*Recent Trends in Intensive Computing*
*M. Rajesh et al. (Eds.)*

*doi:10.3233/APC210196*

# SU-CCE: A Novel Feature Selection Approach for Reducing High Dimensionality

A B Pawar [a, 1], M A Jawale [a], Ravi Kumar Tirandasu [b] and Saiprasad Potharaju [a]

[a]*Department of Computer Engineering, Sanjivani College of Engineering, Kopargaon, Maharashtra, India*
[b]*Department of Computer Science & Engineering, Koneru Laksmaiah Education Foundation, Vaddeswaram, AP, India*

**Abstract.** High dimensionality is the serious issue in the preprocessing of data mining. Having large number of features in the dataset leads to several complications for classifying an unknown instance. In a initial dataspace there may be redundant and irrelevant features present, which leads to high memory consumption, and confuse the learning model created with those properties of features. Always it is advisable to select the best features and generate the classification model for better accuracy. In this research, we proposed a novel feature selection approach and Symmetrical uncertainty and Correlation Coefficient (SU- CCE) for reducing the high dimensional feature space and increasing the classification accuracy. The experiment is performed on colon cancer microarray dataset which has 2000 features. The proposed method derived 38 best features from it. To measure the strength of proposed method, top 38 features extracted by 4 traditional filter-based methods are compared with various classifiers. After careful investigation of result, the proposed approach is competing with most of the traditional methods.

**Keywords.** Data mining, classifier, feature selection, preprocessing.

## 1. Introduction

Data mining is the adulatory area of research since a decade in many fields including medical investigations, stock market analysis, business, education, transportation, etc. With the data mining approaches the more insights of the data can be analyzed for future prediction in order to get positive results. For any field of study data is crucial part. Data may be in different forms (text, audio, video, image etc.) from field to field. Before applying data mining techniques, the type and format of data need to be preprocessed [1]. The collected data may have missing values, missing classes, imbalanced and high dimensional in nature. Missing values can be addressed using by filling the values using various techniques like average values, binning. Missing classes can be filled by calculating the distance with already classes. Imbalanced issue can be addressed with applying the oversampling and under sampling techniques [2].

[1] A B Pawar, Department of Computer Engineering, Sanjivani College of Engineering, Kopargaon, India Email: anil.pawar1983@gmail.com.

High dimensionality can be addressed using feature reduction and feature selection algorithms.In this research article, high dimensionality issue of preprocessing is addressed by proposing novel feature selection approach using correlation (Cor) coefficient and Symmetrical uncertainty.

If a greater number of independent features (may be hundreds to thousand) existed in a dataspace is called high. dimensional data [3]. Search space for the given problem will be increases as number of features are more in the dataspace. This will create a problem to face curse of dimensionality. This type of datasets consumes more amount of memory during the classification model generation. In case of lazy learners, it is very difficult for generating classification model because model is not created in advance. The classification model generated over high dimensional data may perform adversely and classifier may get confuse in classifying an unknown instance.

This high dimensionality issue is very common thing which will appear in machine learning datasets. There are some techniques existed for minimizing the number of features needed for classification as high dimensional data consists of irrelevant and redundant features. Principal component (PCA) analysis is widely used feature reduction process to identify the principal components in a classification which identifies attributes that are orthogonal to others. Apart from feature reduction, feature selection is another possible solution. Filter, wrapper, and embedded are three types of feature selection modes. Filter mode gives the rank to each feature in the dataset. Depending on the threshold (number of features to be considered to generate the model (n)), top ‘n’ features will be selected as per the rank given. ReliefF, Gain ratio, Chi-Squared, Information Gain, and Symmetrical Uncertainty are some of the existed methods based on filter approach. Wrapper mode gives the subset of features directly. It uses the searching algorithm (BFS, DFS. Genetic, etc) and learning algorithm for generating subset of feature. Embedded combines both these approaches [4]. This paper is based on filter methods only.

