The Importance of Feature Selection in Classification

Mrs.K. Moni Sushma Deep
Department of Information Technology
Anil Neerukonda Institute of Technology and Sciences
Visakhapatnam, Andhra Pradesh

Mr. P.Srinivasu
Department of Computer Science
Anil Neerukonda Institute of Technology and Science
Visakhapatnam, Andhra Pradesh

email: monisushmakavila@gmail com.email:ursrinivasu@gmail.com

Abstract: Feature Selection is an important technique for classification for reducing the dimensionality of feature space and it removes redundant, irrelevant, or noisy data. In this paper the feature are selected based on the ranking methods.(1) Information Gain (IG) attribute evaluation, (2) Gain Ratio (GR) attribute evaluation, (3) Symmetrical Uncertainty (SU) attribute evaluation.

This paper evaluates the features which are derived from the 3 methods using supervised learning algorithms K-Nearest Neighbor and Naïve Bayes. The measures used for the classifier are True Positive, False Positive, Accuracy and they compared between the algorithm for experimental results. we have taken 2 data sets Pima and Wine from UCI Repository database.

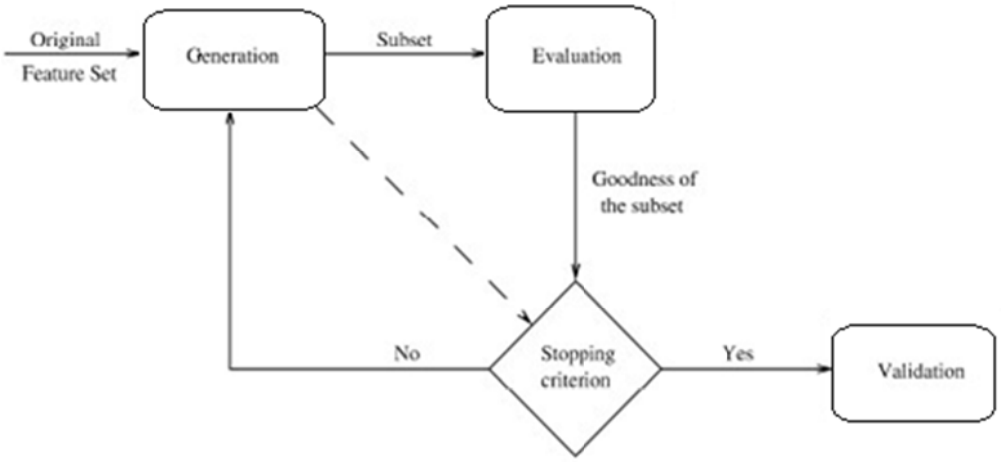
Keywords: Feature Selection, Naïve Bayes, K-Nearest Neighbor and Classification Accuracy

1. INTRODUCTION

Classification is a processes of grouping objects based on some criteria. Feature selection is an important technique to get better accuracy from the classification. Classification is a data mining technique that assigns objects in a collection to target categories or classes. There are many classification algorithms, but we have taken Naïve Bayes and KNN algorithms for evaluation.

1.1 FEATURE SELECTION:

Feature Selection is also useful as part of the data analysis process, as shown in which features are important for prediction and how these features are related. In which algorithm can be seen as the combination of a searching techniques for the proposed of new feature subsets, along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate.[12].



Feature selection Processes with validation

The feature ranking and feature selection techniques have been proposed in the machine learning literature.The purpose of these techniques is to discard irrelevant or redundant features from a given feature vector.[4,1].In this paper, we consider evaluation of the practical usefulness of the following ranking methods:

- Information Gain (IG) attribute evaluation,
- Gain Ration (GR) attribute evaluation,
- Symmetrical Uncertainty (SU) attribute evaluation.

[1]Entropy is a commonly used in the information theory measure, which characterizes the purity of an arbitrary collection of examples. It is in the foundation of the IG, GR and SU attribute ranking methods. The entropy measure is considered as a measure of system's unpredictability. The entropy of Y is.

$$H(Y) = -\sum_{y \epsilon Y} P(y) \log_2 (P(y)) \qquad (1)$$

Where p(y) is the marginal probability density function for the random variable Y. If the observed values of Y in the training data set S are partitioned according to the values of a second feature X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y prior to partitioning, then there is a relationship between features Y and X. Then the entropy of Y after observing X is:

$$H\left(\frac{Y}{X}\right) = -\sum_{x \epsilon X} P(x) \sum_{y \epsilon Y} P(\frac{y}{x}) \log_2 \left(P\left(\frac{y}{x}\right)\right) \qquad (2)$$

where p(y/x) is the conditional probability of y given x.

