ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 2 , February 2015

Feature Subset Selection Method For High Dimensional Data Using Kruskal's Algorithm

G.Geethanjali[1], V.Sharmila[2], P.Balamurugan[3]

M.E. Final Year Student, Department of CSE, K.S.R. College of Engineering, Namakkal, Tamil Nadu, India[1].

Associate Professor, Department of CSE, K.S.R. College of Engineering, Namakkal, Tamil Nadu, India[2,3].

ABSTRACT: Feature Selection is to selecting the useful features from the original dataset for improve the more accurate results. This method removes irrelevant and redundant features. It finds a similarity computation based on the entropy and conditional entropy values. After computing similarity computation to applied Approximate Relevancy(AR) algorithm which will find the relevance between the attribute and class labels from that computation most relevant attributes will be selected and create graph according to that relevant features. After calculating relevant features to form the spanning tree using kruskal's algorithm, removing all redundant features for which it has an edge in tree. Finally, to select best subset of the features from the original dataset.

KEYWORDS: Feature Selection, AR Relevancy, Redundancy, Entropy, Conditional Entropy.

I. INTRODUCTION

Data mining is the process of discovering interesting patterns or knowledge from the large amount of data. Feature selection is frequently used as a preprocessing step to machine learning. It is a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion. In recent years, data has become increasingly larger in both number of instances and number of features in many applications such as genome projects, text categorization, image retrieval, and customer relationship management. This enormity may cause serious problems to many machine learning algorithms with respect to scalability and learning performance. High dimensional data can contain high degree of irrelevant and redundant information which may greatly degrade the Performance of learning algorithms. Therefore, feature selection becomes very necessary for machine learning tasks when facing high dimensional data nowadays. Feature selection has some of the characteristics

- To discover quality patterns
- Provides lower computational cost
- Save time and space
- Leads to better learning performance

II. RELATED WORKS

Khali Benabdeslem et al.,[1] proposed a Constrained Semi-supervised Feature Selection with Redundancy elimination(CSFSR).Constraint Laplacian scoring function is used based on the pair wise Constraints. The drawback of the paper is some of the constraint sets are produced less accuracy.

Daoqiang Zhang et al.,[2] proposed to use another form of supervision information for feature selection using constraint score, pair wise constraints, which specifies whether a pair of data samples belong to the same class (must-link constraints) or different classes (cannot-link constraints).Based on that the scoring function has calculated. It does not deal with large datasets.

Mohammed Hindawi et al.,[3] proposed Constrained Selection for Feature Selection(CSFS). This aims to grasp the most coherent constraints extracted from labelled part of data. which specifies whether a pair of data samples belong to the same class (must-link constraints) or different classes (cannot-link constraints).It has some drawbacks some constraint sets are provide less accuracy and Does not deal with large datasets.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 2 , February 2015

Daoqiang Zhang et al.,[4] proposed a simple algorithm called SSDR (Semi-Supervised Dimensionality Reduction). In that constraint score and laplacian score are calculated. Constraint score is used for labelled data and laplacian score is used for unlabeled data. The drawbacks of the paper is problematic for high dimensional data and decrease the algorithm performance.

III. PROPOSED METHOD

Proposed Feature Selection algorithm is used to select subset of the features from the original dataset. This algorithm effectively removed irrelevant features and redundant features. In our proposed algorithm involves, (i) To find a Symmetrical Uncertainty between features (ii) Construct the graph using correlation coefficient (iii) Eliminate the unwanted edges using Kruskal's Algorithm (iv) Select the best set of features using AR Relevance

