Analysis of Intrusion Detection System Based on K-means Algorithm and Particle Swarm Optimization

Kalaimathi B[1], Annapoorani. G[2] and Gayathri N[3]

[1]*Assistant Professor, Department of CSE , MVJ College of Engineering*
Channasandra, Near ITPL, Bangalore-67, India

[2]*Assistant Professor, Department of CSE, MVJ College of Engineering*
Channasandra, Near ITPL, Bangalore-67, India

[3]*Assistant Professor& HOD,Department of ISE, MVJ College of Engineering*
Channasandra, Near ITPL, Bangalore-67, India

Abstract:

In this modern world internet security has been one of the most threatening problems in the world. Intrusion detection is the important process of monitoring the events in the network system to find out the hackers. Clustering is the most important unsupervised learning process used to find the structures or patterns in a collection of unlabeled data. A intrusion detection system model based on PSO optimization and K-means was analysed in this paper.

Keywords: Intrusion Detection System, Cluster, Particle swarm optimization, K-means algorithm.

1. Introduction:

Clustering analysis is an important part of the Data Mining research; it is an important method of the unsupervised learning. It divides the data into certain polymerization classes according to the attribute of the data. We need effective intrusion detection systems to protect our computers from unauthorized or malicious actions. Data mining can improve variants detection rate, control false alarm rate, and reduce false dismissals .Data mining based on intrusion detection systems can be roughly categorized into major two Groups misuse detection and anomaly detection. Network intrusion detection is the process of monitoring the events occurring in a computing system or network and analysing them for signs of intrusions, defined as attempts to compromise the confidentiality.

The intrusion attacks can be divided into four categories: Probe (e.g. IP sweep, vulnerability scanning), denial of service (DoS) (e.g. mail bomb, UDP storm), user-to-root (U2R) (e.g. buffer overflow attacks, root kits) and remote-to-local (R2L) (e.g. password guessing, worm attack).Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar, and the members from different clusters are quite different from each other[1]. Therefore clustering methods can be useful for classifying log data and detecting intrusion.

Clustering algorithms can be categorized into four main groups: partitioning algorithm, hierarchical algorithm, density-based algorithm and

grid-based algorithm. Partitioning algorithms construct a partition of a database of N objects into a set of K clusters. Usually they start with an initial partition and then use an iterative control strategy to optimize an objective function. PSO algorithm converge rapidly .It is easy to adjust parameter and can be applied to the condition when there is large number of samples and the dimensions of samples are large. K-means clustering algorithm is an effective method has been proved for apply to the intrusion detection system but it is part of the optimal solution.

2. K-means Algorithm:

Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar, and the members from different clusters are quite different from each other groups: partitioning algorithm, hierarchical algorithm, density-based algorithm and grid-based algorithm[7]. Partitioning algorithms construct a partition of a data base of N objects into a set of K clusters. Usually they start with an initial partition and then use an iterative control strategy to optimize an objective function.

Idea of Algorithm:

Given the D-dimensional data set
$X= \{x_i/x_i \varepsilon R^d \ \ i=1,2,\ldots\ldots\ldots N\}$
Clusters are w_1, w_2, w_3 ….w_k. To define K centroids (c_1, c_2, c_3…….c_k), one for each cluster,
$C_i = 1/n_i \sum x$ n_i is the number of datasets in the cluster.
The basic K-means algorithm[1] consists of the following steps:
(1) Assigns the clustering number K.
(2) Random select K points C_1, C_2… C_K as the initial clustering centers from the given point set $\{x_1, x_2 \ldots x_N\}$
(3) Select C_1, $C_2 \ldots C_K$ as the clustering centers and divide the set $\{x_1, x_2 \ldots x_N\}$ according to the following regulation:
If d (x_i, c_p) <d (x_i, c_q) p,
q=1, 2…k and p! =q,
then x_i K_p (Kp is a class and its center is Cp)
(4) Recalculate the new clustering centers
C_1^1 ,$C2^1$… C_K^1 according to the equation
$C_1^1=1/|ki|\sum x_j$
i =1, 2… K
|ki| is the point number in ki
(5) If $c_i^1=c_i$ i=1, 2……, k (or the algorithm has achieved the hypothesis biggest iterative times) then terminate the algorithm, else make $Ci=C_1^1$ and return to point 3

