Feature Usability Index and Optimal Feature Subset Selection

Debdoot Sheet and Jyotirmoy Chatterjee
School of Medical Science and Technoloy
Indian Institute of Technoloy Kharagpur
Kharagpur, 721302, India

Hrushikesh Garud
Texas Instruments (I) Pvt. Ltd.
Bangalore, 560093, India

ABSTRACT

Feature usability index is introduced here as a measure for evaluating classification efficacy of features. It is defined using measures of homogeneity, class specificity, and error in decision making. Homogeneity measures the extent of outlying observations, class specificity assesses the separation between distributions of different labeled classes, and error in decision making is computed using overlap in posteriori decision boundary. This is followed by feature ranking and optimal feature subset selection through ordering of features based on feature usability index and involves a complexity of $O(D\log D)$ for D features. The results validating classifier independent feature ranking and optimal feature subset selection are also presented aong with a comparative analysis using χ^2 statistics for feature selection.

General Terms

Pattern Reconition, Machine Intelligence, Data Mining

Keywords

Feature ranking, feature selection, knowledge discovery, knowledge engineering, pattern recognition

1. INTRODUCTION

Pattern recongition is defined as an act of taking in raw data and making an action based on the representative pattern in it [6, 22]. Over the past millions of years, humans have evolved an astoundingly complex process that underlies this act of perception and decision [4]. Accordingly there have been manifold developments in building human like intelligent decision making machines, along with greater exploration and understanding of salient concepts involved in naturally existing pattern recognition systems.

Traditionally, the process of pattern recognition starts with identification of linguistic features by heuristic inspection of an application expert. These linguistic features are measured appropriately using analytical methods for creating a library of features (attributes) to be used in pattern recognition applications. Every analytical method used for quantifying a linguistic feature indicates at a specific physical essence. Unless explicitly specified, given an object, its linguistic feature, is expressed in an attribute vector $\mathbf{x} = \left[x^{(1)}, x^{(2)}, \cdots, x^{(d)}, \cdots, x^{(D)}\right]$, where each element $x^{(d)}$ of $\mathbf{x}$ is an analytical value contributed by measurement of the d^{th} feature. The N samples of $\mathbf{x}$ are $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$, and their corresponding class labels are $y_1, y_2, \cdots, y_N$. These class labels are finite and $y_n \in \{\omega_1, \omega_2, \cdots, \omega_K\}$ for a K class classification problem.

In a classification or discrimination process we assume that we have a set of patterns of known class $\{(\mathbf{x}_n, y_n), n = 1, 2, \ldots, N\}$ referred to as the training or design set and is used to design the classifier. A classifier is mathematically expressed as a hypothesis or a combination of hypotheses used for mapping an input pattern $\hat{\mathbf{x}}_n$ to a label $\hat{y}_n$.

$$h_\zeta : \hat{\mathbf{x}}_n \to \hat{y}_n \tag{1}$$

This hypothesis $\left(h_\zeta\right)$ is referred to as an inducer in the training phase and as predictor in testing and deployment phase. Proper classification is achieved when $\hat{y}_n \equiv y_n$. Contrary to this action, an improper classification is associated with a loss due to error in decision making. Proper training of the inducer using a salient training set of data ensures reduction in loss due to error in decision making. For a given architecture, this involves selecting the best subset of features $\hat{\mathbf{x}}$ for which $J\left(h_\zeta : \hat{\mathbf{x}}\right) = \max_{\{\hat{\mathbf{x}}\}} \left\{J\left(h_\zeta : \hat{\mathbf{x}}\right)\right\}$, where $J(\cdot)$ is the performance measure of a hypothesis or classifier and includes either of error rates, accuracy, estimates of probability of class membership to true probability, etc. In a training set with D features, the dimension of the hypotheses $\left(h_\zeta : \hat{\mathbf{x}} \to \hat{y}\right)$ to be induced for testing to obtain the best perfroming feature subset is $\sum_{d=1}^{D} \binom{D}{d} \approxeq 2^D$. Thus a linear change in the dimension of the feature vector $\hat{\mathbf{x}}$ leads to an exponential change in dimension of the hypotheses space.

