Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.3, 2019, pp.35-37

DOI: https://doi.org/10.51983/ajcst-2019.8.3.2739

A Research Travelogue on Feature Sub Set Selection Algorithms

R. Ravikumar[1] and M. Babu Reddy[2]
[1]Research Scholar, [2]Assistant Professor
[1&2]Department of Computer Science, Krishna University, Machilipatnam, Andhra Pradesh, India
E-Mail: ravikumar.racha@gmail.com, m_babureddy@yahoo.com
(Received 7 October 2019; Revised 19 October 2019; Accepted 4 November 2019; Available online 16 November 2019)

Abstract **- In machine learning as the dimensionality of the data rises, the amount of data required to provide a reliable analysis grows exponentially. To perform dimensionality reduction on high-dimensional micro array data, many different feature selection and feature extraction methods exist and they are being widely used. All these methods aim to remove redundant and irrelevant features so that classification of new instances will be more accurate. Analyzing microarrays can be difficult due to the size of the data they provide. In addition the complicated relations among the different genes make analysis more difficult and removing excess features can improve the quality of the results. Feature selection has been an active and fruitful field of research area in pattern recognition, machine learning, statistics and data mining communities. The main objective of this paper is feature selection is to choose a subset of input variables by eliminating features.**
Keywords: **Classification, Clustering, Feature Selection, Machine Learning, SVM**

I. INTRODUCTION

The point of picking a subset of good highlights concerning the objective ideas, include subset choice is a viable route for diminishing dimensionality, expelling unessential information, enhancing result intelligibility and expanding learning precision. Highlight subset determination is the most ideal route for disposing of unessential information and repetitive information, decreasing dimensionality, expanding exactness. There are different component subset choice strategies in machine learning applications and they are characterized into four classifications: Embedded, wrapper, channel and half and half methodologies [1] [12]. Here we proposed a FAST calculation in view of the MST technique. The FAST calculation works in two stages. In the first step, highlights are partitioned to different groups by utilizing diagram theoretic bunching techniques [9] [14].

Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learning results. Feature selection in supervised learning has a main goal of finding a feature subset that produces higher classification accuracy [3]. As the dimensionality of a domain expands, the number of features N increases. Finding an optimal feature subset is intractable and problems related feature selections have been proved to be NP-hard [5]. At this juncture, it is essential to describe the traditional feature selection process, which consists of four basic steps, namely,

1. Subset generation,
2. Subset evaluation,
3. Stopping criterion, and
4. Validation.

Subset generation is a search process that produces candidate feature subsets for evaluation based on a certain search strategy [6] [8]. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation. If the new subset turns to be better, it replaces best one.

II. RELATED WORK

Many real world databases are grown up highly unpredictable and unbelievable. Every piece of information attracts the attention, creates a need and insists to collect and store it in the database in its original form. Automatically no one can escape from the tremendous growth of data facts, often referred a “feature”. Then feature selection becomes a major process in data classification after data preprocessing. The main challenge is here to remove the unwanted features from the dataset that were irrelevant to the task performed. It could be extremely useful in reducing the dimensionality of the data [11] [14].

In machine learning as the dimensionality of the data rises, the amount of data required to provide a reliable analysis grows exponentially [2]. To perform dimensionality reduction on high-dimensional micro array data, many different feature selection and feature extraction methods exist and they are being widely used. All these methods aim to remove redundant and irrelevant features so that classification of new instances will be more accurate [6]. Analyzing microarrays can be difficult due to the size of the data they provide. In addition the complicated relations among the different genes make analysis more difficult and removing excess features can improve the quality of the results [10]. In this work, the feature selection methods are applied on learning tasks. Feature selection has been an active and fruitful field of research area in pattern recognition, machine learning, statistics and data mining communities. The main objective of feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information [3] [5]. Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learning

results. Feature selection in supervised learning has a main goal of finding a feature subset that produces higher classification accuracy [5]. As the dimensionality of a domain expands, the number of features N increases. Finding an optimal feature subset is intractable and problems related feature selections have been proved to be NP-hard.