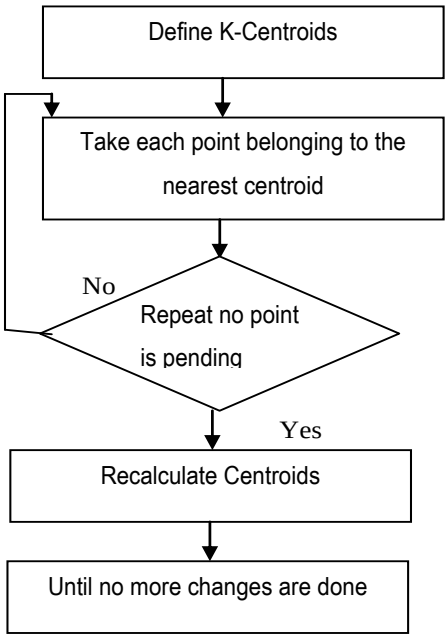

Figure1:K-means algorithm flow chart

3. Intrusion detection system in k-means algorithm:

Due to data preprocessing different features of raw data are on different scales[3]. This causes bias toward some larger features over other smaller features. This leads to intrusion in clusters. To solve the problem a measurement is performed as follows
1) First calculate the mean absolute eviation hf:
$h_f =1/n(|x_{1f}-m_f|+|x_{2f}-m_f|+\ldots\ldots\ldots\ldots+|x_{nf}-mf|)$
where x_{1f},x_{2f}…..x_{nf} are n measuring values of variants ,tf is the mean value of the variant f, that is
$t_f=(x_{1f}+x_{2f}+\ldots\ldots\ldots+x_{nf})/n;$
2) Calculate the standardized measurement:
$Y_{if}=x_{if}-tf/h_f$
3) Then convert every instance in the training sets to a new one based on previous algorithm.
This is the method that transforms the standardized space based on statistical information retrieved from the training sets and also to detect intrusion.

4. Particle swam optimization algorithm:

PSO is an effectively global optimization algorithm, it guides optimization search by the Swarm Intelligence, which comes from cooperation

and competition between particles, Compares with the evolutionary algorithm, PSO retains the global search strategy based on population, its operation is simple, and solution of each generation population has Dual advantages of Self-learning and learning from others. So it can find the optimal solution by lesser iterative times.

Particle Swarm Optimization (PSO) is a swarm intelligence method developed by Kennedy and Eberhart in 1995 .The behavior of PSO mimics the social interaction between individuals such as interactions between the birds in flocks trying to locate an optimal food source. The direction of the movement of each bird is controlled by its current location, the best food location it ever found, and the best food location any bird in the flock ever found. The basic process of the PSO[5] algorithm is given by:

Step 1: (Initialization) randomly generate initial particles.
Step 2: (Fitness) Measure the fitness of each particle in the population.
Step 3: (Update) Compute the velocity of each particle. V=m/h; M represent at which time the connection can open. It represent at which the connection can request.
Step 4: (Construction) For each particle, move to the next position.
Step 5: (Termination) Stop the algorithm if the termination criterion is satisfied; return to Step 2 otherwise

