www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 2 Feb. 2017, Page No. 20403-20409
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.43

Knowledge Discovery using Data Mining

Divya Uppal Sakhuja, Dr. V.K.Pathak
Lecture
APIC Raipur(C.G.)
divya.cvru@gmail.com
Hod(Computer Science)
Govt. P.G. College Dhamtari (C.G.)
vkpath21162@yahoo.co.in

Abstract: - The fast developing computer science and engineering techniques has made the information easy to capture process and store in databases. Information or data are considered as elementary variable facts. Knowledge is considered as a set of instructions, which describes how these facts can be interpreted and use[1]. Data describes the actual state of the world; however knowledge describes the structure of the world and consists of principal and laws. How to gather, store and retrieve data is considered in database. Soft computing (SC) is an evolving collection of methodologies, which aims to exploit tolerance for imprecision, uncertainty, and partial truth to achieve robustness, tractability, and low cost. SC provides an attractive opportunity to represent the ambiguity in human thinking with real life uncertainty. Soft computing has recently been playing an important role in advanced knowledge processing. An advanced learning method using a combination of perception and motion has been introduced. Emergent, self-organizing, reflective, and interactive (among human beings, environment, and artificial intelligence) knowledge processing is considered by using soft computing and by borrowing ideas from bio-information processing [2].

Soft computing (SC)was proposed for construction of new generation artificial intelligence (high machine intelligence quotient (HMIQ), human-like information processing) and for solving nonlinear and mathematically unmodeled systems (tractability) (TR) [3].

Data mining is a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes 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[4][5][6].

I. Introduction:

Due to the increasing use of very large databases and data warehouses, mining useful information and helpful knowledge from transactions is evolving into an important research area. Most of conventional data mining algorithms identify the relation among transactions with binary values. Transactions with quantitative values are, however, commonly seen in real world applications.

Through data mining, we can able to effectively extract data in the form of knowledge discovery which provides useful helping guide for information processing that can be utilized in varieties of applications. It is the most sought after field in recent scenario and its importance can not be ignored at all as effective data analysis outputs to extensive information utilization in almost all the fields and a proper data mining provides the appropriate and effective result. Different types of data mining techniques are augmented for powerful data mining applications that can ranges from Super market data to science, research, medical, media, web, entertainment and a lot of other fields which are implemented through data warehouses and Online Analytical Processing along with different data mining models. In this paper we have focused on data mining works with respect to current research approaches in variety of fields.

II. Steps of knowledge discovery using data mining:

The process of knowledge discovery is the process of information extraction from very large databases. Its importance is described along with several techniques and considerations for selection the most appropriate technique for extracting information from particular data set. We are using 5 steps for knowledge discovery:-

1. **Data Cleaning:-** This is the first step in which the noise, meaningless data and inconsistent data is removed. Only reliable and useful data is required.

DOI: 10.18535/ijecs/v6i2.43

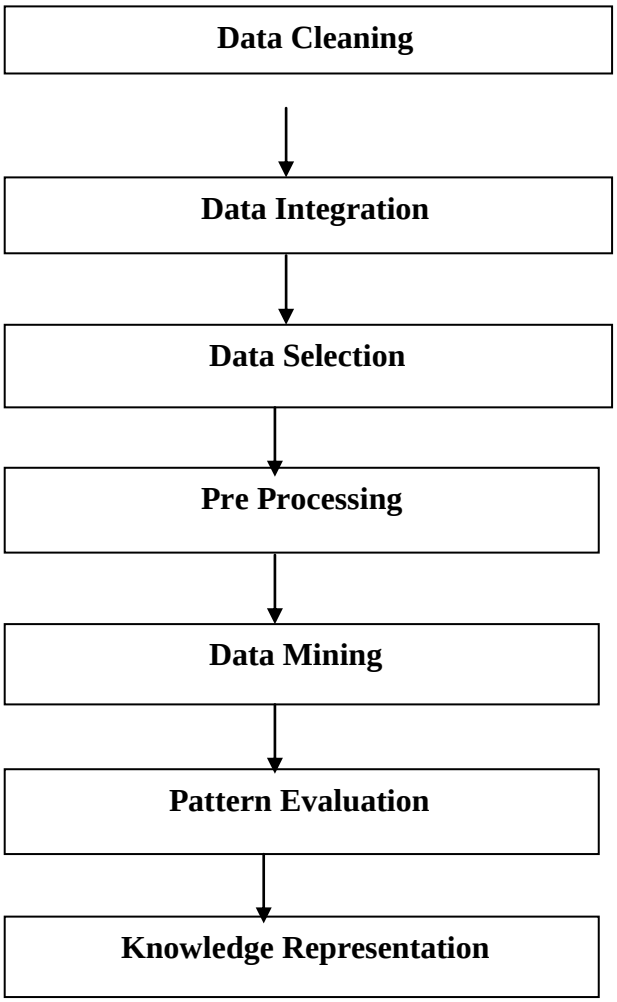
2. **Data Integration:-** Integration is an important step because in this step we have to combine or integrate data from different sources.
3. **Data Selection:-** It involves identification or extraction of relevant data for analysis. Data relevant to the analysis task are retrieved from the database.
4. **Pre processing:-** It involves preparing the data set by resolving problems like missing data, skewed data,

