Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

A NOVEL SUBSET SELECTION FOR CLASSIFICATION OF DIABETES DATASET BY ITERATIVE METHODS

D.UDHAYAKUMARAPANDIAN*[1], RM. CHANDRASEKARAN[2] AND A.KUMARAVEL[3]

[1]Research Scholar, Department of Computer Science and Engineering. Annamalai University, Chidambaram-608 002, India,
[2]Professor, Department of Computer Science and Engineering. Annamalai University, Chidambaram-608 002, India,
[3]Professor and Dean, Department of Computer Science and Engineering. Bharath University, Selaiyur, Chennai-600073, India ,

ABSTRACT

Search methods applied to data mining techniques help us to analyze a data set. These methods are used for reducing the size of the search space in order to select the relevant attribute for identification of diabetes .The research community in diabetes is very much depends on practical prediction and classification of diabetes parameters based on qualified dataset. The main intention in this context is to deal with a large data set with high accuracy. For this purpose models are built using weka tool under supervised learning algorithm. It is necessary to reduce the data dimension before constructing the models and thus the search methods for selection of attributes are followed. Those models are to be applied to predict the possible test cases for evaluation

KEY WORDS: **Data mining ,Classification, Diabetes data set, Search Methods , Tree, Meta boost, Bayes**

D.UDHAYAKUMARAPANDIAN
Research Scholar, Department of Computer Science and Engineering.
Annamalai University, Chidambaram-608 002, India,

***Corresponding author**

1.INTRODUCTION

Obtaining accurate model by aggregating the classification models is the trend established in bioinformatics field. Work of this nature is found more, similar to that of author's in [4, 5]. Designing the experiments for discovering the patterns covers variety of configurations for iterative steps in the literature2. In this paper we improve the accuracy of the model for a diabetes dataset by applying the iterative steps as shown in Fig 1. The idea of iterations using various types of learning models including meta classifiers is novel in this case. We describe the dataset attributes in section 2 and in section 3 the descriptions of learning methods. Finally the experimental results and conclusion are given in section 4.

