Proceedings of the 2020 IISE Annual Conference
L. Cromarty, R. Shirwaiker, P. Wang, eds.

Important Variables using Novel Thresholding Approach for Pan-Cancer Analysis

Kevin Kaufman-Ortiz and Wandaliz Torres-García
University of Puerto Rico-Mayagüez Campus
Mayagüez, Puerto Rico 00681

Abstract

Numerous breakthroughs in cancer have been achieved and it is now widely recognized as a genetic disease. Still many of the molecular mechanisms are unknown, creating a gap to achieve the cure. Omics datasets are being generated at an accelerated rate and becoming more reliable with biotechnology advances, offering an enormous amount of data to study many unknowns in cancer biology and clinical informatics. These datasets are highly heterogeneous and suitable data mining techniques are needed to uncover the molecular drivers of cancer. Henceforth, we developed a feature-selection model that effectively analyzes large amounts of data from The Cancer Genome Atlas and extracts connections and patterns to molecularly understand different cancer subtypes. Our proposed scoring method gave a maximum accuracy of 91.83% to distinguish nineteen cancer types evaluated using a random forest classification model. Whereas existing methods based on statistical metrics - mean absolute deviation, standard deviation, and interquartile range- showed maximum performance of 88.33%, 79.11%, and 64.77%, respectively. This result is encouraging since there are not many methods available to automatize the process of selecting important variables when scores are given to each. PR, ERAlpha, GATA3, FASN, among other proteins were selected as important to correctly classify cancer types.

Keywords
Data mining, Feature Selection, Feature Extraction, Cancer Genomics

1. Introduction
Understanding diseases such as cancer is of great importance to our communities. In fact, in the US alone about 38.4% of its population will be diagnosed with these deadly diseases based on data statistics from the National Cancer Institute in the years of 2013-2015 [1]. Nonetheless, biotechnology advances have enabled the collection of massive amounts of molecular information to better understand the biological mechanisms of cancer initiation and development. Though many discoveries are a result of this explosion of data and computational efforts to extract meaningful knowledge through initiatives such as The Cancer Genome Atlas (TCGA), more data analytics efforts are needed to tackle the interpretation of this firehose of information from very complex systems.
Many researchers have focused on finding genomic differences and similarities across different cancer types a field commonly known as Pan-Cancer analysis. This allowed for the generation of a genomic landscape of patterns to help generate new therapies and extend existing treatments in a particular cancer type to another [2]. Existing pan-cancer studies have studied correlations between protein expression patterns for 11 different cancer types [3, 4]. In 2013, Li et al developed a user-friendly computational platform named The Cancer Protein Atlas (TCPA) to gather, analyze and visualize proteomics data [4]. A year later, Akbani et al studied differentially expressed proteins across 11 cancer types using an integrative approach that revealed that protein levels are highly linked to cancer types and that there exist protein pairs such as MYH11 and RICTOR that are highly correlated across different tumor types [3]. Yet, many of the molecular mechanisms underlying these diseases are still not clearly understood creating a gap to achieve the cure. Thus, it is important to understand the mechanisms that characterize the different cancer types to see which treatments may overlap and transfer easily from one cancer type to another.
Data analytics efforts have taken a predominant role to extract knowledge from a large amount of data and understand diseases such as cancer. In particular, there is a step in the data mining process called feature selection aiming to rank important variable predictors such as proteins that can describe groups of interests such as the diverse cancer types. There exist tons of these methods, each using a different scoring criterion and yielding different results depending on the application. Existing feature selection methods are mainly categorized as filters, wrappers and embedded methods. Many of these are often implemented in practice and the results from the best-performing approach are often utilized

for biological interpretation. Wrapper methods evaluate the performance of a given feature subset using data mining algorithms while embedded methods perform a search for an optimal subset of features that are built in the classifier structure thus making them specific to a given learning algorithm [4]. These convoluted methods (wrapper and embedded) tend to provide good performance in practice but tend to be computationally inefficient in many cases when compared to filter techniques. In particular, wrappers have a higher risk of overfitting than filters [5].

