International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Extracting Subset of Relevant Features for Breast Cancer to Improve Accuracy of Classifier

Rajesh Saturi, Raju Dara, P.Prem Chand

Abstract: ***Data mining is the essential step which identifies hidden patterns from large repositories. Medical diagnosis became a major area of current research in data mining. Machine learning technique which use statistical methods to enable machine to improve with experiences and identify hidden patterns in data like regression algorithms, clustering algorithms, classification algorithms, neural networks(ANN,CNN,DL),recommender system algorithms, Apriori algorithms, page ranking algorithms, text search and NLP(natural language processing) etc.., but due to lack of evaluation, these algorithms are unsuccessful in finding a better classifier for images to estimate accuracy of classification in medical image processing. Classification is an supervised learning which predicts the future class for an unknown object. The main purpose is to identify an unknown class by consulting with the neighbor class characteristics. Clustering can be known as unsupervised learning as it label the objects based on the scale of similar characteristics without consulting its class label. Main principle of clustering is find the distance like nearby and faraway based on their similarities and dissimilarities and groups the objects and hence can be used to identify outliers (which are far away from from the object).***

Feature extraction, variable selection is a method of obtaining a subset of relevant characteristics from large dataset. Too many features of a class may affect the accuracy of classifier. Therefore, feature extraction technique can be used to eliminate irrelevant attributes and increases the accuracy of classifier. In this paper we performed an induction to increase the accuracy of classifier by applying mining techniques in WEKA tool. Breast Cancer dataset is chosen from learning repository to analyze and an experimental analysis was conducted with WEKA tool using training dataset by applying naïve bayes, bayesnet, and PART, ZeroR, J48 and Random Forest techniques on the Wisconsin's dataset on Breast cancer. Finally presented the best classifier where the accuracy is more.

Keywords: breast cancer, UCI machine learning, classifier, supervised learning, features .

I. INTRODUCTION

Many Databases consists of hidden patterns that can be used for qualitative analysis using some artificial intelligence methods. Classification and prediction are the two techniques can be used to build a model which describes about important classes or predicts, to which class data objects it belongs.

Manuscript published on 30 September 2019.
*Correspondence Author(s)
Rajesh saturi ,Research Scholar, Department of CSE,University College of Engineering,Osmania University, Hyderabad-500007, Telangana State, India,Email id:rajeshsaturi.oucse@gmail.com.
Dr. Raju Dara, Professor,Dept., of Computer Science & Engineering,Vignana Bharathi Institute of Technology,Hyderabad, Telangana .
Prof.P.Prem Chand,Professor, Department of CSE,University College of Engineering,Osmania University, Hyderabad-500007, Telangana State, India.

These techniques can help provide us with a better understanding of large dataset. Classification predicts categorical i.e...Discrete, unordered labels, whereas prediction models continuous valued function. Different methods and technologies have been identified but most of the algorithms are based on memory related to small data size. Supervised learning is the process of finding a model (or function) that describes and distinguishes different classes in two phases

a) **Learning phase** : Training data set is analyzed by a classification algorithm where a classifier is built to describe a predetermined set of concepts by analyzing training data set which consists database objects and their associated class labels. For example an object, Y is represented by an n-dimensional attribute vector, Y = (y1, y2, yn), where each tuple, Y, is assumed to a predefined class and y1, y2, yn are the instances of that class. These data objects/data tuples can be referred to as samples, examples, instances. These individual tuples are training sets, selected from the database for analysis.

b) **Classification phase**: Test dataset is used to estimate the accuracy of classifier by classification rules. These rules can be applied to new data tuples to identify to which class does those objects belongs, if the accuracy of classifier is acceptable, because it predicts the objects to which class it belongs by knowing the class label whereas clustering is performed without consulting the class label hence it is called as unsupervised learning .This can be viewed as a mapping function, X = f (y), that predicts the associated class label of X for a given tuple y.

A researcher wants to know about specific treatment from which one of three, a patient must receive for breast cancer. This can be done by construction a classifier to predict labels such as “treatment X,” “treatment Y,” or “treatment Z” should be given to breast cancer patient. These categorical can also be represented as discrete values such as 1 for “treatment X,” 2 for “treatment Y,” and 3 for “treatment Z” The derived model is represented in the form of classification (IF-THEN) rules, or decision trees. Decision tree is a tree like structure; where each node performs a test on an attribute and each branch represents an outcome of the test, leafs represents class distributions. These trees can be easily converted to classification rules.

