K L University

(Deemed to be University estd, u/s. 3 of the UGC Act, 1956)

(NAAC Accredited "A" Grade University)

KONERU LAKSHMAIAH EDUCATION FOUNDATION

Research Progress Review : 04

Sai Prasad Potharaju

Reg No: 15303175

Batch : X

Supervisor: Dr. M. Sreedevi

Research Identification

1	Engineering Discipline	Computer science and Engineering
2	Major Area of Research	Data Mining
3	Minor Area of Research	Preprocessing (Feature Selection)

Data Mining Stages

Data Collection

Data Preprocessing

Apply Data Mining Techniques

Interpret and Visualization

Data Mining Techniques

Association Rule Mining

Classification

Regression

Clustering

Data Preprocessing

Noisy and Missing values

Mislabeled

Imbalanced

High Dimensionality

High Dimensionality (Research Problem)

Drawbacks

Consumption of more memory

Lower the classification performance

Confuse the learning model

Solutions

Apply Feature Selection or Feature Extraction Techniques

Existed Feature Selection Modes

Filter Based

Based on the Information Theory

Assigns the Rank to each feature

Ex: Information Gain, Gain Ratio, Chi2, Symmetrical Uncertainty

Wrapper

Derives the Subset of Feature set

Use the Searching Algorithm

Embedded

Research Objectives

- To Study and analysis of different preprocessing techniques including imabalancing and high dimensionality.
- To propose a feature selection methodology to address the high dimensionality issue using symmetrical uncertainity.
- To propose a novel feature selection methodology to address the high dimensionality issue using correlation coefficient and symmetrical uncertainity.
- To propose an unsupervised feature selection methodology to address the high dimensionality issue using clustering and filter based methods.
- To evaluate proposed methods and compare with the existing methods using various classifiers.

Proposed Methodology : 1

Based on correlation coefficient and Symmetrical Uncertainty

1. **Find out the Symmetric Uncertainty (SU) value of each feature, such that all features will be in descending order of its SU value.**
2. **Choose the middle feature SU value as Threshold (T).**
3. **Generate the Correlation Coefficient Symmetrical matrix (CCE(X_i,Y_i)) of initial data set .**
4. **Transform the above matrix to weighted binary matrix (WB) as per the below steps**

```
for(i=1 to n)
 for(j=1 to n)
    if(CCE($X_i$,$Y_i$)>T)
    WB($X_i$,$Y_i$)=1
    else
    WB($X_i$,$Y_i$)=0
  End
End
```

Proposed Methodology : 1

Based on correlation coefficient and Symmetrical Uncertainty (Contd…)

5. Calculate the total weight of each feature W(F) as per below steps.

```
for(i=1 to n)
    for(j=1 to n)
    $W(F_i)= WB(X_i,Y_i)$
    End
End
```

6. Group the features which are having same weight(W(F))

```
    $Cluster_i=\{F_{i1}, F_{i2},...F_{ik}\}$  /* i is the cluster id, increment i by 1 until all features are formed */
```

7. Choose the best feature (feature which has maximum SU value) from each cluster and form the final candidate subset

```
for(i=1 to last cluster)
    $F_i$= MAX SU(cluster $_i$)
    Candidate Feature set (CFS)<- $F_i$
End
```

Example

Assume there are ten features (a, b, c, d, e, f, g, h ,i, j) in a sample data set . SU value of each feature is given in below Table.1

Table 1 . SU value of sample data set features

SU	Feature No	Fid
.19	10	j
.19	8	h
.19	7	g
.18	9	i
.15	2	b
.09	1	a
.07	4	d
.06	3	c
.06	5	e
.02	6	f

Proposed Methodology : 1

Correlation Coefficient Symmetrical matrix (CCE(X_i,Y_j)) of the data set is given in below.

Table 2. Correlation Coefficient Symmetrical matrix (CCE(X_i,Y_j))

Feature Id	a	b	c	d	e	f	g	h	i	j
a	1	-0.08	-0.03	-0.15	-0.16	-0.05	-0.11	0.31	-0.28	0.29
b	-0.08	1	0.05	0.09	-0.11	-0.04	-0.13	-0.28	0.21	-0.37
c	-0.03	0.05	1	-0.07	0.05	-0.01	0.27	-0.1	0.12	-0.07
d	-0.15	0.09	-0.07	1	0.29	0.01	0.09	-0.23	0.29	-0.31
e	-0.16	-0.11	0.05	0.29	1	0.12	0.23	-0.12	0.56	-0.27
f	-0.05	-0.04	-0.01	0.01	0.12	1	0.01	0.04	0.03	-0.03
g	-0.11	-0.13	0.27	0.09	0.23	0.01	1	0.05	0.27	-0.14
h	0.31	-0.28	-0.1	-0.23	-0.12	0.04	0.05	1	-0.43	0.46
i	-0.28	0.21	0.12	0.29	0.56	0.03	0.27	-0.43	1	-0.47
j	0.29	-0.37	-0.07	-0.31	-0.27	-0.03	-0.14	0.46	-0.47	1

