

# A Survey on Classification Techniques in Internet Environment

Akarshika Rawat, Ankita Choubey

Shri Ram Institute of Science and Technology, Jabalpur, Madhya Pradesh, India

## ABSTRACT

With a perspective of conceited dimensionality, gear feeling of qualities calculations we have resort on introduction alternative systems in double dealing to perform viable arrangement in microarray quality expression information sets. However, the full in the midst of face contrasted with the territory of tests makes the designation of interchange computationally hard and inclined to blunders. So there is the need of legitimate element determination with stochastic optimization. In microarray information investigation, measurement decrease is a critical thought in the development of a successful classification calculation in light of the fact that the example size is too huge. Legitimate arrangement can be valuable for identifying genetic markers or biomarkers. Bunching is additionally valuable since it can aggregate qualities in light of their relationship so as to mine significant examples from the quality expression information. Our paper fundamental inspiration is to overview in the bearing for finding the advancement in this course so that example quality order and choice can be moved forward.

**Keywords :** Gene Expression Dataset, Classification, Clustering, Feature Selection

## I. INTRODUCTION

In the pick-up infrequent epoch, DNA microarray technology has mature a vital requisites in genomic tick, and has been introduced an enduring quarters in biology by shifting experimental approaches from single gene studies to genome level analyses . Prior studies shot at shown go off at a tangent microarray gene childbirth information is useful for phenotype batch of cancer diseases[1]. In spite of that, collection abuse gene delivery materials have a sly panhandler in return of the colophon in microarray text routine, which has the surely swaggering dimensionality (large number of genes) with a small number of samples in the data set[2][3][4][5]. And it is definitely foremost but operas to trade name which genes at odds with most to grouping. Extrinsic these challenges, detail choice Techniques have been introduced to feign a small subset of genes as features for classification. Characteristic possibility (gene selection) is ingenious for duo of arguments in the designation of tumor classification shoot up the gene transport data, such as hoop-la classification preciseness, reducing the allege in a clinical setting and gaining significant insight into the mechanism of disease[3][4][5][6].

Microarray reproduce class has been studied extensively good M techniques in machine learning and pattern recognition. Category cog such as weighted voting (WV)[7], k-nearest neighbor (k-NN)[8] , support vector machine (SVM)[9] , and Fisher's linear discriminate analysis (LDA)[10] venture been used for microarray facts category. Putting, these equipment have remote been sprightly for identifying biomarkers, which are substances (genes in the present context) used for detecting whether a patient has got a particular disease or not , . In a microarray authentication, the in the midst of genes attainable is forth greater than that of samples, a well-known problem called the curse of dimensionality. Gene confinement data is useful for phenotype classification of cancer diseases [11]. But, classification using gene delivery data has a first impoverish because of the characteristics in microarray data set, which has the most assuredly high dimensionality (large number of genes) with a small number of samples in the data set. And it is very memorable but grueling to identify which genes contribute most to classification. Exterior this challenge, detail variant techniques has been introduced to select a small subset of genes as features for classification. Prospect surrogate (gene selection) is cutting for brace of thinking in the distribution of tumor classification using the gene expression data, such as

improving classification accuracy, reducing the cost in a clinical setting and gaining significant insight into the mechanism of disease. Association rule can be useful for better prediction [12][13].

We provide here a brief survey on different methodology. Other sections are arranged in the following manner: Section 2 introduces methodology; Section 3 describes about Literature Review; section 4 shows the analysis; Section 5 describes Conclusion.

## II. METHODS AND MATERIAL

In the above regard there are lot of research had been done with several methodology. Here we discuss some of the methodology.

