RESEARCH ARTICLE OPEN ACCESS

A Combined Approach for Feature Subset Selection and Size Reduction for High Dimensional Data

Anurag Dwivedi, Poonam Sharma
Assistant Professor Dept. of Computer Science and Engineering SRGI, Jhansi
M. Tech Scholar Dept. of Information Technology SATI, Vidisha (M.P)

Abstract: selection of relevant feature from a given set of feature is one of the important issues in the field of data mining as well as classification. In general the dataset may contain a number of features however it is not necessary that the whole set features are important for particular analysis of decision making because the features may share the common information's and can also be completely irrelevant to the undergoing processing. This generally happen because of improper selection of features during the dataset formation or because of improper information availability about the observed system. However in both cases the data will contain the features that will just increase the processing burden which may ultimately cause the improper outcome when used for analysis. Because of these reasons some kind of methods are required to detect and remove these features hence in this paper we are presenting an efficient approach for not just removing the unimportant features but also the size of complete dataset size. The proposed algorithm utilizes the information theory to detect the information gain from each feature and minimum span tree to group the similar features with that the fuzzy c-means clustering is used to remove the similar entries from the dataset. Finally the algorithm is tested with SVM classifier using 35 publicly available real-world high-dimensional dataset and the results shows that the presented algorithm not only reduces the feature set and data lengths but also improves the performances of the classifier.
Keywords: feature selection, data reduction, clustering, fuzzy clustering, and minimum span tree.

I. Introduction

Feature selection involves identifying a subset of very useful features from the large data set that produces compatible results as the original whole set of features [18]. Feature selection is a critical subject in data mining, particularly in high dimensional applications. The selection of relevant feature is a complex problem, and finding the ideal subset of variables is viewed as NP-hard [3].Feature selection may be extremely useful approach for reducing dimensionality, removing irrelevant data and improving learning accuracy. Extensive high-dimensional data are generally sparse and contain numerous classes/groups. For instance, vast content data in the vector space show frequently contains numerous classes of documents contains a huge number of features, this property has turned into a principle instead of the special case that mostly the clustering of high dimensional data happen in subspaces of data, so subspace grouping techniques are needed in high-dimensional data clustering. Numerous subspace grouping techniques have been proposed to handle high dimensional data, used for finding clusters from subspaces of data, rather than the whole data space. These techniques can be broadly categorized two groups one is called the hard subspace clustering that searches the exact sub set of features while other is called the soft subspace clustering which assigns the weights to features. Numerous high-dimensional data sets are the mixture of extracted data from different prospective which causes unwanted features insertion for any specific analysis. In this paper, we propose a new data dimension and length reduction method for high-dimensional data set. The proposed algorithm utilizes the entropy and joint entropy estimation to detect the information gain from each feature and minimum span tree to group the similar features with that the fuzzy c-means clustering is used to remove the similar entries from the dataset.

II. Literature Review

This section presents the brief review of the related literatures available on same topic. R. Ruiz et al [6] proposed hybrid approaches to provide the possibility of efficiently applying any subset evaluator, with a wrapper model for feature subset selection problem for classification tasks.Alexandros Kalousis et al [10] presented the stability of feature selection algorithms based on the stability of the feature preferences that they express in the form of weights-scores, ranks, or a selected feature subset. Finally the examination is performed by a number of measures to quantify the stability of feature preferences and propose an empirical way to estimate them. Guangtao Wang et al [2] proposed a propositional FOIL rule based algorithm FRFS, which not only contain relevant features and excludes

irrelevant and redundant ones but also considers feature interaction, is proposed for selecting feature subset for high dimensional data. FRFS first combine the features appeared in the antecedents of all FOIL rules, achieving a candidate feature subset which excludes redundant features and reserves interactive ones. Then, identifies and removes irrelevant features by evaluating features in the candidate feature subset with a new metric Cover Ratio, and obtains the final feature subset. The supervised wrapper-based feature subset selection in datasets with a very large number of attributes. Pablo Bermejo et al [1]. A Fast Correlation-Based Filter Solution is presented Lei Yu [13] introduce a novel theory predominant correlation, and propose a fast filter technique which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. The efficiency and effectiveness of their method is demonstrated through extensive comparisons with other methods using real-world data of high dimensionality. The author also proposed Relevance and Redundancy based technique in [17] which show that feature relevance alone is insufficient for efficient feature selection of high-dimensional data and define feature redundancy and propose to perform explicit redundancy analysis in feature selection.

