International Journal of Engineering

Journal Homepage: www.ije.ir

TECHNICAL NOTE

A New Hybrid Framework for Filter based Feature Selection using Information Gain and Symmetric Uncertainty

P. Kalpana*, K. Mani

Department of Computer Science, Nehru Memorial College, Puthanampatti, Tiruchirappalli-Dt,Tamil Nadu, India

PAPER INFO

Paper history:
Received 29 November 2016
Received in revised form 12 February 2017
Accepted 10 March 2017

Keywords:
Irrelevant Redundant
Median Based Discretization
Information Gain
Symmetric Uncertainty
Accuracy
Naive Bayesian Classifier

ABSTRACT

Feature selection is a pre-processing technique used for eliminating the irrelevant and redundant features which results in enhancing the performance of the classifiers. When a dataset contains more irrelevant and redundant features, it fails to increase the accuracy and also reduces the performance of the classifiers. To avoid them, this paper presents a new hybrid feature selection method using information gain and symmetric uncertainty. The proposed work uses median based discretization for converting the quantitative features into qualitative one, information gain in finding the relevant features and symmetric uncertainty to remove the redundant features. As the proposed work uses both relevance and redundant analyses the predictive accuracy of the Naive Bayesian classifier has been improved. Further the efficiency and effectiveness of the proposed methodology is analyzed by comparing with other existing methods using real-world datasets of high dimensionality.

doi: *10.5829/idosi.ije.2017.30.05b.05*

1. INTRODUCTION

Many factors affect the success of the data mining and machine learning algorithms. Among them, quality of data is the most prominent one. So, it is necessary to make the data as qualitative as possible. The data is said to be of high quality if it does not contain any irrelevant, redundant, unreliable and noisy elements which can be achieved using Feature Selection (FS). In FS methods, a subset of the features which are supposed to have a high impact are selected [1]. It is the most common pre-processing method [2] and is classified into three major types viz., filters, wrappers and embedded. Filter model depends on the characteristics of the data such as consistency, distance, dependency, information and correlation and it is independent of the learning algorithm [3]. Wrapper model uses the specific learning algorithm itself to assess the quality of the selected features [4]. Embedded model uses a learning algorithm which performs FS in the process of training [5]. The filter approach is computationally more efficient but ignores the fact that, the feature selection may depend on the learning algorithm. On the other hand, the wrapper method is computationally more demanding but it takes dependencies of the feature subset on the learning algorithm into account.

The definition of feature relevance divides the features into strongly relevant, weakly relevant and irrelevant. Removing the irrelevant and weakly relevant features in the original feature space considerably enhances the performance of the classifier in terms of time and accuracy. To remove the irrelevant and weakly relevant features, this paper proposes a filter based FS method which uses the entropy based measure called Information Gain (*IG*). It is noted that among the relevant features, some may be redundant because *IG* cannot handle the redundant features. As the data mining process becomes very inefficient and complex due to feature redundancy, it is also essential to remove them. Thus, this work combines the redundant analysis for removing the redundant features using Symmetric Uncertainty (*SU*).

*Corresponding Author's Email: parasuramankalpana@gmail.com (P. Kalpana)

Please cite this article as: P. Kalpana, K. Mani, A New Hybrid Framework for Filter based Feature Selection using Information Gain and Symmetric Uncertainty, International Journal of Engineering (IJE), TRANSACTIONS B: Applications Vol. 30, No. 5, (May 2017) 659-667

2. REVIEW OF LITERATURE

Research in improving accuracy of the classifier is quite common issue and is in high demand today. An important problem related to mining large data sets both in dimension and size is of selecting a subset of original features. FS methods have been applied to classification problems to select a reduced feature set that makes the classifier faster and more accurate. Several feature selection algorithms have been proposed in the literature and this section presents a brief overview of them and provides a stronger lead to the proposed work.

Hall [6] has introduced a new approach to FS called CFS. First, it computes feature-class correlation, then feature-feature correlation and finally selects the feature subset using best first search. For discrete class problems, the measure SU has been used and proved that it outperforms the well known algorithm ReliefF. Yu and Liu [5] have introduced a new framework for FS called Fast Correlation Based Filter (FCBF), which combines both feature relevance and redundant analysis. It uses SU for both relevance and redundant analysis and achieved a high degree of dimensionality reduction. Further, the predictive accuracy of the selected features is enhanced.

