IJRSET August 2015 Volume 2, Issue 6 Pages: 72-77
ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

MICROARRAY GENE EXPRESSION MINING USING PARTICLE SWARM OPTIMIZATION AND MODIFIED K-MEANS AND K-NEAREST ALGORITHM

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

Abstract:-

These days, microarray gene expression data are playing an essential role in cancer classifications. However, due to the availability of small number of effective samples compared to the large number of genes in microarray data, many computational methods have failed to identify a small subset of important genes. Therefore, it is a challenging task to identify small number of disease-specific significant genes related for precise diagnosis of cancer sub classes. In this paper, particle swarm optimization (PSO) method along with adaptive K-nearest neighborhood (KNN) based gene selection technique are proposed to distinguish a small subset of useful genes that are sufficient for the desired classification purpose. A proper value of K would help to form the appropriate numbers of neighborhood to be explored and hence to classify the dataset accurately. Thus, a heuristic for selecting the optimal values of K efficiently, guided by the classification accuracy is also proposed.

The fuzzy c-means clustering algorithm (FCM) is applied extensively. However, it can easily be trapped in a local optimum, and also strongly depends on initialization. Therefore, a method of fuzzy clustering by using genetic algorithm is proposed in this paper. Genetic algorithm refers to choose the number of cluster centers and the data that are cluster centers firstly, and clustering analysis is processed by FCM consequently. Experiment results show that the method can search global optimum partly to make the clustering analysis more rational.

Keywords: - [PSO, Modified K –Means, K-Nearest]

1. INTRODUCTION

This proposed technique of finding minimum possible meaningful set of genes is applied on three benchmark microarray datasets, namely the small round blue cell tumor (SRBCT) data, the acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) data and the mixed-lineage leukemia (MLL) data. Results demonstrate the usefulness of the proposed method in terms of classification accuracy on blind test samples, number of informative genes and computing time. Further, the usefulness and universal characteristics of the identified genes are reconfirmed by using different classifiers, such as support vector machine (SVM).

The proposed method use to fuzzy based genetic algorithm for clustering and classified the genes our proposed algorithm is more efficient for accuracy and time period ,clustering rang ,classification time is

high and most efficient . Focused in ranging from old nearest neighbor analysis to support vector machine manipulation for the learning portion of the classification model. We don't have a clear picture of supervised classifier (Supervised Multi Attribute Clustering Algorithm) which can manage knowledge attributes coming two different knowledge streams. Our proposed systems take the input from multiple sources, create an ontological store, cluster the data with attribute match association rule and followed by classification with the knowledge acquired.

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. This phenomenon may render many data mining tasks (e.g., clustering) ineffective and fragile because the model becomes vulnerable to the presence of noise.

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. In the meantime to obtain feedback of logical and scientific from the needs and expectations of customers control not only can be used as a means to monitor and control in organizations but in a more comprehensive look can be one of the main prerequisites for institutional planning process Therefore, a continuous awareness from the ultimate satisfaction of each customer or Intermediary in continuous improvements promotes of providing and securing blood products is Very important. It is obvious that planning and operating organization without the knowledge of

Their customers' expectations were unable to accountability to the changing and growing needs of its customers and eventually will stagnate and break .It must be said that essential to a democratic society is to accountable clear and appropriate of organizations for the people. Every administrative organization formed to respond to a range of social needs and during its life is continuous so long to respond to the social needs of its period and whenever social conditions change its nature, they eliminate or structure may be modified adapted to new social conditions. The other side of it can be concluded from client satisfaction is consider as a criterion to determine the effectiveness and efficiency of the organization

2. PARTICLE SWARM OPTIMIZATION (PSO)

Particle Swarm Optimization (PSO) is a biologically inspired computational search and optimization method developed in 1995 by Eberhart and Kennedy based on the social behaviors of birds flocking or fish schooling. A number of basic variations have been developed due to improve speed of convergence and quality of solution found by the PSO. On the other hand, basic PSO is more appropriate to process static, simple optimization problem. Modification PSO is developed for solving the basic PSO problem. The observation and review 46 related studies in the period between 2002 and 2010 focusing on function of PSO, advantages and disadvantages of PSO, the basic variant of PSO, Modification of PSO and applications that have implemented using PSO. The application can show which one the modified or variant PSO that haven't been made and which one the modified or variant PSO

Particle swarm optimization (PSO) is a biologically inspired computational search and optimization method developed in 1995 by Eberhart and Kennedy based on the social behaviors of birds flocking or fish

schooling. Recently, there are many variants of PSO, and it may always grow rapidly. The process of PSO algorithm in finding optimal values follows the work of an animal society which has no leader. Particle swarm optimization consists of a swarm of particles, where particle Represent a potential solution (better condition). Particle will move through a multidimensional search space to find the best position in that space (the best position may possible to the maximum or minimum values).

