IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 13, Issue 3 (Jul. - Aug. 2013), PP 73-79
www.iosrjournals.org

A Modified KS-test for Feature Selection

K Subrahmanyam[1], Nalla Shiva Sankar[1], Sai Praveen Baggam[2],
Raghavendra Rao S[3]

Abstract: *A central problem in machine learning is identifying a representative set of features from which to construct a classification model for a particular task. Feature selection, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. The central hypothesis is that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other. In this paper a fast redundancy removal filter is proposed based on modified Kolmogorov-Smirnov statistic, utilizing class label information while comparing feature pairs. Results obtained from this algorithm are compared with other two algorithms capable of removing irrelevancy and redundancy, such as Correlation Feature Selection algorithm (CFS) and simple Kolmogorov Smirnov-Correlation Based Filter (KS-CBF).*

The efficiency and effectiveness of various methods is tested with two of the standard classifiers such as the Decision- Tree classifier and the K-NN classifier. In most cases, classification accuracy using the reduced feature set produced using the proposed approach equaled or bettered accuracy obtained using the complete feature set and other two algorithms.

I. Introduction

Feature selection, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. It is a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion. In recent years, data has become increasingly larger in both number of instances and number of features in many applications such as genome projects, text categorization, image retrieval, and customer relationship management. This enormity may cause serious problems to many machine learning algorithms with respect to scalability and learning performance. For example, high dimensional data (i.e., data sets with hundreds or thousands of features) can contain high degree of irrelevant and redundant information which may greatly degrade the performance of learning algorithms. Therefore, feature selection becomes very necessary for machine learning tasks when facing high dimensional data nowadays.

Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. Feature selection evaluation methods fall into two broad categories, Filter model and Wrapper model [2].The Filter model relies on general characteristics of the training data to select some features without involving any learning algorithm. The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected. As for each new subset of features, the wrapper model needs to learn a hypothesis (or a classifier). It tends to find features better suited to the predetermined learning algorithm resulting in superior learning performance, but it also tends to be more computationally expensive and less generality than the Filter model. When the number of features becomes very large, the Filter model is usually chosen due to its computational efficiency. Filters have the advantage of fast execution and generality to a large family of classifiers than wrappers [13].

Figure 1 provides a depiction of a simple classification process where a Feature Selection process that uses a filter is involved. The training and testing datasets after the dimensionality reduction process is fed to the ML (Machine Learning) algorithm. In some cases, accuracy on future classification can be improved; in others, the result is a more compact, easily interpreted representation of the target concept. In this work, we have employed a Filter model for the evaluation of features selected.

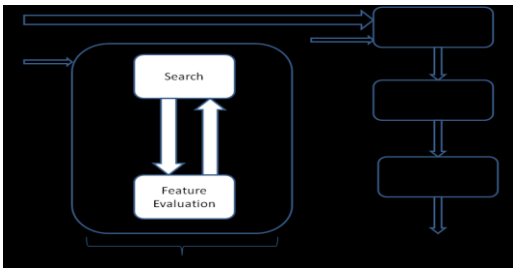


Figure 1: Classification Process that involves Feature Selection stage with Filter approach

In the following sections, we will be describing about various approaches used in our work along with our proposed approach. Other approaches exist such as Rank wrapper algorithm [6], Relief algorithm [.Following it, the datasets used and results obtained are mentioned.

II. Theoretical Framework

2.1. Correlation Based Feature Selection:

This algorithm [1] is based on information theory and uses symmetrical uncertainty (SU) as the filter for the evaluation of the feature set selected. This algorithm involves certain concepts such as mutual information [3], entropy, information gain and symmetrical uncertainty. The process used here in finding correlations between various attributes is different from that used in FCBF [11].In the first step, it processes the given training dataset and initial feature set and removes all the irrelevant features by finding the strength of prediction of feature-to-class. In the second step, it uses this relevant feature set and training dataset to remove all the redundant features and finally presents the significant feature set that is well supervised and uncorrelated with other features.