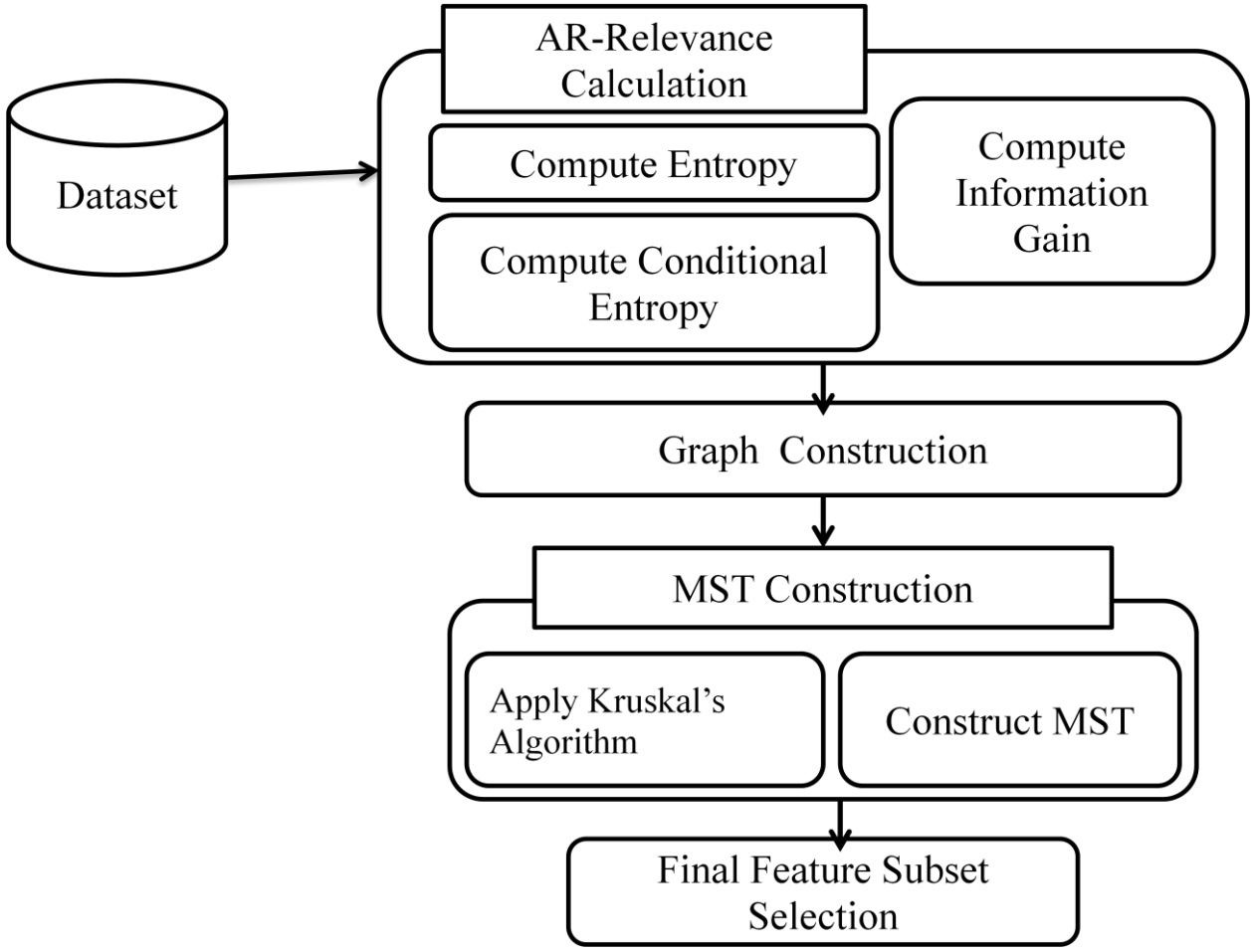

Fig. 1. Feature Subset Selection

A. Entropy and Conditional Entropy Calculation

In Entropy and Conditional Entropy Calculation, 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. Entropy value is calculated using the class labels and Conditional Entropy is calculated based on the features and class labels.

$$H(x) = -\sum p(x) \log_2 p(x) \quad (1)$$

$$H(x|y) = -\sum p(y) \sum p(x|y) \log_2 p(x|y) \quad (2)$$

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 2 , February 2015

From (1) and (2) p(x) is the probability density function and p (x|y) is the conditional probability function.

B. Similarity Calculation and MST Construction

In information theory, entropy is the average amount of information contained in each message received. Here, message stands for an event. In Entropy characterizes the uncertainty about source of the information. The source is also characterized by the probability distribution of the samples drawn from it. The idea is that the less likely an event is, the more information it provides when it occurs. For some other reasons it makes sense to define information as the negative of the logarithm of the probability distribution. The probability distribution of the events, coupled with the information amount of every event, forms a random variable whose average value is the average amount of information, entropy generated by this distribution.

$$SU(x,y) = \frac{2 \times Gain(x|y)}{H(x)+H(y)} \quad (3)$$

From (3) Gain(x|y) is calculated as,

$$Gain(x|y) = H(x) - H(y) \quad (4)$$

From (4) H(x) is the Entropy and H(x|y) is the Conditional Entropy. Here (3) is calculated using Entropy and Gain values.

C. AR Relevancy Calculation

The relevance between the feature F_i and the target concept C is referred to as the AR-Relevance of F_i and C, and denoted by SU (F_i, C). If SU (F_i, C) is greater than a predefined threshold, then say that F_i is a strong AR Relevance feature.

