International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 05 | Aug-2015 www.irjet.net p-ISSN: 2395-0072

Feature Subset Selection using Rough Sets for High Dimensional Data

R Indra Srinivas

Faculty, Department of Information Science, BMS College of Engineering, Bangalore, India

---

Abstract - *Feature Selection (FS) is applied to reduce the number of features in many applications where data has multiple features. FS is an essential step in successful data mining applications, which can effectively reduce data dimensionality by removing the irrelevant (and the redundant) features. It has been effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving comprehensibility. The removal of irrelevant and redundant information often improves the performance of machine learning algorithms. FS techniques aim at reducing the number of unnecessary features in classification rules. The features are measured by their necessity in heuristic FS techniques. The proposed framework uses filter method to remove irrelevant features, clustering-based method to remove redundant features and Rough Set Theory (RST) with greedy heuristics for feature subset selection.*

Keywords: Feature Selection, Filter method, Rough Set, Boundary Region, High Dimensional Data, K-means Clustering, Microarray, Symmetric Uncertainty.

1. Introduction

Data Mining is a multidisciplinary effort to extract nuggets of knowledge from data. The proliferation of large data sets within many domains poses unprecedented challenges to data mining. Not only the data sets getting larger, but also new types of data have also evolved, such as data streams on the Web, microarrays in Genomics and Proteomics, and networks in Social Computing and System Biology. Researchers and practitioners are realizing that in order to use data mining tools effectively, FS is an integral component to successful Data Mining [10].

When data objects that are the subject of analysis using Machine Learning techniques are described by a large number of features, it is often beneficial to reduce the dimension of the data. Dimension reduction can be beneficial not only for reasons of computational efficiency but it can also improve the accuracy of the analysis. The set of techniques that can be employed for dimension reduction can be partitioned in two important ways; they can be separated into techniques that apply to supervised or unsupervised learning and techniques that either entails Feature Selection or Feature Extraction.

Feature Selection is a critical step for high-dimensional data classification. The benefits of FS are several-fold and dependent on the applications. For creating classification models, feature selection can often improve predictive accuracy and comprehensibility [5]. For many Bioinformatics applications, FS is a critical procedure for identifying important biomarkers.

The representation of raw data often uses many features, only some of which are relevant to the target concept. Since relevant features are often unknown in real-world problems, we must introduce many candidate features. Unfortunately redundant features degrade the performance of learners both in speed (due to high dimensionality) and predictive accuracy (due to irrelevant information). The situation is particularly serious in constructive induction, as many candidate features are generated in order to enhance the power of the representation language. FS is the problem of choosing a small subset of features that ideally is necessary and sufficient to describe the target concept.

Feature Selection is typically a search problem for finding an optimal or suboptimal subset of m features out of original M features. FS is important in many Pattern Recognition problems for excluding irrelevant and redundant features. It allows reducing system complexity as well as processing time resulting in improvement of the recognition accuracy. For large number of features, exhaustive search for best subset out of 2M possible subsets is infeasible.

The most widely used Feature Subset Selection techniques are Filter and Wrapper methods. The Wrapper algorithm use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The Wrapper methods are computationally expensive and tend to overfit on small training sets.

The Filter methods are independent of learning algorithms, offering good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed, some feature evaluation function is used rather than optimizing the classifier's performance. Many feature evaluation functions have been

used particularly functions that measure distance, information, dependency, and consistency. Wrapper methods are usually slower than Filter methods but offer better performance. The Filters methods are usually a good choice when the numbers of features are very large. Hence Filter method is used in the proposed method to remove irrelevant features.

Cluster analysis divides data into meaningful or useful groups (clusters). If meaningful clusters are the goal, then the resulting clusters should capture the "natural" structure of the data. Cluster analysis has long been used in a wide variety of fields: Psychology and other Social Sciences, Biology, Statistics, Pattern Recognition, Information Retrieval, Machine Learning, and Data Mining.

Rough Set Theory (RST) is an extension of set theory for study of the intelligent systems characterized by insufficient and incomplete information. An undefinable subset is approximately represented by two definable subsets, called lower and upper approximations. Rough Set Theory (RST) is a good candidate for classification applications. Various efforts have been made to improve the efficiency and effectiveness of classification with Rough Sets.

The concepts in Rough Set Theory (RST) are used to define the necessity of features. The measures of necessity are calculated by the functions of lower and upper approximation. These measures are employed as heuristics to guide the feature selection process. There are at least two types of heuristics, namely significance oriented method and support oriented method that has appeared in literature. The heuristic favours significant features, i.e., features causing the faster increase of the positive region. Zhong's [3] heuristic considers the positive region as well as the support of rules.

2. Related Work

There are many feature subset selection algorithms, where some can effectively eliminate irrelevant features but fail to handle redundant features. Some of other feature subset selection algorithms can eliminate the irrelevant while taking care of the redundant features. The proposed algorithm falls into the second group.

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy, and redundant features do not redound to getting a better predictor. As redundant features provide mostly information which is already present in other feature(s) [1].