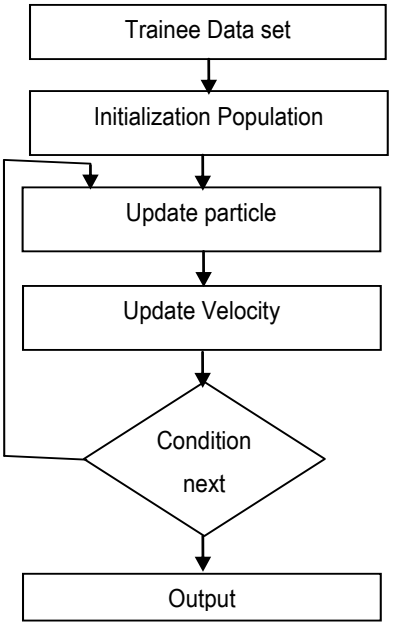

Figure 2:PSO Algorithm flowchart

In PSO, a number of simple entities called particles are placed in the search space of the target problem, and each particle evaluates its position based on a given objective function calculating the fitness value. Each particle then determines its movement (velocity) through the problem search space by taking the history of its own current position and best position achieved by the whole swarm. Furthermore, the movement of a particle is affected by its inertia, and other constants. However, the whole swarm after several iterations is likely to move close to the global best solution.PSO[5] updates the particles positions inside the problem search space using the following equations:
Xi (t+ 1) =Xi (t) +Vi (t+ 1)
Where Xi is the position of particle i,t is the iteration number, and Vi is the velocity of particle I given by the following equation:
Vi (t+ 1) =WVi (t) +r1c1 [XPiXi (t)] +r2c2[XGXi(t)]
Where W is inertia weight,r1andr2 are randomly generated numbers uniformly distributed between 0 and 1,c1,c2 are constant coefficients, XPi is the current best position of particle i and XG is the current global best position of the whole swarm.

5. Particle Swarm Optimization for Intrusion Detection systems:

A module of the system namely Red Teams emulates the behavior of hackers[5]. The Red Teams component employs PSO techniques in their intrusion methodology. The acquired results can dynamically help the IDS reconfigure on-the-fly in order to be more effective. Since most of the PSO based IDS are hybrid anomaly detection systems, it is possible to categorize them according to the additional ML method that is employed. We distinguish
(a)Hybrid PSO-Neural Network Systems,
(b)Hybrid PSO-SVMSystems,
(c)Hybrid PSO-K-means Systems

a)PSO & neural network hybrid approaches:

Artificial Neural Networks (ANN)[2] is one of the most popular soft computing techniques for data classification. PSO is a technique which is used extensively in combination with various types of ANN for improving the performance of the system. During the training phase the PSO is executed recursively to train the network. In PSO-ANN approach, So corresponds to synaptic weigths are fed to ANN during the test phase. To detect intrusion in system PSO-ANN approach is divided into two processes:

i) ANN, a system component does the classification process

ii) PSO is used to improve the critical parameters and train the synaptic weights.

Advantages:

A hybrid PSO-ANN system but also introduce an evolutionary mutation algorithm as an extra step in order to

(a) Protect PSO from trapping into local minima
(b) Increase the diversity of the population
(c) Expand the scope of the search

(b) Hybrid PSO-SVM Systems:

Similarly to ANN, another technique frequently used in combination with PSO[5] is Support Vector Machines (SVM). SVM is based on structural risk minimization of statistical learning theory and shows good learning ability and generalization skill in high dimensional or noisy datasets, two attributes highly appreciated in intrusion detection They used two different flavors of PSO the Standard Particle Swarm Optimization (SPSO) and Binary Particle Swarm Optimization (BPSO) for seeking optimal SVM parameters and extracting a feature subset respectively.

i) In BPSO each particle features and parameters values should be taken as first then that results and training datasets are fed to the SVM classifiers.

ii) In SPSO, the classification process and optimum features happen simultaneously .so the dataset features and the crucial SVM parameters are represented by each particle position.

Advantage: Hybrid PSO-SVM System is more accurate to detect intrusion.

c) hybrid PSO-K-means Systems

In PSO-K-means systems[6], each particle's position is the set of D dimensional centroids produced by the K-Means algorithm. Thus, each particle's position can be represented as an array:

$$\begin{vmatrix} Z_{11}\ Z_{12}\ .\ Z_{1D} \\ Z_{21}\ Z_{22}\ .\ Z_{2D} \\ .\ .\ .\ . \\ Z_{k1}\ Z_{k2}\ .\ Z_{kD} \end{vmatrix}$$

where D is the number of the dimensions of the dataset (therefore the centroids dimension) and k represents the

number of clusters. Initially, data points are assigned to k clusters in a random manner. Then the centroids are calculated and the position of each particle is deduced. For each particle, the fitness function evaluates the position and if necessary the Pbest and Gbest values are updated along with that of velocity and position.

