Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 6, Issue 1; January-March, 2019, pp. 49-54
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

Research Challenges in Time Series Data Mining: An Overview

**Nipan Kumar[1], Reena Mahali[2], Mr. Aira Kharvel Parida[3],
Mr. Pradeepta Kumar Dash[4], Dr. Sanjay Kumar Padhi[5] and
Dr. Subhendu Kumar Pani[6]**

[1]Research Scholar,
[2]Research Scholar,
[3]Lecturer, Dept. of IMSc. ETC, BJB Autonomous College
[4]HOD, Dept. of IMSc.ETC, BJB Autonomous College
[5]Associate Prof., Dept. of CSE, KIST, Bhubaneswar
[6]Associate Prof., Dept. of CSE, Orissa Engineering College

ABSTRACT—*In the context of data mining the feature size is very large and it is believed that it needs a bigger population. Hence, this translates directly into higher computational load. This work presents the review of the application of Deep Learning to financial series Data Mining and Learning Analytics. Data and information have become major assets for most of the organizations. The success of any organization depends largely on the extent to which the data acquired from business operations is utilized. Classification is an important task in KDD (knowledge discovery in databases) process. It has several potential applications.*

Keywords*: Data Mining, Particle Swarm Intelligence, Knowledge Discovery Databases.*

1. INTRODUCTION

A. DATA MINING AND TIME VARYING DATABASES.

In different kinds of information databases, such as scientific data, medical data, financial data, and marketing transaction data; analysis and finding critical hidden information has been a focused area for researchers of data mining[1][2][4]. How to effectively analyze and apply these data and find the critical hidden information from these databases, data mining technique has been the most widely discussed and frequently applied tool from recent decades. Although the data mining has been successfully applied in the areas of scientific analysis, business application, and medical research and its computational efficiency and accuracy are also improving, still manual works are required to complete the process of extraction. Data mining is considered to be an emerging technology that has made revolutionary change in the information world. The term `data mining' (often called as knowledge discovery) refers to the process of analysing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system[3]. Technically, "data mining is the process of finding correlations or patterns among dozens of fields in large relational databases". Therefore, data mining consists of major functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyse data using application tools and techniques, and meaningfully presents data to provide useful information.

Two important aspects of time series data mining can be identified as forecasting and classification. Time-series forecasting has been performed predominantly using statistical-based methods, for example, the linear autoregressive (AR) models because of their flexibility to model many stationary processes. These include the well-known ARMA (autoregressive moving average) model and its extensions by Weron and Misiorek for short-term time series forecasting. The ARMA model assumes a linear relationship between the lagged variables and produces only a coarse approximation to real-world complex systems and generally fails to accurately predict the evolution of nonlinear and non-stationary processes[22,23,24,25].

An Artificial Neural Network (ANN) [19] is a very simplified computational model of the Central Nervous System. It is a labelled oriented graph where the vertices are arranged in layers. The vertices are quite simple computational units that correspond to simplified biological neurons. They integrate, in a nonlinear fashion, the signals they receive from other units. According to the signal that they receive they compute their own activation level. The graph edges correspond to the synapses between units. They encode the influence of a given unit over any other unit that receives a connection from the former. Each given connection can either increase or decrease the receiving unit's activation and this influence is represented by the weight associated to the edge. Just like the Central

Nervous System, an ANN is a learning system; it learns to perform a task from data that exemplifies the task being performed. An ANN learning algorithm is a procedure to adapt the network's weights in a way that the network starts to perform a computational task, as expressed by the association of the activation of units in the input layer to the corresponding activation of units in the output layer. ANN learning algorithms comprise the full range of paradigms: unsupervised, supervised and reinforcement learning.

B. DATA MINING PROCESS

Data Mining is an iterative process consists of the following list of stages:

Data cleaning

Data integration

Data selection

Data transformation

Data mining

Pattern evaluation

Knowledge presentation

Data cleaning: This task handles missing and redundant data in the source file. The real world data can be incomplete, inconsistent and corrupted. In this process, missing values can be filled or removed, noise values are smoothed, outliers are identified and each of these deficiencies are handled by different techniques.

Data integration: Data integration process combines data from various sources. The source data can be multiple distinct databases having different data definitions. In this case, data integration process inserts data into a single coherent data store from these multiple data sources.

In the data selection process, the relevant data from data source are retrieved for data mining purposes.

Data transformation: This process converts source data into proper format for data mining. Data transformation includes basic data management tasks such as smoothing, aggregation, generalization, normalization and attributes construction.

