Computers in Biology and Medicine 79 (2016) 68–79

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

CAFÉ-Map: Context Aware Feature Mapping for mining high dimensional biomedical data

Fayyaz ul Amir Afsar Minhas[a,*], Amina Asif[a], Muhammad Arif[b,*]

[a] Biomedical Informatics Research Laboratory, Department of Computer & Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Islamabad, Pakistan

[b] Department of Computer Science, Umm Al-Qura University, Makkah, Saudi Arabia

ABSTRACT

Feature selection and ranking is of great importance in the analysis of biomedical data. In addition to reducing the number of features used in classification or other machine learning tasks, it allows us to extract meaningful biological and medical information from a machine learning model. Most existing approaches in this domain do not directly model the fact that the relative importance of features can be different in different regions of the feature space. In this work, we present a context aware feature ranking algorithm called CAFÉ-Map. CAFÉ-Map is a locally linear feature ranking framework that allows recognition of important features in any given region of the feature space or for any individual example. This allows for simultaneous classification and feature ranking in an interpretable manner. We have benchmarked CAFÉ-Map on a number of toy and real world biomedical data sets. Our comparative study with a number of published methods shows that CAFÉ-Map achieves better accuracies on these data sets. The top ranking features obtained through CAFÉ-Map in a gene profiling study correlate very well with the importance of different genes reported in the literature. Furthermore, CAFÉ-Map provides a more in-depth analysis of feature ranking at the level of individual examples.

Availability: CAFÉ-Map Python code is available at:
http://faculty.pieas.edu.pk/fayyaz/software.html#cafemap .

The CAFÉ-Map package supports parallelization and sparse data and provides example scripts for classification. This code can be used to reconstruct the results given in this paper.

1. Introduction

Biomedical devices and experiments generate large amount of high dimensional data which needs proper analysis for mining relevant information. Examples include gene expression profiling [1–4], mass spectrometry [5–8], medical images from ultrasound, magnetic resonance or computerized Tomography [9,10], etc. In this domain, the objective of data analysis is to identify diagnostically or biologically significant features such as genes, spectral components or regions of interest in images. This information can be obtained as part of a machine learning or data mining system through feature selection or ranking techniques in conjunction with supervised classification [11]. Classification allows computational prediction of medical disorders or identification of biologically interesting phenomenon based on the features identified during the feature selection or ranking process [12,13].

Before further discussion on feature selection techniques, it is important to point out the challenges and difficulties in applying feature selection and classification approaches in the analysis of biomedical data. Biomedical data is typically "tall and thin", i.e., very high dimensional but with only a small number of samples [11,14,15]. For example, a single gene profiling experiment can easily have tens of thousands of genes whose expression is characterized for hundreds of subjects or cell types [16–21]. Typically, a large number of features or attributes of the data can be irrelevant for prediction. This issue of high dimensionality is exacerbated by the presence of only a small number of data samples. Obtaining medical or biological samples is typically labor intensive, time consuming and expensive. In addition to feature selection, it is desirable in supervised classification of biomedical data that the classification scheme generates some information about the reasons due to which an example has been classified in a certain way. This information can be utilized to gain a deeper understanding of the mechanics of a disease or biological phenomenon. For most classification approaches such as Support Vector Machines (SVMs) [22], Neural Networks, ensemble systems, etc., it is not straightforward to extract this information due to their black or grey box nature and this problem

* Corresponding author.
E-mail addresses: afsar@pieas.edu.pk (F.u.A.A. Minhas), a.asif.shah01@gmail.com (A. Asif), mahamid@uqu.edu.sa (M. Arif).

http://dx.doi.org/10.1016/j.compbiomed.2016.10.006
Received 9 May 2016; Received in revised form 5 October 2016; Accepted 10 October 2016

has received significant attention in recent research [23–27]. This is particularly true in classification problems in which the classification boundary is not linear.

Feature ranking or selection techniques can be used to identify features important for a classification problem. For references, the interested reader is referred to a number of excellent reviews on feature selection and ranking techniques [11,28–35]. However, existing feature selection techniques do not provide ranking or selection of features *for individual instances*. Typically, these approaches identify features which are important at the *global* level. To the best knowledge of the authors, no previously published feature selection or ranking technique has the ability to identify important features in individual training and test examples in a context aware or local manner. We illustrate the motivation behind this idea with an illustrative example in the Figure below (Fig. 1).

For this simple L-shaped synthetic data set comprising of two classes, any existing feature selection technique will rank both features as important. However, it is interesting to notice that the relative importance of the two features for correct classification is dependent upon the local context as well. For instance, along the vertical part of the classification boundary in this figure, feature x_1 is more important in comparison to feature x_2. Similarly, along the horizontal component of the classification boundary, feature x_2 is responsible for discrimination. Though most existing feature selection techniques *implicitly* consider the fact that the role of a given feature in determining the classification boundary varies over different parts of the feature space, their output cannot be interpreted in such a manner. *Context aware* feature selection or ranking can lead to a more detailed analysis of the roles of different features in different parts of the feature space as well as in identifying what features are relatively more important in comparison to others for individual examples.

In this work, we present a contextual feature ranking and classification algorithm called *Context Aware Feature Mapping* or CAFÉ-Map. The novelty of CAFÉ-Map comes from its unique ability to quantify the relative importance of features in any region of the feature space. This is achieved by associating a local or context aware weight vector with each classification example. The mathematical and algorithmic formulation of CAFÉ-Map leads to the selection of a minimal set of *local* or context dependent features for every individual example while ensuring high classification accuracy, computational efficiency and interpretability. CAFÉ-Map is strongly grounded in the theory of Structural Risk Minimization and is designed to handle the challenges of biomedical data discussed earlier [36].

CAFÉ-Map can be of great use in the biomedical domains where data analysts are interested in understanding individual classification instances. For instance, in microarray based classification, examples of the same class can have different sets of differentially expressed genes. Unlike global feature ranking or selection techniques, CAFÉ-Map can reveal the set of genes that are important for classification of individual examples. Note that, due to its context aware nature, CAFÉ-Map can rank genes differently for individual examples. Similar to global feature ranking, CAFÉ-Map can also identify a single set of features that are important for classification. This can be accomplished by a simple average of absolute values of contextual weight vectors of individual examples. CAFÉ-Map can also be used to group examples with similar differential expression profiles by clustering over top ranked components of their local weight vectors. This information can then form the basis for identifying causative or correlative relationships in examples of the same group, e.g., gender, age, race, disease progression, cell type, etc.