Association Rule mining is one of the important and most popular matter mining techniques. Federation head up mining gluteus Maximum be efficiently used in any decision making processor decision based leadership generation. In data mining appointment in consequently so we courage find the frequent patterns to know the effective patterns from the huge data. Change we find positive and negative rules [14]. If we agree to the beyond everything phenomena change we come to the point that the rule generation is also huge. In this compounding we metaphysical join aspects of optimization techniques by which we can optimize the association rules. So hybridization is needed [15]. Turn to course mining is a efficacious movement capable of identifying in a used of objects (called items) those which demonstrate similar behavior. For event, in a Stock Exchange, clientele object encode are kept as storekeeper business each includes a set of items purchased together. Analyzing the used of merchant may discuss to fact mosey are frequently purchased together. The Ant Colony Optimization algorithm is mainly inspired by the experiments run by Goss et al. which using a grouping of real ants in the real environment. They study and observe the behavior of those real ants and suggest that the real ants were able to select the shortest path between their nest and food resource, in the existence of alternate paths between the two. This ant behavior was first formulated and arranged as Ant System (AS) by Dorigo et al.. Based on the AS algorithm, the Ant Colony Optimization (ACO) algorithm was proposed. In ACO algorithm, the optimization problem can be expressed as a formulated

graph  $G = (C; L)$ , where C is the set of components of the problem, and L is the set of possible connections or transitions among the elements of C.

Genetic algorithms work with a population of the potential solutions. In computing terms, genetic algorithms map strings of numbers to each potential solution. Each solution becomes an individual in the population, and each string becomes a representation of an individual. There should be a way to derive each individual from its string representation. The genetic algorithm then manipulates the most promising strings in its search for an improved solution. This algorithm follows the following cycle [14].

- Creation of a population of strings.
- Evaluation of each string.
- Selection of the best strings.
- Genetic manipulation to create a new population of strings.

Clustering is a nearby which similar records are grouped collectively. In any case this is unabridged to take the do away with user a high level view of what is going on in the database. Clustering is contemporarily old to scrap segmenting - which most marketing people tell, is useful for coming up with a bird's eye vision of the business. K-means clustering is a detail mining/machine savoir faire algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The clustering algorithm is pair of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields. In perpetuity purpose foot be orientation of as animal self-styled by many quality vector in an n dimensional space, n being the number of all features used to describe the objects to cluster.

### Literature Review

In 2001, Amadou Toure et al. explore the use of gene expression data (ged) in discriminating two types of very similar cancers - acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). They explore the role of the feature vector in classification. Each feature vector consists of 6817elementswhich are gene expression data for 6817 genes. They show in this preliminary experiment that learning using neural network is possible when the input vector contains the

correct number of gene expression data. This result is very promising because of the nature of the data (available in large amount and more new information becomes available with better technology and better understanding of the problem).

In 2005, Wai-Ho Au et al. presents an attribute clustering method which is able to group genes based on their interdependence so as to mine meaningful patterns from the gene expression data. It can be used for gene grouping, selection, and classification. According to the authors by clustering attributes, the search dimension of a data mining algorithm is reduced. The reduction of search dimension is especially important to data mining in gene expression data because such data typically consist of a huge number of genes (attributes) and a small number of gene expression profiles (tuples).

According to the authors it is for the aforementioned reasons that gene grouping and selection are important preprocessing steps for many data mining algorithms to be effective when applied to gene expression data. They define the problem of attribute clustering and introduce methodology to solving it. Their proposed method group's interdependent attributes into clusters by optimizing a criterion function derived from an information measure that reflects the interdependence between attributes. By applying their algorithm to gene expression data, meaningful clusters of genes are discovered. The grouping of genes based on attribute interdependence within group helps to capture different aspects of gene association patterns in each group. Significant genes selected from each group then contain useful information for gene expression classification and identification.

In 2007, Gianluca Bontempi suggest that because of high dimensionality, machine learning algorithms typically rely on feature selection techniques in order to perform effective classification in microarray gene expression data sets. They interprets features election as a task of stochastic optimization, where the goal is to select among an exponential number of alternative gene subsets the one expected to return the highest generalization in classification. They propose an original blocking strategy for improving feature selection which aggregates in a paired way the validation outcomes of several learning algorithms to assess a gene subset and compare it to others. The rationale of the approach is

that, by increasing the amount of experimental conditions under which we validate a feature subset, we can lessen the problems related to the scarcity of samples and consequently come up with a better selection. They shows that the blocking strategy significantly improves the performance of a conventional forward selection for a set of 16 publicly available cancer expression data sets. The experiments involve six different classifiers and show that improvements take place independent of the classification algorithm used after the selection step.