III. Terminology Explanations

This section explains the terms and operation used in the proposed algorithm.

A. Symmetric Uncertainty:

It is derived from the mutual information by normalizing it to the entropies of variables, and can be used as the measure of correlation between either two features or a feature and the target classes. Mathematically it is defined as,

$$SU(X,Y) = \frac{2 \times Gain(X|Y)}{H(X) + H(Y)}, \dots\dots\dots\dots\dots\dots\dots (3.1)$$

Where $H(X)$ is the entropy of the variable X, and is calculated as follows:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x), \dots\dots\dots\dots\dots\dots (3.2)$$

Here $p(x)$ is the probability of the occurrence of value x of a feature f with the domain X, and can be calculated as

$$p(x) = \frac{occurance\ of\ x}{size\ of\ dataset}, \dots\dots\dots\dots\dots\dots\dots (3.3)$$

B. Information Gain

It represents the mutual information which can be gained by one variable by observing the other variable. Practically it is measured as the reduction in entropy of a certain variable by the knowledge of other, for example let we have to calculate the information gain about the variable Y by some other variable X then it will be represented as $Gain(X|Y)$ and can be calculated as

$$\begin{aligned} \text{Gain(Y|X)} &= \text{H(X)} - \text{H(X|Y)} \\ &= \text{H(Y)} - \text{H(Y|X)}, \dots\dots\dots (3.5) \end{aligned}$$

Where $H(Y|X)$ is the conditional entropy and interpreted as the remaining entropy of variable Y if the value of another variable X is known, on the basis of probability it can be given as

$$\begin{aligned} &H(Y|X) \\ &= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x), \dots\dots\dots\dots (3.6) \end{aligned}$$

Where $p(y|x)$ is conditional probability of occurrence of value y of the feature f_i with domain Y together with occurrence of value x of the feature f_j with domain X. As the equation (3.6) shows information gain is a symmetrical measure hence $G(Y|X) = G(X|Y)$.Since the information gain is a symmetrical measure hence according to equation (3.1) the symmetrical uncertainty must also be symmetrical. The value of symmetrical uncertainty varies in the interval of [0,1], the '1' represents the complete relativeness between two variables while '0' shows the complete irrelevance.

C. Fuzzy C-Means Clustering

The clustering is defined as the process of grouping the depending upon the specific measure. It is generally defined as hard clustering or soft clustering. The fuzzy clustering fall in the second category where each data point belongs to more than one cluster an attachments with the clusters is given by membership value. The Fuzzy C-Means (FCM) algorithm is one of the most widely used fuzzy clustering algorithms. The FCM algorithm attempts to partition a finite collection of elements $X = \{x_1, x_2, x_3, \dots\dots x_n\}$ into a collection of C fuzzy clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of C cluster centers V, such that

$$V = v_i, i = 1,2,3, \dots\dots C$$

and a partition matrix U such that

$$U = u_{ij}, i = 1,2,3, \dots\dots, C, j = 1,2,3, \dots, n$$

Where u_{ij} is a numerical value in $[0, 1]$ which tells the degree to which the element x_j belongs to the i^{th} cluster.The following is a linguistic description of the FCM algorithm, which is implemented Fuzzy Logic.

Step 1: Select the number of clusters $C(2 \le c \le n)$, exponential weight$\mu(1 < \mu < \infty)$, initial partition matrixU^0, and the termination criterionϵ. Also, set the iteration index $l$to 0.
Step 2: Calculate the fuzzy cluster centers $\{v_i^l | i = 1,2,3 \dots C\}$ by usingU^l.
Step 3: Calculate the new partition matrix$U^{l+1}$by using $\{v_i^l, i = 1,2,3 \dots C\}$.
Step 4: Calculate the new partition matrix $\Delta = \|U^{l+1} - U^l\| = \max_{ij} |u_{ij}^{l+1} - u_{ij}^l|$. If$\Delta > \epsilon$, then set $l = l + 1$ and go to step 2.If$\Delta \le \epsilon$, then stop.

