Comparison of Efficient and Rand Index Fitness Function for Clustering Gene Expression Data

P.S. Patheja, Akhilesh A. Waoo, and Ragini Sharma

BIST, Bhopal, India
{Pspatheja,raginishrma}@gmail.com,
akhilesh_waoo@rediffmail.com

Abstract. This paper illustrates a comparative study of Efficient Fitness Function and Rand Index Fitness Function, to show how Efficient Fitness Function can give better results when used to cluster gene expression data. Variance which is the main limitation of Rand Index can be improved with Efficient Fitness Function. The results are evaluated by finding the precision value (i.e. sensitivity and specificity) of the dataset. Genetic Weighted K-Mean Algorithm (GWKMA) which is used here is a hybridization of Weighted K-Mean Algorithm (WKMA) and Genetic Algorithm. WKMA is used to perform optimal partition of data. Genetic Algorithm is then applied to get the best fit gene from clusters through the fitness function, on which genetic operators like selection, crossover and mutation are performed.

Keywords: Genetic Algorithm, Fitness Function, Clustering, Gene Expression Data, Variance.

1 Introduction

Clustering is a technique to bring together data having similar characteristics into a group/cluster, so that no two clusters will have similar data. Clustering methods are useful for data reduction, for developing classification schemes and for suggesting or supporting hypothesis about the structure of the data. A generic description of the clustering objective is to maximize homogeneity within each cluster while maximizing heterogeneity among the different clusters.

In this paper we present an Efficient Fitness Function which is implemented in Genetic Algorithm [9] which helps to find the fittest gene which can be transferred next generation. A comparative study of previously implemented Rand Index Fitness Function and proposed Efficient Fitness Function is done here. Both of the functions are applied on GWKMA for the same gene expression dataset [14]. A dataset is a tabular data which gives information about the expression level of genes. The expression level of each gene is calculated with the help of microarray technology by considering the genes (specified in rows) and by samples (specified in columns). The data which we get is known as Gene Expression data. Each value is known as datum. These values may be real numbers or integers. The clustering is performed on Gene Expression data to group the genes with similar characteristics.

N. Meghanathan et al. (Eds.): CCSIT 2012, Part III, LNICST 86, pp. 160–167, 2012.

To validate our data, synthetic dataset i.e. yeung's et al. dataset is applied. This dataset contains 400 genes and 17 attributes. The reason behind selecting yeung's dataset is that it contains data without repeated measurements and with low noise level. The results are evaluated by finding the precision value (i.e. sensitivity and specificity) of the dataset.

In GWKMA for performing optimal partition of clusters, first Weighted K-Means Algorithm [8] with WKM operator or cost function is used. On these clusters Genetic Algorithm is applied. Genetic algorithm is a heuristic searching algorithm to find fittest gene with the help of the fitness function and performs genetic operations like selection, crossover and mutation on it. Weighted K-Mean Algorithm when merged with Genetic Algorithm not only improves the precision value but also helps the algorithm to converge fast. For our reference, henceforth in this paper we will be refereeing Genetic Weighted K-Mean Algorithm with Rand Index fitness function as GWKMA and Genetic Weighted K-Mean Algorithm with Efficient fitness function as MGWKMA i.e. Modified GWKMA.

2 Related Work

Our purpose is to cluster Gene Expression Data. The Gene Expression Data is formed by calculating the expression level of the Genes using microarray technology. As the gene expression data is a table where rows represents genes and column represents samples, the clustering can be classified as gene based clustering and sample based clustering. Amir Ben-Dor [17] uses appropriate stochastic error model on the input and tries to overcome with the problem of clustering multicondition gene expression pattern, and thus recovers the cluster structure with high probability. K.Y Yeung [15] provides a systematic framework which helps to access the results of various clustering algorithms using one experimental condition. The results analysis is done on simulated Gene Expression data using single link, k-means (with random initialization), CAST and random algorithms. In our paper, the fitness function which we are using is an Efficient Fitness function which is a squared error function used to minimize the variance in the cluster. Erkan Bes.dok [20] has compared various squared error functions as fitness function in improving the camera calibrations. Normalized-Root-Mean-Squared-Error is used by Mohammed Awad [22] as a fitness function to deal with the problem of function approximation from a given set of input/output (I/O) data. The problem consists of analyzing training examples, so that we can predict the output of a model given new inputs, which is solved using Radial Basis Function Neural Networks (RBFNNs) and Genetic Algorithms (GAs). M.K. Deshmukh, C. Balakrishna Moorthy [21] uses squared error function as a fitness function for estimation of wind energy potential at a site using Genetic Algorithm (GA) to Neural Network Mode.

3 Fitness Functions

Fitness biologically can be expressed as a measure of reproductive efficiency of chromosomes. In Genetic Algorithm, fitness acts as some measure of goodness to be maximized. By using fitness function one can state, how fit a specific chromosome is. Better the fitness, more the chances to go to next generation.