In 2007, Li Jiangeng et al. proposed novel hybrid approach that combines gene ranking, heuristic clustering analysis and wrapper method to select marker genes for tumor classification. In their method, they firstly employed gene filtering to select the informative genes; secondly, they extracted asset of prototype genes as the representative of the informative genes by heuristic K-means clustering; finally, employed SVMRFE to find marker genes from the representative genes based on recursive feature elimination. The performance of their method was evaluated by AML/ALL microarray dataset.

In 2010. Yukyee Leung et al. [17] discuss about Filters and wrappers which are two prevailing approaches for gene selection in microarray data analysis. Filters make

Use of statistical properties of each gene to represent its discriminating power between different classes. The computation is fast but the predictions are inaccurate. Wrappers make use of a chosen classifier to select genes by maximizing classification accuracy, but the computation burden is formidable. According to the authors the main drawback of this single-filter-single-wrapper (SFSW) approach is that the classification accuracy is dependent on the choice of specific filter and wrapper. They proposed multiple-filter multiple-wrapper (MFMW) approach that makes use of multiple filters and multiple wrappers to improve the accuracy and robustness of the classification, and to identify potential biomarker genes.

In 2008, Mohd Saberi Mohamad et al. discuss that recent advances in microarray technology allow scientists to measure expression levels of thousands of genes simultaneously in human tissue samples. According to the authors these technology has been

increasingly used in cancer research because of its potential for classification of the tissue samples based only on gene expression levels. A major problem in these microarray data is that the number of genes greatly exceeds the number of tissue samples. Moreover, these data have a noisy nature. So they select a small subset of informative genes that is most relevant for the cancer classification. To achieve this they proposed a hybrid approach.

In 2010, D. G. Stavrakoudis et al proposes the use of a local feature selection scheme, for the effective selection of relevant features, when designing Genetic Fuzzy Rule-Based Classification Systems (GFRBCSs). The method relies in providing the genetic search with deterministic information about the quality of each feature with respect to its classification ability, directing the evolution in selecting the most useful features. To evaluate their method, they propose a learning algorithm that iteratively generates the final fuzzy rule base, extracting one rule at a time, as directed by a boosting algorithm.

In 2010, Tinghua Wang presents an effective feature selection method for support vector machine (SVM). Unlike the traditional combinatorial searching method, features election is translated into the model selection of SVM which has been well studied. This method is to tune the parameters of the Gaussian ARD (Automatic Relevance Determination) kernel via optimization of kernel polarization, and then to rank all features in decreasing order of importance so that more relevant features can be identified. The proposed method is tested on two UCI data sets to demonstrate its effectiveness.

In 2011, Patrick et al. suggest that due to the complexity of the underlying biological processes, gene expression data obtained from DNA microarray. Technologies are typically noisy and have very high dimensionality and these make the mining of such data for gene function prediction very difficult. To tackle these difficulties, they propose to use an incremental fuzzy mining technique called incremental fuzzy mining (IFM). By transforming quantitative expression values into linguistic terms, such as highly or lowly expressed, IFM can effectively capture heterogeneity in expression data for pattern discovery. It does so using a fuzzy measure to determine if interesting association patterns exist between the linguistic gene expression levels. Based on these patterns, IFM can make accurate gene function

predictions and these predictions can be made in such a way that each gene can be allowed to belong to more than one functional class with different degrees of membership. Gene function prediction problem can be formulated both as classification and clustering problems, and IFM can be used either as a classification technique or together with existing clustering algorithms to improve the cluster groupings discovered for greater prediction accuracies.

In 2011, Debahuti Mishra et al. [12] compared the results of two approaches for selecting biomarkers from Leukemia dataset. The first approach for feature selection is by implementing-means clustering and signal-to-noise ratio (SNR) method for Gene ranking, the top scored genes from each cluster is selected and given to the classifiers. The second approach uses signal to noise ratio ranking only for feature selection. For validation of both the approaches, they have used k nearest neighbor (in), support vector machine (SVM), probabilistic Neural Network (PNN) and Feed Forward Neural Network (funny). After comparing the final results of two approaches they have got 100%, 96% and 96% accuracy with SVM, KNN and PNN respectively in first approach with five numbers of genes. Whereas, performance of FNN is 2.17 with 10 numbers of genes. In second approach we have got 96%, 96% and 62% accuracies for SVM, KNN and PNN respectively for 5 numbers of genes and the performance of FNN is 2.52 for 10 genes.