irrelevant fields, removal of outlying points etc[7][8]. This step consisting of two steps, like data cleaning and data integration. Data cleaning consists of some basic operations like normalization, noise removal etc. Data integration includes integrating multiple, heterogeneous data sets generated from different sources

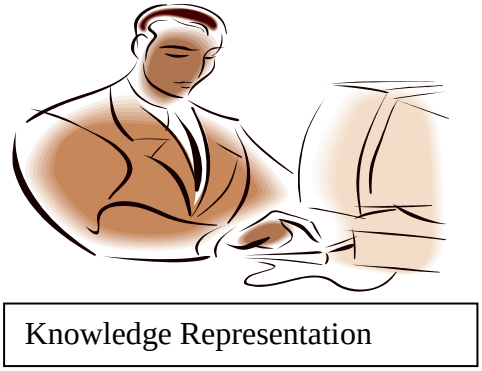
5. Data mining: It involves application of knowledge discovery algorithms to the cleaned , transformed data in order to extract meaningful patterns from the data. In this intelligent methods are applied in order to extract data patterns.

6. Pattern Evaluation: In this step evaluation of pattern has been done.

7. Knowledge Representation:- This step involves representation of discovered knowledge in proper format.



III. Steps of knowledge Discovery



IV. Methods of Data Mining:

I will use two steps Methodology:-

DOI: 10.18535/ijecs/v6i2.43

a. **Pre Processing**
b. **Genetic Algorithm**

a. **Pre-processing:-** Pre-processing is combinations of 3 methods:-
   i. Data Cleaning:- **DATA CLEANING** - a two step process including ***DETECTION*** and then ***CORRECTION*** of errors in a data set. Every dataset contains some errors, and every analyst experiences a rite of passage in wasting days drawing wrong conclusions because the errors have not been first rooted out. Up to half of the time needed for analysis is typically spent in "cleaning" the data. This time is also, typically, underestimated. Often, once a clean dataset is achieved, the analysis itself is quite straightforward.
   ii. Data Reduction:- Obtains reduced representation in volume but produces the same or similar analytical results. It reduces amount of data or we can say it is used to remove unwanted data.
   iii. Data Integration:- It combines data from multiple sources into a coherent data store as in data warehousing. These sources may include multiple databases, data cubes or flat files.

b. **Genetic Algorithm:-** After pre-processing I will use genetic algorithms as second step of my methodology. Genetic algorithms were formally introduced in the United States in the 1970s by John Holland at University of Michigan. The continuing price/performance improvements of computational systems has made them attractive for some types of optimization. In particular, genetic algorithms work very well on mixed (continuous *and* discrete), combinatorial problems. They are less susceptible to getting 'stuck' at local optima than gradient search methods. But they tend to be computationally expensive.

To use a genetic algorithm, you must represent a solution to your problem as a *genome* (or *chromosome*). The genetic algorithm then creates a population of solutions and applies genetic operators such as mutation and crossover to evolve the solutions in order to find the best one(s). The three most important aspects of using genetic algorithms are:

(1) Definition of the objective function.

(2) Definition and implementation of the genetic representation.

(3) Definition and implementation of the genetic operators.