Entropy: Entropy as given by Shannon is a measure of the amount of uncertainty about a source of messages [5]. The entropy of variable Y before and after observing values of another variable X can be described by:

$$H(Y) = - \sum p(yi) \log (p(yi))$$

And

$$H(Y/X) = - \sum p(xj) \sum p(yi/xj) \log (p(yi/xj))$$

Here p (yi) is the prior probabilities for all values of random variable *Y* and p (yi|xj) is the conditional probability of y*i* given *xj*. By treating *Y* as classes and *X* as features in a data set, the entropy is 0, i.e., without any uncertainty at all if all members of a feature belong to the same class. On the other hand, members in a feature set are totally random to a class if the value of entropy is 1. The range of entropy is between 0 and 1.

Information Gain:

The amount by which the entropy of X decreases reflects additional information about Y provided by X and is called information gain, given by

$$\text{Gain, } I(Y; X) = H(Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$
$$= H(Y) + H(X) - H(X, Y).$$

However, information gain is biased if feature with more values[4], which the features with greater numbers of values will gain more information than those with fewer values even if the former ones are actually less informative than the latter ones. Also, the range of information gain is not from 0 to 1. Its values should be normalized in order to ensure they are comparable and have the same affect.

Symmetrical Uncertainty:

Because of the limitation provided by the usage of Information gain, we use another heuristic called Symmetrical Uncertainty and is given by:

$$SU(Y; X) = 2[I(Y; X) / (H(X) + H(Y))]$$

It averages the values of two uncertainty variables, compensates for information gain's bias toward features with more values, and normalizes its values to the range [0, 1]. A value of 1 indicates that knowing the value of either one completely predicts the value of the other and a value of 0 indicates that *X* and *Y* are independent each other.

Algorithm 1:
Table 1: Correlation Feature Selection Algorithm

1. //Remove irrelevant features
2. Input original data set D that includes features X and target class Y
3. For each feature Xi
   Calculate mutual information SU(Y; Xi)
4. Sort SU(Y; Xi) in descending order
5. Put Xj whose SU(Y; Xi)>0 into relevant feature set Rxy
6. //Remove redundant features
7. Input relevant features set Rxy
8. For each feature Xj
   Calculate pair wise mutual information SU (Xj; Xk) for all j≠k
9. Sxy= ∑ (SU (Xj; Xk))
10. Calculate means μR and μS of Rxy and Sxy, respectively
    W= μS/μR
11. R=W. Rxy- Sxy
12. Select Xj whose R>0 into final set F

2.2. Kolmogorov-Smirnov Test:

Equivalence of two random variables may be evaluated using the Kolmogorov-Smirnov (KS) test [12]. In statistics, the Kolmogorov–Smirnov test (K–S test) is a nonparametric test for the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test).

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution (in the two-sample case) or that the sample is drawn from the reference distribution (in the one-sample case).

The **empirical distribution function** F_n for n independent and identical observations X_i is defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}$$

where $I_{X_i \le x}$ is the indicator function, equal to 1 if $X_i \le x$ and equal to 0 otherwise.

The **Kolmogorov–Smirnov statistic** for a given cumulative distribution function F(x) is

$$D_n = \sup_x |F_n(x) - F(x)|$$

where sup x is the supremum of the set of distances.

The cumulative distribution function of K is given by

$$\Pr(K \le x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty}e^{-(2i-1)^2\pi^2/(8x^2)}.$$

Under null hypothesis that the sample comes from the hypothesized distribution F(x),

$$\sqrt{n}D_n \xrightarrow{n\to\infty} \sup_t |B(F(t))|$$

in distribution, where B(t) is the Brownian bridge.