In 2012, PradiptaMaji et al. [13] suggest that one of the major tasks with gene expression data is to find co-regulated gene groups whose collective expression is strongly associated with sample categories. In this regard they proposed a gene clustering algorithm to group genes from microarray data. It directly incorporates the information of sample categories in the grouping process for finding groups of co-regulated genes with strong association to the sample Categories, yielding a supervised gene clustering algorithm. The average expression of the genes from each cluster acts as its representative. Some significant representatives are taken to form the reduced feature set

To build the classifiers for cancer classification. The mutual information is used to compute both gene-gene redundancy and gene-class relevance. The performance of the proposed method, along with a comparison with

existing methods, is studied on six cancer microarray data sets using the predictive accuracy of naive Bayes classifier, K-nearest neighbor rule, and support vector machine. An important finding is that the proposed algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

In 2012, Nikhil Jain et al. [14] discuss about Association rule mining. They suggest that association rule play important rule in market data analysis and also in medical diagnosis of correlated problem. For the generation of association rule mining various technique are used such as A priori algorithm, FP-growth and tree based algorithm. Some algorithms are wonder performance but generate negative association rule and also suffered from Superiority measure problem. They proposed a multi-objective association rule mining based on genetic algorithm and Euclidean distance formula. In this method we find the near distance of rule set using Euclidean distance formula and generate two class higher class and lower class .the validate of class check by distance weight vector.

In 2012, Preeti Khare et al. [5] discusses that importance of data mining is increasing exponentially since last decade and in recent time where there is very tough competition in the market where the quality of information and information on time play a very crucial role in decision making of policy has attracted a great deal of attention in the information industry and in society as a whole. They use density minimum support so that they reduce the execution time. A frequent superset means it contains more transactions then the minimum support. It utilize the concept that if the item set is not frequent but the superset may be frequent which is consider for the further data mining task. By this approach they can store the transaction on the daily basis, then they provide three different density zone based on the transaction and minimum support which is low (L), Medium (M), High (H). Based on this approach they categorize the item set for pruning. Their approach is based on A priori algorithm but provides better reduction in time because of the prior separation in the data, which is useful for selecting according to the density wise distribution in India.

In 2012, Smruti Rekha Das et al. [6] discusses about Support vector machine (SVM) which has become an

Increasingly popular tool for machine learning tasks involving classification, regression or novelty detection. SVM is able to calculate the maximum margin (separating hyper-plane) between data with and without

The outcome of interest if they are linearly separable. To improve the generalization performance of SVM classifier optimization technique is used. According to the authors Optimization refers to the selection of a best element from some set of available alternatives. Particle swarm optimization (PSO) is a population based stochastic optimization technique where the potential solutions, called particles, fly through the problem space by following the current optimum particles. They used Principal Component Analysis (PCA) for reducing features of breast cancer, lung cancer and heart disease data sets and an empirical comparison of kernel selection using PSO for SVM is used to achieve better performance. This paper focused on SVM trained using linear, polynomial and radial basis function (RBF) kernels and applying PSO to each kernels for each data set to get better accuracy.

In 2012, Jian-Bo Yang et al. [3] proposes a new feature selection method using a mutual information-based criterion that measures the importance of a feature in a backward selection framework. It considers the dependency among many features and uses either one of two well-known probability density function estimation methods when computing the criterion. The proposed approaches compared with existing mutual information-based methods and another sophisticated filter method on many artificial and real world problems. The numerical results show that the proposed method can effectively identify the important features in datasets having dependency among many features and is superior, in almost all cases, to the benchmark methods.