A. Information Gain

Given the entropy as a criterion of impurity in a training set S, we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases. This measure is known as IG. It is given by

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \qquad (3)$$

B. Gain Ration:

The Gain Ratio is the non-symmetrical measure that is introduced to compensate for the bias of the IG. GR is given by

$$GR = \frac{IG}{H(X)} \qquad (4)$$

C. Symmetrical Uncertainty:

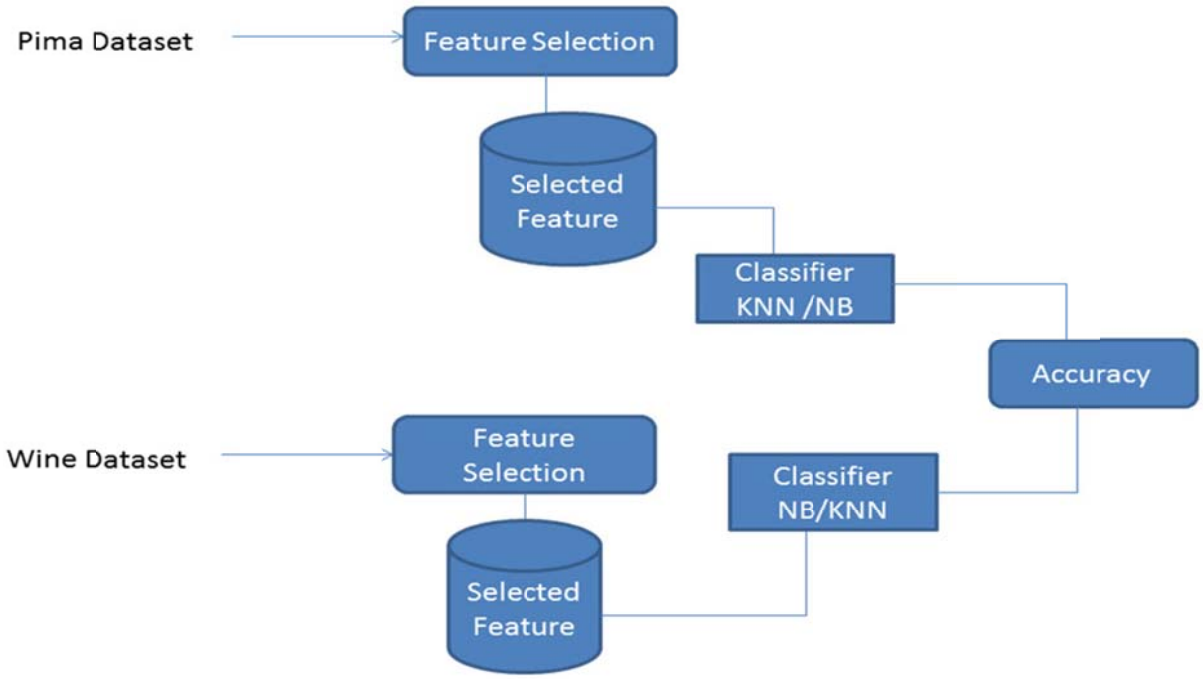
The Symmetrical Uncertainty criterion compensates for the internet bias for IG by dividing it by the sum of the entropies of X and Y. It is given by

$$SU = 2 \frac{IG}{H(Y) + H(X)} \qquad (5)$$

1.2 Classification

Classificationis a data mining function that assigns objects in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. This task begins with a data set in which the class assignments are identified.Classification models are tested by comparing the predicted values to known target values in a set of test data. The goal of the predictive models is to construct a model by using the results of the known data and is to predict the results of unknown data sets by using the constructed mode.

Design of classifier architecture:



Architecture for the Classifier Design

Classification Evaluation:

This evaluation can be classified into different classification. They are

1) Naïve Bayes
2) K-nearest neighbor algorithm.