1.1 Dimensionality Reduction and Feature Selection

Since linear reduction in feature space leads to an exponential decline in size of the inducer training set, there has been considerable developments in this respect. Methods of dimensionality reduction includes principal component analysis, which seeks to project data in a least square sense through linear kernel transformaton. This method performs well with statistically dependent measurements [6, 22]. Another approach for reduction in dimension of feature space is through feature slection [10–13].

Feature selection as an area of interest within pattern recognition, deals with selection of a subset or list of attributes or variables used in construction of a model describing observations. The purpose of this stage includes reducing data dimensionality through removal of irrelevant and redundant features, reducing the amount of data needed for learning, improving predictive accuracy of a classification hypothesis, and increasing comprehensibility of the constructed hypothesis [10].

Although literatures suggest several existing methods of feature selection [13] yet, Forman states "…feature selection is still in its infancy and major opportunities await" [10].

1.2 Organization of the Paper

The paper presents feature usability index as a measure of classification efficacy of features in Section 2. This measure is consequently used for feature ranking followed by optimal feature subset selection in Section 3. Experimental validation of the technique using three classifier architectures and two dataset are also presented along with a comparison of feature subset selction using χ^2 statistics based test in Section 4. Section 5 discusses the performance of experimental trials and their outcomes. Finally we conclude about the properties of our method and scope for further development in Section 6.

2. FEATURE USABILITY INDEX

In this work a fusion of several methods used by domain experts from statistics, image processing and decision theory are used for reducing complexity of the hypotheses space by reducing dimensionality of the feature space. Feature usability index presented in this section is used as a measure of expressing classification efficacy of a feature. We represent usability index of the d^{th} feature $\mathscr{F}^{(d)}$ as $f^{(d)}$, computed using measures of homogeneity, class specificity, and error in decision making. The following subsections elaborate further.

2.1 Homogeneity

An important aspect of data quality analysis not addressed in feature selection approaches till data [11] is in understanding homogeneity of observations. Observations are more homogeneous if they have less density of outliers. Although they are not completely nonexistant, yet complete lack of outliers in a set of observations can lead to model misspecifications, and incorrect results [3]. Various approaches in outlier identification and their subsequent rejection are generally categorized for data generated out of *univariate* and *multivariate* distributions [3], *parametric* and *non-parametric* methods of identification [3], and *predictable* [18, 23] as well as *unpredictable* [1, 2, 7] number of outliers.

The method of expressing homogeneity of observations is based on [23] and starts with computing the *one-outlier scatter ratio of a sample*

$$s_n^{(d)} = \frac{|a^{(d)}|}{|a_n^{(d)}|} \tag{2}$$

where $|a^{(d)}|$ is the *internal scatter* of samples belonging to each of the K classes, and $|a_n^{(d)}|$ is the analogous quantity with the n^{th} observation omitted.

Literatures have used higher order statistics based on *one-outlier scatter ratio* for rejecting outlying observations [2]. Here we define and use *one outlier ratio* as the ratio of the number of outlying observations to the number of coherrent observations for identifying density of outlying observations.

$$O^{(d)} = \frac{\text{card}\left\{\left(|s_n^{(d)} - 1| > 0.01\right)\right\}}{\text{card}\left\{\left(|s_n^{(d)} - 1| \leq 0.01\right)\right\}} \quad \forall n = 1, 2, \cdots, N \tag{3}$$

where card$\{\cdot\}$ denotes the set cardinality opeartor. $O^{(d)}$ provides an essence of the density of outlying observations with respect to the d^{th} feature extraction technique.

The numerical expression of this measure has a theoretical bound of $[0, \infty)$ with the minimum value pertaining to observations with no outliers. Higher values tending towards maximal boundary indicate at abundance of outliers in the data.

2.2 Class Specificity

Class specificity of observations for a particular feature indicates its discrimination potential. It is generally associated with a high value of *between class scatter* and a low value of *within class scatter* [6]. This characteristic of a feature is inspired by [17], and for a multi-class problem it is expressed as the minimum of the ratio of *between class scatter* to *within class scatter* of the observations analyzed over all classes. For the d^{th} feature, it is expressed as,

$$S^{(d)} = \min_{(y_j, y_k)} \left\{ \frac{\left|\mu^{(d)}|_{y_j} - \mu^{(d)}|_{y_k}\right|^2}{\left|\left(\sigma^{(d)}|_{y_j}\right)^2 + \left(\sigma^{(d)}|_{y_k}\right)^2\right|} \right\} \tag{4}$$

where $(y_j, y_k) \in \{\omega_1, \omega_2, \cdots, \omega_K\}$ and $y_j \neq y_k$; $\mu^{(d)}|_{y_k}$ and $\sigma^{(d)}|_{y_k}$ are the *mean* and *standard deviation* of the observations corresponding to the d^{th} feature having class label y_k.

