INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 2 (2016) PAGES 533-537
RECEIVED : 02/03/2016. PUBLISHED : 10/03/2016
March 10, 2016

Detecting Anamolies by Particle Swarm Optimization technique for Cancer Dataset

[1]Ms. V. Prasanna, [2]Mr. N. Suresh Kumar, and [3]Dr. M. Thangamani
[1,2]Research Scholar, Anna University, Chennai, apr0883@gmail.com, nsuresh2@gmail.com
[3]Assistant Professor, Kongu Engineering College, Perundurai-638052,
manithangamani2@gmail.com

ABSTRACT

Data mining mechanism has widely been applied in business and manufacturing companies across many industrial sectors. Data mining is defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data". Data mining encompasses a number of different technical approaches, such as classification, clustering, data summarization, finding dependency networks, analysing changes and detecting anomalies. The proposed system detects the anamolies in cancer dataset using particle swarm optimization and results are taken based on number of data vs. Accuracy.

1. INTRODUCTION

One technique used in data mining is Classification where the desired output is a set of Rules or Statements that characterize the data. Classification is the tasks of generalizing known structure apply to new data. The classification rule is used to predict the class label which is used for making decision in many applications. The classification rule is constructed based on Genetic Algorithm (GA) and A Particle Swarm Optimisation (PSO). The PSO shows more accuracy when compared to GA, so it is considered as suitable candidate for classification task. In order to perform efficient classification, the anomalies have to be removed. Outlier detection is mainly used to remove anomalous observation from data.

The major objective of proposed work is to identify outliers during classification using PSO algorithm.Classification is a supervised learning and is one of the most studied data mining technique. The main goal is to predict the class Ci=f(X1 …Xn),where X1...Xn are input attributes. There is one distinguished attribute called as dependent attribute. The input to the classification algorithm is a data set of training records with several attribute. The process of generating rules from the given example set is called rule induction.

Problem Statement: In order to perform efficient classification, we need to remove anomalies. Outlier detection is mainly used to remove anomalous observation from data. The major objective of proposed work is to identify outliers during classification using PSO algorithm.

2. EXISTING WORK

The classification rule can be constructed based on GA and PSO.GA invented to mimic some of the processes observed in natural evolution. Many people, biologist included, are astonished that life at the level of complexity that we observe could have evolved in the relatively short time suggested by the fossil record. The idea with GA is to use this power of evolution to solve optimization problems. The father of the original GA was John Holland who invented it in the early 1970's.

Particle swarm optimization was introduced by Kennedy and Eberhart (1995)[10,11]. It has roots in the simulation of social behaviors using tools and ideas taken from computer graphics and social psychology research. The rules that govern the movement of the particles in a problem's search space can also be seen as a model of human social behavior in which individuals adjust their beliefs and attitudes that conform to those of their peers (Kennedy & Eberhart 1995).

Ammar W Mohemmed, Mengjie Zhang[1] has proposed a novel PSO to automatically optimize the distance measures for outlier detection. Kennedy J, Eberhart R[10] used particle swarm algorithm as an optimization technique on discrete binary value, it works by adjusting trajectories through manipulation of each coordinate of a particle. Ishibuihi H, Murala T and Turksen B [6] use three methods based on genetic algorithms for finding a set of non-dominated solutions. BakerJ E[2] adaptive selection method to inhibit premature convergence that occurs in genetic algorithms.

3. PSO FOR OUTLIER DETECTION

The outlier detection problem is converted into an optimisation problem. PSO based approach to outlier detection is then applied which expands the scope of PSO and enables new insights into outlier detection. PSO is used to automatically optimise the key distance measures instead of manually setting the distance parameters via trial and error, which is inefficient and often ineffective.

The PSO approach is examined and compared with a commonly used detection method, the results show that the new PSO method is more efficient and significantly outperforms than other methods. In this approach, PSO is used to find the point that has the minimum k/r ratio. The value of r that results in the minimum k/r is used to compute the ratio for the other points and rank them accordingly to identify the top outlier.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 2 (2016) PAGES 533-537
RECEIVED : 02/03/2016. PUBLISHED : 10/03/2016
March 10, 2016

OUTLIER DETECTION BASED PSO ALGORITHM

Initialise the particles, where each particle encodes {ID, r}

While iteration _ Max Iterations do

For each particle do

Evaluate the particle

1. Compute k for the data point of ID

2. Calculate the fitness function according to EQ. (1)

Using k and r

Update particle's best values ID best, r best

Update swarm's best ID gbest, r gbest

End for

For each particle do

Calculate velocity and position

End for

End while

Using rgbest to compute k/rgbest for all the data points

Sort the point.

4. EXPERIMENT RESULTS

This research, the removal of outliers using PSO algorithm has been done and performance is evaluated by means of Accuracy and Efficiency. Both the metrics has been improved when compared to the existing Genetic algorithm, PSO algorithm.

To measure the effectiveness of the approach, the experiment is conducted on breast cancer datasets. All experiments were performed on Pentium IV Core 2 Duo, PC with 3 GB of main memory, running Windows XP. All procedures were coded in Java 1.6.0.

