DR : SARA EL-METWALL

NAME:YOSRA EMAD KHATTAB

AN EFFICIENT CANCER CLASSIFICATION USING MID VALUE K-MEANS AND NAÏVE BAYES.

ABSTRACT:

In the field of bioinformatics, cancer classification is important for cancer cell diagnosis . The main goal is to use machine learning to find the smallest set of genes possible.The proposed approach employs a flexible Naive Bayes classifier at first. However, due to major computational issues, the Naive Bayes is not appropriate for the classification of large datasets. Km-Naive Bayes (Naive Bayes combined with mid k-means clustering) is a quick algorithm designed to speed up both training and prediction of Naive Bayes classifiers by using the cluster centers obtained from the k-means clustering. The new techniques namely weighted Mid K-means-Naive Bayes is Implemented to improve accuracy and reduce misclassification. The algorithm was evaluated with different classifier which were applied on the same database. They classification accuracy has 88.8%

INTRODUCTION:

Cancer is a disease characterised by an uncontrolled, irregular growth that destroys and invades healthy body tissues nearby or elsewhere in the body. it is With the help of feature selection, most challenges to reduce the size of the data inside the microarray expression indeed of convenience in diagnosis of small genes during a specific type of cancer. The most challenging aspect of using this expression data is its high dimensionality, which requires a large number of samples for certain genes. This can be solved by using data preprocessing techniques such as feature extraction or feature selection to reduce the dataset size and improve accuracy. The high-density DNA microarray calculates thousands of genes at the same time, so the organic results are more accurate. phenomenon profiles are used for the cancer classification. This classification is completed so as to predict the sort of cancer person has in order that accurate treatments are often given to the person on time . Naïve Bayes is successfully applied to the cancer identification problems. thanks to its generalization ability and has found success in many applications. This paper presents a replacement technique for cancer Classification Using Weighted k-means Naive Bayesfor predicting cancer cells in living organism by the technique of ANOVA (Analysis Of Variance). Weighted K-means (SWKM) to define weights of varied features so as to urge un-class clusters provided a number of the objects has class labels . Such clustering solutions are often used for classifying unknown objects on cancer classification. from the quality optimization method point of view ELM for classification and Naïve Bayes are equivalent but ELM has ELM for classification tends to achieve better generalisation efficiency than conventional Nave Bayes due to its unique separability function, which is analysed in theory and further checked by simulation results.

RELATED WORKS:

Using organic phenomenon evidence, Chen and Li (2009) proposed a Naive Bayes ensemble for cancer classification. In this research, the author proposes a Naive Bayes (Nave Bayes) ensemble classification system. To begin, the Wilcoxon rank sum test is used to filter out irrelevant genes from the dataset.The training set is then used to train one Nave Bayes, which is then checked by the training set to ensure accurate prediction outcomes. Those samples with a low confidence level or an error prediction result are chosen to mentor the second Nave Bayes, and the second Nave Bayes is also checked once more. Similarly, the third Nave Bayes is derived from samples that could not be correctly identified using the second Nave Bayes. a high level of assurance The ensemble classifier is formed by the three Nave Bayes s. Finally, the ensemble classifier is fed the testing package. Majority voting is often used to determine the final test prediction results. Murat et al. (2009) use artificial neural networks and Naive Bayess to diagnose early prostate cancer. The aim of this research is to create a classifier-based efficient classification expert system for early diagnosis of organs in limitation stages so that informed decisions can be made without surgery.For gene collection, Sepulveda et al. used a possibility measurement. Genes whose appearance concepts are a clear indicator of the category separation, given a training data set has been chosen. Mutual information is used in this paper to choose insightful genes because of its nonlinearity, robustness, scalability, and strong empirical results. The proposed organic phenomenon data classification using Nave Bayes and Mutual Information (MI) has many goals. The first is to select informative genes using mutual information techniques. to train and test a Nave Bayes classifier model with the chosen genes and different kernel settings, as well as to check the model's accuracy Using the standard Leave-One-Out Cross-Validation (LOOCV) process, the established classifier model's generalisation ability was tested. Harbi and Smith (2006) suggested an administered grouping model based on the k implies estimation. Strengthening was reenacted to assess loads for the highlights. The clus ter and thus the classifier were then formed using a weighted Euclidean distance metric. Since simulated annealing is needed, this algorithm is computationally intensive. It took a lot of iteration to get to the desired weights. Yang and Yuan (2009) suggested formal semi-supervised discriminant analysis, which uses the class's secret information structures to quantify intra-class discrepancies within an analogous class. Huang et al. (2011) go on to say that the ELM for classification is an aspect of the qua lity optimization process, and that it is extended to a specific type of "generalised" SLFNs support vector network. It demonstrates that, inside the ELM learning system, The maximum margin of Nave Bayes Nave Bayes is a different approach to deal with guided example arrangement that has been successfully extended to a wide variety of example recognition issues. Nave Bayes could be the best classification algorithm for dealing with high-dimensionality feature spaces accurately and efficiently [15]. Nave Bayes is a statistical approach that produces simple results and a very efficient algorithm that generates a binary classifier. Constructing a hyper-plane separating class members from nonmembers within the input space is a simple way to build a binary classifier. A nonlinear decision function within the input space is used in Nave Bayes to map the information into a better dimensional feature space. separating it into a hyperplane with the highest margin. Last but not least, Nave Bayes solves a simple convex o ptimization problem.