III. OVERVIEW OF FEATURE SELECTION ALGORITHMS

This process is repeated until a given stopping condition is satisfied. Ranking of features determines the importance of any individual feature, neglecting their possible interactions. Ranking methods are based on statistics, information theory, or on some functions of classifier outputs. Algorithms for feature selection fall into two broad categories namely

1. Wrappers that use the learning algorithm itself to evaluate the usefulness of features and
2. Filters that evaluate features according to heuristics based on general characteristics of the data.

Feature selection is the essential preprocessing step in Data mining. Several feature selection algorithms are available [1] [7]. Each algorithm has its own strength and weakness.

TABLE I SUMMARY OF FEATURE SELECTION ALGORITHMS

Algorithm	Type	Approaches Used	Merits	Demerits
Relief	Filter	Relevance Evaluation	It is scalable to data set with increasing dimensionality.	It cannot eliminate the redundant features.
Correlation-based Feature Selection	Filter	Uses Symmetric Uncertainty (for calculating Feature Class and Feature Feature correlation)	It handles both irrelevant and redundant features and It prevents the reintroduction of redundant features.	It works well on smaller datasets It cannot handle numeric class problems
Fast Correlation Based Filter	Filter	Uses predominant correlation as a good measure, based on symmetric uncertainty (SU).	It hugely reduces the dimensionality	It cannot handle feature redundancy.
Interact	Filter	Uses symmetric uncertainty and Backward Elimination Approach	It improves the accuracy.	Its mining performance decreases, as the dimensionality increases.
Fast Clustering-Based Feature Subset Selection	Filter	Graph-Theoretic Clustering Method used for clustering and a best feature is chosen from each cluster.	Dimensionality is hugely reduced	Works well only for Microarray data.
Condition Dynamic Mutual Information Feature Selection	Filter	Mutual Information	Better Performance	Sensitive to noise
Affinity Propagation – Sequential Feature Selection	Wrapper	Affinity Propagation clustering algorithm applied to get the clusters SFS applied for each cluster to get the best subset	Faster than Sequential Feature Selection	Accuracy is not better than SFS
Evolutionary Local Selection Algorithm	Wrapper	K-Means Algorithm used for clustering	Covers a large space of possible feature combinations	As the number of features increases, the cluster quality decreases.
Wrapper Based Feature Selection using SVM	Wrapper	Sequential Forward Selection for feature selection SVM for evaluation	Better Accuracy and Faster Computation	
Two-Phase Feature Selection Approach	Hybrid	(Filter) Artificial Neural Network Weight Analysis used to remove irrelevant features. (Wrapper) Genetic Algorithm used to remove redundant features	Handles both irrelevant and Redundant features. Improves Accuracy	
Hybrid Feature Selection	Hybrid	(Filter) Mutual Information (Wrapper)Wrapper model based feature selection algorithm which uses Shepley value	Improves Accuracy	High Computation Cost for high dimensional data set

Some algorithms involve only in the selection of relevant features without considering redundancy. Dimensionality increases unnecessarily because of redundant features and it also affects the learning performance. And some algorithms select relevant features without considering the presence of noisy data. Presence of noisy data leads to poor learning performance and increases the computational time. Our study concludes that there is a need for an effective unified framework for feature selection which should involve in the selection of best feature subset without any redundant and noisy data [13].

To overcome the disadvantages of existing algorithms, genetic algorithms will be used. The main advantages of this proposed system are

1. It can select high-quality feature subsets for a particular classifier.
2. It creates new variables as combinations of others to reduce the dimensionality of the selected features.
3. Low computational cost
4. Low time complexity
5. Improved efficiency and accuracy of predicting the features.

Different feature selection and feature extraction methods were compared. Their advantages and disadvantages were also discussed. In addition, several methods that incorporate prior knowledge from various sources which is a way of increasing the accuracy and reducing the computational complexity of existing methods have been presented.

