International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 1, January-2014, pp: (319-324), Impact Factor: 1.252, Available online at: www.erpublications.com

Feature Selection for High Dimensional Data

Shailesh Singh Panwar
Department of Computer Science and Engineering, Graphic Era University, Dehradun, India

Abstract: Feature selection is an active research area in pattern recognition, statistics and data mining community. Idea behind feature selection is to choose a subset of input variables, eliminating features with little or no predictive information. Feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. This can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Rough set theory (RST) can be used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information. In this paper, feature selection technique has been used in high dimensional data for removing irrelevant features and producing high accuracy for post processing data.

Keywords: Feature Selection, Clustering, Rough set theory, Quick Redacts, Fast Correlation Based Filter.

I. INTRODUCTION

The main aim of feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features [1]. In real world problems FS is a must due to the abundance of noisy, irrelevant or misleading features. For instance, by removing these factors, learning from data techniques can benefit greatly. Given a feature set size n, the task of FS can be seen as a search for an optimal feature subset through the competing 2n candidate subsets. The definition of what an optimal subset is may vary depending on the problem to be solved. Although an exhaustive method may be used for this purpose, this is quite impractical for most datasets. Usually FS algorithms involve heuristic or random search strategies in an attempt to avoid this prohibitive complexity. However, the degree of optimality of the final feature subset is often reduced. From the perspective of selection strategy, feature selection algorithm broadly fall into three models: filter, wrapper or embedded.

The filter model evaluates features without involving any learning algorithm. The wrapper model requires a learning algorithm and uses its performances to evaluate the goodness of features. An important technique for dimensionality reduction to various areas, including computer vision, text mining and bioinformatics. If the evaluation procedure is tied to the task (e.g. clustering) of the learning algorithm, the FS algorithm employs the wrapper approach. This method searches through the feature subset space using the estimated accuracy from an induction algorithm as a measure of subset suitability. Although wrappers may produce better results, they are expensive to run and can break down with very large numbers of features.

This is due to the use of learning algorithms in the evaluation of subsets, some of which can encounter problems when dealing with large datasets. Feature selection algorithms may be classified into two categories based on their evaluation procedure. If an algorithm performs FS independently of any learning algorithm (i.e. it is a completely separate pre-processor), then it is a filter approach. In effect, irrelevant attributes are filtered out before induction. Filters tend to be applicable to most domains as they are not tied to any particular induction algorithm [12].

II. DATA REDUCTION

Data reduction is the transformation of numerical or alphabetical digital information derived empirical or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts. By data reduction reduce massive data-set to a manageable size without significant loss of information represented by the original data.

The advantages of data reduction are results are shown in a compact form and easy to understand. The graphical or pictorial representations can be used. Overall patterns can be seen. In this comparisons can be made between different sets of data. The quantitative measures can be used. The disadvantages are original data are lost and the process is irreversible.

Data Reduction Reduce the size of massive data-set to a manageable size without significant loss of information represented by the original data and also reduces the communications costs and decrease storage requirements. Data reduction also has some more scopes.

First is Primary Storage which reduces physical capacity for storage of active data. Second is Replication, reduce capacity for disaster recovery and business continuity. Third one is Data Production; reduce capacity for backup with longer retention periods. Fourth is Archive, reduce capacity for retention and preservation. Fifth is Movement/Migration of data, reduce bandwidth requirements for data-in-transit.