D. MST Construction

A Minimum Spanning Tree for a weighted graph is a spanning tree with minimum weight. Kruskal's algorithm is the greedy algorithm in graph theory that finds the Minimum Spanning Tree(MST) for a connected weighted graph. A MST has (v-1) edges where v is the number of edges in the graph. Here Minimum Spanning Tree is used reduced the edges from the original graph.

E. Kruskal's Algorithm

Create a forest F(a set of trees),where each vertex in the graph is a separate tree. This Algorithm follows as, Sort all the edges in decreasing order of their weight, Pick the smallest edge and Check if it forms a cycle with the spanning tree. If cycle is not formed includes the edge else, discard it.

F. Relevant Subset Feature Selection

After building the Tree, the next step is to remove the edges whose weight is smaller than the Approximate Relevance. It checks the condition and eliminates the edges according to that, for find the relevant subsets.

G. Redundant Analysis and Feature Subset Selection

After removing all the unnecessary edges, the forest is obtained. Each sub tree represents a cluster. Features in each cluster are redundant, so representatives are chosen from the each cluster which one has the greatest relevance with that class. After calculating the relevant features to form the spanning tree to from the graph using kruskal's algorithm. Then Removing all redundant features for which it has an edge in spanning tree. Finally, to select best subset of the features from the original dataset.

H. Correlation Measures

The most known measure that can be used to calculating the relationship between two features F_r and F_c is the linear correlation coefficient. It is defined as follows:

$$\rho(F_r,F_c) = \frac{\sum_i (\bar{f}_{ri} - f_r)(\bar{f}_{ci} - f_c)}{\bar{\quad} \quad \bar{\quad}}$$

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 2 , February 2015

$$\sqrt{\sum_i (f_{ri} - f_r)^2 \sum_i (f_{ci} - f_c)^2} \qquad (5)$$

From (5) where f_r and f_c are the means of the feature vectors f_r and f_c respectively. We choose to use the mutual information (MI) between two features F_r and F_c. MI quantifies the dependence between the joint distribution of both features. Under the hypothesis that the joint distribution of F_r and F_c is multi-variate normal, the mutual information can be directly related to the correlation coefficient ρ

$$I(F_r, F_c) = -1/2 \log (1 - \rho^2(F_r, F_c)) \qquad (6)$$

From (6) the weight is calculated and form the graph using kruskal's algorithm Minimum Spanning Tree is formed and then eliminate redundant features.

TABLE 1. Sample Dataset

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

IV. SAMPLE CALCULATION

Similarity Uncertainty (SU) calculation is calculated based on the Entropy and Conditional Entropy. Entropy is calculated based on the class labels (Play Tennis). Conditional Entropy is calculated based on the attribute values (outlook, temperature, humidity, wind).

Entropy Calculation (Play Tennis)

$H(x)+H(y) = -\sum p(x) \log_2 p(x)$

$+ -\sum p(y) \log_2 p(y)$

$5/14*\log_2(5/14) - 9/14*\log_2(9/14)$

$=$ Entropy(5/14,9/14) = 0.9403

Conditional Entropy Calculation(Outlook)

$H(x|y) = -\sum p(y) \sum p(x|y) \log_2 p(x|y)$

$5/14*$Entropy(3/5,2/5)+4/14*Entropy(1,0)+

5/14 * Entropy(3/5,2/5) = 0.6935

Information Gain(Outlook)

Gain(x|y) = H(x)-H(x|y)

IG(Outlook) = 0.9403 – 0.6935 = 0.2468

Similarity Calculation(Outlook)

SU(x,y) = IG / H(x)+H(y)

SU(Outlook) = 0.2468 / 0.9403

SU(Outlook) = 0.26247

This calculation is calculated based on the sample dataset. Similarly other attribute values are calculated which has the highest value which will be taken as the best relevant feature. After find the relevant feature the graph is formed using the relevant feature, then using kruskal's algorithm the redundant features are eliminated by forming Minimum Spanning Tree. Finally, best subsets of the features are selected.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 2 , February 2015

V.PERFORMANCE ANALYSIS

TABLE 2. Evaluation Results

(Heart Disease Dataset)

Algorithm	Original Features	Selected Features	Accuracy
Constrained Semi Supervised Feature Selection With Redundancy Elimination	44	25	80.5
Feature Subset Method For High Dimensional Data Using Kruskal's Algorithm	44	18	86