Data mining: In Data mining process, intelligent methods are applied in order to extract data patterns. Pattern evaluation is the task of discovering interesting patterns among extracted pattern set. Knowledge representation includes visualization techniques, which are used to interpret discovered knowledge to the user.

Pattern Evaluation: During data mining, a large number of patterns may be discovered. However, all those patterns may not be useful in a particular context. It is highly required to assess the usefulness of the discovered patterns based on some criteria, so that truly useful and interesting patterns representing knowledge can be identified.

Knowledge Presentation: Finally, the mined knowledge has to be presented to the decision-maker using suitable techniques of knowledge representation and visualization.

2. EVOLUTIONARY TECHNIQUES FOR DATA MINING

1. GA For Data Mining:

Genetic algorithm is basically used in search, optimization and document. Evolutionary computing (EC) is an exciting development in computer science. It amounts to building, applying and studying algorithms based on the Darwinian principle of natural selection. Genetic algorithm is one of the components of EC. The common underlying idea behind GA is as follows: given a population of individuals, the environmental pressure causes natural selection (survival of the fittest) and here by the fitness of the population is growing. It is easy to see such process as optimization. Given an objective function to be maximized we can randomly create a set of candidate solutions and use the objective function as an abstract fitness measure (the higher the better) based on this fitness some of the better candidate are chosen to seed the next generation by applying recombination and mutation. Recombination is applied to two selected candidates, the so called parents and results in one or two new candidates, the children. Mutation is applied to one candidate and results in one new candidate. Applying recombination and mutation leads to set of new candidates, the offspring. Based on their fitness these offspring compete with the candidates for a place in the next generation. This process can be iterated until a solution is found or a previously set time limit is reached. The general scheme of a genetic algorithm can be given as below:

```
INITIALISE population with random individuals;
EVALUATE each candidate;
REPEAT UNTIL (TERMINATION CONDITION is satisfied)
SELECT genitors;
RECOMBINE pairs of genitors;
MUTATE the resulting offspring;
EVALUATE new born candidate;
SELECT individuals for the next generations;
END OF REPEAT.
```

Search and retrieval: This technique is used to relate to other related home pages and relevant document is retrieved. Query optimization

2. Pso for Data Mining

The original PSO was designed as a global version of the algorithm [9], that is, in the original PSO algorithm, each particle globally compares its fitness to the entire swarm population and adjusts its velocity towards the swarm„s global best particle. There are, however, recent versions of local/topological PSO algorithms, in which the comparison process is locally performed within a predetermined neighbourhood topology [7] [8] [9]. Unlike the original version of ACO the original PSO is designed to optimize real-value continuous problems, but the PSO algorithm has also been extended to optimize binary or discrete problems [10] [11] [12]. The original version of the PSO algorithm is essentially described by the following two simple ―velocity‖ and ―position‖ update equations, shown in 7 and 8 respectively.

$$vid(t+1)= vid(t) + c1\ R1(pid(t) - xid(t)) + c2\ R2\ (pgd(t) - xid(t))$$

$$xid(t+1) = xid(t) + vid(t+1)$$

Where:

 vid represents the rate of the position change (velocity) of the ith particle in the dth dimension, and t denotes the iteration counter.

 xid represents the position of the ith particle in the dth dimension. It is worth noting here that xi is referred to as the ith particle itself, or as a vector of its positions in all dimensions of the problem space. The n-dimensional problem space has a number of dimensions that equals to the numbers of variables of the desired fitness function to be optimized.

 pid represents the historically best position of the ith particle in the dth dimension (or, the position giving the best ever fitness value attained by xi).

Algorithm 1: Basic flow of PSO

1) Initialize the swarm by randomly assigning each particle to an arbitrarily initial velocity and a position in each dimension of the solution space.

2) Evaluate the desired fitness function to be optimized for each particle„s position.

3) For each individual particle, update its historically best position so far, Pi, if its current position is better than its historically best one.

4) Identify/Update the swarm„s globally best particle that has the swarm„s best fitness value, and set/reset its index as g and its position at Pg.

5) Update the velocities of all the particles using above first equation.

6) Move each particle to its new position using above second equation .

7) Repeat steps 2–6 until convergence or a stopping criterion is met (e.g., the maximum number of allowed iterations is reached; a sufficiently good fitness value is achieved; or the algorithm has not improved its performance for a number of consecutive iterations).