Expression bound of this measure is $[0, \infty)$ with minimum indicating at complete overlap of the distributions and the value tending towards maximum due to large separation between them.

2.3 Error in Decision Making

Error in decision making generally arise due to overlap in *posteriori* decision boundary. This is directly associated with the risk involved in misclassifying observationa. Here class overlap in Bayesian *posteriori* decision boundary [6, 22], to the strength of decision making, involved in deciding based on the d^{th} feature is expressed as the risk factor and is quantified accordingly.

$$R^{(d)} = \frac{\int_{x^{(d)}} P_{\min}\left(y_k|x^{(d)}\right) dx^{(d)}}{\int_{x^{(d)}} P_{\max}\left(y_k|x^{(d)}\right) dx^{(d)}} \qquad (5)$$

where $P_{\min}\left(y_k|x^{(d)}\right) = \min_{y_k \in \{\omega_1,\omega_2,\cdots,\omega_K\}} \left\{\left(P\left(y_k|x^{(d)}\right)\right)\right\}$, $P_{\max}\left(y_k|x^{(d)}\right) = \max_{y_k \in \{\omega_1,\omega_2,\cdots,\omega_K\}} \left\{\left(P\left(y_k|x^{(d)}\right)\right)\right\}$, and $P\left(y_k|x^{(d)}\right)$ is the Bayesian *posteriori* probability of belongingness of an observation $x^{(d)}$ to a class with label $y_k \in \{\omega_1, \omega_2, \cdots, \omega_K\}$.

This measure has an expression bound of $[0,1]$ with maximum indicating at complete overlap in posteriori decision boundary while maximum indicating at zero overlap.

2.4 Feature Usability Index

Classification efficacy of a feature is expressed using usability index computed using the prior evaluated characteristics of features. The form of expression is simple and presented as in Eq. 6 in consideration of bounds of the prior expressions.

$$f^{(d)} = \frac{S^{(d)}}{O^{(d)} \times R^{(d)}} \qquad (6)$$

where $O^{(d)}$, $S^{(d)}$, and $R^{(d)}$ are respective measures reflecting homogeneity of observations, their class specificity, and error in decision making based on *posteriori* decision boundary.

Feature usability index expressed based on these three measures has an expression bound of $[0,\infty)$ with minimum corresponding to the worst feature and vice-versa.

3. FEATURE RANKING AND OPTIMAL FEATURE SUBSET SELECTION

3.1 Ranking using Feature Usability Index

A hypothesis designed using features with observations which exhibit a high degree of homogeneity and class specificity, and are less plagued with error in decision making, performs the best. However, since all of these characteristics are not expressed with equal potential in all the features, we devise a ranking scheme based on feature usability index. For every feature $\mathscr{F}^{(d)}$, its usability index is associated as $\left\{\left(\mathscr{F}^{(d)}|f^{(d)}\right)\right\}\ \forall\ d = 1,2,\cdots,D$. This is followed by ranking of features based on feature usability index using *m-ordering* system [2].

$$\mathscr{F}^{(d)}_{\text{ordered}} = \arg\left\{\Phi_{\text{m-order}}\left(\left\{f^{(d)}\right\}\right)|\mathscr{F}^{(d)}\right\} \qquad (7)$$

The ranked features are accordingly available based on their classification efficacy. In this application the ranked set is the ordered set of features $\mathscr{F}^{(d)}_{\text{ordered}}$ and is used as the guiding criteria for feature subset generation for subsequent selection using *wrapper* model [16].

