as witnessed by its overlap with transductive inference. Semi-Supervised learning is the half-way between Supervised and Unsupervised Learning. In this majority of the patterns are unlabelled, they are present in Test set and knowed labeled patterns are present in Training set. Using these training set, we assign the labels for test set. Here our Proposed method is using Nearest Neighbour Classifier for Semi-Supervised learning we can label the unlabelled patterns using the labeled patterns and then compare these method with the traditionally Existing methods as graph mincut, spectral graph partisan, ID3,Nearest Neighbour Classifier and we are going to prove our Proposed method is more scalable than the Existing methods and reduce time complexity of SITNNC(Selective Incremental Approach for Transductive Nearest Neighbour Classifier) using Leaders Algorithm.

Artificial Intelligence for Drug Development, Precision Medicine, and Healthcare covers exciting developments at the intersection of computer science and statistics. While much of machine-learning is statistics-based, achievements in deep learning for image and language processing rely on computer science’s use of big data. Aimed at those with a statistical background who want to use their strengths in pursuing AI research, the book: · Covers broad AI topics in drug development, precision medicine, and healthcare. · Elaborates on supervised, unsupervised, reinforcement, and evolutionary learning methods. · Introduces the similarity principle and related AI methods for both big and small data problems. · Offers a balance of statistical and algorithm-based approaches to AI. · Provides examples and real-world applications with hands-on R code. · Suggests the path forward for AI in medicine and artificial general intelligence. As well as covering the history of AI and the innovative ideas, methodologies and software implementation of the field, the book offers a comprehensive review of AI applications in medical sciences. In addition, readers will benefit from hands on exercises, with included R code.

This book constitutes the refereed proceedings of the 4th International Conference on Soft Computing in Data Science, SCDS 2018, held in Bangkok, Thailand, in August 2018. The 30 revised full papers presented were carefully reviewed and selected from 75 submissions. The papers are organized in topical sections on machine and deep learning, image processing, financial and fuzzy mathematics, optimization algorithms, data and text analytics, data visualization.

If you are ready to dive into the MapReduce framework for processing large datasets, this practical book takes you step by step through the algorithms and tools you need to build distributed MapReduce applications with Apache Hadoop or Apache Spark. Each chapter provides a recipe for solving a massive computational problem, such as building a recommendation system. You’ll learn how to implement the appropriate MapReduce solution with code that you can use in your projects. Dr. Mahmoud Parsian covers basic design patterns, optimization techniques, and data mining and machine learning solutions for problems in bioinformatics, genomics, statistics, and social network analysis. This book also includes an overview of MapReduce, Hadoop, and Spark. Topics include: Market basket analysis for a large set of transactions Data mining algorithms (K-means, KNN, and Naive Bayes) Using huge genomic data to sequence DNA and RNA Naive Bayes theorem and Markov chains for data and market prediction Recommendation algorithms and pairwise document similarity Linear regression, Cox regression, and Pearson correlation Allelic frequency and mining DNA Social network analysis (recommendation systems, counting triangles, sentiment analysis)

Distance-based algorithms are machine learning algorithms that classify queries by computing

