# Discriminative Feature Selection by Nonparametric Way with Cluster Validation

**Choudari Anantha[1], Dharmaiah Devarapalli[2]**

[1,2]Department of Computer Science & Engineering, Vignan's Institute Of Information Technology, Duvvada, Viskhapatnam-530049.A.P, India

**Abstract**: ***Feature Selection is the preprocessing process of identifying the subset of data from large dimension data. To identifying the required data, using some Feature Selection algorithms. Like Relief, Parzen-Relief algorithms, it attempts to directly maximize the classification accuracy and naturally reflects the Bayes error in the objective. Proposed algorithmic framework selects a subset of features by minimizing the Bayes error rate estimated by a nonparametric estimator. As an example, we show that the Relief algorithm greedily attempts to minimize the Bayes error estimated by the k-Nearest-Neighbor (kNN) method. In particular, by exploiting the proposed framework, we establish the Parzen-Relief (P-Relief) algorithm based on Parzen window estimator. The Relief algorithm is a popular approach for feature weight estimation. Many extensions of the Relief algorithm are developed. Because of the randomicity and the uncertainty of the instances used for calculating the feature weight vector in the Relief algorithm, the results will fluctuate with the instances, which lead to poor evaluation accuracy. To solve this problem, a feature selection algorithm parzen+relief based algorithm is proposed. It takes both the mean and the variance of the discrimination among instances and weights into account as the criterion of feature weight estimation, which makes the result more stable and accurate.***

**Keywords**: Raw data, Bayes errors, Relief+knn, Parzen+relief, Cluster quality

## 1. Introduction

Feature selection methods can be grouped into two categories from the point of view a method's output. One category is about ranking feature according to same evaluation criterion; the other is about choosing a minimum set of features that satisfies an evaluation criterion. In this work we are using Discriminative optimal criterion (DOC)**[11]**[14], DoC is pragmatically advantageous because it attempts to directly maximize the classification accuracy and naturally reflects the Bayes error in the objective. Many feature subset selection algorithms are existed but not all of them are suitable for a given feature selection problem **[1]**. As an example, we show that the Relief algorithm greedily attempts to minimize the Bayes error estimated by the k-Nearest Neighbor (kNN) method. The Bayes error rate gives a statically lower bound on the error achievable for a given classification problem and associated choice of features [7]. In particular, by exploiting the proposed framework, we establish the Parzen-Relief (P-Relief) algorithm based on Parzen window estimator, and the MAP-Relief (M-Relief) which integrates label distribution into the max-margin objective to effectively handle imbalanced and multiclass data. Relief is a acknowledged feature re-weighting algorithm [8].

Typical estimation measures can be divided into filters and wrappers. Filter based feature selection methods are in general faster than wrapper based methods. As one of the filter based feature selection methods, the Relief algorithm is an effective, simple, and widely used approach to feature weight estimation [13]. The weight for a feature of a measurement vector is defined in terms of feature relevance. Relief algorithm solves a convex optimization problem with a margin based objective function [9]. These two probabilities are of the value of a feature being different conditioned on the given nearest miss and nearest hit, respectively. Thus, Relief usually performs better than the other filter based approaches due to the feedback of the nearest-neighbor classifier. Relief methods are naturally successful attribute estimators [10].

In this paper, a novel relief[12] feature selection algorithm based on Mean-Variance model is proposed. Both the mean and the variance of the samples discrimination are considered as the criterion of feature weight estimation. In this way, the results are more stable and accurate. In this paper, we first present a theoretically optimal criterion for feature selection, namely the discriminative optimal criterion (DoC), as a complementarity to the representative one (referred to as representative optimal criterion (RoC)).

Compared to RoC, DoC is practically favorable if our ultimate goal is for the purpose of supervised classification. For example, taking feature weighting as an example search strategy, we show that the Relief algorithm attempts to greedily minimize the nonparametric Bayes error that is estimated by k-nearest-neighbor (kNN) method. In this paper, an alternative algorithm, called Parzen-Relief (PRelief), is proposed, which resembles the standard Relief algorithm, but instead of kNN, it uses the Parzen window method to estimate the Bayes error. We show that the empirical performance of Parzen-Relief usually outperforms Relief. In addition, we find that the Relief makes an implicit assumption that the class distribution is balanced among every "one-versus-rest" split of the data.