Filter methods tend to be very fast and scalable. In many applications, they can provide as good results as those from wrappers and embedded approaches [5]. These filters calculate relevance scores for each feature that permits the feature ranking by importance. However, to the best of our knowledge, there are not many approaches that establish where to cut and extract important variables optimally. Many studies use statistical metrics, percentiles, and trial-and-error thresholds to narrow the list of important features. For example, when the scores are p-values from multiple hypotheses one could establish a particular significance level after correction for experiment-wise error but even these have their limitations. Reducing the list of important variables if their adjusted p-values are less than 0.01 will mean that a feature with a p-value of 0.009 will be included in the importance list but a variable with a p-value equal to 0.0101 will not when in reality they might be both important to evaluate further. Hence, there is a need for an automatic method that can extract important features from a list of scored features by looking at the changes in their scores which is focused on this work to help uncover important proteins that can characterize different cancer types.

2. Methodology

Data scientists commonly follow a general data-to-knowledge framework to solve a particular problem. Often, this framework consists of data selection, data processing, data transformation, data mining, integration and evaluation. In this work, the process is broken down into analogous steps depicted in **Figure 1** and described in this section.

Data Selection	Preprocessing	Feature Selection	Cut-off Threshold Methods	Performance Evaluation
TCGA Protein Expression Data	• Removed rows with missing values (NA) • Removed Irrelevant Columns	• Correlation Based Feature Selection (CFS) • Correlation (CA) • Gain Ratio (GRA) • Info Gain (IGA) • One R (ORA) • Symmetrical Uncertainty (SUA)	• Standard Deviation (SD) • Interquartile Range (IQR) • Mean Absolute Deviation (MAD) • Tolerance Below Deviation (TBD)	Classification Model using Random Forest

Figure 1: General methodology framework.

2.1 Data Selection & Preprocessing

The focus of this research is to find distinct patterns across different types of cancer to uncover new knowledge regarding the relationship of these diseases using data mining techniques. This Pan-Cancer analysis will provide a comprehensive set of molecular patterns that are distinctive to each type of cancer. These patterns will be learned TCGA data repository which is a comprehensive platform of molecular data gathered through consistent and coherent protocols reducing factors of variability due to experimentation procedures [3]. There are many types of molecular assays available but for the purposes of this work, we have focused on data at the proteomic level since it can describe functional aspects of the tumorigenesis as embodied in the central dogma. Protein expression data using reverse-phase protein arrays (RPPA) is available within TCGA for a large number of tumor samples [4]. RPPA is a quantitative technology based on targeting specific antibodies to improve sensitivity and to assess several protein markers at the time across multiple samples in a cost-effective [6]. Once data is gathered from TCGA, we proceed to preprocess it for further use. Preprocessing is a stage where several important actions on the raw data are performed to facilitate its analysis (i.e. cleaning, reorganization). Irrelevant columns such as age and rows containing at least one missing value (known as NA) were removed, resulting in a data structure of 217 columns and 4,979 rows representing proteins and patient samples, respectively (see Table 1). There was information available for all 4,979 patient tumor samples regarding their diagnosed type of cancer.

Table 1: Resulting dataset description after pre-processing.

Categorical Response Variable	Number of Patient Samples	Numeric Predictor Variables	Number of Proteins
Cancer Type	4979	RPPA Protein Expression Levels	217 (e.g. ERAlpha)

2.2 Feature Selection

Feature selection methods are useful to determine which predictor variables (i.e. proteins) are relevant to the problem at hand (cancer subtype classification). In this work, a protein that is found highly over-expressed or under-expressed when compared from one cancer type to another could mean that this protein has a significant role in the pathways that govern particular cancers. To detect this over- or under- expression, many scoring feature selection methods will provide a high score. The scoring methods used in this work are: the Correlation Based Feature Selection (abbreviated as CFS), Correlation Attribute (CA), Gain Ratio (GRA), Info Gain (IGA), One R (ORA), Relief F (RFA) and Symmetrical uncertainty (SUA), all provided through the software Weka. For example, IGA is a common entropy-based metric aiming to measure the amount of information gained by a particular feature. Moreover, in Figure 2 we provide an example of OneR scores in the column labeled "Average Merit" sorted from largest to smallest. More information on how each score is calculated can be found on Weka's SourceForge site.