distances between these queries and a number of internally stored exemplars. Exemplars that are closest to the query have the largest in uence on the classi cation assigned to the query. Two speci c distance-based algorithms, the nearest neighbor algorithm and the nearest-hyperrectangle algorithm, are studied in detail. It is shown that the k-nearest neighbor algorithm (kNN) outperforms the rst- nearest neighbor algorithm only under certain conditions. Data sets must contain moderate amounts of noise. Training examples from the di erent classes must belong to clusters that allow an increase in the value of k without reaching into clusters of other classes. Methods for choosing the value of k for kNN are investigated. It is shown that one-fold cross-validation on a restricted number of values for k su ces for best performance. It is also shown that for best performance the votes of the k-nearest neighbors of a query should be weighted in inverse proportion to their distances from the query. Principal component analysis is shown to reduce the number of relevant dimen- sions substantially in several domains. Two methods for learning feature weights for a weighted Euclidean distance metric are proposed. These methods improve the performance of kNN and NN in a variety of domains. The nearest-hyperrectangle algorithm (NGE) is found to give predictions that are substantially inferior to those given by kNN in a variety of domains. Experiments performed to understand this inferior performance led to the discovery of several improvements to NGE. Foremost of these is BNGE, a batch algorithm that avoids construction of overlapping hyperrectangles from di erent classes. Although it is generally superior to NGE, BNGE is still signi cantly inferior to kNN in a variety of domains. Hence, a hybrid algorithm (KBNGE), that uses BNGE in parts of the input space that can be represented by a single hyperrectangle and kNN otherwise, is introduced. The primary contributions of this dissertation are (a) several improvements to existing distance-based algorithms, (b) several new distance-based algorithms, and (c) an experimentally supported understanding of the conditions under which various distance-based algorithms are likely to give good performance.

This book is devoted to a novel approach for dimensionality reduction based on the famous nearest neighbor method that is a powerful classification and regression approach. It starts with an introduction to machine learning concepts and a real-world application from the energy domain. Then, unsupervised nearest neighbors (UNN) is introduced as efficient iterative method for dimensionality reduction. Various UNN models are developed step by step, reaching from a simple iterative strategy for discrete latent spaces to a stochastic kernel-based algorithm for learning submanifolds with independent parameterizations. Extensions that allow the embedding of incomplete and noisy patterns are introduced. Various optimization approaches are compared, from evolutionary to swarm-based heuristics. Experimental comparisons to related methodologies taking into account artificial test data sets and also real-world data demonstrate the behavior of UNN in practical scenarios. The book contains numerous color figures to illustrate the introduced concepts and to highlight the experimental results.

A Study of Distance-based Machine Learning Algorithms

The Handbook of Statistical Analysis and Data Mining Applications is a comprehensive professional reference book that guides business analysts, scientists, engineers and researchers (both academic and industrial) through all stages of data analysis, model building and implementation. The Handbook helps one discern the technical and business problem, understand the strengths and weaknesses of modern data mining algorithms, and employ the right statistical methods for practical application. Use this book to address massive and complex datasets with novel statistical approaches and be able to objectively evaluate analyses and solutions. It has clear, intuitive explanations of the principles and tools for solving problems using modern analytic techniques, and discusses their application to real problems, in ways accessible and beneficial to practitioners across industries - from science and engineering, to medicine, academia and commerce. This handbook brings together, in a single

