Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Bio-Inspired Approach to Data Mining: A Review

Poonam Kataria*
Department of CSE,SUSCET
India

Rahul Sharma
Department of IT,GNDEC
India

Abstract— Data mining has become a powerful tool to extract hidden patterns of data and is gaining importance as it helps in decision making in all spheres of life. Its various techniques allows users to get refined result of their query. Clustering is the technique which allows search engine to group similar kind of data. Diversification techniques ranks the data according to the priority. There are number of bio-inspired approaches which can be implemented to solve hard and complex problems in data mining. In this paper, various swarm intelligence approaches will be discussed.

Keywords— Data Mining, Clustering , Diversification, Bio-Inspired, PSO , GA

I. INTRODUCTION

A. Data Mining:

Data Mining is a generic term that is used to find hidden patterns of data(tabular, spatial, temporal, spatio-temporal etc.) . These hidden patterns is the outcome of types of data and various interestingness criteria. Data Mining[1] is the process of selection, exploration and modelling of large quantities of data to discover regularities or relation that are at first unknown with the aim of obtaining clear and useful results for the owner of database. The basic work of data mining can be categorised using the technique of clustering and classification.

B. Clustering:

Clustering is the process in which data of similar type is grouped together and then classified. Classification and clustering are different but they are dependent as classification is based on which are is chosen as cluster region and kind of dataset which will be applied on selected region. Various Clustering techniques[2] based on their taxonomic representation are available which can be categorised as follows:-

1. **Agglomerative *vs* divisive**:- This aspect relates to algorithmic structure and operation.
2. **Monothetic *vs* polythetic**:- This aspect relates to sequential and subsequent use of features in clustering process
3. **Hard *vs* Fuzzy:** Hard clustering algorithm allocates each pattern to a single cluster during its operation whereas fuzzy clustering assigns a degree of membership in several clusters to each input pattern
4. **Deterministic *vs* Stochastic:-** The most relevant partitioned approach designed to optimize square error function.

C. Diversification:

Web has become a richest source of information and is largest container of data. Whenever a users fires a query a long list of result is generated and to refine the result, search engine uses various diversification techniques for ranking the data. To some extent, today's search engines, such as Google and Yahoo!, apply some diversification technique to their top-ranking results.[3]

1) Max-sum Diversification. The first objective combines the sums of the relevance and diversity measure as a weighted sum.

2) Max-min Diversification. The second objective targets at maximising the sum of the minimum relevance and minimum dissimilarity within the set.

3) Average Dissimilarity Diversification. Their third objective adds the original relevance for a result with the average dissimilarity regarding all other results in the set. The sum over the whole set is to be maximised.

4) Max-sum of max-score Diversification. Similarly to max-sum diversification,maximises the sum of dissimilarity of the result set, but it only produces sets that have the maximal relevance sum. Therefore, it does not find sets with higher diversity scores but slightly lower relevance sum.

5) Max-product Diversification. Based on the already chosen results, Zhaiet al. select the next result by maximising the parameterised product of the relevance of the next result and its dissimilarity to the chosen results.

D. Bio-Inspired Approaches

Bio-inspired is the term that is inspired from nature and works very well to solve a number of problems in computer science. It becomes a motivating force for developing artificial intelligent tool which plays a vital role.

1.Genetic Algorithm:- Genetic Algorithm is class of bio- inspired algorithm that uses technique inspired from biology inheritance, mutation, crossover, selection. This algorithm is implemented as simulation of computer in which population

of abstract representations (called chromosome and genomes) of candidate solutions to an optimization problem. The evolution starts from randomly generated individuals and in each generation fitness of every individual is evaluated and modified to form a new population. The new population so generated is used in next iteration of algorithm. The algorithm is terminated when a satisfactory fitness level is achieved

2 Particle Swarm Optimization:-Particle Swarm Optimization is a swarm intelligence method which is inspired from group of flocking birds. Each single solution is called particle in search space. All of the particles have fitness value which are evaluated by fitness function to be optimized. Each particle has velocity which directs the flight of particles. The two values that are calculated are pbest and gbest, where pbest is best solution that is achieved so far and gbest is best value obtained by any particle. Particle fly around in a multidimensional search space. During the flight, each particle adjust its position according to its own experience and experience of neighbouring particle, making use of best position encountered by itself and its neighbour.

3 Ant Colony Optimization:- Ant Colony Optimization is another swarm intelligence method which is inspired by behaviours of ant which tend to live in colony rather than individual. The main idea behind this algorithm is finding the shortest path between the food and their nest.

II. RELATED WORK

One of the work in which bio-inspired algorithm has been applied is done by Velu[4]. In this work authors have proposed artificial intelligent tools of data mining for diabetes detection. The study of classification of diabetic patients was main focus of research work. Diabetic patients were classified by data mining techniques for medical data obtained from Pima Indian Diabetes(PID) data set. This paper presented analyses made while comparing EM Algorithm, H-means Clustering and GA for clustering analyses in a very simple way using WEKA tools. The data records were filtered during pre-processing to make it compact and free from redundancy. The EM algorithm was used for sampling which consist of two steps i.e. determination of expectation and maximizing it. After that h-means+ clustering was done by performing iterations in order to reduce error function. Algorithm iterations removed empty clusters if found during execution. At last Genetic Algorithm was used which searched best solution within a collection of large number of solutions of a problem being solved. The simulation results were noticeable that h-means+ algorithms performed little better compared to EM. The clustering algorithm was run for local data sets and validated the results by comparing to PID. Another data mining tool of Genetic Algorithm was used to optimize the chromosome was performed.

