Proceedings of the International Conference on Intelligent Computing and Communication for Smart World (I2C2SW-2018)
14-15 December 2018

Application of Rough Set Based Reduction for Network data set

V.R. Saraswathy
Department of ECE
Kongu Engineering College
Perundurai Tamil Nadu
msaraswathy@gmail.com

M.Prabhu Ram
Department of Mechanical Engineering
Amrita Vishwa Vidyapeetham
Amritapuri
prabhu.5187@gmail.com

A.Vennila
Department of ECE
Kongu Engineering College
Perundurai
vennila.ece@kongu.ac.in

S.G.Dravid
Department of ECE
Kongu Engineering College
Perundurai
dravidgovind1998@gmail.com

***Abstract*— The modern technologies in all the fields constantly generate a large amount of data. The data if it is represented in an understandable form will influence the real world in all respects. The tremendous increase in the data size makes the analysis of the data more tedious. Hence the retrieval of useful information using systems with human approach is essential in today's scenario. Hence feature reduction using Quick reduct , an application of Rough set theory is used to reduce feature set size and identify the useful features based on semi-supervised learning. Particle swarm optimization is used for Quick reduct feature reduction process. The algorithm is applied for network data set.**

Keywords— Network instrusion detection, semi-supervised learning, rough set theory, particle swarm optimization

I. INTRODUCTION

The networking applications developed in different environment with different users create the need for protection system for improving the security. The network systems should avoid unauthorized users. The continuous monitoring of the intruders manually for large networks is not possible but the detection should be done in an intelligent manner and faster. The increasing volume of data in large networks and rapid increase in computer networks put the intrusion detection system in a challenging situation to detect the attacks in a reasonable time. Hence, feature reduction is required for reducing the processing time and space. The identification of the significant features is important[5]. However, to cope with this large amount of data, machine learning approaches are used.

II. MACHINE LEARNING METHODS

Machine Learning incorporates artificial intelligence in problem solving. The components of machine algorithm are representation, evaluation and optimization and machine learning acquires the knowledge as humans do. The machine learning deals with the learning from the past experiences in terms of the output generated for a given task to improve the performance of intelligent programs. The relationships hidden in large amounts of data can be found using machine learning which is practically not possible by humans. Machine learning makes the redesigning of the learning algorithm for the new knowledge easier[3].The learning methods(LM) namely unsupervised (US), supervised(S) and semi-supervised (SS) [1]are classified as shown in Fig. 1.

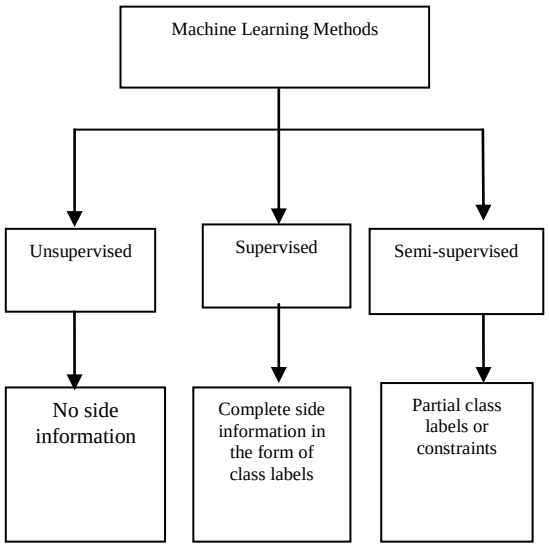

Fig .1. Machine learning methods for clustering

III. FEATURE REDUCTION

The feature reduction technique is based on Rough Set Theory(RST). Quick Reduct(QR) is used for feature reduction. The existing reduct algorithms are for supervised and unsupervised learning methods. The semi-supervised method is proposed for the QR algorithm.

A. Rough Set Theory

RST is important in the areas of machine learning, decision analysis, knowledge discovery from databases. RST uses the characteristics of every object for decision making.

RST uses the decision table which has data about objects of interest characterized in terms of some attributes and consists of condition and decision attributes. The decision table provides information for decision making in terms of decision attributes. The data objects which has same information are termed as indiscernible.

The indiscernibility relation forms the basis of RST. RST consists of a pair of precise sets namely the lower and the upper approximations. The data objects which cannot be

grouped under the approximations results in boundary region.

B. Quick Reduct

The QR algorithm partitions the attributes based on the positive region.

