International Journal of Computational Intelligence and Informatics, Vol. 1 : No. 1, April - June 2011

# Hybridized Oscillating Search Algorithm for Unsupervised Feature Selection

**D. Devakumari**
Department of Computer Science, L.R.G. Government Arts College for Women,
Tirupur, Tamil Nadu, India.
ramdevshri@gmail.com

## Abstract

In feature selection, a search problem of finding a subset of features from a given set of measurements has been of interest for a long time. However, unsupervised methods are scarce. An unsupervised criterion, based on SVD-entropy (Singular Value Decomposition), selects a feature according to its contribution to the entropy (CE) calculated on a leave-one-out basis. Based on this criterion, this paper proposes a Hybridized Oscillating Search feature selection method (HOS) which does not follow a pre defined direction of search (forward or backward). It is a randomized search method which begins with a random subset of features. The proposed HOS method makes use of a sequential feature selection method called Simple Ranking based on CE to get the initial feature subset. Repeated modification of the subset is achieved through up and down swings which form the oscillating cycles. The up swing adds good features to the current subset while the down swing removes worst features from the current subset. After each oscillating cycle, the subset is evaluated by comparing its predictive accuracy with known classification. Common indices like Rand Index and Jaccard Coefficient are used for this purpose. If the last oscillating cycle did not find a better subset, then the process ends with the current subset.

**Keywords**: Unsupervised Feature Selection, Contribution Entropy (CE), Hybridized Oscillating Search (HOS), Simple Ranking (SR), Rand Index, Jaccard Coefficient.

## 1. Introduction

Feature selection involves selecting a particular set of features of the original problem. Feature filtering is a process of selecting features without referring back to the data classification or any other target function. Hence we find filtering as a more suitable process that may be applied in an unsupervised manner [1].

Unsupervised feature selection algorithms are quite different from the major bulk of feature selection studies that are based on supervised methods [2, 3], and compared to the latter are relatively overlooked [4]. Unsupervised studies, unaided by objective functions, may be more difficult to carry out, nevertheless they convey several important theoretical advantages in contrast to supervised feature selection that may be unable to deal with a new class of data [5, 6, 7].

Existing methods of unsupervised feature filtering include ranking of features according to range or variance [2, 8], selection according to highest rank of the first principal component [9, 10] and other statistical criteria. An intuitive, efficient and deterministic principle, depending on authentic properties of the data, which serves as a reliable criterion for feature ranking is based on SVD-entropy, selecting a feature according to its contribution to the entropy (CE) calculated on a leave-one-out basis [11] . It has been demonstrated that this principle can be turned into efficient and successful feature selection methods like simple ranking according to higher CE values (SR), sequential forward selection by accumulating features according to which set produces highest entropy (SFS1), sequential forward selection by accumulating features through the choice of the best CE out of the remaining ones (SFS2), sequential backward elimination (SBE) of features with the lowest CE.

Most of the above mentioned sequential search strategies are based on step-wise adding of features to initially empty feature set, or step-wise removing features from the initial set of all features. One of the search directions, forward or backward, is usually preferred, depending on several factors, the expected difference between the original and the final required cardinality being the most important one. Regardless of the direction, it is apparent, that all these algorithms spend a lot of time testing feature subsets having cardinalities far distant from the required cardinality [12].

Here, a search method is presented, the Hybridized Oscillating Search Feature Selection (HOS). This is a randomized Search method which begins with an initial subset of features and adds / removes features to / from this initial set.

The organization of the paper is as follows: Section 2 presents the background of the proposed work. The proposed work is discussed in section 3. The experimental results are provided in section 4. Analysis and Discussion of the results are provided in section 5. This paper concludes in section 6.

## 2 Background

The Oscillating Search is based on repeated modification of current subset $X_d$ of $d$ features. This is achieved by alternating the down-swings and up-swings. The down – swing removes o worst features from the current set $X_d$ to obtain a new set $X_{d-o}$ at first, then adds o best features to $X_{d-o}$ to obtain a new current set $X_d$. The up-swing adds o good features to the current set $X_d$ to obtain a new set $X_{d+o}$ at first, then removes o bad ones from $X_{d+o}$ to obtain a new current set $X_d$ again. Let us denote two successive opposite swings as an oscillation cycle. Using this notion, the oscillating search consists of repeated oscillation cycles [12].

Every oscillation algorithm assumes the existence of some initial set of d features. Obtaining such an initial set will be denoted as an initialization. Oscillating algorithms may be initialized in different ways: the simplest ways are random selection or the forward selection procedure. From this point of view the oscillating search may serve as a mechanism for tuning solutions obtained in another way [12].