V. Conclusion:

Our aim is to reduced time complexity with the help of genetic algoritm for data with quantitative values.
Secondary objective is achieving prediction of the data-item sets.
Third objective is to find the frequent pattern of data- items sets.
Fourth objective is to eliminate the unwanted data-item with the help of genetic algorithm.

The goal of data mining is to discover the important associations among items such that the presence of some items in a transaction will imply the presence of some other items and also some kind of predictions from data sets. For achieving this purpose I will collect data from different sources then I will use several genetic algorithms to find optimal results. They decomposed the mining process into two phases. In the first phase, we have to represent a solution of problem as a genome or chromosome. In the second phase, the genetic algorithm then creates a population of solutions and applies genetic operators such as mutation and crossover to evolve the solutions in order to find the best one.

The proposed model is more objective and reasonable in determining the object oriented queries with genetic algorithms and focuses in the most important linguistic term for reduced time complexity. With the help of this we can also perform the prediction of any given data sets. This is used to find optimal result.

VI. References:

(i) Books

1. K. Das, K. Bhaduri. Parallel and Distributed Data Mining for Astronomy Applications. A chapter in *Data Mining and Machine Learning for Astronomical Applications*, K. Ali, A. Srivastava, J. Scargle and M. Way (editor), Chapman & Hall/CRC Press. (in communication). 2010.
2. K. Bhaduri, K. Das, K. Sivakumar, H. Kargupta, R. Wolff, R. Chen. Algorithms for Distributed Data Stream Mining. A chapter in *Data Streams: Models and Algorithms*, C. Aggarwal (editor), Springer. pp. 309-332. 2006.
3. **Data Mining: Concepts and Techniques** *Jiawei Han and Micheline*

Kamber, **Publisher:** Morgan Kaufmann (6 Sep 2000).

4. Automating the Design of Data Mining Algorithms: An Evolutionary Computation Approach (Natural Computing Series) authors: Gisele L. Pappa, Alex Freitas Publisher: Springer; 2010 edition (November 5, 2009).
5. Data Mining Using Grammar Based Genetic Programming and (GENETIC PROGRAMMING Volume 3) Man Leung Wong , Kwong Sak Leung Publisher: Springer; 2000 edition (January 1, 2000).

(ii)Reference Papers

[2] Y. Dote, "Soft computing in computational intelligence and information/ intelligent system," in *Proc. IEEE Int. Workshop on Soft Computing in Industry*, Muroran, Japan, 1999, pp. 66–71.

[3] L. A. Zadeh, "Fuzzy logic, neural networks and soft computing," in *Proc. IIEEE Int. Workshop Neuro Fuzzy Control*, Muroran, Japan,1993, p. 1.

[4] FALOUTSOS, C. and LIN, K. 1995. Fastmap: A fast algorithm for indexing, data mining and visualization of traditional and multimedia. In Proceedings of the ACM SIGMOD Conference, 163-174, San Jose, CA.

[5] Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[6] HAN, J. and KAMBER, M. 2001. Data Mining. Morgan Kaufmann Publishers.

[7] BAEZA-YATES, R. 1992. Introduction to data structures and algorithms related to information retrieval. In Frakes, W.B. and Baeza-Yates, R. (Eds.) Information Retrieval, Data Structures and Algorithms, 13-27, Prentice-Hall.

[8] G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

[9] W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus, Knowledge Discovery in Databases: An Overview. In G. Piatetsky-Shapiro et al. (eds.), Knowledge Discovery in Databases. AAAI/MIT Press, 1991.

[10] DUDA, R. and HART, P. 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, NY.

[11] A.K. Jain, M.N. Murty, P.J.Flynn, "Data Clustering: A Review", ACM Computing Surveys ,vol. 31 , no.3, September 1999.

[12] BAKER, L.D. and MCCALLUM, A. K. 1998. Distributional clustering of words for text classification. In Proceedings of the 21st ACM SIGIR Conference, Melbourne, Australia.

[13] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proc. 1993 ACMSIGMOD Int. Conf. Management Data, Washington, DC, May 1993, pp. 207–216.

[14] D. Shalvi and N. De Claris, "Unsupervised neural network approach to medical data mining techniques," Proc. IEEE Int. Joint Conf. Neural Networks, pp. 171– 176, May 1998.

[15] Towell G., Shawlik J.W. (1994): Knowledge-based artificial neural networks. Artificial Intelligence, 70 (1- 2): 119-165 (October).