If F is continuous then under the null hypothesis $\sqrt{n}D_n$ converges to the Kolmogorov distribution, which does not depend on F. This result may also be known as the **Kolmogorov theorem**; the goodness-of-fit test or the **Kolmogorov–Smirnov test** is constructed by using the critical values of the Kolmogorov distribution. The null hypothesis is rejected at level α if

$$\sqrt{n}D_n > K_\alpha,$$

where K_α is found from

$$\Pr(K \le K_\alpha) = 1 - \alpha.$$

The asymptotic power of this test is 1.The various test parameters for the KS-test and minimum estimations are given in [9] and [10].

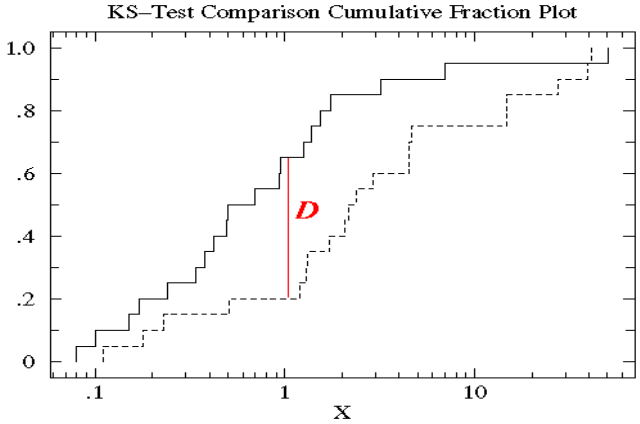
We will be using the KS two-sample test for analyzing the redundancy of two features. So, a brief description about it is given.

Two-Sample Kolmogorov Smirnov Test:

The Kolmogorov–Smirnov test may also be used to test whether two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov–Smirnov statistic is

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

where $F_{1,n}$ and $F_{2,n'}$ are the empirical distribution functions of the first and the second sample respectively.



The null hypothesis is rejected at level α if

$$\sqrt{\frac{nn'}{n+n'}} D_{n,n'} > K_\alpha.$$

It should be noted here that the two-sample test checks whether the two data samples come from the same distribution. This does not specify what that common distribution is (e.g. normal or not normal).

KS test
Steps taken to calculate KS include: 1. Discretization of both features Fi, Fj into k bins. 2. Estimation of probabilities in each bin. 2.1. Calculation of cumulative probability distributions for both features. 2.2. Calculate KS statistic

Table 2: KS test for two features Fi, Fj

Algorithm 2:

Algorithm K-S CBF
Relevance analysis 1. Calculate the *SU*(*X,C*) relevance indices and create an ordered list *S* of features according to the decreasing value of their relevance. **Redundancy analysis** 2. Take as the feature *X* the first feature from the *S* list 3. Find and remove all features for which *X* is approximately equivalent according to the K-S test 4. Set the next remaining feature in the list as *X* and repeat step 3 for all features that follow it in the *S* list.

Table 3: A two-step Kolmogorov-Smirnov Correlation Based Filter (K-S CBF) algorithm.

III. Modified KS-CBF Test:

The proposed algorithm introduces the concept of binning the input dataset into n bins. Redundancy is calculated in each of the bins individually and the set of redundant features for each bin are stored. Finally, the set of features that are common in all the bins (here, it can also be taken as an input parameter to decide at run-time as required) are considered as redundant for the input dataset and they can be eliminated.

In each bin, again we divide the available set of records into a particular number of partitions for which the actual KS-test is applied. The set of redundant features in each of these partitions is mixed up with those for the other partitions in that bin. So, now we can expect a good redundant subset to be produced from each bin. The union operation ensures that the possibility of redundancy has been checked for every feature in its entirety. The intersection operation ensures that a feature which is actually non-redundant will not be claimed as being redundant. Now, since we perform an intersection of the redundant features obtained from all the bins, the final

feature subset produced will not contain these features and could sufficiently represent a significant subset of features that can be used for the classification process.

Algorithm 3:
Modified K-S CBF Algorithm:

Relevance analysis
1. Order features based on decreasing value of SUC (f, C) index which reflects their decreasing value of their relevance.