Both Parzen-Relief and MAPRelief are of the same computational complexity as the standard Relief algorithm; And we are having Leaders and sub-Leaders algorithm, it is an efficient hierarchical clustering algorithm, suitable for large data sets[6]. It uses incremental clustering principles to generate a hierarchical structure for finding the subgroups/sub clusters within each cluster. Two major features of Leaders–Sub-leaders are: effective clustering and prototype selection for pattern classification. In this paper, a new clusters algorithm is proposed, which makes use of a clustering quality index is proposed for this application. Clustering algorithms organize data elements into groups based on their similar properties [2].

Clustering is the process of grouping the data into classes [3]. The main idea of clustering is the outliers based on the leader algorithm instead of reclustering the entire sliding data set by calculating the threshold with the average method and maximal resemblance of leaders [4]. Clustering of data has been of intense need for any organization in the fast and various researchers in the field of data mining [5].

Organization of the paper follows Introduction, Methodology, Implementation, Results and Conclusion. Methodology contains Relief+kNN and Parzen+relief algorithms. Working of methodology explains in implementation part. Results area consist output of the algorithms. Conclusion tells how these algorithms are worked.

## 2. Methodology

In this paper we are taking input as a large amount of raw data in Microsoft excel file. That dataset contains number of rows and large amount of columns. After that we are performing normalization operation on the data set for better outputs because normally dataset having some missing values, irrelevant values and multiclass problems. Here normalization takes all the values into single scale.

In this paper relationship is the difference. Three distance measures are used in this paper, that are Euclidean, Manhattan, cosine.

### A. Analysis of the Paper

Analysis of the paper follows below flow chart. These are the following steps to execute this paper

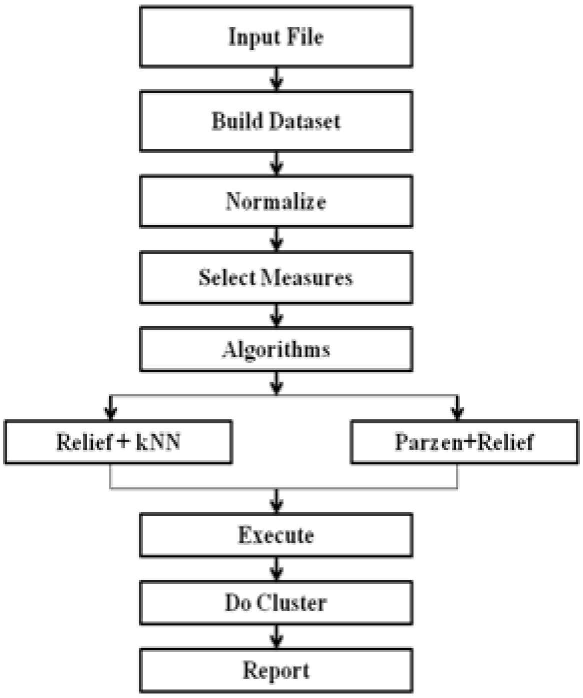


**Figure 2.1:** Analysis of the paper ”flow chart”

1) **Input File**
   Here input file takes from the external source and it acts as dataset to the algorithm. That dataset should be large amount of raw data and it is the quantitative dataset.
2) **Build Dataset**
   Build dataset is used to count the number of rows and columns of the dataset and it gives output as Dataset Build Over, Dataset size: 60, Select number of Dimensions: 5726. This output depends upon the selected input file. The above specified numbers are example of one file.
3) **Normalize**
   Normalization on dataset process gives better outputs because normally dataset having some values, irrelevant values and multiclass problems. So dataset requires normalization.
4) **Select Measures**
   Distance measures

Distance measures formulas very often, especially when measuring the distance in the plain, we use the formula for the Euclidean distance. According to the Euclidean distance formula, the distance between two points in the plane with coordinates (x, y) and (a, b) is given by
$dist((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$

Manhattan distance

Definition: The distance between two points measured along axes at right angles. In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is |x1 - x2| + |y1 - y2|.
Next we are applying Relief+kNN algorithm on the dataset.