IV. Proposed algorithm

The proposed algorithm can be explained as following , Let the D be the high dimensional dataset and can be expressed as

$$D = \begin{Bmatrix} d_{11}, d_{12}, d_{13}, \dots \dots \dots \dots d_{1n} \\ d_{21}, d_{22}, d_{23}, \dots \dots \dots \dots d_{2n} \\ \cdot \\ \cdot \\ \cdot \\ d_{m1}, d_{m2}, d_{m3}, \dots \dots \dots \dots d_{mn} \end{Bmatrix}, T \begin{Bmatrix} t_c \\ t_c \\ \cdot \\ \cdot \\ \cdot \\ t_c \end{Bmatrix} \dots . (4.1)$$

Hence the dataset has n dimensions and m entries which targets to class$t_c, t_c \in T, T = \{t_1, t_2, \dots t_C\}$ where C is the total number of target classes. The objective of the problem is to find a subset $D'$of data D such that D' have the $m' \times n'$ dimensions and $m' < m, n' < n$.

At the first step of the algorithm the relation between the each feature and the target class is estimated by calculating the symmetric uncertainty. Let the symmetric uncertainty between i^{th}feature and target classes is given by $SU(F_i, C)$. Since the $SU(F_i, C)$ shows the predictability of target classes by the feature F_i this can be used as first measure to remove unwanted features by defining that the feature F_i is important if and only if satisfies

$$SU(F_i, C) > \theta \dots \dots \dots \dots \dots \quad (4.2)$$

Where θ is the user defined constant can be seen as the minimum required relation between feature and target class.
After performing the above discussed operation the dimension of feature will change let it be n_1 then

$$D_1 = \begin{Bmatrix} d_{11}, d_{12}, d_{13}, \dots \dots \dots \dots d_{1n_1} \\ d_{21}, d_{22}, d_{23}, \dots \dots \dots \dots d_{2n_1} \\ \cdot \\ \cdot \\ \cdot \\ d_{m1}, d_{m2}, d_{m3}, \dots \dots \dots \dots d_{mn_1} \end{Bmatrix}, T \begin{Bmatrix} t_c \\ t_c \\ \cdot \\ \cdot \\ \cdot \\ t_c \end{Bmatrix}, n_1 < n \dots \dots \dots \dots \dots \dots \dots \dots \dots . (4.3)$$

In the second step of algorithm the features which shares the same information are detected by calculating the symmetric uncertainty amongst each other's, and defined as $SU(F_i, F_j), i \ne j$. The concept of using the symmetric uncertainty is similar to first step, hence the features with higher values of $SU(F_i, F_j)$ may considered as identical features now a $SUFF$ matrix is calculated as

$$\begin{Bmatrix} \sim, SU(F_1, F_2), \dots \dots \dots . \dots \dots SU(F_1, F_{n'}) \\ SU(F_2, F_1), \sim, \dots \dots \dots \dots \dots . SU(F_2, F_{n'}) \\ \cdot \\ \cdot \\ \cdot \\ SU(F_m, F_1), SU(F_m, F_2), \dots \dots \dots \dots . \dots, \sim \end{Bmatrix} \dots . (4.4)$$

The $SUFF$ matrix is used to calculate the Minimum Span Tree (MST) such that the value of every elements of the matrix $SUFF$ is taken as the bonding between the corresponding features. However the MST can group every feature which have relation strength ($SUFC$) greater than zero. To eliminate the loosely connected features another condition is applied in which if the feature has greater relation with target classes ($SUFC$) than any other feature ($SUFF$)then the link between these feature is removed by modifying the $SUFF$ as below

$$SU(F_i, F_j) = 0, if\ SU(F_i, C) > SU(F_i, F_j) \dots (4.5)$$

After modifying the $SUFF$ according to equation (4.5) the MST is reconstructed and the features still found connected are considered as similar features and replaced by single representative feature depending upon their target class relation or $SUFC$ value as follows
Let the $F_t = \{F_a, F_b, F_c\}$ be the connected feature in the MST then the representative feature (F_r) will be selected as

$$F_r = F_t(i), \underset{i}{\text{argmax}}\{SU(F_a, C), SU(F_b, C), SU(F_c, C)\} \quad (4.6)$$

Now we have the most useful features set and the new dataset can be presented as$D_{m \times n''}$, here the $n'' < n$ hence the dimension of feature is reduced although the entries in the dataset is still same which needed to be reduced. In the proposed system the third step is used for this purpose which groups the similar data points using the fuzzy c-means clustering and then the points having large membership value with any one cluster is replaced by cluster centroid.

