Cancer Classification on Micro Array Gene Expression Data using Hybrid Feature Selection Techniques Correlation Coefficient and Particle Swarm Optimization

Snehith Reddy Kandadi1, Madala Kristu Raju[2], Siddula Shiva[3], Kesoram Konapaneni[4], Sai Prasanth Grandisiri[5]

ABSTRACT- Detection and confirmation of cancer have been one of the most transpiring clinical applications in micro array gene expression data. However, it remains an extremely difficult job. There are many reasons for this problem to arise. We do not have many samples that can be used for training. On the other side, we have plenty number of genes. In this research paper, we propose a different approach where hybrid feature selection takes place with the help of correlation coefficient and particle swarm optimization. This process of feature selection and classification is performed on multiclass data sets like Lymphoma and SRBCT. Machine learning classifiers are introduced after the process of feature selection is completed. Experimental results conclude that the proposed approach achieves higher accuracy by decreasing the number of levels used on gene expression data. It also uses less features as compared to the old traditional method using classifiers like random forests, decision trees, decision stumps and hybridization.

***Key Words*: Feature Selection, Correlation Coefficient, Particle Swarm Optimization**

INTRODUCTION-

Hybridization method is employed to come up with DNA samples in microarray organic phenomenon data. This can be reviewed in two methods. within the first method, during the hybridization RNA is obtained from samples which are taken from the tissues. We can observe many numbers in one experiment. this can be noted as the prominent advantage of DNA microarray technology. production of the protein helps us detect different styles of memberships. this is often achieved because the organic phenomenon level refers to a specific protein production gene. Microarray datasets are used to solve the problems of cancer classification. Analysis of genes which are different are used to assign to classes which are different This process improves the understanding of basic biological processes within the system. We can review the activity of thousands of genes in a single go. Microarray data analysis is a very useful tool for many purposes like disease treatment and prevention. The utmost purpose of the classification is to create a good model which will identify differentially expressed genes and will even be accustomed identify classes within the unknown samples. a number of the challenges within the microarray data are the littlest number of coaching and testing data available, the upper dimensionality of the info and also the variations that would sneak in experiments performed to estimate the degree of organic phenomenon. the 2 main tasks within the analysis of organic phenomenon microarray are feature Selection and classification. If we want to achieve higher levels of accuracy, we need to use hybrid microarray classification techniques. Microarray organic phenomenon data contains many thousands of genes or feature information. Feature selection is a way of selecting genes that are expressed differently.

Naïve Bayes Theorem:

It is a classification algorithm which is used for binary and multi-class classifications. It is the most efficient algorithm for binary classifications. The name Naïve Bayes comes to this algorithm because, in this algorithm, the probability calculation is made simple. It is very basic which helps us to track back the calculations. Instead of trying to calculate the value of every attribute, it assumes that they are conditionally different. It seems like it does not really work on real data but, the results prove otherwise. The probabilities are stored in a file. They are of two kinds, Class Probabilities - these are the frequency of instances in each class and Conditional Probabilities – these are the frequency of each attribute in a given class.

Fuzzy Classification:

It is a well-known algorithm where it assigns an object with a class label. It is similar to the decision making which is done by a human. It predicts the class label. It is majorly used on information which is imprecise and incomplete and vague data.

METHODOLOGY:

Correlation Coefficient:

A connection based heuristic assessment work is utilized to process the correlation coefficient utilizing the Correlation-based Feature Selection (CFS). It over-comes the inconvenience of univariate channel moves toward that doesn't consider the communication between highlights. The ID capacity of every one of the ascribes is utilized to assess a subset of characteristics. A multivariate methodology is compelling in distinguishing the relationship that exists among the various qualities in the dataset.

Particle Swarm Optimization:

Particle Swarm Optimization (PSO) is a stochastic developmental calculation dependent on a populace, based on satisfactory socio-mental standards to take care of designing issues dependent on a few factors. Multitude Intelligence is implanted in this strategy. It includes the idea of sharing of data that is simulated by bio roused conduct. It permits particles to procure benefits dependent on disclosures and past experience for food.

CONCLUSION:

Microarray knowledge analysis provides valuable results that contribute towards finding organic phenomenon profile issues. One the foremost necessary applications of Microarray knowledge analysis is cancer classification. The classification gets difficult because we do not have many samples that can be used for training. On the other side, we have plenty number of genes. Hence, we used hybrid feature selection on microarray gene expression data. we have conducted this study so that we can compare the algorithms. hybrid feature selection technique with a filter that helps us identify all the genes which are responsible for cancer classification.

FUTURE WORKS:

The challenge of feature selection lies in various domains because of the abundance of data available for research and also the rate at which new data is being generated day by day. Having the right set of features helps in improving the accuracy of the model, reduces the processing time and also reduces the computational power required to carry out the process. The idea of this project can be extended to multiple domains and can be applied to various data sets. In future we strive to bring a more generalized solution to the problem of feature selection rather than being application oriented or requiring prior knowledge about the dataset.

REFERENCES:

1.Mitra, S., Das, R., Hayashi, Y.: Genetic networks and soft computing. IEEE/ACM Trans. Computer Biology. Bio information. 8(1) (2011)
2. Yang, C.-S., Chuang, L.-Y. C.-H., Yang, C.-H.: A hybrid feature selection method for microarray classification. IAENG Int. J. Computer Sci. 21 (2008)
3. Yang, C.-S., Chuang, L.-Y., Yang, C.-H., IG-GA: a hybrid filter/wrapper method for feature selection of microarray data. J. Med. Biol. Eng. 30(1), 23–28
4. Maji, P., Das, C.: Relevant and significant supervised gene clusters for microarray cancer classification. IEEE Trans. Nano Bioscience.