3. APPLICATION AREAS

There are several applications of data mining. Some common used applications of data mining are given below:

a) Fraud or non-compliance anomaly detection: Data mining isolates the factors that lead to fraud, waste and abuse. The process of compliance monitoring for anomaly detection (CMAD) involves a primary monitoring system comparing some predetermined conditions of acceptance with the actual data or event. If any variance is detected (an anomaly) by the primary monitoring system then an exception report or alert is produced, identifying the specific variance.

For instance credit card fraud detection monitoring, privacy compliance monitoring, and target auditing or investigative efforts can be done more effectively [5].

b) Intrusion detection: It is a passive approach to security as it monitors information systems and raises alarms when security violations are detected. This process monitors and analyzes the events occurring in a computer system in order to detect signs of security problems. Intrusion detection systems (IDSs) may be either host based or network based,according to the kind of input information they analyze [6]. Over the last few years, increasing number of research projects(MADAM-ID, ADAM, Clustering project, etc.) have been applied data mining approaches (either host based or network based) to various problems (construction of operational IDSs, clustering audit log records, etc.) of intrusion detection [13].

c) Lie detection (SAS Text Miner): SAS institute introduced lie-detecting software, called SAS Text Miner. Using intelligence of this tool, managers can be able to detect automatically when email or web information contains lies. Here data mining can be applied successfully toidentify uncertainty in a deal or angry customers and also have many other potential applications [14]. Many other market mining tools are also available in real practice viz. Clementine, IBM's Intelligent Miner, SGI's MineSet, SAS's Enterprise Miner, but all pretty much the same set of tools.

d) Market basket analysis (MBA): Basically it applies data mining technique in understanding what items are likely to be purchased together according to association rules, primarily with the aim of identifying cross-selling opportunities. Sometimes it is also referred to as product affinity analysis. MBA gives clues as to what a customer might have bought if an idea had occurred to them. So, it can be used in deciding the location and promotion of goods by means of combo-package and also can be applied to the areas like analysis of telephone calling patterns, identification of fraudulent medical insurance claims, etc. [15].

e) Aid to marketing or retailing: Data mining could help direct marketers by providing useful and accurate trends on purchasing behavior of their customers and also help them in predicting which products their customers may be interested in buying. In addition, trends explored by data mining help retail-store managers to arrange shelves, stock certain items, or provide a certain discount that will attract their customers. In fact data mining allows companies to identify their best customers, attract customers, aware customers via mail marketing, and maximize profitability by means of identifying profitable customers [16].

f) Customer segmentation and targeted marketing: Data mining can be used in grouping or clustering customers based on the behaviors (like payment history, etc.), which in turn helps in customer relationship management (epiphany) and performs targeted marketing. Usually it becomes useful to define similar customers in a cluster, holding on good customers, weeding out bad customers, identify likely responders for business promotions.

g) Phenomena of "`beer and baby diapers"': This story of using data mining to find a relation between beer and diapers is told, retold and added to like any other legend. The explanation goes that when fathers are sent out on an errand to buy diapers, they often purchase a six-pack of their favorite beer as a reward. An article in The Financial Times of London (Feb. 7, 1996) stated, "The oft-quoted example of what data mining can achieve is the case of a large US supermarket chain which discovered a strong association for many customers between a brand of babies nappies (diapers) and a brand of beer [17].

h) Financial, banking and credit or risk scoring: Data mining can assist financial institutions in various ways, such as credit reporting, credit rating, loan or credit card approval by predicting good customers, risk on sanctioning loan, mode of service delivery and customer retention (i.e. build profiles of customers likely to use which services), and many others. A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. In addition, data mining can also assist credit card issuers in detecting potentially fraudulent credit card transaction. In general, data mining methods such as neural networks and decision trees can be a useful addition to the techniques available to the financial analyst [18].

i) Medicare and health care: Applying data mining techniques, it is possible to find relationship between diseases, effectiveness of treatments, to identify new drugs, market activities in drug delivery services, etc. However, a pharmaceutical company can analyze its recent sales to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. Such dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sale situation.