3.1 Rand Index Fitness Function

Rand Index [12] is a statistical measure to find similarity between two data clusters. The values of Rand Index can vary from 0 to 1, where the value 0 states that the two clusters do not show any similar characteristics while the value 1 states that the characteristics of two clusters are similar. For understanding the Rand Index Fitness Function equations, let's say that we have a set of n elements E= $\{R_1,\ldots R_n\}$ and two partitions of E to compare P=$\{ p_1,\ldots p_r\}$ and Q=$\{q_1,\ldots q_s\}$. We have the equation, with a, b, c and d as the pairs of elements in E.

$$RI = \frac{a+d}{a+b+c+d}$$

Where,

- a- Pairs of elements that are in same sets in P and same sets in Q.
- b- Pairs of elements that are in different sets in P and different sets in Q.
- c- Pairs of elements that are in same set in P & different sets in Q.
- d- Pairs of elements that are in different set in P & in same sets in Q.

3.2 Efficient Fitness Function

Efficient Fitness Function is basically a Squared Error Function to improve the variance of a cluster. Statically it can be represented as a way to measure the difference between the values implied by an estimator and the true values of the quantity being estimated. This difference occurs due to randomness. We have the equation for Squared error function as

$$E=\sum_{N=1}^{N} \sum_{x \in Cn} ||x-mn||^2$$

Where N being the number of clusters, mn the centre of clusters Cn and x represents the data.

4 Limitations of Rand Index Fitness Function

Rand Index Fitness Function is an existing function which is used in GWKMA for clustering large scale gene expression data. Efficient fitness function is the proposed function for the same GWKMA. Efficient fitness function not only gives precise results but also overcomes the limitations of Rand Index. The limitations of Rand Index Fitness Function are:

4.1 Variance

Variance is the means of measurement for the scattered data in clusters, which specifies to minimize the internal feature variation in a cluster or to maximize the variation between different clusters. The cluster variance needs to be taken care when the data is selected randomly. In this paper we are creating clusters using Weighted K-Mean Algorithm, which creates the cluster randomly. In this case there are chances of scattering of data among various clusters. On these clusters when Rand Index fitness function is applied, it is seen that Rand Index fitness function does not give precise results, it's index is quite low when more data is scattered among the cluster i.e. when inter cluster variance is more, as specified in [4]. To overcome this limitation, we use Efficient Fitness Function instead of Rand Index. As the error is squared in Efficient Fitness Function, it neutralizes the appearance of opposite (polarity) internal features in a cluster and hence is able to improve the variance (by minimizing the internal feature variance in a cluster or maximizing the variance between different clusters).

4.2 Equally Weighted Type I and Type II Errors

As Rand Index is also considered as a measure of percentage of correct decision made by an algorithm, it can be computed with the equation

$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

Where,

- True positive (TP)- Equation with hit
- True negative (TN)- Equation with correct rejection
- False positive (FP)- Equation with false alarm, Type I error
- False negative (FN)- Equation with miss, Type II error

One issue with Rand Index is that in this False Positive and False Negative is equally weighted, which may be undesirable for some clustering algorithms. As False Positive and False Negative represents Type I and Type II errors, it is not desirable to provide equal weights to both of them.

4.3 Fuzzy and Non-fuzzy Partition

Rand Index can be used for both fuzzy and non-fuzzy partitions, but it has been seen that it is properly defined only for the comparison of a fuzzy partition with a non-fuzzy reference partition, which is not the basic requirement of GWKMA.

5 Experimental Setup

For the implementation of Genetic Weighted k-Means Algorithm with the fitness functions, the experimental setup includes both hardware and software requirements. The algorithm is implemented with MATLAB and Simulink software on LINUX platform using Fedora-9 (Sulphur) as operating system with (2.6.25 – 14.fc9.i686) kernel and Red Hat Nash version. The algorithm can be implemented using the MATLAB software in Windows family to avoid the changes in the resulted data from the virus problem we are using MATLAB in Linux platform. The hardware support required for the implementation includes, at least 60 GB of the Hard disk as even though the capacity of MATLAB software is 3.5 GB while installing it will extract or decompress the compressed files and retrieve the capacity of the hard disk about 45 GB. The memory requirement is1GB RAM, Intel Core TM 2 Duo Processor. The gene expression dataset used for implementation is Yeung's synthetic dataset as these dataset contains values without repeated measurements and with low noise levels. For the implementation the dataset is stored in the form of a text file and then used. The Software System Attribute includes portability. The algorithm is written using Matlab, therefore it can be ported to any Operating System running Matlab.