To decide on the best and worst features, some unsupervised feature selection criterion has to be used.

Let us consider a dataset of n instances and $m$ features $A_{[nXm]} = \{A_1, A_2, \ldots, A_i, \ldots, A_n\}$ , where each instance, or observation, $A_i$ is a vector of m measurements or features. The objective is to obtain a subset of features of size $m_c < m$, that, in a sense to be defined below, best represents the data. Alter et al., [13] have defined a SVD (singular value decomposition) based entropy of the dataset. Denote by $S_j$ the singular values of the matrix $A$. $S_j^2$ are then the eigen values of the $n \times n$ matrix $A * A^t$. Let us define the normalized relative values (Wall, M., Rechtsteiner, A. and Rocha, L., 2003):

$$V_j = {s_j}^2 / \sum_k {s_k}^2 \qquad (1)$$

and the resulting dataset entropy (Alter, O., Brown, P.O. and Botstein, D., 2000):

$$E = -\frac{1}{\log(N)} \sum_{j=1}^{N} V_j \log(V_j) \qquad (2)$$

This entropy varies between 0 and 1. E = 0 corresponds to an ultra ordered dataset that can be explained by a single eigenvector (problem of rank 1), and E = 1 stands for a disordered matrix in which the spectrum is uniformly distributed.

The contribution of the $i^{th}$ feature to the entropy $(CE_i)$ is defined by a leave-one-out comparison according to

$$CE_i = E\,(A_{[nXm]}) - E\,(A_{[nX(m-1)]}) \qquad (3)$$

where the $i^{th}$ feature was removed in $A_{[nX(m-1)]}$ . Let us define the average of all CE to be $c$. We distinguish then, between three groups of features:

(i) $CE_i > c$, features with high contribution

(ii) $CE_i = c$, features with average contribution

(iii) $CE_i < c$, features with low (usually negative) contribution

Let $m_c$ represent the number of features whose CE value is greater than the average of all the CE values. Then Entropy maximization can be implemented in three different ways [7]:

i) Simple ranking (SR) – Select $m_c$ features according to the highest ranking order of their $CE_i$ values.

ii) Sequential Forward Selection (SFS) – Choose the first feature according to the highest CE. Recalculate the CE values of the remaining features and select the second feature according to the highest CE value. Continue the same way until $m_c$ features are selected.

iii) Sequential Backward Elimination (SBE) – Eliminate the feature with the lowest CE value. Recalculate the CE values and iteratively eliminate the lowest one until $m_c$ features remain.

## 3 Proposed Work

Hybridized Oscillating Search feature selection method (HOS) does not follow a pre defined direction of search (forward or backward). It is a randomized search method which begins with a random subset of features. The proposed HOS method makes use of a sequential feature selection method called Simple Ranking based on the Contribution Entropy (CE) value [7] to get the initial feature subset. Repeated modification of the subset is achieved through up and down swings which form the oscillating cycles. The up swing adds good features to the current subset while the down swing removes worst features from the current subset. After each oscillating cycle, the subset is evaluated by comparing its predictive accuracy with known classification[14, 15]. Common indices like Rand Index and Jaccard Coefficient can be used for this purpose. If the last oscillating cycle did not find a better subset, then the process ends with the current subset. The pseudo code of HOS method is given in Fig. 1

Let $Y$ be the given data set with $D$ features.

Let ADD(o) represents adding of $o$ features and REMOVE(o) represents removing of $o$ features.

Let $R$ represents the Rand Index score and $J$ represents the Jaccard Coefficient score of $Y_D$.

1. Calculate the CE value for each feature in $Y$.
2. Find the initial sub set $X_d$ of d features using Simple Ranking (SR) method.
3. Calculate Rand score $R_1$ or Jaccard score $J_1$ for $X_d$.

4. While ($R_1 > R$) || ($J_1 > J$) do
    a) Perform REMOVE(o) to remove features in $X_d$ with lowest CE value and generate a new feature subset $X_{d-o}$.
    b) $R = R_1$, $J = J_1$, $Xd = X_{d-o}$
    c) Goto step 3.
5. While ($R_1 < R$ || $J_1 < J$) do
    a) Perform ADD(o) to select features from $Y_D - X_d$ with highest CE value and add them to $X_d$ to generate a new feature subset $X_{d+o}$.
    b) $R = R_1$, $J = J_1$, $Xd = X_{d+o}$
    c) Goto step 3.
6. End

Fig. 1 Pseudo code for Hybridized Oscillating Search (HOS) method

## 4 Experimental Results

The HOS method of feature selection is experimented with four different data sets from the UCI machine learning repository [www.archive.ics.uci.edu]. The experimental procedure and the results obtained are explained below for each of the chosen data set.

