International Journal of Engineering Research and General Science Volume 3, Issue 1, January-February, 2015
ISSN 2091-2730

Feature Selection Technique Using Homogeneity based cluster

Ligendra Kumar Verma[1], Kesari Verma[2] and Priyanka Tripathi[3]

[1]Department of Computer Science and Engineering, C.V. Raman University, Bilaspur, ligendra@rediffmail.com
[2.]Department of Computer Applications, National Institute of Technology, Raipur, keshriverma@gmail.com,
[3.] Department of Computer Applications, National Institute of Technology Raipur ptripathi@hotmail.com

ABSTRACT

Feature subset selection is an important problem in knowledge discovery, not only for the insight gained from determining relevant features variables, but also for the improved understandability, scalability, and, possibly, accuracy of the resulting models by reducing computational cost. It is an important challenge in many classification problems, especially for complex data like images when the number of features greatly exceeds the number of examples available. The research aims to select optimal number of relevant features by eliminating irrelevant features using clustering technique. Clustering techniques are used to form the group of objects based on the characteristics. We have used this characteristic for selecting the features. The main focus of this paper is to search the homogeneity among the features and select the representative features by eliminating remaining features from the set.

Keywords: Feature selection; Feature ranking, Cluster based feature selection

INTRODUCTION

Features are defined as a function of one or more measurements, the values of some quantifiable property of an object, computed so that it quantifies some significant characteristics of the object. A set of features that helps the model to recognize the pattern is called class label. The feature set may contain a set of irrelevant features. The irrelevant input features will induce great computational cost. Feature subset selection is the process of identifying and removing as much irrelevant and redundant information as possible. The reduction of dimensionality the data and may allow learning algorithms to operate faster, accurately more effectively. We propose a novel model for feature selection based on cluster formed by the features of data. A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other [1].

The main idea of feature selection is to choose a relevant subset of input variables by eliminating features or called dimension reduction with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. It reduced the computational cost by taking less time and memory.
The different set of features are shown in Table 1.

Sno	Feature Set	Count
1	Haralick Texture Features [5]	47
2	HoG2x2, HoG3x3	840
3.	LBP	1239
4	Sift	512
5	Gist	512
6	Total	3617

We summarize our contribution as follows.

1. We proposed a new model that eliminates the redundant set of features by using cluster algorithms.
2. In order to find optimal number of features we used silhouette algorithm that represents the disjoint set of features.

Paper is organized as follows. In section 2 we discussed related work done in this area. All the related work related to feature selection is elaborated in section 3. Section 4 describe our proposed method. Section 5 provide the experimental results. Section 6 summarizes the work and draws some conclusions.

RELATED WORK

Two approaches that enable standard machine learning algorithms to be applied to large databases are feature selection and sampling. Both reduce the size of the

database—feature selection by identifying the most salient features in the data; sampling by identifying representative examples [2]. The feature selection problem has been studied by the statistics and machine learning and data mining communities for many years like [3].

Feature selection is categorized into two category.

1. Rank based feature selection
2. Subset based feature selection

Information gains, Gain Ratio, Best First search algorithm, Chi-Square test are some specific techniques that are widely using for feature selection purpose. The details are given as below.

2. Rank Based Feature Selection

Kohavi and John [4] proposed variable ranking method for ranking the features based on their importance. Algorithm 1. demonstrate the layout of the feature ranking algorithm.

Algorithm 1 : Ranking the Features

Input : S ← set of features
Output : N ← Top n ranked features
Method

1. Features ← Evaluation_criteria(D) // Evaluation criteria on that basis the features are evaluated.
2. Rank_features← sort_descending(Features)

Return Top n features

2.1.2 Information Gain. This technique is based on decision tree induction ID3 [5] it uses information gain as its attribute selection measure. This measure is based on pioneering work by Claude Shannon from information theory. If p_i represents the number of times tuples occurred in data D. This attribute minimize the information needed to classify the tuples in the resulting partitions. The information gain is represented by equation 1.

$$\inf(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (1)$$

Splitting attribute measures, that define information needed to exact classify the data is defined by equation 2.

$$\inf_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \inf(D_j) \quad (2)$$

Information gain is difference between original information and information after splitting is defined in equation 3.

$$Gain(A) = \inf(D) - \inf_A(D) \quad (3)$$

In this technique the features which have highest information will be ranked high otherwise low. Using Quinlan C4.5 algorithm [5] the attribute that are in the higher level of the tree are considered for further classification and these features have more importance.