6 Results

The results are evaluated by implementing both Efficient and Rand Index fitness functions on same Genetic Weighted K-mean Algorithm one by one for the same Yeung's synthetic dataset, by varying the size of population (the number of chromosomes used for clustering) and generation (the number of iteration). The results are evaluated on the basis of precision values obtained by calculating the sensitivity and specificity of the algorithm by implementing both fitness functions one by one. The sensitivity is proportion of actual positives which are correctly identified as such, while specificity represents the proportion of negatives which are correctly identified. According to the results obtained by the precision values evaluated for sensitivity and specificity, comparison graphs are plotted to show how the algorithm (MGWKMA) implementing Efficient fitness function gives better results than the algorithm (GWKMA) implementing Rand Index fitness function.

6.1 Comparison Graphs for MGWKMA and GWKMA

Fig.1 Shows a comparison graph between MGWKMA and GWKMA with Population size=10 and Generation =10. It can be clearly seen from the graph that MGWKMA gives good results than GWKMA.

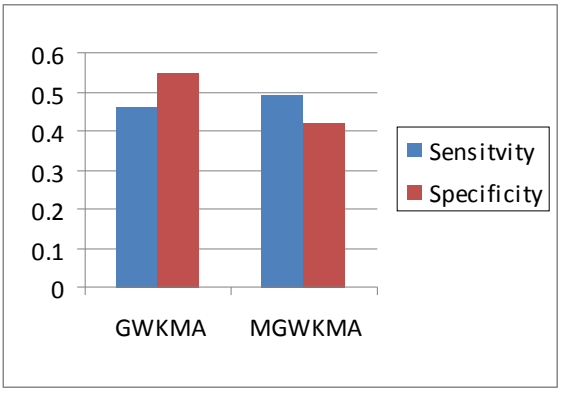

Fig. 1. Comparison graph for Sensitivity & Specificity for P=10, G=10

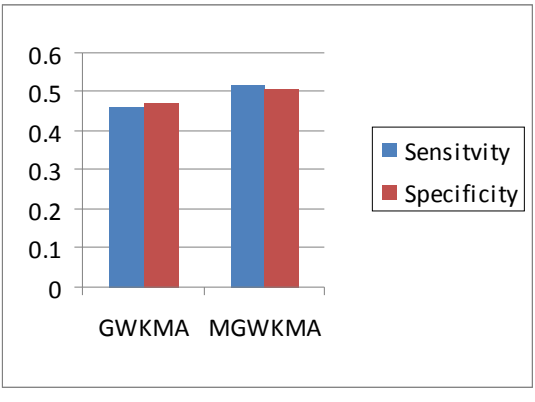

Fig. 2. Comparison graph for Sensitivity & Specificity for P=10, G=20

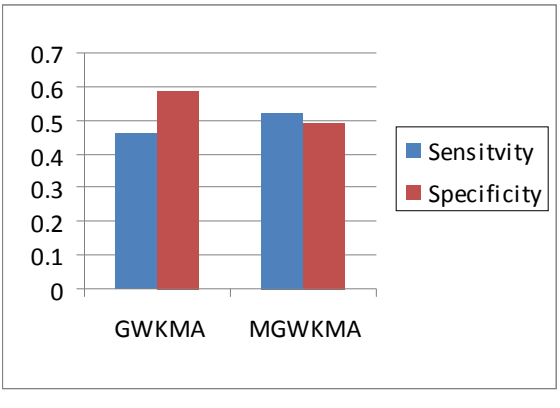

Fig. 3. Comparison graph for Sensitivity & Specificity for P=20, G=10

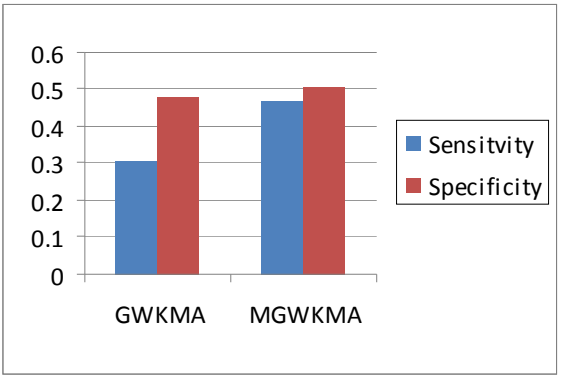

Fig. 4. Comparison graph for Sensitivity & Specificity for P=20, G=20

Fig 2 Shows a comparison graph between MGWKMA and GWKMA with Population size=10 & Generation =20. It is clear from the graph that MGWKMA performs better than GWKMA even though the number of iterations increases.

Fig 3 Shows a comparison graph between MGWKMA and GWKMA with Population size=20 and Generation =10. It is clear from the graph that MGWKMA performs better than GWKMA even though the size of population is increased.