Haindl et al. [7] have introduced a Mutual Correlation based FS method (FSBMC). It consists of two phases. In the first phase, the algorithm computes the mutual correlation between each pair of features in the original feature space *S*. In the second step, the average absolute mutual correlations for each feature with all other features are computed and removed the feature which is mostly correlated with others i.e., the feature which has the largest mean mutual correlation. This process is repeated until the desired size of features is obtained. The removed feature is also discarded from the remaining average mutual correlation for the next iteration [8]. However, it does not consider the continuous features and the feature relevance.

A novel filter has been developed by Biesiada and Duch [9], which consists of both the relevance and redundancy analysis. A very similar process to the FCBF algorithm is employed. The only difference is that it uses Pearson Chi-Square function for redundant analysis. It was proved that the new algorithm works well with linear SVM classifier and also analyzed that it is similar to other correlation measures and much lower than relief F.

Senliol et al. [10] have introduced a new FS called FCBF$^{\#}$ which enables FCBF to select a given size of feature subset in different order than FCBF and proved that the algorithm results in more accurate classifiers. Peter and Somasundaram [11] have introduced a novel FS by joining FCBF and Bayes Theorem. In this method, the Bayes theorem reduces the features returned by FCBF using the conditional probability for each attribute i.e., the attribute which has highest conditional probability is selected and proved that the new algorithm provides better accuracy than the traditional ones.

Mani and Kalpana [12] have introduced a FS method using IG with a novel unsupervised discretization method called Median Based Discretization (MBD) for continuous features and proved that the IG_{MBD} provides more accuracy for Naive Bayesian Classifier (NBC) for the selected features. It has also been proved that the IG_{MBD} selects more relevant features than IG with EWID, IG with EFID and IG with CBD because the accuracy of the selected features using IG_{MBD} is more than that of the others. However, the feature redundancy is not considered.

After an extensive review of literature, it has been identified that the FS methods which incorporates both relevance and redundant analysis enhances the classification accuracy [13]. Further, it is noted that a filter by combining IG and SU is not proposed so far. Thus, this paper has introduced a new hybrid FS method using IG_{MBD} for relevance analysis and SU for redundant analysis. The relevance analysis is performed in a supervised way since it uses the information of the class. The redundant analysis is carried out in an unsupervised manner because during the reduction of redundant features it does not use the class information.

3. MATHEMATICAL BACKGROUND

This section illustrates the mathematical background required to understand the proposed method. It uses the correlation based measures to determine the goodness of features. A feature is said to be good, if it is highly correlated to class and not redundant to any of the other relevant features. To find the correlation between two random variables, there are two basic approaches available viz., classical linear correlation and information theoretic correlation. As the real world datasets contain both numerical and non-numerical data, this work uses the information theoretic correlation.

3. 1. Information Gain It is a univariate and entropy-based FS method based on Claude Shannon on information theory and it determines the feature relevance between the attribute (*X*) and class label (*C)* in a supervised way [5] and it is calculated as:

$$IG(X,Y) = H(X) - H(X \mid Y) \tag{1}$$

where:
H(X) is the entropy of *X* and it is computed as:

$$H(X) = -\sum_{i} P(x_i)\log_2(P(x_i)) \tag{2}$$

H(X|Y) is the entropy of *X* after observing *Y* and it is computed as:

$$H(X \mid Y) = -\sum_j P(y_j) \sum_i P(x_i \mid y_j) \log_2 (P(x_i \mid y_j)) \quad (3)$$

Here,
$P(x_i)$ is the prior probability for all values of X and $P(x_i|y_i)$ is the posterior probability of X given the values of Y.

3. 2. Symmetric Uncertainty It is normalized information theoretic correlation measure used to determine the dependency of features and it is computed as:

$$SU = 2 \times \frac{IG(X \mid Y)}{H(X) + H(Y)} \quad (4)$$

SU takes the normalized value in the range [0, 1] due to the correlation factor 2. If *SU=1*, then X and Y are said to be highly correlated which means that the knowledge of one attribute completely predicts the other. If *SU=0* then X and Y are uncorrelated [5, 14].

4. CLASSIFICATION TECHNIQUES

The classification process consists of three phases viz., Model Construction (Learning), Model Evaluation (Testing) and Model Use (Classification) [2]. In the first phase, learning is performed on training data, which associates the class information and the classifier is built with the training set. In the second phase, the predictive accuracy of the model is computed using the test data and the final phase classifies the new instances [3, 6].

