International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:7, 2016

Automatic Threshold Search for Heat Map Based Feature Selection: A Cancer Dataset Analysis

Carlos Huertas, Reyes Juarez-Ramirez

Abstract—Public health is one of the most critical issues today; therefore, there is great interest to improve technologies in the area of diseases detection. With machine learning and feature selection, it has been possible to aid the diagnosis of several diseases such as cancer. In this work, we present an extension to the Heat Map Based Feature Selection algorithm, this modification allows automatic threshold parameter selection that helps to improve the generalization performance of high dimensional data such as mass spectrometry. We have performed a comparison analysis using multiple cancer datasets and compare against the well known Recursive Feature Elimination algorithm and our original proposal, the results show improved classification performance that is very competitive against current techniques.

Keywords—Feature selection, mass spectrometry, biomarker discovery, cancer.

I. Introduction

WITH the emerge of new technologies, the ability to generate more data has been increased, causing that big data is now present in many areas of research, including text mining, information retrieval and bioinformatics [1], [2]. From the machine learning point of view, this increase in the data resolution (i.e., dimensions), causes that in order to produce reliable models, an exponential growth of samples need to be fed into the algorithms, if such amount of samples are not reached, we fall under the *"curse of dimensionality"* [3], this is the case of biomedical data such as microarrays where there is very little amount of samples but the features for each of those can be as high as 450,000, which implies high computational costs and prone to overfitting [2]. In order to deal with high dimensional data, it is required to reduce the number of features as usually most of them are not relevant to the target [4], more important, in areas such as bioinformatics a reduced subspace of features can help biologists to discover important insights about the disease under study [5].

According to the World Health Organization, cancer is among the leading causes of morbidity and mortality worldwide. In the *World Cancer Report 2014* it is reported that in 2012, the worldwide burden of cancer rose to an estimated 14 million new cases per year, a figure expected to rise to 22 million annually within the next two decades. Over the same period, cancer deaths are predicted to rise from an estimated 8.2 million annually to 13 million per year. Globally, in 2012 the most common cancers diagnosed were those of the lung (1.8 million cases, 13.0% of the total), breast (1.7 million, 11.9%), and large bowel (1.4 million, 9.7%). Public health is a critical issue, therefore there is a need to find faster and more accurate methods to detect diseases as early as possible, since late detection usually leads to high death rate [6].

Carlos Huertas is with the Department of Computer Science at Autonomous University of Baja California, Tijuana, B.C, Mexico (e-mail: chuertas@uabc.edu.mx).

With the development of technologies such as mass spectrometry, it has been possible to improve detection and understand diseases in a better way [7]. The idea behind the analysis of mass spectrometry data is to find biomarkers (e.g. features in the data) that help scientist to differentiate between normal and abnormal samples [8]. In a typical dataset built upon mass spectrometry, we have values of the abundance of a specific mass-to-charge (m/z) interval, therefore, we have as much features as intervals measured with the spectrometer, as resolution increases, so does the noise and the presence of irrelevant intervals in the data; nonetheless, the usage of mass spectrometry has greatly accelerated the discovery of biomarkers [9], although some studies have reported that very few biomarkers appear when a new set of data is used [10], which leads to the conclusion that there is some sort of overfitting and still a long road to go in the improvement of feature selection for biomarker discovery.

Due to the nature of mass spectrometry, a reduction in the dimensionality of the data is a very important processing step before building models, and a long number of techniques have been proposed since 2000 [11], [12]. The two main approaches to achieve a reduction of data dimensionality are [13], [14]: 1) *Feature extraction*, where the key idea is to transform the high dimensional features into a whole new space of lower dimensionality, e.g. Principal Component Analysis (PCA) [15], and 2) *Feature selection*, which main goal is to find the smallest number of features that still describes the data in a reasonable way [16]. In this work we focus in feature selection, these algorithms are divided in three groups [17], [18]: 1) *Filters*: perform an analysis of the data without the need of any external classifier to perform the selection. 2) *Wrappers*: using a classifier as evaluator, the selection is guided by the performance of the classifier. 3) *Embedded*: which can be seen as a combination of a filter approach and a second step of classifier performance analysis.