[16] Carpenter, G.A., Grossberg, S., and Rosen, D.B. 1991. Fuzzy art: Fast stable learning and Categorization of analog patterns by an adaptive resonance system. Neural Networks, 4, 759-771.

[17] L. A. Zadeh, "Fuzzy logic, neural networks, and soft computing," Commun. ACM, vol. 37, pp. 77–84, 1994.

[18] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison- Wesley,1989.

[19] E. Noda, A. A. Freitas, and H. S. Lopes, "Discovering interesting prediction rules with a genetic algorithm," Proc. IEEE Congr. Evolutionary Comput. CEC '99, pp. 1322–1329, July 1999.

[20] I.W. Flockhart and N. J. Radcliffe, "A genetic algorithm-based approach to data mining," in Proc. 2nd Int. Conf. Knowledge Discovery DataMining (KDD- 96). Portland, OR, Aug. 2–4, 1996, p. 299.

[21] Sushmita Mitra, Sankar K. Pal, Pabitra Mitra, "Data Mining in Soft Computing Framework: A Survey", IEEE Transactions On Neural Networks, Vol. 13, No. 1, January 2002.

[22] Adewole Philip, Akinwale Adio Taofiki, Otunbanowo Kehinde, "A Genetic Algorithm for Solving Travelling Salesman Problem", International Journal of Advanced Computer Sciences and Applications, Volume 2, No. 1, January 2011.

[23] Agustin-Blas, L.E.; Salcedo-Sanz, S.; Ortiz-Garcia, E.; Perez-Bellido, A.; Portilla-Figueras, A.; Dept. of Signal Theor. & Commun., Univ. de Alcala, Madrid, "Assignment of Students to Preferred Laboratory Groups Using a Hybrid Grouping Genetic Algorithm", Hybrid Intelligent Systems, 2008. HIS '08. Eighth International Conference.

[24] Christodoulopoulos, C.E.; Papanikolaou, K.A., "A Group Formation Tool in an E-Learning Context", Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference.

[25] Hak Koon Yeoh, Mohamad Iskandr Mohamad Nor, "An algorithm to form balanced and diverse groups of students", 2009 Wiley Periodicals, Inc. Comput Application Engineering Education 19: 582–590, 2011.

[26] Kusum Deep, Hadush Mebrahtu, "Combined Mutation Operators of Genetic Algorithm for the Travelling Salesman problem", International Journal of Combinatorial Optimization Problems and Informatics, Vol. 2, No.3, Sep. - Dec. 2011, pp. 1-23.

[27] Li Liu, Murat Kantarcioglu, , Bhavani Thuraisingham, The applicability of the perturbation based privacy preserving data mining for real-world data, Data & Knowledge Engineering, Volume 65, Issue 1, April 2008, Pages 5-21.

[28] Seung-Woo Kim, Sanghyun Park, Jung-Im Won, Sang-Wook Kim, Privacy preserving data mining of sequential patterns for network traffic data**,** Information Sciences, Volume 178, Issue 3, 1 February 2008, Pages 694-713.

[29] Jose Zubcoff, Juan Trujillo, A UML 2.0 profile to design Association Rule mining models in the multidimensional conceptual modeling of data warehouses, Data & Knowledge Engineering, Volume 63, Issue 1, October 2007, Pages 44-62.

[30] George Gigli, Éloi Bossé, George A. Lampropoulos, An optimized architecture for classification combining data fusion and data mining, Information Fusion, Volume 8, Issue 4, October 2007, Pages 366-378.

[31] Michael Böhlen, Linas Bukauskas, Poul Svante Eriksen, Steffen Lilholt Lauritzen, Artūras Mažeika, Peter Musaeus, Peer Mylov, 3D visual Data Mining—goals and experiences, Computational Statistics & Data& Analysis, Volume 43, Issue 4, 28 August 2003, Pages 445-469.

