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

Efficient Method for Periodic Outlier Detection in time Series Databases

M.Pramod
M. Tech (SE),
Department of CSE,
SNIST, Hyderabad.

Dr.Prasanta Kumar Sahoo
Professor,
Department of CSE,
SNIST, Hyderabad.

Abstract:

Periodic pattern detection in time-ordered sequences is a vital data processing task that discovers within the statistic all patterns that exhibit temporal regularities. Periodic pattern mining contains a sizable amount of applications in real life; it helps understanding the regular trend of the info on time, and permits the forecast and prediction of future events. a remarkable connected and very important drawback that has not received enough attention is to find outlier periodic patterns during a statistic. Outlier patterns are outlined as those that are completely different from the remainder of the patterns; outliers aren't noise. Whereas noise doesn't belong to the info and it's largely eliminated by pre-processing, outliers are actual instances within the knowledge however have exceptional characteristics compared with the bulk of the opposite instances. To propose a replacement spectral cluster technique known as correlation protective compartmentalization (CPI) that is performed within the correlation similarity live area the documents are projected into a low-dimensional linguistics area.

Keywords:

Outlier periodic patterns, performance, periodicity detection, suffix tree, surprising patterns, surprising periodicity, time series, unusual periods.

Introduction:

Outliers are unusual patterns that rarely occur, and, thus, have lesser support (frequency of appearance) in the data. Outlier patterns may hint toward discrepancy in the data such as fraudulent transactions, network intrusion, change in customer behaviour, recession in the economy, epidemic and disease biomarkers, severe weather conditions like tornados, etc. We argue that detecting the periodicity of outlier patterns might be more important in many sequences than the periodicity of regular, more frequent patterns.

In this paper, we present a robust and time efficient suffix tree -based algorithm capable of detecting the periodicity of outlier patterns in a time series by giving more significance to less frequent yet periodic patterns. Several experiments have been conducted using both real and synthetic data; all aspects of the proposed approach are compared with the existing algorithm Info Miner; the reported results demonstrate the effectiveness and applicability of the proposed approach.

Previous Method:

A typical and widely used distance measure is the Euclidean distance. The k-means method is one of the methods that use the Euclidean distance, which minimizes the sum of the squared Euclidean distance between the data points and their corresponding cluster centers. Since the document space is always of high dimensionality, it is preferable to find a low dimensional representation of the documents to reduce computation complexity. Latent semantic indexing (LSI) is one of the effective spectral clustering methods, aimed at finding the best subspace approximation to the original document space by minimizing the global reconstruction error (Euclidean distance). However, because of the high dimensionality of the document space, a certain representation of documents usually resides on a nonlinear manifold embedded in the similarities between the data points. Unfortunately, the Euclidean distance is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents.

Limitations:

It is not able to effectively capture the nonlinear manifold structure embedded in the similarities between them. It is a dissimilarity measure which describes the dissimilarities rather than similarities between the documents. The existing one not works with high dimensionality of the document space.

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

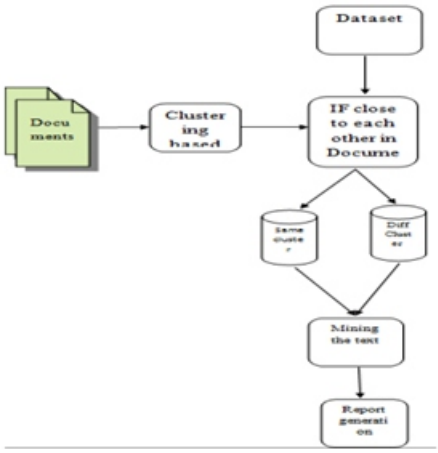
Proposed method:

An effective document clustering method must be able to find a low-dimensional representation of the documents that can best preserve the similarities between the data points. Locality preserving indexing (LPI) method is a different spectral clustering method based on graph partitioning theory. In probability theory and statistics, correlation indicates the strength and direction of a linear relationship between two random variables which reveals the nature of data represented by the classical geometric concept of an ―angle.‖ It is a scale-invariant association measure usually used to calculate the similarity between two vectors. The LPI method applies a weighted function to each pair wise distance attempting to focus on capturing the similarity structure, rather than the dissimilarity structure, of the documents.

Advantages:

The correlation as a similarity measure can capture the intrinsic structure embedded in high-dimensional data, especially when the input data is sparse. In many cases, correlation can effectively represent the distributional structure of the input data which conventional Euclidean distance cannot explain. It effectively captures the non-linear manifold structure embedded in the similarities between them.