5) **Algorithms**

Relief Algorithm

- Take a data set with n instances of p features, belonging to two known classes.
- Within the data set, each feature should be scaled to the interval [0 1] (binary data should remain as 0 and 1).
- The algorithm will be repeated m times.
- Start with a p-long weight vector (W) of zeros. At each iteration, take the feature vector (X) belonging to one random instance, and the feature vectors of the instance closest to X (by Euclidean distance) from each class.
- The closest same-class instance is called 'near-hit', and the closest different-class instance is called 'near-miss'. Update the weight vector such that
  $W_i=W_{i-1}-(x_i\text{-nearHit}_i)^2 + (x_i\text{-nearMiss}_i)^2$
- Thus the weight of any given feature decreases if it differs from that feature in nearby instances of the same class more than nearby instances of the other class, and increases in the reverse case.
- After m iterations, divide each element of the weight vector by m. This becomes the relevance vector. Features are selected if their relevance is greater than a threshold τ.

Kira and Rendell's experiments showed a clear contrast between relevant and irrelevant features, allowing τ to be determined by inspection. However, it can also be determined by Cebysev's inequality for a given confidence level (α) that a τ of 1/sqrt(α*m) is good enough to make the probability of a Type I error less than α, although it is stated that τ can be much smaller than that.

### K-Nearest Neighbor Algorithm

KNN is an non parametric lazy learning algorithm. That is a pretty concise statement. When you say a technique is nonparametric, it means that it does not make any assumptions on the underlying data distribution. This is pretty useful, as in the real world , most of the practical data does not obey the typical theoretical assumptions made (eg: gaussian mixtures, linearly separable etc) . Non parametric algorithms like KNN come to the rescue here. It is also a lazy algorithm. What this means is that it does not use the training data points to do any *generalization*. In other words, there is no explicit training phase or it is very minimal. This means the training phase is pretty fast . Lack of generalization means that KNN keeps all the training data. More exactly, all the training data is needed during the testing phase. (Well this is an exaggeration, but not far from truth). This is in contrast to other techniques like SVM where you can discard all non support vectors without any problem. Most of the lazy algorithms – especially KNN – make decision based on the entire training data set. In k-NN classification, the input is a class membership. An object is classified by a majority vote of its neighbors, with the object being specified to the class most common among its k nearest neighbors. If k=1, then the object is simply specified to the class of that single nearest neighbor.

- In k-NN regression, the output is the value for the object. This value is the average of the values of its k nearest neighbours.

### 6) Execute

Execute gives the all values of object-object similarity in the form of cost matrix.

### 7) Do Cluster

Cluster gives better performance of algorithms. By taking the similarity threshold value.

### 8) Report

Finally Report gives total duration of clustering and number of clusters. There are depends on no of dimensions of dataset.

## 3. Implementation

Working of this paper deals in implementation. First take the input file from external source. This file consists the large amount of quantitative raw data. After that performing normalization operation on dataset. Because normalization shows all data into single scale measurement. Next perform the distance measure operations like Euclidean, Manhattan and cosine. After that performing relief+kNN and parzen+relief algorithm, it gives different set of relevant features and next perform the do cluster operation. Finally it gives clusters and cluster report.