Redundancy analysis
2. Pass the dataset with relevant features for KS test measure.
3. Discretize the dataset into n bins each containing approximately same number of records.
4. For each bin B_i
    4.1. Form k data partitions each approximately containing the same number of records.
    4.2. For each of the k partitions $P_1, P_2, \ldots, P_k$
        4.2.1. Initialize Fi with the first feature in the F-list.
        4.2.2. Find all features for which Fi forms an approximate redundant cover using K-S test.
        4.2.3. Set the next remaining feature in the list as Fi and repeat above step for all features that follow it in the F list.
    4.3. Take the union of all these redundant features into Bi
5. Get the common features that are redundant in all bins.
6. Remove those features and get the significant subset of features.

Table 4: A modified Kolmogorov-Smirnov Correlation Based Filter algorithm.

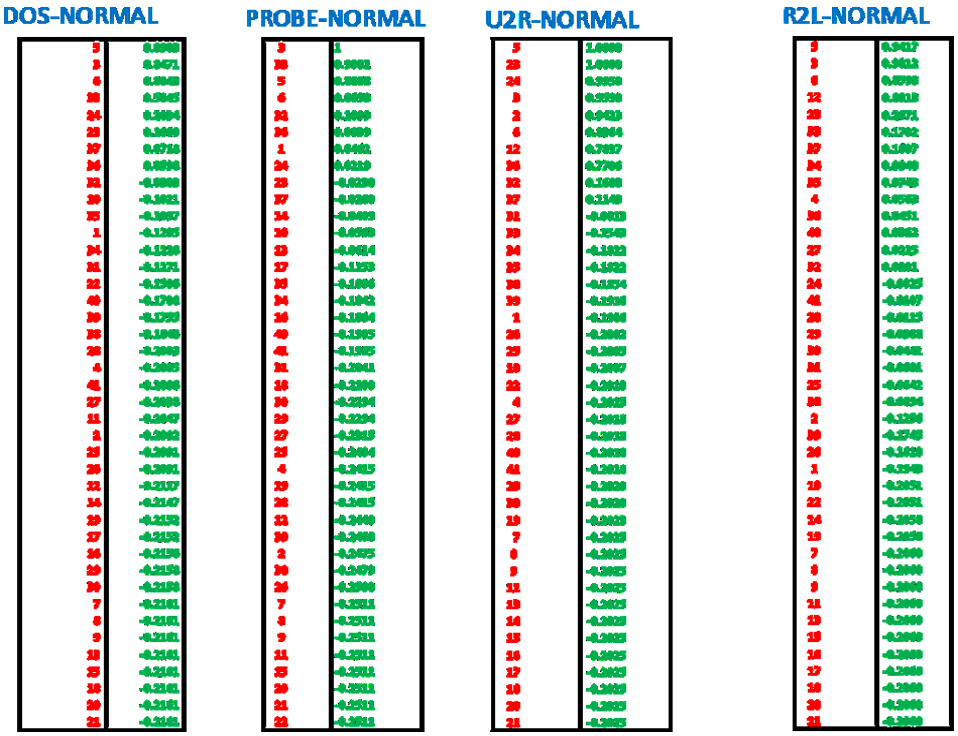
Dataset	No. Features	No. Instances	No. Classes	Class Distribution	Balanced Dataset
Ionosphere	34	351	2	126/225	No
Pima	8	768	2	500/268	No
Wdbc	31	569	2	212/357	No
Wine	13	178	2	59/71	No
Dos-Normal	41	4765	2	2371/2394	No
Probe-Normal					
U2R-Normal	41	4346	2	1996/2350	No
R2L-Normal					
	41	326	2	52/274	No
	41	2137	2	1062/1075	No

In the first step, a Symmetrical Uncertainity filter has been applied to remove the irrelevant features. The dataset containing this filtered subset is now passed to the second step to perform the redundancy analysis and obtain the set of redundant features. These redundant features are removed from the remaining feature set and finally the dataset with significant features is only considered for further analysis.

