IJRSET November 2015 Volume 2, Issue 12 Pages: 32-36
ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

DATA MINING BY ADOPTING PARTICLE SWARM OPTIMIZATION

[1] M.LAKSHMI DURGA, [2] P. LALITHA
M.Phil Research Scholar. Assistant Professor,
Dept Of Computer Application,
Hindustan College Of Arts And Science,
Coimbatore.

Abstract:-

The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions. With increasing dimensionality, these algorithms soon become computationally intractable and therefore inapplicable in many real applications. Secondly, the specificity of similarities between points in a high dimensional space diminishes. Continuous improvement processes based on the principles of total quality management that including customer orientation, quality orientation and affairs implementation as shape of team is always from interest principle of dynamic and successful organizations. PSO is a stochastic, population-based evolutionary algorithm particularly suitable for solving multi-variable optimization problems. It embeds a kind of swarm intelligence that is based on socio-psychological principles and provides insights into social behavior contributing to engineering applications.

Keywords: - [Data Mining, PSO, Adaptive K-Nearest, K-Means]

1. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

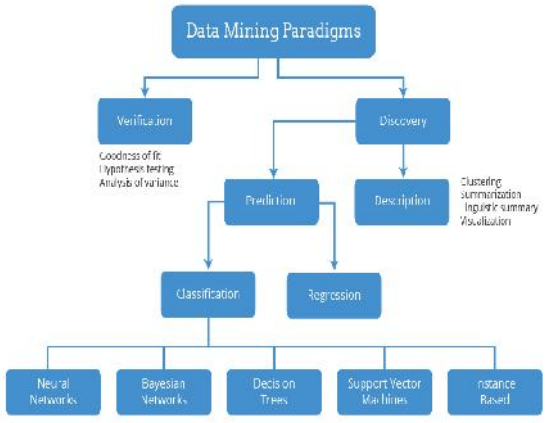

Figure: 1 Data mining paradigms

2. METHODOLOGY

PSO is a stochastic, population-based evolutionary algorithm particularly suitable for solving multi-variable optimization problems. It embeds a kind of swarm intelligence that is based on socio-psychological principles and provides insights into social behavior contributing to engineering applications (Kennedy & Eberhart, 1995). Population dynamics as defined in classical PSO simulates bio-inspired behavior i.e. a bird flock's behavior, which involves sharing of information and allows particles to gain profit from the discoveries and previous experience while in quest of food. In PSO-based applications, each particle represents a candidate solution and flies through the search space. The position of a particle is biased by the best position visited using its own knowledge and the position of the best particle regarded by the knowledge of neighboring particles. When the neighborhood of a particle is the entire swarm, the particle is said to be the global best particle.

3. PSO ALGORITHM

Particle swarm optimization (PSO) is a stochastic global optimization technique developed by Beernaert and Kennedy in 1995 based on social behavior of birds [2]. In PSO a set of particles or solutions traverse the search space with a velocity based on their own experience and the experience of their neighbors. During each round of traversal, the velocity, thereby the position of the particle are updated based on the above two parameters. This process is repeated till an optimal solution is obtained. According to the original PSO the particle velocity and position are updated according to the following equations.

$$v_k^{in+1} = v_k^{in} + c_1 r_1 \left(pbest_k^{in} - p_k^{in}\right) + c_2 r_2 \left(gbest^{in} - p_k^{in}\right) \quad (1)$$

$$x_k^{in+1} = x_k^{in} + v_k^{in+1} \quad (2)$$

where v k i n and p k i n are the velocity and position of kth particle in ith dimension during nth iteration, pbest is the best position experience by the particle upto that iteration and gbest is the best position experience by all particles upto that iteration. The best positions of a particle are evaluated according to a fitness function.c1, c2 are called acceleration constants usually equal to 2 and r1 and r2 are random numbers uniformly distributed in (0, 1). Thus these constants are a measure of inertia experienced by the particle. The PSO developed by Eberhart and Kennedy is suited for continuous optimization problems.

