(1)
Mining disease integrated ontology

Taysir Hassan A. Soliman, Marwa Hussein, and Mohamed El-Sharkawi

Abstract:

Ontology has become a very vital issue to solve important issues regarding human diseases through data integration of chemical and biological data. Mining such data discovers highly important knowledge about diseases can give an important insight to arrive to new drug targets and assist in personalized medicine. In the current paper, a mining technique for diseases is developed based on integrated ontology and association rule mining algorithm. To perform mining, the semantic web, as a knowledge representation methodology is used to integrate data. In addition, an Ontology Association Rule Mining algorithm (OARM) is developed since existing algorithms cannot be applied because of the ontology nature of data containing several types of relations. To test our performance, prostate cancer data is obtained from NCI, which is related to 279 genes and 89 genes (from prostate cancer pathway).

Keywords:

Association Rule Mining , Disease-related information ontology , Gene Ontology , Semantic Web

Published In:

IEEE 12th International Conference on Bioinformatics and Bioengineering, Cyprus , , pp. 40- 45

(2)

A gene selection approach for classifying diseases based on microarray datasets

Taysir Hassan A. Soliman, Adel A. Sewissy, and Hisham Abdel Latif

Abstract:

Gene Selection is very important problem in the classification of serious diseases in clinical information systems. A limitation of these gene selection methods is that they may result in gene sets with some redundancy and yield an unnecessary large number of candidate genes for classification analysis. In the current work, a hybrid approach is presented in order to classify diseases, such as colon cancer, leukemia, and liver cancer, based on informative genes. This hybrid approach uses clustering (K-means) with statistical analysis (ANOVA) as a preprocessing step for gene selection and Support Vector Machines (SVM) to classify diseases related to microarray experiments. To compare the performance of the proposed methodology, two kinds of comparisons were achieved: 1) applying statistical analysis combined with clustering algorithm (K-means) as a preprocessing step and 2) comparing different classification algorithms: decision tree (ID3), naïve bayes, adaptive naïve bayes, and support vector machines. In case of combining clustering with statistical analysis, much better classification accuracy is given of 97% rather than without applying clustering in the preprocessing phase. In addition, SVM had proven better accuracy than decision trees, Naïve Bayes, and Adaptive Naïve Bayes classification.

Keywords:

ANOVA test , Classification , Clustering , Feature Selection , Gene Selection , Microarray data

Published In:

Computer Technology and Development (ICCTD), 2010 2nd International Conference on , , pp.626- 631

(3)

SS-SVM (3SVM): A New Classification Method for Hepatitis Disease Diagnosis

Mohammed H. Afif, Abdel-Rahman Hedar, Taysir H. Abdel Hamid, Yousef B. Mahdy

Abstract:

Abstract.In this paper, a new classification approach combining support vector machine with scatter search approach for hepatitis disease diagnosis is presented, called 3SVM. The scatter search approach is used to find near optimal values of SVM parameters and its kernel parameters. The hepatitis dataset is obtained from UCI. Experimental results and comparisons prove that the 3SVM gives better outcomes and has a competitive performance relative to other published methods found in literature, where the average accuracy rate obtained is 98.75%.

Keywords:

Support Vector Machine; Scatter Search; Classification; Parameter tuning

Published In:

International Journal of Advanced Computer Science and Applications , Vol. 4 , No. 2

(4)

Support Vector Machines with Weighted Powered Kernels for Data Classification

Mohammed H. Afif, Abdel-Rahman Hedar, Taysir H. Abdel Hamid, and Yousef B. Mahdy

Abstract:

Abstract. Support Vector Machines (SVMs) are a popular data classification method with many diverse applications. The SVMs performance depends on choice a suitable kernel function for a given problem. Using an appropriate kernel; the data are transform into a space with higher dimension in which they are separable by an hyperplane. A major challenges of SVMs are how to select an appropriate kernel and how to find near optimal values of its parameters. Usually a single kernel is used by most studies, but the real world applications may required a combination of multiple kernels. In this paper, a new method called, weighted powered kernels for data classification is proposed. The proposed method combined three kernels to produce a new combined kernel (WPK). The method used Scatter Search approach to find near optimal values of weights, alphas and kernels parameters which associated with each kernel. To evaluate the performance of the proposed method, 11 benchmark are used. Experiments and comparisons prove that the method given acceptable outcomes and has a competitive performance relative to a single kernel and some other published methods

Keywords:

Support Vector Machine, Scatter Search, Classification

Published In:

Advanced Machine Learning Technologies and Applications Communications in Computer and Information Science , Volume 322 , pp 369-378

(5)

Utilizing Support Vector Machines in Mining Online Customer Reviews

Taysir Hassan A. Soliman, Mostafa A. Elmasry, Abdel Rahman Hedar, and Magdy M. Doss

Abstract:

As e-commerce is increasingly becoming popular, the number of customer reviews that a product receives grows rapidly. However, for popular products, many online product reviews exist but for other reviews product reviews are very few. These online discussions about particular products may help other online users to make a decision in buying/ not buying those products, like in amazon.com and ebay.com. Since an enormous number of unstructured and ungrammatical reviews on a product exist, opinion mining is getting a crucial research area for better decision making of buying products. In this paper, we apply an opinion mining approach to summarize the unstructured and ungrammatical users' reviews, based on Support Vector Machine (SVM). Two levels of classification is applied: 1)Features classification and 2) Polarity classification for every feature class. Our approach has been tested on Amazon data with dataset of 535 sentences, where a summary is obtained and analysis of precision (93.15%) and recall (92.41%) illustrate the accuracy of the proposed system.