In most cases, classification accuracy using this reduced feature set produced equaled or bettered accuracy using the complete feature set. However, the dimensionality of the feature set has been reduced to a better extent then compared to the simple KS-CBF algorithm. This can give a good computational gain over the simple KS-CBF when testing with a classifier as the new feature set contains less number of dimensions. In few cases, it is however observed that the proposed algorithm has been producing results that are quite less efficient when tested with a classifier than that compared with the simple KS-CBF results. However, this can still be handled and improved by varying the values of number of bins which is taken as an input parameter. Feature selection can sometimes degrade machine learning performance in cases where some features were eliminated which were highly predictive of very small areas of the instance space. The size of the dataset also plays a role in this and as we are further dividing into smaller bins, it sometimes affects the process when there are no sufficient records in a bin to perform the test. Also, when the information available in the dataset is very random with a large range of distinct feature values, then also this algorithm could produce very good results than others.

It has been observed that, as the number of bins during the test is increased, the number of redundant features is increasing and the final feature set produced is getting smaller. However, in some cases, this has been leading to a slight decrease in detection rates. Yet, the performance gain obtained is very much high. So, the effect of slight decrease in detection rates can be negotiated with the rapid increase in performance.

IV. Datasets and Results:



The results obtained with the above features are tested with two of the standard classifiers- Decision Tree and K-NN classifier.

Table 8.1 : Selected Features of KDD 99 datasets

Data Set	No: of features	Correlation Feature Selection	Simple KS-test	Modified KS-test
Normal-DoS	41	1,5,10,23,24,25,,31,33,35,38,39	2-6, 12, 23, 24, 31-37,	2-6,12,23,32,36
Normal-Probe	41	4,24,27-32,40,41	3-6, 12, 23, 24, 27, 32-37, 40	3-6,12,23,32,33,37
Normal-U2R	41	1,10,13,14,17,23,35-37	1,3,5,6,32-34,24,36,37	1,3,5,6,24,32-34,36
Normal-R2L	41	1,10,22,23,31,32,34-37	1,3,5,6,10,22-24,31-37	2-6,24,32,33,36,37

Table 8.2 : Selected feature sets for various UCI datasets

Dataset	No: of features	Correlation Feature Selection	Simple KS-test	Modified KS-test
Ionosphere	34	6,12,22,24,25,27,29,32-34	4,25,28	4,25,28
Pima	8	2,5,6,8	2-8	2,6,7
Wdbc	31	8,9,11,18,21,27-31,	1,4,5,8,9,12,14-18,21,24,25,28	1,4,5,8,9,14-16,18,19,21-23,28
Wine	14	1,2,6,9,10,12	1,2,7,10,12,13	1,2,7,10,12,13

Table 8.3 : Detection Rates with various feature subsets for KDD99 dataset

Data Set	K-NN Classifier				Decision Tree			
	Full Set	Correlation Feature Selection	Simple KS-test	Modified KS-test	Full Set	Correlation Feature Selection	Simple KS-test	Modified KS-test
Normal-DoS	99.92	99.55	99.73	99.39	99.40	99.73	99.41	99.41
Normal-Probe	97.37	91.5	97.28	97.28	100	90.44	100	100
Normal-U2R	96.80	92.0	93.89	99.20	100	95.2	100	100
Normal-R2L	99.28	83.73	98.40	98.92	97.72	94.2	97.37	98.95

Table 8.4 : Detection Rates with various feature subsets for UCI datasets

Data Set	K-NN Classifier				Decision Tree			
	Full Set	Correlation Feature Selection	Simple KS-test	Modified KS-test	Full Set	Correlation Feature Selection	Simple KS-test	Modified KS-test
Ionosphere	80.82	87.67	84.24	84.24	86.98	84.24	69.18	69.36
Pima	74.02	75.32	71.42	75.65	59.7	54.5	50.0	56.49
Wdbc	67.36	92.05	67.36	67.78	85.77	85.35	93.31	93.61
Wine	95.55	99.77	96.65	97.77	99.33	100	97.78	97.78

