International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V68I10P208

Volume 68 Issue 10, 48-51, October 2020
© 2020 Seventh Sense Research Group®

Compound Feature Generation And Boosting Model For Cancer Gene Classification

S. Jafar Ali Ibrahim [a,1], A. Mohamed Affir [b,2], M. Thangamani [c,3], S. Nallusamy [d,4]

[a] *Associate Professor, Department of IT,QIS College of Engineering and Technology, Ongole, AP, India*

[b] *Manager- R&D, Holeteq Group, Doha, Qatar*

[c] *Associate Professor, Department of Information Technology, Kongu Engineering College, Tamilnadu, India*

[d] *Professor & Dean, Department of Mechanical Engineering, Dr. M.G.R. Educational and Research Institute, Chennai - 600 095, Tamil Nadu, India*

[1]jafartheni@gmail.com, [2]md.affir@gmail.com, [3]manithangamani2@gmail.com, [4]ksnallu@gmail.com

***Abstract** - The huge-data processing applications are conducted utilizing data mining or deep learning approaches. In data processing and deep learning systems, computational complexity is the key problem. High dimensional data analysis requires immense computing time and computer capital. For improved visuality, optimization of data, elimination of noise and comprehensible factors and generalization, dimensionality restriction methods are implemented. The dimensionality reduction activities monitor the data output. In the high dimensional data world, feature selection models are implemented to minimize complexity. Throughout the potential selection process, sub-set filtering with significance element is considered. In the function selection process, quantitative techniques are implemented. The poor results of the T-test configuration are found. F-test models disable the unnecessary functions. To test the apps, Q-statistics activities are added. For the practical enhancement cycle, the booster algorithm is used. For the classification method, the Naïve Bayes algorithm is used. Dynamic characteristics are identified with the filtering methods of the applications. The retrieval of characteristics is implemented in the microscope data values to catch complex properties. The method for integrating feature discovery with abstraction is added to the compound object creation. Many percentage-based attribute associations are introduced for app incorporation. The boosting approach is combined with the production of compound functions. The classification is performed using the algorithm Naïve Bayes with function values produced.*

Keywords - *High Dimensional Data Classification, Feature Selection, Feature Extraction, Feature Generation and Naïve Bayesian Classifier*

I. INTRODUCTION

The collection of function subsets is an efficient way to cut down complexity, delete redundant details, improve learning precision, and increase results understandability in order to pick a sub-set of appropriate characteristics for appropriate concepts. Many approaches for choosing sub-sets for machine learning have been suggested and researched. They can be classified into four major categories: Integrated, Shell, Filter and Hybrid. The built-in methods integrate functional filtering as part of the teaching phase and are typically unique to defined learning algorithms. Integrated methods are representations of conventional machine learning techniques such as decision trees or artificial neural networks. In order to evaluate the quality of the chosen subsets, wrapper approaches use the predictive preciseness of the prescribed learning algorithm, which is normally large. Nevertheless, the conceptual model and the computational sophistication of the chosen features are reduced. The filtering approaches have strong generality independent of learning algorithms. Their code sophistication is low, but their precision is not promised. The hybrid approaches are a mixture of a filter and wrapper approach to raising the breadth of the quest that is taken into account by the following wrapper. They mostly merged filter and wrapper methods with a different learning algorithm with equal time difficulty of the filter processes to obtain the best possible results. The wrapper mechanisms are computationally inefficient and appear to overwhelm limited amounts of instruction. Besides the tautology of the filter methods, the number of features is typically a strong choice. So, in this paper we must concentrate on the filter process. The use of cluster analyzes has proven to be more powerful than conventional feature selection algorithms for filter feature selection methods. The distributional clustering of terms was used by Pereira et al., Baker et al., and Dhillon, etc. to the aspect of text results.

II. ASSOCIATED WORK

The design of the task sub-set may be described as a phase in which the most appropriate and repetitive aspects are found and eliminated. That is since (i) obsolete characteristics do not add to forecasting precision and (ii) redundant characteristics do not refreshed a reasonable indicator with the bulk of knowledge that is already usable in the other characteristics. For the numerous sub-set filtering algorithms for apps, some may efficiently remove obsolete apps, but some of them may also not manage redundant features. The second category contains the new FAST algorithm. Throughout the past, looking for appropriate functionality has become a priority of the function section. One well known example is the relief that weights each element by its capacity to distinguish instances against specific goals on the basis of the task of distance criteria [1].