4. 1. Naive Bayesian Classifier It is a statistical classifier based on the Bayes theorem. Let D be a set of training tuples, each associates the class label. Suppose there are m classes, $C_1, C_2, \ldots, C_m$. The role of this classifier is to predict that the given tuple X belongs to the class having the highest posterior probability contained on X [15]. i.e., the tuple X belongs to C_i iff $P(C_i \mid X) > P(C_j \mid X)$ for $1 \le j \le m$ and $j \ne i$. $P(C_i \mid X)$ is computed as

$$P(C_i \mid X) = \frac{P(X \mid C_i) P(C_i)}{P(X)} \quad (5)$$

The accuracy is the most common measure used to evaluate the performance of a classifier which is the percentage of records that are classified correctly by the classifier in the test data [1, 16].

5. PROPOSED METHODOLOGY

It consists of four phases. Phase 1 is used to convert the continuous attributes into discrete using Median Based Discretization (*MBD*). Phase 2 computes the feature-class correlation using *IG* and selects the more relevant features if its *IG* is more than the threshold value where median of *IG*'s is considered as threshold value. Phase 3 computes the feature-feature correlation using *SU* for the selected relevant features obtained from phase 2. In the final phase, the average correlation is computed for each relevant feature with all other remaining relevant features and removes the worst feature which has the highest mean correlation. Phase 4 is repeated until the desired k-size subset of features is obtained. The attribute which is removed in phase 4 is also discarded from the average calculation in the subsequent iteration. The framework of the proposed methodology is named as GainSUn (Information Gain and Symmetric **Un**certainty). The steps involved in the proposed methodology are shown in the procedure ***GainSUn***.

```
procedure GainSUn(S, δ, k, S_best)
// S(F_1,F_2,...,F_n,C) -A training dataset
δ                -A user defined threshold
k                -Size of the optimal subset
S_best           -Optimal subset, an output
                  parameter //
begin {main}
   S_a_d ← call MBD(S);
   S_list ← Relvance_IG(S_a_d, δ);
   m ← length(S_list);
   if m <> k and m >2 {
      ffc_t ← call FF_Corr_SU(S_list,m);
      S_best ← call Remove_worst_feature(S_list,ffc_t,k); }
   return S_best;
end {main}

procedure MBD(S : training dataset)
//This procedure is used to convert the continuous
features into discrete//
{
   for each continuous attribute F_i in S
   {
      a) compute median M;
         i.sort the values of a continuous feature F_i in
           ascending order;
         ii.for each unique value x_i in F_i, calculate the
           frequency of occurrence f and cumulative
           frequency cf;
        iii.Mid ← (N+1) / 2 where N = ∑f;
        iv.M ← the item which has cf ≥ Mid;
      b) perform discretization
         F_i_des ← {low,high};
         for each x_i ∈ F_i
            if x_i > M then x_i ← F_i_des [1];
            else x_i ← F_i_des [0]; }
      return S;
   }
}
```

procedure **Relevance_IG**(S_{ad}: discretized training dataset, δ: threshold value)
//This procedure is used to find the relevance features//
{
S_{list} = Ø;
for $i \leftarrow 1$ to n do{
$IG_{Fi,c} \leftarrow$ calculate $IG_{Fi,c}$ for each F_i in S_{ad};
if ($IG_{Fi,c} > \delta$) then append F_i to S_{list};
}
return S_{list} ; //relevant feature set //
}

procedure **FF_Corr_SU**(S_{list} : relevant feature set, *m:*size of the relevant set)
//This procedure is used to find the feature-feature correlation//
{
for $i \leftarrow 1$ to m do{
for $j \leftarrow 1$ to m do{
if $i \neq j$ then
$SU_{ij} \leftarrow$calculate $SU(F_i,F_j)$;
if $i = j$ then
$ffc\_table[i,j] \leftarrow 0.0f$;
else{
$ffc\_table[i,j] \leftarrow SU_{ij}$;
$ffc\_table[j,i] \leftarrow SU_{ij}$;
}
}
}
return *ffc_table*;
}
procedure **Remove_worst_feature**(S_{list},*k)*
//This procedure is used to remove the worst features in the relevant list of feature set//
{
$x \leftarrow$length (S_{list});
$m \leftarrow x-1$;
repeat:
len $\leftarrow m$;
for $i \leftarrow 1$ to x do{

$F_{i_worst} \leftarrow$calculate $\sum_{i=1, j=1, i\neq j}^{x} ffc\_table[i, j]$;