In 2012, Xiao Zhang et al. [8] discuss that several clustering algorithms have been suggested to analyze genome expression data, but fewer solutions have been implemented to guide the design of clustering-based experiments and assess the quality of their outcomes. Cluster validity framework provides insights into the problem of predicting the correct the number of clusters. They present several validation techniques for gene expression data analysis. Normalization and validity aggregation strategies are proposed to improve the prediction about the number of relevant clusters. The

results obtained indicate that this systematic evaluation approach may significantly support genome expression analyses for knowledge discovery applications.

In 2012, ShangGAO Et Al. suggest that genes are encoding regions that form essential building block within the cell and lead to proteins which are achieving various functions. However, some genes may be mutated due to internal or external factors and this is a main cause for various diseases. So author suggested that it is important to identify mutated genes as disease biomarkers. They addresses this problem by introducing comprehensive framework that incorporates the two stages of the process, namely

Feature selection and sample classification. In fact, high dimensionality in terms of the number of genes and small number of samples distinguishes gene expression data as an ideal application for the proposed framework. Reducing the dimensionality is essential to efficiently analysis the samples for effective knowledge discovery. The target is to find the reduction level or compact set of features which once used for knowledge discovery will lead to improved performance and acceptable accuracy. For the first stage, they concentrate on four feature selection techniques, namely chi-square from statistics, frequent pattern mining and clustering from data mining, and community detection from network analysis. The effectiveness of the feature reduction techniques is demonstrated in the second stage by coupling them with classification techniques, namely associative classification, support vector machine and naive Bayesian classifier. The results reported for four cancer datasets demonstrate the applicability and effectiveness of their proposed framework.

### III. RESULTS AND DISCUSSION

#### Problem Formulation and Analysis

Different people with the same type of diseases may have a different set of missing or damaged genes, with differing implications for prognosis and treatment of the disease. So proper classification can be very informative and clinically useful for more detailed analysis. In [4] authors suggest that gene expression data analysis, conventional clustering algorithms often encounter the problem related to the nature of gene expression data, which is normally wide and shallow. In other words,

data sets usually contain a huge number of genes (attributes) and small number of gene expression profiles (tuples). This characteristic of gene expression data often compromises the performance of conventional clustering algorithms. So there is the need of methodology to group attributes that are interdependent or correlated with each other.

Three issues in make a wrapper feature selection algorithm a highly challenging task:

- The search in a high-dimensional space: This is known to be an NP-complete problem.
- The assessment of the quality of a feature set on the basis of a small set of samples: This is made difficult by the high ratio between the dimensionality of the problem and the number of measured samples.
- The choice of the best feature configuration on the basis of uncertain assessments.

Indebted to high-handed gift of microarray data, it is NP hard to find the superb optimal gene subset among all combinations of genes. Notwithstanding unlike heuristic search strategies can be employed, these are still too computationally expensive. Colander methods which were spoken gene ranking methods in gene expression data area, attempt to find predictive subsets of the genes by using a simple criterion computed from the empirical distribution, and the top-genes were selected as a feature subset. The most a lot used gene selection methods are based on statistical tests or information theory to rank the genes. Every time gene is evaluated individually and assigned a score reflecting its correlation with the class according to certain criterion in these gene ranking methods. The gene is independent of any learning methods in ranking gene methods. Profit they have better generalization property and computational efficiency. Be that as it may, there is a problem that these selected genes are often highly correlated. So in [6] they suggested a hybrid approach of clustering and wrapping.

Filters and wrappers are two prevailing approaches for gene selection in microarray data analysis. Filters beg take into consideration of statistical properties of each gene to represent its discriminating power between different classes [7]. The in conformity with is changeless but the predictions are inaccurate. Wrappers feel sorry compliantly by of a picked out classifier to select genes by maximizing classification accuracy, but

the computation burden is formidable. Filters and wrappers attack been attached in in front of studies to maximize the classification accuracy for a chosen classifier with respect to a filtered set of genes. The drawback of this single-filter-single-wrapper (SFSW) approach is that the classification accuracy is dependent on the choice of specific filter and wrapper. As per our analysis we draw the following table:

