ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

An Innovative Method To Deal With Data Noise Where Multiple Measurements Are Collected Using Repeated Experiments

Mallikarjuna Lokireddy
M.Tech Student,
Nova College of Engineering and Technology.

Ch. Rajajacob
Guide,
Nova College of Engineering and Technology.

Abstract:

A distributed system is a software system in which components located on networked computers communicate and coordinate their actions by passing messages. The components interact with each other in order to achieve a common goal. Three significant characteristics of distributed systems are: concurrency of components, lack of a global clock, and independent failure of components. Examples of distributed systems vary from SOA-based systems to massively multiplayer online games to peer-to-peer applications.

A computer program that runs in a distributed system is called a distributed program, and distributed programming is the process of writing such programs. There are many alternatives for the message passing mechanism, including RPC-like connectors and message queues. An important goal and challenge of distributed systems is location transparency.

Distributed computing also refers to the use of distributed systems to solve computational problems. In distributed computing, a problem is divided into many tasks, each of which is solved by one or more computers, which communicate with each other by message passing.Noisy data is meaningless data.

The term was often used as a synonym for corrupt data, but its meaning has expanded to include data from unstructured text that cannot be understood by machines. We studied and implemented An Innovative method to deal with data Noise where multiple measurements are collected using repeated experiments. We suggest a type of OPSM, where every data item is represented by a set of values achieved from replicated experiments.

Keywords: OPSM, Data Noise, Repeated experiments, Data mining, Classification, bioinformatics, mining methods and Algorithms.

Introduction:

Data mining an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices.

These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

Data mining involves six common classes of tasks:

Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".Regression – attempts to find a function which models the data with the least error.Summarization – providing a more compact representation of the data set, including visualization and report generation.

The word distributed in terms such as "distributed system", "distributed programming", and "distributed algorithm" originally referred to computer networks where individual computers were physically distributed within some geographical area. The terms are nowadays used in a much wider sense, even referring to autonomous processes that run on the same physical computer and interact with each other by message passing. While there is no single definition of a distributed system,[6] the following defining properties are commonly used:

- There are several autonomous computational entities, each of which has its own local memory.

- The entities communicate with each other by message passing.

Data Noise::

Noisy data is meaningless data. The term has often been used as a synonym for corrupt data.

However, its meaning has expanded to include any data that cannot be understood and interpreted correctly by machines, such as unstructured text. Any data that has been received, stored, or changed in such a manner that it cannot be read or used by the program that originally created it can be described as noisy.Noisy data unnecessarily increases the amount of storage space required and can also adversely affect the results of any data mining analysis. Statistical analysis can use information gleaned from historical data to weed out noisy data and facilitate data mining.Noisy data can be caused by hardware failures, programming errors and gibberish input from speech or optical character recognition (OCR) programs. Spelling errors, industry abbreviations and slang can also impede machine reading.

EXISTING SYSTEM:

To synthesize additional replicates, for each gene and each experiment, we follow standard practice to model the values by a Gaussian distribution with the mean and variance equal to the sample mean and variance of the 4 replicates. The expression values of new replicates were then sampled from the Gaussian. New columns are synthesized by randomly drawing an existing column, fitting Gaussians as described, and sampling values from it. This way of construction mimics the addition of knockout experiments of genes in the same sub pathways of the original ones.

Disadvantage:

1.The database with the tuple data does not be maintained confidentially.
2.The existing systems another person to easily access database.

PROPOSED SYSTEM:

Propose a generic mining algorithm. We further propose a series of techniques to speed up two time-dominating components of the algorithm. We show the effectiveness and efficiency of our methods through a series of experiments conducted on real microarray data. The conventional order-preserving sub matrix (OPSM) mining problem was motivated and introduced to analyze gene expression data without repeated measurements.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

They proved that the problem is NP hard. A greedy heuristic mining algorithm was proposed, which does not guarantee the return of all OPSM's or the best OPSM's.

Advantage:

1.We can find duplicate records and same group types.
2.Easily we can catch genes relationship from database.