### 4.1 Lung Cancer Data set

This data set contains 56 features and 32 instances. The initial feature subset selected through Simple Ranking (SR) method contains 24 features. The Rand index value is calculated for this initial subset of features and they are found to be more than the index value calculated with all features. Hence features with lowest CE value are to be removed from the initial set. Feature number 8 of the initial set has the lowest CE value 0.0002 and it is removed. Now the feature subset contains 23 features and the indices are calculated again. They are again found to be greater than the previous values and hence features with lowest CE value are to be removed from the current subset. Features 31 and 32 of the current subset have lowest CE value 0.0005 and are removed. Now the feature subset contains 21 features and the indices are calculated again. They are found to be lesser than the previous values and hence the removed features 31 and 32 are added to the current feature set and this becomes the finally selected feature subset. The results are tabulated in Table. 1

Table 1. Experimental results for Lung Cancer Data set

| | No. of Features | Rand Index |
|---|---|---|
| Known Classification | 56 | 0.6379 |
| All Features | 56 | 0.6049 |
| Initial Feature subset | 24 | 0.6543 |
| Iteration I | 23 | 0.6927 |
| Iteration II | 21 | 0.6601 |
| Iteration III | 23 | 0.6927 |

### 4.2 Cardiac Tomography Data set

This data set contains 44 features and 187 instances. The initial feature subset selected through Simple Ranking (SR) method contains 17 features. The Rand index value is calculated for this initial subset of features and they are found to be less than the index values calculated with all features. Hence features with CE value smaller than 0.0538 (the lowest CE value in the initial feature subset) from the original feature set are to be added to the initial set. Feature number 36 of the original set has the CE value 0.0317 and it is added to the initial subset. Now the feature subset contains 18 features and the indices are calculated again. They are again found to be lesser than the previous values and hence features with CE value smaller than 0.0317 (the lowest CE value in the current subset) are to be added to the current subset. Feature 15 of the original set has the CE value 0.0102 and it is added to the current subset. Now the feature subset contains 19 features and the indices are calculated again. They are again found to be lesser than the previous values and hence features with CE value smaller than 0.0102 (the lowest CE value in the current subset) are to be added to the current subset. Feature 18 of the original set has the CE value 0.0032 and it is added to the current subset. Now the feature subset contains 20 features and the indices are calculated again. Now the Rand index value is same as the previous iteration which means that the feature subset with 19 features is the finally selected feature subset. The results are tabulated in Table 2:

Table 2. Experimental results for Cardiac Tomography Data set

| | No. of Features | Rand Index |
|---|---|---|
| Known Classification | 44 | 0.5938 |
| All Features | 44 | 0.5938 |
| Initial Feature subset | 17 | 0.5604 |
| Iteration I | 18 | 0.5681 |
| Iteration II | 19 | 0.5762 |
| Iteration III | 20 | 0.5762 |
| Iteration IV | 19 | 0.5762 |

### 4.3 Dermatology Data set

This data set contains 33 features and 366 instances. The initial feature subset selected through Simple Ranking (SR) method contains 26 features. The Rand index value is calculated for this initial subset of features and they are found to be more than the index values calculated with all features. Hence features with lowest CE value are to be removed from the initial set. Features 12, 16 and 20 of the initial set have the lowest CE value 0.0013 and they are removed. Now the feature subset contains 23 features and the indices are calculated again. They are found to be lesser than the previous values and hence the removed features 12, 16 and 20 are added to the current feature subset. Now the feature subset contains 26 features and this becomes the finally selected feature subset. The results are tabulated in Table 3:

Table 3. Experimental results for Dermatology Data set

| | No. of Features | Rand Index |
|---|---|---|
| Known Classification | 33 | 0.7755 |
| All Features | 33 | 0.6056 |
| Initial Feature subset | 26 | 0.6358 |
| Iteration I | 23 | 0.5826 |
| Iteration II | 26 | 0.6358 |

### 4.4 Ionosphere Data set

This data set contains 34 features and 351 instances. The initial feature subset selected through Simple Ranking (SR) method contains 16 features. The Rand index value is calculated for this initial subset of features and they are found to be less than the index values calculated with all features. Hence features with CE value smaller than 0.0053 (the lowest CE value in the initial feature subset) from the original feature set are to be added to the initial set. Feature number 31 of the original set has the CE value 0.0012 and it is added to the initial subset. Now the feature subset contains 17 features and the indices are calculated again. They are found to be more than the previous values. Hence feature 31 with lowest CE value is to be removed from the current set. But then the Rand Index value will become less. Hence the feature set with 17 features is the selected subset. The results are tabulated in Table 4:

