A FEATURE SELECTION APPROACH TO INCREASING CLASSIFICATION PERFORMANCE

INTRODUCTION :

- System can support any degree of dataset as input.
- System focuses on classification task of data mining and high dimenstional issue of preprocessing .
- Create groups of similar weight data and select best from group.
- Drop unnecessary data using CCE and SU.
- Gives best features set.

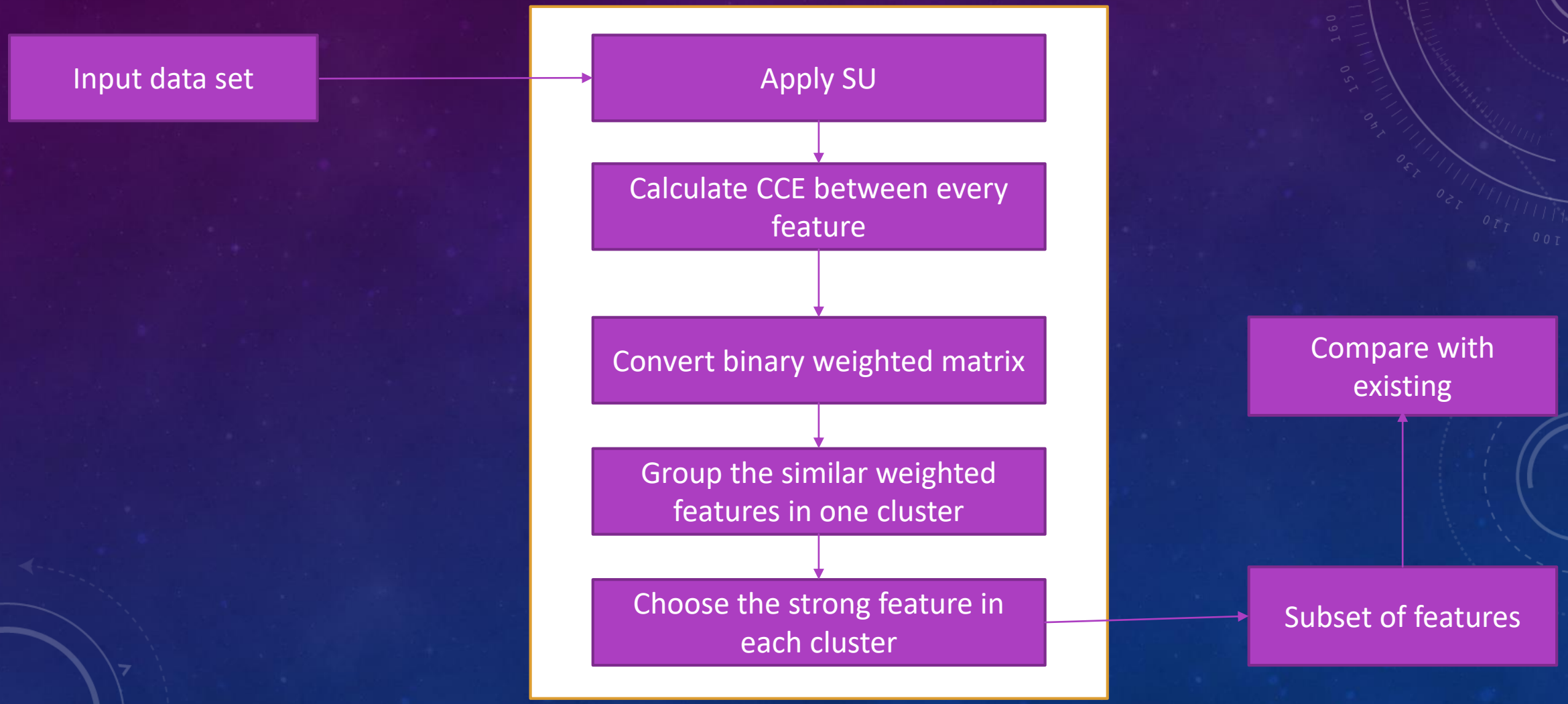
SCOPE:

1. Feature selection using CCE and SU.
2. Make groups of same priority and out of them select best.
3. removes redundant and independent attributes with class level.
4. CCE is considered for relationship between two variables.
5. SU is considered to fix minimum value of weight

OBJECTIVE :

- Extracting the better features.
- To increase the classification performance.
- Reduce high dimensionality.
- To reduce dimensionality using CCE and SU.
- To devolope methodology to address the high dimensionality issue using clustering and filter based methods.
- To devolope and compare with the existing methods using various classifiers.