IV. CONCLUSION

Transfer learning aims to use acquired knowledge from existing (source) domains to improve learning performance on a different but similar (target) domains. Feature-based transfer learning builds a common feature space, which can minimize differences between source and target domains. However, most existing feature-based approaches usually build a common feature space with certain assumptions about the differences between domains. The number of common features needs to be predefined. In this research, we propose an optimization of learning tasks using feature selection method using particle swarm optimization (PSO), ACO, or Genetic Algorithm, mPSO, where a new fitness function is developed to guide the suitable algorithm to select a number of original features and shift source and target domains to be closer. Classification performance is used in the proposed fitness function to maintain the discriminative ability of selected features in both domains. The use of classification accuracy leads to a minimum number of model assumptions.

REFERENCES

[1] H. Liu and Z. Zhao, "Manipulating data and dimension reduction methods: Feature selection", *in Encyclopedia of Complexity and Systems Science. Berlin, Germany: Springer*, pp. 5348–5359, 2009.

[2] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining", *in Proc. JMLR Feature Sel. Data Min.*, Hyderabad, India, Vol. 10. pp. 4–13, 2010.

[3] Y. Zhai, Y. S. Ong, and I. W. Tsang, "The emerging 'big dimensionality", *IEEE Comput. Intell. Mag.*, Vol. 9, No. 3, pp. 14–26, Aug. 2014.

[4] P. Praveen, and B. Rama, "A Novel Approach to Improve the Performance of Divisive Clustering-BST", In: S. Satapathy, V. Bhateja, K. Raju, B. Janakiramaiah (eds) Data Engineering and Intelligent Computing. *Advances in Intelligent Systems and Computing*, Vol. 542, *Springer*, Singapore, 2018.

[5] V. Bolón-Canedo, N. Sánchez-Maroño and A. AlonsoBetanzos, "A review of feature selection methods on synthetic data", *Knowledge and information systems*, Vol. 34, No.3, pp. 483-519, 2013.

[6] V. Bolón-Canedo, N. Sánchez-Maroño and A. AlonsoBetanzos, "Recent advances and emerging challenges of feature selection in the context of big data", *Knowledge and information systems*, Vol. 86, pp. 33-45, Sept. 2015

[7] Saroj and Jyoti, "Multi-Objective Genetic Algorithm Approach to Feature Subset Optimization", *IEEE International Advance Computing Conference (IACC)*, 2014.

[8] R. Ravi Kumar, M. Babu Reddy and P. Praveen, "A review of feature subset selection on unsupervised learning", *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, pp. 163-167, 2017. DOI: 10.1109/AEEICB.2017.7972404, 2017

[9] Vasily Sachnev and Hyoung Kim, "Binary Coded Genetic Algorithm with Ensemble Classifier for Feature in JPEG Steg analysis", *IEEE Nineth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, April 2014.

[10] Hoai Bach Nguyen, Bing Xue, Ivy Liu and Mengjie Zhang. "Filter Based Backward Elimination in Wrapper based PSO For Feature Selection Classification", *IEEE Congress On Evolutionary Computation (CEC)*, July 2014.

[11] P. Praveen and B. Rama, "An empirical comparison of Clustering using hierarchical methods and K-means", *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, pp. 445-449, 2016. DOI: 10.1109/AEEICB.2016.7538328

[12] Shima Kashef and Hossein Nezamabadi "A new Feature Selection Algorithm based on binary ant colony optimization", *IEEE 5th conference on Information and Knowledge Technology (IKT)*, 2013.

[13] P. Praveen, C. J. Babu and B. Rama, "Big data environment for geospatial data analysis", *2016 International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, pp. 1-6, 2016. DOI: 10.1109/CESYS.2016.7889816

[14] P. Praveen, B. Rama and T. Sampath Kumar, "An efficient clustering algorithm of minimum Spanning Tree", *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, pp. 131-135, 2017. DOI: 10.1109/AEEICB.2017.7972398