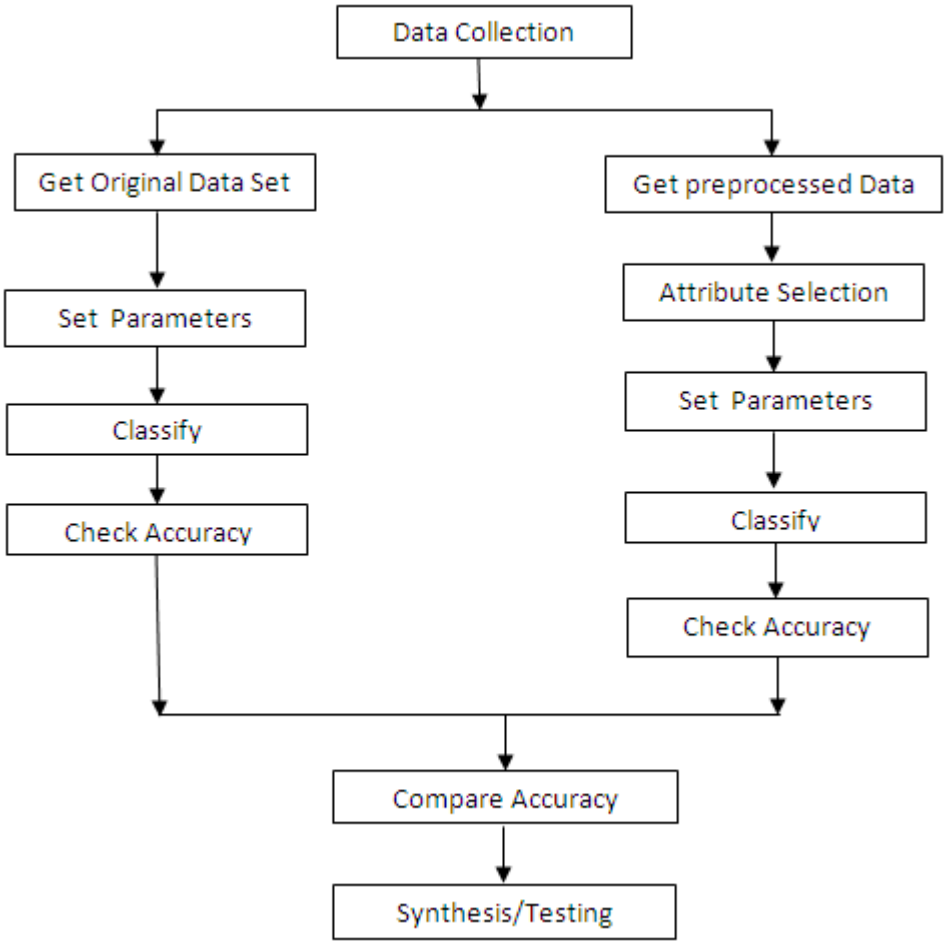

Figure.1
Iterative steps in Data Mining

2. DATA PREPARATION

In this section, we dwell the collection of data and format in which the data has to be presented for mining experiments following the iterative steps in Fig 1.We use java based implementation namely Weka tool from University of Waikato, Newzealand.

2.1 DATASET

The datasets for these experiments are from [8]. The original data format has been slightly modified and extended in order to get relational format.

2.1.1 DATASET DESCRIPTION

The database of diabetes describes a set of eight attributes[11] as shown in the below list 2.2. The class attribute has binary values 'tested negative' and 'tested positive'. The number of instances in this database is 768.

2.2 List of description of attributes.

For each attribute (all numeric-valued), the description and the units are shown:

1. Number of times pregnant
2. Plasma glucose concentration at 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1) ' tested negative' or 'tested positive'

2.3 Brief statistical analysis

Attribute number	Mean	Standard Deviation
1.	3.8	3.4
2.	120	32.0
3.	69.1	19.4
4.	20.5	16.0
5.	79.8	115.2
6.	32.0	7.9
7.	0.5	0.3
8.	33.2	11.8

2.4 Related work in Diabetes Dataset

For the long time the research in diabetes prediction have been conducted. The main objectives are to predict what variables are the causes, at high risk, for diabetes and to provide a preventive action toward individual at increased risk for the disease. Several variables have been reported in literature as important indicators for diabetes prediction. However obtaining the accuracy for recommendation for assisting the physician is a paramount issue. Increased awareness and treatment of diabetes should begin with prevention. Much of the focus has been on the impact and importance of preventive measures on disease occurrence and especially cost savings resulted from such measures. A risk score model is constructed by Lindstrom and Tuomilehto (2003) which includes Age, BMI, waist circumference, history of antihypertensive drug treatment, high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables as categorical variables. A sequential neural network model is obtained by Park and Edington (2001) for indicating risk factors, in the final model, as well as cholesterol, back pain, blood pressure, fatty food, weight index or alcohol index. Concaro et al, (2009) present the application of a data mining technique to a sample of diabetic patients. They consider the clinical variables such as BMI, blood pressure, glycaemia, cholesterol, or cardio-vascular risk in the model

3. Methods Description

Here we select a standard set of methods for predicting from the data set described above. We consider three types of classifiers for our study, such as tree based, Bayes approach based, and Meta level based classifiers. The following sections describe briefly the methods for classifier and results of such methods are tabulated further. Then final results are interpreted

3.1 Tree Classifiers

Supervised Learning is performed conducted using tree classifiers .We select four types of tree classifiers as shown below.

3.1.1 AD Tree

An alternating decision tree (ADTree) is a machine learning method for classification. It generalizes decision trees and has connections to boosting. Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data. Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Boosting is a machine learning meta-algorithm for reducing bias in supervised learning.

3.1.2 Decision StumpOne of the tree classifier is a decision stump, is a machine learning model consisting of a one-level decision tree as described in [3] . That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes. A decision stump makes a prediction based on the value of just a single input feature

3.1.3 J48

This method description is given from the tool descriptor found in [3]The first number is the total number of instances (weight of instances) reaching the leaf. The second number is the number (weight) of those instances that are misclassified. If your data has missing attribute values then you will end up with fractional instances at the leafs. When splitting on an attribute where some of the training instances have missing values, J48 will divide a training instance with a missing value for the split attribute up into fractional parts proportional to the frequencies of the observed non-missing values. This is discussed in the Witten & Frank Data Mining book as well as Ross Quinlan's original publications on C4.5.

3.1.4 Random Forest

One of the ensamble method described in [3] is 'Random forests'which are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost , but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

3.2 Bayes Classifiers

These types of classifiers includes probability measure for the class values and comes under supervised learning.