3.2 Optimal Feature Subset Selection

Wrapper model of optimal feature subset selection presented here is specific to the hypothesis used. This method includes subset generation using SFS and subset evaluation is done based on the accuracy of experiments tried out with the chosen hypothesis. Algorithm 1 illustrates the wrapper model used in this application.

```
Input: Ordered feature set $\left\{\mathscr{F}^{(d)}_{\text{ordered}}\right\}$
Output: Optimal feature subset $\left\{\mathscr{G}^{(d)}\right\}$ specific to a classifier
Initialization: Shift the top element of $\left\{\mathscr{F}^{(d)}_{\text{ordered}}\right\}$ to $\left\{\mathscr{G}^{(d)}\right\}$, and set $A_0 = 0$;
Compute the classifier accuracy $A_1 = J\left(h_\zeta : \mathbf{x} \to y|\left\{\mathscr{G}^{(d)}\right\}\right)$ with current $\left\{\mathscr{G}^{(d)}\right\}$;
while $A_k > A_{k-1}$ do
    Shift the top element of $\left\{\mathscr{F}^{(d)}_{\text{ordered}}\right\}$ to $\left\{\mathscr{G}^{(d)}\right\}$;
    Compute the classifier accuracy $A_{k+1} = J\left(h_\zeta : \mathbf{x} \to y|\left\{\mathscr{G}^{(d)}\right\}\right)$;
    if $A_{k+1} > A_k$ then
        increment $k$;
    else
        end process and exit;
    end
end
```

Algorithm 1: Optimal feature subset selection

The optimal feature subset $\left\{\mathscr{G}^{(d)}\right\}$ obtained using this method is specific to a classifier of choice and hence varies across different classifier models.

4. EXPERIMENTS

The results are presented for experiments conducted for feature ranking based optimal feature subset selection using two data sets. The data sets are obtained from the UCI Machine Learning repository [`http://archive.ics.uci.edu/ml/`], and pertain to disease diagnosis. Validation of this idea of classifier specific optimal feature subset selection is substantiated using results of experiments performed using Bayesian Classifier with multivariate normal distribution and minimum risk rule(Bayesian MVN), Fisher's Linear Discriminant Analysis (FLDA), and a Support Vector Machine with linear kernel and least square separating hyperplane(SVM) [6, 22].

4.1 About the Datasets

4.1.1 Dataset 1: Wisconsin Diagnostic Breast Cancer

The Wisconsin Diagnostic Breast Cancer database was originally contributed by the University of Wisconsin, Madison. It consists of ten real valued attributes used for diagnosis of breast cancer into benign and malignant classes nuclear features extracted from using cytological images of fine needle aspirates of breast lesions. Past usage involves [14, 21].

4.1.2 Dataset 2: Pima Indian Diabetes

The Pima Indian Diabetes database was originally contributed by the National Institute of Diabetes and Digestive and Kidney Diseases. It consists of eight characteristic real valued attributes pertaining to genetic pedigree, clinical and biochemical investigation and used for diagnosis of diabetes. Past usage involve aftificial intelligence based techniques for predicting onset of diabetes [20].

4.2 Design of Experiment

Generally a large sample size results in better performance of a complex classifier. Several literatures suggest the ratio of sample size to number of features in the design of a pattern recognition system giving the general guidance of having 5-10 times more samples per class than number of features [8, 9, 22].

Once sufficient observations are gathered, they are partitioned into *train* and *test* sets. This is done to serve two purposes: one in which a *train* set is used in hypothesis design for tuning of parameters, and the other in which a classifier's performance is independently assessed using a *test* set. Generally the hypotheses assessed with the best performance using the *train* set are used in deployment on the classifier.

An important point worth mention is *randomization* of data collection. This is done to reduce the effect of bias. A general approach is randomization of the order of data, rather than collecting all data from all classes in experimentation. A well adopted method of experimentation used for reporting of classifier performance is through multiple randomized experimentation, followed by reporting of the average performance metric as a result of such random experiments [15, 22]. In this work multiple number of experiments are conducted, with randomly chosen *training* and *testing* samples. One trial consists of 10,000 randomized experiments done in order to avoid experimental bias. The accuracy with a feature subset for a particular classifier is the mean of the accuracy obtained from all the experiments.

4.3 Results

4.3.1 *Dataset 1: Wisconsin Diagnostic Breast Cancer*

The ten features are ranked using a classifier independent method, followed by selection of the optimal feature subset specific to a classifier. Figure 1 presents a consolidated account of the values of $O^{(d)}$, $S^{(d)}$ and $R^{(d)}$ for each feature, along with $f^{(d)}$. These are arranged according to their ranks based on feature usability index, and optimal performing feature subsets are selected using Algorithm 1 for a specific classifier. The average performance of three different classifiers over 10,000 random trials using subsets selected are presented in Figure 2.