Our work is mainly focused on the filter approach, in a previous work [19], we presented the Heat Map Based Feature Selection (HmbFS) algorithm and performed a comparison with other filter algorithms such as Chi2 [20], Fcbf [21] and Relief-F [22]. One of the potential weakness of HmbFS is that the selection threshold although very intuitive, it can be dataset dependent, in this work we present an automatic approach to determine the best threshold for a dataset through an efficient embedding approach, we discuss the overall design of HmbFS as well as how the automatic selection of threshold

is applied and perform experimental comparison to find out how our proposed search methodology improves over the original HmbFS with default threshold and vs the Recursive Feature Elimination (RFE) algorithm as well.

The remaining of this paper is structured as follows: In Section II, we review some of the related works. In Section III, we analyze the original HmbFS algorithm and how the new embedding architecture helps to find the optimal threshold. In Section IV, the testing protocol is discussed and we report our results. In Section V, we review our conclusions and future work. Finally the list of references for further review are provided.

II. Related Works

Since 2000 [11], [12], we have seen the development of new techniques to tackle the inherent increase in data dimensionality, for bioinformatics, and specially for mass spectrometry data, Support Vector Machines (SVM), has been successfully employed not only as a classification method but as a feature selection as well with prominent results.

In 2009, Liu [23] showed a proposal to use wavelet feature extraction that in conjunction with SVM could be used for biomarker detection and therefore improve disease classification. Although this particular work focus more in feature extraction than feature selection, it is a relevant work because it presents early works that confirm that the usage of SVM its a viable option for these problems.

In 2010, Abeel et al. [24] presented an approach that uses multiple runs of SVM in combination with Recursive Feature Elimination (RFE), the idea behind this work is to use multiple subsamples of the data to create a set of candidates features, later the results are assembled and a final set of features is reported. In this work we perform a direct comparison with RFE, but since it is computationally demanding, we did not split the data as this work proposes.

In 2011, Kim et al. [25] presented a work which tackles the common issue of univariate analysis where no interaction between features is considered; to handle this, each feature is evaluated against a series of significance tests, then the features that managed to successfully pass the tests are considered for further interaction. In their experiments they reported an improvement over RFE with SVM, but since the number of features to be selected need to be set in advance it may be a drawback in most datasets.

In 2014, Gonzalez and Belanche [26] proposed an algorithm called eTAFS that combines a simulated annealing and incremental joint entropy to select subsets of features. According to the authors, the proposed combination of techniques is more accurate than competing methods while being fast, effective and requires no critical parameters to tune. Although we do not have access for this algorithm for direct comparison, we have included similar datasets to have some degree of comparison.

In 2016, Lei et al. [27] presented a very interesting approach focused in high-dimensional data, where the main idea was to develop a model-free selection mechanism via the a gradient function, selecting those features with a gradient substantially non-zero. However, the real life examples evaluated by authors suggests their approach is more focus on number of instances than number of features.

In summary, we can see that the problem of dimensionality reduction has not been solved and it is still in research interest, multiple proposals have reached successful performance, but there is still many open problems to be solved, one of those particular problems is the parameter tuning stage, in some scenarios, a good algorithm can lead to disastrous results if executed with the wrong parameters, although HmbFS does not require critical parameter tuning, it is important to ensure maximum possible performance with fewer user interaction.

In the following section we discuss particular details about HmbFS and our proposed search scheme for optimal threshold.

III. HmbFS and Threshold Search

In order to build a model in a supervised approach, we use data in form of a dataset which includes instances and features, these features are used by a classifier to fit the training data and make further predictions for unknown instances to discriminate among classes (e.g., normal vs cancer). Since the features are the input to the classifier, it is obvious that the classifier performance is dependent on the features quality. In theory, Yu [28] proposed that the more features we have, the more discriminative power among classes we achieve, this statement is certainly true, as long as all those features are good features, this is exactly where the problem lies, not all features are good, especially in high dimensional scenarios.

A potential ambiguity issue arises regarding what a "*good feature*" is, we use John et al. [29]'s proposal to identify three types of features as strongly relevant, relevant and irrelevant, the formal representation is shown below:

Let F be the full set of features, F_i a particular feature and $S_i = F - \{F_i\}$ a subset of features where F_i has been removed, given this we have three groups as follows:

A feature F_i is strongly relevant iff:

$$Pr(C \mid F_i, S_i) \neq Pr(C \mid S_i) \tag{1}$$

A feature F_i is weakly relevant iff:

$$\begin{gathered} Pr(C \mid F_i, S_i) = Pr(C \mid S_i) \text{ and } \exists S_i' \subset S_i \\ \text{such as } Pr(C \mid F_i, S_i') \neq Pr(C \mid S_i') \end{gathered} \tag{2}$$

A feature F_i is irrelevant iff:

$$\forall S_i' \subset S_i \quad Pr(C \mid F_i, S_i') = Pr(C \mid S_i') \tag{3}$$

Given these definitions, we now can define HmbFS main objective which is to find as much irrelevant features as possible in order to reduce the dimensionality of the data in the seek for optimal classifier performance. Our Heat Map Based Feature Selection (HmbFS) algorithm was presented in detail for a previous work [19], in this section we provide a summary about its design.