Let the fuzzy C-Means clusters the given data into k groups then after clustering the data can be presented as

$$Membership\ Matrix = \begin{Bmatrix} M_{11}, M_{12}, M_{13}, \dots\dots\dots\dots M_{1m} \\ M_{21}, M_{22}, M_{23}, \dots\dots\dots\dots M_{2m} \\ . \\ . \\ . \\ M_{k1}, M_{k2}, M_{m3}, \dots\dots\dots\dots M_{km} \end{Bmatrix} \dots (4.7)$$

Where M_{ij} shows the membership of j^{th} entry (data point) to i^{th} cluster.

$$D_1(i) = C_i, \underset{i}{\text{argmax}}\{M_{ij}, 1 \le j \le k\} \dots\dots\dots (4.8)$$

How many points will be substituted is depends upon the user defined minimum merging similarity. The complete algorithm can be described in following steps

Step 1: calculate $SUFF$ matrix and $SUFC$ matrix for given dataset D having target classes T.
Step 2: on the basis of $SUFC$ values reject the features having $SUFC$ value lesser than threshold θ.
Step 3: recalculate the $SUFF$ matrix for reduced features $SUFF'$.
Step 4: construct the minimum span tree (MST).
Step 5: remove the branches of MST having $SUFC(i, C) > SUFF(i, J)$.
Step 6: selected the isolated features from MST and generate representative for non-isolated features.
Step 7: form new dataset with only features selected in step 6.
Step 8: performed fuzzy c-means clustering and take the cluster centroid as a representative for data points having greater membership with it.
Step 9: Train the classifier from the dataset obtained and test for accuracy.

V. Simulation Results

In this section, we present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy. The dataset used has 36 features all with domain size of 2 while the target class also has a domain size of 2. The total number of entries in the dataset is 3196. For the testing of the selected features quality the classification test is performed using the probabilistic neural network

Table 1: Results for Dataset Size vs. processing time.

Dataset Size (%)	Processing Time (Seconds)
40	3.6514
60	4.6715
80	5.5949
100	6.5218

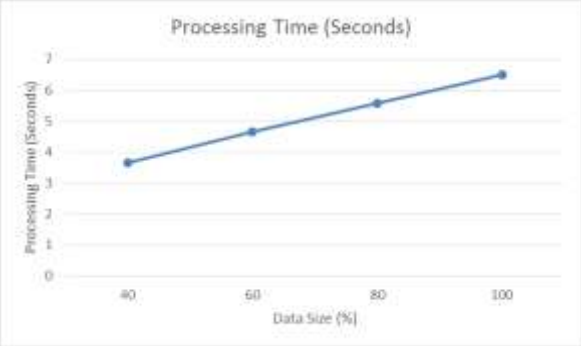

Figure 1: Plot of the table 1 data (impact of data size on processing time.

Table 2: Results for Dataset Size vs. Reduced Data Size of Previous Method and Proposed Method.

Dataset Size (%)	Data Size	
	Previous Method	Proposed Method
40	1278	1151
60	1918	1726
80	2557	2301
100	3196	2876

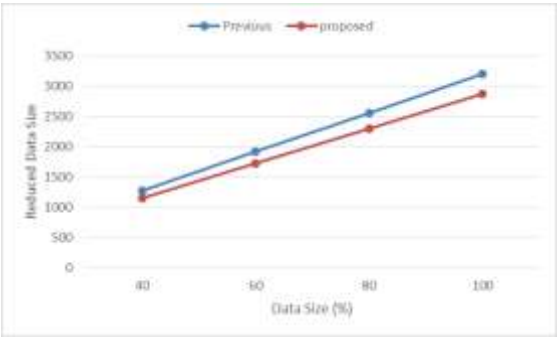

Figure 2: Plot of the table 2 comparison for data size reduction of Previous and Proposed Method.