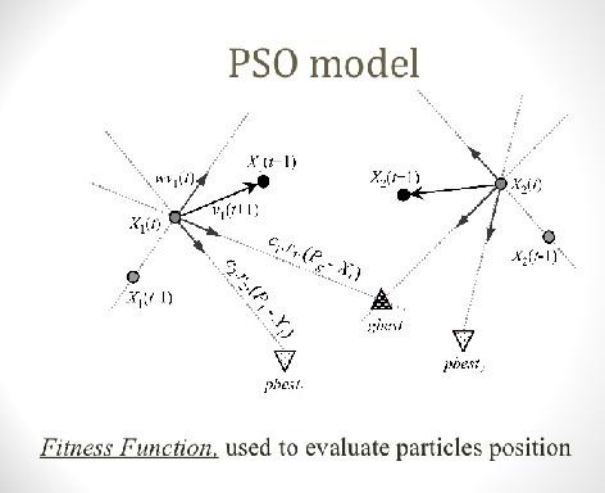

Figure: 2 Particle swarm optimization

4. ADAPTIVE K-NEAREST NEIGHBORHOOD

If the entire data sample G is expressed as {G[i, j]}, where i is the number of genes and j is the number of samples, a k-fold cross validation technique (Simon, Subramanian, Li, & Menezes, 2011) is employed to divide the entire data sample into k equal subsets of samples. Of the k subsets of samples, a single subset of samples is retained as the validation data for testing and the remaining (k _ 1) subsets of samples are used as training data. This process then repeated k times, with each of the k subset of samples used exactly once as the validation data. Each time value of K for KNN classifier is varied from 3 to 20.

$$\text{CV training accuracy} = \frac{1}{k}\sum_{m=1}^{k} Tr_m \quad (5)$$

$$\text{CV test accuracy} = \frac{1}{k}\sum_{m=1}^{k} Ts_m \quad (6)$$

$$\text{CV training accuracy} = \frac{1}{k}\sum_{m=1}^{k} Tr_m \quad (5)$$

$$\text{CV test accuracy} = \frac{1}{k}\sum_{m=1}^{k} Ts_m \quad (6)$$

where *Tr* and *Ts* are the training accuracy and test accuracy for each fold respectively.

For each *K*, *K* numbers of single estimations (SE) are calculated as:

$$SE(K) = \frac{1}{2}\sum_{K=3}^{20}\left(\frac{1}{k}\sum_{m=1}^{k} Tr_m + \frac{1}{k}\sum_{m=1}^{k} Ts_m\right) \quad (7)$$

5. MODIFIED K-NEAREST NEIGHBOR

The main idea of the presented method is assigning the class label of the data according to K validated data points of the train set. In other hand, first, the validity of all data samples in the train set is computed. Then, a weighted KNN is performed on any test samples. Fig. 1 shows the pseudo code of the MKNN algorithm. In the rest of this section the MKNN method is described in detail, answering the questions, how to compute the validity of the points and how to determine the final class label of test samples. Validity of the Train Samples In the MKNN algorithm, every sample in train set must be validated at the first step. The validity of each point is computed according to its neighbors. The validation process is performed for all train samples once. After assigning the validity of each train sample, it is used as more information about the points. To validate a sample point in the train set, the H nearest neighbors of the point is considered. Among the H nearest neighbors of a train sample x, validity(x) counts the number of points with the same label to the label of x. The formula which is proposed to compute the validity of every points in train set is (1).