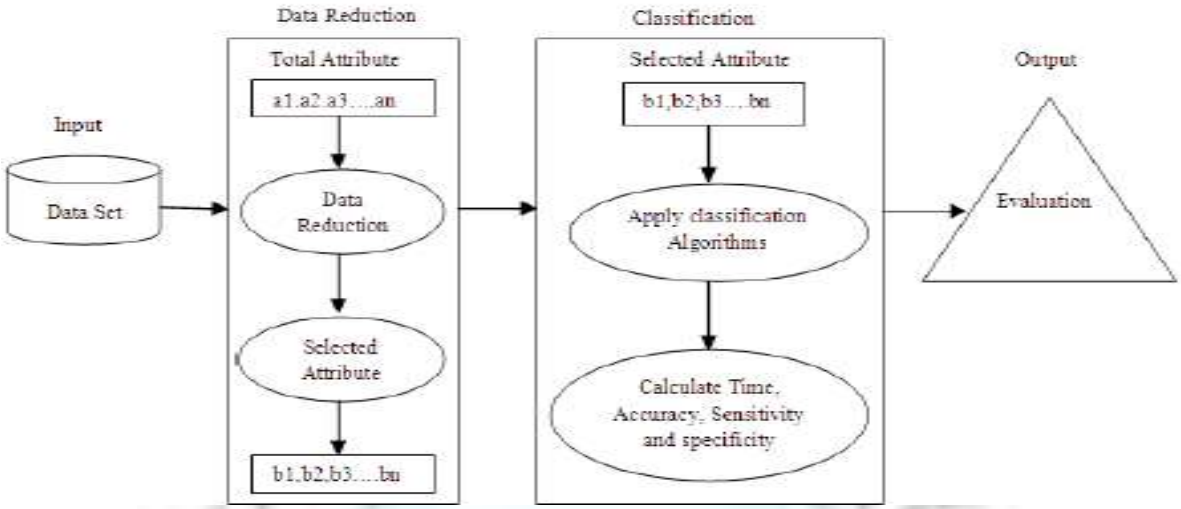

Figure 1: Data Reduction

The advantages of data reduction are results are shown in a compact form and easy to understand. The graphical or pictorial representations can be used. Overall patterns can be seen. In this comparisons can be made between different sets of data. The quantitative measures can be used. The disadvantages are original data are lost and the process is irreversible.
Dimensionality Reduction:-Feature selection (i.e., attribute subset selection) is selecting a minimum set of attributes (features) that is sufficient for the data mining task. Heuristic methods is step-wise forward selection and step-wise backward elimination. It is combining forward selection and backward elimination.

II. LITERATURE SURVEY

D.W. Aha, "Feature Weighting for Lazy Learning Algorithms" Learning algorithms differ in the degree to which they process their inputs prior to their use in performance tasks. Many algorithms eagerly compile input samples and use only the compilations to make decisions. Others are lazy: they perform less recompilation and use the input samples to guide decision making. The performance of many lazy learners significantly degrades when samples are defined by features containing little or misleading information [1]. A. Appice, M. Ceci, S. Rawles, and P. Flach, "Redundant Feature Elimination for Multi-Class Problems," We consider the problem of eliminating redundant Boolean features for a given data set, where a feature is redundant if it separates the classes less well than another feature or set of features [2]. A. Argyriou, T. Evgeniou, and M. Pontil, "Convex Multi-Task Feature Learning," We present a method for learning sparse representations shared across multiple tasks. This method is a generalization of the well-known single-task 1-norm regularization. It is based on a novel non-convex regularize which controls the number of learned features common across the tasks. The types of method shown in Fig 2 and Fig 3 for the feature selection process [3].

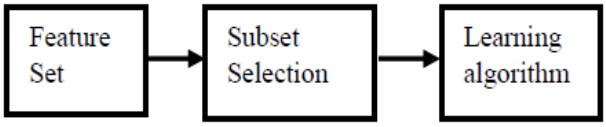

Figure 2 Filter

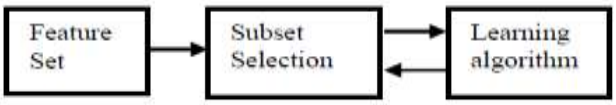

Figure 3 Wrapper

M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," One of the central problems in machine learning and pattern recognition is to develop appropriate representations for complex data. We consider the problem of constructing a representation for data lying on a low-dimensional manifold embedded in a high-dimensional space [4].

C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," How to selecting a small subset out of the thousands of genes in microarray data is important for accurate classification of phenotypes. Widely used methods typically rank genes according to their differential expressions among phenotypes and pick the top-ranked genes. We observe that feature sets so obtained have certain redundancy and study methods to minimize it. We propose a minimum redundancy - maximum relevance (MRMR) feature selection framework [5].