$w[i] \leftarrow F_{i\_worst} / len$;
}
$max \leftarrow$ maximum($w[i]$) $1 \leq i \leq x$;
feature_remove $\leftarrow i$;
remove $F_{feature\_remove}$ from S_{list};
$m = m-1$;
if $m = k$ then return(S_{list});
else{
for $y \leftarrow 1$ to x do{
for $j \leftarrow 1$ to x *do*{
if *feature_remove*= j then
ffc_table [y,j]=0.0f;
if *feature_remove* = y then
ffc_table [j,y]=0.0f;
}
}
go to repeat;
}

5. 1. Proposed Work - An Example In order to understand the relevance of the proposed methodology the Statlog heart dataset from the UCI machine learning repository has been taken[1] . The general information of the dataset is shown in Table 1.
The steps involved in the proposed work are as follows.

Step 1: Discretization of numeric attributes into nominal

For sample illustration the age field of the dataset has been taken and it has the values [*70, 67, 57, 64, 74, 65, 56, 59, 60, 63,..., 58, 60, 58, 49, 48, 52, 44, 56, 57, 67*]. The *M* for the data is 55. The attribute values which are ≤ *M* is discretized as "*low*" and "*high*" otherwise. Thus the discretized age attribute contains ['*h*',' *h*',' *h*',' *h*', '*h*', '*h*', '*h*',' *h*',' *h*', ' *h*',...,'*h*',' *h*',' *h*', '*l*', '*l*', '*l*', '*l*', '*h*', '*h*', '*h*'] where '*h*' is *high* and '*l*' is *low*. Similar calculation can be performed for other numerical attributes too.

Step 2: Select the relevant features using IG

This step uses the values of the nominal attributes without any changes, discretized values using MBD for numerical attributes and finds the *IG* for all the features. The target variable *'class'* contains two values viz., *presence (p)* and *absence (a)*, each containing 120 and 150 records respectively. The class label of the dataset contains ['*p*', '*a*', '*p*', '*a*', '*a*', '*a*', '*p*', '*p*', '*p*', '*p*',...,'*a*', '*p*', '*p*', '*a*', '*p*', '*a*', '*a*', '*a*', '*a*', '*p*']. Using (2), the entropy of class is computed as

H (Class) = -((120/270)×$\log_2$(120/270) +(150/290) ×$\log_2$(150/270)) = 0.9911

The discretized *'age'* attribute contains two unique values *low* and *high* containing 138 and 132 instances each. Among the low instances, 44 instances are identified as '*p*' and 94 as '*a*'. Similarly, among the high instances of age 76 and 56 are identified as '*p*' and '*a*' respectively. Using (3), the entropy of class after observing age is computed as

H(Class|Age)=(138/270)×(-(44/138) × $\log_2$(44/138) -(94/138 × $\log_2$(94/138)) + (132/270) × (-(76/132) × $\log_2$ (76/132) -(56/132) × $\log_2$ (56/132)) = 0.9424

Using (1) the IG (Class, Age) is calculated as

IG(Class,Age)=H(Class)-H(Class|Age)=0.9911- 0.9424 = 0.0487.

As IG is symmetric, IG(Class,Age) = IG(Age, Class) = 0.0487. Similar calculations can also be performed for other attributes in the dataset and it is shown in Table 2.

1. UCI Machine Learning Repository - Center for Machine Learning and Intelligent System. [online] http://archive.ics.uci.edu (accessed 10 October 2015).

The *M* for $IG(F_i)$ is 0.0487 and it is fixed as threshold value. Thus, the attributes such as *cp, mhr, eia, op, nmv* and *thal* are selected as relevant ones because their IG's are > 0.0487.

Step 3: Finding the F-F correlation

The *cp* and *mhr* attributes have the discretized values ['*h*', '*l*', '*l*', '*h*', '*l*', '*h*', '*l*', '*h*', '*h*', '*h*',..., '*l*', '*h*', '*l*', '*l*', '*l*', '*l*', '*l*', '*h*', '*h*'] and ['*l*', '*h*', '*l*', '*l*', '*l*', '*l*', '*l*', '*l*', '*h*', '*h*',..., '*h*', '*l*', '*h*', '*h*', '*h*', '*h*', '*h*', '*l*', '*l*', '*l*'] respectively.