Microarray datasets has huge number of features ranging from few hundreds to few thousand, generating a classification model on such huge dataset is challenging job. Such type of dataset analysis required a proper feature selection approach in order to reduce the memory consumption problem and increasing the classification accuracy [5]. In this research colon microarray dataset, which has 2000 features is examined with the some of the existing methods and proposed method. The same method also tested on other datasets also for generalizing it.

In this research, one of the filter-based method Symmetrical uncertainty (SU) is used along with correlation coefficient (CCE) for proposing a new method. The relation between SU and Cor are given in next section. The existed literature related to the feature selection is articulated in second part. The proposed approach is described in methodology section. The experimental result analysis output and dataset description is explained in fourth section. The article is concluded with future recommendations.

## 2. Related Work

The proposed method is based on Symmetrical uncertainty (SU) which is filter-based approach and correlation coefficient (CCE). In the existed literature, SU and Cor is applied by many researchers on various datasets. Some of the researchers considered SU and Cor separately, few are considered combined. SU can be defined as

$$SU=2*IG/(H(F1)+H(F2)) \qquad .\ (1)$$

IG is Information Gain; *H(F1)* is Entropy of F1; *H(F2)* is Entropy of F2. The value of SU ranges from 0 to 1. The featurewith high SU score is considered as strong feature.

A distributed feature selection approach based on SU is proposed by the authors to reduce the features of 6 microarray datasets [6]. A quarter feature selection approach based on SU is proposed by the researcher and examined on various datasets [7]. In their approach, researchers divided the dataset in 4 clusters (25% of features). Top 25% of features derived by some of the traditional methods are compared with their approach by applying various classifiers. One of the cluster of features derived by the researchers performing better than the existed methods. Assembling approach based on SU is proposed by the authors and examined on the sonar target classification [8]. Authors divided the sonar dataset in various clusters and applied bagging and boosting to measure the classification accuracy.

Another statistical measure used in this research is Correlation Coefficient (CCE), which is used to measure the strongness /dependency between two features. The CCE between two features can be [-1, +1]. +1 indicates, two variables are strongly/positively correlated, so one attribute can be considered and other can be ignored. CCE of two random variables X and Y can be derived as below equation

$$CCE = n\,(\sum xy) - (\sum x)\,(\sum y) \,/\, \surd[n\sum x2-\sum x2)]\ [n\sum y2-\sum y2)] \qquad .\ (2)$$

A FAST feature selection-based SU and CCE is proposed and tested over several datasets [9]. Authors achieved 85% success rate over traditional methods using FAST. Authors constructed the graph based on the values of SU and CCE, later applied the prims algorithm to find minimum spanning tree. The result of minimum spanning tree is considered as reduced subset of features. Our proposed system is inspired from FAST subset feature selection with different approach.

There some literature related to the colon cancer is available. Authors proposed feature selection approach based on clustering concept over the colon cancer to classify the gene expression and they achieved the best accuracy than some of the existing methods [10]. Authors proposed a method based on Particle Swarm Optimization (PSO) feature selection algorithm along with the Support Vector Machine classifier algorithm to get the best features from colon cancer microarray data, their method competing with other existing algorithms in accuracy [11]. Feature selection practices like genetic algorithm and mutual information are applied by the researchers for testing the cancer microarray data [12]. By using their approach most expected cancer connected genes are determined from large microarray dataset. A Multi-Objective Binary PSO (MOBPSO)algorithm is suggested for inspecting the cancer gene expression data [13].

After applying feature selection process in the pre- processing, classification algorithms are applied using selected features. For this research also various tree, lazy, rule, and functional classifiers are applied to test the accuracy of selected features. The selected features are compared with the features derived by some of the existing methods such as ReliefF, Gain ratio, Chi-Squared, Information Gain. These are based on the concept of information theory [14].

## 3. Proposed Methodology

In this research, a novel feature selection technique is proposed based on two statistical components called correlation coefficient and symmetrical uncertainty. The proposed technique is based on the steps given in algorithm.