In order to achieve feature selection, HmbFS uses a two stage process: a) compression and b) selection.

A. Compression Stage

This stage is responsible of building a compressed version of the original dataset that is used for the actual feature selection stage, the key idea is to group continuous features (e.g., feature 1, feature 2 and feature 3) to build a RGB color pattern, the interaction between features produces different colors that are used to build a full heat map that represents a generalized version of the dataset. Since this process is scale sensitive, all features are first scaled to a 0-255 interval, using the Scikit-learn [30] library this can be achieved as follows:

```
#HmbFS 0-255 Feature Normalization
mms = preprocessing.MinMaxScaler(
  feature_range=(0, 255))
new_feats = mms.fit_transform(
            original_feats,
            dataset_classes)
```

Once each feature has been normalized, they can be used to build colors for the heatmap, e.g., features values $F_1 = 242$, $F_2 = 20$ and $F_3 = 35$ will create a *bright red* color (in RGB format), while the values $F_1 = 12$, $F_2 = 5$ and $F_3 = 55$ would create a very *dark blue*. In order to build a generalized representation, the heatmap is not built with the raw generated colors, but with a quantized version, reduced to 16 basic colors. This can be easily achieved by a distance search to find the basic color that most likely represents the real color. The following code assigns *red* and *black* for the feature-values examples we provided above.

```
#Color quantization, from real to 16 colors
 def getBaseColor(self, R, G, B):
  distance=[768]*16
  for x in range(0,16):
   bC = self.baseColors[x]
   distance[x]= ( abs(bC[0]-R) + abs(bC[1]-G) +
                  abs(bC[2]-B) )
  return distance.index(min(distance))
```

Once the color quantization is completed, the new compressed dataset is ready for feature selection analysis which is described in the next step.

B. Feature Selection Stage

Since the data is already grouped as the new compressed dataset has one third ($\frac{1}{3}$) of the original features, a multivariate analysis is inherent even when only single features (as a single feature now is represented by three of the original features) are analyzed. The main idea is that different classes must be represented by different color patterns, and the probability of a given color pattern must be greater for some class than another in order to be selected as useful. The resultant selection formula is presented below:

$$\begin{gathered}\text{Useful } F_i \text{ iff:} \\ \exists\{C_j, C_k\} \subset C \mid \\ [Pr(C_j \mid Mo(C_j \mid F_i))] > [Pr(C_k \mid Mo(C_j \mid F_i)) * T_h]\end{gathered} \tag{4}$$

In order to mark a feature as useful, we first find the *mode* (Mo), i.e., the most common color that a particular feature F_i exhibits for a given class C_j, then we compare if this color pattern belongs more likely to the class C_j than for a class C_k. In order to decide if the difference is significant, we use a threshold Th (default 1.5) that prevents selection of features with minimal difference. The following code shows the selection process:

```
#Search useful features
for Cj in range(0,number_of_classes):
 for Ck in range(0,number_of_classes):
  if Cj != Ck:
   if ( (max(data[Cj])/sum(data[Cj]) ) >
      ( Th * (data[Ck][data[Cj].argmax()]
        / sum(data[Ck]))) ):
```

When the if-condition is met, that feature is marked as useful, and the process continues until all features have been evaluated. After the process is completed, the selected features are mapped to the original feature space, e.g., if features 3, 8 and 56 are selected, that means that in the original space, the selected features will be 7,8,9 (because compressed F_3 = F_7, F_8 and F_9), 22,23,24 and 166,167,168.