2.3 Cut-off Threshold Methods

Once these scores are sorted, a cut-off threshold must be established to reduce the number of features. Establishing this threshold is not an easy task and among the most common methods are the use of graphical methods, arbitrary percentile levels, and statistical metrics. Very often visual methods result in inconsistent results if the experiment were to be replicated while using percentile levels can leave out features with a similar score. In 2016, Pramokchon and P. Piamsa-nga used several simple statistical methods to determine this cut-off threshold [7]. Their methodology seeks to find the *outlier cut-off threshold (*θ*)* using *The Standard Deviation (SD) Method*, *Interquartile Range (IQR) Method*, and *Mean Absolute Deviation (MAD) Method* are described in Equations 1, 2, and 3, respectively.

$$\theta_{upper} = \mu + \alpha * \sigma \quad (1)$$

where, μ = mean of score, σ = Standard deviation, α = confidence coefficient (common α: **1.5** or 3)

$$\theta_{upper} = Q3 + \alpha * IQR \quad (2)$$

where, $Q3$ = Upper Quartile, IQR = $Q3$ - $Q1$, α = confidence coefficient (common α: **1.5** to 5)

$$\theta_{upper} = M + \alpha * MAD \quad (3)$$

where, M = median, $MAD = 1.483 * median(|x_i - M|)$, α = confidence (common α: 2, **2.5** or 3)

Moreover, the proposed algorithm called Tolerance Below Deviation (TBD) has been developed to eliminate the "artistic factor" when determining a cut-off value for which to compare scores provided by the scoring methods implemented with Weka. The algorithm shown in Figure 2 has been named after the parameter created. To use the algorithm, there are several steps to take. First, the merit or score determined by Weka must be sorted from largest to smallest. Then, the differences between consecutive scores and the standard deviation of these differences must be calculated. Subsequently, the analyst must determine the number of scores' differences that fall below the calculated standard deviation consecutively. This is demonstrated in the "Runs Below Deviation" column shown in Figure 2. The TBD parameter must be chosen by the user. This number can be determined by user experience or by implementing a sensitivity analysis to determine the best value for it. For visualization purposes, the differences (light blue) and the standard deviation of the differences (dark blue) have been graphed in Figure 2. The general idea of this algorithm is to establish a cut-off value where the differences between ranked features converge to a certain point.

2.4 Performance Evaluation

To evaluate and compare the performance of the proposed methods with those existing from the literature, we built classification models to distinguishing nineteen different cancer types (i.e. categorical response) based on the protein expression levels (numeric predictors). Several classifier modalities could have been implemented, in this work we only used random forest models using R [8] to evaluate the predictive performance of the different protein subsets extracted from the different cut-off threshold methods. The main parameters in random forest are the number of variables evaluated before determining a split (mtry) and the number of trees used before averaging the results (ntree). These parameters, mtry and ntree, can be optimized by iterating a set of possible value combinations. The execution time to optimize these parameters depends on the number of variables that are determined in the feature extraction stage, see Exec. time column in Table 3. Among the performance metrics evaluated in this study are: (1) the execution time of the parameter tuning/running of the algorithms, (2) the accuracy of the model calculated as 1 minus the out-of-bag (OOB) error, and (3) the area under the curve (AUC) of the receiver operating characteristic (ROC) [9].

Protein	Avg. Merit	Differences	Runs Below Deviation
PR	31.46	2.27	0
ERALPHA	29.20	0.66	0
FASN	28.54	0.63	0
PKCALPHA	27.91	0.05	1
CLAUDIN7	27.86	0.86	0
PKCALPHA_pS657	27.00	0.63	0
ERK2	26.37	0.16	1
AR	*26.21*	*0.08*	***2***