IG(*cp, mhr*)= 0.0550

H(*cp*)=-(141/270× $\log_2$(141/270)+129/270× $\log_2$ (129/270)) = 0.9986

H(*mhr*) = -(135 /270 × $\log_2$ (135/270)+135/270 × $\log_2$(135/270))=1.0000

The SU for *cp* and *mhr* is computed using (4) as

SU(*cp,mhr*) = SU(*mhr,cp*) = 2 × (0.0550 / (0.9986 + 1.0000)) = 0.0550

Similarly the *F-F* correlation is computed for each relevant feature with all other relevant features obtained in step 2 and they are shown in Table 3.

Step 4: Removing the worst features

In order to remove the worst feature, the average of SU's are computed for each relevant feature with all other relevant features and they are listed in Table 4.

From Table 4, it is identified that '*eia*' has the highest average SU, so it is removed. After removing the '*eia*' attributes, the process is terminated because the current subset contains the desired number of *5* features viz., *'cp', 'mhr', 'op', 'nmv' and 'thal'*.

The desired number of features (*k*) for the final subset is *k=number of relevant features - round (number of relevant features × % of features to be reduced)*.

In this case, *k = 6 - round (6×0.10) = 5*. Thus, the features *'cp', 'mhr', 'op', 'nmv' and 'thal'* are the selected features using the proposed method.

After FS, the original features (15), the relevant features (6) and the optimal subset of selected features (5) using the proposed method are fed into Naive Bayesian Classifier (NBC) to determine the predictive accuracy.

It is observed that for all attributes the accuracy of NBC is 83.7037%. For the identified relevant features the accuracy is 85.1852%. After removing the redundant attributes using proposed methodology, the accuracy is 83.3333%.

TABLE 1. Description of the Statlog heart dataset

No. of Attributes	Numerical: 13 Nominal: 1
Class type	Nominal
Features	age, sex, cp, bp, sc, fbs, re, mhr, eia, op, slope, nmv, thal and class
No. of instances	270
No. of instances in each class type	presence :120 absence :150

TABLE 2. Computed feature-class correlation for the Statlog heart dataset

Feature (F_i)	**IG (F_i)**
age	0.0487
sex	0.0000
cp	0.1904
bp	0.0068
sc	0.0268
fbs	0.0002
re	0.0000
mhr	0.0867
eia	0.1299
op	0.2434
slope	0.0026
nmv	0.1659
thal	0.2030

TABLE 3. Feature-Feature correlation using SU

Feature	**cp**	**mhr**	**eia**	**op**	**nmv**	**thal**
cp	---	0.055	0.148	0.040	0.040	0.077
mhr	0.055	---	0.089	0.078	0.049	0.047
eia	0.148	0.089	---	0.060	0.032	0.078
op	0.040	0.078	0.060	---	0.024	0.060
nmv	0.040	0.049	0.032	0.024	---	0.049
thal	0.077	0.047	0.078	0.060	0.049	---

TABLE 4. Average SU's for each selected relevant feature with all other relevant features

Attribute	**Average SU**
cp	0.0601
mhr	0.0531
eia	**0.0678**
op	0.0437
nmv	0.0320
thal	0.0518

6. EXPERIMENTAL STUDY

For analyzing the proposed work, 8 datasets have been taken from UCI Machine Learning Repository [2]. Each dataset contains both continuous and nominal features. The missing values for each attribute in the datasets are filled with their corresponding mean. The detailed specification of the datasets is shown in Table 5. The proposed methodology has been implemented using python. The number of original features, selected

[2] UCI Machine Learning Repository - Center for Machine Learning and Intelligent System. [online] http://archive.ics.uci.edu (accessed 10 October 2015).

features after relevance analysis and redundant analysis are shown in Table 6 and they are fed into NBC for determining the efficiency of the proposed method using WEKA tool with 10-fold cross validation. The accuracy of the NBC after relevance and redundant analysis in comparison with the original features is shown in Table 7. Its graphical representation is shown in Figure 1.

From Table 7, it is observed that the accuracy of NBC is increased for the datasets viz., Pima Indian Diabetes, Statlog Heart, Eeg, Weather and Ann-train. Similarly, it is decreased for the datasets viz., Breast Cancer, Lung Cancer and SPECTF_train with the selected features after relevance analysis. It is also identified that the accuracy of NBC after redundant analysis is increased for the datasets viz., Eeg and Pima Indian Diabetes. However, for the datasets Ann-train and Statlog heart it is slightly decreased. Similarly, the accuracy of NBC for the Weather dataset remains same after relevance and redundant analysis.