C. Finding Th Automatically

It is clear that the threshold Th plays a very important role in the selection process, although very intuitive to tune (higher Th = more aggressive reduction), it is still the user responsibility to select a proper value, in order to fix this inconvenient, we propose a sequential stratified-cross validated search, where different threshold are analyzed with automated cross validation in a similar fashion as RFE does, with the main difference that with HmbFS, the steps in selection are not arbitrary reductions as it happens with RFEcv, where we are required to select a number of features to reduce in each step, e.g., we could try reducing 10 features each iteration, or 15, or 50, or 100, while in HmbFScv the steps are more intuitive, e.g., a $Th = 3.0$ is twice as rigorous in the selection than a $Th = 1.5$, however, if the features are good enough, changing the threshold would not lead to different selections which helps to provide more stable results. Fig. 1 shows the proposed architecture.

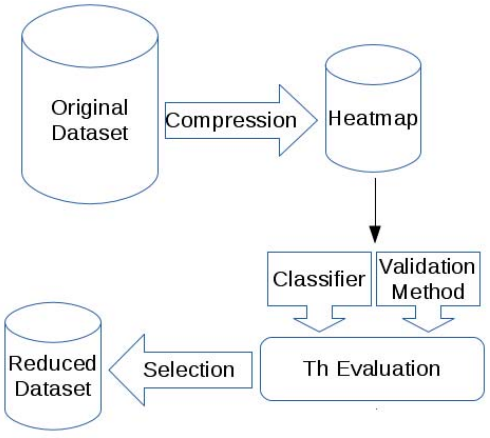

Fig. 1 Proposed architecture for optimal threshold selection

In order to find an optimal threshold Th, we included two key components in the process, a classifier that its responsible of building models with the subset of features that HmbFS

selected, and a validation method that tests the built models to ensure that the selected threshold is the optimal for the dataset. Other techniques such as RFE uses an arbitrary reduction step, such as 10% of features reduced each step, this greatly differs from HmbFS because the number of features is never required. In the following section we present some of the validation tests we performed to see how the proposed methodology improves over the default HmbFS with a fixed threshold.

IV. Tests and Results

In order to test our proposal, we have selected datasets that covers multiple types of cancer, specifically, we have selected a dataset from Alon [31] for colon cancer, Chowdary [32] for breast cancer, Gravier [33] for central nervous system tumors and, finally a dataset from Tian [34] for myeloma classification. All our experiments can be replicated using our provided code and data from a Github account[1].

We have performed a classification experiment using a stratified 10-fold approach, the idea behind stratified splits is to penalize the algorithms if results are biased towards the majority class. We evaluated an online classifier: Passive Aggressive (PA) [35], a classifier with built-in feature selection: Random Forest (RF) [36], and a classifier commonly used in related literature: Support Vector Machines (SVM) with linear kernel [37]. The feature selection stage was accomplished by HmbFS (the original version, with $Th = 1.5$), HmbFS with automatic threshold selection and cross-validation (this work, HmbFScv) and comparison with a popular approach Recursive Feature Elimination with cross-validation (RFEcv), the cross-validation stage was carried out by Logistic Regression, which is not included in the evaluation results to avoid overfitting, every classifier was executed without feature selection, and later with the 3 different approaches to evaluate the usefulness of each approach, the mean accuracy for every experiment is reported in Tables I-IV.

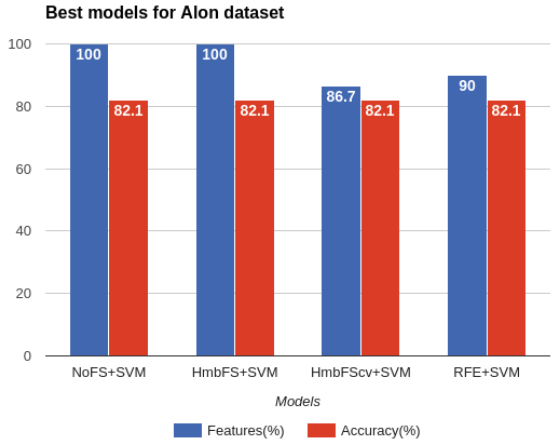

Fig. 2 Comparison of best models for the Alon dataset

The Alon dataset, being the lowest dimensionality of the datasets reviewed presented a great challenge for the feature selection algorithms. For instance our original proposal HmbFS was unable to remove any feature, leaving all the original 2,000 features and therefore producing the exact same results as if no feature selection (No FS) would have been employed; however, with the modified version (HmbFScv), it was possible to reduce over 10% of the features, not only without loss of precision, but with a huge improvement in stability, as the Random Forest (RF) performance was increased from 74.0% to 81.9%. In the case of RFEcv, the reduction of features is very similar to HmbFScv but the improvement is less stable as the RF performance did not improve as much. However, as an overall overview as it can be seen in Fig. 2, for the Alon dataset, feature selection did not improve the best possible score, which was achieved by Support Vector Machines (SVM) without any feature selection, however, it is worth noting that a reduced set of features helps to understand the data as an added benefit, even if no classification performance is achieved.