4.3.2 *Dataset 2: Pima Indian Diabetes*

The eight features are ranked using a classifier independent method, followed by selection of the optimal subset specific to a classifier. Figure 3 presents a consolidated account of the values of $O^{(d)}$, $S^{(d)}$ and $R^{(d)}$ for each feature, along with $f^{(d)}$. These are arranged according to their ranks based on feature usability index, and optimal performing feature subsets are selected using Algorithm 1 for a specific classifier. The average performance of three different classifiers over 10,000 random trials using subsets selected are presented in Figure 4.

5. DISCUSSION

Cover [5] has observed that statistical dependence among measurements (features) can cause the best *k*-element subset not to be composed of the *k* best measurements, and even conditionally independent measurements (features) can exhibit such behavior. Ideally a feature ranking scheme should be such that the *k* best features would form the best *k*. Under such circumstances, during the SFS stage the accuracy levels achieved increase monotonically for first *k* iterations to reach the global maxima. Any further addition to the subset give accuracy lower than the global maximum. So, while evaluating the performance of a ranking scheme, the observed behavior of accuracy achieved in SFS should be compared with this ideal situation.

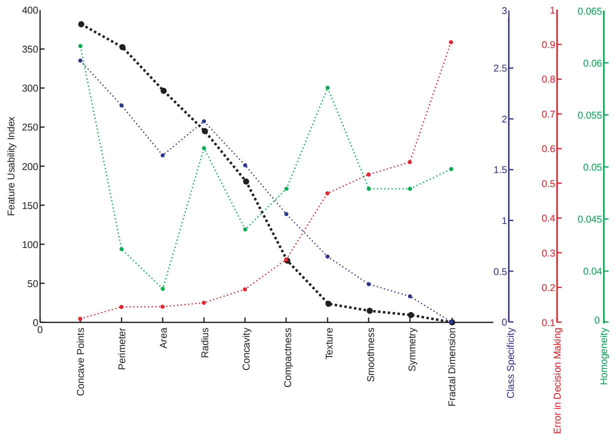

Figure 1: Plot illustrating feature usability index and constituent criteria for ten different features in Dataset 1

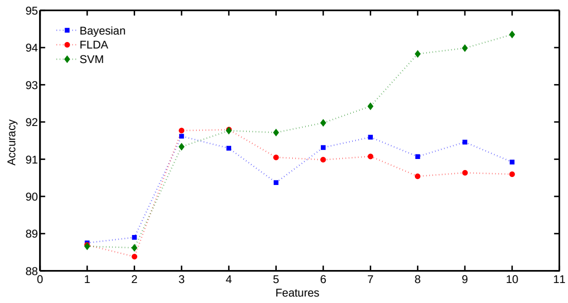

Figure 2: Accuracy of the experimental trials obtained with feature ranking based subset selection with Dataset 1. Performance of the different classifiers are color coded corresponding to legends. The independent axis (x-axis) of the plot presents number of features in a subset formed by incremental inclusion of ranked features starting from the best ranked feature (*concave points*), following (*concave points, perimeter*), and similarly to form incrementally increasing feature subsets.

5.1 Computational Complexity of the Method

Computational complexity involved in optimal feature subset selection is an important analysis requirement for comparing across different feature selection algorithms. Table 1 presents a consolidated account of the complexities involved in different algorithms, as a function of the total number of features (D). The *m-ordering* stage being a sorting mechanism limits the complexity to $O(D \log D)$.

The method described here achieves the minimum complexity as compared to other popular feature selection algorithms for classifi-

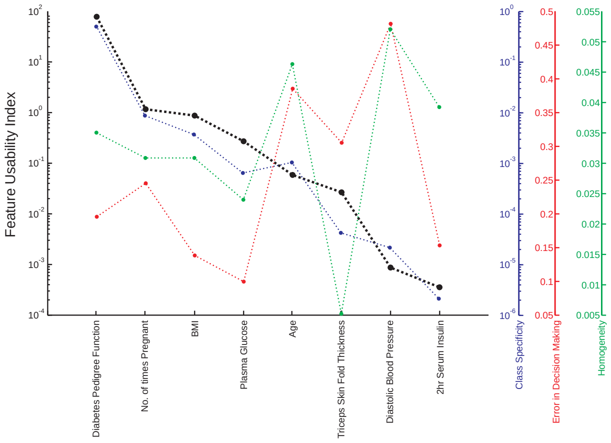

Figure 3: Plot illustrating feature usability index and constituent criteria for eight different features in Dataset 2