Finally, the K-Means algorithm runs in order to optimize the new generation of particles. The algorithm converges to local optimum with very low probability and has high convergence speed.

6.Comparison of k-means and Particle Swarm Optimization:

Advantages of K-mean clustering

•K-mean clustering is simple and flexible.
•K-mean clustering algorithm is easy to understand and implements.

Disadvantages of K-mean clustering

•In K-mean clustering user need to specify the number of cluster in advanced

K-mean clustering algorithm performance depends on a initial centroids that why the algorithm doesn't have guarantee for optimal solution

Advantages of PSO

•PSO based on the intelligence and it is applied on both scientific research and engineering.
•PSO has no mutation and overlapping calculation. The search can be take place by the speed of the particle. Most optimist particle can able to transmit the information onto the other particles during the development of several generations, and the speed of researching is faster.[8]
•PSO accepts the real number code, and that is decided directly by the solution. Calculation in PSO is simpler and efficient in global search

Disadvantages of PSO

•It is slow convergence in refined search stage and weak local search ability.
•The method cannot work on the problems of non-coordinate systems like the solution of energy field and the moving rules for the particles in the energy field

7. Conclusion:

Survey of K-mean clustering algorithm is most widely used clustering local search method for detecting intrusion in systems. K-mean which is depending on initial condition, which causes the algorithm, may converge to suboptimal solution. K-means method is an effective algorithm for partioning large data set. One basic point to be taken into account is that the use of PSO has significantly boosted the performance of all the machine learning techniques in which it was applied. So Particle swarm optimization is more likely to find near optimal solution. Compare to K-means as we have discussed in this paper, the PSO has more approaches to detect intrusion in the systems.

8. References:

[1] International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-3, February 2013 A Survey on K-mean Clustering and Particle Swarm Optimization Pritesh Vora, Bhavesh Oza

[2]Swarm intelligence in intrusion detection: A survey C. Kolias a,b,*, G. Kambourakis a,b, M. Maragoudakis a,b a Laboratory of Information and Communication Systems Security, University of the Aegean, Samos GR-83200, Greece b Department of Information and Communication Systems Engineering, University of the Aegean, Samos GR-83200, Greece,2011

[3]The Application on intrusion detection based on k-means algorithm.Meng Jianling Shang Haikum Bian Ling Department of computer science and technology college,North China Electic Power University,Hebei Boading 2009International Conference on information technology and applications.

[4]2008 International Conference on Intelligent Computation Technology and Automation The Clustering Algorithm Based on Particle Swarm Optimization Algorithm Pei Zhenkui1,2 , Hua Xia1 , and Han Jinfeng1 1College of Computer and Communication ngineering, China University of Petroleum,Dongying 257061, China 2School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044

[5] The Clustering Algorithm based on particle swarm optimization algorithm Pei Zhenkui,Hua Xia,and Han Jinfeng ,2008 International Conference on Intelligent Computation Technology and Automation.

[6] 2008 International Conference on Intelligent Computation Technology and Automation The Clustering Algorithm Based on Particle Swarm Optimization Algorithm Pei Zhenkui1,2 , Hua Xia1 , and Han Jinfeng1 1College of Computer and Communication ngineering, China University of Petroleum,Dongying 257061, China 2School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044

[6]S. Gao, J.Y. Yang, "New Clustering Method Based on Particle Swarm Algorithm", Journal of Nanjing University of Aeronautics &Astronautics, 2006, pp.62-65.

[7] Bradley,Fayyad.Refining Initial Point for K-means Clustering.Proceedings of the Fifteenth International Conference on Machine Learning,1998