A Review of Data Security Primitives in Data Mining

Asmita Singh, Anchal Pokharana

Department of CE, Poornima University, Jaipur, Rajasthan, India

Abstract—*This paper has discussed various issues and security primitives like Spatial Data Handing, Privacy Protection of data, Data Load Balancing, Resource Mining etc. in the area of Data Mining.A 5-stage review process has been conductedfor 30 research papers which were published in the period of year ranging from 1996 to year 2013. After an exhaustive review process, nine key issues were found "Spatial Data Handing, Data Load Balancing, Resource Mining ,Visual Data Mining, Data Clusters Mining, Privacy Preservation, Mining of gaps between business tools & patterns, Mining of hidden complex patterns." which have been resolved and explained with proper methodologies. Several solution approaches have been discussed in the 30 papers. This paper provides an outcome of the review which is in the form of various findings, found under various key issues. The findings included algorithms and methodologies used by researchers along with their strengths and weaknesses and the scope for the future work in the area.*

Keywords—Data load balancing , Privacy , D3M , AKD , Data Hiding.

I. INTRODUCTION

To generate information it requires massive collection of data. The data can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia data and hypertext documents. With enormous amount of data stored in files, database and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in Decision making. Data mining is a set of activities used to find new, hidden or unexpected patterns in data or unusual patterns in data. Data mining tools predict future trends and behaviors,allowing businesses to make proactive, knowledge-drivendecisions. The automated, prospective analyses offered bydata mining move beyond the analyses of past events provided by retrospective tools typical of decision supportsystems. Data mining tools can answer business questionsthat traditionally were too time consuming to resolve.They search databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses. Internal auditors need to be aware of the different kinds of data mining tools available and recommend the purchase of a tool that matches the organization's current detective needs. Data mining commonly involves four classes of tasks[7].

Classification - Arranges the data into predefined groups .For example an email program might attempt to classify an email as legitimate or spam. Common algorithms include Decision Tree Learning, Nearest neighbour, naive Bayesian classification and Neural network[2].

Clustering - Is like classification but the groups are not predefined, so the algorithm will try to group similar items together[2].

Regression - Attempts to find a function which models the data with the least error[2].

Association rule learning - 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"[2].

II. REVIEW PROCESS ADOPTED

This review process approach was divided into five stages in order to make the process simple and adaptable. The stages were:-

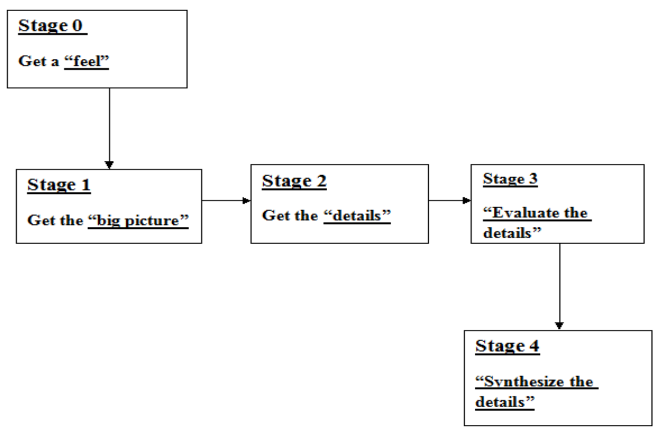

Fig1: Review Process Adopted

Stage 0: Get a "feel":

This stage provides the details to be checked while starting literature survey with a broader domain and classifying them according to requirements.

Stage 1: Get the "big picture"

The groups of research papers are prepared according to common issues & application sub areas. In order to understand the paper, it is necessary to find out the answers to certain questions by reading the Title, Abstract, introduction, conclusion and section and sub section headings.

Stage 2: Get the "details"

Stage 2 deals with going in depth of each research paper and understand the details of methodology used to justify the problem, justification to significance & novelty of the solution approach, precise question addressed, major contribution, scope & limitations of the work presented.

Stage 3: "Evaluate the details"

This stage evaluates the details in relation to significance of the problem, Novelty of the problem, significance of the solution, novelty in approach, validity of claims etc.