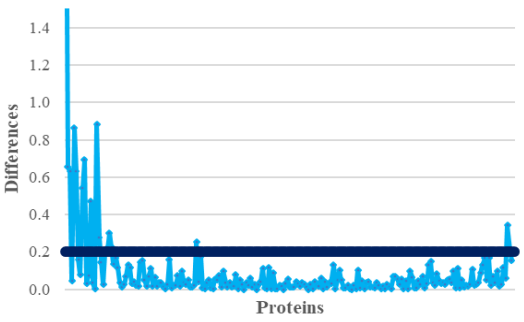

Figure 2: Demonstration of Tolerance Below Deviation (e.g. TBD = 2), Standard Deviation = 0.200

3. Results

3.1 Evaluation of Cut-off Threshold Methods

To optimize the selection of the TBD parameter in the proposed approach, the OOB error and execution time to classify the nineteen cancer types was measured across a range of possible TBD values (1-15) using Random Forest across six different scoring FS methods. The number of trees at each of these random forest models were set to 5000 at which their error rate reached steady state while the square root of the number of variables (n) was used as the recommended mtry value. To minimize both OOB error rate and execution time, TBD = 7 was chosen as a good trade-off point where most scoring methods gave an accuracy of around 90% (solid lines) as shown in Figure 3. After tuning the TBD parameters, we proceeded to compare its performance with other cut-off methods. The proposed method, TBD, consistently showed smaller OOB error rates than those from SD, IQR, and MAD methods across all scoring FS approaches as shown in Table 2. These methods consistently selected a very small number of proteins.

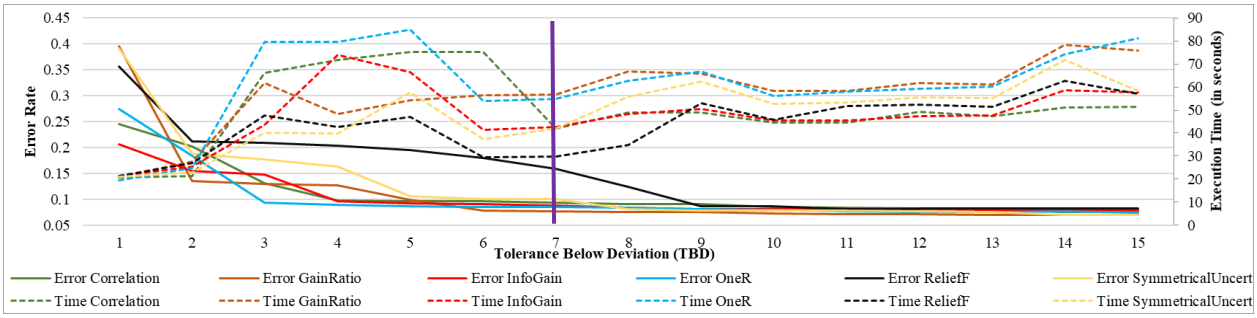

Figure 3: Sensitivity analysis plots.

Table 2: Comparison of TBD against existing SD, IQR, and MAD methods using OOB error rate from Random Forest classifiers. The variable p is the number of proteins kept after the cut-off.

Attribute Evaluator	Standard Deviation (SD)		Interquartile (IQR)		Mean Absolute Deviation (MAD)		Tolerance Below Deviation (TBD)	
	p	OOB Error	p	OOB Error	p	OOB Error	p	OOB Error
Correlation	2	56.88%	0	--	2	56.82%	20	9.46%
GainRatio	4	32.68%	0	--	6	21.23%	29	8.17%
InfoGain	6	20.73%	2	53.75%	12	12.77%	21	9.08%
OneR	6	20.89%	2	53.69%	13	12.45%	29	8.58%
ReliefF	6	25.51%	4	35.23%	16	11.67%	11	17.90%
SymmetricalUncert	5	23.68%	2	53.83%	10	15.67%	18	9.86%

Although accuracy has mostly been the basis of discussion, a desirability function (See Eq. 4) was constructed to compare how well the model estimates several metrics using a weighted average. The model metrics considered were the prediction accuracy, sensitivity and (AUC), and the amount of time the model takes to run (execution time). The weights 40%, 40%, and 20% were assigned to accuracy, AUC, and execution time, respectively. The desirability

values from scoring methods with TBD outperformed the CFS desirability of 0.563 in all cases showcasing the superiority of the TBD approach. Additionally, the ReliefF scoring method was eliminated from further analysis as it showed low accuracy and desirability values (see Figure 4 & Table 3).

$$D = Weight_1 * Accuracy + Weight_2 * AUC - Weight_3 * Scaled\_ExecutionTime \quad (4)$$

Table 3: Results with parameter tuning using TBD = 7 including CFS.