METHODOLOGY :

Using the Nave Bayes classification methodology, this system focuses on cancer prediction. The ANOVA test is used in the Nave Bayes methodology to group the sample amount. Sequential data The Nave Bayes approach solves the problem.by using a previous categorization methodology while consuming, and by providing the highest level of precision. During this time Cancer classification is divided into two parts in this study. All genes in the training data set are first sorted using a scoring methodology, and then genes with high scores are kept. Second, the ability to classify genes is important. Two gene combinations are chosen and put to the test using Naive Bayes is a significantly superior classifier.

Step 1: Gene importance ranking - using methods for change analysis of variance (ANOVA), calculate the significant positioning of each quality. ANOVA (Analysis of Variance) (Analysis Of Variance)ANOVA is a methodology that is frequently used.in knowledge analysis, and to draw interesting conclusions P-values were supported by data. The ANOVA is seen to be reliable since it presupposes that all sample populations are normally distributed with equal variance and that all observations (samples) are independent of one another. The method was chosen as the approach for this paper. ANOVA is a statistical method for comparing two or more groups, and it uses a (sample) to return one p value after each comparison.a considerable amount of value for organisations that aren't the same as yousome others.

Step 2: Finding the minimum gene subset- Finding the base quality subset this progression endeavors to arrange the data set with single quality in the wake of choosing a few top qualities inside the significant positioning rundown each chose quality is given as a contribution to the classifier . When good accuracy isn't achieved, it's necessary to classify the data set using every feasible two-gene combination inside the data set. genes that have been chosen Despite the lack of precision, this is a good start. The technique is done for all three genes. combinations, and so on, until a high level of precision is achieved. 'The' A subsequent classifier is used to check for two-gene mutations.During this research, we used many combinations.

EXPERIMENTAL RESULTS:

MATLAB is used to execute the experiments on various blood cell pictures. On publicly available cancer datasets, the proposed methodology was tested. Cancer datasets were utilised as benchmark datasets to demonstrate the effectiveness of the algorithm. Lymphoma, Leukemia, and SRBCT are the five DNA microarray gene expression data sets. The dataset was divided into two parts: training and testing. In terms of classification accuracy, precision, recall, area under the curve, and execution time, they performed well. It's difficult to pick just a few features. In terms of precision, the The proposed method gives a reliable and high-quality forecast.The findings of the last test are compared to those of the other classifier. This strategy is used to find the smallest gene subsets that can assure a high level of accuracy in the categorization of the complete data set.