Algorithm steps:

1. Derive the SU score of every feature and ignore the feature whose score is zero, then place remaining features it in it's descending order of SU score. Elect the middle feature's SU score as Threshold (*T*).
2. Create the CCE Symmetrical matrix *(CCE(*$X_i$*,*$Y_i$*))* of initial dataset .
3. Transfigure the *CCE(*$X_i$*,*$Y_i$*)* matrix to weighted binary matrix (WB) as per the steps given below.

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

4. Calculate the total weight of each feature by summing up all 1's related to each feature.
5. Group the features which are having the same weight(W(*F*))
6. *Cluster*$_i$*={*$F_{i1}$*,* $F_{i2}$*,...*$F_{ik}$*}*

   /* i is the cluster id, increment i by 1 until all features are formed */
7. Select the strong feature from each cluster. (i.e a feature whose SU score is greater than all of its features are nominated as strong feature)

Example

Consider the features *f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12)* in primary data set.

1. SU score of every feature is given in Table 1. (As per step 1)

**Table 1**. SU score of all features in primary data set.

| SU | Rank | Fid |
|---|---|---|
| .19 | 1 | *f10* |
| .19 | 2 | *f8* |
| .19 | 3 | *f7* |
| .18 | 4 | *f9* |
| .15 | 5 | *f2* |
| .09 | 6 | *f1* |
| .07 | 7 | *f4* |
| .06 | 8 | *f3* |
| .06 | 9 | *f5* |
| .02 | 10 | *f6* |
| 0 | 11 | *f11* |
| 0 | 12 | *f12* |

* Ignore *f11* and *f12,* as their SU score is zero

2. *Threshold (T) = .15,* as ‘b’ is the middle feature. (As perstep 2)
3. $CCE(X_i, Y_i)$ matrix of the primary data set is given in below Table 2. (As per setp3)

**Table 2.** $CCE(X_i,Y_i)$ Matrix

| *Feature Id* | *f1* | *f2* | *f3* | *f4* | *f5* | *f6* | *f7* | *f8* | *f9* | *f10* |
|---|---|---|---|---|---|---|---|---|---|---|
| *f1* | 1 | 0.21 | 0.25 | 0.22 | -0.23 | 0.01 | -0.23 | 0.06 | 0.07 | 0.09 |
| *f2* | 0.21 | 1 | -0.15 | -0.9 | 0.03 | 0.04 | -0.09 | 0.05 | 0.04 | 0.07 |
| *f3* | 0.25 | -0.15 | 1 | 0.08 | 0.05 | 0.24 | -0.09 | 0.26 | 0.01 | 0.13 |
| *f4* | 0.22 | -0.9 | 0.08 | 1 | 0.27 | 0.06 | 0.2 | -0.21 | 0.02 | 0.03 |
| *f5* | -0.23 | 0.03 | 0.05 | 0.27 | 1 | 0.03 | 0.06 | -0.1 | 0.06 | 0.17 |
| *f6* | 0.01 | 0.04 | 0.24 | 0.06 | 0.03 | 1 | 0.04 | 0.19 | 0.03 | -0.08 |
| *f7* | -0.23 | -0.09 | -0.09 | 0.2 | 0.06 | 0.04 | 1 | 0.03 | 0.19 | 0.02 |
| *f8* | 0.06 | 0.05 | 0.26 | -0.21 | -0.1 | 0.19 | 0.03 | 1 | 0.04 | 0.02 |
| *f9* | 0.07 | 0.04 | 0.01 | 0.02 | 0.06 | 0.03 | 0.19 | 0.04 | 1 | -0.06 |
| *f10* | 0.09 | 0.07 | 0.13 | 0.03 | 0.17 | -0.08 | 0.02 | 0.02 | -0.06 | 1 |