Keywords:

Opinion mining, E-commerce, sentiment analysis, support vector machines, reviews classification, opinion visual summary.

Published In:

Proceedings of 22th International Conference on Computer Theory and Applications ICCTA 2012, Alexandria, Egypt , NULL , NULL

(6)

Mining Multi Drug-Pathways via A Probabilistic Heterogeneous Network Multi-labelClassifier,

Taysir Hassan A. Soliman

Abstract:

Mining drug networks is a very important research issue to discover hidden relations between multi drug-entities relations, such as multi drug-pathways, multi drug-targets, and multi drug-diseases. One very important relation is the drug-pathway, where drugs affect the human body through their pathways. In this paper, a probabilistic Heterogeneous Network Multi-label Classifier (HNMC) is proposed to classify multi drug-pathways relations. Data is collected from Drugbank.ca [1], Kegg (keg drug, Kegg diseases, Kegg pathways, Kegg orthologs, Kegg brite) [2] and small molecular pathways [3,4]. For drug-pathways data, two datasets are considered: one is based on Drug-Drug Interaction (DDI) and the other is based on Drug-Pathways Interactions (DPI). HNMC has proved its efficiency with an average of 90% precision, 92.35% recall, 92% accuracy, and 96% ROC

Keywords:

Multi Drug-Pathways Prediction , Probabilistic Heterogeneous Networks and Multi-label Classification

Published In:

Bonfring International Journal of Research in Communication Engineering, , Vol. 4, No. 2 , pp.10-16

(7)

Systems biology analysis of hepatitis C virus infection reveals the role of copy number increases in regions of chromosome 1q in hepatocellular carcinoma metabolism

Ibrahim E. Elsemman, Adil Mardinoglu, Saeed Shoaie, Taysir H Soliman and Jens Nielsen

Abstract:

Hepatitis C virus (HCV) infection is a worldwide healthcare problem; however, traditional treatment methods have failed to cure all patients, and HCV has developed resistance to new drugs. Systems biology-based analyses could play an important role holistic analysis of the impact of HCV on hepatocellular metabolism. Here, we integrated HCV assembly reactions with a genome-scale hepatocyte metabolic model to identify the metabolic targets for HCV assembly and the metabolic alterations that occur between different HCV progression states (cirrhosis, dysplastic nodule, and early and advanced hepatocellular carcinoma (HCC)) and healthy liver tissue. We found that diacylglycerolipids were essential for HCV assembly. In addition, the metabolism of keratan sulfate and chondroitin sulfate was significantly changed in the cirrhosis stage, whereas the metabolism of acyl-carnitine was significantly changed in the dysplastic nodule and early HCC stages. Our results explained the role of the upregulated expression of BCAT1, PLOD3 and six other methyltransferase genes involved in carnitine biosynthesis and S-adenosylmethionine metabolism in the early and advanced HCC stages. Moreover, GNPAT and BCAP31 expression was upregulated in the early and advanced HCC stages and could lead to increased acyl-CoA consumption. By integrating our results with copy number variation analyses, we observed that GNPAT, PPOX and five of the methyltransferase genes (ASH1L, METTL13, SMYD2, TARBP1 and SMYD3), which are all located on chromosome 1q, had increased copy numbers in the cancer samples relative to the normal samples. Finally, we confirmed our predictions with the results of metabolomics studies and proposed that inhibiting the identified targets has the potential to provide an effective treatment strategy for HCV-associated liver disorders.

Keywords:

NULL

Published In:

Molecular BioSystems , 2016 , NULL

(8)

Sentiment analysis of Arabic slang comments on facebook

Taysir Hassan Soliman, M.A. Elmasry, A. Hedar, M.M. Doss

Abstract:

ABSTRACT Social networks have become one of our daily life activities not only in socializing but in e-commerce, e-learning, and politics. However, they have more effect on the youth generation all over the world, specifically in the Middle East. Arabic slang language is widely used on social networks more than classical Arabic since most of the users of social networks are young-mid age. However, Arabic slang language suffers from the new expressive (opinion) words and idioms as well as the unstructured format. Mining ...