Redundancy relief becomes unstable when two neutral yet strongly associated features are discarded as they are also likely to be heavily weighted. Relief-F expands relief that enables this process, but still cannot understand redundant functions, to function with noisy and incomplete databases and cope with multi-class problems. Redundant characteristics influence the speed and precision of learning algorithms along with redundant functions and can thus therefore be removed. For CFS, FCBF and CMIM take redundant functions into account. The theory for CFS is that a reasonable subset of apps includes apps that are directly related, but not connected to, the target. FCBF is a quick filter approach that determines the relevant features without parallel correlation analysis as well as the similarity between relevant features. CMIM selects features sequential to optimize their reciprocal knowledge, based on the answer of some already defined function. Our proposed Quick algorithm uses a machine learning approach to pick items, unlike these algorithms. In the sense of text grouping, hierarchical clustering was recently introduced in word search. Distribution-related clusters have been used to classify terms into classes on the basis of either their involvement or the distribution by Baker and McCallum of the class-related marks. Dhillon, entre, has suggested a modern knowledge theoretical controversial algorithm for word clustering, since the distributional clustering of word is agglomerate in nature and results in sub-optimal word fragments and high computing costs.

Butterworth et al. suggested the usage of a special Barthelemy-metric for cluster characteristics and instead allowed use of the cluster hierarchical dendrogram to pick the most important attributes. Unfortunately, the Barthelemy-Montjardet distance cluster assessment measure does not imply a subset of functionality that enables linear regressions to boost their original output accuracy. In fact, the precision achieved is poorer in contrast to other forms of function collection. Hierarchical clustering was often used to pick spectral data functionality. The sub-set selection algorithm for regression has been introduced by the Van Dijk and the Van Hullefor [2].

A methodology which combined hierarchically restricted spectral variables clustering with the reciprocal knowledge option of clusters was implemented [3]. Their system of classification is identical to Van Dijk and Van Hullefor, with the distinction that former clusters only include consecutive characteristics. In order to eliminate unnecessary functions, both approaches employed clustering. Our suggested FAST algorithm uses a minimal covering tree approach to cluster functionality, somewhat different from other hierarchical clustering related algorithms. In the meanwhile, the data points are not considered to be clustered around centres or divided by a normal geometric curve. In comparison, our suggested FAST will not restrict itself to any unique data forms.

III. CLASSIFICATION ELEMENT COLLECTION AND BOOSTER

For several functional implementations, for example, data processing, deep learning, and study of gene expression, the use of high-dimensional data is becoming much more popular. Typical micro array data open to the public has hundreds of thousands of features with limited samples and the features taken into consideration in the data review are that. There is a basic problem in the statistical analysis of results, with a vast number of features and limited sample size. It turns out that the simple, common linear Fisher analysis may be as weak as random variance, as the number of features increases. Many of the characteristics of high dimensional microarray samples are unrelated to the goal trait, so the proportion of applicable characteristics or the ratio of up-regulated or down regulated genes is just 2 percent to 5 percent, compared with acceptable normal tissue. The detection of specific features promotes learning and enhances prediction accuracy. Nevertheless, changes in structure results, especially in biomedical studies, should be considered to be fairly resilient as domain experts would spend significant time and energy in this limited collection of selected features. Therefore, the range suggested would not only have the high but also the moderate predictive capacity.

Over the last two decades, several FS works has been done, and the study remains a hot subject in the area of machine learning [4-7]. One frequently used method is to first discrete continuous features in this pre-processing phase and to pick appropriate features by using Mutual Information (MI). That's because it is fairly easy to locate the related properties dependent on a discretized MI, although it is actually a wonderful task to extract specific characteristics directly from a great range of functions with continuous values. For FS problems may be utilized approaches used for the identification of predictive predictor problems, such as forward collection, retrograde exclusion and its adaptation. Many popular FS algorithms in higher dimensional challenges have used an automated approach of sorting, but not a reverse elimination approach as reverse elimination processes with a range of features are unworkable. A big difficulty in the forward collection is that a flip of the original function judgment will lead to a totally different subset of functions, and thus the chosen function set's reliability is very poor even though the specification is extremely accurate. The consistency problem in FS is defined as this. Throughout this sector, the work is fairly new [8-12] and a difficult area of research is to establish an effective approach of obtaining a stronger and more robust sub-set of features. In this article, Q-statistics was suggested to test the output of a classifier FS algorithm. The estimation of the classificatory and the reliability of the chosen elements were calculated hybridly.