## 4. Results

To run the project, first appear the below scenario

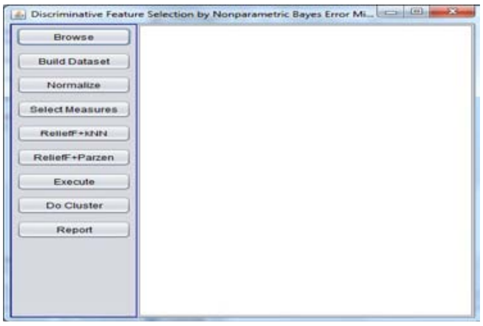
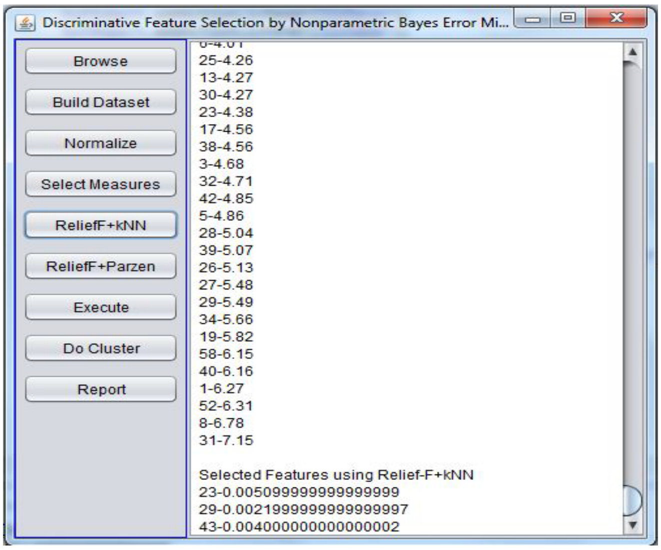


**Figure 4.1:** Browse

After appearing this window, select browse button to load the specified input file. Next applying some operations and then apply the relief+knn algorithm on the dataset. The below scenario will display



```
0-4.01
25-4.26
13-4.27
30-4.27
23-4.38
17-4.56
38-4.56
3-4.68
32-4.71
42-4.85
5-4.86
28-5.04
39-5.07
26-5.13
27-5.48
29-5.49
34-5.66
19-5.82
58-6.15
40-6.16
1-6.27
52-6.31
8-6.78
31-7.15

Selected Features using Relief-F+kNN
23-0.0050999999999999
29-0.0021999999999999997
43-0.0040000000000000002
```

**Figure 4.2:** Relif +kNN

After performing the do cluster operation, the output will appear as given bellow

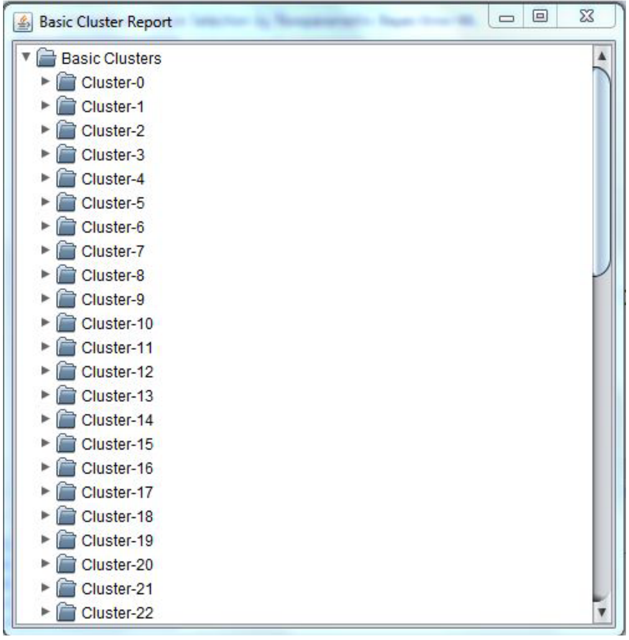
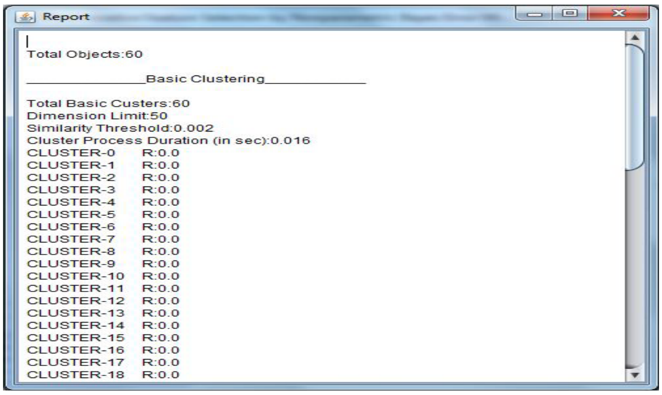