However, along with irrelevant features, redundant features also affect the speed and accuracy of learning algorithms, and thus should be eliminated as well. Correlation-based Feature subset Selection (CFS) [7], Fast Correlation-Based Filter (FCBF) [9], and Conditional Mutual Information Maximization (CMIM)[8] are examples that take into consideration the redundant features. CFS [7] is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other. FCBF [9] is a fast filter method which can identify relevant features as well as redundancy among relevant features without pairwise correlation analysis. CMIM [8] iteratively picks features which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked. The proposed method employs the clustering-based method to choose features and Rough Sets to evaluate and choose the best feature subset.

3. Proposed Method

FS should be able to identify and remove as much of the irrelevant and redundant information as possible. Moreover, "good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other." [4]

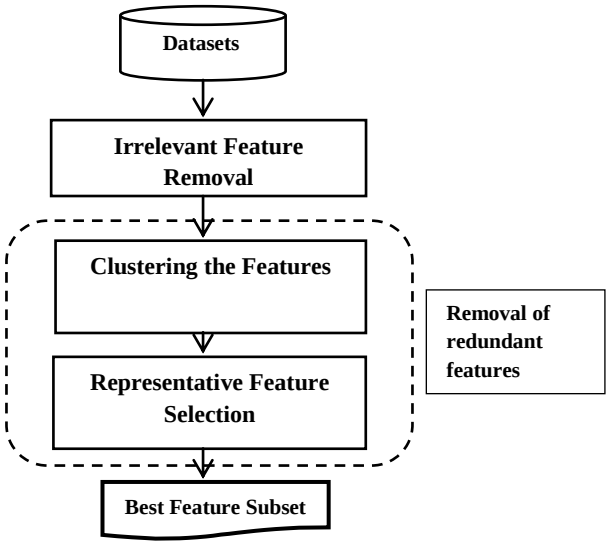

Figure 1: Framework of the proposed Feature Subset Selection.

Relevant features have strong correlation with target concept so are always necessary for a best subset, while redundant features are not because their values are completely correlated with each other. Thus, notions of feature redundancy and feature relevance are normally in

terms of feature correlation and feature-target concept correlation.

The Symmetric Uncertainty (SU) is derived from the Mutual Information by normalizing it to the entropies of feature values or feature values and target classes, and has been used to evaluate the goodness of features for classification by a number of researchers: Hall [7], Hall and Smith [10], Yu and Liu [9]. Therefore, symmetric uncertainty is chosen as the measure of correlation between either two features or a feature and the target concept.

The Symmetric Uncertainty (Press et al., 1988) is defined as follows:

$$SU(X,Y)=2 * Gain(X|Y) \div H(X) + H(Y)$$

where H(X) is the entropy of a discrete random variable X. Gain(X|Y) is the amount by which the entropy of Y decreases. Where H(X|Y) is the conditional entropy which quantifies the remaining entropy (i.e., uncertainty) of a random variable X given that the value of another random variable Y is known.

Symmetric Uncertainty treats a pair of variables symmetrically, it compensates for information gain's bias toward variables with more values and normalizes its value to the range [0,1]. A value 1 of SU(X,Y) indicates that knowledge of the value of either one completely predicts the value of the other and the value 0 reveals that X and Y are independent.

The proposed method initially removes irrelevant features using filter method. It ranks the features based on the correlation of the individual features with respect to the target class. Ranking of features based on the scores calculated using Symmetrical Uncertainty measure which ranges from 0 to 1. The feature score equal to 0 are irrelevant and feature score equal to 1 are strongly relevant features.

Features in different clusters are relatively independent; the clustering-based strategy of the proposed method has a high probability of producing a subset of useful and independent features. The representative features are clustered using K-means clustering method. The idea is to find the best division of N samples by K clusters such that the total distance between the clustered samples and their respective centers (that is, the total variance) is minimized [6].

Four different types of classification algorithms are employed to classify data sets before and after FS. They are 1) the probability-based Naive Bayes (NB), 2) the Tree-based C4.5, 3) the instance-based lazy learning algorithm IB1, and 4) the Boundary Region Rough Set, respectively.

Naive Bayes utilizes a probabilistic method for classification by multiplying the individual prob- abilities of every feature-value pair. This algorithm assumes independence among the features and even then provides excellent classification results.

Decision tree learning algorithm C4.5 is an extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees, rule derivation, and so on. The tree comprises of nodes (features) that are selected by information entropy.

Instance-based learner IB1 is a single-nearest- neighbour algorithm, and it classifies entities taking the class of the closest associated vectors in the training set via distance metrics. It is the simplest among the algorithms used in our study.

Rough Set Model was introduced by Z. Pawlak in 80's to satisfy the need for a formal framework to manage imprecise knowledge expressed in terms of data acquired from experiments. Imprecise refers to the fact that the granularity of knowledge causes indiscernibility. These imprecise concepts can be defined approximately with available knowledge using two precise concepts called lower approximation ($\underline{RX}$) and upper approximation ($\overline{RX}$).