$$Validity(x) = \frac{1}{H}\sum_{i=1}^{H} S(lbl(x), lbl(N_i(x)))$$

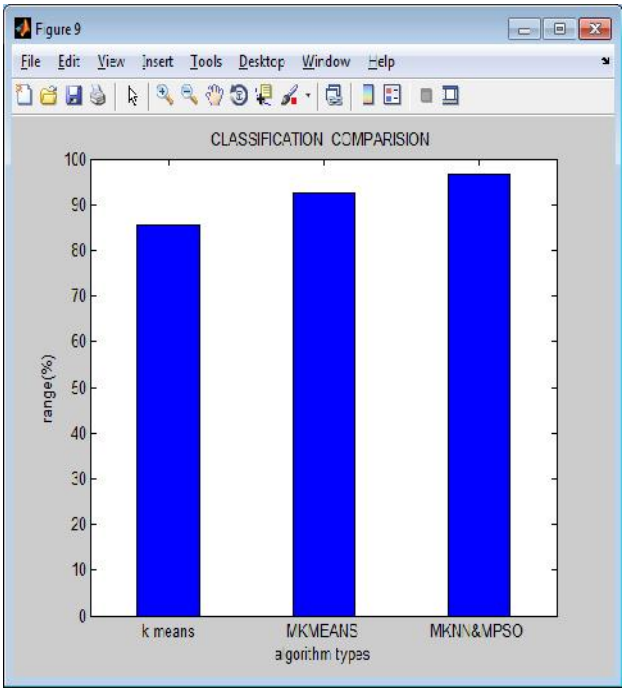

Figure : 3 Classification Comparision

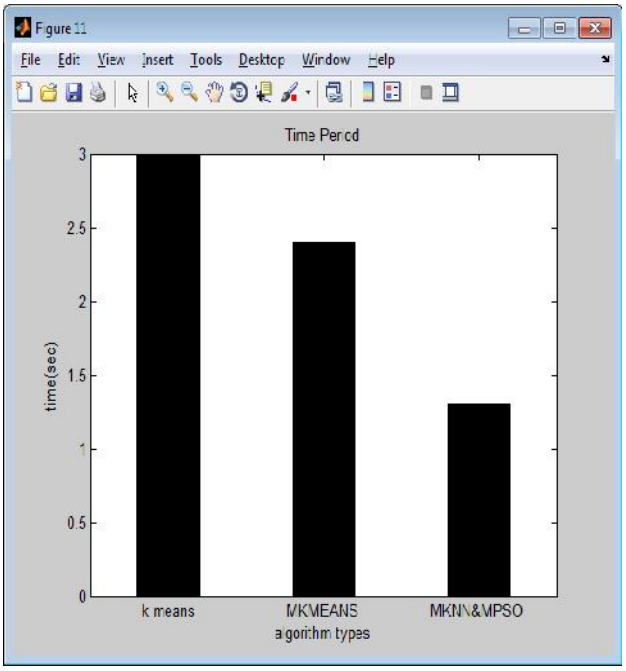

Figure:4 Time Period Comparision

CONCLUSION

An approach of the K-Means clustering based on genetic algorithm has been proposed, to a certain degree, which overcomes the defects that is sensitive to initial value and it is easily able to be trapped in a local optimum. The practicability of this approach is analyzed in principle, and its practical effect is confirmed by experiment in technical. The experiment results show that the global distribution characteristic of the space clustering centers which are found during the process of the K-Means clustering analysis by utilizing GA is properly kept, so clustering effect is more rational. Association Rule Mining is used for intrusion detection in this paper. Use of K-Means logic overcomes the sharp boundary problem caused by the association rules. Thus K-Means association rules can be mined to find the abstract correlation among different security features. Using genetic algorithms with the K-Means data mining method may result in the tune of the K-Means Membership functions to improve the performance and select the set of features available from the audit data that provide the most information to the data mining component. These algorithms are mostly used for optimization problems.

REFERENCES

[1].D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10), 906–914.

Ganesh Kumar, P., Aruldoss Albert Victoire, T., Renukadevi, P., & Devaraj, D. (2012).

[2].Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm. Expert Systems with Applications, 39, 1811–1821.

[3].Horng, J. T., Wu, L. C., Liu, B. J., Kuo, J. L., Kuo, W. H., & Zhang, J. J. (2009). An expert system to classify microarray gene expression data using gene selection by decision tree. Expert Systems with Applications, 36, 9072–9081.

[4].Ji, G., Yang, Z., & You, W. (2011). PLS-based gene selection and identification of Tumor-specific genes. IEEE Transactions on Systems, Man and Cybernetics – Part C:Applications and Reviews, 41(6), 830–841.

[5].Kennedy, J., & Eberhart, R. C. (1995). Particle swarm optimization. In Proceedings of the IEEE international conference on neural networks (Vol. 4, pp. 1942–1948).

[6].Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., et al. (2001). Classification and diagnostic prediction of cancers using gene expressing profiling and artificial neural network. Nature Medicine, 7, 673–679.

[7].Lee, C. P., Lin, W. S., Chen, Y. M., & Kuo, B. J. (2011). Gene selection and sample Classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. Expert Systems with Applications, 38, 4661–4667.

[8].Li, H. et al. (2013). Genetic algorithm search space splicing particle swarm optimization as general-purpose optimizer. Chemometrics and Intelligent Laboratory Systems, 128, 153–159.

[9].Li, X., & Shu, L. (2009). Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis. Expert Systems with Applications, 36,7644–7650.

[10].Li, S., Wu, X., & Tan, M. (2008). Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Computing, 12, 1039–1048.

[11] Y. Ma, H. Lu and L. He, A genetic encoding scheme for fuzzy clustering, Journal of Air Force Radar Academy, vol.16, no.1, pp.40-41, 2002.

[12] X. Dai and M. Li, A dynamic clustering method based on genetic algorithms, Systems Engineering Theory and Practice, no.10, pp.108-116, 1999.
[13] T. J. Ross, Fuzzy Logic with Engineering Applications, Electronics Industry Press, 2003.
[14] X. Wu and Z. Lin, Matlab Auxiliary Fuzzy System Designs, Xidian University Press, 2002.
[15] W. Pedrycz, Collaborative and knowledge-based fuzzy clustering, International Journal of Innovative Computing, Information and Control, vol.3, no.1, pp.1-12, 2007.
[16] M. Sato-Ilic, General class of weighted fuzzy cluster loading models, International Journal of Innovative Computing, Information and Control, vol.4, no.5, pp.1023-1032, 2008.
[17] K. Zou, J. Hu and X. Kong, The structure optimized fuzzy clustering neural network model and its application, International Journal of Innovative Computing, Information and Control, vol.4, no.7,pp.1627-1634, 2008.