An Ameliorated Methodology for Feature Subset Selection on High Dimensional Data using Precise Relevance Measures

Kaveri B.V.
Department of Computer Science,
Bangalore Institute of Technology.
V V Puram, Bangalore, Karnataka,
India.

Asha T., PhD
Department of Computer Science,
Bangalore Institute of Technology.
V V Puram, Bangalore, Karnataka,
India.

ABSTRACT
Attribute subset selection refers to the method of choosing the set of attributes that best describes the dataset. The attributes obtained from the attribute subset selection method when applied to machine learning operations such as clustering, classification etc., should provide the same result as that of the original dataset. The method employed for attribute subset selection must be efficient in terms of selecting the relevant attributes and must also be accurate in terms of eliminating the redundant attributes.

With the aim of satisfying the above two goals we have designed a feature subset selection method using the precise relevance measures. We first efficiently select the relevant attributes using the relevance measure "symmetric uncertainty (SU)". The selected relevant attributes are, then divided into clusters based on "graph-theoretic" clustering method using the relevance measure "conditional mutual information (CMI)". Then the relevance measure "symmetric uncertainty" is used to select the attributes that are strongly related to the target class and also which best represents each cluster, thus giving us an accurate and independent subset of features. The above developed method not only produces smaller more accurate subset of features but also improves the performance of the machine learning operations such as naive base classifier

General Terms
Features subset selection, Machine learning techniques, clustering method, MST (minimum spanning tree), Naïve byes classifier

Keywords
Relevant feature, Redundant features, relevance measures, symmetric uncertainty *(SU),* conditional mutual information *(CMI).*

1. INTRODUCTION
Attribute subset selection refers to choosing the best set of attributes to describe the dataset without losing any information for clustering or classification. Attribute subset selection is an effectual way for dimensionality reduction, irrelevant attribute removal which increases the accuracy of the machine learning algorithm and for enhancing result clarity [30]. Many attribute subset selection approaches have been premeditated and estimated for machine learning techniques. They can be alienated into four general categories: The Filter, Wrapper, Embedded and Hybrid approaches.

The Filter methods are based on performance estimation metric which is calculated directly from the data to reduce the number of attributes, they more traditionally used as they are independent of the learning algorithm [4] [20] [14]. It has low computational intricacy, but there is no guarantee for the learning algorithm's correctness. In wrapper methods integrity of the selected subsets is influenced by the accuracy of the learning algorithm, the correctness of the learning algorithms is typically high but it has large computational intricacies and the attributes that are chosen have limited usage[4] [5]. The embedded methods integrate attribute selection as an element of the training process and are typically explicit to given learning algorithms, and thus might be more efficient than the other three approaches [1]."Decision-trees" or "Artificial Neural Networks" are examples of embedded methods [15]. The hybrid methods are a permutation of wrapper and filter methods [16] [5] [22] [20] [29]. The filter method reduces the search space that is measured by the following rapper method. The wrapper has good precision for a given learning algorithm whereas the filter method has good generality and minimum computational cost therefore the two approaches is united to get the best of both methods.

Pragmatic study shows that with respect to the filter methods, the utilization of cluster analysis is more effectual than other conservative analysis approaches [17] [6] [3]. In cluster analysis predominantly "graph-theoretic" method has been used widely in many applications which also show good outcome. Clustering based "graph-theoretic" method is as follows: First we use the instances to calculate a take up graph, then the edge that is greater or lesser than its neighbors (following a given norm) are eliminated from the graph resulting in a forest and each tree of the forest represents a cluster. In our algorithm we use clustering based "graph-theoretic" the clustering algorithm used is "minimum spanning tree (MST)" and the relevance measure "conditional mutual information *(CMI)*". We use MST because it does not presume that data points are grouped about the centers or divided by a standard geometric curve and have been broadly used in practice.

In this paper, we propose an attribute subset selection method using the precise relevance measures based on "graph-theoretic" method which uses "MST" and attempts at removing the irrelevant as well as the redundant features. The proposed methodology has a high possibility of producing a subset of useful and standalone attributes in an efficient and accurate manner. We even show that the subset of features obtained increases the accuracy of the machine learning algorithm –"Naïve Base Classifier" (NBC).

