

Soft computing in data mining, a young and interdisciplinary area of computer science used as a problem solving field

Niyati Gupta, Monali Jindal, Palak Makhija, Shruti Agarwal

Abstract— With the originations in science and technology and the advanced tools there has been an outburst outgrowth in transitory and stored data and hence to transfer the huge measure of data into utile information. Mining refers to the evocation of data while soft computing covers the processing of information. In this paper we examine ,data mining as a practical method for solving various issues like providing goal of enhanced yield in semiconductor manufacturing ,privacy security , use in banking industry, application in broadcasting effect to figure equivalent surface, for deliberately harmful code detection, dupery detection in accounting, in geosciences data, in heart disease diagnosis and treatment. Our examination manifests diverse positive applications of data mining, proving it to be one of the most versatile technologies for problem figuring as well as for providing a wide range of implementation. And in all of the above implementations the usefulness of the distinctly separate methodologies of soft computing are highlighted. Mostly, fuzzy sets are employed for pattern recognition. Neural networks do not include any estimation of the parameters and demonstrate good discovering and abstraction potentialities in data-rich environments.

Index Terms— Data mining, soft computing, code detection, pattern-recognition, fuzzy –sets, neural networks.



1. Introduction

Data mining is a lately, struck expression for merging ideas from statistics and computer science (database methods and gaining machine knowledge) applied to large databases in business ,science, and engineering. Basically, there are two questions raised for consideration, to what place data mining applications are deployed and they are: - more beneficial employment of data assets and customer relationship. A diverseness has been depicted in the definitions of data mining but the one which is acknowledged is “Data mining is the process of ascertaining purposeful new correlations, trends and patterns by separating big quantities of data put in repositories, practising statistical and mathematical techniques and pattern recognition technologies”.

The several applications of data mining can be categorized into two: micro, single mining component and macro, multiple element server-based systems.

The other terminology which holds a great importance here is soft computing which for both Knowledge Discovery and Data Mining introduces theoretical go about and practical computing methods broadening the package of issues that data mining can figure out and solve in an efficient manner. Adverting to this synergistic

assembly, the introductory virtues of data mining and soft computing paradigms are indicated and new data mining execution and implementation twinned to a soft computing approach for knowledge discovery is posed

2. Intervention and diagnosis of heart disease using Data Mining techniques

In the past ten years, heart attacks have been indicating the most important headings as a ground of deaths all over the world [1]. The European Public Health Alliance accounted that stroke, heart attacks, and other circulatory diseases report for 41% of all deaths [2].

Various researchers, to assist health care professionals in the diagnosis of heart disease are using statistical and data mining tools. The data mining techniques being used by health care professionals for curing hearts diseases are: naïve bayes, neural network and decision tree. And bagging algorithm, kernel density, support vector machine and automatically defined groups are other included data mining techniques.

An important question that must be settled and is in dispute is the exact and accurate treatment and diagnosis given to the person who requires the medical help .researchers ,in the recent past have began looking into data mining techniques to handgrip the ramification and

problems of treatment processes for them who provides healthcare.

In spite of the fact that health care professionals get help in the diagnosis of heart disease by data mining techniques, the use of data mining techniques to recognize a desirable treatment for heart disease patients has obtained less attention. Also, excellent outcomes has been shown in the diagnosis of heart diseases by employing hybrid data mining techniques, hence for picking out the suitable treatment for heart disease patients, applying hybrid data mining techniques needs advance investigations.

3. Data Mining on petro physical data, logging data, seismic data and geological data

Data mining is a process of taking out implicit, previously not known, but potentially important information and knowledge from a mass amount of incomplete, ambiguous, noisy and haphazard data in the practical application [3-5].

Fitting in to the data mining techniques, to retrieve the connections and predict reservoirs the petro physical data are employed; to appraise the fuzzy reservoirs and accredit the in effect reservoirs in difficult to analyse and understand geological conditions the logging data will be used; the space mining leads of the 3D seismic data; the charts and text mining leads of the geological data. In the exploration the oil and natural gas data mining takes in the ways of data analysis and the representing mathematical model to operate the exploration data, and acquire the possible knowledge.