The number of instances taken for genetic, PSO, PSO (outlier removal) are 286 and accuracy is improved in PSO (Outlier Detection) compared to Genetic Algorithm, PSO Algorithm, Enhanced PSO shown in fig.1

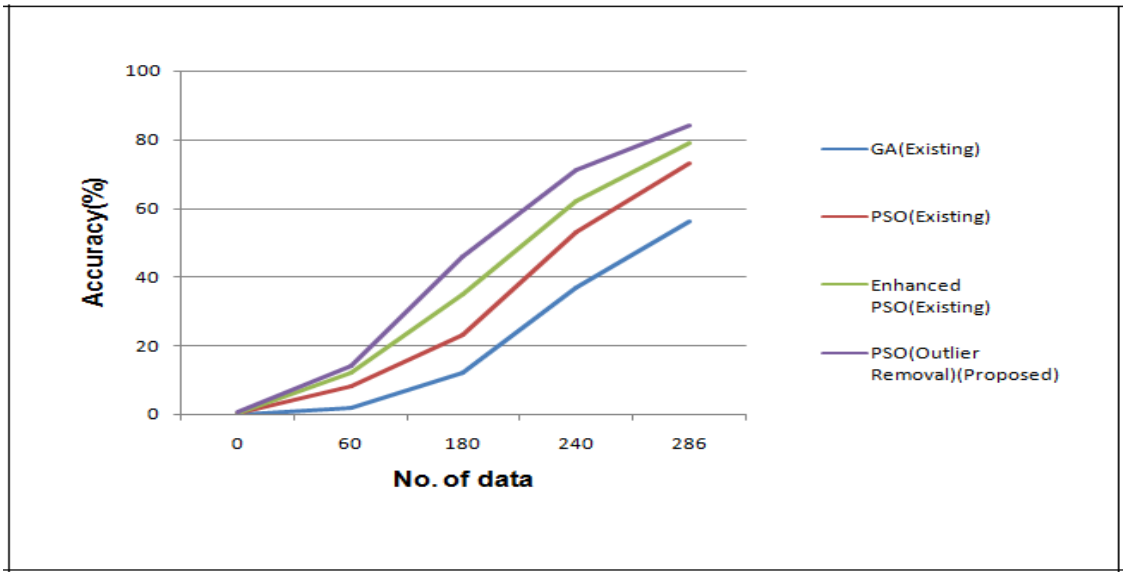

Fig. 1 Number of Data Vs Accuracy

5. CONCLUSION AND FUTURE WORK

The classification rule is used for decision making in many organization. The classification rule has been generated by using GA and PSO.PSO has been considered as best candidate for classification task when comparing its accuracy with GA.In many classification algorithms, which were used earlier does not have clear rule to predict outliers. In order to predict clear rule for identifying outliers, PSO algorithm will be used during classification, which improves accuracy and efficiency. Future work may involve in improving the accuracy and efficiency in PSO algorithm (Outlier removal) for effective classification rule generation.

REFERENCES

1. Ammar W Mohemmed, Mengjie Zhang, Will Browne (2010), "Particle Swarm Optimisation for Outlier Detection".
2. BakerJ E(2009), "Adaptive selection methods for genetic algorithms". Proceedings of the 1st IEEE International Conference on Genetic Algorithms, pp. 101-111.
3. Booker L(2008), "Improving search in genetic algorithm**",** Genetic algorithms and Simulated Annealing, Davis (ed.), Morgan Kaufmann Publishers. pp. 61-73**.**
4. Fayyad U, Piatetsky G Shapiro and Smyth P (2008) From data mining to knowledge discovery: an overview, in: "Advances in Knowledge Discovery and Data Mining", Cambridge, pp.1–34.

5. Hisao Ishibuchi, Tomoharu Nakashima, and Tetsuya Kuroda (2006) ".A hybrid fuzzy genetics-based machine learning algorithm:hybridization michigan approach and pittsburgh approach",.Pattern Recognition, vol. 37, no. 6, pp.1287- 1298.
6. Ishibuihi H, Murala T and Turksen B (2006). "Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems'': Fuzzy Sets and Systems, vol. 89, pp. 134-150.
7. Ishibuchi H, Nakashima T and Murata T (2008) "Performance evaluation of fuzzy classifier systems of multidimensional Pattern Classification Problems", IEEE Transactions on Fuzzy Systems, vol. 29, pp. 601618.
8. Jiawei Han and Micheline Kamber (2006)," Datamining concepts and techniques " (www.infibeam.com).
9. Jaume Bacardit, Natalio Krasnogor (2005) "Smart crossover operator with multiple parents for a Pittsburgh learning classifier system", IEEE Transaction on Pattern analysis and Machine Intelligence, vol. 15, no. 11, pp.1148- 1160.
10. Kennedy J, Eberhart R (2007)" A discrete binary version of the particle swarm Algorithm". IEEE Conference on Systems, Man, and Cybernetics, pp.4104-4109.
11. Kennedy J, Eberhart R(2003). "Particle swarm optimization". Proceedings on IEEE International Conference on Neural Networks, Perth, Australia.