2. RELATED WORK

Generally, Feature subset selection methods focus on identifying the relevant features. A good example is Relief [9], it is based on distance-based criteria function. Relief is not effective at identifying redundant features [11]. Relief-F [12] extends Relief, which has the ability to deal with noisy, incomplete data sets and to deal with multi-class problems, but it still does not have the ability to identify redundant features. However, along with irrelevant features, redundant features also need to be identified and eliminated as it affects the efficiency and accuracy of machine learning algorithms, [11], [10]. CFS [7] is based on the hypothesis that a good feature subset is one that contains features highly correlated with the target class, yet uncorrelated with each other, thus identifying the redundant features as well. FCBF [23][25] is a fast filter method which can identify relevant and redundant features without pair wise correlation analysis. CMIM [2] iteratively picks features which maximize their mutual information with the target class, from the set of features that have been already picked. In our feature subset selection method we use graph theoretic clustering, although most feature subset selection methods are based on hierarchical clustering for word selection in context of text classification (e.g., [17], [6], and [3]). Hierarchical clustering also has been used to select features on spectral data. Van Dijk and Van Hullefor [21] proposed a hybrid filter/wrapper feature subset selection algorithm for regression. Krier et al. [13] presented a methodology combining hierarchical constrained clustering of spectral variables and selection of clusters by mutual information. Both methods employed agglomerative hierarchical clustering to remove redundant features. Dhillon et al. [3] proposed a new information-theoretic divisive algorithm for word clustering and applied it to text classification. Work on conditional mutual information [2] has been done for feature subset selection. We have derived our basic work from Qinbao et al. [26] and introduced the use of the relevance measure "symmetric uncertainty" [18] and conditional mutual information [2] in feature subset selection algorithm. Qinbao et al. [26] has done work on feature subset selection using graph theoretic clustering using the relevance measure symmetric uncertainty.

3. PROPOSED FEATURESUBSET SELCTION MATHOD

3.1 Basic Framework and Definitions

In our feature subset selection method the basic frame work can be divided into two parts. The first part is the irrelevant feature removal and the second part is the redundant features removal thus leaving us with the useful accurate subset of features. The first part of removing the irrelevant features is simple and is achieved using the relevance measure "symmetric uncertainty" and using a predefined relevance threshold[26] .Second part is redundant features removal which can be in turn be divided into three steeps. The first sub step is the construction of the minimum spanning tree using the subset of relevant features and the relevance measure "conditional mutual information". Second sub step is to divide the minimum spanning tree into clusters using the relevance measure "conditional mutual information". The third sub step is to select the feature that best represents each cluster using the relevance measure "symmetric uncertainty"[26]. Thus we get an accurate subset of features that are independent and closely related to the target class. Since our basic frame work revolves around removal of irrelevant and redundant features using the relevance measures "Symmetric uncertainty" and "Conditional mutual information" accordingly. We introduce the following definitions.

Definition 1:
(Relevant feature)[8] – $F_i$is relevant to the target concept C if and only if there exists some s', and C, such that, for probability $p(S'_i = s'_i, F_i = f_i) > 0, p(C = c|S'_i = s'_i, F_i = f_i) \neq p(C = c|S'_i = s'_i)$otherwise, it is an irrelevant feature. Here F is the full set of features, $f_i \in F$ be one of the features, $S_i = F - \{F_i\}$ and$S'_i \subseteq S_i, s'_i$ is the value assignment of all the featuresS'_i, f_ia value assignment of features F_i, and c value assignment of the target concept C.

Definition 2:
(Redundant feature) [24] - Let S be a set of features, a feature in $S$is redundant if and only if it has a Markov Blanket within S.

Definition 3:
(Markov blanket) [11] -Given a feature $F_i \in F$, let $M_i \subset F(F_i \notin M_i), M_i$ is said to be a Markov blanket for $F_i$if and only if $p(F - M_i - \{F_i\}, C|F_i, M_i) = p(F - M_i - \{F_i\}, C|M_i)$.

Definition 4:
(Symmetric uncertainty) (SU) [18]-

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

$Gain(X|Y)$ is the amount by which the entropy of Y decreases. It reflects the additional information about Y provided by $X$and is called the information gain [29] which is given by

$$Gain(X|Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

where $(X|Y)$ is the conditional entropy which quintiles the remaining entropy (i.e. uncertainty) of a random variable X given the value of another random variable Y. Suppose $p(x)$ is the prior probabilities for all values of X and (x/y) is the posterior probabilities of given the values of, $H(X|Y)$ is defined by

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x\ l\ y)\ log_2\ p(x\ l\ y)$$

Information gain is a symmetrical measure therefore the order of the two variables i.e. (X,Y) or (Y,X) will not affect the measure of the value.