TABLE I
Alon Dataset Accuracy(%)

Methods	No FS	HmbFS	HmbFScv	RFEcv
Features	2,000	2,000	1,733	1,800
PA	77.1	77.1	77.1	77.1
RF	74.0	74.0	81.9	74.8
SVM	82.1	82.1	82.1	82.1
Average	77.7	77.7	80.4	78.0

TABLE II
Gravier Dataset Accuracy(%)

Methods	No FS	HmbFS	HmbFScv	RFEcv
Features	2,905	171	171	1
PA	70.8	73.8	73.8	75.0
RF	68.9	70.9	70.9	63.6
SVM	75.2	73.2	73.2	74.5
Average	71.6	72.6	72.6	71.0

[1]https://github.com/nxgtr/ICMLPR2016

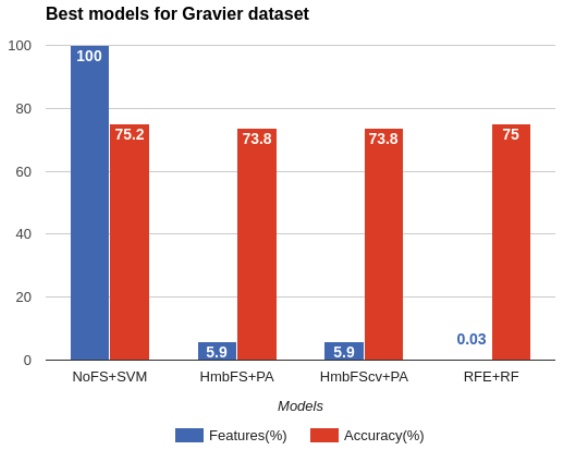

Fig. 3 Comparison of best models for the Gravier dataset

Although the Gravier dataset presents a similar dimensionality as Alon, the results are indeed data-dependent as in this case, we got very big reductions, HmbFS in its original fixed threshold version or the automatic approach, both converged to the same selection, 171 features (over 90% reduction) even though RFEcv managed to find a golden feature capable of getting an average accuracy of 71.0% only 0.6% below the NoFS benchmark. However, in terms of better selection stability and classification performance HmbFS achieved the highest averaged accuracy achieving improvements up to 3% when using the online classifier.

TABLE III
CHOWDARY DATASET ACCURACY(%)

Methods	No FS	HmbFS	HmbFScv	RFEcv
Features	22,283	234	57	2222
PA	94.3	97.3	91.3	94.3
RF	88.5	97.2	94.3	97.1
SVM	95.3	96.3	97.3	97.3
Average	92.7	96.9	94.3	96.2

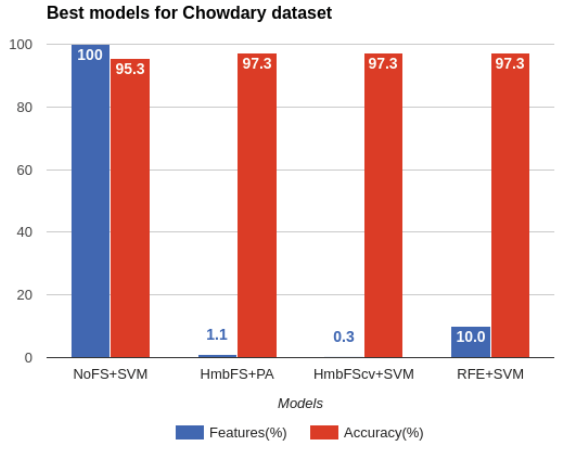

Fig. 4 Comparison of best models for the Chowdary dataset

The Chowdary dataset presented two main challenges, first its high dimensionality suggest a lot of potential noisy features, second, its relative high performance classification without the need of feature selection, using SVM it was possible to achieve 95.3% accuracy even with all the features. The RFEcv achieved at 90% reduction in features, with only 2,222 it achieved improvements as big as 8.6% when paired with RF, however using HmbFS we can notice that there is still a lot of noise left in the dataset, as with only 1.05% of features, it got the highest average at 96.9% achieving also the highest overall model performance of 97.3% with the simple PA. On the other hand, HmbFScv produced an outstanding reduction, as 99.74% of features were removed, still, it was possible to match the best model performance with SVM, as well as improve overall performance over the NoFS benchmark.