The rest of the paper is organized as follows: Section 2 describes the formulation of CAFÉ-Map and how it can be used to achieve a context dependent ranking of features in a given classification domain. Section 3 presents an empirical comparison of its classification performance with different existing feature selection techniques for a number of toy problems as well as widely used real world biomedical data sets. Section 4 presents the conclusions.

2. Materials and methods

2.1. Mathematical formulation

As discussed in the Introduction section, CAFÉ-Map can identify important features in individual training and test examples in a *context aware* or *local* manner. This is in contrast to existing classification and feature selection techniques which can only produce a global ranking of features. The fundamental idea behind CAFÉ-Map is to use a manifold encoding based locally linear classification function whose L1-norm is regularized as part of a structural risk functional by an efficient iterative optimization algorithm. In this section, we present the detailed formulation of CAFÉ-Map based on this principle. For this purpose, we assume a classification data set of N examples $\{(\boldsymbol{x}_i, y_i) | i = 1 \ldots N\}$ in which each example is represented by a d-dimensional feature vector $\boldsymbol{x}_i \in \Re^d$ and $y_i \in \{-1, +1\}$ indicates the label of that example. These labels are available for training examples only. For classification, CAFÉ-Map uses a discriminant function $f(\boldsymbol{x}) = \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{x}$ with a *context dependent* or *localized* weight vector function $\boldsymbol{w}(\boldsymbol{x})$: $\Re^d \rightarrow \Re^d$. Note that we have omitted the bias term from the discriminant function for simplicity. The learning problem is to calculate $\boldsymbol{w}(\boldsymbol{x})$ from training data to correctly predict the score of any test example $\boldsymbol{x}$. Classical classification approaches like Linear Support Vector Machine (SVMs) use a context independent or global weight vector $\boldsymbol{w}(\boldsymbol{x}) = \boldsymbol{w} \in \Re^d, \forall \boldsymbol{x}$ in their discriminant function. In linear SVMs, the magnitude of different components of the global weight vector, $|w_j|$, can be used to rank the importance of different features in the classification problem [22], [37]. However, these methods are limited to linearly separable classification problems and can only produce feature rankings at the global level. The use of non-linear kernels does allow non-linear classification boundaries but it makes feature ranking or interpretation very difficult. In contrast, the use of a local weight vector $\boldsymbol{w}(\boldsymbol{x})$ leads to a *locally linear* classifier which can solve classification problems with non-linear boundaries without using kernel functions or feature transformations [22], [38–40]. We propose and demonstrate that the magnitude of the components of $\boldsymbol{w}(\boldsymbol{x})$ can reveal context sensitive importance of different features. This is achieved in CAFÉ-Map by reducing the number of non-zero or large valued components of the local weight vector function through L1-norm regularization of its discriminant. Like the locally linear SVMs proposed by Ladicky and Torr, a locally linear representation of the CAFÉ-Map discriminant function is obtained through local encoding of data as discussed below [33].

2.1.1. Local encoding of data

Local codings for manifold learning represent an example $\boldsymbol{x}$ as a linear combination of K *a priori* chosen d-dimensional *anchors*

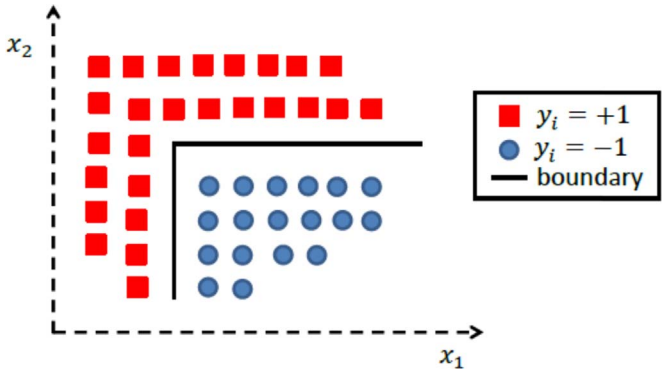

Fig. 1. Concept diagram of CAFÉ-Map. The importance of the features is different in different parts of the feature space in the L-shaped data set.

represented by the d × K matrix $\boldsymbol{V}=[\boldsymbol{v}_1\ \boldsymbol{v}_2\ \cdots\ \boldsymbol{v}_K]$. This can be written mathematically as:

$$\boldsymbol{x}\approx V\boldsymbol{\gamma}(\boldsymbol{x}) \tag{1}$$

In the above equation, $\boldsymbol{\gamma}(\boldsymbol{x})=[\gamma_1(\boldsymbol{x})\ \gamma_2(\boldsymbol{x})\ \ldots\ \gamma_K(\boldsymbol{x})]^T$ is the local coordinate representation of $\boldsymbol{x}$. The anchors are simply sampling points in the feature space. The set of anchors can be obtained by randomly selecting a subset of K examples in the given dataset or applying K-means clustering on it and using the K cluster centers as anchors. Conceptually, $\boldsymbol{\gamma}(\boldsymbol{x})$ is a description of $\boldsymbol{x}$ in terms of the feature representations of a small number of nearby anchors such that the re-projection error $\|\boldsymbol{x}-V\boldsymbol{\gamma}(\boldsymbol{x})\|$ is small. CAFÉ-Map uses Locality-constrained linear coding (LLC) to obtain $\boldsymbol{\gamma}(\boldsymbol{x})$ for all examples in the given data set [41]. For a description of other encoding techniques in the literature, the interested reader is referred to recent papers and reviews on the subject [15], [41–51]. LLC is algorithmically attractive due to its accuracy and the existence of an analytical solution for its underlying optimization problem. LLC produces an accurate and sparse mapping of a given example to local coordinates $\boldsymbol{\gamma}(\boldsymbol{x})$ by minimizing re-projection error and enforcing locality and regularization on $\boldsymbol{\gamma}(\boldsymbol{x})$. With $\boldsymbol{d}(\boldsymbol{x}_i)=[d(\boldsymbol{v}_1,\ \boldsymbol{x}_i)\ \ldots\ d(\boldsymbol{v}_K,\ \boldsymbol{x}_i)]^T$ defined as a vector of distances values $d(\bullet,\boldsymbol{x}_i)$ of a given example $\boldsymbol{x}_i$ from all the anchors in $\boldsymbol{V}$, the LLC constrained optimization problem aims to find $\boldsymbol{\gamma}(\boldsymbol{x}_i)$ such that $\sum_{k=1}^{K}\gamma_k(\boldsymbol{x}_i)=1$ for all examples. This can be written mathematically as follows:

$$\min_{\boldsymbol{\gamma}(\boldsymbol{x})}\sum_{i=1}^{N}\|\boldsymbol{x}_i-V\boldsymbol{\gamma}(\boldsymbol{x}_i)\|^2+\gamma\|\boldsymbol{d}(\boldsymbol{x}_i)\odot\boldsymbol{\gamma}(\boldsymbol{x}_i)\|^2$$

The first term of the LLC minimizes the re-projection error across all examples. The second term involves the norm of the element wise multiplication ($\odot$) of the distance and local coordinate vectors. This term, weighted by the hyper-parameter $\gamma>0$, enforces locality and sparsity by reducing the values of the components of $\boldsymbol{\gamma}(\boldsymbol{x}_i)$ corresponding to faraway anchors.

2.1.2. Local encoding of weight function

In the above section, we discussed how to represent a given example in terms of local coordinates. Now, we provide an approximation of the local weight function $\boldsymbol{w}(\boldsymbol{x})$ in terms of $\boldsymbol{\gamma}(\boldsymbol{x})$and the anchors. This approximation paves the way for formulating the optimization problem for CAFÉ-Map. If we assume each component of our weight vector function $\boldsymbol{w}(\boldsymbol{x})$ to be Lipschitz continuous, i.e., each of the d components of $\boldsymbol{w}(\boldsymbol{x})$ is bounded in how fast it can change with a change in $\boldsymbol{x}$, then, based on the properties of local codings, we can approximate $\boldsymbol{w}(\boldsymbol{x})$ as [33]:

$$\boldsymbol{w}(\boldsymbol{x})\approx W\boldsymbol{\gamma}(\boldsymbol{x}) \tag{2}$$

Here, $\boldsymbol{W}=[\boldsymbol{w}(\boldsymbol{v}_1)\ \boldsymbol{w}(\boldsymbol{v}_2)\ \cdots\ \boldsymbol{w}(\boldsymbol{v}_K)]$ is a $d\times K$ weight matrix. The k^{th} column of this matrix is obtained by applying the weight vector function $\boldsymbol{w}(\bullet)$ to anchor $\boldsymbol{v}_k$ in $\boldsymbol{V}$. With this approximation, we can rewrite the discriminant function of our classifier as $f(\boldsymbol{x})=\boldsymbol{w}(\boldsymbol{x})^T\boldsymbol{x}=(W\boldsymbol{\gamma}(\boldsymbol{x}))^T\boldsymbol{x}$. With this approximation, the CAFÉ-Map learning problem can be reformulated as finding a matrix $\boldsymbol{W}$ that produces correct scores for all training examples. A solution to this problem is obtained through Structural Risk Minimization as discussed in the next section.

2.1.3. Structural risk minimization based feature selection

Like Support Vector Machines (SVMs) and other large margin classifiers, CAFÉ-Map is also based on the principle of structural risk minimization (SRM) [36]. SRM states that, in order to provide good generalization, a classifier should reduce the empirical error on its prediction through a loss function while minimizing its complexity by regularization. In contrast to existing locally linear variants of support vector machines, CAFÉ-Map uses an L1-norm unsquared regularizer $\|\boldsymbol{W}\|_1=\sum_{i=1}^{d}\sum_{j=1}^{K}|\boldsymbol{W}_{ij}|$ to obtain a sparse local weight function. With an empirical loss term $L(X,\ Y;f)$ and a regularization parameter λ, the complete learning problem of CAFÉ-Map can be written as:

$$min_{\boldsymbol{W}}P(\boldsymbol{W})=\lambda\|\boldsymbol{W}\|_1+L(X,\ Y;f) \tag{3}$$

The empirical loss term $L(X,\ Y;f)=\sum_{i=1}^{N}l(f(\boldsymbol{x}_i),\ y_i)$ measures the error between the prediction from the classifier $f(\boldsymbol{x})$ and the desired target label y for all examples. CAFÉ-Map uses the logistic loss function $l(f(\boldsymbol{x}),y)=log(1+exp(-yf(\boldsymbol{x})))$ due to its continuous nature which aids in the optimization procedure discussed in the next section. A change of the loss function to a square loss or an ε-insensitive loss function can lead to the solution of a regression problem. Other extensions, such as ranking or multiple instance learning, are also possible by simply changing the loss function.

2.1.4. Optimization algorithm for CAFÉ-Map

The CAFÉ-Map optimization problem with L1-norm regularization is solved by an efficient stochastic coordinate descent algorithm proposed by Shalev-Shwartz and Tewari [52]. Reasons for choosing this optimization algorithm in CAFÉ-Map include its good convergence and run time characteristics, ease of implementation and parameter free nature.

This stochastic coordinate descent algorithm initializes W to a zero matrix. At each iteration of the algorithm, a coordinate (j, k) is picked uniformly at random from the weight matrix such that $j\in[1,\ d]$ and $k\in[1,\ K]$. Then the weight component $w_{jk}=W(j,\ k)$ is updated in a direction opposite to the partial derivative of the CAFÉ-Map objective function with respect to w_{jk} with a step size of $\frac{1}{\beta}$. Specifically, the update can be written as: $w_{jk}\leftarrow w_{jk}-\frac{1}{\beta}\left(\frac{\partial L(X,\ Y;W)}{\partial w_{jk}}+\lambda\frac{\partial\|W\|_1}{\partial w_{jk}}\right)$.

