Efficient Feature Subset Selection Using Kruskal's Process

Mr. R. Gnana Prakash , Mr. S. Kannudurai

Department of Computer Science and Engineering,
Kalasalingam Institute of Technology (Affiliated to Anna University),
Krishnankovil, Virudhunagar District, Tamil Nadu.

***Abstract*— In this paper, we proposed a new way of efficient feature subset selection. Feature subset selection is a process of finding the subset of most useful features for a dataset which produces the results similar to that of the entire dataset. A feature selection algorithm can be evaluated from its efficiency and effectiveness. The required by an algorithm to obtain a subset of useful features from a dataset is defined as the efficiency of that algorithm. Similarly, the effectiveness of the algorithm is defined as the quality of the subset of features obtained from the original dataset. According to these factors, an efficient feature subset selection algorithm (FAST) using kruskal's process is proposed and experimentally evaluated in this paper.**

I. INTRODUCTION

Feature selection not only aims on selecting a subset of relevant features with respect to the target classes, it also concentrates on reducing dimensionality, removing redundant features, increasing learning clarity, and enhancing result quality.

II. MACHINE LEARNING

So many feature subset selection methods are proposed and studied for machine learning applications. They can be classified into four basic types: The Embedded, Wrapper, Filter, and Hybrid methods. The embedded method states the feature selection as a part of the training process and it is specific to the learning algorithms, and therefore it may be more efficient than the other machine learning methods. Machine learning algorithms such as decision trees or artificial neural networks are the best examples of embedded approaches.

The wrapper methods focus on the predictive accuracy of a predetermined learning algorithm to find the perfection of the selected subsets, the accuracy of the learning algorithm is usually high. However, the selected features generality is limited and this method is highly complex for computation. The filter methods are the independent ones among the learning algorithms, with better generality for the selected features. They are very much low in complexity for computation, but the accuracy of the features that are selected using the learning algorithms is not guaranteed. The hybrid methods are the results of a combination of filter and wrapper methods.

Filter method is used to reduce the search space. The wrapper method considers it and proceeds the further processing. The computations using the wrapper methods are highly expensive and therefore they are not suitable for finding subsets from smaller datasets. The filter methods are not only good in their observation. They are perfect for usage when the numbers of features are very high in their dimension.

With respect to the filter machine learning approach, the usage of cluster verification has been demonstrated to be more effective than traditional feature selection algorithms. Hybrid approaches are formed by combining filter and wrapper methods together, because they can achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. These are the various approaches for machine learning in finding a relevant feature subset for a given data set.

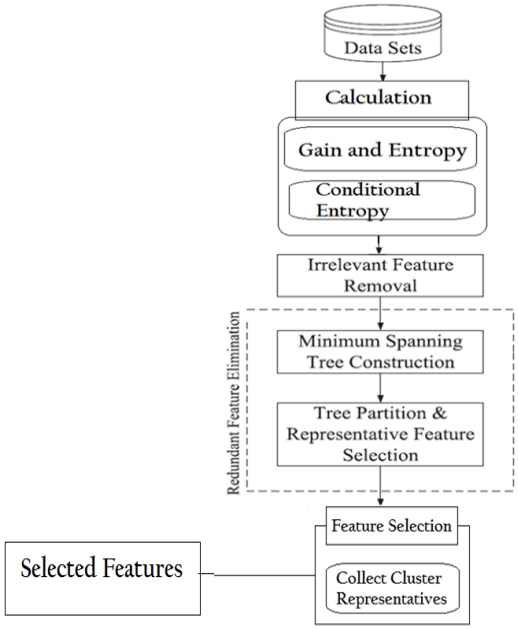
III. FEATURE SELECTION

Feature Selction, overall called variable determination, characteristic decision or variable subset determination, is the system of selecting a subset of imperative attributes for usage in model improvement. The central supposition when using a trademark determination technique is that the data holds various dull or immaterial attributes. Abundance attributes are those which outfit no more information than the at present picked qualities, and superfluous aspects give no helpful information in any setting. Feature Selection routines are a subset of the more general field of trademark extraction. Feature Selection makes new aspects from limits of the first ever offers, since trademark decision gives back a subset of the attributes. Trademark decision methods are consistently used as a piece of territories where there are various aspects and generally few examples.