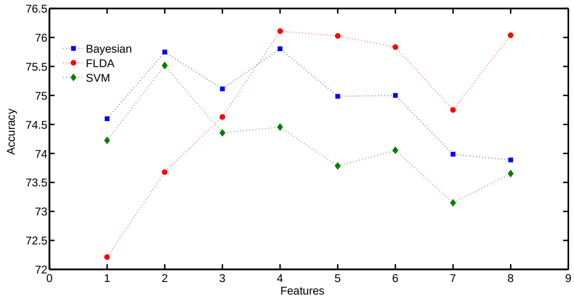

Figure 4: Accuracy of the experimental trials obtained with feature ranking based subset selection with Dataset 2. Performance of the different classifiers are color coded corresponding to legends. The independent axis (x-axis) of the plot presents number of features in a subset formed by incremental inclusion of ranked features starting from the best ranked feature (*diabetes pedigree function*), following (*diabetes pedigree function, no. of times pregnant*), and similarly to form incrementally increasing feature subsets.

cation with labeled samples.

5.2 Experimental Observations

5.2.1 Dateset 1: Wisconsin Diagnostic Breast Cancer

During SFS stage it has been observed that the accuracy levels achieved with Bayesian MVN and FLDA increase monotonically to achieve global maximum accuracy with a subset upto third and fourth features respectively, after which it decreases to levels lower than the global maxima. For Bayesian MVN three best features form the optimal feature subset with 91.62% accuracy, while for FLDA four best features form the optimal feature subset with 91.79% accuracy. Whereas for SVM the accuracy levels achieved increase monotonically to achieve global maxima with 94.35% accuracy when the entire feature set is used for classification; thus the entire feature set forms the optimal subset here.

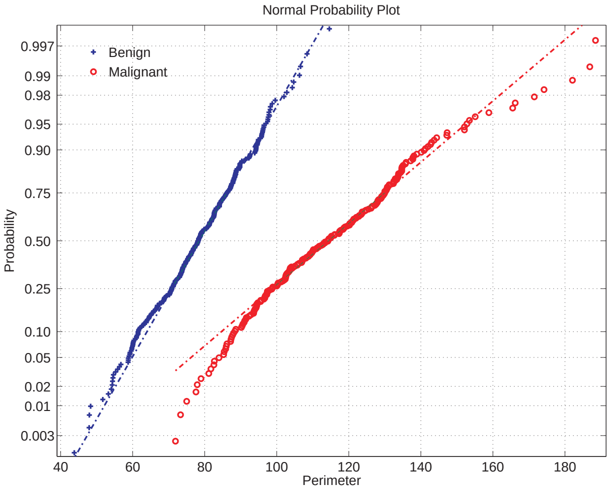

Figure 5: Log-normal probability plot for the class labeled observations of the feature *Perimeter* in Dataset 1

For the purpose of comparing performance of the proposed method with an existing method, χ^2 test of significance for feature selection is choosen. The method is based on [11, Chapter 14]. χ^2 index with 95% margin and 1 degree of freedom is used here for selecting the optimal subset of features containing the five features *texture, perimeter, area, smoothness,* and, *concave points*. The comparitive performance is presented in Table 2. Overall average accuracy with this subset through one trial of 10,000 random experiments is 86.27% for Bayesian MVN, 89.66% for FLDA, and 82.87% for SVM. The optimal feature subset selected through this method has *concave points, perimeter*, and *area* as the constituents increasing accuracy of the feature set while *texture*, and *smoothness* lead to detrioration of performance.

In figures 5 and 6 we provide the log-normal probability distribution of the observations for best and worst ranked features, viz., *Perimeter* and *Fractal Dimension* respectively. These figures provide visual representations of the characteristics that are quantified by class homogeneity, class specificity, and probability of error in misclassification. Ideally the cumulative probability of observations obeying a normal distribution follows the dashed line. It is observed that for both the features, observations for both the labeled classes follow normal distribution, except at the tails. These observations which deviate from the normal distribution at the extremities are statistically termed as outliers, and are responsible for erroneous inference. For the best ranked feature *Perimeter*, number of such observations is less than that for the worst ranked feature *Fractal Dimension*. It can also be seen that the observations of the labeled classes are widely separated for *Perimeter*, while they completely overlap for *Fractal Dimension*.