The partial derivative of the empirical loss term $L(X,\ Y;f)=\sum_{i=1}^{N}l(f(x_i),\ y_i)$ with respect to w_{jk} is given by $g_{jk}=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial l(z_i,\ y_i)}{\partial w_{jk}}$. The gradient of the discriminant function score $z_i=f(x_i)=(W\gamma(x_i))^T x_i$ with respect to w_{jk} is $\frac{\partial z_i}{\partial w_{jk}}=x_{ij}\gamma_k(x_i)$. Here, x_{ij} is the value of the feature j for example x_i and $\gamma_k(x_i)$ is its local coordinate corresponding to anchor k. Therefore, the gradient of the logistic loss function with respect to w_{jk} can be written as: $\frac{\partial l(z_i,\ y_i)}{\partial w_{jk}}=\frac{\partial\log(1+\exp(-y_i z_i))}{\partial w_{jk}}=-\frac{y_i x_{ij}\gamma_k(x_i)}{1+\exp(y_i z_i)}$. Thus, $g_{jk}=-\frac{1}{N}\sum_{i=1}^{N}\frac{y_i x_{ij}\gamma_k(x_i)}{1+\exp(y_i z_i)}$.

The partial derivative of the L1-regularization term $\lambda\|W\|_1$ with respect to w_{jk} is given by

$$\frac{\partial\lambda\|W\|_1}{\partial w_{jk}}=\frac{\partial\lambda|w_{jk}|}{\partial w_{jk}}=\begin{cases}\lambda & \text{if } w_{jk}>0\\ -\lambda & \text{if } w_{jk}<0\\ 0 & \text{otherwise}\end{cases}.$$

As a consequence, the step update $w_{jk}\leftarrow w_{jk}-\frac{1}{\beta}\left(\frac{\partial L(X,\ Y;W)}{\partial w_{jk}}+\frac{\partial\lambda\|W\|_1}{\partial w_{jk}}\right)$ can be written as:

$$w_{jk}\leftarrow\begin{cases}w_{jk}-\frac{1}{\beta}(g_{jk}+\lambda) & \text{if } w_{jk}-\frac{1}{\beta}(g_{jk}+\lambda)>0\\ w_{jk}-\frac{1}{\beta}(g_{jk}-\lambda) & \text{if } w_{jk}-\frac{1}{\beta}(g_{jk}-\lambda)<0\\ 0 & \text{otherwise}\end{cases}$$

Thus, $w_{jk}\leftarrow 0$ if $\left(w_{jk}-\frac{g_{jk}}{\beta}\right)\in\left[-\frac{\lambda}{\beta},\ \frac{\lambda}{\beta}\right]$. It is easy to notice that a large λ would forces a large number of components of the weight matrix to zero.

The parameter β in the coordinate descent algorithm is taken to be the upper bound on the second derivative of the logistic loss function. For normalized data in which all examples have unit norm, this parameter is set to $\frac{1}{4}$.

The complete CAFÉ-Map training algorithm with all the optimization steps is given below. This algorithm reduces the run time by requiring weight-update based changes to the function scores of examples instead of re-evaluating them every time.

The algorithm is run for a pre-specified number of epochs or

iterations T over the training data. Customization of this algorithm for use in CAFÉ-Map lead to a run time with an expectation upper bound of $\frac{NKd\beta\|W^*\|_2^2}{\varepsilon}$ to reach an ε-accurate solution, where $W^*=\text{argmin}_W P(W)$ is the optimal solution. This run time can be significantly reduced in case of sparse input data and sparse coding. The algorithm also provides a theoretical guarantee on the upper bound of the error between the weight matrix obtained after T iterations W(T) and W* that decreases hyperbolically with T. For proofs, the interested reader is referred to the paper by Shalev-Shwartz and Tewari [52].

Input:

Data Set: N training examples with associated labels $\{(\boldsymbol{x_i}, y_i)|i = 1\ldots N\}$

Parameter values: K, γ, λ

Output:

An optimal weight matrix $\boldsymbol{W}$ that can be used to obtain the local weight vector $\boldsymbol{w}(\boldsymbol{x})$ and discriminant score $f(\boldsymbol{x})$ for any $\boldsymbol{x}$

Algorithm Description

Select K Anchor points from the given data set

Obtain local coding $\boldsymbol{\gamma}(\boldsymbol{x})$ for every example $\boldsymbol{x}$ through LLC with parameter γ

//optimize Eq.(3) using the optimization algorithm as follows:

Let $\boldsymbol{W} = 0 \in \mathfrak{R}^{d\times K}$, $z = 0 \in \mathfrak{R}^N$

For t = 1, 2... until convergence

Sample $j \in [1, d]$ and $k \in [1, K]$ uniformly at random

Compute the derivate $g_{jk}=-\frac{1}{N}\sum_{i=1}^{N}\frac{y_i x_{ij}\gamma_k(\boldsymbol{x_i})}{1+exp(y_i z_i)}$

If $w_{jk}-\frac{1}{\beta}(g_{jk}+\lambda)>0$ then

$W_{jk} \leftarrow W_{jk} - \frac{1}{\beta}(g_{jk} + \lambda)$

else if $w_{jk}-\frac{1}{\beta}(g_{jk}-\lambda)< 0$ then

$W_{jk} \leftarrow W_{jk} - \frac{1}{\beta}(g_{jk} - \lambda)$

else

$w_{jk} \leftarrow 0$

End if

Let the change in w_{jk} be denoted by Δw_{jk}

If $\Delta w_{jk}\neq 0$, then For all examples $i = 1\ldots N$ for which $x_{ij}\neq 0$ and $\boldsymbol{\gamma_k}(\boldsymbol{x_i})\neq 0$