**Table 1:** Analysis

S.No	Methodology	Length/Size	Result
1	Neural Network[23]	10-300	Approx. 55 %
2	SOM + IFM [31]	Cluter1-Cluster 10	Average 82 %
3	3NN [32]	20	87 %
4	SVM [33] K-NN NB	Breast 1 Data set	87.1 68.3 52.7
5	SVM Naive Bayes Classification	Leukemia	95% 93%
	SVM Naive Bayes Classification	Colon	77 % 81 %
	SVM Naive Bayes Classification	Prostate	90 % 91.2 %

#### IV. CONCLUSION AND FUTURE SUGGESTIONS

In this paper we survey about gene classification and clustering with optimization trends. We discuss different methodology. We also survey related research in the direction of rule optimization and provide the analysis. Based on this we suggest some further suggestions which are as follows:

1. Rule optimization can be applied using ACO or PSO.
2. After applying rule optimization, we can obtain reduced rules which are helpful in the case gene classification so it can improve the efficiency.
3. Rule optimization is also done after partitioning.
4. Hybrid method with some wrapper technique, clustering and classification.

#### V. REFERENCES

- [1] V. Blaschke C., Oliveros, J.C. and Valencia, A., Mining functional information associated with expression arrays , Functional and Integrative Genomics, Springer, Berlin, 2001, 1(4), pp.256-268.
- [2] Golub T R, Slonim D K, Tamayo P, et al, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring , Science, AAAS, New York, 1999, 286(15), pp.531-537.
- [3] Inza, I., Larranaga, P., Blanco, R. and Cerrolaza, A.J., Filter versus wrapper gene approaches in DNA microarray domains , Artificial Intelligence in Medicine, ELSEVIER, Amsterdam, 2004, 31(2), pp.91-103.
- [4] C.H.Ooi and P.Tan, Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, Bioinformatics, Oxford University Press, Oxford, 2003, 19(1), pp. 37-44.
- [5] Li, J., Zhang, C. and Olihara, M., A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression , Bioinformatics, Oxford University Press, Oxford, 2004, 20(15), pp.2429-2437.
- [6] Jaeger, J., Sengupta, R. and Ruzzo, W.L., Improved gene selection for classification of microarrays , Pac. Symp. Biocomput, Hawaii, USA, 2003, pp. 53-64.
- [7] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, vol. 286, no. 5439, pp. 531-537, 1999.
- [8] L. Li, C.R. Weinberg, T.A. Darden, and L.G. Pedersen, Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/KNN Method, Bioinformatics, vol. 17, no. 12, pp. 1131-1142, 2001.
- [9] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, Support Vector Machine Classification and Validation of Cancer Tissue Samples Using

- Microarray Expression Data, *Bioinformatics*, vol. 16, no. 10, pp. 906-914, 2000.
- [10] M.M.Xiong, L.Jin, W.Li, and E.Boerwinkle, Tumor Classification Using Gene Expression Profiles, *Bio techniques*, vol. 29, pp. 1264-1270, 2000.
- [11] JogendraKushwah, Divakar Singh, Classification of Cancer Gene Selection Using Random Forest and Neural Network Based Ensemble Classifier, *International Journal of Advanced Computer Research (IJACR)*, Volume-3 Number-2 Issue-10 June-2013.
- [12] Sachinsohra, NarendraRathod, An Improved Single and Multiple association Approach for Mining Medical Databases, *International Journal of Advanced Computer Research (IJACR)*, Volume 2, Number 2, June 2012.
- [13] Shashank Singh, ManojYadav, Hitesh Gupta, Finding the Chances and Prediction of Cancer through Apriori Algorithm with Transaction Reduction, *International Journal of Advanced Computer Research (IJACR)* Volume 2 Number 2 June 2012.
- [14] Anshuman Singh Sadh, NitinShukla, Association Rules Optimization: A Survey, *International Journal of Advanced Computer Research (IJACR)*, Volume-3 Number-1 Issue-9 March-2013.
- [15] Anshuman Singh Sadh, NitinShukla, Apriori and Ant Colony Optimization of Association Rules, *International Journal of Advanced Computer Research (IJACR)*, Volume-3 Number-2 Issue-10 June-2013.