Attribute Evaluator	# Proteins	Cut-off	Mtry, Ntree	Accuracy (Acc.)	ROC AUC	Exec. time	Desirability Function (D)
CfsSubsetEval	153	10 folds	9, 4000	95.02%	95.81%	2.4 days	0.563
Correlation w/ TBD	20	≥0.157	2, 5000	90.54%	91.23%	1.6 hours	0.722
GainRatio w/TBD	29	≥0.240	3, 5000	91.83%	94.06%	3.7 hours	0.731
InfoGain w/TBD	21	≥0.592	3, 5000	90.92%	93.12%	1.7 hours	0.730
OneR w/TBD	29	≥21.519	2, 5000	91.42%	93.20%	3.5 hours	0.726
ReliefF w/ TBD	11	≥0.071	3, 5000	82.10%	85.81%	1.4 hours	0.667
SymmetricalUncert w/TBD	18	≥0.196	2, 5000	90.14%	92.46%	2.2 hours	0.723
Union of Proteins (w/o CFS and ReliefF)	42	--	4, 5000	93.41%	95.00%	9.5 hours	0.721

3.2 Pan-Cancer Analysis Important Variables

To explore the importance of those extract variables using the cut-off methods we evaluated their prevalence across methods. As shown in Table 4, nine proteins were commonly found to characterize the cancer types across all scoring FS methods. Though, the union of proteins across all scoring methods with TBD reached a high 93% accuracy and low execution time of 9.5 hours. A heatmap of these 42 variables demonstrates their relationship. For example, cancers like Ovarian (PAAD) and Stomach (STAD) tend to have overexpression of the MHY11 protein in their tumors.

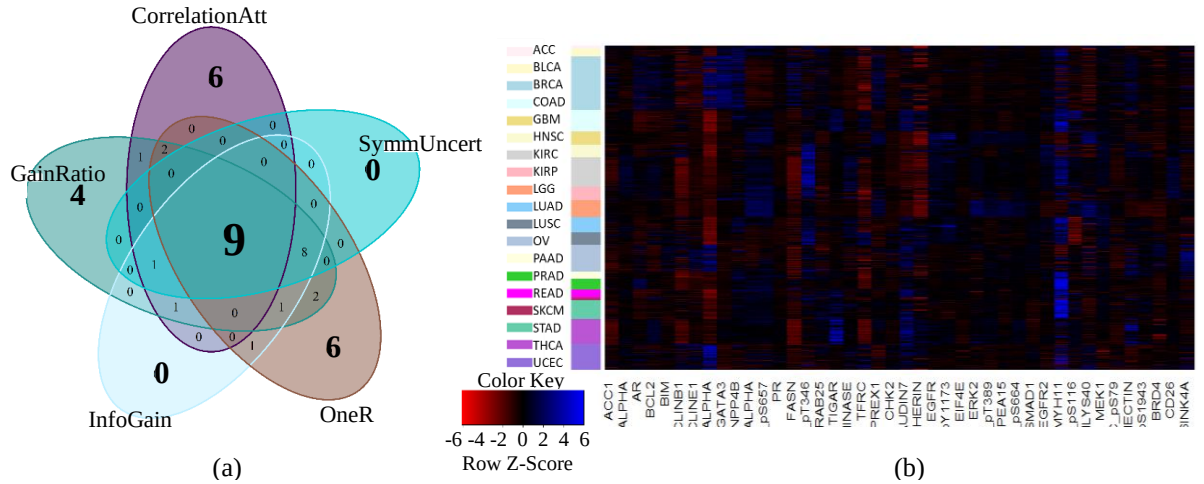

Figure 4: Important proteins (a) Venn diagram showing the relationship of proteins between feature selection methods using the TBD method. (b) Heatmap showing the relationships in protein expression across the 19 cancer types.