2.1.3 Gain Ratio. C4.5 [5] a successor of ID3[6] uses, an extension to information gain known as gain ration. It applies a kind of normalization to information gain using split information defined in equation 5.

$$splitInfo(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \frac{(|D_j|)}{(|D|)} \quad (4)$$

The gain ratio can be defined by equation 5. Intrinsic information: entropy of distribution of instances into branches by using equation 4.

$$GainRatio(S,A) = \frac{Gain(S,A)}{IntrinsicInfo(S,A)} \quad (5)$$

2.1.4 Random Forest Filter Breiman et. al [7] has proposed random forest algorithm, it is an ensemble approach that work as form of nearest neighbor predictor. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator [10]. Ensembles are divide-and-conquer tree based approach used to improve performance of classifier. The ensemble method is that a group of weak learners that group together and work as a strong learner to take the decision for unknown attributes.

2.1.5 Best First Search. Best first search [9] is an Artificial Intelligence search technique which allows backtracking in search path. It is a hill climbing, best first search through the search space by making change in current subsets.

2.2 Feature Subset Selection

In this approach subsets of features are selected, subset feature selection is an exhaustive search process. If data contain N initial features there exist 2^N possible subsets. Selection of features from 2^N possible subsets is an exhaustive search process that is call heuristic search algorithm.

Subsets of features are selected and analyzed their classification accuracy, if it is increasing, that feature is selected otherwise rejected, a new set of feature are participate in evaluation process.

Many feature selection routines used a wrapper approach [4] to find appropriate variables such that an algorithm that searches the feature space repeatedly fits the model with different predictor sets. The best predictor set is determined by some measure of performance. The objective of each of these search routines could converge to an optimal set of predictors. The layout of subset feature selection method is shown in Algorithm 2.

Algorithm 2: Subset feature selection

S ← All subsets {}
For each subset $s \in S$
Evaluates (s)
Return {subset}

2.3 Recursive Feature Elimination [3]

Recursive feature elimination method is based on the concept that the features are eliminated recursively till the optimal set of features are not selected from the whole set. Random forest, backword subset selection algorithm using caret, Boruta is one of the well-known techniques in R [8].

3. PROPOSED WORK

In this section we give the framework of recommended approach.

Definition 1 (Cluster): Clusters are group of similar objects that helps to discover distribution of patterns and interesting co-relation in large dataset.

Example 1 (Formation of Cluster and Initialization). Illustrate of k-means clustering for feature selection is shown in shown in Figure 1. Initialize the random features as seeds is called centroid of feature. The difference from other features with all cetnroid are computed. The feature is assigned to the cluster which have minimum distance. After each features are assigned to cenntroid, the centroid is updated. The process terminates when no more changes occurs based on grouping of features. The details of process is shown in Figure 1 and Algorithm 2.

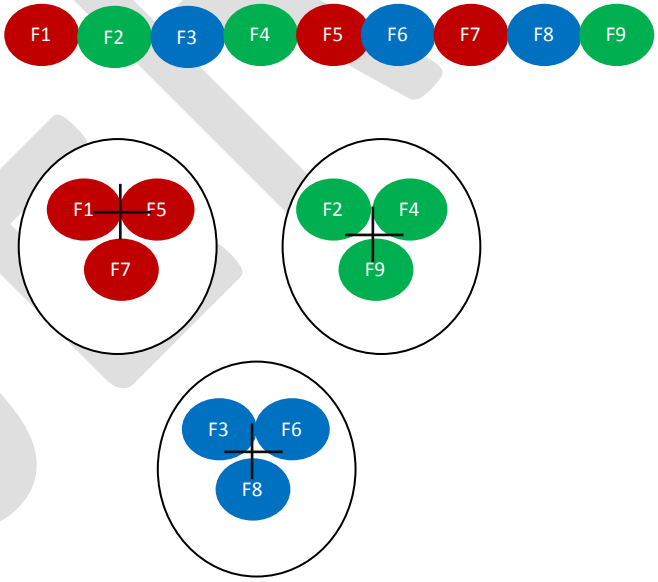

Figure 1 : Clustering based approach for feature selection

Algorithm 2 K-means algorithm for feature selection

Input : Dataset D.