R. Duangsoithong, "Relevant and Redundant Feature Analysis with Ensemble Classification," Feature selection and ensemble classification increase system efficiency and accuracy in machine learning, data mining and biomedical informatics. This research presents an analysis of the effect of removing irrelevant and redundant features with ensemble classifiers using two datasets from UCI machine learning repository. Accuracy and computational time were evaluated by four base classifiers; NaiveBayes, multilayer preceptor, support vector machines and decision tree. Eliminating irrelevant features improves accuracy and reduces computational time while removing redundant features reduces computational time and reduces accuracy of the ensemble [6].

J.G. Dy et al., "Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images," This paper describes a new hierarchical approach to content-based image retrieval called the "customized-queries" approach (CQA). Contrary to the single feature vector approach which tries to classify the query and retrieve similar images in one step, CQA uses multiple feature sets and a two-step approach to retrieval. The first step classifies the query according to the class labels of the images using the features that best discriminate the classes. The second step then retrieves the most similar images within the predicted class using the features customized to distinguish "subclasses" within that class. Needing to find the customized feature subset for each class led us to investigate feature selection for unsupervised learning [7].

J.G. Dy and C.E. Brodley, "Feature Selection for Unsupervised Learning," In this paper, we identify two issues involved in developing an automated feature subset selection algorithm for unlabeled data: the need for finding the number of clusters in conjunction with feature selection, and the need for normalizing the bias of feature selection criteria with respect to dimension. We explore the feature selection problem and these issues through FSSEM (Feature Subset Selection using Expectation-Maximization (EM) clustering) and through two different performance criteria for evaluating candidate feature subsets: scatter separability and maximum likelihood. We present proofs on the dimensionality biases of these feature criteria, and present a cross-projection normalization scheme that can be applied to any criterion to ameliorate these biases. Our experiments show the need for feature selection, the need for addressing these two issues, and the effectiveness of our proposed solutions [8].

G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," Machine learning for text classification is the cornerstone of document categorization, news filtering, document routing, and personalization. In text domains, effective feature selection is essential to make the learning task efficient and more accurate. This paper presents an empirical comparison of twelve feature selection methods (e.g. Information Gain) evaluated on a benchmark of 229 text classification problem instances that were gathered from Reuters, TREC, OHSUMED, etc. The results are analyzed from multiple goal perspectives-accuracy, F-measure, precision, and recall since each is appropriate in different situations [9].

III. EXPERIMENTAL RESULTS

Rough set theory (RST) can be used as a tool to discover data dependencies and to reduce the number of attributes contained in a dataset using the data alone, requiring no additional information (Pawlak, 1991, Polkowski, 2002).

Over the past ten years, RST has become a topic of great interest to researchers and has been applied to many domains. Given a dataset with discredited attribute values, it is possible to find a subset (termed a reduct) of the original attributes using RST that are the most informative; all other attributes can be removed from the dataset with minimal information loss. From the dimensionality reduction perspective, informative features are those that are most predictive of the class attribute [10][11].

Let I= (U,A) be an information system, where U is a nonempty set of finite objects and A is a non-empty, finite set of attributes such that a: $U \rightarrow V_a$ for every a Є A. V_a is the set of values that attribute a may take. The information table assigns

a value a(x) from V_a to each attribute a and object x in the universe U. With any P A there is an associated equivalence relation IND (P):

$$\mathbf{IND\ (P) = \{(x,y)\ \epsilon\ U^2 \mid Va\ \epsilon\ P,\ a(x) = a(y)\}}$$

The relation IND (P) is called a P-indiscernibility relation.
The partition of U is a family of all equivalence classes of IND (P) and is denoted by U/IND (P). If (x,y) Є IND(P), then x and y are indiscernible by attributes from P. Let the information table be:

Table 1 Information System

	Age	LEMS
X1	16-30	50
X2	16-30	0
X3	31-45	1-25
X4	31-45	1-25
X5	46-60	26-49
X6	16-30	26-49
X7	46-60	26-49

After decision table is generated for the information table.