SYSTEM ARCHITECTURE



PROPOSED METHODOLOGY WORKING

CSV File as input

X	y		
43	99		
	10	20	
21	65		
25	79		
10	12	58	
42	75		
57	87		
59	81		
10			

Data Preprocessing
Noisy and Missing values
Mislabeled Imbalanced

X	y		
43	99		
	10	20	
21	65		
25	79		
10	12	58	
42	75		
57	87		
59	81		
10	“a6”		

- 3 rows are drop due to invalid data
- 6 rows are selected for next processing

Generate the Correlation coefficient symmetrical matrix

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

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

Find SU for each feature and sort decending order by SU

SU	Feature No	Feature Name
.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

$SU=(2*IG)/(H(X)+X(Y))$

$H(X)=-\int p(x)\log(p(x))\partial x$

Supervised Feature Selection - Information Gain

$$IG(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i)$$

$$+ p(t)\sum_{i=1}^{m} p(c_i \mid t)\log p(c_i \mid t) + p(\bar{t})\sum_{i=1}^{m} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t})$$

c_i represents the i th category, $P(c_i)$ is the probability of the i th category.

$P(t)$ and $P(\bar{t})$ are the probabilities that the term t appears or not in the documents.

$P(c_i|t)$ is the conditional probability of the i th category given that term t appeared, and $P(c_i|\bar{t})$ is the conditional probability of the i th category given that term t does not appeared.

Choose the middle feature's SU value as t

SU	Feature No	Feature Name
.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

t=.15

Transform CCE matrix to binary matrix as compare each value to t if less represent as 0 else 1

t=0.15

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

Feature Id	a	b	c	d	e	f	g	h	i	j
a	1	0	0	0	0	0	0	1	0	1
b	0	1	0	0	0	0	0	0	1	0
c	0	0	1	0	0	0	1	0	0	0
d	0	0	0	1	1	0	0	0	1	0
e	0	0	0	1	1	0	1	0	1	0
f	0	0	0	0	0	1	0	0	0	0
g	0	0	1	0	1	0	1	0	1	0
h	1	0	0	0	0	0	0	1	0	1
i	0	1	0	1	1	0	1	0	1	0
j	1	0	0	0	0	0	0	1	0	1

Calculate the total weight using binary matrix row wise

Feature Id	a	b	c	d	e	f	g	h	i	j
a	1	0	0	0	0	0	0	1	0	1
b	0	1	0	0	0	0	0	0	1	0
c	0	0	1	0	0	0	1	0	0	0
d	0	0	0	1	1	0	0	0	1	0
e	0	0	0	1	1	0	1	0	1	0
f	0	0	0	0	0	1	0	0	0	0
g	0	0	1	0	1	0	1	0	1	0
h	1	0	0	0	0	0	0	1	0	1
i	0	1	0	1	1	0	1	0	1	0
j	1	0	0	0	0	0	0	1	0	1

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

Group the feature which having same weight

Select the one feature from group by first appearance in SU Table

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

SU	Feature No	Feature Name
.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

Group id	Weight	F_name	Selected feature
1	1	f	f
2	2	{b, c}	b
3	3	{a,d,h,j}	j
4	4	{g,e}	g
5	5	i	i

Final candidate feature set:

$$CFS=\{ f, b, j, g, I \}$$

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

CONCLUSION:

- It reduce the data set dimensionality by selecting the best features in it to boosts-up the classification performance.
- .method was compared with four existing filter based methods namely, information gain(IG), Chi-Square (Chi), Grain Ratio (GR), and Relief.
- For testing the our proposed method, six different classifiers Jrip, Ridor, J48, Simple cart, Naive Bayes, IBk are applied on Ten different real time data sets.
- method displayed better results than existing IG and GR on 7 data sets,

FUTURE SCOPE

The same technique can be implemented using Hadoop framework .

Final candidate feature set:

$$CFS=\{ f, b, j, g, I \}$$

REFERENCES

- Correlation Coefficient Based Candidate Feature Selection Framework Using Graph Construction

  SP Potharaju, M Sreedevi

  Gazi University Journal of Science 31 (3), 775-787