5.2.2 Dataset 2: Pima Indian Diabetes

In experiments on this dataset, it has been found that two best ranked features form the optimal subset for Bayesian MVN with 75.55% accuracy, and SVM with 75.51% accuracy, whereas for FLDA the four best features form the optimal subset with 76.11% accuracy.

The eight different features in the PIMA Indian Diabetes dataset exhibit large variations in their characteristic properties. It can be

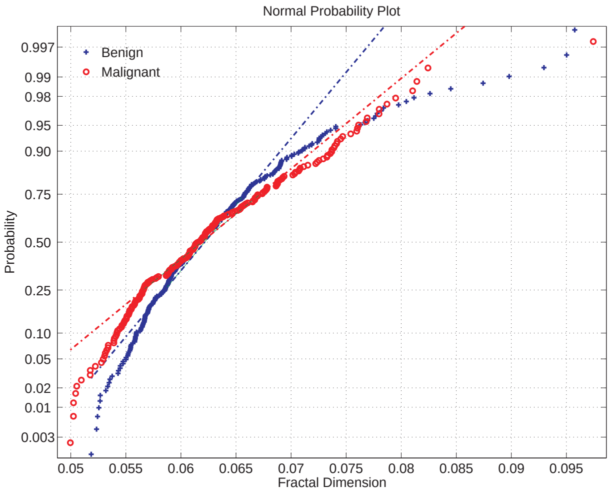

Figure 6: Log-normal probability plot for the class labeled observations of the feature *Fractal Dimension* in WDBC

inferred from the characteristics of the features presented in Fig. 4.4 that *class specificity* varying in the range of 10^{-6} to 100 plays an important role in deciding the value of *feature usability index*. However, other factors *error in decision making* and *homogeneity* which vary in a limited range of 0 to 1, also have their relevant effects. Accordingly in Fig. 4.5 it is observed that large variations in *homogeneity* and *error in decision making* between the features *age, triceps skin fold thickness, diastolic blood pressure, 2hr serum insulin* affect the performance of the feature groups at 6, 7, 8. It can be observed that performance of FLDA decreases between 6 and 7 while increases between 7 and 8, due to associated increase and subsequent decrease in *error in decision making* and *homogeneity* respectively. Similar behavior for other classifiers viz., Bayesian MVN and SVM are also evident.

6. CONCLUSION

Feature usability index is presented as an objective measure for evaluating classification efficacy of features. This measure is evaluated based on available observations and using measures of homogeneity, specificity, and error in decision making computed independently for each feature. Homogeneity has an expression bound of $[0,\infty)$ with the minimum value corresponding to the feaures with no outlying observations. Class specificity has an expression bound of $[0,\infty)$ with minimum value corresponding to completely overlapped distributions. Error in decision making has an expression bound of $[0,1]$ with maximum value corresponding to completely overlapped decision boundary. Feature usability index is accordingly expressed based on these three measures and for the d^{th} feature is represented as $\mathscr{F}^{(d)}$ with an expression bound of $[0,\infty)$ where minimum value corresponding to the worst feature and viceversa. The method of feature ranking uses a *m-ordering* scheme for ranking of features based on objective score represented by feature usability index. Optimal feature subset contributing to maximum performance of a classifier is selected from the ranked feature set using SFS. The overall complexity of searching for an optimal feature subset from a candidate set of D features is $O(D\log D)$, which is the minimum when compared to other existing techniques proposed in literature. Experimental validations of the claim are also provided in this chapter and are based on disease diagnostic database sourced from UCI Machine Learning repository.

For the purpose of adequacy and sample sufficiency, the ratio of sample size to the number of features in computing the measures of individual features is about 5 - 10 times more samples per class than the total number of features [8, 9, 22]. This criteria introduces limitation in usage of the proposed feature usability index in datasets with number of features far higher than the number of observations. Subtle modifications may be incorporated appropriately for usage in such experimental conditions.

7. ACKNOWLEDGMENTS

The authors acknowledge pedaogical contributions from Prof. Ajoy K. Ray, Dr. Chandan Chakraborty and Prof. Pranab K. Datta from Indian Institute of Technology Kharagpur in development of this conjecture, and support of M/s Texas Instruments (I) Pvt. Ltd., Bangalore, India in supporting this work in part under the Texas Instruments (India) and Indian Institute of Technology Research Scheme.