Architecture:



Weka:

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.

Weka is free software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality [4]. Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code [5]. The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

I.Free availability under the GNU General Public License

II.Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.

III.A comprehensive collection of data preprocessing and modeling techniques.

IV.Ease of use due to its graphical user interfaces Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection [10]. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling [4].

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

Literature Survey :
Text Classification Using Clustering:

This paper addresses the matter of learning to the classify texts by exploiting knowledge derived from every coaching job and testing sets. To accomplish this, cluster is used as a complementary step to text classification, and is applied not only to the coaching job set but put together to the testing set. This approach permits North yankee country to estimate the location of the testing examples and conjointly the structure of the whole dataset, that is not possible for associate inductive learner. The incorporation of the info succeeding from cluster to the easy BOW illustration of the texts is foretold to boost the performance of a classifier. Experiments conducted on tasks and datasets provided among the framework of the ECDL/PKDD2006 Challenge Discovery on bespoken spam filtering, demonstrate the effectiveness of the projected approach. The experiments show substantial enhancements on classification performance notably for tiny coaching job sets.

Fast Fuzzy Feature Clus Tering For Text Classification:

Feature bunch could be a powerful technique to cut back the spatiality of feature vectors for text classification. During this paper, quick Fuzzy Feature bunch for text classification is planned. it\'s supported the framework planned by Jung-Yi Jiang, RenJiaLiou and Shie-Jue Lee in 2011. The word within the feature vector of the document is sorted into the cluster in less iteration. The numbers of iterations needed to get cluster centers square measure reduced by remodeling clusters center dimension from n dimension to 2-dimension. Principle element Analysis with slit modification is employed for dimension reduction. Experimental results show that, this technique improve the performance by considerably reducing the quantity of iterations needed to get the cluster center. constant is being verified with 3 benchmark datasets.

Ant-Based Clustering Algorithms: A Brief Survey:

Ant-based bunch may be a biologically galvanized knowledge bunch technique. bunch task aims at the unsupervised classification of patterns in numerous teams. bunch drawback has been approached from completely different disciplines throughout last years.

In recent years, several algorithms are developed for resolution numerical and combinatorial optimization issues. Most promising among them area unit swarm intelligence algorithms.Bunch with swarm-based algorithms is rising as an alternate to a lot of standard bunch techniques. These algorithms have recently been shown to supply smart ends up in a large kind of real world applications. throughout the last 5 years, analysis on and with the ant-based bunch algorithms has reached a really promising state. During this paper, a short survey on ant-based bunch algorithms is delineated . We have a tendency to additionally gift some applications of hymenopteron primarily based bunch algorithms.

Related Work:

1. Classification
2. Stemming
3. CPI implementation
4. Frequency Calculation
5. Clustering

Classification:

Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done “manually” (or “intellectually”) or algorithmically. The intellectual classification of documents has mostly been the province of library science, while the algorithmic classification of documents is used mainly in information science and computer science. The problems are overlapping, however, and there is therefore also interdisciplinary research on document classification.

Stemming :

The Porter algorithmic program (or _Porter stemmer‘) could be a method for removing the somebody morphological and in flexional endings from words in English. Its main use is as a part of a term normalisation method that\'s typically done once fixing info Retrieval systems. In linguistic morphology and knowledge retrieval, stemming is that the method for reducing inflected (or typically derived) words to their stem, base or root type—generally a word form. The stem needn\'t be the image of the morphological root of the word; it\'s typically comfortable that connected words map to an equivalent stem, notwithstanding this stem is not in itself a sound root.

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

A stemmer for English, for instance, ought to establish the string "cats" (and presumably "cat like", "catty" etc.) as supported the basis "cat", and "stemmer", "stemming", "stemmed" as supported "stem". A algorithmic program reduces the words "fishing", "fished", "fish", and "fisher" to the basis word, "fish". On the opposite hand, "argue", "argued", "argues", "arguing", and "argus" scale back to the stem "argu" (illustrating the case wherever the stem is not itself a word or root) however "argument" and "arguments" scale back to the stem "argument".

CPI Implementation:

A set of documents x1; x2; . . . ; xn 2 IRn. Let X denote the document matrix. The algorithm for document clustering based on CPI can be summarized. Construct the local neighbor patch, and compute the matrices MS and MT .Project the document vectors into the SVD subspace by throwing away the zero singular values. The singular value decomposition of X can be written as X ¼ U_V T . Here all zero singular values in _ have been removed. Accordingly, the vectors in U and V that correspond to these zero singular values have been removed as well. Thus the document vectors in the SVD subspace can be obtained by ~X ¼ UTX.