Table 3: Results for Dataset Size vs. Selected Features of Proposed Method

Dataset Size (%)	Number of Features
40	14
60	17
80	15
100	13

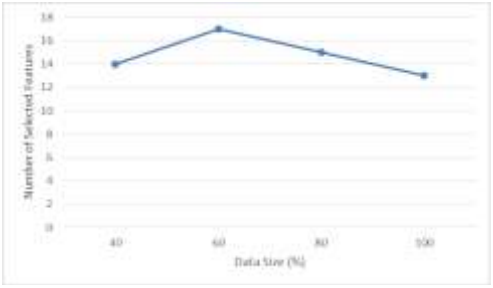

Figure 3: Plot of the table 3 data size vs. number of selected features.

Table 4: Results for Dataset Size vs. Classification Accuracy of Previous and Proposed Method.

Dataset Size (%)	Accuracy	
	Previous Method	Proposed Method
40	74.5	77.8
60	80.2	82.4
80	77.6	83.1
100	81.9	83.3

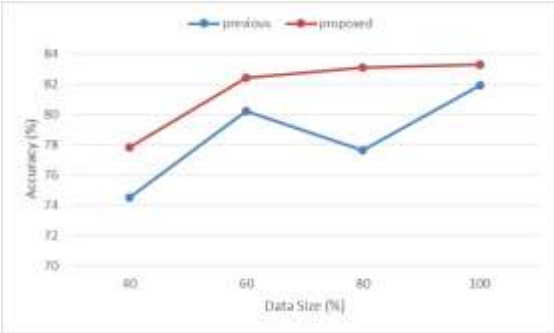

Figure 4: Plot of the table 4

VI. Conclusion

In this paper, we present a novel information theory and fuzzy clustering based feature subset selection algorithm in combination with data size reduction, which is very applicable, especially to high-dimensional data. This algorithm is developed for not only identifying and removing irrelevant and redundant features, but also dealing with interactive features. We first defined relevant, redundant and interactive features based on symmetric uncertainty then based on these definitions, we presented the feature selection algorithm, which involves four steps (1) redundant feature exclusion and interactive feature reservation and (2) the irrelevant feature identification (3) minimum span tree formation for similar features grouping (4) fuzzy C-means clustering for data size reduction. We also explained the concept behind the redundant as well as irrelevant features and reserve interactive features with appropriate expression formations. Finally the test with real world data sets show that our proposed algorithm has moderate reduction capability. Meanwhile, it also reduces the data size and obtains the best average accuracies for all the neural network based classification algorithms.