In Rough Set Theory (RST), an information table is defined as a tuple T = (U, A) where U and A are two finite, non-empty sets, U the universe of primitive objects and A the set of attributes. Each attribute or feature a ∈ A is associated with a set V_a of its value, called the domain of a. We may partition the attribute set A into two subsets C and D, called condition and decision attributes, respectively. Let P ⊂ A be a subset of attributes. The indiscernibility relation, denoted by IND(P), is an equivalence relation defined as:

IND(P) = {(x, y) ∈ U × U: ∀a ∈ P, a(x) = a(y)},

where a(x) denotes the value of a feature of object x. If (x, y) ∈ IND(P), x and y are said to be indiscernible with respect to P. The family of all equivalence classes of IND(P) (Partition of U determined by P) is denoted by U/IND(P). Each element in U/IND(P) is a set of indiscernible objects with respect to P. Equivalence classes U/IND(C) and U/IND(D) are called condition and decision classes.

For any concept X ⊆ U and attribute subset R ⊆ A, X could be approximated by the R-lower approximation and R-upper approximation using the knowledge of R.

The lower approximation of X is the set of objects of U that are surely in X, defined as:

$\underline{RX}$ = ∪ {E ∈ U/IND(R): E ⊆ X}.

The upper approximation of X is the set of objects of U that are possibly in X, defined as:

$\overline{RX}$ = ∪ {E ∈ U/IND(R): E ∩X ≠φ}.

The boundary region is defined as:

$BND_R(X) = \overline{RX} - \underline{RX}$

If the boundary region is empty, i.e., $\underline{RX} = \overline{RX}$, concept X is said to be R-definable. Otherwise X is a Rough Set with respect to R.

If the boundary region of X is the empty set, i.e., $BND_R(X)$ = Ø, then X is crisp (exact) with respect to B; in the opposite case, i.e., if $BND_R(X) \neq Ø$; X is referred to as rough (inexact) with respect to B. Thus, the set of elements is rough (inexact) if it cannot be defined in terms of the data, i.e. it has some elements that can be classified neither as member of the set nor its complement in view of the data.[2]

The positive region of decision classes U/IND(D) with respect to condition attributes C is denoted by $POS_c(D) = U\underline{RX}$. It is a set of objects of U that can be classified with certainty to classes U/IND(D) employing attributes of C. A subset R ⊆ C is said to be a D-reduct of C if $POS_R(D) = POS_C(D)$ and there is no R'⊂ R such that $POS'_R(D) = POS_C(D)$. In other words, a reduct is the minimal set of attributes preserving the positive region. There may exist many reducts in an information table.

4. Results and analysis

The proposed method was tested on four types of microarray datasets: Colon, Lung, Breast and Ovarian Cancer datasets, where F is features and I is instances in the dataset. The results are tabularized in Table1.

Table 1: Features in subsets

Datasets	F	I	F in subsets
Colon	2004	152	280
Lung	24482	97	227
Breast	12601	203	344
Ovarian	15155	253	196

To improve the performance of the feature subset selection four different classifiers are used and their accuracy is tabulated in Table 2.

Table 2: Accuracy by classifiers

Datasets	Naïve Bayes	IB1	C4.5	Rough Set
Colon	73.1	70	75.6	80.3
Lung	80	80.3	82	82
Breast	81	84	80	81.7
Ovarian	90	91.3	91	94

5. Conclusion and Future Work

The proposed approach removes irrelevant and redundant features using filter and clustering-based method. A cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. It uses Rough Set approach to obtain best subset features with good accuracy when compared to other classifiers. This method obtains best proportion of selected features, the best runtime, and best classification accuracy for Naive Bayes, IB1 and C4.5 also. The proposed Feature Subset Selection can be enhanced to handle text and image features. The refined algorithm could be used with RVM for generating good classification rules.

REFERENCES

[1] 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.
[2] Zdzisław Pawlak, " Rough set theory and its applications" Journal of Telecommunication and Information Technology, March 2002.
[3] Ning Zhong, Juzhen Dong," Using Rough Sets with Heuristics for Feature Selection", Journal of Intelligent Information Systems, 2001

[4] M.A. Hall and L.A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper," Proc. 12th Int'l Florida Artificial Intelligence Research Soc. Conf., pp. 235-239, 1999.

[5] Pengyi Yang, Wei Liu, Bing B. Zhou1," Ensemble-based wrapper methods for feature selection and class imbalance learning", 17th Pacific-Asia Conference, PAKDD 2013.

[6] Wang Kay Ngai, Ben Kao, Chun Kit Chui," Efficient Clustering of Uncertain Data" ICDM'06 IEEE Computer Society, 2006.

[7] M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato, 1999.

[8] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," J. Machine Learning Research, vol. 5, pp. 1531-1555, 2004.

[9] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.

[10] H. Liu and H. Motoda. Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic Publishers, 1998.

BIOGRAPHIES

R Indra Srinivas , Department of Information Science and Engineering, BMS College of Engineering, Bangalore- 560019.