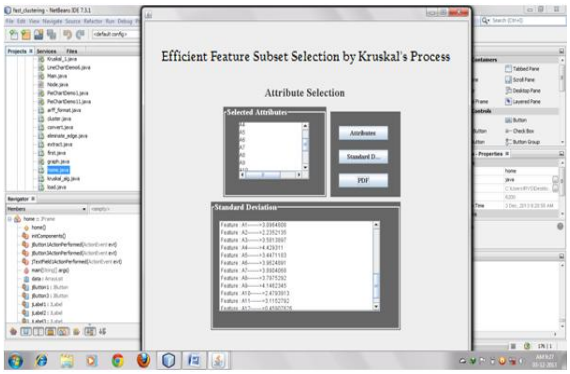
The model case is the use of trademark decision in researching DNA microarrays, where there are various numerous aspects, and several tens to a few samples. Unimportant qualities, in addition to abundance attributes, strongly impact the precision of the taking in machines. Thusly, trademark subset decision should have the ability to distinguish and evacuate however a great part of the unimportant and abundance information as could be normal. Keeping these at the highest point of the necessity record, we enhance a novel count which can viably and reasonably oversee dull aspects and get an uncommon trademark subset.

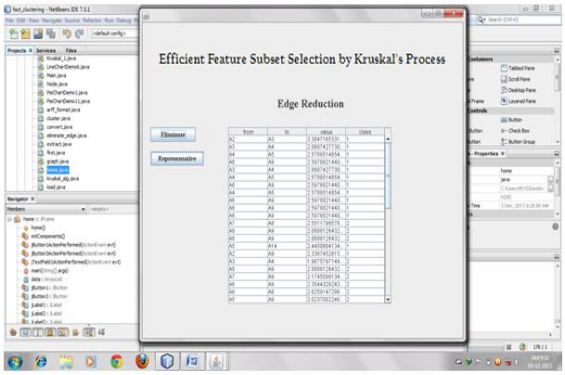
The past procures aspects paramount to the target thought by murdering unimportant ones, and the later clears abundance attributes from critical ones by method of picking operators from differing trademark clusters, and thusly handles the last subset. The superfluous trademark clearing is clear once the right relevance measure is portrayed or picked, while the tedious trademark transfer is a spot of complex.

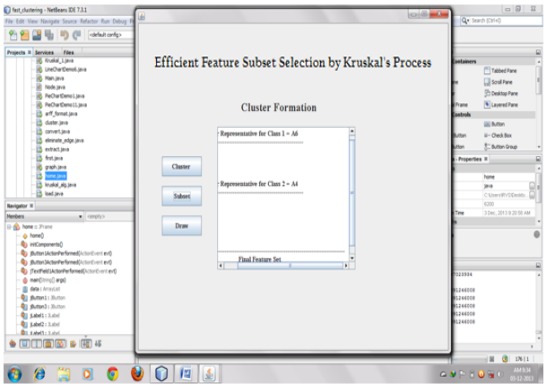
IV. DATA FLOW FRAMEWORK



V. RESULTS







VI. CONCLUSION

In this project, we have presented a novel clustering-based feature subset selection algorithm for high dimensional data. We have defined the performance of the proposed algorithm. The evaluation is done on some publicly available image, micro array and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record. Generally, the proposed algorithm obtained the best proportion of selected features, the best runtime, and the best classification accuracy for Naive Bayes, and RIPPER, and the second best classification accuracy for IB1.

REFERENCES

[1] H. Almuallim and T.G. Dietterich, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.

[2] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[3] D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.

[4] J. Biesiada and W. Duch, "Features Election for High-Dimensional data a Pearson Redundancy Based Filter," Advances in Soft Computing, vol. 45, pp. 242-249, 2008.

[5] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf. Data Mining, pp. 581-584, 2005.

[6] C. Cardie, "Using Decision Trees to Improve Case-Based Learning," Proc. 10th Int'l Conf. Machine Learning, pp. 25-32, 1993.

[7] P. Chanda, Y. Cho, A. Zhang, and M. Ramanathan, "Mining of Attribute Interactions Using Information Theoretic Metrics," Proc. IEEE Int'l Conf. Data Mining Workshops, pp. 350-355, 2009.

[8] S. Chikhi and S. Benhammada, "ReliefMSS: A Variation on a Feature Ranking Relieff Algorithm," Int'l J. Business Intelligence and Data Mining, vol. 4, nos. 3/4, pp. 375-390, 2009.

[9] W. Cohen, "Fast Effective Rule Induction," Proc. 12th Int'l Conf. Machine Learning (ICML '95), pp. 115-123, 1995.

[10] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.