III. LITERATURE REVIEW

Parpinelli[5] proposed an algorithm for data mining called Ant Miner which extracts classification rules from data. ACO algorithms involve simple agents (ants) that cooperate with one another to achieve an emergent, unified behavior for the system as a whole, producing a robust system capable of finding high-quality solutions for problems with a large search space. Author fouind that Ant-Miner seemed particularly advantageous when it was important to minimize the number of discovered rules and rule terms (conditions), in order to improve comprehensibility of the discovered knowledge. Also he added that it can be argued that this point is important in many (probably most) data mining applications, where discovered knowledge will be shown to a human user as a support for intelligent decision making.

Binitha[6] surveyed various bio-inspired algorithms and described that EA and SI Algorithms perform with heuristic population-based search procedures that incorporate random variation and selection. According to author it has been witnessed that the applications and growth of natural computing in the last years is very drastic and has been applied to numerous optimization problems in computer networks, control systems, bioinformatics, data mining, game theory, music, biometrics, power systems, image processing, industry and engineering, parallel and distributed computing, robotics, economics and finance, forecasting problems, applications involving the security of information systems etc. Biologically inspired computing still has much room to grow since this research community is quite young. There still remain significantly challenging tasks for the research community to address for the realization of many existing and most of the emerging areas in technology. Author also added there are great opportunities in exploring a new approach/algorithm. For this it requires collaboration of researchers from different communities like computer science, artificial intelligence, biology, ecology, social science etc. in order to have a broader and deeper view and analysis of each micro level steps/interactions there by having much more significant and outstanding results. Nevertheless, nature-inspired algorithms are among the most powerful algorithms for optimization which is going to have a wide impact on future generation computing.

Keshavamurthy[7] proposed an approach which improved the evolutionary technique such as genetic algorithm by improving the fitness function parameters .Author presented the work that improved the rule based genetic algorithm classifier by improving upon the fitness function parameter modification. Also, it compared the results with the probabilistic approach such as Naïve Bayes which always gives better results and very efficient in case there is no attribute dependency in the problem, which is not true in most of the real world problem .

Saraswathi**[8]** proposed a novel method for detecting the onset of Alzheimer's disease (AD) from Magnetic Resonance Imaging (MRI) scans. It used a combination of three different machine learning algorithms in order to get improved results and is based on a three-class classification problem. The three classes for classification considered in this study are normal, very mild AD and mild and moderate AD subjects. The machine learning algorithms used are: the Extreme Learning Machine (ELM) for classification, with its performance optimized by a Particle Swarm Optimization (PSO) and

a Genetic algorithm (GA) used for feature selection. A Voxel-Based Morphometry (VBM) approach is used for feature extraction from the MRI images and GA is used to reduce the high dimensional features needed for classification. The GA-ELM-PSO classifier yields an average training accuracy of 94.57 % and a testing accuracy of 87.23 %, averaged across the three classes, over ten random trials. The results clearly indicate that the proposed approach can differentiate between very mild AD and normal cases more accurately, indicating its usefulness in detecting the onset of AD.

Relan**[9]** proposed a model in which they for the discovery of biomarkers in the retinal vasculature it is essential to classify vessels into arteries and veins. We automatically classify retinal vessels as arteries or veins based on colour features using Gaussian Mixture Model, an Expectation-Maximization (GMM-EM) unsupervised classifier, and a quadrant-pair wise approach. Classification is performed on illumination-corrected images. 406 vessels from 35 images were processed resulting in 92% correct classification (when unlabelled vessels are not taken into account) as compared to 87.6%, 90.08%, and 88.28%. The classifier results were compared against two trained human graders to establish performance parameters to validate the success of classification method. The proposed system results in specificity of (0.8978, 0.9591) and precision (positive predicted value) of (0.9045, 0.9408) as compared to specificity of (0.8920, 0.7918) and precision of (0.8802, 0.8118) for (arteries, veins) respectively. The classification accuracy was found to be 0.8719 and 0.8547 for veins and arteries, respectively.

Ming**[10]** proposed that Event-Related Potential (ERP) has being the most popular method in evaluating brain waves of schizophrenia patients. ERP is one of the electroencephalography (EEG), which is measured the change of brain waves after giving patients certain stimulations instead of resting state. However, with traditional statistical analysis method, both P50 and MMN showed significant difference between controls and patients but not in Gamma band. Gamma band is a 30-50 Hz auditory stimulation which had been suggested may be abnormal in schizophrenia patients. Their data are recruited from 5 schizophrenia patients and 5 controls in National Taiwan University Hospital have been tested with this platform. The results showed that detection rate is 88.24% and we also analyzed the importance of features, including Standard Deviation (SD) and Total Variation (TotalVar) in different stage of wavelet transform. Therefore, this proposed methodology could serve as a valuable clinical decision support for physiologists in evaluating schizophrenia.