4. LITERATURE REVIEWS

Over the past few decades, artificial neural networks (ANNs) those exhibit superior performance on classification and regression problems in machine learning domain have attracted tremendous attention in the time series forecasting community. Compared to statistics-based forecasting techniques, neural network approaches have several unique characteristics, including: 1) being both nonlinear and data driven; 2)having no requirement for an explicit underlying model (nonparametric); and 3) beingmore flexible and universal, thus applicable to more complicated models. Thus, neural networks have been used extensively for a wide range of applications in time series forecasting varying from financial, economic, to energy systems, earthquakes, and weather. These models do not need prior assumptions on the form of nonlinearity and are known as universal approximators (Park and Sandberg, 1991) since they can approximate any continuous function to an arbitrary precision. A recent review of NN models for time series forecasting has been provided by Zhang (2012). Feed-forward Neural Network models (FNNs) parameterized with a back-propagation algorithm has been employed for nonlinear time series forecasting (Lapedes and Farber, 1987; French et al., 1992). They are known to outperform traditional statistical methods such as regression and Box–Jenkins approaches in functional approximation, but they assume the dynamics underlying time series are time-invariant. FNNs with recurrent feedback connections have also been attempted for time series forecasting (20). Such dynamic Recurrent NN (RNN) models allow forecasting of nonlinear time series occurring in various fields (21).

5. RESEARCH TRENDS AND ISSUES

Time series data mining has been an ever growing and stimulating field of study that has continuously raised challenges and research issues over the past decade. We discuss in the following open research issues and trends in time series data mining for the next decade.

Stream analysis. The last years of research in hardware and network research has witnessed an explosion of streaming technologies with the continuous advances of bandwidth capabilities. Streams are seen as continuously generated measurements which have to be processed in massive and fluctuating data rates. Analyzing and mining such data flows are computationally extreme tasks .Several papers review research issues for data streams mining or management. Algorithms designed for static datasets have usually not been sufficiently optimized to be capable of handling such continuous volumes of data. Many models have already been extended to control data streams, such as clustering classification segmentation or anomaly detection . Novel techniques will be required and they should be designed specifically to cope with the ever flowing data streams.

Convergence and hybrid approaches. A lot of new tasks can be derived through a relatively easy combination of the already existing tasks. For instance, three approaches, polynomial, DFT and probabilistic, to predict the unknown values that have not fed into the system and answer queries based on forecast data. This approach is a combination of prediction query by content over data streams. This work shows that future research has to rely on the convergence of several tasks. This could potentially lead to powerful hybrid approaches.

Embedded systems and resource-constrained environments. With the advances in hardware miniaturization, new requirements are imposed on analysis techniques and algorithms. Two main types of constraints should absolutely be met when hardware is inherently limited. First, embedded systems have a very limited memory space and cannot have permanent access to it. However, most method use disk-resident data to analyze any incoming information's. Furthermore, sensor networks (which are frequently used in embedded systems) usually generate huge amounts of streaming data. So there is a vital need to design space efficient techniques, in terms of memory consumption as well as number of accesses. An interesting solution has been recently proposed in . The algorithm is termed auto cannibalistic, meaning that it is able to dynamically delete parts of itself to make room for new data. Second, as these resource-constrained environments are often required to be autonomous, minimizing energy consumption is another vital requirement. [Bhargava et al. 2003] has shown that sending measurements to a central site in order to process huge amounts of data is energy inefficient and lack scalability.

Data mining theory and formalization. A formalization of data mining would drastically enhance potential reasoning on design and development of algorithms through the use of a solid mathematical foundation. The examined the possibility of a more general theory of data mining that could be as useful as relational algebra is for database theory. They studied the link between data mining and Kolmogorov complexity by showing their close relatedness. They conclude from the undesirability of the latter that data mining will never be automated, and therefore stating that "data mining will always be an art". However, a mathematical formalization could lead to global improvements of both reasoning and the evaluation of future research in this topic.

Parameter-free data mining. One of the major problems affecting time series systems is the large numbers of parameters induced by the method. The user is usually forced to "fine-tune" the settings in order to obtain best performances. However, this tuning highly depends on the dataset and parameters are not likely to be explicit. Thus, parameter-free systems is one of the key issue that has to be addressed.

Adaptive mining algorithm dynamics. Users are not always interested in the results of a simple mining task and prefer to focus on evolution of these results in time. This actually represents the dynamics of a time series data mining system. This kind of study is of particular relevance in the context of data streams.

Exhaustive benchmarking. A wide range of systems and algorithms has been proposed over the past few years. Individual proposals are usually submitted together with specific datasets and evaluation methods that prove the superiority of the new algorithm. As for most scientific research, trying to find the solution to a problem often leads to raising more questions than finding answers. We have thus outlined several trends and research directions as well as open issues for the near future. The topic of time series data mining still raises a set of open questions and the interest of such research sometimes lies more in the open questions than the answers that could be provided.