TABLE 5. Summary of bench-mark datasets

Dataset	No. of Attributes	No. of Instances	No. of Classes
Pima Indian Diabetes	9	768	2
Breast Cancer	11	699	2
Statlog Heart	14	270	2
Eeg	15	14979	2
Weather	5	14	2
Ann-train	22	3772	3
Lung Cancer	57	32	2
SPECTF_train	45	80	2

TABLE 6. Number of original features, selected features after relevance and redundant analysis

Dataset	No. of attributes		
	Before FS	After FS with	
		Relevance analysis	Redundant analysis
Pima Indian Diabetes	9	4	4
Breast Cancer	11	5	5
Statlog Heart	14	6	**5**
Eeg	15	7	**6**
Weather	5	2	2
Ann-train	22	10	**9**
Lung Cancer	57	28	**25**
SPECTF _train	45	20	**18**
Total	178	82	**74**

TABLE 7. Accuracy of the NBC with original features, selected features after relevance and redundant analysis

Dataset	Accuracy (%) of NBC with		
	all features	selected features after	
		Relevance analysis	Redundant analysis
Pima Indian Diabetes	76.3021	**77.4740**	**77.4740**
Breast cancer	**95.9943**	95.2790	95.2790
Statlog heart	83.7037	**85.1852**	83.3333
Eeg	48.0406	**48.9285**	**53.1744**
Weather	64.2857	**71.4286**	**71.4286**
Ann-train	95.6522	**95.7317**	95.3075
Lung Cancer	84.3750	84.3750	**90.6370**
SPECTF_ train	76.2500	76.2500	**82.7500**
Average	78.0755	79.3315	**81.1730**

It is further noted that the accuracy of NBC with the original features for the Lung Cancer and SPECTF_train remains same with selected features after redundant analysis. On an average, the accuracy of NBC using the selected features after redundant analysis is enhanced by 3.0975% than the original features.

Further, it is noticed that when there are large number of features in the dataset, there is more possibility of redundant attributes. From Table 6, it is observed that for the datasets with large number of features, there is a drastic decrease in the number of selected features (approximately more than 50% i.e., Ann-train – from 22 attributes to 9, Eeg – from 15 to 6 attributes, Statlog heart – from 14 to 5 attributes, Lung Cancer - from 57 to 25 and SPECTF_train - from 45 to 18) using the proposed method.

Thus, the proposed methodology is more suitable for high dimensional datasets. As the proposed method removes irrelevant and redundant features there is no possibility of occurring them in the optimal subset of selected features. On an average, the predictive accuracy of the NBC for the optimal subset of selected features has been enhanced using this hybrid method.

The efficiency and effectiveness of the proposed methodology has been analyzed by comparing it with the representative features selection methods viz., Gain Ratio (GR), Information Gain (IG), Chi Squared Attribute Eval, One-R, ReliefF and Symmetric Uncertainty (SU). The experiments for the existing methods are carried out with the WEKA environment using the same threshold as in GainSUn. The selected features of each datasets using the existing methods are fed into the NBC.

The accuracies obtained from the experimental study are shown in Table 8.