[32] Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., "A Data Mining Engine based on Internet, Emerging Technologies and Factory Automation," Proceedings ETFA '03, IEEEz Conference,16-19Sept.2003.WebSite:www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8955

[33] Bianca V. D.,Philippe Boula de Mareüil and Martine Adda-Decker, "Identification of foreign-accented French using data mining techniques, Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI)".Website www.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf

[34] Bianca V. D.,Philippe Boula de Mareüil and Martine Adda-Decker, "Identification of foreign-accented French using data mining techniques, Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI)".Website www.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf

[35] Halteren, H. V., Oostdijk N., "Linguistic profiling of texts for the purpose of language verification, The ILK research group, Tilburg centre for Creative Computing and the Department of Communication and Information Sciences of the Faculty of Humanities, TilburgUniversity,TheNetherlands."Website: www.ilk.uvt.nl/~antalb/textmining/LingProfColingDef.pdf

[36] Antonie, M. L., Zaiane, O. R.,Coman, A., "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, August 26, 2001.

[37]. Pei, M., Goodman, E.D., and Punch, W.F. "Pattern Discovery from Data Using Genetic Algorithms", *Proceeding of 1st Pacific-Asia Conference Knowledge Discovery & Data Mining (PAKDD-97)*. Feb. (1997).

[38]. Pei, M., Punch, W.F., and Goodman, E.D. "Feature Extraction Using Genetic Algorithms", *Proceeding of International Symposium on Intelligent Data Engineering and Learning'98 (IDEAL'98)*, Hong Kong, Oct. (1998).

[39]. Punch, W.F., Pei, M., Chia-Shun, L., Goodman, E.D., Hovland, P., and Enbody R. "Further research on Feature Selection and Classification Using Genetic Algorithms", In *5th International Conference on Genetic Algorithm* , Champaign IL, pp 557-564, (1993).

[40]. Siedlecki, W., Sklansky J., A note on genetic algorithms for large-scale feature selection, *Pattern Recognition Letters*, Vol. 10, Page 335-347, (1989).

[41]. Skalak D. B. (1994). Using a Genetic Algorithm to Learn Prototypes for Case Retrieval an Classification. *Proceeding of the AAAI-93 Case-Based Reasoning Workshop*, pp. 64-69. Washigton, D.C., American Association for Artificial Intelligence, Menlo Park, CA, 1994.

[42]. Vafaie H and De Jong K. "Robust feature Selection algorithms". *Proceeding 1993 IEEE Int. Conf on Tools with AI*, 356-363. Boston, Mass., USA. Nov. (1993).

[43] S. Brin, R. Motwani, J. Ulman, and S. Tsur, Dynamic itemset counting implication rules for market basket data. In Proc. Of ACM SIGMOD Intl. Conf. on Management of Data, May 1997.

[44] D. Cheung, T. Vincent, and W. Benjamin, "Maintenance of discovered knowledge: A case in multilevel association rules," in proc. of international conference on "Knowledge discovery and data mining (KDD)," August 1996.

[45] H. Toivonen. Sampling large databases for association rules. In Proc. Of Intl. Conf. on very Large Databases ((VLDB), 1996.

[46] J. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," artificiall intelligence, 40: 11-61, 1989.

[47] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-211.

[48] Cai, W. and Li L., "Anomaly Detection using TCP Header Information, STAT753 Class Project Paper, May 2004.". Web Site:http://www.scs.gmu.edu/~wcai/stat753/stat753report.pdf.

[49] Nandi, T., Rao, C. B. and Ramchandran, S., "Comparative genomics using data mining tools, Journal of Bio-Science, Indian Academy of Sciences, Vol. 27,No. 1, Suppl. 1, page No. 15-25, February 2002".

[50] Khreisat, L., "Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study". proceedings of The 2006 International Conference on Data Mining, DMIN'06, pp 78-82, Las Vegas, Nevada, USA, June 26-29, 2006

[51] Onkamo, P. and Toivonen, H., "A survey of data mining methods for linkage disequilibrium mapping", Henry Stewart Publications 1473 - 9542. Human Genomics. VOL 2, NO 5, Page No. 336- 340, MARCH 2006.

[52] Smith, L., Lipscomb, B., and Simkins, A., "Data Mining in Sports: Predicting Cy Young Award Winners". Journal of Computer Science, Vol. 22, Page No. 115-121,April 2007.

[53] Deng, B., Liu, X., "Data Mining in Quality Improvement". USA.ISBN1-59047-061-3.WebSite http://www2.sas.com/proceedings/sugi27/Proceed27.pdf

[54] Cohen, J. J., Olivia, C., Rud, P., "Data Mining of Market Knowledge in The Pharmaceutical Industry". Proceeding of 13th Annual Conference of North-East SAS Users Group Inc., NESUG2000, Philadelphia Pennsylvania, September 24-26 2000.