**Figure 4.3:** Basic Clustering

Finally the output of the gives the report, that report is



```
|
Total Objects:60

_________________Basic Clustering_______________

Total Basic Custers:60
Dimension Limit:50
Similarity Threshold:0.002
Cluster Process Duration (in sec):0.016
CLUSTER-0     R:0.0
CLUSTER-1     R:0.0
CLUSTER-2     R:0.0
CLUSTER-3     R:0.0
CLUSTER-4     R:0.0
CLUSTER-5     R:0.0
CLUSTER-6     R:0.0
CLUSTER-7     R:0.0
CLUSTER-8     R:0.0
CLUSTER-9     R:0.0
CLUSTER-10    R:0.0
CLUSTER-11    R:0.0
CLUSTER-12    R:0.0
CLUSTER-13    R:0.0
CLUSTER-14    R:0.0
CLUSTER-15    R:0.0
CLUSTER-16    R:0.0
CLUSTER-17    R:0.0
CLUSTER-18    R:0.0
```

**Figure 4.4:** Report

The graph shows, discrimination between relief+kNN and parzen+relief. In this we are showing the number of features selected by using two algorithms with respect to threshold value.

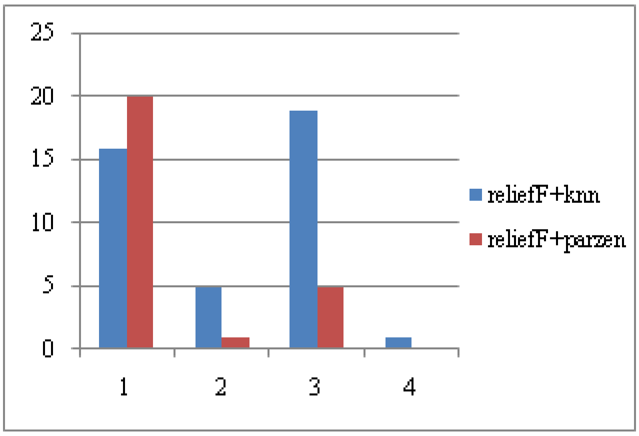


**Figure 4.5:** Relieff+Knn And Parzen+Relief Algorithm

It shows the cluster quality using relief+kNN algorithm.

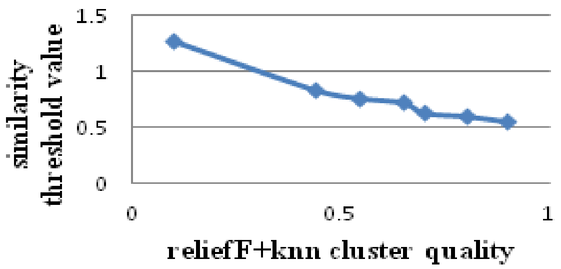


**Figure 4.6:** relief+kNN Algorithm

It shows the cluster quality using relief+parzen algorithm.

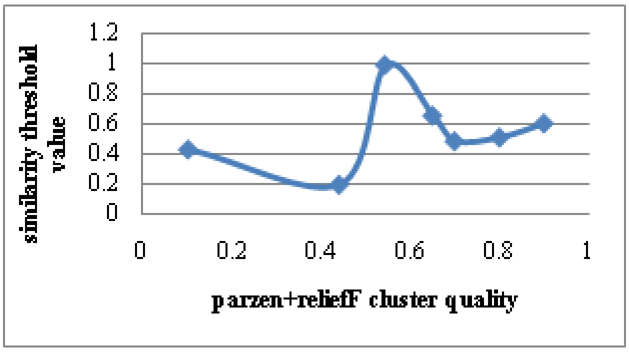


**Figure 4.7:** Relief+Parzen Algorithm

## Tables With Some More Results

In the below tables shows the values of given threshold values and selected features and weights of the belonging features with respect to the distance metrics like Euclidean, Manhattan, Cosine etc.