Let DS: I = (U, A Ụ {d}). D Є A is the decision attribute (instead of one we can consider more decision attributes). The element of A are called the condition attributes. Equivalence class is generated after the decision table. Let
IS= (U, A) be an information system, then with any B C A there an associated equivalence relation be

$$\mathbf{IND_{IS}\ (B) = \{(x, x') \in U^2 \mid Va \in B,\ a\ (x) = a\ (x')\}}$$

Let X ⊆U is a target set. X can be approximated using only the information contained within P by constructing the P-lower and P-upper approximations of X

$$\underline{P}X = \{x \mid [x]_p \subseteq X\}$$
$$\overline{P}X = \{x \mid [x]_p \cap X \neq \emptyset\}$$

Table 2: Decision Table

	Age	LEMS	Walk
X1	16-30	50	Yes
X2	16-30	0	No
X3	31-45	1-25	No
X4	31-45	1-25	Yes
X5	46-60	26-49	No
X6	16-30	26-49	Yes
X7	46-60	26-49	No

The reduction of attributes is achieved by comparing equivalence relations generated by sets of attributes. Attributes are removed so that the reduced set provides the same quality of classification as the original. A reduct is defined as a subset R of the conditional attribute set C such that $_R$(C) . A given dataset may have many attribute reduct sets, so the set R of all reducts is defined as:

$$R = \{X: X \subseteq C, \gamma_R(D) = \gamma_C(D)\}$$

The intersection of all the sets in R is called the core, the elements of which are those attributes that cannot be eliminated without introducing more contradictions to the dataset. In RSAR, a reduct with minimum cardinality is searched for; in other words an attempt is made to locate a single element of the minimal reduct set R min<= R

$$R_{min} = \{X \mid X \in R, \forall Y \in R, |X| \leq |Y|\}$$

The problem of finding a minimal reduct of an information system has been the subject of much research. The most basic solution to locating such a reduct is to simply generate all possible reducts and choose any with minimal cardinality. Obviously, this is an expensive solution to the problem and is only practical for very simple datasets. Most of the time only one minimal reduct is required, so all the calculations involved in discovering the rest are pointless [13]. To improve the performance of the above method, an element of pruning can be introduced. By noting the cardinality of any pre-discovered reducts, the current possible reduct can be ignored if it contains more elements. However, a better approach is needed - one that will avoid wasted computational effort. The Quick Reduct algorithm given in figure 2, attempts to calculate a minimal reduct without exhaustively generating all possible subsets. It starts of withan empty set and adds in turn, one at a time, those attributes that result in the greatest increase in dependency, until this produces its maximum possible value for the dataset. Algorithm steps involved are as

QUICKREDUCT (C, D)
C, the set of all conditional features;
D, the set of decision features.
R { }
Do
T← R
∀x ∈(C – R)
If γR {x} (D) >γT (D)
T← R U {x}
R ←T
Until γR(D) ==γ C(D)
return R

Rough set theory is a new mathematical approach to imprecision, vagueness and uncertainty. In an information system, every object of the universe is associated with some information. Objects characterized by the same information are indiscernible with respect to the available information about them. Any set of indiscernible objects is called an elementary set. Any union of elementary sets is referred to as a crisp set- otherwise a set is rough (imprecise, vague). Vague concepts cannot be characterized in terms of information about their elements.

A rough set is the approximation of a vague concept by a pair of precise concepts, called lower and upper approximations. The lower approximation is a description of the domain objects which are known with certainty to belong to the subset of interest, whereas the upper approximation is a description of the objects which possibly belong to the subset. Relative to a given set of attributes, a set is rough if its lower and upper approximations are not equal.The main advantage of rough set analysis is that it requires no additional knowledge except for the supplied data. Rough sets perform feature selection using only the granularity structure of the data Note that an intuitive understanding of Quick Reduct implies that, for a dimensionality of n, (n2+n)/2 evaluations of the dependency function may be performed for the worst case dataset. According to the Quick Reduct algorithm, the dependency of each attribute is calculated, and the best candidate chosen. The next best feature is added until the dependency of the reduct candidate equals the consistency of the dataset (1 if the dataset is consistent). This process, however, is not guaranteed to find a minimal reduct. Using the dependency function to discriminate between candidates may lead the search down a non-minimal path. It is impossible to predict which combinations of attributes will lead to an optimal reduct based on changes in dependency with the addition or deletion of single attributes. It does result in a close-to-minimal reduct, though, which is still useful in greatly reducing dataset dimensionality. Below performance results have been shown for the medical dataset as heart diseases using the quick reduct algorithm.