3.2.1 Naïve Bayes

This belongs to the class implemented in [3] for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an Updateable Classifier you need the Updateable Classifier functionality, use the Naïve Bayes Updateable classifier. The Naïve Bayes Updateable classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

3.2.2 Bayes Net

Bayes Network learning using various search algorithms and quality measures. Base class for a Bayes Network classifier. Provides data structures and facilities common to Bayes Network learning algorithms like K2 and B.

3.2.3 Naive Bayes Simple

This belongs to the class for building and using a multinomial Naive Bayes classifier.This can be described as found in the tool downloaded from [3]. A Bayesian network, Bayes network, belief network, Bayesian model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

3.2.4 Naive BayesUpdateable

Class for a Naive Bayes classifier using estimator classes. This is the updateable version of Naïve Bayes. This classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

3.3 Meta Classifiers

Most of the time, the aggregation of more than one classifier has better performance. Such combinational methods are shown below.

3.3.1 Adaboost

Class for boosting a nominal class classifier using the Adaboost M1 method. Only nominal class problems can be tackled. Often dramatically improves performance, but sometimes over fits.

3.3.2 Bagging

Class for bagging a classifier to reduce variance. Can do classification and regression depending on the base learner. Generate B bootstrap samples of the training data: random sampling with replacement. Train a classifier or a regression function using each bootstrap sample For classification: majority vote on the classification results. For regression: average on the predicted values. Reduces variation. Improves performance for unstable classifiers which vary significantly with small changes in the data set, e.g., CART. Found to improve CART a lot, but not the nearest neighbor classifier.

3.3.3 Dagging

This meta classifier creates a number of disjoint, stratified folds out of the data and feeds each chunk of data to a copy of the supplied base classifier. Predictions are made via averaging, since all the generated base classifiers are put into the Vote meta classifier. Useful for base classifiers that are quadratic or worse in time behavior, regarding number of instances in the training data.

3.3.4 Logit Boost

This classifier is for performing additive logistic regression. This class performs classification using a regression scheme as the base learner, and can handle multi-class problems. This method belongs to the type of meta classifiers.

4. Training and testing with selected classifiers

The above twelve classifiers are considered for learning the diabetes dataset given in 2.1. The ‘a’ tables show the accuracies without selection of attributes, whereas ‘b’ tables show the accuracies with selection of attributes. We use best first search, greedy search for selecting the subset evaluation of attributes using Weka tool and the output shows the reduced attribute set {2,6,7,8}.

a. Results with all attributes

Tree Classifiers	Accuracy
AD Tree	72.9167
Decision Stump	71.875
J48	73.8281
Random Forest	73.8281

b. Results with selected attributes

Tree Classifiers	Accuracy
AD Tree	73.9583
Decision Stump	71.875
J48	74.8698
Random Forest	73.9583

Table 1
Tree Classifier Iterations to get Maximum Accuracy

a. Results with all attributes

Meta Classifiers	Accuracy
Adaboost	74.349
Bagging	74.4792
Dagging	73.5677
Logit Boost	74.0885

b. Results with selected attributes

Meta Classifiers	Accuracy
Adaboost	74.349
Bagging	75
Dagging	73.0469
Logit Boost	74.6094

Table 2
Meta Classifier Iterations to get Maximum Accuracy

a. Results with all attributes

Bayes	Accuracy
Naïve Bayes	76.3021
Bayes Net	74.349
NaveBayes Simple	76.3021
NaiveBayesUpdateable	76.3021

b. Results with selected attributes

Naïve Bayes	Accuracy
Naïve Bayes	77.474
Bayes Net	75.5208
NaveBayes Simple	77.3438
NaiveBayesUpdateable	77.474