It can be seen that the proposed approach selects a less number of significant feature subset in many cases than the other two algorithms. Also, the accuracy and efficiency of this approach was much better in most of the cases. In few cases, the accuracy slightly reduced but it is not that much far away from the other methods and this method outperformed both the CFS and KS-test in terms of efficiency i.e. in terms of execution performance. As compared to the results in [4] and [6], the results obtained in our experiment are good and encouraging.

V. Conclusion:

A modified KS-test which can efficiently select good feature subsets has been proposed in this paper. The proposed algorithm has the computational demands that are very much similar to the traditional KS-test and is proportional to the total number of bins. A comparative test with some widely used feature selection algorithms showed its better performance both in terms of efficiency and accuracy. The statistical significance of 0.05 has been used in test which can be varied. The number of bins and number of partitions should be given depending on the number of records in the incoming dataset. According to our observations, good results will be obtained if a bin is made to contain atleast 40 records.

Since a filter approach is used, the results can be well suited to any classifier process. The results in our experiment have been tested with C4.5 and K-NN classifier. Various variants of the Kolmogorov-Smirnov test exist and the algorithm may be used with other indices for relevance indication.

Future Work:

In future we would like to compare and contrast the ability of the modified KS-Test against the features obtained using the rough set theory, Support Vector Machines, and Decision Trees. We have identified the problem of Intrusion Detection as our subject matter and we will apply these techniques using GANN weight extraction algorithm [14] to find the accuracy of the Intrusion Detection Systems based on these techniques. Further we would be developing an extension for SNORT using the best of these four techniques according to our findings.

References:

[1]. Te-Shun Chou, Kang K.Yen, JunLuo, NikiPissinou and KiaMakki Correlation Based Feature Selection for Intrusion Detection Design.

[2]. Feature Subset Selection: A Correlation Based Filter Approach Mark A. Hall, Lloyd A. Smith ([mhall, las]@cs.waikato.ac.nz)Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[3]. A Novel Information Theory for Filter Feature Selection BoyenBonov, FransiscoEscolonov and MiguelAnguel.

[4]. Ensemble Fuzzy Belief Intrusion Detection Design AT Florida International University Miami, Florida BY Te-Shun Chou.

[5]. Conditional Entropy Metrics for Feature Selection by Iain Bancarz.

[6]. Feature Selection for Supervised Classification: A Kolmogorov-Smirnov Class Correlation-Based Filter Marcin Blachnik, Włodzisław Duch, Adam Kachel, Jacek Biesiada.

[7]. J. Biesiada and W. Duch, "Feature Selection for High-Dimensional Data: A Kolmogorov-Smirnov Correlation-Based Filter Solution," Proceedings of the 4th International Conference on Computer Recognition Systems.

[8]. Data Mining: Concepts and Techniques, 2nd edition: Jiawei Han and Micheline Kamber.

[9]. The Kolmogorov-Smirnov Test When Parameters are estimated from data: Hovhannes Keutelian, Fermilab.

[10]. Minimum Kolmogorov-Smirnov test Statistic Parameter Estimates: Michael D:Weber, Lawrence M.Leemis and Rex K.Kincaid .

[11]. L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proceedings of the Twentieth International Conference on Machine Leaning, pp. 856-863, Washington, D.C.

[12]. Feature Selection For High Dimensional Data: A Kolmogorov-Smirnov Correlation-Based Filter: Jacek Biesiada and Wlodzislaw Duch

[13]. Correlation Based Feature Selection For Machine Learning: A Dissertation work by Mark A.Hall, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

[14]. Artificial Neural Network Classifier for Intrusion Detection in Computer Networks,ICN-2010,Karnataka, India.