Fig 1: decision tree

II. FEATURE SELECTION

Feature extraction plays a key role in classification and clustering analysis. Feature Selection is done by reducing the number of attributes for a selected class from dataset by using different techniques. This reduced data set can improve accuracy of a classifier compared with original dataset. Different areas of feature selection include image processing, biomedical, text classification, and system monitoring etc..,. Feature selection methods are classified as filter, wrapper, and hybrid. The outcome of these methods differs in time and accuracy. In medical diagnosis accuracy is very essential hence one can integrate preprocessing techniques in classification tasks to achieve the target.

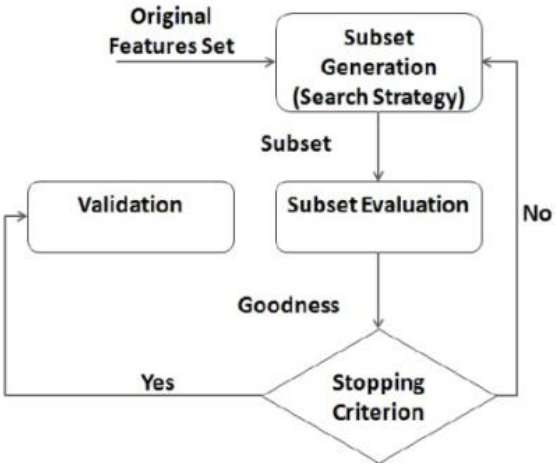


Fig:2 Feature selection process

III. PROPOSED METHOD

An experiment has been conducted with data mining tool using classification algorithms on the Wisconsin's dataset on Breast cancer. The architecture of proposed method is illustrated in the Fig3. Initially, a dataset is loaded in to a tool which has to be pre-processed and then develop a rough-set based feature selection algorithm which can be applied on dataset with full features to obtain dataset with selected features. The dataset obtained with selected features are split into training and testing dataset. Then a model is built by the training dataset using supervised algorithms. Finally new incoming data can be tested with newly build classifier to obtain the accuracy of the classifier.

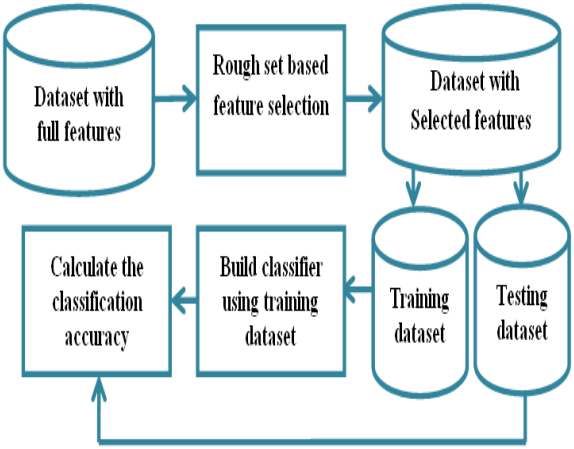


Fig 3: Classification

Step 1: Find and separate unique rows from selected dataset

Step 2: Each instance is compared with all other instances to find similar characteristics.

Step 3 concatenated comparability matrix is constructed.

Step 4 Get the lower and upper bound of the dataset.

Lower Bound=Union(X & Ind (Data Set): X is subset of Y)

Where Y is the given data set, Ind is the indiscernibility relation

Ind (Data Set) = {{1}, {2, 5}, {3}, {4}, {6}}

Y Yes= {1, 2, 3, 6}

Y No= {4, 5)

Lower Bound= {1, 3, 6}

Upper Bound= Union(X & Ind (Data Set): X intersection Y! =empty)

Upper Bound= {1, 2, 3, 5, 6}

Difference between lower and upper bounds :{ 2, 5}

Hence, it is a rough set because the difference set is non-empty

Step 5 The sum of occurrence of each feature in comparability matrix.

Where, 1's occurrence=6

2's occurrence=4

3's occurrence=9

Step 6 finds the most important relevant features.

Step 7 obtains new dataset with major features and finds accuracy of classification.

IV. IMPLEMENTATION

The main purpose of work is to build a new model based on the selection of attributes such that the new built in model can be used to classify new incoming data accurately. Classification has been broadly studied in the areas of neural networks, machine learning, statistics, and expert systems over decades. Classification methods include:

- Bayesian classification algorithms
- Random forest
- Decision tree algorithms
- NN(Neural networks)
- SVM(Support vector machines)
- Genetic Algorithms
- Rule based algorithms etc...