6. CONCLUSION

To overview, evolution, parameter and the applications of GA and PSO are presented in a simple way. Although PSO has been used mainly to solve unconstrained, single objective optimization problems, PSO algorithms have been developed mainly to solve constrained problems, multi objective optimization problems and problems with dynamically changing landscapes and to find multiple solutions.

REFERENCES

[1]Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[2]Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning, Vol. 42, No.3, pp.203-231, 2001.

[3]Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005.

[4]Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.

[5]Goldschmidt P S, Compliance monitoring for anomaly detection, Patent no. US 6983266 B1, issue date January 3, 2006, Available at: www.freepatentsonline.com/6983266.html

[6]Bace R, Intrusion Detection, Macmillan Technical Publishing, 2000.

[7] J. Kennedy, R. C. Eberhart, and Y. Shi, Swarm Intelligence, Morgan Kaufmann, San Francisco, CA, 2001.

[8] J. Kennedy, Small worlds and mega-minds: Effects of neighborhood topology on particle swarm performance, In Proceeding of the 1999 Conference on Evolutionary Computation, pp. 1931-1938, 1999.

[9]J. Kennedy and R. Mendes, Population structure and particle swarm performance, Proceeding of the 2002 Congress on Evolutionary Computation, Honolulu, Hawaii, May 2002.

[10]J. Kennedy and R. C. Eberhart, A discrete binary version of the particle swarm algorithm, in Proceeding of the 1997 Conference on Systems, Man, and Cybernetics, pp. 4104-4109, 1997.

[11] C. K. Mohan and B. Al-kazemi, Discrete particle swarm optimization, Proceedings of the Workshop on Particle Swarm Optimization, Indianapolis, IN, 2001.

[12]D. K. Agrafiotis and W. Cedeño, Feature selection for structure-activity correlation using binary particle swarms, Journal of Medicinal Chemistry,Vol. 45, pp. 1098-1107, 2002

[13] Smyth P, Breaking out of the Black-Box: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in DataMining and Knowledge Discovery (DMKD-2001), held on May 20 (2001), Santra Barbara, California, USA.

[14]SAS Institute Inc., Lie detector software: SAS Text Miner (product announcement), Information Age Magazine, [London, UK], February 10 (2002), Available at: http://www.sas.com/solutions/fraud/index.html.

[15]Berry M J A and Linoff G S, Data mining techniques: for marketing, sales, and relationship management, 2 nd edn (John Wiley; New York), 2004.

[16]Delmater R and Hancock M, Data mining explained: a manager's guide to customer-centric business intelligence, (Digital Press, Boston), 2002.

[17]Fuchs G, Data Mining: if only it really were about Beer and Diapers, Information Management Online, July 1, (2004), Available at: http://www.information-management.com/news/1006133-1.html.

[18]Langdell S, Use of data mining in financial applications, (Data Analysis and Visualization Group at NAG Ltd.), Available at: http://www.nag.co.uk/IndustryArticles/DMinFinancialApps.

[19].Haykin, S. (2008) Neural Networks and Learning Machines. 3. ed, Pearson.

[20]De Groot, C. and Wuertz, D. (1991) Analysis of univariate time series with connectionist nets: a case study of two classical examples. Neurocomputing, 3(4), 177–192.

[21].Grudnitski, G. and Osburn, L. (1993) Forecasting S&P and gold futures prices: an application of neural networks. Journal of Futures Markets, 13(6), 631–643.

[22] R. Chandra, “Competition and collaboration in cooperative coevolution of Elman recurrent neural networks for time-series prediction,” Neural Networks and Learning Systems, IEEE Transactions on, vol. 26, pp. 3123–3136, 2015.

[23].R. Nand and R. Chandra, “Coevolutionary feature selection and reconstruction in neuro-evolution for time series prediction,” in Artificial Life and Computational Intelligence - Second Australasian Conference,ACALCI 2016, Canberra, ACT, Australia, February 2-5, 2016, Proceedings,2016, pp. 285–297.

[24].S. B. Taieb and A. F. Atiya, “A bias and variance analysis for multistepahead time series forecasting,” 2015.

[25]H. Zheng, X. Geng, D. Tao, and Z. Jin, “A multi-task model for simultaneous face identification and facial expression recognition,” Neurocomputing,vol. 171, pp. 515 – 523, 2016.