Update $z_i=z_i+\Delta w_{jk} x_{ij}\boldsymbol{\gamma_k}(\boldsymbol{x_i})$

2.1.5. Local feature ranking and interpretation of CAFÉ-Map

Once the weight matrix $\boldsymbol{W}$ has been obtained using the optimization algorithm described in the last section, we can then calculate the local weight vector $\boldsymbol{w}(\boldsymbol{x})=W\boldsymbol{\gamma}(\boldsymbol{x})$ and the discriminant function score $f(\boldsymbol{x})=\boldsymbol{w}(\boldsymbol{x})^T\boldsymbol{x}$ for any example $\boldsymbol{x}$. L1-norm regularization of $\boldsymbol{W}$ in CAFÉ-Map forces most components of $\boldsymbol{w}(\boldsymbol{x})$ to zero. As a consequence, the local weight vectors across examples can be used in different ways for feature ranking. Specifically, we can rank the importance of different features for an individual example $\boldsymbol{x}$ by using the absolute values of different components of the context dependent weight vector given by $|\boldsymbol{w}_j(\boldsymbol{x})|$, $j = 1\cdots d$. A global feature ranking can also be obtained by simply averaging the absolute values of local weights across examples, i.e., $|\boldsymbol{w}_j|=\frac{1}{N}\sum_{i=1}^{N}|\boldsymbol{w}_j(\boldsymbol{x_i})|$, $j = 1\cdots d$. If the data is normalized, then the norm of the local weight vector $\|\boldsymbol{w}(\boldsymbol{x})\|$ can also be used to rank the importance of different training examples for classification.

A deeper look at the scoring function of CAFÉ-Map $f(\boldsymbol{x})=\boldsymbol{w}(\boldsymbol{x})^T\boldsymbol{x}$ reveals that the score of an example is, in essence, the projection or correlation of the feature vector of that example with its local weight vector. For correct classification, we require $yf(\boldsymbol{x}) > 0$. As a consequence, the learning problem of CAFÉ-Map can be interpreted as follows: CAFÉ-Map finds a minimum L1-norm local weight vector $\boldsymbol{w}(\boldsymbol{x_i})$ such that the projection or correlation of $\boldsymbol{x_i}$ with the vector $y_i\boldsymbol{w}(\boldsymbol{x_i})$ is positive. Therefore, $y_i\boldsymbol{w}(\boldsymbol{x_i})$ can be visualized as a sparse or reduced variant of $\boldsymbol{x_i}$ containing only those components that are important for the given classification problem. For the special case in which $\boldsymbol{\gamma}(\boldsymbol{x})=\boldsymbol{x}$, the CAFÉ-Map formulation results in large margin locally linear discriminant analysis with its scoring function given by $f(\boldsymbol{x})=\boldsymbol{x}^T\boldsymbol{W}^T\boldsymbol{x}$ [53].

2.2. Handling class imbalance, bias and sparsity

The basic formulation presented earlier can be improved to handle imbalanced data classification problems. This can be achieved by introducing an example-specific weighting factor to the loss function that assigns greater importance to correct classification of the under-represented class. Specifically, we modify the loss term $L(X, Y;f)$ to $\sum_{i=1}^{N} c_i l(f(\boldsymbol{x_i}), y_i)$ by introducing user-specified factors $c_i>0$ for all examples $i = 1\cdots N$ such that $\sum_{i=1}^{N} c_i=1$. Class imbalance is adjusted by setting $c_i=\frac{1}{2N^+}$ for all N^+ positive examples and $c_i=\frac{1}{2N^-}$ for the N^- negative examples in the training set of $N = N^+ + N^-$ total examples. The gradient term is modified accordingly. These factors can also be used to introduced prior or domain knowledge in the classifier.

The formulation of CAFÉ-Map omits the bias term in the discriminant function for simplicity. The addition of the bias term simply requires an additional feature for all examples with a value of 1.0. The objective function and the associated gradient evaluations are accordingly updated.

The Python implementation of CAFÉ-Map is optimized for handling sparse data. It prevents unnecessary computation for sparse data by preventing gradient calculation and score updates if either x_{ij} or $\boldsymbol{\gamma_k}(\boldsymbol{x_i})$ is zero.

Table 1
Results of CAFE-Map for Different Real World Data Sets. The average number of non-zero local weights and the percentage of selected features (in parenthesis) obtained after CAFÉ-Map training is shown for each data set. Also shown is the associated AUC-ROC and Balance Accuracy value with the standard deviation given in parenthesis. The average run time for multiple cross-validation runs in seconds is also given for different data sets. For comparison, we also provide best value of accuracy obtained by existing techniques cited as references.

Dataset Details			**CAFÉ-Map Results**				**Comparison**	
Name	Samples	Features	Active Features	AUC-ROC	Mean Accuracy	Average Time (s)	Best reported Accuracy	References
Arcene	88 (P) 112(N)	10,000	55 (0.6%)	94	86 (1)	409	86 (3)	[54],[56–61]
Prostate	52 (T) 50 (N)	2,135	38 (1.8%)	98	96 (6)	354	96 (8)	[62–67]
Lymphoma	58 (D) 19 (F)	7,129	426 (6.0%)	94	98 (5)	451	95 (8)	[16],[62],[65],[67–70]
Breast	84 (L) 44 (N)	47,293	437 (0.9%)	88	92 (6)	497	89 (5)	[62]