Modules:

1.Patient Module.
2.Doctor Module.
3.User Module.
4.Admin Module.

Patient Module:

Department Manager's Patient Module allows administrators to access and manage their own patient billing records at their desktop. It tracks patient diagnosis information for each admission or visit as well as a historical account of services provided to each patient. This module helps to collect complete and relevant patient information. The system automates the patient administration functions to have better and efficient patient care process.

Doctor Module:

In this module, doctor register and login. He inserts patient details and view patient records and he change his password and the data will be passed from the co-coordinator and thus it will be submitted to the End Users (Data Users).

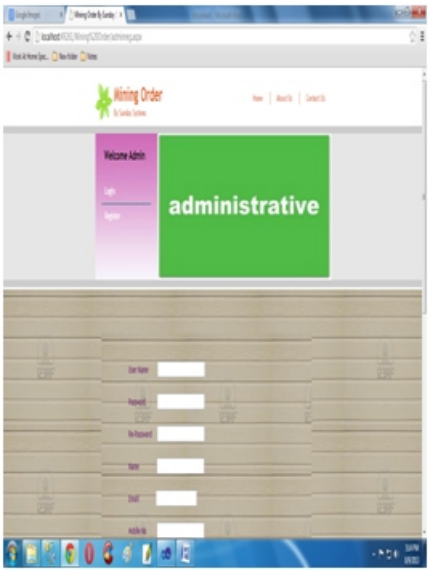
User Module:

In this module, the Users are classified into two types they are, Data Users and Data Owner Depends on the restriction the data will be passed to the Co-coordinator.The co-coordinator pass the details via broker and the data will be checked with the secret key and thus it will display for the users.

Admin Module:

In this module, to arrange the database based on the patient and doctor details and records. The admin needs to register and register the Organization and Users Forms.

Screen Shots:



Conclusion:

Due to elevated level of noise in distinctivemicroarray data, it is typically moresignificant to evaluate the comparativeexpression levels of different genes atdissimilar time points rather than theircomplete values.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Monthly Peer Reviewed Open Access International e-Journal

The difficulty of Order-Preserving Submatrix pertains to a matrix of statistical data values. It is connected to problems of pattern-based subspace clustering as well as sequence mining all of which search for patterns in particular subspaces or subsequences. In gene expression circumstance, orderpreserving submatrix match up to groups of genes that have comparable activity patterns, which might suggest shared regulatory mechanisms as well as protein functions.The studied and implemented system shows better results when compared to the existing system.

Regerences:

[1] C. K. Chui, B. Kao, K. Y. Yip, and S. D. Lee. Mining order-preserving submatrices from data with repeated measurements.In ICDM '08, pages 133–142, 2008.

[2] Beginning ASP.NET 4: in C# and VB by Imar Spaanjaars.

[3] Programming ASP.NET 3.5 by Jesse Liberty, Dan Maharry, Dan Hurwitz.

[4] Beginning ASP.NET 3.5 in C# 2008: From Novice to Professional, Second Edition by Matthew MacDonald.

[5] M.-L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations," Proceedings of the National Academy of Sciences of the United States of America, vol. 97, no. 18, pp. 9834–9839, 2000.

[6] G. Li, Q. Ma, H. Tang, A. Paterson, and Y. Xu. Qubic: a qual-itative biclustering algorithm for analyses of gene expressiondata. Nucleic acids research, 37(15), 2009.

[7] J. Liu and W. Wang. Op-cluster: Clustering by tendency in high dimensional space. In ICDM '03, 2003.

[12] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. IEEE/ACM TCBB, 1(1):24–45, 2004.

[8] G. Pandey, G. Atluri, M. Steinbach, C. L. Myers, and V. Kumar. An association analysis approach to biclustering. InSIGKDD '09, pages 677–686, 2009.

[9] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, and et al. Prefixs-pan: Mining sequential patterns efficiently by prefix-projectedpattern growth. In ICDE '01, 2001.

[10] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, and et al. Mining se-quential patterns by pattern-growth: The prefixspanapproach. IEEE TKDE, 16:1424–1440, 2004.