TABLE IV
TIAN DATASET ACCURACY(%)

Methods	No FS	HmbFS	HmbFScv	RFEcv
Features	12,625	3,609	2,931	1
PA	64.4	71.1	68.6	60.5
RF	75.9	78.7	79.8	68.3
SVM	72.6	77.6	77.0	50.1
Average	71.0	75.8	75.1	59.6

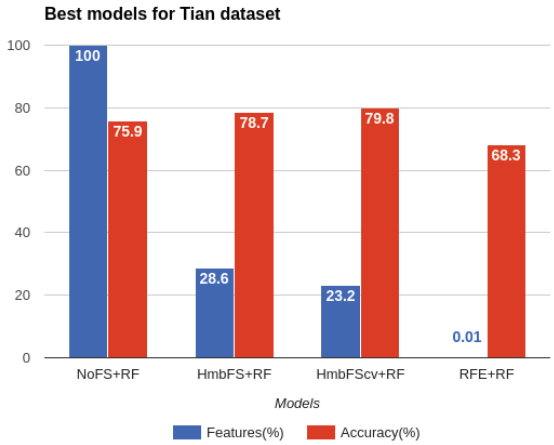

Fig. 5 Comparison of best models for the Tian dataset

The Tian dataset with high dimensionality allowed great reductions to be performed, starting with a reference benchmark of 75.9% accuracy with RF and no feature selection, it presented a considerable challenge, however HmbFS managed to outperformed that benchmark for every algorithm reaching a maximum of 77.5% with SVM with less than 30% of the original features, following the same path, with the improved HmbFScv it was possible to reduce even more to around 23% of the original features and producing an even better model at 79.7% with RF. In the case of RFEcv, the over-reduce effect seems to be recurrent behavior in very noisy datasets, in some cases it produce very good results, but for the Tian dataset it impacted too much to the classifiers, getting as low as 50.1% in the case of SVM.

We completed a total of 48 experiments to test HmbFS and HmbFScv (automatic Th selection), each dataset was tested under 12 different setups with combinations of feature and learning algorithms. Among the results, we can see that

as the dimension increases, so does the potential noise, the Alon dataset with 2,000 features was very difficult to reduce, however, Chowdary dataset with over 22,000 lead us to huge reductions overall, for instance RFEcv reduced the data to only 10% and still improved over the 22,283 of the original space, and the other impressive reduction goes for HmbFScv with a reduction to only 57 features. In terms of predictive power improvement, HmbFS and HmbFScv produced very competitive results vs RFEcv, for Alon dataset although best model remained at 82.1%, the smaller subset of features goes for HmbFScv. For Gravier dataset, overall best average goes for HmbFS/HmbFScv although notable best model is achieved with RFEcv and PA. For Chowdary dataset HmbFS and RFEcv achieved similar performance, with a slight advantage for HmbFS at much lower features, and HmbFScv pushed the reduction even more while still maintaining the best overall model at 97.3%. Finally, for the Tian dataset, HmbFScv got the most overall model at 79.8% while RFEcv exhibit issues with the selection.

V. Conclusions and Future Work

In this paper we have presented HmbFScv, an automatic search to select the optimal threshold for the original HmbFS algorithm, the results have been satisfactory although not decisive. Using automatic threshold selection it was possible to reduce even more the feature space, sometimes by a huge margin as happened with Chowdary dataset, in all cases the new approach selected less or at least the same amount of features as the original proposal. There seems to be a special case for RFEcv for noisy datasets where it over-reduce it as it happened with Gravier and Tian datasets, this behavior is worth additional research.

The proposed HmbFScv modification is competitive with the original HmbFS algorithm achieving similar classification performance and small subset of features as well, however this gain is currently achieved by a more computation expensive process, if the dataset is too big, a fixed threshold selection may be more suitable and still provide competitive results.