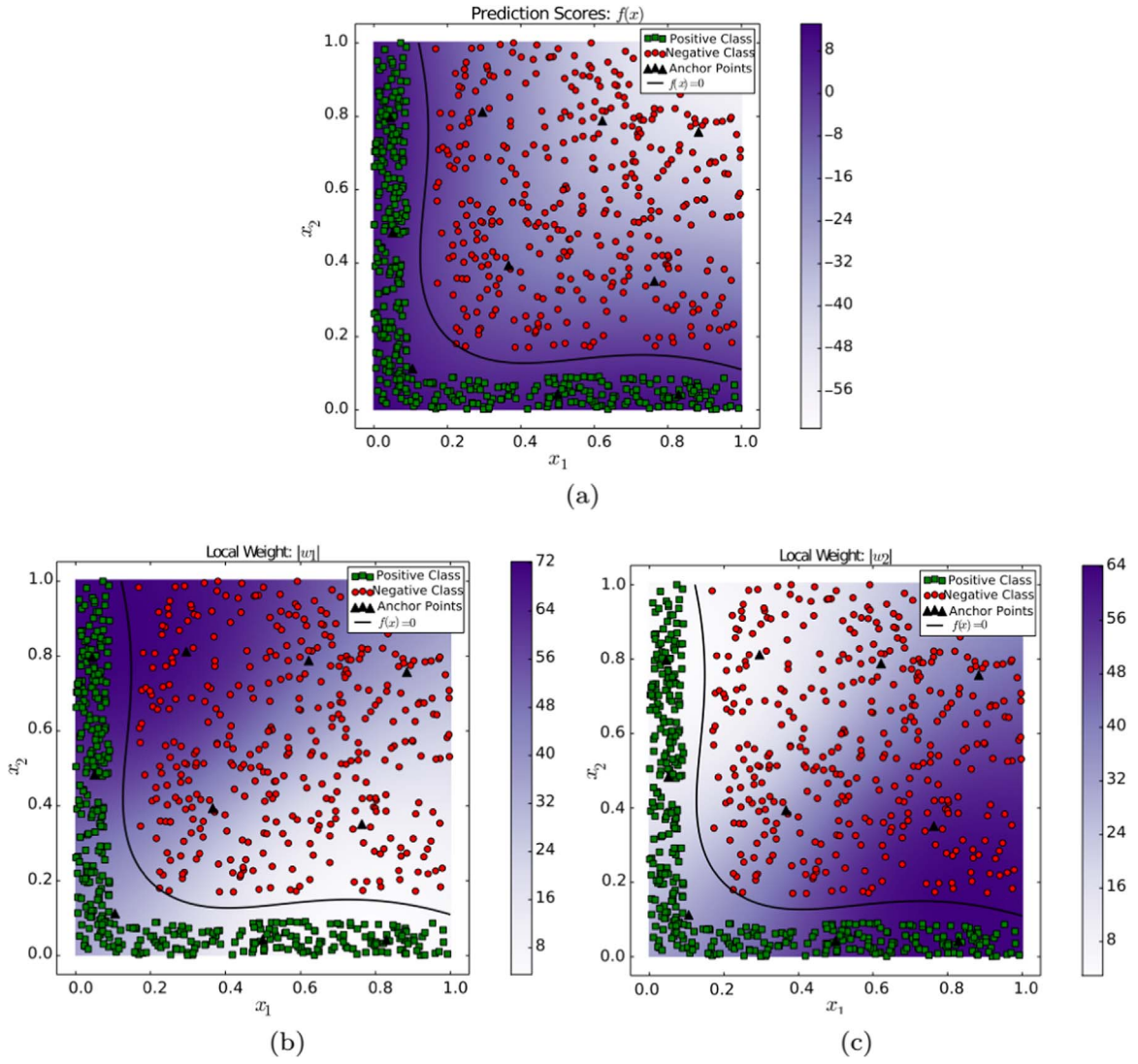

Fig. 2. Results of CAFÉ-Map with the L-shaped data set: (a) The scatter plot of the data and the decision boundary obtained with CAFÉ-Map shown as a black line. (b) The absolute value of the first component of the local weight vector $\boldsymbol{w}(\boldsymbol{x})=W\gamma(\boldsymbol{x})$. (c) The absolute value of the second component of the local weight vector.

2.3. Experimental evaluation

2.3.1. Datasets

We have used two groups of data sets for demonstration and analysis of the performance of CAFÉ-Map: Toy datasets and real world biomedical datasets. Here, we provide a brief description of these data sets.

2.4. Toy data sets

We have used four different toy datasets to demonstrate the behavior of CAFÉ-Map. These data sets are separable with binary labels and allow us to understand the feature selection process in CAFÉ-Map. These data sets include: L-shaped, 2×2 Checkerboard pattern, Linear Interpolation set and a circular pattern. The number of positive and negative examples in these datasets is equal. The circular, L-shaped and 2×2 checkerboard patterns are two-dimensional whereas the linear interpolation data set is 50 dimensional. Each data set has 100 positive and 100 negative examples. The coordinates of each point are used as features. For each data set, we trained on the whole data set and show the plot of the data along with the prediction scores and the classification boundary obtained from CAFÉ-Map. We also plot the absolute values of the local weight vector at different points in the feature space. This visualization allows us to find the local importance of different features in the feature spaces of these datasets.

2.5. Real world data sets

For the performance assessment and analysis of CAFÉ-Map, we have used four high dimensional real world datasets: the mass-spectrometry Arcene data set [54] and three different microarray datasets for Diffuse large B-cell Lymphoma [16], Prostate [17] and Breast [18–21] cancers. Arcene's task is to distinguish cancerous vs. normal patterns. The arcene dataset contains 100 training and 100 validation examples. Each example is represented by its 10,000 mass spectrometry spectrum features. All the datasets microarray cancer data sets were obtained from Glaab et al. [55]. Each of these data sets consists of expression measurements for a number of genetic probes for different types of tumors or control samples. Glaab et al. have used fold change filtering and thresholding for preprocessing as explained in the supplementary material of their paper. Each example has been normalized to unit norm. The number of features and examples used in evaluating the performance of CAFÉ-Map for each of these data sets is given in Table 1. The prostate cancer data set has 52 tumor and 50 control samples with 2135 genes. The Lymphoma data set has 7129 genes with 58 Diffuse large B-cell lymphoma samples and 19 follicular

