Estimating The Surveillance Of DNA Based On Various Diseases Using Classification Algorithm

[1]R. Shalini, [2]Dr. K. Rameshkumar
[1]Research Scholar, Computer Science Department, Masss College of Arts and Science, Kumbakonam ,Tamilnadu, India
[2] Research Supervisor, Computer Science Department, Masss College of Arts and Science, Kumbakonam Tamilnadu, India
[1,2] Affiliated to Bharathidasan University
spprajinbala@yahoo.com
rameshkumark.dr@gmail.com

***Abstract*— Classification algorithm based on rule based classifier by embedding the concept of dynamic threshold value which is executed with minimum time and space. The algorithm is mainly designed to avoid the complication which exists in the multiclass classification which uses the threshold value to predict the class. Many classification algorithms exist in data mining for classifying the medical dataset. But not all the classifier supports multiclass classification with better accuracy. Certain classification algorithm highly support for binary class classification where there are two resultant classes with high accuracy, when it is applied on multiclass classification which has the more than two resultant classes, it will not produce the classification result with best accuracy. Our proposed algorithm is designed to improve the classification performance of binary as well as multiclass classification. The proposed algorithm well suits for medical dataset for classifying the data into multiclass. The proposed algorithm produces better accuracy on various kinds of medical datasets with minimum time and space. The algorithm also supports to classify the other synthetic datasets.**

***Keywords*— Classification algorithm, RBA, KDD, APS, Threshold value.**

I. INTRODUCTION

The Rule Based Algorithms (RBA) produce less classification performance on the existing datasets compared with other classifier. Enhancing of rule based algorithm will improve its classification percentage. Our proposed algorithm introduces the concept of dynamic threshold value which enhances the classification performance of RBA and produced the better accuracy on various kinds of datasets.

Fixing of dynamic threshold value increases the classification percentage of rule based algorithm than the static threshold value. The algorithm also executed with less time and consumes less space. Medical data sets are more sensitive in nature. Predicting and classifying the causes for particular disease is important in medical science for perfect treatment of the patient in the right time. In this research the RBA is used for predicting the classes of medical and other synthetic dataset.

II. RBA WITH DYNAMIC THRESHOLD VALUE

An Threshold value is necessary for RBA for accurate classification of the data. There are two ways of fixing the threshold values static and dynamic. The sensitivity and the specificity are calculated by the rule based classification algorithm and the values are checked against the user fixed threshold value to predict the class. So for the threshold values are fixed as a static one which is constant. The threshold value is a user choice for predicting the class. The support and confidence value of the algorithm compared with the fixed threshold value to predict the class. When a threshold value is fixed statically for a classification, the algorithm concludes the class when it reaches the fixed value or the nearby value. It works well for binary class classification. For the dataset with 'N' number of classes, fixing static threshold value affects the classification percentage. When there are two nearby values to the threshold value there is a complication in prediction of the class. So the alternative method for static is dynamic and is used for fixing the threshold value to avoid complication in multiclass classifier.

The medical data set with multiple classes especially needs dynamic way of fixing the threshold values. Fixing of dynamic threshold value improves the classification performance reasonably and also it reduces the misclassification rate. For sensitive data each time the threshold values get changed for each instance. The RBA with dynamic threshold value is mandatory for predicting the class accurately. The modified RBA algorithm is implemented in two phases; in the first phase the algorithm preprocess the data by using the concept discretization to avoid the data complexity of the dataset. In the second phase the modified rule based classifier with predefined rule and dynamic threshold value is implemented for classification.

III. PREPROCESSING PHASE

The proposed algorithm has a two phases. In the first phase it works as a preprocessing phase for ordering the attributes and the second phase is classifies the data by calculating the dynamic threshold value. The preprocessing of the real world data is prime important for further usage. The data mining methodologies will works on data accurate only when the data is preprocessed. The medical information are collected from various medical sources and are stored which intern used for classification purposes. The medical data which are used to monitoring and analyzing the health condition of the population. The Knowledge Discovery in Databases (KDD) methodology seems to be attractive on analyzing of large databases.

In the KDD process, the preprocessing step (data cleaning and handling of missing values) is paramount since it conditions the quality of the results obtained by data mining procedures and represents about 80% of the whole project time. Real world data are generally incomplete like lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. And it may consist of Noisy: containing errors or outliers and is inconsistent containing discrepancies in codes or names.