resource, all the information a beginner will need to understand the tools and issues in data mining to build successful data mining solutions. Written "By Practitioners for Practitioners" Non-technical explanations build understanding without jargon and equations Tutorials in numerous fields of study provide step-by-step instruction on how to use supplied tools to build models Practical advice from successful real-world implementations Includes extensive case studies, examples, MS PowerPoint slides and datasets CD-DVD with valuable fully-working 90-day software included: "Complete Data Miner - QC-Miner - Text Miner" bound with book
Im Zeitalter des Internet of Things (IoT) erzeugen Edge-Geräte in jedem Sekundenbruchteil gigantische Datenmengen. Dabei besteht das Hauptziel dieser Netzwerke darin, aus den gesammelten Daten sinnvolle Informationen abzuleiten. Gleichzeitig werden gewaltige Datenmengen in die Cloud übertragen, was extrem teuer und zeitaufwändig ist. Es ist somit notwendig, effiziente Mechanismen für die Verarbeitung dieser gewaltigen Datenmengen zu entwickeln, und dafür sind effiziente Datenverarbeitungstechniken erforderlich. Nachhaltige Paradigmen wie Cloud Computing und Fog Computing tragen zu einem geschickten Umgang mit Themen wie Leistung, Speicher- und Verarbeitungskapazitäten, Wartung, Sicherheit, Effizienz, Integration, Kosten, Energieverbrauch und Latenzzeiten bei. Allerdings werden ausgefeilte Analysetools benötigt, um die Anfragen in einer optimalen Zeit zu bearbeiten. Daher wird derzeit eifrig an der Entwicklung eines effektiven und effizienten Rahmens geforscht, um den größtmöglichen Nutzen zu erhalten. Bei der Verarbeitung der gewaltigen Datenmengen steht das maschinelle Lernen besonders hoch im Kurs und wird in zahlreichen Disziplinen angewandt, auch in den sozialen Medien. In Machine Learning Approach for Cloud Data Analytics in IoT werden sämtliche Aspekte des IoT, des Cloud Computing und der Datenanalyse ausführlich erläutert und aus verschiedenen Perspektiven betrachtet. Das Buch präsentiert den neuesten Stand der Forschung und fortschrittliche Themen. So erhalten die Leserinnen und Leser aktuelle Informationen und können das gesamte Spektrum der Anwendungen von IoT, Cloud Computing und Datenanalyse erfassen.
Fourth International Conference on Information and Communication Technology for Competitive Strategies targets state-of-the-art as well as emerging topics pertaining to information and communication technologies (ICTs) and effective strategies for its implementation for engineering and intelligent applications.
With the ever-growing power of generating, transmitting, and collecting huge amounts of data, information overloadis nowan imminent problemto mankind. The overwhelming demand for information processing is not just about a better understanding of data, but also a better usage of data in a timely fashion. Data mining, or knowledge discovery from databases, is proposed to gain insight into aspects ofdata and to help peoplemakeinformed,sensible,and better decisions. At present, growing attention has been paid to the study, development, and application of data mining. As a result there is an urgent need for sophisticated techniques and toolsthat can handle new ?elds of data mining, e. g. , spatialdata mining, biomedical data mining, and mining on high-speed and time-variant data streams. The knowledge of data mining should also be expanded to new applications. The 6th International Conference on Advanced Data Mining and Appli-
tions(ADMA2010)aimedtobringtogethertheexpertsondataminingthrou- out the world. It provided a leading international forum for the dissemination of original research results in advanced data mining techniques, applications, al- rithms, software and systems, and di?erent applied disciplines. The conference attracted 361 online submissions from 34 di?erent countries and areas. All full papers were peer reviewed by at least three members of the Program Comm- tee composed of international experts in data mining ?elds. A total number of 118 papers were accepted for the conference. Amongst them, 63 papers were selected as regular papers and 55 papers were selected as short papers.
Nearest neighbor search is a fundamental requirement of many machine learning algorithms