**Table 4.1:** Distance Metrics Euclidean, Manhattan, Cosine

| Threshold value | Selected features | Weights | | |
|---|---|---|---|---|
| | | Euclidean | Manhattan | Cosine |
| 0.002 | 1 | 0.204 | 0.116 | — |
| | 2 | 0.0194 | 0.0176 | 0.0072 |
| | 10 | 0.0197 | 0.0425 | 0.0536 |
| | 45 | 0.0184 | 0.0593 | 0.0844 |
| 0.003 | 10 | 0.0644 | 0.0595 | 0.0581 |
| | 12 | 0.0374 | 0.0253 | 0.0132 |
| | 13 | 0.06426 | 0.0504 | 0.0331 |
| | 14 | 0.00313 | — | — |
| | 23 | 0.0032 | — | 0.0048 |
| | 45 | 0.0849 | 0.0862 | 0.09123 |
| 0.004 | 10 | 0.0480 | 0.0532 | 0.0370 |
| | 13 | 0.0244 | 0.1186 | 0.1258 |
| | 45 | 0.1337 | 0.1573 | 0.1454 |

Here this table shows the threshold values and selected measures and features selected. Clearly it shows the different selected features w.r.t. distance measures.

**Table 4.3:** Threshold Values And Selected Measures And Features Selected.

| Threshold value | Selected measures | | | Features selected |
|---|---|---|---|---|
| | Euclidean | Manhattan | Cosine | |
| 0.002 | | | | 8 |
| | | | | 4 |
| | | | | 6 |
| 0.003 | | | | 6 |
| | | | | 4 |
| | | | | 5 |
| 0.004 | | | | 3 |
| | | | | 5 |
| | | | | 5 |
| 0.0076 | | | | 21 |
| | | | | 14 |
| | | | | 12 |
| 0.0083 | | | | 26 |
| | | | | 16 |
| | | | | 25 |
| 0.0543 | | | | 1 |
| | | | | 13 |
| | | | | 12 |
| 0.0005 | | | | 27 |
| | | | | 25 |
| | | | | 25 |

This table shows the cluster quality by giving the similarity threshold value for separate the relevant features into similar groups.

**Table 4.4:** Similarity Threshold Value For Separate The Relevant Features

| Threshold for selection of useful features | Similarity threshold value | Cluster quality | |
|---|---|---|---|
| | | relief+knn | relief+parzen |
| 0.1 | 0.90 | 0.5558 | 0.6066 |
| 0.12 | 0.80 | 0.6032 | 0.5131 |
| 0.203 | 0.70 | 0.6319 | 0.4881 |
| 0.180 | 0.65 | 0.7237 | 0.6579 |
| 0.254 | 0.543 | 0.7040 | 0 |
| 0.30 | 0.44 | 0.8371 | 0.10 |
| 0.497 | 0.10 | 1.2744 | 0.09 |

## 5. Discussion

In this paper we are applying algorithms on discontinuous raw data. This paper mainly uses the combinations of Relief and kNN algorithms. The output of this paper is clustered data. There is a scope for performing same operations on continuous data, that is another scenario.

## 6. Conclusion

In this work, comparing two feature weighting algorithms. So the selected relevant features are showing in clusters by using some clustering algorithms for better validation, limitations of the well known clustering techniques for large datasets and details of the proposed clustering method, Leaders-Subleaders have been presented. Our experimental results on numerical datasets show that Leaders-Subleaders algorithm performs well. The representatives of the Subleaders help in improving the classification accuracy. Devies-Bouldin index showed a good performance to the results were equivalent, even with different radius.