Stage 3+: "Synthesize the detail"

Stage 3+ deals with evaluation of the details presented and generalization to some extent. This stage deals with synthesis of the data, concept & the results presented by the authors

III. VARIOUS ISSUES IN THE AREA

After reviewing 30 research papers on Data Mining we have found following issues, which have been listed as under. The issues are:

1) **Spatial Data Handling and Mining**
2) **Gap between various hidden patterns and business tools**
3) **Problem of decision making in heterogeneous data bases**
4) **Problem of Resource Mining**
5) **Problem of mining of visually interactive data**
6) **Problem of mining of data clusters**
7) **Mining of data in terms of load balancing and data fittability**
8) **Problem of protecting and preserving data**
9) **Mining of various complex and hidden patterns of data**

IV. ISSUE WISE DISCUSSION

Issue 1:- Spatial Data Handling and Mining.

Some approaches were used for this issue which is spatial clustering,spatial classification, spatial characterization and spatio-temporal association rule mining are performed for spatial data mining. Data model of the spatial data cube has also been proposed..The selection in spatial data cube performed at cuboids level. For better sup By these solution approaches , spatial data can be properly handled and mined.

Issue 2:- Gap between various hidden patterns and business tools.

Task driven data mining, domain driven data mining (D3M) and actionable knowledge discovery (AKD) database are the approaches that have been given . Task driven data mining system involves seven elements such as data warehousing, data pre -processing, feature subset selecting, modelling, model evaluating, model updating and model releasing

D3M solves the developing problem of areas by many intelligence methods. It ensures decision making at the same time from the different fields

Issue 3:- Problem of decision making in heterogeneous data bases.

The technique of "Intelligent data mining system" for bio database solves the problem of evaluation and analysis of bio data and decision making process. It collects data from distributed data bases and provides integrated data, which is used with other data for analysis purpose then extract valid, relevant information from bio databases.

Issue 4:- Problem of Resource Mining.

Model driven data mining effectively mines the categories of data in oil and gas exploration and production. Various methodologies like model driven data mining,Intrusion Detection, Predictive Data Mining, Descriptive Data Mining, Clustering, E-commerce, Web mining and Business Intelligence perfectly explain and mine the resources efficiently.

Issue 5:- Problem of mining of visually interactive data.

A mechanism of bootstrapping data mining with visualization has been provided. A smooth interface between visualization and data mining is built & a flexible tool to explore and query temporal data derived from raw multimedia data.

Issue 6:- Problem of mining of data clusters.

A data clustering method named BIRCH has been demonstrated which is highly efficient for clustering large sized data bases. Another approach i.e. the CLARANS algorithm is used to cluster the set of compound objects. This algorithm is relational in the sense that it takes relational data as input and does the proper mining of data clusters.

Issue 7:- Mining of data in terms of load balancing and data fittability.

Data type generalization process has been devised. Map ND strategy is used for solving the problem. Parallel data mining for data mapping has been used.Development of data mining language,data mining query languages like DMQL and TDML has been done for mining relational databases.

Issue 8:-Problem of preserving and protecting data.

Two approaches namely "perturbation approach" and "k-anonymity" have been proposed. K-anonymity requires each record in an anonymized table to be indistinguishable with at least *k*-1 other records within the dataset. In perturbation approach,distribution of each data dimension is reconstructed independently. Two techniques namely "Cryptographic techniques & Randomized Response techniques" have also been proposed.

Issue 9:-Mining of various complex and hidden patterns of data.

Artificial neural network system is devised to formulate such problems. Artificial neural network provide robustness&data parallelism in processing .Various neural network techniques like maps, neuro fuzzy logic,adaptive resonance theory, neuro-computing&natural intelligent systems have been given.

V. ISSUE WISE SOLUTION APPROACHES USED

The solution approaches under the various issues have been shown in the **Table I to IX,** which includes additional information like hardware, software, variable/parameters used along with results obtained. The same table also describes the comparative analysis between various solution approaches.