2. NAÏVE BAYES ALGORITHM:

A supervised algorithm is adopted here to build model using naïve Bayes. This section gives a brief overview of this algorithm.This classifier is based on the Bayes theorem. It can achieve relatively good performance on classification tasks. Naïve Bayes classifier greatly simplifies learning by assuming that features are independent given the class variable. In simple terms, a naïve Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature. In spite of their naïve design and apparently over simplified assumptions, naïve bayes classifiers have worked quite well in many complex real world situations. An advantage of the naïve bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

More formally, this classifier is defined by discriminate function.

$$fi(X) = N\prod j = 1\ P(xj|ci)P(ci)$$

where $X=(x_1,x_2,\ldots..x_N)$ denotes a feature vector and c_j, j=1,2,…..N, denote possible class labels.

3. K- NEAREST NEIGBOUR ALGORITHM:

Nearest neighbor classifies are based on learning by analogy. The training samples are described by n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample."Closeness" is defined in terms of Euclidean distance, where the Euclidean distance between two points, X=(x1,x2,….xn) and Y=(y1,y2,….yn) is

$$d(X,Y) = \sqrt{\Sigma^{n}_{i=1}\ (x_i-y_i)^2}$$

Nearest neighbor classifiers are instance based-based or lazy learners in that they store all of the training samples and do not build a classifier until a new samples needs to be classified. Therefore, they require efficient

indexing techniques. As expected, lazy learning methods are faster at training than eager methods, but slower at classification since all computation is delayed to that time. Unlike decision tree induction and backpropagation, nearest neighbor classifiers can also be used for prediction, that is to return a real-value prediction for a given unknown sample. In this case the classifier returns the average value of the real-valued labels associated with the k nearest neighbors of the unknown sample.

4. DATA SET USED

These datasets are taken by using the UCI Repository. We are taken 2 datasets like Pima and Wine datasets from the Repository.

4.1 PIMA DATASET

To train up the network we used PIMA dataset that contain 768 records and 8 attributes and one class variable.

This data set collects information from patients who are all females over 21-year old of Pima Indian heritage.

The attributes are:

1. Number of times pregnant
2. Plasma glucose concentration a 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)

From this we taken 507($\frac{2}{3}^{rd}$ of preprocessed dataset) are used for training and rest 254 (remaining $\frac{1}{3}^{rd}$ of preprocessed dataset) are tested. The preprocessed dataset contain total of 7 attributes which includes 6 features and 1 class attribute. There 6 attributes are fed as inputs to the input layer.

4.2 WINE DATASET:

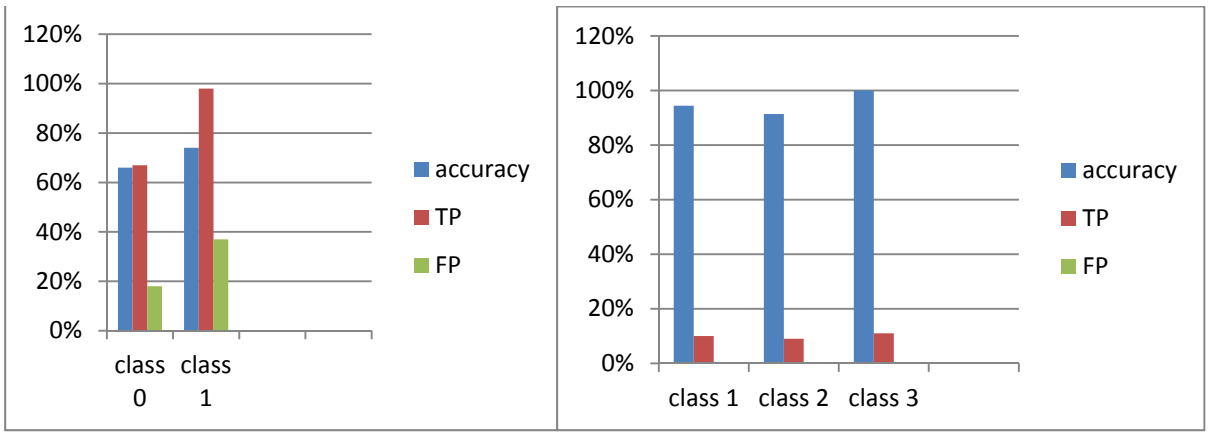
To train up the network we used WINE dataset that contain 178 records and 13 attributes and one class variable.