D*efinition 5*:
(Conditional mutual information) (*CMI*) [2]-

Conditional Mutual Information provides an extension to Mutual Information. It measures correlation between two independent features, when value of a third feature is known. It is used to evaluate inter-feature correlation within a selected subset. This helps reduce redundancy among the selected features. Conditional Mutual Information between a target class X and independent variables Y and Z is given by

$$\begin{aligned} I(X;Y|Z) &= H(X|Y) - H(X|Y,Z) \\ &= H(y,z) - H(Z) - H(X,Y,Z) \\ &\quad + H(Y,Z) \end{aligned}$$

[31]For discrete random variables this can be simplified as

$$I(X;Y|Z)$$

$$= \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} pX,Y,Z(x,y,z)\ log \frac{pZ(z)pX,Y,Z(x,y,z)}{pX,Z(x,z)pY,Z(y,z)}$$

where the marginal, joint, and/or conditional probability mass functions are denoted by p with the appropriate subscript.

3.2 Sequence of Feature Subset Selection Method

The proposed feature subset selection method works as per the following steps:

(I) Irrelevant feature removal.
(II) Redundant feature removal.
  (II) a. Construct a MST from the subset relevant features.
  (II) b. Partition the MST.
  (II) c. Select representative features from partitioned MST.

Step (I) Irrelevant feature removal.
We first take the data set D with m features $F = \{F_1, F_2, \dots, F_m\}$, class C and a pre defined threshold θ as input. For the given input we compute $SU(F_i, C)$ value for each feature $F_i (1 \le i \le m)$. The features whose $SU(F_i, C)$values are greater than the predefined threshold θ comprise the target class-relevant feature subset $F' = \{F'_1, F'_2, \dots, F'_k\} (k \le m)$. Thus we have the subset of relevant features.

Step (II) Redundant feature removal.
This is further subdivided into (II) a. (II) b. (II) c. as follows.

Step (II) a. Construct a MST from the subset of relevant feature.
Here we first compute the weighted graph $G(V, E)$using the subset of relevant features, then from the weighted complete graph$G(V, E)$we construct the minimum spanning tree (MST). To do the above we first calculate $CMI(F'_i, F'_j)$value for each pair of features $F'_i$and$F'_j (F'_i, F'_j \in F' \wedge i \ne j)$. Then, setting features F'_i and F'_j as vertices and $CMI(F'_i, F'_j)(i \ne j)$as the weight of the edge between vertices $F'_i$and F'_j , a weighted complete graph $G(V, E)$is constructed where $V = \{F'_i | F'_i \in F' \wedge i \in [1, k]\}$and $E = \{(F'_i, F'_j) | (F'_i, F'_j \in F' \wedge i, j \in [1, k] \wedge i \ne j\}$. Here G is an undirected graph. The complete graph G reflects the correlations among all the target class-relevant features. Here, graph G has k vertices and $k(k-1)/2$ edges. For high dimensional data, the graph G becomes dense and the weights of the different edges are highly interconnected, this makes the decomposition of complete graph NP-hard [28]. Therefore from graph G, we build a MST, such that the sum of the weights of the edges is the minimum connecting all vertices, using the classic Prim algorithm [19].

Step (II) b. Partitioning of the MST into individual clusters
After building the MST, we first remove the edges$E = \{(F'_i, F'_j) | (F'_i, F'_j \in F' \wedge i, j \in [1, k] \wedge i \ne j\}$, whose weights are smaller than both of the $CMI(F'_i, C)$ and$CMI(F'_j, C)$, T_1 and T_2.Assuming the set of vertices in any one of the final clusters to be $V(T)$, we have the property that for each pair of vertices $(F'_i, F'_j \in V(T)), CMI(F'_i, F'_j) \ge CMI(F'_i, C) \vee CMI(F'_i, F'_j) \ge CMI(F'_j, C)$ always holds. We also have property where $S = \{F_1, F_2, \dots, F_i \dots, F_{k<|F|}\}$is a cluster of features. If $\exists F_j \in S, CMI(F_j, C) \ge CMI(F_i, C) \wedge CMI(F_i, F_j) > CMI(F_i, C)$ isalways corrected for each$F_i \in S (i \ne j)$, then F_i are redundant features with respect to the given F_j, this property guarantees the features in $V(T)$ are redundant.