VI. ISSUE WISE DISCUSSION ON RESULTS

Table I Spatial Data Handling and Mining

Solution Approach	Results	Ref
Spatial classification, & Characterization, Spatio-temporal association- rule mining.	Extracts spatial patterns and knowledge from a spatial database.	[3]
Model of Spatial data cube	Supports both spatial and non -spatial data and mines data at global level	[1]

Table II Gap between Various Hidden Patterns and Business Tools

Solution Approach	Results	Ref.
Domain Driven Data Mining and Actionable Knowledge Discovery Database .	D3M construct next-generation methodology which solves the real world problems.	[6]
Task Driven Data Mining and .	It rationally combines domain knowledge with mining methods.	[8]

Table III Problem of Decision Making in Heterogeneous Data bases

Solution Approach	Results	Ref
Sensor based intelligent mining & Environmental monitoring ontology based reasoning architecture	Early warning systems which helps in mining of geological data	[12]
Knowledge Discovery Database .	Evaluation and analysis of valid , relevant information from bio databases.	[12]

Table IV Problem of Resource Mining

Solution Approach	Results	Ref
Model driven data mining is used	Determines the reliability and practicality of the mining outcome	24

Table V Problem of Mining of Visually Interactive Data

Solution Approach	Results	Ref
A mechanism of bootstrapping data & smooth interface between visualization and data mining	Examine and synthesize information into new ideas and hypotheses &test the insights gained from visualization.	[15]

Table VI Problem of Mining of Data Clusters

Solution Approach	Results	Ref
Clustering algorithm using IBM I-Miner has been used	Consistency of data is maintained.	[4]

BIRCH data clustering method has been used.	Gives correct output at the first scan of data	[4]

Table VII Mining of Data in Terms of Load Balancing and Data Fittability

Solution Approach	Results	Ref
MapND strategy &Parallel data mining	Minimum time cost has been achieved..	[10]

Table VIII Problem of Protecting and Preserving Data

Solution Approach	Results	Ref
Bottom up generalization technique ,Cryptographic technique & Randomized response techniques	Proper Surveying of the relationships between data forms, and then analysis has been done	[16]

Table XI Mining of Various Complex and Hidden Patterns of Data

Solution Approach	Results	Ref
Artificial Neural network techniques like maps, neuro fuzzy logic, parallel distributed processing, natural intelligent systems	Good robustness, , adaptive parallel processing, distributed storage and high degree of fault tolerance have been achieved.	[18]

VII. COMMON FINDINGS

Issue 1:- Spatial Data Handling and Mining

- ❖ The best solution Approach is " Proposal of data model of spatial data cube" because dimensions and measure of the spatial data cube are extended to support both spatial and non-spatial data.
- ❖ The worst Approach is spatial classification and spatio-temporal association mining because they require time consuming computations and available analytical operations are limited in them.

Issue 2:- Gap between various hidden patterns and business tools

- ❖ The best approach is Task driven data mining because it is independent of type of data & is operational and depends upon the tasks carried out on data.
- ❖ The worst approach is of domain driven data mining method. it is driven by the data & depends entirely on the domain knowledge of extracted data.

Issue 3:- Problem of decision making in heterogeneous data bases.

- ❖ The best approach is Ontology based approach to intelligent data mining for sensor networks because the ability of the sensor networks to collect information accurately enables building both real-time detection and early warning systems.
- ❖ Worst approach is Traditional data analysis techniques because of insufficiency and could not support and handle huge and complex biological data.

Issue 4:- Problem of Resource Mining.

- ❖ In Fourth Issue the best approach is Model driven data mining because effectively mines petro physical data, geological data, seismic data and logging data by mining actionable knowledge
- ❖ The worst approach is Temporal association mining Because they require time consuming computations and available analytical operations which are lesser in numbers.

Issue 5:- Problem of mining of visually interactive data

- ❖ In Fifth Issue the best approach is Model driven data mining because effectively mines petro physical data, geological data, seismic data and logging data by mining actionable knowledge
- ❖ The worst approach is Temporal association mining Because they require time consuming computations and available analytical operations which are lesser in numbers.

Issue 6:- Problem of mining of visually interactive data

- ❖ In Sixth Issue the best approach is bootstrapping mechanism because it allows users to easily examine and synthesize & test the insights gained from visualization.

  The worst approach is Information visualization. Because it involves problems like navigation between spaces and transferability that need to be satisfied.