A. *Methodology*

In our proposed algorithm the ordering of the features has been done. For uncertain data the numbers of features are unknown, and the numbers of values are increase with time. It is necessary to include new features and its values without affecting the existing original data and classification .The ordering of features and its value makes classification task easier. The total number of features can be calculated and updated easily when there found new unknown features.

There are varieties of techniques used in data mining to fill the missing values in the dataset. Filling the missing values using preprocessing technique used for other kind of corporate dataset is applicable and it doesn't affect the application much even it increases wrong prediction rate and It can be rectify without affecting the client side. But the sensitive dataset like medical dataset will get affected when the missing values are filled with wrong values. So in our research we use Not Applicable (NA) parameter to fill the missing values which is taken as a special value while classification is carried out. The NA value for missing value doesn't has any impact and will not affect the classification.

The NA value is replacing by its original value at any time and is used for classification. After the preprocessing is completed the values of each attributes are arranged in order by using concept discretization. By implementing the methods for replacing the categorical value into numerical simplifies the dataset and reduce the memory space of the dataset and the new values for any attribute can be added without affecting the classification. We incorporate all the preprocessing methodology, the data is made available for the classification phase and the proposed classification algorithm is implemented and the performance is estimated by comparing with the existing models.

B. *Algorithm*

The algorithm steps represented here are used only for converting the multivariate values into standard numerical values, and it arranges the attribute and its values in order to simplify the dataset. The two values were used; one to represent the features and other represent its possible data values. The value which is used to represent the feature gets changed for every new feature by adding a constant value. The outer loop is used to assign the values for each feature in the dataset and the inner loop will assign the constant value for each possible data values for a particular feature and it gets incremented till the algorithm assigns the value for the last possible data value. The value which is used to represent the feature and the possible data value together forms the converted data table.

IF A[i] = 1 to n then for each attribute is assigned a value as A [1] = X and,

A [2] = X + constant value and then the X gets the new value in each step which is assigned as X= X + constant value.

These steps are represented in the steps 2, 3, 11, and 12 represent

The attribute A[1] has the values 1to m, The numerical value 1 to m is assigned for the data value of attribute A[1] which further represented as value j. The A[i] and the j value are concatenated to form the data table as a[i] [j].It can be represented as Aij

In this algorithm the plus operator serves in two purposes, one for incrementing the value and other for concatenation. In step 12 for each attribute Ai, it assigns the fixed value for indicating attribute position and the possible value it gets which further referred as Aij. Those two values are concatenated to form an equivalent numeric value rather adding those two values. But in step 14 the value of X is added with the constant value which is fixed already by the algorithm to indicate the next attribute. In step 10 the addition operator served as an incrementing operator.

C. *Algorithm: Preprocessing*

Step 1: Input: CSV data file
Step 2: Const = increment value
Step 3: X = constvalue
Step 4: Y = 0
Step 5: Ai = categorical value
Step 6: Ak = Numerical value
Step 7: For each j in 1 to ncols
Step 8: For i in 1 to nrows
Step 9: select distinct (Ai)
Step 10: Y=Y+1
Step 11: For each distinct Ai
Step 12: assign Ai=X+Y
Step 13: next i
Step 14: X=X + constvalue
Step 15: next j
Step 16: Replace all Ai
Step 17: Stop

The data table with categorical values will produce the complete details and description of the attributes. By having the table with categorical data the understanding of application area is made simple. Having the converted table it reduces the memory space of the data but understanding of the application area is not possible. But the drastic reduction of memory space of the algorithm it can be compromised.

The algorithm assigns the numerical values for the given categorical value is represented in this section with sample data. There are 15 attribute to be analyzed to predict the heart disease. Amongst "Resting electrographic results" is a one of the attribute and is getting three possible values as Normal , Having_ST_T wave abnormal, Left ventricular hypertrophy, It is the 6th attribute in the Data table. The attributes are assigned any predefined constant value by an algorithm.

When the original data consisting of combination of categorical, numerical and continuous values the numerical value again treated as a categorical and is assigned a uniform numerical value by an algorithm. The attributes with continuous value the range is fixed and the value is assign for the range. In our research the data with most categorical values are taken for analysis. It is not possible to avoid continuous value which are there in the existing dataset are used range values. Table 1 and table 2 shows sample of 10 attributes out of 52 attributes of APS dataset which are randomly chosen and its related values. And the table 3 and table 4 have its corresponding converted numerical values and how it is ordered. In our research the existing medical datasets from the UCI repository are taken and the same criterion is implemented for categorical values.