Proposed Methodology : 1

Transformed Weighted binary matrix of above matrix and weight of each feature is given below

Table 3. Weighted binary matrix

Feature Id	a	b	c	d	e	f	g	h	i	j		Weight	Feature
a	1	0	0	0	0	0	0	1	0	1		3	a
b	0	1	0	0	0	0	0	0	1	0		2	b
c	0	0	1	0	0	0	1	0	0	0		2	c
d	0	0	0	1	1	0	0	0	1	0		3	d
e	0	0	0	1	1	0	1	0	1	0		4	e
f	0	0	0	0	0	1	0	0	0	0		1	f
g	0	0	1	0	1	0	1	0	1	0		4	g
h	1	0	0	0	0	0	0	1	0	1		3	h
i	0	1	0	1	1	0	1	0	1	0		5	i
j	1	0	0	0	0	0	0	1	0	1		3	j

Proposed Methodology : 1

Form the clusters and select the best feature in each cluster

Cluster Id	Weight	FID	Selected Feature From each cluster
1	1	f	f
2	2	b	b (As SU value of ‘b’ is maximum than other features in Cluster)
	2	c	
3	3	a	J (As SU value of ‘j’ is maximum than other features in cluster)
	3	d	
	3	h	
	3	j	
4	4	d	d (As SU value of ‘d’ is maximum than other features in cluster)
	4	e	
5	5	e	I As SU value of ‘i’ is maximum than other features in cluster
	5	i	

Proposed Methodology : 1

Form the final candidate feature set (CFS)

CFS= {f, b, j, d, i}

Proposed Methodology : 1

Experiment

To examine the proposed framework, ten (10) real-time benchmark data sets are taken into consideration.

Data sets description

Data set ID	Name of the Data Set	# Instances	# Features	# Features Selected	# Class
1	Ionosphere	351	34	13	2
2	Dermatology	366	34	13	6
3	Biodegradation	1055	41	23	2
4	Cardiotocography	2126	22	12	3
5	Lung Cancer	33	56	15	3
6	Libras Movement	360	90	21	15
7	Connectionist Bench(Sonar)	208	60	28	2
8	Spambase	4601	57	16	2
9	Breast Cancer(WDBC)	569	30	15	2
10	Musk (V 2)	476	166	54	2

Proposed Methodology : 1

Click for Result

Proposed Methodology 2

Based on the Project Allocation Strategy

1. Generate the weight and Rank of each feature using SU

2 Remove the features, whose weight is Zero (0) as it can’t influence the learners.

Follow the below steps to form the subset of features in 4 Quarters.

Step 1: Arrange the first 4 features in descending order of Ranks from left to right in Level 1

Step 2: Arrange the next 4 features in descending order of Ranks from right to left in Level 2.

Step 3: Repeat the Step 1 then step 2 for next Levels until the all features are arranged.

Step 4: Group, all vertically first order features of all levels in First Quarter, Second order features of all levels in Second Quarter, and so on.

Step 5: Balance the number of features of each quarter by removing last feature from the quarter which has an extra feature ,if not balanced.

Proposed Methodology 2

For Example /Experiment/ Result Click

Proposed Methodology 2

Variation 1 (SONAR Target)

For Example /Experiment/ Result Click

Proposed Methodology 2

Variation 2 (Microarray Datasets)

For Example /Experiment/ Result Click

Conclusion

In this research study , a novel cluster of feature selection frameworks based on Symmetrical Uncertainty(SU), correlation coefficient are proposed. The new approaches could generate finite clusters, in which each cluster has finite number of features without duplication. All the cluster of features are evaluated with existing feature selection methods such as Gain Ratio Attribute Evaluator, Chi Square Feature selection, Information Gain. For evaluating the accuracy of each cluster rule based, tree based ,Lazy learners (KNN) are applied .After complete analysis , it has been noticed that, clusters formed by proposed methods are competing with regular methods.

Research Profile

- List of Publications **Click**
- Scopus Author's Profile **Click**
- Google Scholar Profile **Click**

Thank You

Suggestions are welcome

Q & A