4. Convert the $CCE(X_i, Y_i)$ matrix to weighted binary matrix(WB) and calculate the total weight of each feature.

**Table 3.** Weighted Binary Matrix

| *Feature Id* | *f1* | *f2* | *f3* | *f4* | *f5* | *f6* | *f7* | *f8* | *f9* | *f10* | *Sum of Weight* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *f1* | *1* | *1* | *1* | *1* | *0* | *0* | *0* | *0* | *0* | *0* | *4* |
| *f2* | *1* | *1* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *2* |
| *f3* | *1* | *0* | *1* | *0* | *0* | *1* | *0* | *1* | *0* | *0* | *4* |
| *f4* | *1* | *0* | *0* | *1* | *1* | *0* | *1* | *0* | *0* | *0* | *4* |
| *f5* | *0* | *0* | *0* | *1* | *1* | *0* | *0* | *0* | *0* | *1* | *3* |
| *f6* | *0* | *0* | *1* | *0* | *0* | *1* | *0* | *1* | *0* | *0* | *3* |
| *f7* | *0* | *0* | *0* | *1* | *0* | *0* | *1* | *0* | *1* | *0* | *3* |
| *f8* | *0* | *0* | *1* | *0* | *0* | *1* | *0* | *1* | *0* | *0* | *3* |
| *f9* | *0* | *0* | *0* | *0* | *0* | *0* | *1* | *0* | *1* | *0* | *2* |
| *f10* | *0* | *0* | *0* | *0* | *1* | *0* | *0* | *0* | *0* | *1* | *2* |

5. Group the features which are having the same weightCluster 1 with weight 4: ***{f1, f3, f4}***

Cluster 2 with weight 3: ***{f5,f6,f7,f8}***
Cluster 1 with weight 2: ***{f2, f9, f10}***

6. Select the strong feature from each clusterStrong subset of cluster is ***{f1, f8, f10}***

## 4. Experimental Results and Discussion

The entire research experiment was carried out in the software laboratory of computer engineering department terminal. It was configured with Linux distribution and installed necessary packages to run proposed algorithmic steps via terminal. The proposed algorithm is experimented on Colon cancer microarray dataset. It has 2000 features and 2 classes. As per the steps proposed in the algorithm, firstly SU is applied and determined the features whose score is greater than zero. We could get 138 features whose SU score is greater than zero. Then, CCE matrix of those 138 features is constructed. Then, weight of each feature is calculated. After completion of the process, we could get subset of 38 features. To know the strength of the proposed method, top 38 features derived by some of the existing filter-based algorithms are taken into consideration. With these 38 features, classification models are created. For this, Jrip, Ridor (rule based), Simple cart(sc), J48 (tree based), Naive Bayes and instant based

IBK are used. This experiment is done with popular machine learning tool WEKA with all its default settings. For generating Correlation coefficient matrix, the popular statistical program R is used. Below table 4 has classification result accuracy.

**Table 4.** Classification Result Analysis

| | Jrip | Ridor | J48 | SC | NB | IBK |
|---|---|---|---|---|---|---|
| IG | 77.41 | 74.19 | 80.64 | 77.41 | 80.64 | 82.25 |
| Chi | 79.03 | 70.96 | 80.64 | 77.41 | 83.87 | 82.25 |
| Gr | 77.41 | 74.19 | 80.64 | 77.41 | 82.25 | 82.25 |
| Rel | 77.41 | 74.19 | 80.64 | 77.41 | 80.64 | 82.25 |
| Proposed | **82.25** | 70.96 | **91.93** | **83.87** | **88.87** | 79.03 |
| ALL features (SU > 0) | 75.80 | 64.51 | 82.25 | 75.80 | 53.22 | 77.41 |

The accuracy of the features derived by Information gain (IG), ReliefF (Rel), and Gain ration (Gr) is same because, top 38 features derived by those are same. The comparative analysis of proposed method and existing algorithms with various classifiers is given in Fig. 1.