we have made review of the different methods of PSO algorithm. Basic particle swarm optimization has advantages and disadvantages, to overcome the lack of PSO. There are several basic variant of PSO. The basic variants as mentioned above have supported controlling the velocity and the stable convergence. At the other hands, modified variant PSO help the PSO to process other conditions that cannot be solved by the basic PSO.The observation and review is made to show the absolute function of PSO, advantages and disadvantages of PSO, the basic variant of PSO, Modification of PSO and applications that have implemented using PSO. The application can show which one the modified or variant PSO that haven't been made and which one the modified or variant PSO that will be developed.

3. MODIFIED K-MEANS CLUSTERING

The k-means algorithm is one of the most widely used clustering algorithms and has been applied in many fields of science and technology. One of the major problems of the k-means algorithm is that it may produce empty clusters depending on initial center vectors. For static execution of the k-means, this problem is considered insignificant and can be solved by executing the algorithm for a number of times. In situations, where the k-means is used as an integral part of some higher level application, this empty cluster problem may produce anomalous behavior of the system and may lead to significant performance degradation. This paper presents a modified version of the k-means algorithm that efficiently eliminates this empty cluster problem.

In cluster analysis, the k-means algorithm can be used to partition the input data set into k partitions (clusters). However, the pure k-means algorithm is not very flexible, and as such of limited use (except for when vector quantization as above is actually the desired use case!). In particular, the parameter k is known to be hard to choose (as discussed below) when not given by external constraints. In contrast to other algorithms, k-means can also not be used with arbitrary distance functions or be use on non-numerical data.

The basic k-means algorithm is commonly measured by any of intra-cluster or inter-cluster criterion. A typical intra-cluster criterion is the squared-error criterion (Equation 1). It is the most commonly used and a good measure of the within-cluster variation across all the partitions. For the current work, we consider intra-cluster squared-error function to evaluate the present scheme of clustering. In basic k-means algorithm, a set D of d dimensional data is partitioned into K clusters, starting with a set of K randomly generated initial center vectors.

The process iterates through the following steps:

• Assignment of data to representative centers upon minimum distance, and

• Computation of the new cluster centers.

The process stops when cluster centers (or the metric M) become stable for two consecutive iterations. The basic k-means algorithm is greedy in nature.

K-means is the most popular and easy-to-understand clustering algorithm. The main idea of K-means is summarized in the following steps:

Arbitrarily choose k objects to be the initial cluster centers/centroids;

Assign each object to the cluster associated with the closest centroid; Compute the new position of each centroids by the mean value of the objects in a cluster; and Repeat Steps 2 and 3 until the means are fixed.

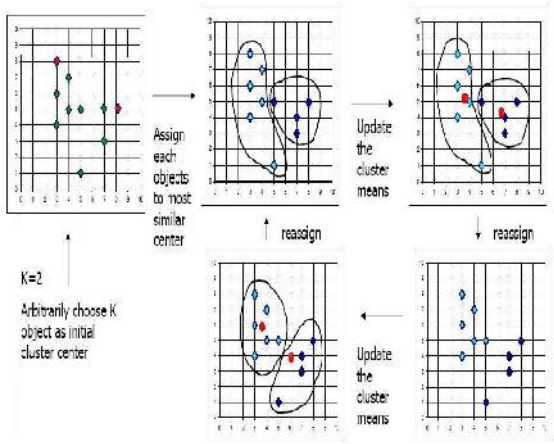

Figure 1: The process of K-means clustering algorithm.

4. 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.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).

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

Where H is the number of considered neighbors and lbl(x) returns the true class label of the sample x. also, Ni(x) stands for the ith nearest neighbor of the point x. The function S takes into account the similarity between the point x and the ith nearest neighbor. The (2), defines this function.

$$S(a,b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$$

Applying Weighted KNN Weighted KNN is one of the variations of KNN method which uses the K nearest neighbors, regardless of their classes, but then uses weighted votes from each sample rather than a simple majority or plurality voting rule. Each of the K samples is given a weighted vote that is usually equal to some decreasing function of its distance from the unknown sample. For example, the vote might set be equal to 1/(de+1), where de is Euclidian distance. These weighted votes are then summed for each class, and the class with the largest total vote is chosen. This distance weighted KNN technique is very similar to the window technique for estimating density functions. For example, using a weighted of 1/(de+1) is equivalent to the window technique with a window function of 1/(de+1) if K is chosen equal to the total number of training samples [19].