and is essential to fuzzy information retrieval. The utility of efficient database search and construction has broad utility in a variety of computing fields. Applications such as coding theory and compression for electronic communication systems as well as use in artificial intelligence for pattern and object recognition. In this thesis, a particular subset of nearest neighbors is consider, referred to as c-approximate k-nearest neighbors search. This particular variation relaxes the constraints of exact nearest neighbors by introducing a probability of finding the correct nearest neighbor c, which offers considerable advantages to the computational complexity of the search algorithm and the database overhead requirements. Furthermore, it extends the original nearest neighbors algorithm by returning a set of k candidate nearest neighbors, from which expert or exact distance calculations can be considered. Furthermore thesis extends the implementation of c-approximate k-nearest neighbors search so that it is able to utilize the burgeoning GPGPU computing field. The specific form of c-approximate k-nearest neighbors search implemented is based on the locality sensitive hash search from the E2LSH package of Indyk and Andoni [1]. In this paper, the authors utilize the exceptional properties of the Leech Lattice [2], as a subspace quantizer for the locality sensitive hash families. The Leech Lattice is unique in that it provides the closest lattice packing of equal sized spheres in 24 dimensional space. In addition, advances from coding theory provide a very favorable decoding algorithm for finding the nearest lattice center to a query point in euclidean 24 dimensional space [3] [4]. The multilevel construction of the Leech Lattice provides an excellent opportunity for parallelization as it contains the minimization of many independent sub-lattice decodings resulting from the lattices exceptional symmetry among lattices. These decodings are additionally highly floating point computationally intensive, and because of which suggest a favorable implementation on GPGPU architectures such as NVIDIAs CUDA based framework. Furthermore, the overall construction of a locality sensitive hash based, nearest neighbors search algorithm, is able to be parallelized fairly efficiently as the hash decodings are completely independent of one another. The goal of this thesis is to present a CUDA optimized parallel implementation of a bounded distance Leech Lattice decoder [4] for use in query optimized c-approximate k-nearest neighbors using the locality sensitive hash framework of E2LSH. The system will be applied to the approximate image retrieval of SIFT transformed [5] image vectors.[1] A. Andoni, Nearest Neighbor Search: the Old, the New, and the Impossible. INSTITUTE OF [2] J. Leech, "Notes on sphere packings," [3] . J. Forney, G.D., "A bounded-distance decoding algorithm for the leech lattice, with generalizations, "[4] O. Amrani and Y. Beery, "Efficient bounded-distance decoding of the hexacode and associated decodersfor the leech lattice and the golay code," [5] D.G. Lowe, "Object recognition from local scale-invariant features,"

Technologies using identification by radio frequencies (RFID) are experiencing rapid development and healthcare is a major application area benefiting from it. Highly pervasive RFID enables remote identification, tracking and localization of the medical staff, patients, medications and equipment, thus increasing safety, optimizing in real-time management and providing support for new ambient-intelligent services. This thesis describes and evaluates an algorithm that enables object localization and tracking using passive RFID tags. This thesis also describes scenarios of how this technology can be used as a part of building a smart trauma resuscitation room by tracking the equipments. The main contribution of this thesis is the adaptation of the Weighted K-Nearest Neighbor Algorithm as a localization technique to track objects in a confined and crowded space by using passive RFID tags. The input parameter to the algorithm is the received signal strength indicator (RSSI), which gives a measure of back-scattered radio frequencies from passive tags. While using RFID technology special attention has to be given to the placement of antennas to get the optimum result. Therefore, we analyzed various antenna placement configurations with mean error and error consistency as the two performance parameters. The detection of multiple tags and human

occlusion are two major concerns while tracking tags in a confined space with many team members collaborating on solving a problem. The RF signal can be interrupted by people walking around randomly and holding multiple (tagged) instruments at the same time. While the algorithm worked fine when tracking multiple tags, we had to modify the experimental set-up and attach an antenna onto the ceiling (which we call a vertical antenna), so that even if all the wall antennas are blocked we get at least one input parameter to base our localization decision on. We evaluated the algorithm for different combinations of configurations and number of neighbors, and achieved the following results. The best results were obtained for the 3 antennae (placed orthogonally) configuration considering the 4 nearest neighbors wherein a mean error rate of 15% of the maximum possible error was achieved under ideal conditions. We tested the algorithm for different human occlusion scenarios i.e. blocking 1 or 2 wall antennas, standing in random positions and then roaming in the field area randomly. The mean error rate for the standing scenario was measured as 20% of the maximum possible error and 18% in the case of roaming configuration. The error was found to be consistently within our defined maximum error for 100% of the recorded readings. The results obtained were found to be satisfactory for our application where, more than the exact location of the object, knowing whether the object is within a particular region is good enough for the users to know what task is being carried out in the trauma bay. Also the algorithm holds good in an indoor environment having a lot of factors and materials which affect the RF signal disrupting accurate calculation of the location co-ordinates. The algorithm does not require extensive data collection prior to implementation which makes it easily deployable in any environment. Apart from the problems mentioned there are some other factors like materials on which the tags are attached and orientation of tags which were found to be potential hindrances for accurate localization. Acceptable solutions to these problems form a part of our future work.