Keywords:

NULL

Published In:

International Journal of Computers & Technology , Vol. 12, No. 5 , pp. 3470-3478

(9)

A Hadoop Extension for Analysing Spatiotemporally Referenced Events

Mohamed S Bakli, Mahmoud A Sakr, Taysir Hassan A Soliman

Abstract:

A spatiotemporally referenced event is a tuple that contains both a spatial reference and a temporal reference. The spatial reference is typically a point coordinate, and the temporal reference is a timestamp. The event payload can be the reading of a sensor (IoT systems), a user comment (geo-tagged social networks), a news article (gdelt), etc. Spatiotemporal event datasets are ever growing, and the requirements for their processing goes beyond traditional client-sever GIS architectures. Rather, Hadoop-like architectures shall be used. Yet, Hadoop does not provide the types and operations necessary for processing such datasets. In this paper, we propose a Hadoop extension (indeed a SpatialHadoop extension) capable of performing analytics on big spatiotemporally referenced event dataset. The extension includes data types and operators that are integrated into the Hadoop core, to be used as natives. We further optimize the querying by means of a spatiotemporal index. Experiments on the gdelt event dataset demonstrate the utility of the proposed extension.

Keywords:

Spatiotemporal data, Hadoop, Geo-events, Movement analysis

Published In:

International Conference on Advanced Intelligent Systems and Informatics. , (Vol 639) , (pp.905-914)

(10)

A spatiotemporal algebra in Hadoop for moving objects

Mohamed S. Bakli, Mahmoud A. Sakr, Taysir Hassan A. Soliman

Abstract:

Spatiotemporal data represent the real-world objects that move in geographic space over time. The enormous numbers of mobile sensors and location tracking devices continuously produce massive amounts of such data. This leads to the need for scalable spatiotemporal data management systems. Such systems shall be capable of representing spatiotemporal data in persistent storage and in memory. They shall also provide a range of query processing operators that may scale out in a cloud setting. Currently, very few researches have been conducted to meet this requirement. This paper proposes a Hadoop extension with a spatiotemporal algebra. The algebra consists of moving object types added as Hadoop native types, and operators on top of them. The Hadoop file system has been extended to support parameter passing for files that contain spatiotemporal data, and for operators that can be unary or binary. Both the types and operators are accessible for the MapReduce jobs. Such an extension allows users to write Hadoop programs that can perform spatiotemporal analysis. Certain queries may call more than one operator for different jobs and keep these operators running in parallel. This paper describes the design and implementation of this algebra, and evaluates it using a benchmark that is specific to moving object databases.

Keywords:

Spatiotemporal algebra, Hadoop, MapReduce, moving objects

Published In:

Geo-spatial Information Science , (Vol 21 - No 2) , (PP.102-114)

(11)

HadoopTrajectory: a Hadoop spatiotemporal data processing extension

Mohamed Bakli, Mahmoud Sakr, Taysir Hassan A. Soliman

Abstract:

The recent advances in location tracking technologies and the widespread use of location-aware applications have resulted in big datasets of moving object trajectories. While there exists a couple of research prototypes for moving object databases, there is a lack of systems that can process big spatiotemporal data. This work proposes HadoopTrajectory, a Hadoop extension for spatiotemporal data processing. The extension adds spatiotemporal types and operators to the Hadoop core. These types and operators can be directly used in MapReduce programs, which gives the Hadoop user the possibility to write spatiotemporal data analytics programs. The storage layer of Hadoop, the HDFS, is extended by types to represent trajectory data and their corresponding input and output functions. It is also extended by file splitters and record readers. This enables Hadoop to read big files of moving object trajectories such as vehicle GPS tracks and split them over worker nodes for distributed processing. The storage layer is also extended by spatiotemporal indexes that help filtering the data before splitting it over the worker nodes. Several data access functions are provided so that the MapReduce layer can deal with this data. The MapReduce layer is extended with trajectory processing operators, to compute for instance the length of a trajectory in meters. This paper describes the extension and evaluates it using a synthetic dataset and a real dataset. Comparisons with non-Hadoop systems and with standard Hadoop are given. The extension accounts for about 11,601 lines of Java code.

Keywords:

Spatiotemporal, Hadoop, 3DR-tree, Trajectory data management, Big data

Published In:

Journal of Geographical Systems , NULL , NULL

(12)

Developing an Efficient Spectral Clustering Algorithm on Large Scale Graphs in Spark

Ahmed I. Taloba Marwan R. Riad Taysir Hassan A. Soliman

Abstract:

Recently, most of the data can be represented by graph structures, such as social media, Protein-Protein Interaction, transportation system, systems biology,..., etc. Many researches have been achieved to cluster very large graphs but more efficient algorithms are required since such a process takes a long time and requires more memory. In this paper, we propose an Efficient Spectral Clustering Algorithm on Large Scale Graphs in Spark (ESCALG), using map reduce function and shuffling phases in Dijkstra's algorithm. In addition, ESCALG depends mainly on a sparse matrix as a data structure, which less time in execution. Then, GraphX is applied to deal with graph data processing and in GraphX used Pregel in computing shortest path. To test the performance of ESCALG, it is compared with Large-Scale Spectral Clustering on Graphs and Standard Spectral Clustering Algorithms using seven datasets, where ESCALG proved high efciency in terms of memory and time performance.

Keywords:

Spectral Clustering , Apache Spark, Large scale Graph Clustering

Published In:

2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS) , NULL , 292-298