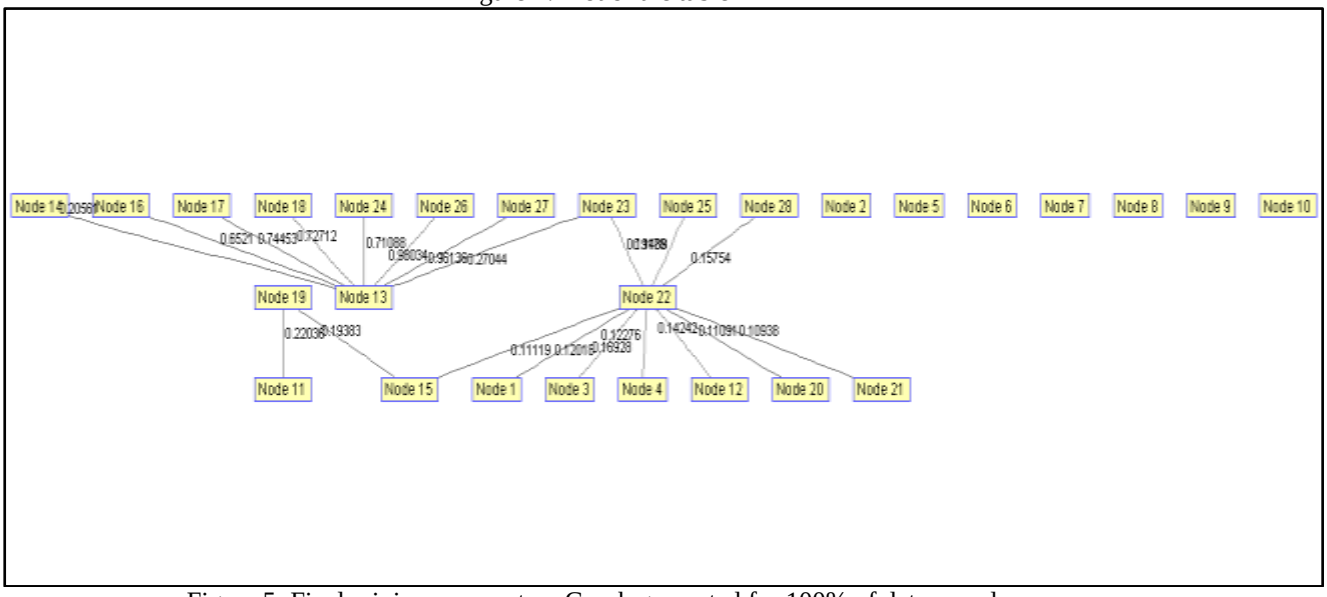

Figure 5: Final minimum span tree Graph generated for 100% of data samples

References

[1] Pablo Bermejo, Luis de la Ossa, Jos´e A. G´amez, Jos´e M. Puerta "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking",Knowledge-Based Systems Volume 25, Issue 1, February 2012, Pages 35–44Special Issue on New Trends in Data Mining.

[2] Guangtao Wang, Qinbao Song, Baowen Xu, Yuming Zhou "Selecting feature subset for high dimensional data via the propositional FOIL rules" Pattern Recognition 46 (2013) 199–214.

[3] Sebastián Maldonado, Richard Weber, Fazel Famili "Feature selection for high-dimensional class-imbalanced datasets using Support Vector Machines", Information Sciences 286 (2014) 228–246.

[4] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Senior Member, IEEE, Jennifer Spielman, Lorraine O. Ramig "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease", IEEE Trans Biomed Eng. 2012 May;59(5):1264-71.

[5] Alok Sharma, Seiya Imoto, and Satoru Miyano "A top-r Feature Selection Algorithm for Microarray Gene Expression Data",Computational Biology and Bioinformatics, IEEE/ACM Transactions on (Volume:9 , Issue: 3), 22 March 2012

[6] R. Ruiz, J.C. Riquelme, J.S. Aguilar-Ruiz, M. García-Torres "Fast feature selection aimed at high-dimensional data viahybrid-sequential-ranked searches", Expert Systems with Applications 39 (2012) 11094–11102.

[7] Xiaojun Chen, Yunming Ye, Xiaofei Xu, Joshua ZhexueHuang "A feature group weighting method for subspace clusteringof high-dimensional data", Pattern Recognition 45 (2012) 434–446.

[8] Yuzong Liu, Kai Wei, Katrin Kirchhoff, Yisong Song, Jeff Bilmes "Sub modular Feature Selection for High-Dimensional Acoustic Score Spaces", Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on26-31 May 2013

[9] Mohak Shah, Mario Marchand, and Jacques Corbeil "Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data", Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume: 34, Issue: 1), 17 November 2011

[10] Alexandros Kalousis, Julien Prados, Melanie Hilario "Stability of Feature Selection Algorithms:a study on high dimensional spaces", Knowledge and Information Systems table of contents archive Volume 12 Issue 1, May 2007.

[11] Qiang Cheng*, Hongbo Zhou, and Jie Cheng "The Fisher-Markov Selector: Fast Selecting Maximally Separable Feature Subset for Multi-Class Classification with Applications to High-Dimensional Data"Pattern Analysis and Machine Intelligence, IEEE Transactions on (Volume:33 , Issue: 6)19 April 2011.

[12] Yongjun Piao, Minghao Piao, Kiejung Parkand Keun Ho Ryu "An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data", Vol. 28 no. 24 2012, pages 3306–3315.

[13] Lei Yu, Huan Liu "Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

[14] Lance Parsons, Ehtesham Haque, Huan Liu "Subspace Clustering for High Dimensional Data: A Review", ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets Homepage table of contents archive Volume 6 Issue 1, June 2004.

[15] Daphne Koller Mehran Sahami "Toward Optimal Feature Selection", Technical Report. Stanford Info Lab.

[16] Isabelle Guyon, Andr´e Elisseeff "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1157-1182.

[17] Lei Yu, Huan Liu "Efficient Feature Selection via Analysis of Relevance and Redundancy",Journal of Machine Learning Research 5 (2004) 1205–1224.

[18] Qinbao Song, Jingjie Ni, and Guangtao Wang "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 1, January 2013.