As part of future work it can be divided in two areas:

A. Improved Experimentation

We need to collect more datasets and include more scenarios to test HmbFS, we are currently gathering datasets as well as including more feature selection algorithms, more classifiers to benchmark results, and the addition of multiple metric analysis that can help understand better the outcomes than the simpler accuracy measure

B. Visualization

Since HmbFS works by building a heatmap representation of the data, it is possible to generate datasets visual images that are mapped to the original space, this can help in the seek of regions of interest, not only from the feature point of view, but for instance selection as well.

References

[1] J. Deng, A. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 785–792.

[2] S.-Y. Kung and M.-W. Mak, *Feature Selection for Genomic and Proteomic Data Mining*. John Wiley & Sons, Inc., 2008, pp. 1–45. (Online). Available: http://dx.doi.org/10.1002/9780470397428.ch1

[3] R. "Bellman, *"Dynamic Programming"*, "1" ed. "Princeton, NJ, USA": "Princeton University Press", "1957".

[4] A. Y. Ng, "On feature selection: Learning with exponentially many irrelevant features as training examples," in *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann, 1998, pp. 404–412.

[5] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003. (Online). Available: http://dl.acm.org/citation.cfm?id=944919.944968

[6] J. Yu and X.-W. Chen, "Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data," *Bioinformatics*, vol. 21, no. 1, pp. 487–494, Jan. 2005. (Online). Available: http://dx.doi.org/10.1093/bioinformatics/bti1030

[7] S. Datta and L. M. DePadilla, "Feature selection and machine learning with mass spectrometry data for distinguishing cancer and non-cancer samples," *Statistical Methodology*, vol. 3, no. 1, pp. 79 – 92, 2006, bioinformatics. (Online). Available: http://www.sciencedirect.com/science/article/pii/S157231270500064X

[8] P. R. Srinivas, M. Verma, Y. Zhao, and S. Srivastava, "Proteomics for cancer biomarker discovery," *Clinical Chemistry*, vol. 48, no. 8, pp. 1160–1169, 2002. (Online). Available: http://www.clinchem.org/content/48/8/1160.abstract

[9] M. D. I. C. W. E. C. L. H. S. O. T. E. R. Kuschner, Karl W., "A bayesian network approach to feature selection in mass spectrometry data," *BMC Bioinformatics*, vol. 11, 2010. (Online). Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3098056/

[10] D. I. Malyarenko, W. E. Cooke, B.-L. Adam, G. Malik, H. Chen, E. R. Tracy, M. W. Trosset, M. Sasinowski, O. J. Semmes, and D. M. Manos, "Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques," *Clinical Chemistry*, vol. 51, no. 1, pp. 65–74, 2005. (Online). Available: http://www.clinchem.org/content/51/1/65.abstract

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, 2002. (Online). Available: http://dx.doi.org/10.1023/A:1012487302797

[12] O. Chapelle, S. Keerthi, O. Chapelle, and S. Keerthi, "Multi-class feature selection with support vector machines," in *Proceedings of the American Statistical Association*, 2008.

[13] B. Dittmann and S. Nitz, "Strategies for the development of reliable qa/qc methods when working with mass spectrometry-based chemosensory systems," *Sensors and Actuators B: Chemical*, vol. 69, no. 3, pp. 253 – 257, 2000, proceedings of the International Symposium on Electronic Noses. (Online). Available: http://www.sciencedirect.com/science/article/pii/S0925400500005049

[14] U. Depczynski, V. Frost, and K. Molt, "Genetic algorithms applied to the selection of factors in principal component regression," *Analytica Chimica Acta*, vol. 420, no. 2, pp. 217 – 227, 2000. (Online). Available: http://www.sciencedirect.com/science/article/pii/S000326700000893X

[15] M. Suganthy and P. Ramamoorthy, "Principal component analysis based feature extraction, morphological edge detection and localization for fast iris recognition."

[16] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 14, pp. 131 – 156, 1997. (Online). Available: http://www.sciencedirect.com/science/article/pii/S1088467X97000085

[17] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 12, pp. 245 – 271, 1997. (Online). Available: http://www.sciencedirect.com/science/article/pii/S0004370297000635

[18] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 74–81. (Online). Available: http://dl.acm.org/citation.cfm?id=645530.658297

[19] C. Huertas and R. Juárez-Ramírez, "Heat map based feature selection: A case study for ovarian cancer," in *Applications of Evolutionary Computation - 18th European Conference, EvoApplications 2015, Copenhagen, Denmark, April 8-10, 2015, Proceedings*, 2015, pp. 3–13.