The journal instead proposes a booster for the function subset collection of a specific FS algorithm. Booster's fundamental concept is to generate several data sets on sample space while utilizing the initial data collection. The FS-algorithm is then used to generate separate subset of features for each of these data samples. The combination of the chosen subsets is the subset function of the FS algorithm Booster. Empirical evidence suggests that the algorithm booster not only improves the value of the Q-statistic but also the classifier's predictive accuracy incorporated. Numerous studies have been conducted based on resampling technology in order to generate various classification problem sets [13] and some of the studies have used feature space re-sampling [14-18]. The aim of all these studies is to predict the classification accuracy without taking into account the consistency of the selections.

IV. PROBLEM STATEMENT

The attributes that are weakly important are found with the T-test model. F-test model removes the inconsequential characteristics. To assess the characteristics, Q-statistics measures are used. For the feature continuous improvement, the booster algorithm is used. For classification the Naïve Bayes algorithm is used. In the existing system, the following issues are identified. The framework does not support extracting features operations. There is no support for heterogeneous feature-based integration transactions. Low features extraction levels and restricted classification accuracy levels.

V. SCHEME OF FEATURE GENERATION AND DATA CLASSIFICATION BOOSTERS

Features are used to achieve the high-dimensional data classification. The characteristics are evaluated with statistical measures. With the boosting process, the features are enhanced. The system consists of six main modules. It is data pre-process, feature collection, the generation of compounds, feature analysis, strengthening process and classification algorithm. For cleaning of data, the data pre-processing module is used. Operations are conducted during the selection process of the functionality. The collection of functions and extraction processes is combined in the composite generation module. The functional analysis module assesses quality. The boosting process is designed to improve the functionality. The gene data is classified as a module of data classification.

A. Pre-Processing of Data

The data values of the gene micro-array are collected as text files. For the delivery of textual data to the Oracle servers, data complements are used. Under the data cleaning process, the noisy and redundant data values are corrected. The missing values are changed using the data replacement process of aggregation.

B. Feature Selection

The feature choice operations area unit dispensed with the information connection identification method. The T-test model is applied to get infirm relevant options. The moot options area unit is known as exploitation the F-test model. The connection and redundancy factors area unit analyzed within the feature selection method.

C. Generation of Composite Feature

The function selection process identifies the original features. The role extraction method exposes the transformed characteristics. The selection of functions and the extraction of functions is integrated into the compound function generation process. The advanced characteristics are discovered using the CFG algorithm Composite Feature Set Generation.

D. Feature Analysis

The optimal features are identified using the feature analysis model. The q-statistics is used for the feature analysis process. The q-statistics measure is calculated for the discovered features. The selected features are passed to the boosting process. The feature selection process is carried out with three algorithms. The Minimal-redundancy-maximal-relevance (mRMR) algorithm considers the redundancy and relevancy factors in the feature selection process. The Fast Correlation-Based Filter (FCBF) algorithm uses the correlation factors. The data partitioning is applied in the Fast clustering based feature Selection algorithm (FAST).

E. Feature Analysis

The best features are defined using the function analysis model. For the feature analysis process the q-statistics are used. For the functions discovered, the q-statistics measure is computed. In the boosting phase the chosen features are moved on. Three algorithms are used to select the feature. The algorithm (mRMR) deals with the redundancy-maximum value (MRMR) considerations in the feature selection process. The algorithm of Fast Correlation-Based Filter (FCBF) uses the correlation factors. The data segmenting is implemented in the FAST clustering-based feature Selection algorithm.

F. Process Boosting

To enhance the selected function, the boosting mechanism is implemented. In the boosting process, the q statistical measurement is strengthened. In the feature improvement process the booster approach is employed. The process of re-sampling is used on the area of the sample.

G. Process of Classification

In the classification process, the extent of cancer severity is determined. The Naive Bayes (NB) classifier is used in the gene analysis process. The method of learning is used for the recognition of the class models. The testing procedure is used to designate the class attributes for the gene data.