## References

[1] Guangtao Wang, Qinbao Song, Heli Sun, Xueying Zhang, "A Feature Subset Selection Algorithm Automatic Recommendation Method." Journal of Artificial Intelligence Research 47 (2013) 1-34 Submitted 10/2012; published 05/2013

[2] Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emani Chukanlo, "A Survey of Hierarchical Clustering Algorithms." The Journal of Mathematics and Computer Science Vol .5 No.3 (2012) 229-240

[3] Yogita Rani, Dr.Harish Rohil, "A Study of Hierarchical Clustering Algorithm." International Journal of Information and Computation Technology. ISSN 0974- 2239 Volume 3, Number 11 (2013), pp. 1225-1232 © International Research Publications House

[4] N.Sudhakar Reddy, KVN Sunitha, "Efficient Clustering Algorithm for Large Dataset." Volume 2, Issue 1, January 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[5] Ankita Choubey, Dr.Sadhna K.Mishra, "Enhanced Clustering Algorithm for Processing Online Data" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278- 8727Volume 12, Issue 5 (Jul. - Aug. 2013), PP 24-29

[6] Srinivasulu M, Kotilinswara Rao, Baji Mohammed, RVSP Kumar, Data Clustering With Leaders and Subleaders Algorithm." IOSR Journal of Engineering (IOSRJEN) e- ISSN: 2250-3021, p-ISSN: 2278-8719, Volume 2, Issue 11 (November2012), PP 01-07

[7] Kagan Tumer, Joydeep Ghosh, "Bayes Error Rate Estimation Using Classifier Ensembles"

[8] Ali Mustafa Qamar, Eric Gaussier, "RELIEF Algorithm and Similarity for k-NN". International Journal of Computer Information Systems and Industrial Management Applications. ISSN 2150-7988 Volume 4 (2012) pp. 445-458 MIR Labs.

[9] Yijun Sun, Jian Li, "Iterative RELIEF for Feature Weighting." Interdisciplinary Center for Biotechnology Research,Department of Electrical & Computer Engineering, University of Florida, Gainesville, FL 32611, USA

[10] Marko Robnik-Sikonja, Igar Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF." University of Ljubljana, Faculty of Computer and Information Science, Trˇzaˇska 25, 1001 Ljubljana, Slovenia

[11] Shuang-Hong Yang and Bao-Gang Hu, Senior Member, IEEE, "Discriminative Feature Selection by Nonparametric Bayes Error Minimization."IEEE Transactions on knowledge and data engineering, vol.24, no.8, August 2012

[12] Kenji Kira, Larry A.Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm." From: AAAI-92 Proceedings. Copyright ©1992, AAAI. All rights reserved.

[13] Yuxuan SUN, Xiaojun LOU, Bisai BAO, "A Noval Relief Feature Selection Algorithm Based on Mean-Variance Model." Journal of Information & Computational Science 8: 16 (2011) 3921–3929.

[14] K.Saranya, T.Deepa, "Discriminative Clustering Based Feature Selection and Nonparametric Bayes Error Minimization and Support Vector Machine(SVM)." International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.

## Author Profile

**Mr. Choudari Anantha** is currently pursuing his M.Tech(SE) in Department of Computer Science and Engineering at Vignan Institute of Information Technology, Visakhapatnam. He is Received B.Tech (IT) From Swarnandhra College of Engineering and Technology, Narsapuram, West Godavari; A.P. His area of interests includes Data Mining and Information Security.

**Dr. Devarapalli Dharmaiah** is currently working as an Associate Professor in Computer Science and Engineering department, Vignan Institute of Information Technology, Visakhapatnam and the teaching and research experience of about 11 years. He also guided various dissertation works for both UG and PG students of VIGNAN'S IIT and other Colleges. He has taught various subjects of Computer Science and Engineering for both of UG & PG students such as C programming, Data Structures, Java, Operating Systems, Compler Design, Linux, UNIX, Data Structures and Bioinformatics. He has a very good reputation among the students and faculty community for his proficiency in subjects. He is a life member of CSI. He has published many papers in National, International conferences and leading International Journals. His area of interest is Bioinformatics, Neural Networks, Data mining and computer networks. He received Best Engineering Teachers award 2013, PRATIBHA AWARD, Sasatra Award, Best Project Award in Computer Science and Engineering Department.