[55] Elovici, Y., Kandel, A., Last, M., Shapira, B., Zaafrany, O., "Using Data Mining Techniques for Detecting Terror-Related Activities on the Web".WebSite: www.ise.bgu.ac.il/faculty/mlast/papers/JIW_Paper.pdf

[56] Solieman, O. K., "Data Mining in Sports: A Research Overview, A Technical Report, MIS Masters Project, August 2006". Web Site: http://ai.arizona.edu/hchen/chencourse/Osama-DM_in_Sports.pdf

[57] Maciag, T., Hepting, D. H., Slezak, D., Hilderman, R. J., "Mining Associations for Interface Design". Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 4481, pp.109-117, June 26, 2007

[58]Silva1, S., Almeida, J.: Dynamic Maximum Tree Depth A Simple Technique for avoiding Bloat in Tree-Based GP.Biomathematics Group, Instituto de Tecnologia Qu′ımica e Biol′ogica Universidad Nova de Lisboa, PO Box 127, 2780-156 Oeiras, Portugal (2002)

[59]Stoffel, K., Spector, L.: High-Performance, Parallel, Stack-Based Genetic Programming. In: Proceeding of the First Annual Conference, pp. 224–229 (1996)

[60]Luke, S., Panait, L.: Fighting Bloat with Nonparametric Parsimony Pressure. In: Proceeding of the First Annual Conference (2000)

[61]Koza, J.R.: Genetic Programming. Encyclopedia of Computer Science and Technology (8.18), 2–4 (1997)

[62]Aguilar-Ruiz JS, Riquelme JC, Toro M (2003) Evolutionary learning of hierarchical decision rules. IEEE Trans Syst Man Cybern B 33(2): 324–331 CrossRef

[63]Araujo DLA, Lopes HS, Freitas AA (1999) A parallel genetic algorithm for rule discovery in large databases. In: Proceedings of IEEE systems, man and cybernetics conference, vol 3, Tokyo, pp 940–945

[64]Bacardit J (2004) Pittsburgh genetics-based machine learning in the data mining era: representations, generalization, and run-time. PhD thesis, Ramon Llull University, Barcelona

[65]Bacardit J, Krasnogor N (2006a) Empirical evaluation of ensemble techniques for a Pittsburgh learning classifier system. In: 9th International workshop on learning classifier systems (IWLCS 2006), Lecture Notes in Artificial Intelligence. Springer

[66]Bacardit J, Krasnogor N (2006b) Biohel: Bioinformatics-oriented hierarchical evolutionary learning. Nottingham eprints, University of Nottingham

[67]Bacardit J, Stout M, Hirst JD, Sastry K, LloráX, Krasnogor N (2007) Automated alphabet reduction method with evolutionary algorithms for protein structure prediction. In: Proceedings of the 9th annual conference on genetic and evolutionary computation. ACM Press, New York, pp 346–353

[68]Bacardit J, Burke EK, Krasnogor N (2009) Improving the scalability Of rule-based evolutionary learning. Memet Comput 1(1): 55–67 CrossRef

[69]Dehuri S, Mall R (2006) Predictive and comprehensible rule discovery using a multi-objective genetic algorithm. Knowl Based Syst 19: 413–421 CrossRef

[70]De Jong KA, Spears WM (1991) Learning concept classification rules using genetic algorithms. In: Proceedings of the international joint conference on artificial intelligence. Morgan Kaufmann, pp 651–656.

(iii) E-References:

1. *www.cs.uvm.edu/~xwu/KDD/Links.shtml - United States*
2. www.**web-datamining**.net/structure/

3. www.cs.umn.edu/~kumar/dmbook
4. www.ibm.com/developerworks/library/wa-wbdm/
5. **www .datamining**.typepad.com/
6. http://www.cs.helsinki.fi/u/htoivone/pubs
7. www.en.wikipedia.org/wiki/**Web_mining**
8. http://www.tulane.edu/~panda2/Analysis2/datclean/dataclean.htm
9. http://iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf
10. http://www.isical.ac.in/~shubhra/
11. http://www.lon-capa.org/papers/v90-gapaper.pdf