Fig 4 Shows a comparison graph between MGWKMA and GWKMA with Population size=20 and Generation =20. It can be seen from the graph that MGWKMA gives better performance than GWKMA, at high population and generation.

7 Conclusion

In this paper a comparison of Efficient Fitness Function and Rand index fitness function is done for Genetic Weighted K-Means Algorithm used to cluster Gene Expression Data. The dataset used for both is Yeung's Synthetic dataset. An Efficient Fitness Function gives better results in clustering gene expression data than Rand Index fitness function. The limitations of Rand Index Function is learnt and it is shown that Variance (which is a main limitation of Rand Index Function) can be improved using Efficient Fitness Function (as this is a Squared Error Function). A Graphical comparison has been done using Efficient Fitness Function vis-à-vis Rand Index Function in Algorithms for different sets of Population (P) & Generation (G). The comparisons are done by evaluating the precision value by calculating Specificity & Sensitivity. From the comparison it can be concluded that the Algorithm when implemented with Efficient Fitness Function improves the variance in the cluster and gives better results for clustering gene expression data as compared to Genetic Weighted K-Means Algorithm (GWKMA) with Rand Index fitness function.

References

[1] Hartigan, J.: Clustering Algorithms. Wiley, New York (1975)
[2] Obitko, M.: Introduction to Genetic Algorithms (1998)

[3] Krishna, K.K., Murty, M.M.: Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics 29, 1083-4419(99)00770-0
[4] Wehrens, R., Buydens, M.C., Fraley, C., Raftery, A.C.: Model-Based Clustering for Image Segmentation and Large Datasets Via Sampling. Journal Of Classification 21, doi:10.1007/s00357-004-001-8
[5] Maulik, U., Bandyopadhyay, S.: Genetic algorithm-based clustering technique. Pattern Recognition 33, 1455–1456 (2000)
[6] Wu, F.X., Zhang, W.Z., Kusalik, A.J.: A genetic k-means clustering algorithm applied to gene expression data. In: Proceedings of The Sixteenth Canadian Conference on Artificial Intelligence, Halifax, Canada, pp. 520–526 (June 2003)
[7] Kerdprasop, K., Kerdprasop, N., Sattayatham, P.: Weighted K-Means for Density-Biased Clustering
[8] Tho, D.X.: Genetic Algorithms and Application in Examination Scheduling. In: Scholarly Research Paper (2009), doi:10.3239/9783640636723
[9] Srivastava, P.R., Kim, T.H.: Application of Genetic Algorithm in Software Testing. IJSE (2009)
[10] Dudoit, S., Fridlyland, J.: A prediction-based resampling method for estimating the number of clustering in a dataset. BMC Genome Biology 3, research 0036.1- 0036.2 (2002)
[11] Santos, J.M., Embrechts, M.: On the use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification
[12] Jeevanand, E.S., Abdul-Sathar, E.I.: Estimation of residual entropy function for exponential distribution from censored samples. ProbStat Forum (2009) ISSN 0974-3235
[13] Sherlock, G., Boussard, T.H., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., Cherry, J.M.: The Stanford Microarray Database. Nucleic Acids Research 29, 152–155 (2001)
[14] Jiang, D., Tang, C., Zhang, A.: Cluster Analysis for Gene Expression Data: A Survey. IEEE Transactions on Knowledge and Data Engineering 16(11)
[15] Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L.: Model-based clustering and data transformations for gene expression data. Bioinformatics (2001)
[16] Belacel, N., Wang, Q., Cuperlovic-Culf, M.: Clustering Methods for Microarray Gene Expression Data. OMICS 10(4) (2006)
[17] Ben-Dor, A., Shamir, R., Yakhini, Z.: Clustering Gene Expression Patterns. Journal of Computational Biology 6, 281–297
[18] Suresh, R.M., Dinakaran, K., Valarmathie, P.: Model based modified k-means clustering for microarray data. In: International Conference on Information Management and Engineering, vol. 13, pp. 271–273. IEEE (2009)
[19] Sarmah, S., Bhattacharyya, D.K.: An Effective Technique for Clustering Incremental Gene Expression data. IJCSI International Journal of Computer Science Issues 7(3(3)) (2010) ISSN (Online): 1694-0784
[20] Beşdok, E.: 3D Vision by Using Calibration Pattern with Inertial Sensor and RBF Neural Networks Sensors, vol. 9, pp. 4572–4585 (2009), doi: 10.3390/s90604572
[21] Deshmukh, M.K., Moorthy, C.B.: Application Of Genetic Algorithm To Neural Network Model For Estimation Of Wind Power Potential. Journal of Engineering, Science and Management Education 2, 42–48 (2010)
[22] Awad, M.: Optimization RBFNNs Parameters Using Genetic Algorithms: Applied on Function Approximation. International Journal of Computer Science and Security (IJCSS) 4(3)