Table 4. Experimental results for Ionosphere Data set

| | No. of Features | Rand Index |
|---|---|---|
| Known Classification | 34 | 0.5901 |
| All Features | 34 | 0.5901 |
| Initial Feature subset | 16 | 0.5068 |
| Iteration I | 17 | 0.5132 |
| Iteration II | 16 | 0.5068 |
| Iteration III | 17 | 0.5132 |

## 5 Conclusion

A novel principle for unsupervised feature filtering is based on maximization of SVD-entropy. The features are ranked according to their CE values. Based on this principle, four feature selection methods have already been implemented. This paper proposes the Hybridized Oscillating Search feature selection (HOS) method in which no pre defined direction of search (forward or backward) is followed. The proposed HOS method makes use of a sequential feature selection method called Simple Ranking based on the Contribution Entropy (CE) value to get the initial feature subset. Repeated modification of the subset is achieved through up and down swings which form the oscillating cycles. After each oscillating cycle, the subset is evaluated by comparing its predictive accuracy with known classification. Common indices like Rand Index and Jaccard Coefficient are used for this purpose. The proposed algorithm is experimented with bench mark data sets and the results are analysed.

## References


[1] Julia Handl and Joshua Knowles, “Feature Subset Selection in Unsupervised Learning via Multiobjective Optimization”, International Journal of Computational Intelligence Research, Vol.2, No.3, pp. 217-238, 2006.

[2] Guyon, I. and Elisseeff, A, “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research, 3, 1157—1182, 2003.

[3] Liu, H., Li, J. and Wong, L., “A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns”, In R. Lathrop, K.N., S. Miyano, T. Takagi, and M. Kanehisa (ed), 13th International Conference on Genome Informatics, Universal Academy Press, Tokyo Japan, 51-60, 2002.

[4] Huan Liu and Lei Yu, “Towards Integrating Feature Selection Algorithms for Classification and Clustering’, IEEE Transactions on Knowledge and Data Engineering, Vol.17, pp 491-502, 2005.

[5] J. G. Dy and C. E. Brodley, “Feature selection for unsupervised learning”, Journal of Machine Learning Research, pp 845.889, 2004.

[6] P. Mitra, C. A. Murthy, and S. K. Pal, “Unsupervised feature selection using feature similarity”, IEEE Transactions on Pattern Analysis and Machine intelligence, pp 301.312, 2002.

[7] N. Sondberg-Madsen, C. Thomsen, and J. M. Pena, “Unsupervised feature subset selection”, Proceedings of the Workshop on Probabilistic Graphical Models for Classification, pp 71.82, 2003.

[8] Herrero, J., Diaz-Uriarte, R. and Dopazo, J. “Gene expression data preprocessing”, Bioinformatics, 19, 655-656, 2003.

[9] Ding, C.H.Q. “Unsupervised Feature Selection Via Two-way Ordering in Gene Expression Analysis”, Bioinformatics, 19, 1259-1266. 2003.

[10] Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D. and Brown, P. “Gene Shaving as a method for identifying distinct sets of genes with similar expression patterns, Genome Biology, 2000.

[11] Roy Varshavsky, Assaf Gottlieb, Michal Linial and David Horn, “Novel Unsupervised Feature Filtering of Biological Data”, Bioinformatics / bti283, pp 1-5, 2005.

[12] P. Somol and P. Pudil, “Oscillating Search Algorithms for feature selection”, Proceedings of the 15th International Conference on Pattern Recognition, pp 1 – 4, 2000.

[13] Alter, O., Brown, P.O. and Botstein, D. “Singular value decomposition for genome-wide expression data processing and modeling, PNAS, 97, 10101-10106. 2000.

[14] D. Guo, M. Gahegan, D. Peuquet, and A. MacEachren, “Breaking down dimensionality: An effective feature selection method for high-dimensional clustering”, Proceedings of the Third SIAM International Conference on Data Mining, pp 29.42, 2003.

[15] M. Dash and H. Liu, "Handling large unsupervised data via dimensionality reduction", Proceedings of the ACM SIGMOD Workshop on Research Numbers in Data Mining and Knowledge Discovery, 1999.

**Bibliography**

**D. Devakumari**, has received M.Phil degree in the area of Web Server Scheduling from Manonmaniam Sundaranar University in 2003. Currently she is working as Assistant Professor in the Department of Computer Science, L.R.G. Government Arts College for Women, Tirupur, India. She is pursuing her Ph.D. in the area of Data Mining in Mother Teresa Women's University. Five of her research papers have been published in International journals. She has presented four papers in International Conferences. Her research interests include Data Pre-processing and Pattern Recognition.