Issue 7:- Mining of data in terms of load balancing and data fittability

- ❖ In Seventh Issue the best approach is MapND strategy because solve the problem of data load balancing for data mining nodes, and improves the performance of parallel data mining in grid and minimize the time cost.

  The worst approach is Knowledge discovery database technique Because the size of distributed database increase & it results in inflexible results

Issue 8:- Problem of protecting and preserving data

- ❖ In Eighth Issue the best approach is bottom-up generalization technique because incorporates partially the requirement of a targeted data mining task into the process of masking data so that essential structure is preserved in the masked data
- ❖ The worst approach is perturbation approach Because it does not reconstruct the original data values but only distributions, it is inefficient.

Issue 9:- Mining of various complex and hidden patterns of data

- ❖ In Ninth Issue the best approach is Artificial neural network technique because mines the data with utmost accuracy and will make the data noise tolerant
  The worst approach is General framework Because it is inefficient and often results in inconsistent mining of data because it generally applied on uncertain data sets.

VIII. SCOPE OF WORK IN AREA

- ❖ Particle swarm optimization, Ant colony optimization can be integrated with artificial neural network to further enhance the performance of ANN in Data mining.
- ❖ To increase the flexibility to be compatible with data mining, our system allows users to use any programming language to obtain new results. Thus, data researchers can implement new data mining algorithms using their own analysis tools (from Matlab and to C/C++) as far as users write the results into text files with pre-defined formats.
- ❖ The insights from visualization can be used to guide further data mining. Meanwhile, the results from the next round of data mining can be visualized which allows users to obtain new insights and develop more hypotheses with the data.

IX. CONCLUSION

The review of 30 research papers has been carried out in the area of Data Mining to investigate and find out current challenges and scope of work. After the review, we found several issues which should be given proper concern, when the effective mining of data takes place. These papers are a survey of different mining issues that affect the related work that carried out in the area of data mining. Purpose of these methods and techniques is to reduce the mining inefficiencies that occurs while mining of data and to improve system reliability. We have found various nine issues for which specific methods and techniques have been discussed.

The exhaustive review has finally led to extract findings in the area of Data Mining, strengths and weaknesses and scope of work during M. Tech 1st semester Research work.

REFERENCES

[1] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth,1996, "Advances in Knowledge Discovery and data Mining, In IEEE, International conference on data mining, CA, pp. 1-34, 1996.

[2] Jiawei Han, MichelineKamber, 1996 "Data Mining Concepts and Techniques, Second Edition, In IEEE, International conference on data mining, CA, pp. 9-34.

[3] M.S. Chen, J.W. Han and Philip S. Yu, 1996 " Data Mining: "An Overview from a Database Perspective", IEEE conference on Knowledge and Data Engineering pp. 866-883,

[4] Kusiak, A., Kernstine, K.H., Kern, J.A., McLaughlin, K.A., and Tseng, T.L.,2000 "Data Mining: Medical and Engineering Case Studies". The International conference on data mining , pp. 1-7,May 21-23

[5] R. D. Stevens, P. G. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. Paton, C. A. Goble, and A. Brass. Tambis,2000: "Transparent access to multiple bioinformatics information sources.Bioinformatics", 16:200–0.

[6] M. Stundner and J. S. Al-Thuwaini,2001. "How Data-Driven Modelling Methods Like Neural Network scan Help to Integrate Different Types of Data into Reservoir Management", The International conference on data mining SPE68163.

[7] Antonie, M. L., Zaiane, O. R.,Coman, A. 2001, "Application of Data Mining Techniques for Medical Image Classification" ,Proceedings of the Second International Workshop on Multimedia Data Mining"(MDM/KDD 2001) in conjunction with ACM SIGKDD conference, San Francisco.

[8] Panel members ,2002," The Perfect Data Mining Tool: Automated or Interactive", IEEE conference on data mining

[9] Y.Y. Yao,2003 A Step Towards the Foundations of Data Mining, Data Mining and Knowledge Discovery: Theory, Tools, Technology V, B.V. Dasarathy(ed.), The International conference on data mining, pp.254-263 .