1. D_{new} → transpose(D)// Feature becomes horizontal
2. Select K feature as initial centroid
3. **repeat**
4. Form k cluster by assigning each feature to closest centroid
5. Recompute the centroid of each cluster

6. **Until centroid do not change**
7. F ← representative of each cluster

---

4. Experimental Results

The experiments were performed in Windows 7 operating system, MATLAB 2012 in windows environment with 4 GB RAM and 500GB Hard disk 2.8 GHz intel processor.

Pollard et. al. [12] has defined silhouetee width which used to measure the strength of clusters using equation 1.

$$SWi = (b(i)-a(i)) / \max(a(i),b(i))$$

The concept of silhouette width involves the difference between the within-cluster tightness and separation from the rest. Specifically, the silhouette width s(i) for entity where a(i) is the average distance between i and all other entities of the cluster to which i belongs and b(i) is the minimum of the average distances between i and all the entities in each other cluster. The silhouette width values lie in the range from—1 to 1. The valid clusters are shown in Figure 2. The extracted and clustered features are shown in Table 2.

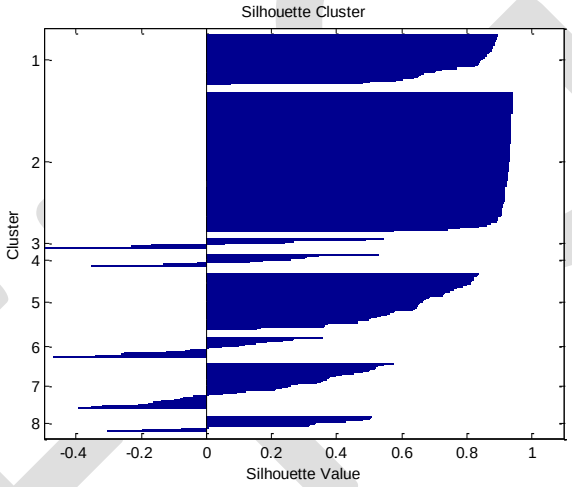

Figure 2.: Silhouetee Cluster Validation

Table 2. Clustered features based on homogeneity

Cluster 1	Cluster 2	Cluster3	Cluster 4	Cluster5	Cluster 6
18, 31,196, 197,198	14	1	8	10	7
199	57	5	13	23	12
	61	11	24	28	15
	260	16	26	36	20
	261	17	30	40	48
		21	32		55
		54	33		
		58			
		448,500			

5. CONCLUSION

In this paper we proposed a new method of feature reduction by creating the cluster of features. The proposed method reduced the number of feature based on user defined k arbitrary number. In order to find the optimal value of k we use silhouette [2] algorithm.

REFERENCES:

[1] Hall, M .A. : Correlation-based Feature Selection for Machine Learning. Ph.D. thesis in Computer Science. University of Waikato, Hamilton, New Zealand (1999)
[2] G. H. John and P. Langley. Static versus dynamic sampling for data mining. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996.
[3] YongSeog Kim, W. Nick Street, and Filippo Menczer, Feature Selection in Data Mining . (Google citation) University of Iowa,
[4] Kohavi, John, G. : Wrappers for feature subset selection, Artificial Intelligence, volume 97 Issue.1-2, p.273-324, doi>10.1016/S0004-3702(97)00043-X (1997).
[5] Quinlan, J.R.: C4.5: Programs for Machine Learning. Machine Learning, 16,235-240 Academic Kluwer Academic Publishers, Boston(1994).
[6] Quinlan, J.R :Induction of decision trees. Machine Learning, Volume 1, Number 1, pp81-106 (1986)
[7] Vladimir, S., Christopher, A.L., Tong, Wang. T. : Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. Lecture Notes in Computer Science Volume 3077, pp .334-343,(2004)
[8] Breiman, L.: Random Forests. Machine Learning, 45, pp. 5–32 (2001).
[9] Rich, E., Knight, K: Artificial Intelligence. McGraw-Hill, (1991)
[10] Ensemble method. http: //scikit-learn.org/stable/modules/ensemble.html#b2001 (Ocober,2014)
[11] POLLARD, K.S., and VAN DER LAAN, M.J. (2002), "A Method to Identify Significant Clusters in Gene Expression Data",
[12] U.C. Berkeley Division of Biostatistics Working Paper Series, p. 107. W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.