Retrieval Number: K15070981119/19©BEIESP
DOI: 10.35940/ijitee.K1507.0981119
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Breast Cancer Diagnosis Dataset

1. Title: Breast Cancer Database

2. Source: Weka tool

3. Previous study:

From attribute 2nd through 10th are used to represent instances. Each instance has possibility of two classes either **benign or malignant**. Two pairs of parallel hyperplanes were found to be consistent with 50% of the data

Attribute	Domain
1.Sample code number	id number
2.Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 – 10
8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitoses	1 – 10
11. Class:	(2 for benign, 4 for malignant)

- Accuracy on remaining dataset is: 93.5%.
- Accuracy on 33% of dataset is : 95.9%

4. Dataset Information:

Number of Instances/tuples: 286

Number of Attributes/domain: 10

5. Missing attribute values: 16

Node caps-3% and breast-quad-1%,

Value of missing attributes is denoted by "?".

6. Class distribution:

- Benign: 457 (65%)
- Malignant: 242 (35%)

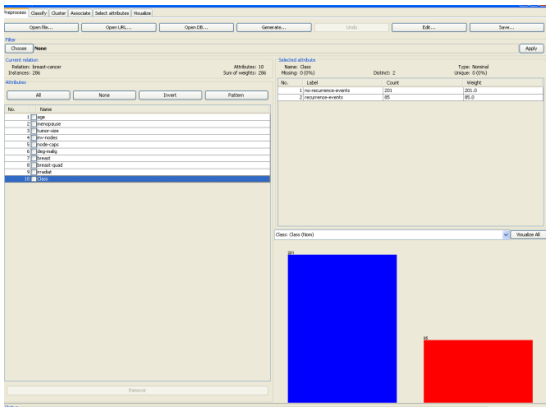
Total 10 attributes: *9 features+1class attribute*

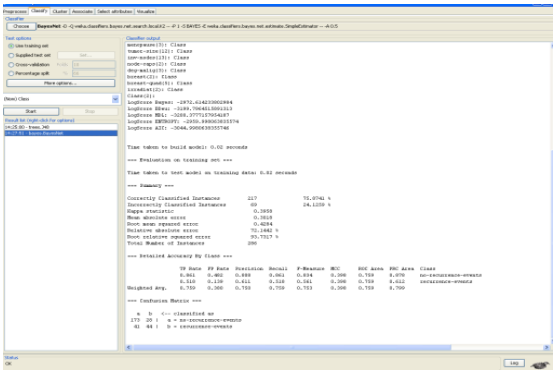
TEST CASES:

Confusion matrix is also called as a matching matrix which can be used to analyze the performance of an algorithm. Column in the matrix represents the instances of predicted class and row represents the instances of actual class vice versa.

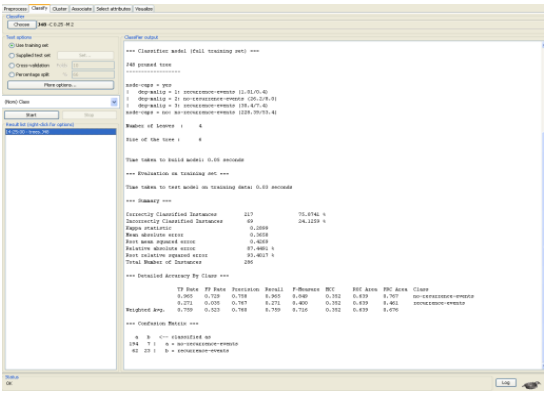
S.NO		Correctly Classified Instances	%	Incorrectly Classified Instances	%	Testing time Build
Bayes Network	Use Training set	217	75.8741 %	69	24.1259 %	0.01 seconds
	Cross validation-Folds-10	206	72.028 %	80	27.972 %	0.00 seconds
	Folds-5	206	72.028 %	80	27.972 %	0.03 seconds
	Folds-2	202	70.6294 %	84	29.3706 %	0 seconds
Naive Bayes	Use Training set	215	75.1748 %	71	24.8252 %	0.01 seconds
	Cross validation-Folds-10	205	71.6783 %	81	28.3217 %	0 seconds
	Folds-5	208	72.7273 %	78	27.2727 %	0 seconds
	Folds-2	207	72.3776 %	79	27.6224 %	0 seconds