1. Class {1,2,3}
2. Alcohol REAL
3. Malic_acid REAL
4. Ash REAL
5. Alcalinity_of_ash REAL
6. Magnesium INTEGER
7. Total_phenols REAL
8. Flavanoids REAL
9. Nonflavanoid_phenols REAL
10. Proanthocyanins REAL
11. Color_intensity REAL
12. Hue REAL
13. OD280/OD315_of_diluted_wines REAL
14. INTEGER

In this we are taken 118($\frac{2}{3}^{rd}$ of preprocessed dataset) are used for training and rest 59 (remaining $\frac{1}{3}^{rd}$ of preprocessed dataset) are tested. The preprocessed dataset contain total of 14 attributes which includes 13 features and 1 class attribute. There 6 attributes are fed as inputs to the input layer.

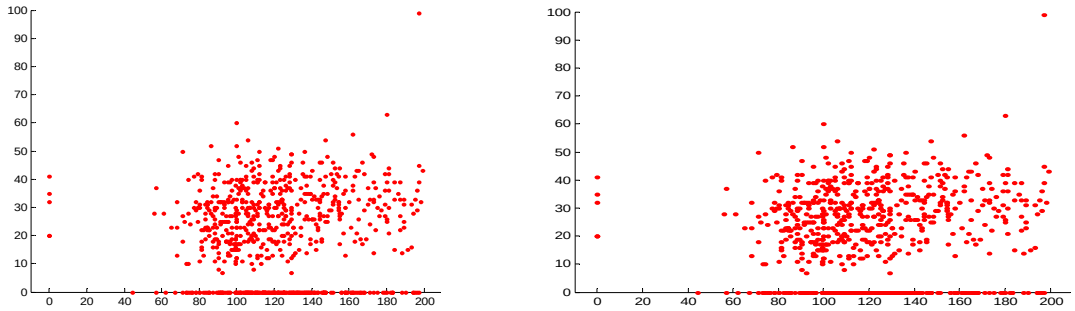
5.EXPERIMENTAL RESULTS FOR NAÏVE BAYES AND KNN:

DATASETS	NAÏVE BAYES ACCURACY	KNN ACCURACY
PIMA	96%	12.76%
WINE	100%	73.03%



pima dataset graph for naïve bayes **wine dataset graph for naivebayes.**

when comparing both the datasets by using naivebayes algorithm its clearly shows that wine dataset is having more accuracy when compared to pima dataset. while testing and training dataset its clearly says that the time escaped are also taking very less time while comparing the pima dataset. In this graphs it clearing shows that how much accuracy it was going to performed.



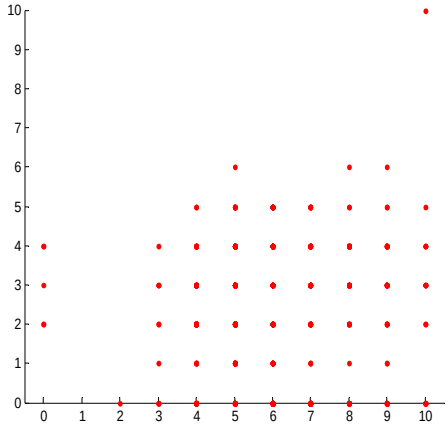
Pima dataset graph for KNN **Wine dataset graph for KNN**

When comparing both the datasets by using K-Nearest Neighbor algorithm its clearly show that wine dataset is having more accuracy when compared to pima dataset. While testing and training dataset it's clearly says that the time escaped are also taking very less time while comparing the Pima dataset. In this graphs it clearly shows accuracy when algorithms are implemented on two data sets..

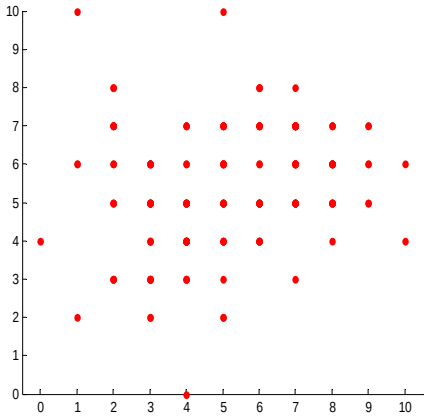
The datasets are compared by using the two algorithms i.e; Naive Bayes and K-Nearest Neighbor in this its clearly says that naïve bayes is having more accuracy when compare to KNN algorithm.

While applying the general feature selection method it says that the accuracy of both dataset pima and wine it says that for pima accuracy is more than knn algorithm and for wine it also having the more accuracy for the KNN algorithm.