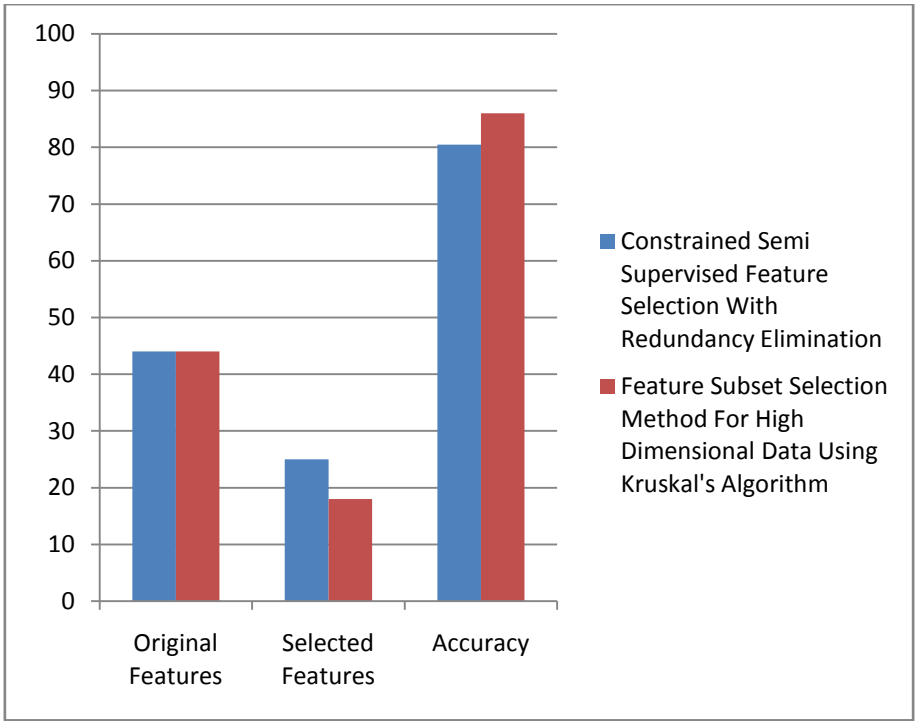

Fig 2. Results on "Heart Disease Dataset"

VI. CONCLUSION

This paper is intended to improve the speed and accuracy of the learning algorithms. This similarity computation method provides best features from the original dataset. This Algorithm effectively removed redundant and irrelevant features. In this method Similarity Calculation is used to find relevant features and redundant features are effectively removed from the dataset using kruskal's algorithm by forming Minimum Spanning Tree(MST). Finally, This proposed algorithm is used to select best subset of the features from the original dataset. The future work is concentrated on to deal with plenty of datasets and applying some optimization techniques.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 2 , February 2015

ACKNOWLEDGMENT

I wish to thank my institution, 'K.S.R. College of Engineering' for giving me the opportunity to write a research paper. A special thanks to my Head of the Department, Dr.A.Rajiv Kannan for encouraging me and to Mrs.V.Sharmila for her support and guidance throughout and without whom, this work is not possible.
Last but not least, I would like to the authors of the various research papers that I have referred to, for the completion of this work.

REFERENCES

[1] Khali Benabdeslem and Mohammed Hindawi, " Efficient Semi-Supervised Feature Selection: Constraint, Relevance, and Redundancy," IEEE Trans. Knowledge and Data mining, vol. 26, no. 5,pp.1131-1143 may 2014.
[2] Daoqiang Zhang, S.Chen, and Z.Zhou, "Constraint score: A new filter method for feature selection with pairwise constraints,"Pattern Recognition., vol. 41, no. 5, pp. 1440 –1451, 2008.
[3] M. Hindawi, K. Allab, and K. Benabdeslem, "Constraint selection based semi-supervised feature selection," in Proc. IEEE ICDM, Vancouver, BC, Canada,2011, pp. 1080–1085.
[4] Daoqiang Zhang, Zhi-Hua Zhou and Songcan Chen," Semi-Supervised Dimensionality Reduction," in Proc. SIAM Int. Conf. Data Mining, Pittsburgh, PA, USA, 2007,pp. 629-634.
[5] Z. Zhao, L. Wang, and H. Liu,"Efficient spectral feature selection with minimum redundancy," in Proc. AAAI, 2010. pp. 673-678.
[6] H.Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria max- dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[7] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection,"IEEE Trans. Knowledge. Data Eng., vol. 25, no. 3, pp. 619–632, Mar. 2013.
[8] L.Yu and H. Liu, "Efficient feature selection via analysis of relevance and the redundancy," J. Mach. Learn. Res., vol. 5, pp. 1205–1224, Oct. 2004.