3.3 Biological Interpretation

Interpreting biological mechanisms from complex data mining methodologies is a hard task to achieve. Results from many well-performing methods are sometimes difficult to infer but extremely needed to advance the science of clinical translation. However, several studies such as the one presented here enable the extraction of potential biomarkers that can aid the development of new therapies. Here, we consistently found nine proteins across different feature selection methods to be relevant to classify the nineteen different cancer types as shown in Figure 4 (a). Based on the National Center for Biotechnology Information (NCBI) annotation repository, all of these proteins have previously been associated with cancer which validates the approach presented in this work. ERALPHA is one of the most relevant proteins in the study of breast cancer as well as many other hormone-related (i.e. PR). Similarly, AR is been widely

studied to understand prostate cancer. Nonetheless, interesting results are those of GATA3 and CYCLINB1 that are linked to many cancer types across different human organs. There are many cancer types in the group of studied cancers that are not well represented such as glioblastoma which could improve the accuracy performance.

Table 4: List of the top 9 commonly relevant proteins

Protein	Protein Name	Cancers associated from NCBI Annotation (https://www.ncbi.nlm.nih.gov/gene/)
AR	Androgen receptor	Prostate, Breast, Bladder
CYCLINB1	Mitotic-specific Cyclin B1	Breast, Cervical, Colorectal, Lung
ERALPHA	Estrogen receptor 1 (alpha)	Breast, Endometrial
GATA3	GATA binding protein 3	Breast, Colorectal, Acute lymphoblastic leukemia, Bladder, Prostate, Renal
PKCALPHA	Protein kinase C alpha	Colon, Endometrial, Pancreatic
PKCALPHA_Ps657	Similar variant to Protein kinase C alpha	Colon, Endometrial, Pancreatic
PR	Progesterone receptor	Breast, Endometrial, Prostate, Ovarian
FASN	Fatty acid synthase	Lung, Ovarian, Bladder, Liver
TIGAR	TP53 induced glycolysis regulatory phosphatase	Lung, Renal

4. Conclusions

This work has presented a new approach to establish an automatic threshold and enabled the extraction of important predictors from scores provided by filter feature selection methods. In practice, this is a difficult problem to tackle and at times the interpretation can be hindered and biased due to trial-and-error cut-off values established. To lessen this impact, the proposed method has shown results that support a more computationally efficient way to automatically select an accurate set of relevant features across all scoring methods with desirability values greater than the 0.563 from the non-scoring FS approach, CFS. Moreover, this methodology found important features to characterize nineteen different cancer types at the proteomic level. Many of these important proteins are highly associated with many cancers validating the usability of this approach. In the future, this approach should be tested across different problems, in particular, in applications with less number of classes than the pan-cancer problem investigated here since it may provide accuracy comparable to top non-scoring methods like CFS in less time.

Acknowledgments

This work was supported by the Puerto Rico Louis Stokes Alliance for Minority Participation (PR-LSAMP).

References

[1] National Cancer Institute, “Cancer Statistics,” *Originally Published by the National Cancer Institute*. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/statistics. [Accessed 20-Jan-2020].

[2] J. N. Weinstein *et al.*, “The cancer genome atlas pan-cancer analysis project.,” *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013.

[3] R. Akbani *et al.*, “A pan-cancer proteomic perspective on the cancer genome atlas,” *Nat. Commun.*, vol. 5, 2014.

[4] J. Li *et al.*, “TCPA: A resource for cancer functional proteomics data,” *Nature Methods*, vol. 10, no. 11. Nature Methods, pp. 1046–1047, 2013.

[5] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[6] K. M. Sheehan *et al.*, “Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma,” *Mol. Cell. Proteomics*, 2005.

[7] P. Pramokchon and P. Piamsa-nga, “Effective Threshold Estimation for Filter-based Feature Selection,” *IEEE*, no. 16, p. 6, 2016.

[8] R. Díaz-Uriarte and S. Alvarez-de-Andrés, “Variable selection from random forests: application to gene expression data,” pp. 1–11, 2005.

[9] T. Fawcett, “ROC Graphs: Notes and Practical Considerations for Data Mining Researchers,” *HP Lab.*, pp. 1–28, 2003.