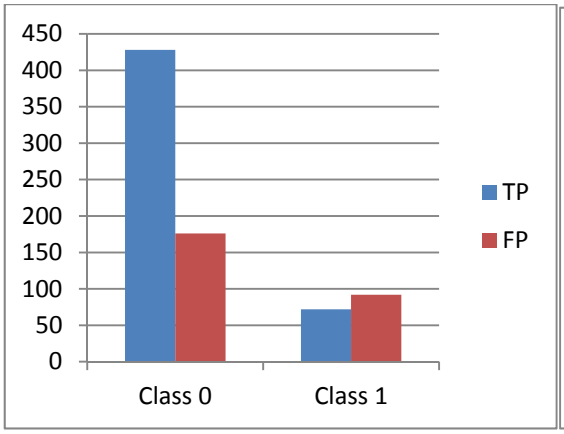
DATASETS	NAÏVE BAYES ACCURACY	KNN ACCURACY
PIMA	67.7%	80.4%
WINE	95.5%	100%



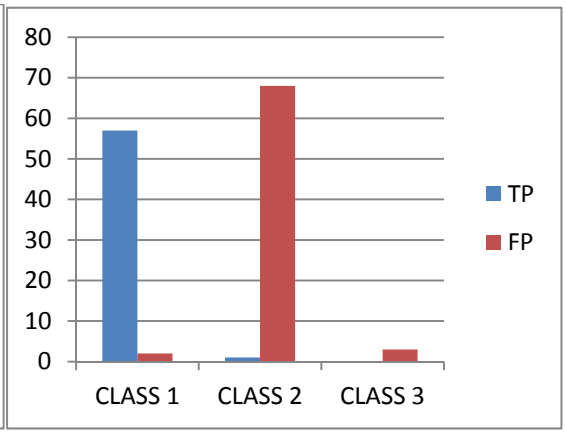
Pima dataset graph for KNN



Wine dataset graph for KNN



PIMA FOR NAÏVE BAYES



WINE FOR NAÏVE BAYES.

While comparing both feature selection and algorithm it says that KNN algorithm is having more accuracy.

6. Conclusion:

The features with feature selection method are used for classification and accuracy is good when we consider these features in the classification process. Hence feature selection plays a vital role in obtaining better accuracy when dimension space of data set is more.

REFERENCES

[1] JasminaNovakovic "The Impact of Feature Selection on the Accuracy of Naïve Bayes Classifier" 18th Telecommunications forum TELFOR 2010 Serbia, Belgrade, November 23-25,2010.
[2] H.Almuallim, and T.G.Dietterich," Learning with many irrelevant features", In:Proc.AAAI-91,Anaheim,CA,1991,547-552.
[3] K.Kiran and I.A.Rendell, "The feature selection problem: traditional methods and a new algorithm",In:Proc,IAAAI-92, San Jose,CA.
[4] Jasmina, Novakovic, Perica strbac, Dusan Bulatovic "Toward Optimal Feature Selection Using Ranking Methods and Classification algorithms", March 2011.
[5] Auburn, Alabama "A Class-specific Ensemble Feature Selection Approach For Classification Problems" May 9, 2009.
[6] C.J.Merz, and P.M. Murphy UCI Repository of machine learning databases, http://www.ics.uci.edu/mlearn/MLRepository.html 1998.
[7] M.Hall, Correlation based Feature Selection for Machine Learning, thesis, The University of Wailkato, Newzealand,1999.
[8] M.A.Jayaram, Asha Gowda Karegowda, A.S. Manjunath,"Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", International Journal of Computer Applications (0975 – 8887) Volume 1 – No. ,2010.
[9] PATLANGLEY,"**Selection of Relevant Features in Machine Learning",**AAAI Technical Report FS-94-02. Compilation copyright 1994.
[10] Yanshan Shi "**Comparing K-Nearest Neighbors and Potential Energy Method in classification problem".**
[11] Włodzisław Duch, Rafał Adamczak and Krzysztof Gra¸bczewski,"A new methodology of extraction, optimization andapplication of crisp and fuzzy logical rules" IEEE TRANSCTIONS ON NEURAL NETWORKS, VOL.1 NO.2, MARCH 2000.
[12] Manoranjan Dash, Huan Liu " Consistency-based search in feature selection" ELSEVIER Artificial Intelligence 151 (2003) 155–176.