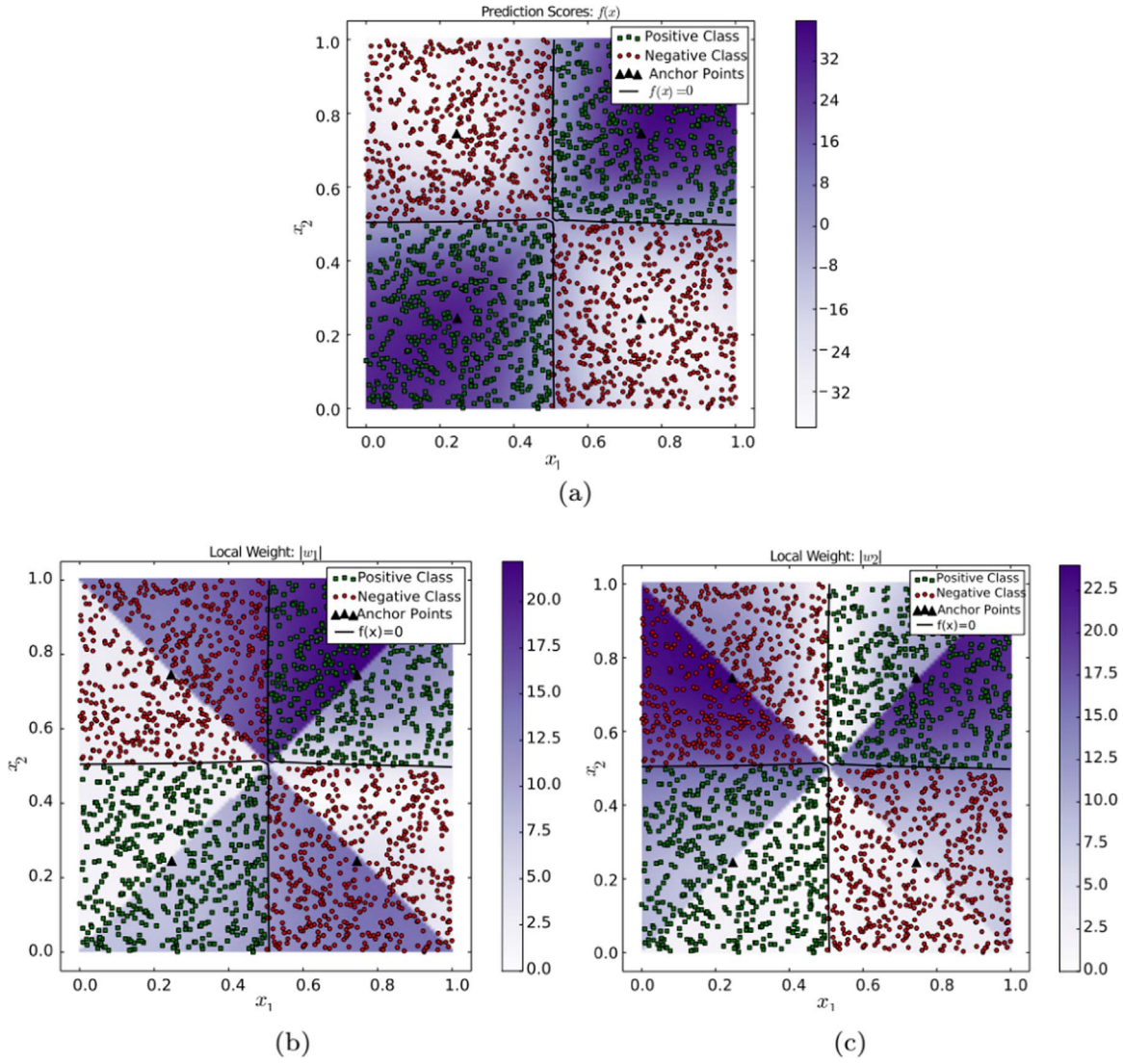

Fig. 3. Results of CAFÉ-Map with the2x2 Checkerboard pattern: (a) The scatter plot of the data and the decision boundary obtained with CAFÉ-Map shown as a black line. (b) The absolute value of the first component of the local weight vector $\boldsymbol{w}(\boldsymbol{x})=W\boldsymbol{\gamma}(\boldsymbol{x})$. (c) The absolute value of the second component of the local weight vector.

lymphoma samples. The breast cancer set consists of 84 Luminal and 44 Non-Luminal samples with 47, 293 genes. We report the prediction performance of CAFÉ-Map for these datasets as well as the interpretation of different features.

2.5.1. Cross-validation protocol and performance metrics

We have used 10-fold stratified cross-validation for computing accuracy metrics such as the mean accuracy and Area under the Receiver Operating Characteristics curves (AUC-ROC) across different folds. We also report the standard deviation of the accuracy metric so that our results can be directly compared with those in the work by Glaab et al.

To aid the reader in interpreting the results of feature selection in CAFÉ-Map, we also report the number of non-zero components of the local weight vector averaged across training examples. We refer to this term as the number of active features defined mathematically as $F = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d} I\left[\boldsymbol{w}_j(\boldsymbol{x}_i)\right]$ with the indicator function $I[\bullet]=1$ if its argument is non-zero and 0 otherwise.

2.5.2. Model selection

CAFÉ-Map requires a value of the regularization parameter λ as well as the parameters associated with the local coding. The LLC algorithm involves selecting its own locality enforcing parameter γ as well as the number of anchors K. These parameters were coarsely optimized for each data set using a stratified cross-validation protocol over the training examples in each of the 10 folds used in performance evaluation. It was observed that the most important parameter in the local encoding was the number of anchors (data not shown). This is due to the fact that for good performance, the feature space needs to be appropriately sampled. The anchors were obtained by randomly selecting K examples in the given data set to act as anchors. The number of randomly selected anchors from each class is proportional to its number of examples in the training set. An alternative approach is to apply K-means clustering on the data and use the cluster centers as anchors. However, we noticed that the exact method used for selecting the anchors, random selection or clustering, did not seem to affect classification performance (results not shown).

3. Results and discussion

3.1. Results on toy datasets

In this section, we discuss the results of training CAFÉ-Map on four different toy datasets. The L-shaped data set is designed in a way that the decision boundary of the ideal classifier is shaped like an “L” with its two features being important in different regions of the feature