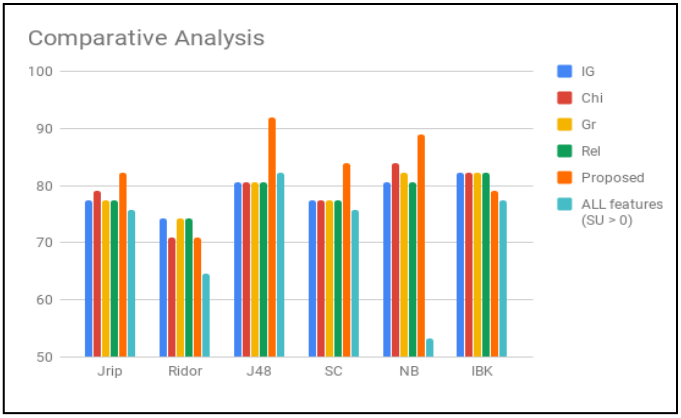

**Figure 1.** Comparative Analysis

Figure 1 show Comparative analysis reveals that, proposed method is performing better than all existing methods with Jrip, J48, and NB classifiers. It also competing with Chi and performing better if all the features are considered using Ridor. The proposed method is performing little lower than all existing, but recorded improved performance with IBK. The proposed method is tested with little variation on 10 real datasets and its performance is analyzed. After careful analysis majority of the cases the proposed method has displayed improved performance than existing methods.

## 5. Conclusion

In this research paper, a novel approach of subset selection of feature selection framework is demonstrated to reduce the dimensionality of a dataset to resolve the issues of selecting the appropriate and required strong features in high dimensionality dataset. In the experiment analysis, initially, Symmetrical Uncertainty and then and correlation coefficient are taken to select the useful, strong features. The developed method is compared with 04 existing filter-based methods as, Chi- Square, Grain Ratio, information gain, and ReliefF. For testing purpose, 06 classifiers Jrip, Ridor, J48, Simple cart, Naive Bayes, IBk are experimented on colon cancer high dimensional dataset, which has 2000 features and later demonstrated with 10 different real time data sets. After rigorous analysis, it is observed, proposed method gives promising better results over existing IG and GR on 8 data sets, also performed better on 8 data sets. Also, found more promising than ReliefF method.

## References

[1] Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, *239*,39-57.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

[3] Rehman, M. H., Liew, C. S., Abbas, A., Jayaraman, P. P., Wah,T. Y., & Khan, S. U. (2016). Big data reduction methods: a survey.*Data Science and Engineering*, *1*(4), 265-284.

[4] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 94.

[5] Al-Rajab, M., Lu, J., & Xu, Q. (2017). Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. *Computer methods and programs in biomedicine*, *146*, 11-24.

[6] Potharaju, S. P., & Sreedevi, M. (2018). Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance. *Clinical Epidemiology andGlobal Health.*

[7] Potharaju, S. P., & Sreedevi, M. (2018). A Novel Cluster of Quarter Feature Selection Based on Symmetrical Uncertainty. Gazi University Journal of Science, 31 (2), 456-470.

[8] Potharaju, S. P., Sreedevi, M., & Amiripalli, S. S. (2019). An Ensemble Feature Selection Framework of Sonar Targets Using Symmetrical Uncertainty and Multi-Layer Perceptron (SU-MLP). In *Cognitive Informatics and Soft Computing* (pp. 247-256). Springer, Singapore.

[9] Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp.856-863).

[10] Nies, H. W., Daud, K. M., Remli, M. A., Mohamad, M. S., Deris, S., Omatu, S., ... & Sulong, G. (2017, June). Classification of Colorectal Cancer Using Clustering and Feature Selection Approaches. . In *International Conference on Practical Applications of Computational Biology & Bioinformatics* (pp. 58-65). Springer, Cham.

[11] Al-Rajab, M., Lu, J., & Xu, Q. (2017). Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. Computer methods and *programs in biomedicine*, *146*, 11-24.

[12] Pavithra, D., & Lakshmanan, B. (2017, June). Feature selection and classification in gene expression cancer data. In *Computational Intelligence in Data Science (ICCIDS), 2017 InternationalConference on* (pp. 1-6). IEEE.

[13] Chandra Sekhara Rao Annavarapu, S. D., & Banka, H. (2016). Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. *EXCLI journal*, *15*,460.

[14] Hall, M. A., & Smith, L. A. (1999). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In *FLAIRS conference* (Vol. 1999, pp. 235-239).