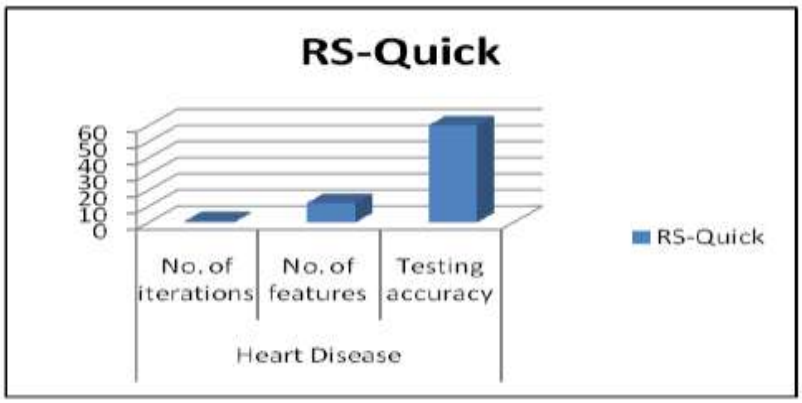

Figure 4: Performance chart

IV. CONCLUSIONS

Feature Selection is an important research direction of rough set application. However, this technique often fails to find better reduct. This project starts with the fundamental concepts of rough set theory and explains basic technique as Quick Reduct. Feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features. Idea behind feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection (FS) is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing theoriginal features. This can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points.

REFERENCES

[1]. D.W. Aha, "Feature Weighting for Lazy Learning Algorithms," Feature Extraction, Construction and Selection: a Data Mining Perspective, pp. 13-32, Springer, 1998.
[2]. A. Appice, M. Ceci, S. Rawles, and P. Flach, "Redundant Feature Elimination for Multi-Class Problems," Proc. 21st Int'l Conf. Machine Learning (ICML), 2004.
[3]. A. Argyriou, T. Evgeniou, and M. Pontil, "Convex Multi-Task Feature Learning," Machine Learning, vol. 73, no. 3, pp. 243-272, 2008.
[4]. M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Proc. Neural Information Processing Systems (NIPS), 2003.
[5]. C. Ding and H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," Proc. IEEE CS Conf. Bioinformatics (CSB), 2003.
[6]. R. Duangsoithong, "Relevant and Redundant Feature Analysis with Ensemble Classification," Proc. Seventh Int'l Conf. Advances in Pattern Recognition (ICAPR), 2009.
[7]. J.G. Dy et al., "Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 3, pp. 373-378, Mar.2003.
[8]. J.G. Dy and C.E. Brodley, "Feature Selection for Unsupervised Learning," J. Machine Learning Research, vol. 5, pp. 845-889, 2004.
[9]. G. Forman, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," J. Machine Learning Research, vol. 3, pp. 1289- 1305, 2003.
[10]. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol. 3, pp. 1157-1182, 2003.
[11]. M. Hall, "Correlation Based Feature Selection for Machine Learning," PhD thesis, Univ. of Waikato, Computer Science, 1999.
[12]. X. He, D. Cai, and P. Niyogi, "Laplacian Score for Feature Selection," Proc. Advances in Neural Information Processing Systems, vol. 18, 2005.
[13]. P. Nithya,and T.Menaka "A Study on performing clusters in transactional Data with different sizes and shapes", International Journal Of Advanced Research in computer Science and Software Engineering, Vol-3,issue 7,july 2013.