The converted data table is used for classification and the performance is estimated. It can be observed that by having the converted table one can easily identify the attribute position in the dataset and number of possible values. When new values are added for a particular attribute it can be identified without executing any complex queries. The conversion table is formed for each of the data set for conversion of categorical values. The conversion table will be altered every time when there is a new values or attributes are identified.

It can be observed that by having the converted table the user can easily identify the attribute position in the dataset and number of possible values for each attributes. When new values are identified and are to be added for a particular attribute it can be identified without executing any complex queries. The conversion table is formed for each of the data set with all attributes and its values. This conversion table is used by an algorithm for conversion of categorical values into numeric. The conversion table will be altered every time when there is a new values or attributes are identified for the dataset.

TABLE I
A SAMPLE TABLE WITH CATEGORIES VALUE

F.Age	F.MH	F.DOM	F.CU	F.NOPREPRGOUT
20-30	My periods are irregular	Living with partner	Regular	1 loss

Below 20	My periods are irregular	Living with partner	Regular	1 loss
31-39	I've had chemotherapy which has stopped my periods.	Living with partner	Regular	>3 loss
40-50	I've taken medication which has stopped my periods	Living with partner	Regular	>2 and <3 loss
Above 50	I'm still having regular periods	Living with partner	Irregular	>2 and <3 loss
Above 50	Any other reasons	Living with partner	Birth Control	>2 and <3 loss
31-39	I'm still having regular periods	Living with partner	Regular	1 loss
31-39	I'm still having regular periods	Living with partner	Irregular	1 loss
Below 20	I've had radiation therapy which has stopped my periods.	Living with partner	Regular	>2 and <3 loss
31-39	I've had chemotherapy which has stopped my periods.	Living with partner	Regular	>3 loss
Above 50	Any other reasons	Living with partner	Irregular	1 loss
40-50	I've taken medication which has stopped my periods	Living with partner	Regular	>2 and <3 loss
Above 50	I'm still having regular periods	Living with partner	Regular	1 loss
Above 50	Any other reasons	Living with partner	Birth Control	>3 loss
31-39	I'm still having regular periods	Living with partner	Regular	1 loss

TABLE 2
A SAMPLE TABLE WITH CATEGORIES VALUE

F.LIGST ST.CLMDA	F.INF	F.SYBL	F.CNG ANA	F.UTRSHP
DNA-Found	HIV,Gram negative infections	Early - Primary < 1 Year	No	abnormal-Uterus didelphys
DNA-Found	mumps,measles,chicken pox,	Early - Secondary <1 Year	No	Abnormal-Uterus partial bicornuate
DNA-Found	Negative	Early - Secondary <1 Year	No	Abnormal-Septate completel uterus
DNA-Found	mumps,measles,chicken pox,	Early - Secondary <1 Year	No	Abnormal-Septate completel uterus
DNA-Not-Found	Negative	Later > 1 year	Yes	Normal
DNA-Not-Found	Negative	Negative	Yes	Normal
DNA-Not-Found	HIV,Gram negative infections	Negative	Yes	Normal
DNA-Not-Found	HIV,Gram negative infections	Negative	Yes	Normal
DNA-Found	mumps,measles,chicken pox,	Early - Primary < 1 Year	No	Abnormal-Uterus complete bicornuate
DNA-Found	Negative	Early - Secondary <1 Year	No	Abnormal-Septate completel uterus

TABLE 3
THE SAMPLE CONVERTED TABLE

F.Age	F.MH	F.DOM	F.CU	F.NOPREPRGOUT
1001	2007	3002	4003	5003
1005	2007	3002	4003	5003
1002	2006	3002	4003	5002
1003	2005	3002	4003	5001
1004	2002	3002	4002	5001
1004	2001	3002	4001	5001
1002	2002	3002	4003	5003
1002	2002	3002	4002	5003
1005	2004	3002	4003	5001
1002	2006	3002	4003	5002
1004	2001	3002	4002	5003
1003	2005	3002	4003	5001
1004	2002	3002	4003	5003
1004	2001	3002	4001	5002
1002	2002	3002	4003	5003
1002	2008	3002	4003	5001
1003	2007	3002	4003	5001
1004	2001	3002	4002	5003
1005	2005	3002	4003	5003
1002	2007	3002	4003	5003
1002	2002	3002	4003	5003
1004	2002	3002	4002	5003