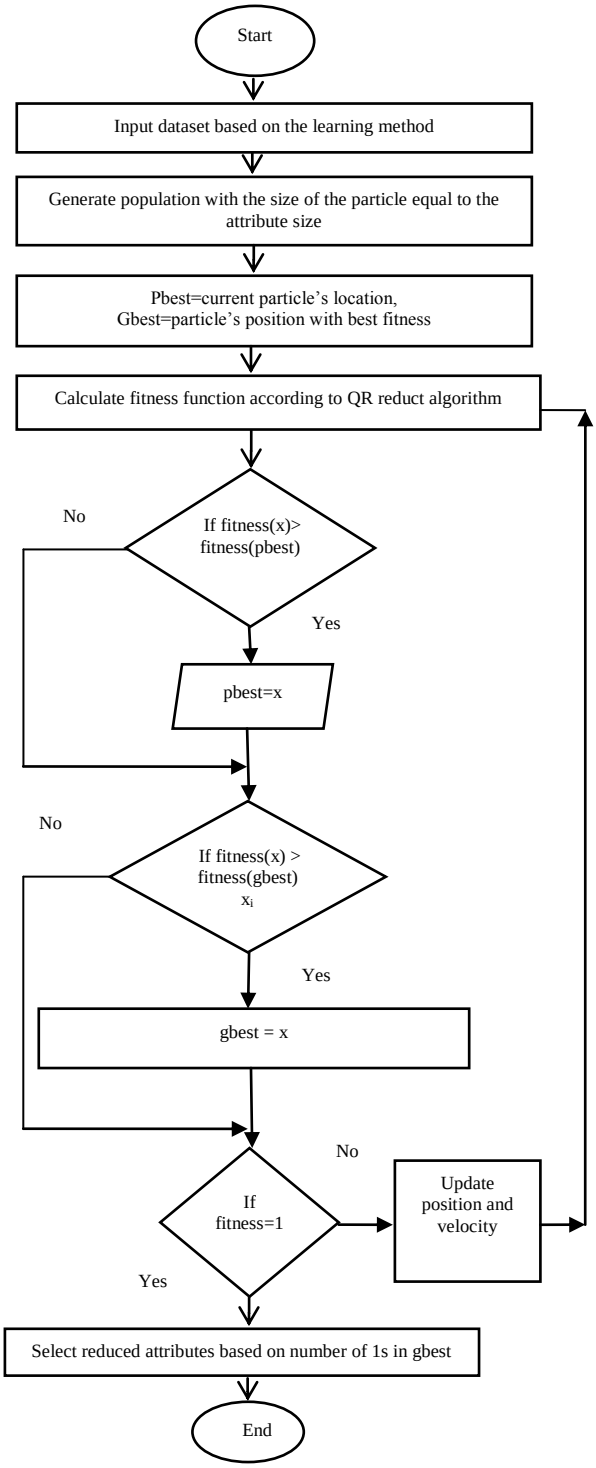

Fig.2. Flow chart for Particle Swarm Optimization for QR algorithm

The positive region of each attribute is calculated. The objects having the same value for an attribute should have the same decision attribute value. Then this defines the positive region for that attribute. The remaining is the negative region which is eliminated. The number of objects defined by the positive region for an attribute determines the dependency of that attribute with the decision attribute. The highest dependency attributes are chosen[6].

Equation (1) gives the dependency degree(γ)

$$\gamma_{Z\cup\{X\}}(B) = \frac{\left|POS_{Z\cup(X)}(B)\right|}{\left|E\right|} \quad (1)$$

Where POS Positive region

E - Objects in the dataset
A - Conditional attributes
B - Decision attribute set
Z - Reduced data set
X - Subset of A

IV. Semi-Supervised Particle Swarm Optimization of Quick Reduct

In QR, the fitness function is given by equation (1). The positive region of each attribute with the other attributes is calculated for finding the fitness value. The feature with the highest fitness is taken. With the possible combinations with that attribute are obtained,the fitness is calculated and the calculation continues till the fitness becomes one. The corresponding particle position represents the reduced attribute. Finally, the number of ones in the gbest value gives the selected features[2]. The flow chart for semi-supervised particle swarm optimized QR (SS-PSO-QR) is given in Fig.2

V. Network Dataset

The data set available for network intrusion detection is KDD CUP '99. The features in this dataset is used for semi-supervised learning. The NSL-KDD used has 39 attributes and 11850 instances[5].

VI. Result And Conclusion

The Particle swarm optimized QR algorithm is applied for the network data set.

The table I shows the result of application of PSO-QR for the three learning methods for network dataset . The results are obtained and tabulated with features selected(FS) and time in secondsT(s). It demonstrates that SS learning results in optimum feature reduced set. The SS learning approach guides the feature reduction compared to US learning and it uses available labeled data.

TABLE I. Optimization of QR algorithm for the three learning methods for Network data set

LM	**Network data set dimension**							
	100		**200**		**300**		**400**	
	FS	**T(s)**	**FS**	**T(s)**	**FS**	**T(s)**	**FS**	**T(s)**
US	18	207.16	20	577.5	23	1122.92	36	3914.62

S	15	142.33	16	322.5	21	523.53	17	716.4
SS	16	372.29	27	921.33	17	1770.02	19	2594.21

REFERENCES

[1] K. Benabdeslem and M. Hindawi, M , 'Efficient Semi-Supervised Feature Selection: Constraint, Relevance, and Redundancy', IEEE Transactions on Knowledge and Data Engineering, vol. 26, no.5, 2014, pp. 1131-1143.

[2] R.C.Eberhart and Y.Shi, 'Particle swarm optimization: Developments, applications and resources', Proceedings of Congress on Evolutionary Computation, vol.1, 2001, pp.81-86.

[3] F.Gharibian and A. Ghorbani, 'Comparative study of supervised machine learning techniques for intrusion detection', In Communication Networks and Services Research, CNSR'07, Fifth Annual Conference on IEEE, 2007, pp. 350-358.

[4] H.Hannah Inbarani, Ahmad Taher Azar and G.Jothi,'Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis', Computer Methods and Programs in Biomedicine, vol.113, no.1,2014, pp. 175-185.

[5] H.G.Kayacik, A.N. Zincir-Heywood and Heywood, M.I , 'Selecting features for intrusion detection: A feature relevance analysis on KDD'99 intrusion detection datasets', In Proceedings of the third annual conference on Privacy, Security and Trust, 2005,pp. 122-127.

[6] C. Velayutham and K.Thangavel,'Unsupervised Quick Reduct Algorithm Using Rough Set Theory',Journal of Electronic Science and Technology, vol. 9, no.3, 2011,pp. 193-201.