Step (II) c. Selecting representative features from individual clusters

Here the feature $F'_i$from the partitioned tree (cluster) is a representative feature of the cluster if and only if $F'_i = argmax_{F_j \in S}(F'_j, C)$ condition is satisfied.

3.3 Algorithm: Feature Subset Selection using the Relevance Measure SI and CMI

Input: $D(F_1, F_2, \dots, F_m, C)$- *the given data set,* θ- *the T-Relevance threshold.*
Output: S - *Selected feature subset.*

//==== part1: Irrelevant Feature Removal ====
For $i = 1$ to m **do**
$SU(F_i, C)$
If $SU(F_i, C) > \theta$**then**
$S = S \cup \{F_i\}$;

//==== Part 2: Minimum Spanning Tree Construction ====
G = NULL; *//*G *is a complete graph*
For each pair of features $\{F'_i, F'_j\} \subset S$**do**
$CMI(F'_i, F'_j)$
AddF'_i and/or $F'_j to\ G$ with $CMI(F'_i, F'_j)$as the weight of the corresponding edge;
$MinSpanTree = prim(G)$; *//Using Prim Algorithm to generate the minimum spanning tree*

//==== Part 3: Tree Partition into clusters ====
Forest= $MinSpanTree$
For each edge $E_{ij} \in$ Forest **do**
If$CMI(F'_i, F'_j) < CMI(F'_i, C) \wedge CMI(F'_i, F'_j) < CMI(F'_j, C)$ **then**
Forest = Forest$-F_{ij}$

//==== Part4: Representative feature selection from each cluster (partitioned tree) ====
$S = \phi$
For each tree $T_i \in$ Forest **do**
$F_R^j = argmax_{F' k \in T_i} SU(F'_k, C)$
$S = S \cup \{F_R^j\}$;
return S;

4. EMPIRICAL STUDY & RESULTS

Our observations focus on enhancing the importance of using the precise relevance measure in the "feature subset selection method". We use 10 textual data sets that are publically available from the UCI- KDD's "Machine learning Repository". Here we feed the subset of features obtained from Qinbao et al.[26] "feature subset selection method" which uses only "symmetric uncertainty" as relevance measure and those obtained from our proposed methodology that uses "symmetric uncertainty" as well as "conditional mutual information" as the relevance measure to the well known " naïve byes" classifier. This is to check which subset of features gives better classification accuracy. Table 1. summarizes the results.

In Table 1. *SU* indicates the subset of features obtained by using *SU* in the "feature subset selection method" [26]. *SU& CMI* indicates the subset of features obtained by using both *SU & CMI* in the "feature subset selection method".

The Table 1., clearly shows that the accuracy of the machine learning algorithm "naïve byes" classifier has increased when we use the subset of features obtained from the "feature subset selection method" that uses both "Symmetric uncertainty" and "Conditional mutual information" as the relevance measure when compared to the previously proposed "feature subset selection method" that uses only *SU*[26].

Table 1. Accuracy of naïve byes classifier

Data ID	Data Base Name	Accuracy of Naïve Byes Classifier	
		SU	SU & CMI
01	Glass	76%	88%
02	Diabetes	70%	81%
03	Wine	74%	86%
04	Wine quality	72%	80%
05	Heart disease	68%	80%
06	Firm teacher clave direction classification	74%	86%
07	Sensor less drive diagnosis	75%	84%
08	Banknote authentication	73%	86%
09	Blood transfusion detection center	70%	82%
10	Parkinsons	74%	85%

5. CONCLUSION

We have developed a novel features subset algorithm with two types of relevance measure one of them is symmetric uncertainty and the other is conditional mutual information. We assessed that conditional mutual information is a better relevance measure to remove the redundant attributes as it takes into consideration the entropy of a combination of 3 variables and this also supported by the results which was run on data sets. Thus the subset of features obtained by using the proposed technique provides a better accuracy rate for the machine learning technique such as "naïve byes" classifier. For feature work we would like to explore more relevance measures and see how different relevance measures used in features subset selection methods effect the performance of different machine learning algorithm.