[10] C. Rosse and J. L. V. Mejino.,2003 : "A reference ontology for biomedical informatics: the foundational model of anatomy." J. of Biomedical Informatics, 36(6):478–500, December

[11] Y.Y. Yao, N. Zhong and Y. Zhao,2004 " A Three-layered Conceptual Framework of Data Mining, IEEE, International conference on data mining ,pp.215-221

[12] R. Mizoguchi., 22(2), 2004 Tutorial on ontological engineering - part 3: Advanced course of ontological engineering. New Generation Comput.

[13] I. H. Witten and E. Frank.,2005 " Data Mining: Practical Machine Learning Tools and Techniques".

Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition.

[14] B. Smith, W. Ceusters, B. Klagges, J. Kohler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse,2005" Relations in biomedical ontologies. Genome Biology".

[15] R. Ramakrishnan, R. Agrawal, J.-C. Freytag, T. Bollinger,C. W. Clifton, S. Dzeroski, J. Hipp, D. Keim, S. Kramer,H.-P. Kriegel, U. Leser, B. Liu, H. Mannila, R. Meo, S. Morishita, R. Ng, J. Pei, P. Raghavan, M. Spiliopoulou,J. Srivastava, and V. Torra. Data mining, 2005: "The nextgeneration. In R. Agrawal, J. C. Freytag, and R. Ramakrishnan,editors, Perspectives Workshop: Data Mining:

[16] The Next Generation, number 04292 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany",Z.Y. He, X.F. Xu, and S.C. Deng, 2005 "Data Mining for Actionable Knowledge: A Survey, Technical Report: In IEEE, International conference on data mining 0501079.

[17] L.B. Cao, L. Lin and C.Q. Zhang, 2005 Domain-Driven In-Depth Pattern Discovery: A Practical Methodology, IEEE conference on data mining .

[18] Q. Yang and X.Wu,2006 " 10 challenging problems in data mining research". International Journal of Information Technology and Decision Making, 5(4):597–604.

[19] L. N. Soldatova and R. D. King.An,2006 "ontology of scientific experiments. Journal of the Royal Society Interface", 3(11):795–803

[20] L.B. Cao and C.Q. Zhang, 2006 "Domain-Driven Actionable Knowledge Discovery in the Real World", International conference on data mining pp. 821-830 .

[21] S.-A. Sansone et al,2007. Metabolomics standards initiative – "ontology working group.work in progress. Metabolomics",3(3):249–256

[22] D. Schober,W. Kusnierczyk, S. E. Lewis, and J. Lomax.,2007 "Towardsnaming conventions for use in controlled vocabularyand ontology engineering. In Proceedings of BioOntologies" SIG, ISMB 2007, pages 29–32

[23] B. Smith and N. Shah ,2007 " Ontologies for biomedicine – how to make them and use them." IEEE conference at ISMB/ECCB

[24] C. F. Taylor et al, 2007. "The minimum information about aproteomics experiment (miape). Nature Biotechnology",(25):887 – 893

[25] M. ˇZ ´akov´a, P. Kremen, F.,Zelezn´y, and N. Lavraˇc.,2008 "Planning to learn with a knowledge discovery ontology". In P. Brazdil, A. Bernstein, and L. Hunter, editors, Proceedings of the Second Planning to Learn Workshop (PlanLearn) at the ICML/COLT/UAI , pages 29–34, 2008

[26] G.Y. Wang and Y. Wang.2008" Domain-Oriented Data-Driven Data Mining: A New Understanding for Data Mining, Journal of Chongqing University of Posta and Telecommunications "(Natural Science Edition): , pp.266-271.

[27] Y.Y. Yao, N. Zhong and Y. Zhao, 2008 "A Conceptual Framework of Data Mining, Studies in Computational Intelligence (SCI) IEEE conference on data mining " 118 , pp.501-515, 2008.

[28] L. N. Soldatova, W. Aubrey, R. D. King, and A. Clare.,2008" The exact description of biomedical protocols. Bioinformatics",24(13).

[29] C. Wang, Q. Wang, K. Ren, and W. Lou, 2009 "Ensuring Data Security in Data Mining ,", IEEE, pp. 1-9 Communications

[30] M. Zang and L.Wu, 2009 " 10 challenging problems in data mining research". International Journal of Information Technology and Decision Making, 5(4):597–604.