8. REFERENCES

[1] F. J. Anscombe and Irwin Guttman. Rejection of outliers. *Technometrics*, 2(2):123–147, May 1960.

[2] V. Barnett. The ordering of multivariate data. *Journal of Royal Statistical Society*, 139(3):318–355, 1976.

[3] Irad Ben-Gal. Outlier detection. In O. Maimon and L. Rockach, editors, *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Acad. Pub., 2005.

[4] P. M. Churchland. Chimerical colors: Some novel predictions from cognitive neuroscience. In A. Brook and K. Akins, editors, *Cognition and the Brain*, pages 309–335. Cambridge University Press, 2005.

[5] T. M. Cover. The best two independent measurements are not the two best. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-4(1):116–117, Jan 1974.

[6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2001.

[7] Ali S. Hadi. Identifying multiple outliers in multivariate data. *Journal of Royal Statistical Society*, 54(3):761–771, 1992.

[8] Anil K. Jain and B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, pages 835–855. North Holland, Amsterdam, 1982.

[9] H. M. Kalayeh and D. A. Landgrebe. Predicting the required number of training samples. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-5(6):664–667, Nov. 1983.

[10] H. Liu, E.R. Dougherty, J.G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, Z. Zhao, L. Yu, and G. Forman. Evolving feature selection. *Intelligent Systems, IEEE*, 20(6):64–76, Nov.-Dec. 2005.

[11] Huan Liu and Hiroshi Motoda. *Computational Methods for Feature Selection*. CRC Press, 2008.

[12] Huan Liu and Rudy Setiono. Incremental feature selection. *Applied Intelligence*, 9:217–230, 1998.

[13] Huan Liu and Lei Yu. Toward integrating feature selection algorithms for classification and clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):491–502, April 2005.

[14] Olvi L. Mangasarian, W. Nick Street, and William H. Wolberg. Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Operations Research*, 43(4):570–577, 1995.

[15] P. Mitra, C.A. Murthy, and S.K. Pal. Unsupervised feature selection using feature similarity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):301–312, Mar 2002.

[16] Linda Mthembu and Tshilidzi Marwala. A note on the separability index. Online, 2008.

[17] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62–66, Jan. 1979.

[18] E. S. Pearson and C. Chandra Sekar. The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28(3/4):308–320, Dec. 1936.

[19] Helene Schulerud and Fritz Albergtsen. Many are called, but few are chosen. feature selection and error estimation in high dimensional spaces. *Computer Methods and Programs in Biomedicine*, 73:91–99, 2004.

[20] J W Smith, J E Everhart, W C Dickson, W C Knowler, and R S Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Annual Symposium on Computer Applications in Medical Care, Proceedings of*, pages 261–265, Nov 1988.

[21] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In R. S. Acharya & D. B. Goldgof, editor, *SPIE Conference Series, Proceedings of*, volume 1905, pages 861–870, Jul 1993.

[22] Andrew Webb. *Statistical Pattern Recognition*. Wiley, 2002.

[23] S. S. Wilks. Multivariate statistical outliers. *Sankhyā*, 25(4):407–426, 1963.

Table 1: Computational Complexity of Feature Selection Algorithms

Method	**Literature**	**Complexity**
Sequential selection (SFS)	[19]	$O\left(2^D\right)$
Genetic Algorithm based search	[11, Chapter 8]	$O\left(D^2\right)$
χ^2 test based selection	[11, Chapter 14]	$O\left(D^2\right)$
Feature Ranking based Selection	**Proposed Method**	$O(D\log D)$

Table 2: Performance Comparison of Feature Selection Techniques for Supervised Classification

Method	**Bayesian MVN**		**FLDA**		**SVM**	
	Constitutent features	Accuracy	Constituent features	Accuracy	Constituent features	Accuracy
χ^2 test based	*texture, perimeter, area, smoothness, concave points*	86.27%	*texture, perimeter, area, smoothness, concave points*	89.66%	*texture, perimeter, area, smoothness, concave points*	82.87%
Proposed method	*concave points, perimeter, area*	91.62%	*concave points, perimeter, area, radius*	91.79%	*concave points, perimeter, area, radius, concavity, compactness, texture, smoothness, symmetry, fractal dimension*	94.35%