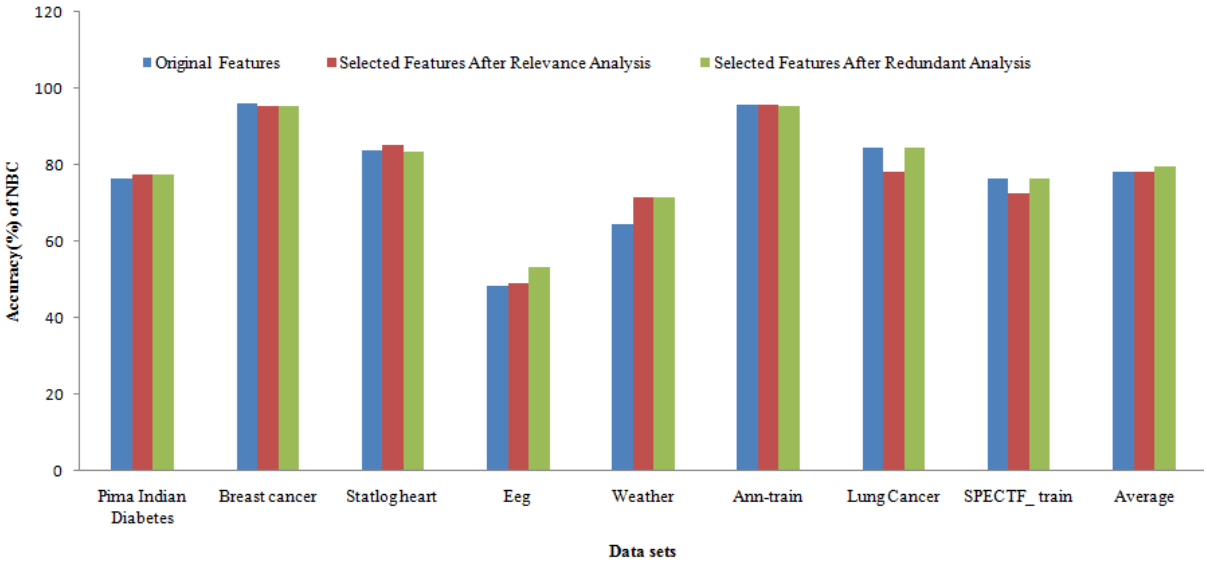

Figure 1. Accuracy comparison of NBC with original and selected features after relevance and redundant analysis

TABLE 8. Comparison of NBC accuracies between GainSUn and existing FS methods

Data sets	Accuracy (%) of NBC with the selected features using						
	GainSUn*	IG#	GR#	ChiSquaredAttribute Eval#	One-R#	ReliefF#	SU#
Pima Indian Diabetes	**77.4740(4)**	75.3906 (4)	75.5208(4)	75.3906 (4)	73.0469(4)	75.3906 (4)	75.3906 (4)
Breast cancer	95.2790 (5)	95.9943 (5)	95.7082(5)	95.9943 (5)	95.7082(5)	**96.1373 (5)**	95.7082 (5)
Statlog heart	83.3333 (5)	**84.4444 (5)**	82.963(5)	**84.4444 (5)**	82.9630(5)	82.5926 (5)	82.9630 (5)
Eeg	53.1744 (6)	48.7149 (6)	46.1379(6)	48.7149 (6)	**53.2813(6)**	47.1193 (6)	46.1379 (6)
Weather	**71.4286 (2)**	57.1429 (2)	57.1429(2)	57.1429 (2)	64.2857(3)	**71.4286 (2)**	57.1429 (2)
Ann-train	95.3075 (9)	95.5992 (9)	95.5992(9)	95.5992 (9)	95.5726(9)	95.4931 (9)	**95.7052 (9)**
Lung Cancer	**90.637(25)**	90.625 (25)	90.625(25)	90.6250 (25)	84.375(25)	84.375(25)	90.625(25)
SPECTF _train	**82.75 (18)**	76.250 (18)	77.500(18)	77.5000 (18)	82.50(18)	78.750(18)	80.000(18)
Average	**81.173(74)**	78.02(74)	77.65(74)	78.18(74)	78.97(75)	78.91(74)	77.96(74)

*Proposed Method, #Existing methods and No. of selected features - enclosed within parenthesis

From Table 8 it is noted that for the datasets viz., Pima Indian Diabetes, Statlog heart, Weather, Ann-train, Lung Cancer and SPECTF_train the accuracy of NBC is enhanced with the optimal subset of features obtained using the proposed methodology. On an average, the accuracy of NBC using GainSun is increased by 3.28% approximately.

7. CONCLUSION

A new hybrid FS method has been proposed in this paper. The proposed work finds irrelevant and redundant attributes and removes them from the original feature space using IG with MBD and SU. The optimal subset of selected features and the original features are fed into the NBC for determining the efficiency of the proposed work by calculating the predictive accuracy. The predictive accuracy of the NBC using the proposed method is enhanced on an average with small subset of selected features rather than the original. This is due to the fact that the resultant final subset does not contain any irrelevant, noise and redundant information. It has also been proved that the accuracy is increased by 3.28% approximately when it is compared with the popular existing FS methods.