Use training set	a	b
	No-recurrence-events	Recurrence-events
NAIVE BAYES	174	27
	44	41
BAYESNET	173	28
	41	44
PART	184	17
	40	45
ZEROR	201	0
	85	0
J48	194	7
	62	23
RANDOM FOREST	198	3
	3	82





Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)



Ranked attributes:

1. Age
2. Menopause
3. Tumor-size
4. Inv-nodes
5. Node-caps
6. Deg-malign
7. Breast
8. Breast-quad
9. irradiat

Selected attributes: 6,4,3,5,9,1,8,7,2 : 9

S.NO		Correctly Classified Instances	%	Incorrectly Classified Instances	%
One R	Use Training set	208	72.7273 %	78	27.2727 %
	Cross validation - Folds-10	188	65.7343 %	98	34.2657 %
	Folds-5	192	67.1329 %	94	32.8671 %
	Folds-2	201	70.2797 %	85	29.7203 %
PART	Use Training set	229	80.0699 %	57	19.9301 %
	Cross validation - Folds-10	204	71.3287 %	82	28.6713 %
	Folds-5	203	70.979 %	83	29.021 %
	Folds-2	186	65.035 %	100	34.965 %
Zero R	Use Training set	201	70.2797 %	85	29.7203 %
	Cross validation - Folds-10	201	70.2797 %	85	29.7203 %
	Folds-5	201	70.2797 %	85	29.7203 %
	Folds-2	201	70.2797 %	85	29.7203 %
J48		217	75.8741 %	69	24.1259 %

V. CONCLUSION

The feature extraction is a method used to obtain subset of similar characteristics from large dataset. Several characteristics of a class may affect the accuracy of classifier. Hence, feature extraction is used to eliminate irrelevant attributes and increase the accuracy of classifier. In this paper performed an induction, and observed the accuracy of classifiers. Breast Cancer dataset is considered from WEKA tool. In future an experiment should be conducted with data mining tool using algorithms such as, k-means clustering, Hierarchical clustering and Density based clustering.

REFERENCES

1. E-Sebakhy et al.., "Evaluation of breast cancer tumor classification with unconstrained functional networks classifier," Computer Systems and Applications, IEEE International Conference, 2006, pp. 281 – 287.
2. Wang, C.M., et al., "Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data". Expert Syst. Appl. 36, 5900–5908 (2009).
3. Howley, et al., "The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data". Knowl.-Based Syst. 19, 363–370 (2006).
4. Guldogan, et al., "Feature selection for content-based image retrieval". Signal, Image and Video Processing, 241 250 (2008).
5. Rodriguez, et al., "Quadratic Programming Feature Selection". Journal of Machine Learning Research 11, 1491–1516 (2010).
6. Wu, O., Zuo, H., Zhu, e al., "Rank aggregation based text feature selection". Web Intelligence, pp. 165–172 (2009).
7. Bouaguel, et al., "An improvement direction for filter selection techniques using information theory measures and quadratic optimization". International Journal of Advanced Research in Artificial Intelligence 1, 7–11 (2012).
8. J. Jossinet (1996) "Variability of impedivity in normal and pathological breast tissue". Med. & Biol. Eng. & Comput, 34: 346-350.
9. JE Silva, et al., "Classification of Breast Tissue by Electrical Impedance Spectroscopy". Med & Bio Eng & Computing, 38:26-30.

ACKNOWLEDGEMENT

Dr. Raju Dara is a professor of Computer Science and Engineering Department at Vignana Bharathi Institute of Technology, Hyderabad. He has 16 years of teaching experience for Graduate and Post Graduate engineering courses. His current research interests are Data Warehousing, Image Processing, and Network Security. He published 30 research papers in international journals as well as international conferences. I will be grateful to the FIST laboratories for providing required computational facilities at the institution, eventually, I extend the deep sense of gratitude to the management, principal, director of R&D of Vignana Bharathi Institute of Technology for supporting in accomplishment of this paper.

Rajesh Saturi, pursuing PhD in computer science &engineering from university college of engineering, osmania university under the esteemed guidance of Prof.P.Prem chand. Presented several papers in spinger,IEEE, national and international journals & conferences.

Prof.Prem Chand is a senior professor in department of computer science & engineering, Active Member of AICTE-NBA, Selection Committee Member at Jawaharlal Nehru Technological University, Andhra University, Acharya Nagarjuna University, Kakatiya University, Indian Space Research Organization. Actively involved in research, supervised 4 research students for the award of their PhD and many more students are pursuing their PhD at O.U, J.N.T.U and A.N.U. Has presented several papers in National and International Conferences.