Frequency Calculation:

The term frequency vector can be computed as follows:
1. Transform the documents to a list of terms after words stemming operations.
2. Remove stop words. Stop words2 are common words that contain no semantic content.
3. Compute the term frequency vector using the TF/IDF weighting scheme.

Clustering:

The experimental results of LPI and CPI on data set are obtained when the number of nearest neighbors is set to seven or eight. For Reuters data sets, the number of nearest neighbours used for LPI and CPI varies from 3 to 24. In all experiments, our algorithm performs better than or competitively with other algorithms.CPI algorithm has two essential parameters: the dimension of optimal semantic subspace and the number of nearest neighbors. Unfortunately, how to determine the optimal dimension of the semantic subspace is still an open problem. In typical spectral clustering, the dimension of semantic subspace is set to the number of clusters.

Results:

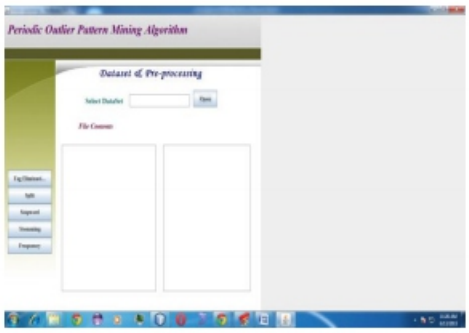


Fig: Home Page

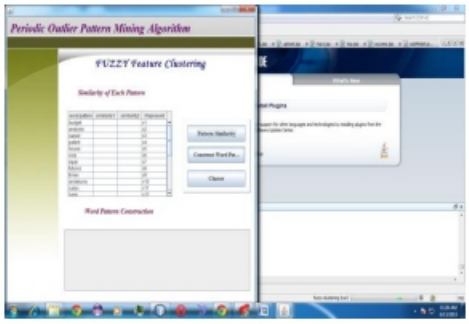


Fig: pattern recognition

Conclusion:

In this paper, we've conferred a completely unique rule for the periodicity detection of outlier, surprising, or uncommon patterns. Our notion of a stunning or uncommon pattern takes under consideration the frequency of a pattern with patterns of comparable length. The rule additionally takes under consideration the coverage space of the pattern and also the chance of pattern prevalence to classify it as associate outlier pattern.

This definition isn't restricted to the assumption that patterns involving less frequent events area unit uncommon patterns, as represented in nor it needs. To propose a new spectral clustering method called correlation preserving indexing (CPI), which is performed in the correlation similarity measure space, the documents are projected into a low-dimensional semantic space.

References :

[1] E. Keogh, J. Lin, and A. Fu, ―HOT SAX: Efficiently finding the most unusual time series subsequence,‖ in Proc. IEEE Int. Conf. Data Mining, Houston, TX, USA, Nov. 2005, pp. 226–233.

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

[2] N. Kumar, N. Lolla, E. Keogh, S. Lonardi, C. A. Ratanamahatana, and L. Wei, ―Time-series bitmaps: A practical visualization tool for working with large time series databases,‖ in Proc. SIAM Int. Conf. Data Mining, Newport Beach, CA, USA, 2005, pp. 531– 535.

[3] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid, ―Periodicity detection in time series databases,‖ IEEE Trans. Knowl. Data Eng., vol. 17, no. 7, pp. 875–887, Jul. 2005.

[4] M. G. Elfeky,W. G. Aref, and A. K. Elmagarmid, ―WARP: Time warping for periodicity detection,‖ in Proc. IEEE Int. Conf. Data Mining, Nov. 2005, pp. 8–15.

[5] J. Han, W. Gong, and Y. Yin, ―Mining segmentwise periodic patterns in time related databases,‖ in Proc. ACM Int. Conf. Knowl. Discov. Data Mining, vol. 8, no. 1, pp. 53–87, Aug. 1998.

[6] P. Indyk, N. Koudas, and S. Muthukrishnan, ―Identifying representative trends in massive time series data sets using sketches,‖ in Proc. Int. Conf. Very Large Data, Sep. 2000, pp. 363–372.

[7] S. Ma and J. Hellerstein, ―Mining partially periodic event patterns with unknown periods,‖ in Proc. IEEE Int. Conf.Data Eng.,Apr. 2001, pp. 205– 214.

[8] J.Yang,W.Wang, and P.Yu, ―InfoMiner+: Mining partial periodic patterns with gap penalties,‖ in Proc. IEEE Int. Conf. Data Mining, Dec. 2002, pp. 725– 728.