In the systematic search and yielding oil and gas, the data primarily is drew up of petro physical data, seismic data, logging data, and geological data.

Beneath static and dynamic circumstances, the change rules of petro physical data should be mined as it has a bang-up importance to the oil and gas production and exploration in scope of gainfulness of regional and local data.

In spite of the fact that, the logging data are accurate to a lesser extent and straight than the petro physical data, they can manifest the formation and fluid information with few mistakes from various angles and have a more universal significance to the formation, recognition and

rating. Therefore, the most spectacular goal of mining logging data is to get the best favourable model anticipating the nature of the fluid and spot the production and exploration of oil and gas reservoir.

Because of the bulky and elaborated structure of seismic data, its mining methods and estimates are extremely dissimilar from petro physical data and logging data. Moreover, the seismic data mining is not a strictly mathematical problem and more distinctly, the normal mining algorithms cannot achieve unjust knowledge and in effect outcomes to profit the production and search of oil and gas.

Geological data is in patterns of symbols, graphs and words to convey structural feature, sedimentary facies, hydrocarbon generation ability, distribution of source rocks thickness and so on, however experiments in laboratory and watching in field. For the dissimilar kinds of geological data, it is necessary to select equating mining methods and ideas to attain mining targets and get legal effectual information.

4. Industry Applications of Data Mining

Data mining utilisation in the industry are directed at meeting two issues that organizations face: improved usage of data assets and customer intimacy.

These uses can be categorised into those who use macro mining, i.e. multi component server-based systems and those that use micro mining, i.e. single mining component, desktop systems. The occupying outcome of this coupling for the data mining community is that the data warehouses cannot be affirmed by the present data mining oblations holding up the readying of applications in yielding environments. The intensity of data is very large, the data types are various, and the data characteristics too contrastive for the living data mining algorithms. Moreover, the pure mining procedure is a much diminished component of the whole application life cycle. The speak will confront the problems associated to the coupling of macro mining with data warehouses, and suggest problems that must be settled for large-scale data mining applications to extend being positioned successfully.

5. Data Mining in Broadcasting Effect

With the evolution of television monitoring and the broadcasting, monitoring data in database have become more and more. Individuals desire that these data are both perceivable and utile to succeeding work. Hence, it is very important to talk about the particular applications of data mining techniques in the broadcasting monitoring. A data mining solution for this problem can be based on neural networks. The result can figure tantamount face of radio broadcasting consequence by data mining.

All the way through the past twenty years, lots of monitoring data have been collected in database. Applying data mining technology can transfer mass data into knowledge and make efficient use of monitoring data [6].

6. Malicious Code Detection and Security Applications in Data Mining

Data mining is the geographic expedition of collection of data and study of data and summing up the significant data [7]. Data mining is the soft computing technique through which we solve the security problems and malicious code detection. This application is used to find out the mistrustful persons who are not trustful and involved in act of terrorism. Spyware is a software which is mistrusted and setup on the computing device without the user cognition and collects the important information around the user. Spyware arises variety of attributes and these are very difficult to blemish and installed themselves on the computer system for a long run [8]. Spyware can be classified namely malware and adware. Malware is the computer program that performs undesirable actions. Examples of malware software are Trojan horse, spyware, virus, worms and key loggers. Adware is a computer program which is generally used for advertisement. They are basically the toolbars and they are usually safe. Spyware can be observed through the antivirus programs. To search the malicious code detection we use data mining methods. Data mining uses many algorithms to detect malicious code [9]. Some of them are RIPER algorithm, Naive Bayes algorithm and Multi-Naive Bayes algorithm. RIPER algorithm put on Portable Executable (PE) data format [9]. The Naive Bayes algorithms apply strings in the binary form. But this method can be easily removed by

encrypting a string of characters. The Multi-Naive Bayes algorithm works on sequence of bytes in a file.

