

Constrained Based Feature Subset Selection Algorithm for High Dimensional Data

G.Geethanjali¹ Mr.P.Prakash²

¹Student of M.E Final Year ²Assistant Professor

^{1,2}Department of Computer Science & Engineering

^{1,2}K.S.R. College of Engineering, Namakkal, Tamilnadu-637215, India

Abstract— Feature Selection is to selecting the useful features from the original dataset for improve the more accurate results. Constrained Based Feature Subset Selection(CFSS) Algorithm Removes irrelevant and redundant features. This method is to find 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, then using Adaptive k++ neighborhood algorithm group those relevant features 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.

Key words: Feature Selection, AR Relevancy, Redundancy, Entropy, Conditional Entropy

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 labeled 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.

Daoqiang Zhang et al.,[1] proposed a simple algorithm called SDDR (Semi-Supervised Dimensionality Reduction). In that constraint score and laplacian score are calculated.Constraint score is used for labeled 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.

I. INTRODUCTION

Feature selection has been an active research area in pattern reorganization, statistics and data mining communication. Nowadays a rapid growth of high dimensional data such as digital images, gene expression microarrays, dimensionality reduction has been a fundamental tool for many data mining tasks.

Dimensionality reduction can be performed by two categories of techniques: Feature extraction or Feature selection. Feature extraction reduces dimensionality by generating a small set of new feature via combining the original features. Feature selection achieves dimensionality reduction by selecting a small set of original features. For feature selection Wrapper model, Filter model and Embedded model used. Wrapper model is based on learning algorithm. The objective function is a pattern classifier, which evaluates feature subsets by their predictive accuracy (recognition rate on test data) by cross-validation. Filter method is independent of used classifier. Feature subsets are evaluated by their information content and distance. It is for high dimensional dataset. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms. In this paper used filter approach for improve accuracy.

II. RELATED WORKS

Jennifer G.Dy et al.,[8] proposed feature selection problem using FSSEM (Feature Subset Selection using Expectation-Maximization (EM) clustering).It only applicable for unsupervised learning used by wrapper approach. It has some drawbacks class labels does not guide to feature selection and it will affect the performance.

Daoqiang Zhang et al.,[2] proposed to use another form of supervision information for feature selection using

III. PROPOSED METHOD

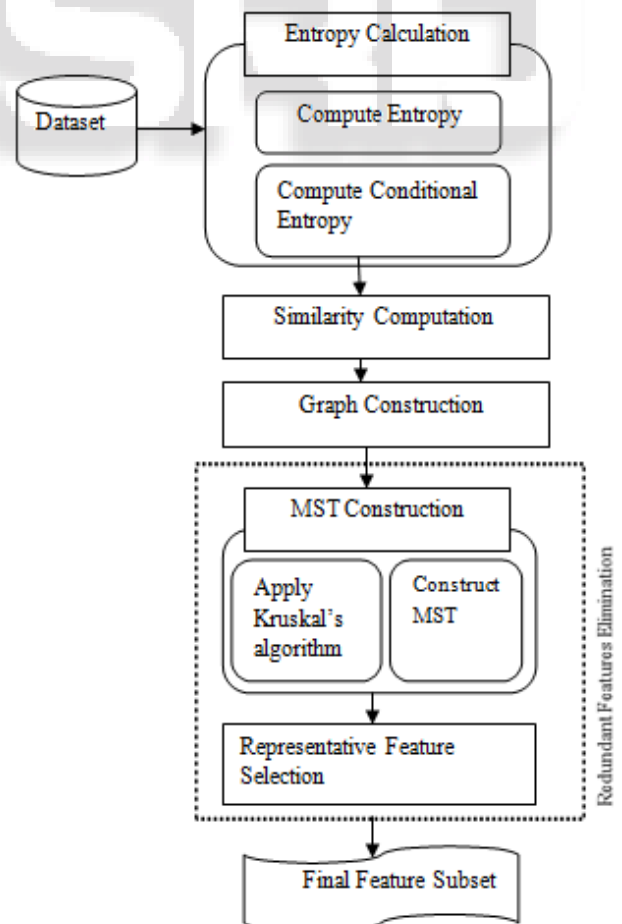


Fig. 1: Constrained Based Feature Subset Selection

Constrained Based Feature Subset Selection (CFSS) 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 CFSS algorithm involves, 1) To calculate Symmetrical Uncertainty between features. 2) To applied Approximate Relevancy (AR) algorithm. 3) Eliminate Redundant Features using Kruskal's Algorithm.

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.

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

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

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.

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

From (3) $\text{Gain}(x|y)$ is calculated as,

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

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.

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. Adaptive K++ Neighborhood Algorithm:

In the proposed system implement the algorithm called Adaptive k++ neighborhood algorithm. In that the algorithm would be related to data structure and could be defined as follows: Two instances are neighbors if they belong to the same cluster. Consequently, each cluster has its own k which is the number of its elements.

H. 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.

I. 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:

$$\rho(F_r, F_c) = \frac{\sum_i (f_{ri} - \bar{f}_r) (f_{ci} - \bar{f}_c)}{\sqrt{\sum_i (f_{ri} - \bar{f}_r)^2 \sum_i (f_{ci} - \bar{f}_c)^2}} \quad (5)$$

From (5) where \bar{f}_r and \bar{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

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

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

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

Table 1: Sample Dataset

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).

A. Entropy calculation (play tennis):

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

$$= \text{Entropy}(5/14, 9/14) = 0.9403$$

B. Conditional Entropy Calculation(Outlook):

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

$$= 5/14 * \text{Entropy}(3/5, 2/5) + 9/14 * \text{Entropy}(1, 0) + 5/14 * \text{Entropy}(3/5, 2/5) = 0.6935$$

C. Information Gain(Outlook):

$$\text{Gain}(x/y) = H(x) - H(x|y)$$

$$\text{IG}(\text{Outlook}) = 0.9403 - 0.6935 = 0.2468$$

D. Similarity Calculation(Outlook):

$$\text{SU}(x,y) = \text{IG} / H(x)+H(y)$$

$$\text{SU}(\text{Outlook}) = 0.2468 / 0.9403$$

$$\text{SU}(\text{Outlook}) = 0.26247$$

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.

V. CONCLUSION

This paper is intended to improve the speed and accuracy of the learning algorithms. Constrained based Feature Subset Selection(CFSS) 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.

VI. ACKNOWLEDGMENT

I extend my sincere thanks to 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 us and to Mr.P.Prakash for his support and valuable guidance throughout this project work and makes this project as a successful one.

Finally, I would like to thank authors of the various research papers that I have referred to, for the completion of this work.

REFERENCES

- [1] 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.
- [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] 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.
- [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] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," J. Mach. Learn. Res., vol. 5, pp. 845-889, Aug.2004.