VI. CONCLUSION

For the operation of feature extraction, the feature engineering methods are applied. The procedures of classification are performed on selected functions. In the classification and feature analysis method, Q-statistics and Naïve bayes algorithms are used. To boost the selection criteria, a compound feature generation scheme is implemented. Classification of the cancer gene is carried out with characteristics. The framework combines practical collection and extraction processes. In the function generation process, static and dynamic features are combined. With low computing overhead, the machine achieves high precision levels.

REFERENCES

[1] Qinbao Song, Jingjie Ni and Guangtao Wang. "*A fast clustering-based feature subset selection algorithm for high dimensional data"*, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, pp. 65-73, 2013.

[2] Van Dijk G. and Van Hulle M.M., "*Speeding up the wrapper feature subset selection in regression by mutual information relevance and redundancy analysis*", International Conference on Artificial Neural Networks, pp. 122-132, 2006.

[3] 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.

[4] S.K. Muruganandham, D. Sobya, S. Nallusamy, Dulal Krishna Mandal and P.S. Chakraborty. "*Study on leaf segmentation using k-means and k-medoid clustering algorithm for identification of disease"*, Indian Journal of Public Health Research and Development, vol. 9, no. 2, pp. 289-293, 2018.

[5] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys. "*Robust biomarker identification for cancer diagnosis with ensemble feature selection methods*," Bioinformatics, vol. 26, no. 3, pp. 392-398, 2010.

[6] A. J. Ferreira and M. A. T. Figueiredo, "*Efficient feature selection filters for high dimensional data*," Pattern Recognised Letters, vol. 33, no. 13, pp. 1794-1804, 2012.

[7] Q. Song, J. Ni, and G. Wang, "*A fast clustering-based feature subset selection algorithm for high-dimensional data*," IEEE Trans. Knowledge. Data Eng., vol. 25, no. 1, pp. 1-14, 2013.

[8] Y. Han and L. Yu, "*A variance reduction framework for stable feature selection*," Statist. Anal. Data Mining, vol. 5, no. 5, pp. 428-445, 2012.

[9] S. Alelyan, "*On feature selection stability: A data perspective*," PhD dissertation, Arizona State Univ., Tempe, AZ, USA, 2013.

[10] D. Dernoncourt, B. Hanczar, and J. D. Zucker, "*Analysis of feature selection stability on high dimension and small sample data,*" Comput. Statist. Data Anal., vol. 71, pp. 681-693, 2014.

[11] N. Meinshausen and P. Buhlmann, "*Stability selection*," J. Roy. Statist. Soc.: Series B (Statist.Methodol.), vol. 72, no. 4, pp. 417-473, 2010.

[12] Sobya, D., Manoj, S. "*Prediction and identification of cancer and normal genes through wavelet transform technique*", *Indian Journal of Public Health Research and Development*, vol. 10, no. 8, pp. 631-637, 2019.

[13] Z. He and W. Yu, "*Stable feature selection for biomarker discovery*," Comput. Biol. Chem., vol. 34, no. 4, pp. 215-225, 2010.

[14] K. M. Ting, J. R. Wells, S. C. Tan, S. W. Teng, and G. I. Webb, "*Feature-subspace aggregating: Ensembles for stable and unstable learners*," Mach. Learn., vol. 82, no. 3, pp. 375-397, 2011.

[15] F. Alonso-Atienza, J. L. Rojo-Alvare, A. Rosado-Mu~noz, J. J. Vinagre, A. Garcia-Alberola, and G. Camps-Valls, "*Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection*," Expert Syst. Appl., vol. 39, no. 2, pp. 1956-1967, 2012.

[16] D. Dembele, "*A flexible microarray data simulataion model,*" Microarrays, vol. 2, no. 2, pp. 115-130, 2013.

[17] S. Jeyabalan, V. Cyril Raj and S. Nallusamy. "*A genetic algorithm based protein signal pathway analysis", Indian Journal of Public Health Research and Development*, vol. 9, no. 1, pp. 402-406, 2018.

[18] Ibrahim, S.J.A. and Thangamani, M. "*Enhanced singular value decomposition for prediction of drugs and diseases with Hepatocellular carcinoma based on Multi-Source Bat Algorithm based random walk"*, Measurement, vol. 141, pp. 176-183, 2019.

[19] Kassahun Azezew Ayidagn, prof. Shilpa Gite "*Analysis of Feature Selection Algorithms and a Comparative study on Heterogeneous Classifier for High Dimensional Data survey*", International Journal of Engineering Trends and Technology (IJETT), V53(2),59-63 November 2017.