7. TDML-Transaction Data Mining Language

Data mining scheme has the power to endure the specific plus intellectual data mining with the objective of effectual and flexible cognition [10, 11]. Data mining also support many languages. For the query there is data mining query language (DMQL) so for the transaction databases data mining suggested data mining language for the transaction databases known as the TDML. This language is basically introduced for the consecutive pattern mining and for the affiliation rule mining. The TDML language uses bit map processing method with the help of buffer which is needed to stores the result. In the TDML we manage the databases that are used for the transaction purposes and through this language it becomes easier to do the transactions. There is a predefined format for the TDML language. The syntax for the TDML is specified through the BNF grammar where ‘[]’ stand for 0 or the 1 condition and ‘{ }’ stand for 0 and more conditions and the words are in the Arial font size symbolize keywords.

```
<TDML> ::= <TDML Statement> { <TDML  
Statement> }
```

```
<TDML Statement> ::=
```

```
Mining<ARM/SPM/target/reduce>
```

```
from files=<f1,f2,...,fn> |
```

```
[buffers=<b1,b2,...,bn>] |
```

```
[systems=<s1,s2,...,sn>]
```

```
with support=<s1,s2,...,sn> |
```

```
conf=<c1,c2,...,cn> |
```

```
[parents=<p1,p2,...,pm>] |
```

```
[levels=<no>] |
```

[itemsets=<i1,i2,...,im>] |

[inputs=<i1,i2,...,im>] |

[minrow=<r1>, maxrow=<r2>, mincol=<c1>]

[save files=<f1,f2,...,fn> |

buffers=<b1,b2,...,bn>]

[using <ddm /gen/ mlevel/ mdim/ online/ Incr/

merge/ stream/ partition>]

Dynamic databases are sustained in the additive online ad for the merge mining [12] alternative assigned in "using" clause

using<incr/online/merge>

8. Conclusion

We have discussed many applications of data mining through these applications we can easily work on different databases, can perform pattern matching. It can also be used in the security and for the code detection purposes. In this research paper we have discussed various problems and their solutions. Data mining can be used for many purposes like medical purpose, banking industry, security etc. This paper shows the use of the various applications in the day-to-day life. We have concluded that data mining as the proficiency for the screening the information. One time the data is screened we can apply different data mining techniques.

9. References

1. World Health Organization. 2007 7-February 2011]; Available: <http://www.who.int/mediacentre/factsheets/fs310.pdf>.
2. European Public Health Alliance. 2010 7-February-2011]; Available from: <http://www.epha.org/a/2352>
3. J.W.Han and M.Kamber. Data Mining Concepts and Techniques, Second Edition, Beijing: China Machine Press, 2006.
4. U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth.From Data Mining to Knowledge Discovery: an Overview, in:

Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy(eds.), AAAI/MIT, Menlo Park, CA, pp. 1-34, 1996.

5. M.S. Chen, J.w. Han and Philip S. Yu. Data Mining: An Overview from a Database Perspective, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 866-883, December 1996.

6. Pang-Ning Tan, Michael Steinbach and Vipin Kumar, *Introduction to Data Mining*, Addison-Wesley Longman, Boston, 2005.

7. <http://www.anderson.ucla.edu/faculty/jason.frand/teaching/technologies/palace/datamining.htm>

8. "The silent epidemic of 2005: 84% of malware on computers worldwide is spy-ware". <http://www.pandasoftware.com/about/press/viewnews.aspx?noticia=5968>.

9.M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo, **Data Mining Methods for Detection of New Malicious Executables**, Proceedings of IEEE Symposium on Security and Privacy. Oakland, pp.38-49, 2001

10. Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaufman Publishers, San Francisco, 2001.

11. Willi Kloggen and Jan M. Zytkow, "Hand Book on Data Mining and Knowledge Discovery", Oxford University Press, 2002.

12. Walid G. Ared, Mohamed G. Elfeky, Ahmed K Elmagarmid, Incremental, Online, and Merge Mining of Partial Periodic Patterns in Time-Series Databases, IEEE TKDE, Vol. 16, No.3, 2004, pp.332-342

IJSER