6. REFERENCES

[1] Conference on Machine Learning, pp 74-81 (2001). Guyon I. and Elisseeff A., An introduction to variable and feature selection, Journal of Machine Learning Research, 3, pp 1157-1182 (2003).

[2] Fleuret F., Fast binary feature selection with conditional mutual Information, Journal of Machine Learning Research, 5, pp 1531-1555 (2004).

[3] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res.,3,pp 1265-1287 (2003).

[4] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp 131-156 (1997).

[5] Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.

[6] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval, pp 96103 (1998).

[7] Hall M.A., Correlation-Based Feature Subset Selection for Machine Learning, Ph.D. dissertation Waikato, New Zealand: Univ. Waikato (1999).

[8] John G.H., Kohavi R. and Pfleger K., Irrelevant Features and the Subset Selection Problem, In the Proceedings of the Eleventh International Conference on Machine Learning, pp 121-129 (1994).

[9] Kira K. and Rendell L.A., The feature selection problem: Traditional methods and a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134 (1992).

[10] Kohavi R. and John G.H., Wrappers for feature subset selection, Artificial Intelligence, Intell., 97(1-2), pp 273-324 (1997).

[11] Koller D. and Sahami M., Toward optimal feature selection, In Proceedings of International Conference on Machine Learning, pp 284-292 (1996).

[12] Kononenko I., Estimating Attributes: Analysis and Extensions of RELIEF, In Proceedings of the 1994 European Conference on Machine Learning, pp 171-182 (1994).

[13] Krier C., Francois D., Rossi F. and Verleysen M., Feature clustering and mutual information for the selection of variables in spectral data, In Proc European Symposium on Artificial Neural Networks Advances in Computational Intelligence and Learning, pp 157-162 (2007).

[14] Langley P., Selection of relevant features in machine learning, In Proceedings of the AAAI Fall Symposium on Relevance, pp 1-5 (1994).

[15] Mitchell T.M., Generalization as Search, Artificial Intelligence, 18(2), pp 203-226 (1982).

[16] Ng A.Y., On feature selection: learning with exponentially many irrelevant features as training examples, In Proceedings of the Fifteenth International Conference on Machine Learning, pp 404-412 (1998).

[17] Pereira F., Tishby N. and Lee L., Distributional clustering of English words, In Proceedings of the 31st Annual Meeting on Association For Computational Linguistics, pp 183-190 (1993).

[18] Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T., Numerical recipes in C. Cambridge University Press, Cambridge (1988).

[19] Prim R.C., Shortest connection networks and some generalizations, Bell System Technical Journal, 36, pp 1389-1401 (1957).

[20] Souza J., Feature selection with a general hybrid algorithm, PhD, University of Ottawa, Ottawa, Ontario, Canada (2004).

[21] Van Dijk G. and Van Hullefor M.M., Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy, International Conference on Artificial Neural Networks (2006).

[22] Xing E., Jordan M. and Karp R., Feature selection for high-dimensional genomic microarray data, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 601-608 (2001).

[23] Yu L. and Liu H., Feature selection for high-dimensional data: a fast correlation-based filter solution, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863 (2003).

[24] Yu L. and Liu H., Redundancy based feature selection for microarray data, In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 737-742 (2004).

[25] Yu L. and Liu H., Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research, 10(5), pp 1205-1224 (2004).

[26] Qinbao Song, Jingjie Ni and Guangtao Wang, IEEE Transactions on knowledge and data engineering Vol: 25 No.:1 (2013).

[27] Arey M.R. and Johnson D.S., Computers and Intractability: a Guide to the Theory of Np-Completeness. W. H. Freeman & Co, (1979).

[28] Quinlan J.R., C4.5: Programs for Machine Learning. San Mateo, California: Morgan Kaufman (1993).

[29] Yu J., Abidi S.S.R. and Artes P.H., A hybrid feature selection strategy for image defining features: towards interpretation of optic nerve images, In Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 8, pp 5127-5132, (2005).

[30] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artificial Intelligence., 159(1-2), pp 49-74 (2004)

[31] Wyner., A. D.:A definition of conditional mutual information for arbitrary ensembles. Information and Control **38** (1):51–59. doi: 10.1016/s0019-9958 (78)90026-8. http://en.wikipedia.org/wiki/Conditional_ mutual_information (1978)