Table 3
Bayes Classifier Iterations to get Maximum Accuracy

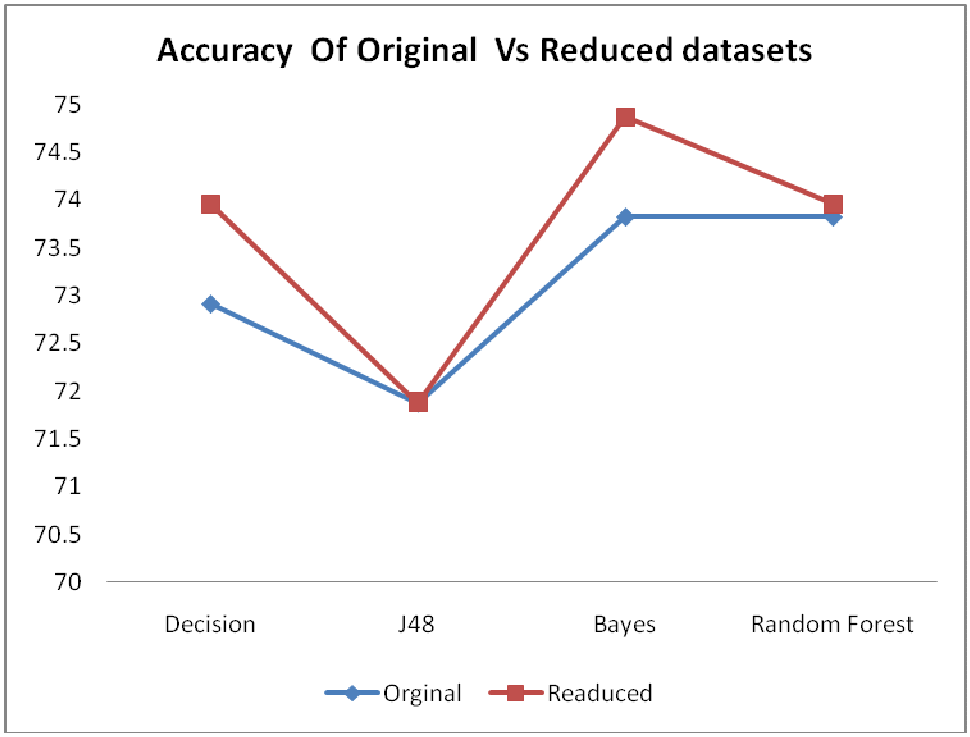

Figure 2
Comparison of original reduced dataset Vs for accuracy

4.1 Results

The above table clearly shows the effect of attribute reduction in the process of classification of given diabetes dataset. The graph in the figure shows the difference in accuracy results.

4.2 CONCLUSION

The experiment carried out clearly shows the improvement of accuracy of the classifiers while reducing the attributes using various searching methods.Diabetes database being the underling data set, the domain expert can use and recommend the reduced set for possible minimal symptoms .The results can be extended to larger data sets with augmented set of attributes.

REFERENCES

1. https://www.waset.org/journals/waset/v68/v68-21.pdf world academy of science, engineering and technology, 2012.
2. H.Dunham, *Data Mining, Introductory and Advanced Topics*, Prentice Hall, 2002First edition **ISBN-13:** 978-0130888921
3. Source about wekahttp://www.cs.waikato.ac.nz/ml/weka/down loaded on 2 feb 2014.
4. A.Kumaravel, Pradeepa.R, Efficient molecule reduction for drug design by intelligent search methods Int J Pharm Bio Sci 2013 Apr; 4(2): (B) 1023 - 1029
5. A.Kumaravel, Udhayakumarapandian.D, Consruction Of Meta Classifiers For Apple Scab Infections , Int J Pharm Bio Sci 2013 Oct; 4(4): (B) 1207 - 1213
6. L. Breiman, " RandomForests,"in*Machine Learning*, vol. 45, pp. 5-32, 2001.
7. Dietterich, T. G., Jain, A., Lathrop, R., Lozano-Perez, T. (1994). A comparison of dynamic reposing and tangent distance for drug activity prediction,Advances in Neural

Information Processing Systems,Vol No 6. San Mateo, CA: Morgan Kaufmann. 216--223.

8. A.Stensvand, T. Amundsen, L. Semb, D.M. Gadoury, and R.C. Seem. 1997. Ascospore release and infection of apple leaves by conidia and ascospores of Venturia inaequalis at low temperatures. Phytopathology 87:1046-1053.
9. Website for attribute description http://archive.ics.uci.edu/ml/machine-learning databases/pima-indians-diabetes,accessed on 10 FEB 2014
10. Bal,HP..Bioinformatics-principles and applications.Tata McGraw-Hill Publishing company Ltd New Delhi.sixth Edition 2008,ISBN:13:978-0-07-058320-7.
11. Bo.Th and Jonassen,I-2002 New feature subset selection procedures for classification of expression profiles.Genome Biology 3:research 00170.-0017.11