Basheer [11] proposed a genetic algorithm-based approach for mining classification rules from large database is presented. For emphasizing on accuracy, coverage and comprehensibility of the rules and simplifying the implementation of a genetic algorithm. The design of encoding, genetic operators and fitness function of genetic algorithm for this task are discussed. Experimental results show that genetic algorithm proposed in this paper is suitable for classification rule mining and those rules discovered by the algorithm have higher classification performance to unknown data.

Vivekanandan [12] proposed a model in which by applying the incremental genetic (IGA) algorithm in a batch mode can mine accurate rules reflecting the current concept. But applying genetic algorithm without monitoring for a change in an incremental manner repeatedly on arriving data will result in an unnecessary increase in the learning cost. There is also another problem, Due to change in the data distribution some of the rules which are generated may be lost when we apply genetic algorithm in an incremental fashion. In this paper a new incremental genetic algorithm is proposed to rectify the above problems. The New IGA applies the Genetic algorithm iteration step only when required, so that learning cost may be reduced. The new method also keeps track of the rules which are generated earlier and which would have been lost due to change in data distribution. In the proposed method each record of the incoming dataset is monitored. If they are correctly classified they are dropped and misclassified records are added to a window. When the window is full, the genetic algorithm is applied to the records in the window and new rules are generated based only on the misclassified examples and on the examples of new classes. The invalid rules are replaced with the newly generated valid rules. The new method ensures that the next iteration of genetic algorithm is called only when there is a concept drift or when there is a change in the data distribution and sufficient number of records is available. This will reduce the learning cost particularly when there is no concept drift or when there is a slow drift and also ensures that no rule is lost due to change in data distribution.

IV. CONCLUSIONS

There are various artificial intelligent tools that helps in data extraction as well in other fields of computing. There is great need to collaborate these bio-inspired approaches with the power of computing so that we tackle many more problems and handle various problems related to all spheres of life very efficiently. So, to use these approaches in data mining becomes a challenging task to do.

REFERENCES

[1] Giudici, P., Applied Data-Mining: Statistical Methods for Business and Industry. West Sussex, England: John Wiley and Sons (2003).

[2] A.K. JAIN,M.N. MURTY AND P.J. FLYNN, "*Data Clustering*", ACM Computing Surveys, Vol.31, No. 3, September 1999

[3] Velu, C.M., and Kashwan, K.R., "Visual Data Mining Techniques for Classification of Diabetic Patients", 3rd International Advance Computing Conference (IACC), pp. 1070-1075, 2013 IEEE.

[4] Enrico Minack, Gianluca Demartini, and Wolfgang Nejdl, "Current Approaches to Search Result Diversitication",L3S Research Center, Leibniz Universität Hannover, 30167 Hannover, Germany

[5] Rafael S. Parpinelli, Heitor S. Lopes, Alex A. Freitas ," Data Mining with an Ant Colony Optimization Algorithm", IEEE Transactions on Evolutionary Computation ,Aug 2002

[6] Binitha S, S Siva Sathya, "A Survey of Bio-Inspired Optimization Algorithms", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May 2012

[7] Keshavamurthy B. N, Asad Mohammed Khan & Durga Toshniwal," Improved Genetic Algorithm Based Classification" published in International Journal of Computer Science and Informatics (IJCSI) ISSN (PRINT): 2231 –5292, Volume-1, Issue-3

[8] Saraswathi, S.Mahanand, B.S. ; Kloczkowski, A. ; Suresh, S. ,"Detection of onset of Alzheimer's disease from MRI images using a GA-ELM-PSO classifier" IEEE Fourth International Workshop on Computational Intelligence in Medical Imaging (CIMI), 2013

[9] D. Relan, T. MacGillivray, L. Ballerini, E. Trucco, "Retinal vessel classification: sorting arteries and veins", in 35th Annual International Conference of the IEEE EMBS Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 2013, pp. 7396-7399.

[10] Ming-Hsien ,Hiesh Lam, Y.-Y.A.,Chia-Ping Shen , Wei Chen , Feng-Shen Lin ,Hsiao-Ya Sung ,Jeng-Wei Lin ,Ming-Jang Chiu , Feipei Lai, "Classification of schizophrenia using Genetic Algorithm-Support Vector Machine (GA-SVM)" Engineering in Medicine and Biology Society (EMBC)published in 35th Annual International Conference of the IEEE(2013)

[11] Basheer M. Al-Maqaleh , Hamid Shahbazkia, "A Genetic Algorithm for Discovering Classification Rules in Data Mining" International Journal of Computer Applications (0975 – 8887)
Volume 41– No.18, March 2012

[12] Vivekanandan P , Nedunchezhian R, "a new incremental genetic algorithm based Classification model to mine data with concept Drift", Journal of Theoretical and Applied Information Technology© 2005 - 2010