TABLE 4
THE SAMPLE CONVERTED TABLE

F.LIGSTST.CLMDA	F.INF	F.SYBL	F.CNGANA	F.UTRSHP
37001	38001	39001	40001	41006
37001	38002	39002	40001	41007
37001	38003	39002	40001	41002
37001	38002	39002	40001	41002
37002	38003	39003	40002	41008
37002	38003	39004	40002	41008
37002	38001	39004	40002	41008
37002	38001	39004	40002	41008
37001	38002	39001	40001	41005
37001	38003	39002	40001	41002
37002	38001	39004	40002	41008
37001	38002	39002	40001	41007
37002	38001	39004	40001	41008
37001	38002	39001	40001	41006

In basic rule based algorithm with static TV the rule is applied on the data set and it checks how many instances satisfy the given rule. Here when we fix the threshold value as a constant when it reaches that value it concludes the class which leads misclassification. In case of sensitive data which uses multiple and complex rules for prediction of multiple classes the TV value can't be fixed as a static one when it uses rule based classification. Every time the TV value gets changed. The rule based algorithm with multiple rules is applied on the test data and percentage of each class is estimated. Each time the percentage achieve for each class gets differ for every instances.

The TV value gets changed for each instance so our proposed algorithm predict the class by using the assessed support count, target value as a factors on Euclidean distance function. When the rule based algorithm with static threshold value is applied on the first instance it might predict the resultant class c1 with p1% which is the maximum support count among the n-1 classes and is the threshold value and other classes on that instance will have less percentage than that. But the algorithm might predict the class c1 with p2% which might be maximum value than p1 as its threshold value in the second instant. Here in both the instance the resultant class is same but the percentage to predict that class got changed.

The proposed algorithm sets the TV and changes its value dynamically each time when it got a new calculated value as the new instance occurs. The new vale might be lesser or greater that the existing value. The TV value is kept unchanged for the new instance only when it got the same value as it previous otherwise the new value obtained by the algorithm is fixed as new TV value. Here the threshold value is determined with respect to the percentage of satisfactory factor of n-1 classes.

The sensitivity is the proportion of correct positive classifications out of the number of true positives. The specificity is total number of correct negative classifications out of the number of true negatives. The concordance is the total number of correct classifications out of the total number of samples the classification algorithms which uses the threshold value is fixed by the user. The constant value is fixed as a threshold value by the user for checking the percentage of sensitivity and specificity calculated by an algorithm.

IV. CONCULSION

Preprocessing of data is mandatory for classification to classify the data with high accuracy. The medical dataset which are collected from the different medical sources has different formats and consisting of lot of missing and irrelevant values. The irrelevant values are removed using preprocessing technique and the data made consistent. The missing values were filled and are used for classification. The preprocessing of APS dataset which has been done manually. The algorithm is used for ordering the attribute and its values are explained in detail.

Researchers developed many rule based algorithms based on the basic rule based algorithm to classify the dataset. But not all the algorithm highly supports the multiclass classification. The proposed rule Based Algorithm with dynamic threshold value is designed to support both binary as well as the multiclass classifications. The prototype and fuzzy rules were used for generating the rules for the classifier. The algorithm is executed in two stages for calculating the threshold value and the assessed support TV to predict the class .Two values are used to accurate prediction of the class and classify the dataset.

REFERENCES

[1] Aneeshkumar.A.S, C.J.Venkateswaran, "Estimating the surveillance of liver disorder using classification algorithms", Int. J. Comput. Applic, vol.57.
[2] *Chen.JJ, Tsai.CT, Young.JF, Kodell.R, "Classification ensembles for unbalanced class sizes in predictive toxicology", SAR and QSAR in Environ Res, 2005*.
[3] Devangi.L, Kotak, Shweta Shukla," Protecting Sensitive Rules Based on Classification in Privacy Preserving Data Mining", International Journal of Engineering Research & Technology , vol.2 Iss.11P.
[4] Devasena, C.Lakshmi,"Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set", Bonfring International Journal of Man Machine Interface, Special Issue Inaugural Special Issue.
[5] Duch.W, **"**Similarity based methods a general framework for classification approximation and association", Control and Cybernetics.