8. REFERENCES

1. Hemati, H., Ghasemzadeh, M. and Meinel, C., "A hybrid machine learning method for intrusion detection", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 29, No. 9, (2016), 1242-1246.
2. Hamidi, H. and Daraee, A., "Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases", ***International Journal of Engineering-Transactions A: Basics***, Vol. 29, No. 7, (2016), 921-930.
3. Han, J., Pei, J. and Kamber, M., "Data mining: Concepts and techniques, Elsevier, (2011).
4. Amr, T., "Survey on feature selection", ***IEEE Transactions on Information Forensics and Security***, Vol. 3, No. 1, (2008), 91-100.
5. Yu, L. and Liu, H., "Feature selection for high-dimensional data: A fast correlation-based filter solution", in ICML. Vol. 3, (2003), 856-863.
6. Hall, M. A., "Correlation-based feature selection of discrete and numeric class machine learning", ***Seventeenth International Conference on Machine Learning***, USA, Morgan Kaufmann Publishers Inc., (2000), 359-366. (2000).
7. Haindl, M., Somol, P., Ververidis, D. and Kotropoulos, C., "Feature selection based on mutual correlation", ***Progress in Pattern Recognition, Image Analysis and Applications***, Vol., 4225, (2006), 569-577.
8. Pino, A. and Morell, C., "Analytical and experimental study of filter feature selection algorithms for high-dimensional datasets", in Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support, Atlantis Press., (2013), 339-349.
9. Biesiada, J. and Duch, W., "Feature selection for high-dimensional data—a pearson redundancy based filter", ***Computer Recognition Systems 2***, Vol. 45, (2007), 242-249.
10. Senliol, B., Gulgezen, G., Yu, L. and Cataltepe, Z., "Fast correlation based filter (FCBF) with a different search strategy", in Computer and Information Sciences, 23rd International Symposium on, IEEE, (2008), 1-4.
11. Peter, T. J. and Somasundaram, K., "Study and development of novel feature selection framework for heart disease prediction", ***International Journal of Scientific and Research Publications***, Vol. 2, No. 10, (2012), 1-7.
12. Mani. K. and Kalpana. P., "A filter-based feature selection using information gain with median based discretization for naive bayesian classifier", ***International Journal of Applied and Engineering Research***, Vol. 10, No. 82, (2015), 280-285.
13. Yu, L. and Liu, H., "Efficient feature selection via analysis of relevance and redundancy", ***Journal of Machine Learning Research***, Vol. 5, No. Oct, (2004), 1205-1224.
14. Rajesh, K. and Sangeetha, V., "Application of data mining methods and techniques for diabetes diagnosis", ***International Journal of Engineering and Innovative Technology (IJEIT)***, Vol. 2, No. 3, (2012).
15. Tang, J., Alelyani, S. and Liu, H., "Feature selection for classification: A review", ***Data Classification: Algorithms and Applications***, (2014), 37-64.
16. Mahdizadeh, M. and M. Eftekhari, "A novel cost sensitive imbalanced classification method based on new hybrid fuzzy cost assigning approaches, fuzzy clustering and evolutionary algorithms", ***International Journal of Engineering (IJE), Transactions B: Applications***, Vol. 28, No. 8, (2015), 1160-1168.

A New Hybrid Framework for Filter based Feature Selection using Information Gain and Symmetric Uncertainty

TECHNICAL NOTE

P. Kalpana, K. Mani

Department of Computer Science, Nehru Memorial College, Puthanampatti, Tiruchirappalli-Dt,Tamil Nadu, India

P A P E R I N F O

Paper history:
Received 29 November 2016
Received in revised form 12 February 2017
Accepted 10 March 2017

Keywords:
Irrelevant Redundant
Median Based Discretization
Information Gain
Symmetric Uncertainty
Accuracy
Naive Bayesian Classifier

چکیده

انتخاب ویژگی، یک تکنیک پیش پردازش مورد استفاده برای از بین بردن ویژگی های بی ربط و زائد است که به افزایش عملکرد طبقه بندی کننده نتیجه می دهد. هنگامی که یک مجموعه داده شامل ویژگی های بی ربط و برکنار شده است، آن نمی تواند دقت را افزایش دهد و همچنین عملکرد طبقه بندی را کاهش می دهد. برای جلوگیری از آنها، این مقاله یک روش انتخاب ویژگی ترکیبی جدید را با استفاده از به دست آوردن اطلاعات و عدم اطمینان متقارن ارائه داده است. کار پیشنهادی از گسسته متوسط بر اساس تبدیل ویژگی های کمی به کیفی، کسب اطلاعات در پیدا کردن ویژگی های مربوط و عدم اطمینان متقارن به حذف ویژگی های برکنار شده استفاده می کند. همانگونه که کار پیشنهادی از هر دو تجزیه و تحلیل ارتباطی و برکنار شده استفاده می کند، دقت پیش بینی طبقه بندی Naive Bayesian بهبود یافته است. بیشتر بهره وری و اثربخشی روش ارائه شده، با استفاده از مقایسه با سایر روش های موجود و با استفاده از مجموعه داده های دنیای واقعی از ابعاد بالا تحلیل می شود.

***doi**: 10.5829/idosi.ije.2017.30.05b.05*