[20] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *In Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, 1995, pp. 388–391.
[21] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," 2003, pp. 856–863.
[22] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning*, ser. ML92. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992, pp. 249–256. (Online). Available: http://dl.acm.org/citation.cfm?id=141975.142034
[23] Y. Liu, "Feature extraction and dimensionality reduction for mass spectrometry data," *Comput. Biol. Med.*, vol. 39, no. 9, pp. 818–823, Sep 2009.
[24] T. Abeel, T. Helleputte, Y. Van 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, Feb. 2010. (Online). Available: http://dx.doi.org/10.1093/bioinformatics/btp630
[25] H. Kim, J. Watkinson, and D. Anastassiou, "Biomarker discovery using statistically significant gene sets." *Journal of Computational Biology*, vol. 18, no. 10, pp. 1329–1338, 2011.
[26] F. Gonzlez and L. A. B. Muoz, "Feature selection for microarray gene expression data using simulated annealing guided by the multivariate joint entropy," *CoRR*, vol. abs/1302.1733, 2013.
[27] L. Yang, S. Lv, and J. Wang, "Model-free variable selection in reproducing kernel hilbert space," *Journal of Machine Learning Research*, vol. 17, no. 82, pp. 1–24, 2016. (Online). Available: http://jmlr.org/papers/v17/15-390.html
[28] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, dec 2004. (Online). Available: http://dl.acm.org/citation.cfm?id=1005332.1044700
[29] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *MACHINE LEARNING: PROCEEDINGS OF THE ELEVENTH INTERNATIONAL*. Morgan Kaufmann, 1994, pp. 121–129.
[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[31] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, jun 1999.
[32] D. Chowdary, J. Lathrop, J. Skelton, K. Curtin, T. Briggs, Y. Zhang, J. Yu, Y. Wang, and A. Mazumder, "Prognostic gene expression signatures can be measured in tissues collected in rnalater preservative," *The Journal of Molecular Diagnostics*, vol. 8, no. 1, pp. 31–39, feb 2006.
[33] Gravier, Eleonore, G. Pierron, A. Vincent-Salomon, N. gruel, V. Raynal, A. Savignoni, Y. De Rycke, J.-Y. Pierga, C. Lucchesi, F. Reyal, A. Fourquet, S. Roman-Roman, F. Radvanyi, X. Sastre-Garau, B. Asselain, and O. Delattre, "A prognostic DNA signature for T1T2 node-negative breast cancer patients." *Genes, Chromosomes and Cancer*, vol. 49, no. 12, pp. 1125–1125, Sep. 2010.
[34] E. Tian, F. Zhan, R. Walker, E. Rasmussen, Y. Ma, B. Barlogie, and J. D. Shaughnessy, Jr., "The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma," *New England Journal of Medicine*, vol. 349, no. 26, pp. 2483–2494, dec 2003.
[35] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 551–585, dec 2006. (Online). Available: http://dl.acm.org/citation.cfm?id=1248547.1248566
[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, oct 2001. (Online). Available: http://dx.doi.org/10.1023/A:1010933404324
[37] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, jun 2008. (Online). Available: http://dl.acm.org/citation.cfm?id=1390681.1442794

M.S. Carlos Huertas received his bachelor and masters degree from the Autonomous University of Baja California, Mexico, in 2010 and 2013 respectively, and he is currently working towards the PhD degree. He has developed multiple machine learning systems mainly for classification and regression problems in industry, besides is creator of the Heat Map Based algorithm for dimensionality reduction, currently he is working towards improving classification performance in bioinformatics, specifically with mass spectrometry data. His research interests include feature selection, model selection, model ensemble and deep learning.

Dr. Reyes Juarez Ramirez has a bachelor degree in Computer Engineering, a master degree in Computer Science and a PhD in Computer Science (2008). Since 2002 he is a Professor-Researcher at Universidad Autonoma de Baja California, Campus Tijuana, Mexico. He is the General Chair of the CONISOFT Conference from 2010 to 2016 and the Technical Coordinator of the Mexican Thematic Network of Software Engineering from 2014 to 2016. He is Technical Coordinator of Accreditation of Undergraduate Engineering Programs and also is a member of the Computer Engineering Technical Committee at CACEI Council. Also he is IEEE Senior Member since 2014. His research areas are: Software Engineering, Human-Computer Interaction, Knowledge Engineering, and Engineering Education and Training. Recently he is adopting intelligent methods and techniques to support his research areas of interest.