In the MKNN method, first the weight of each neighbor is computed using the 1/(de+0.5). Then, the validity of that training sample is multiplied on its raw weight which is based on the Euclidian distance. In the MKNN method, the weight of each neighbor sample is derived according to (3).

$$W(i) = Validity(i) \times \frac{1}{d_e + 0.5}$$

Where W(i) and Validity(i) stand for the weight and the validity of the ith nearest sample in the train set. This technique has the effect of giving greater importance to the reference samples that have greater validity and closeness to the test sample. So, the decision is less affected by reference samples which are not very stable in the feature space in comparison with other samples. In other hand, the multiplication of the validity measure on distance based measure can overcome the weakness of any distance based weights which have many problems in the case of outliers. So, the proposed MKNN algorithm is significantly stronger than the traditional KNN method which is based just on distance.

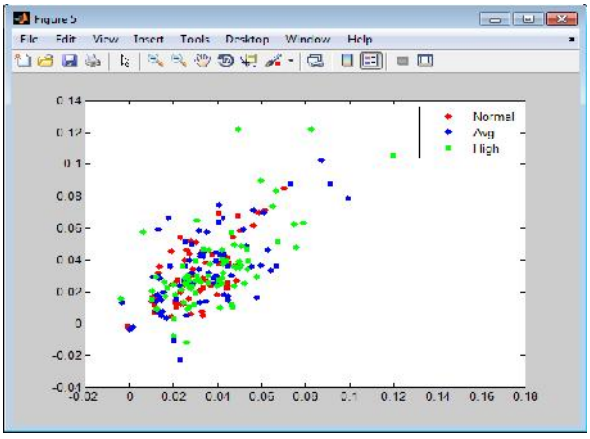

Figure: 2 Data classified maximum range

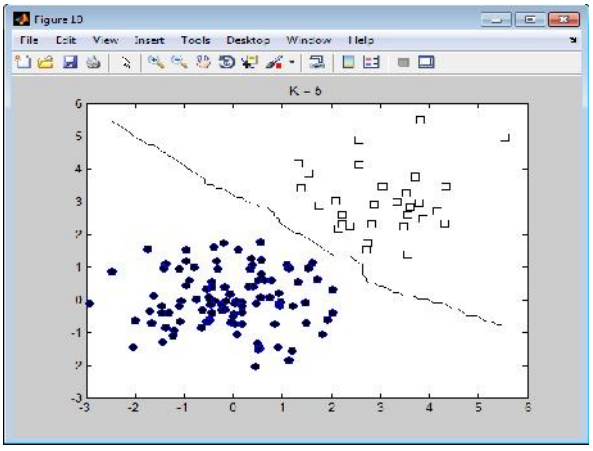

Figure: 3 Data Bi Classification

CONCLUSION

In this review, we discussed the concepts of micro gene expression mining while we outlined their application. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain. In this paper, an approach of the Particle swarm optimization, Modified K means and nearest clustering. The input parameters of any algorithms to get the best results are very carefully taken. This algorithmic implemented on the numeric dataset. The execution time depends upon the how large the no. of iterations, population size, correlation ratio. Efficiency increases with the increase in execution time till a certain limit. But one can do work on time management in future to make it more efficient. The usage of other classification algorithms like machine learning will be explored in future. Gene expression data can be considered to achieve better result in cancer detection and development. Moreover, an efficient algorithm can be developed in order to classify different types of cancer genes from huge amount of gene expression data.

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].Melin, P., & Castillo, O. (2013). A review on the applications of type-2 fuzzy logic in Classification and pattern recognition. Expert Systems with Applications, 40(13),5413–5423.
[12].Melin, P., & Castillo, O. (2014). A review on type-2 fuzzy logic applications in clustering, classification and pattern recognition. Applied Soft Computing, 21,568–577.
[13].Melin, P., Olivas, F., Castillo, O., Valdez, F., Soria, J., Mario, J., et al. (2013). Optimal
design of fuzzy classification systems using PSO with dynamic parameter adaptation through fuzzy logic. Expert Systems with Applications, 40(8),
[14] Y. Shi and M. Mizumoto, An improvement of neuro-fuzzy learning algorithm for tuning fuzzy rules,Fuzzy Sets and Systems, vol.118, no.2, pp.339-350, 2001.
[15] L. O. Hall and I. B. Ozyurt, Clustering with a genetically optimized approach, IEEE Trans, vol.7,no.3, pp.103-112, 1999.
[16] J. Li, X. Gao and L. Jiao, A new feature weighted fuzzy clustering algorithm, Acta Electronica Sinica, vol.34, no.1, pp.89-92, 2006.