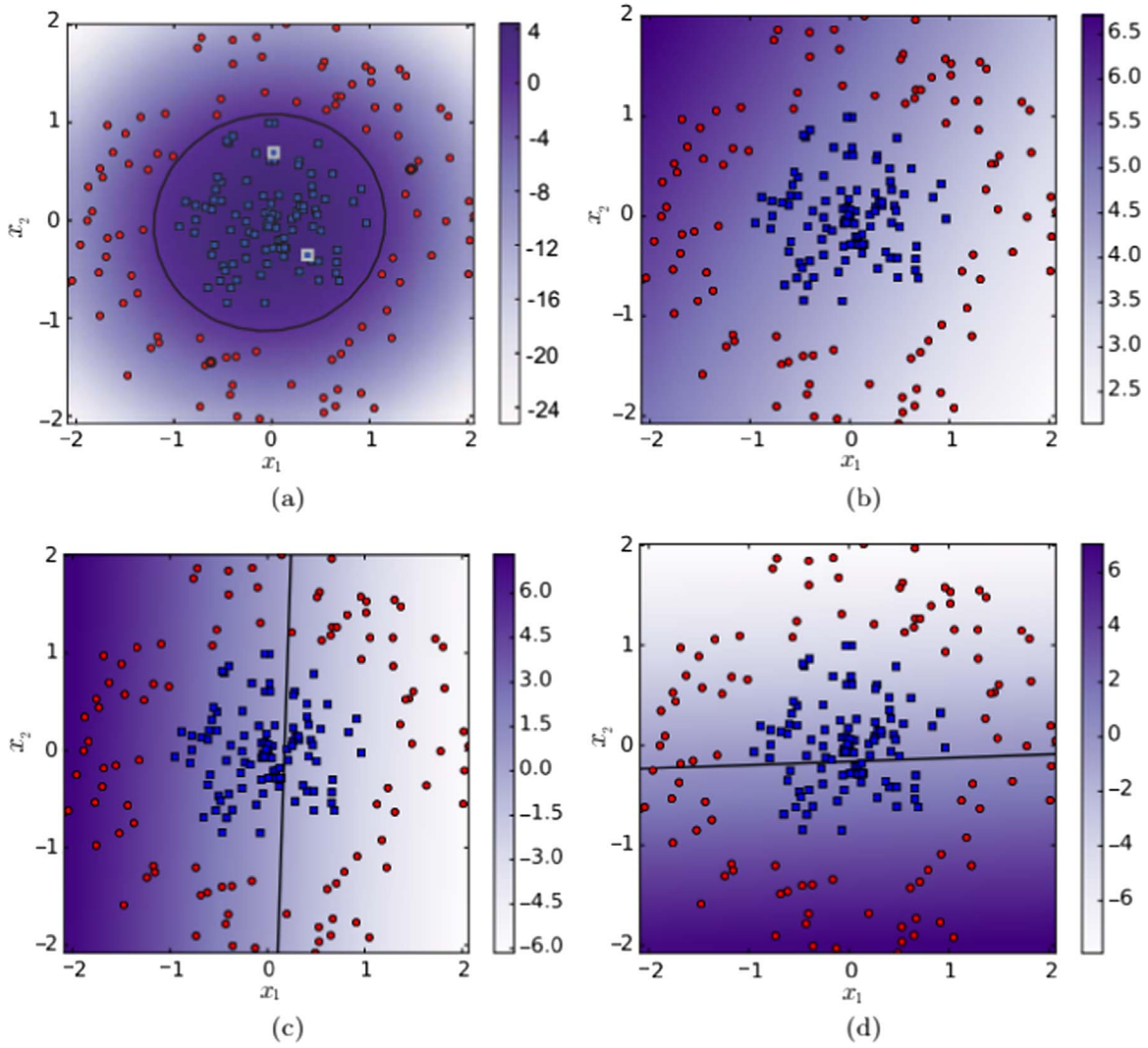

Fig. 4. Results of CAFÉ-Map on the circular data set. (a) The scatter plot of the data set and the boundary obtained by CAFÉ-Map along with the anchor points (highlighted). (b) The local bias component. (c) First component of the local weight vector. (d) The second component of the local weight vector.

space. Fig. 2 shows the results of applying CAFÉ-Map on this data set. It is evident that CAFÉ-Map is able to find a highly accurate classification boundary (Fig. 2(a)). The analysis of the absolute values of the local weights $\boldsymbol{w}(\boldsymbol{x})=W\boldsymbol{\gamma}(\boldsymbol{x})$ in different regions of the feature space shows their relative importance in Fig. 2(b) and (c). It is clear that different features are important in different regions of the feature space with $|w_1(\boldsymbol{x})|$ being higher than $|w_2(\boldsymbol{x})|$ along the vertical decision boundary and vice versa. This allows us to analyze the importance of each feature for any given example. The discriminant boundaries and local weights for the 2x2 checkerboard patterns are shown in Fig. 3. These plots also illustrate the effectiveness of CAFÉ-Map in uncovering the local importance of different features in different parts of the feature space.

We also demonstrate the results of CAFÉ-Map on a circular data set in Fig. 4. It is interesting to note that the locally linear classification boundary obtained by CAFÉ-Map with just 4 anchors is very smooth. The variation of local weight and bias values in different regions of the feature space clearly indicate the context aware nature of CAFÉ-Map for this data set.

To illustrate the working of CAFÉ-Map on higher dimensional and more complex data, we use the 50-dimensional linear interpolation data set in Fig. 5. This artificial data set has been created by setting the value of the j^{th} feature of each positive example from −1.0 for $j = 1$ to +1.0 for $j = 50$ through linear interpolation. Uniform random noise is then added for each example so that the noise to signal ratio is 100%. For the negative examples, the feature values are assigned in a completely opposite manner as shown in Fig. 5. For this data set, the first and last features are more important in comparison to the others. Ideally, a single feature can perform perfect classification for this data set. However, this is typically not possible due to the added noise. Fig. 5(b) and 5(d) show that the only a few components of the local vector are non-zero. It is interesting to note that the local weight values of all examples are almost similar. This is a consequence of the fact that the underlying classification problem is linearly separable. Fig. 5(c) plots the product of the local weight vector of each training example $\boldsymbol{w}(\boldsymbol{x}_i)=W\boldsymbol{\gamma}(\boldsymbol{x}_i)$ with its label y_i. It is easy to visually notice the high positive correlation of $y_i\boldsymbol{w}(\boldsymbol{x}_i)$ with $\boldsymbol{x}_i$. As discussed in the previous section, this high positive correlation is indicative of correct classification.

3.2. Results on real world datasets

We have tested CAFÉ-Map extensively on real world data sets. The results are shown in the table below. In this section we discuss these results and their comparison with other methods. We provide the references used in the comparison.

For the 10K Dimensional Arcene data set we have obtained AUC-ROC score of 94 and accuracy of 86% which is comparable to that

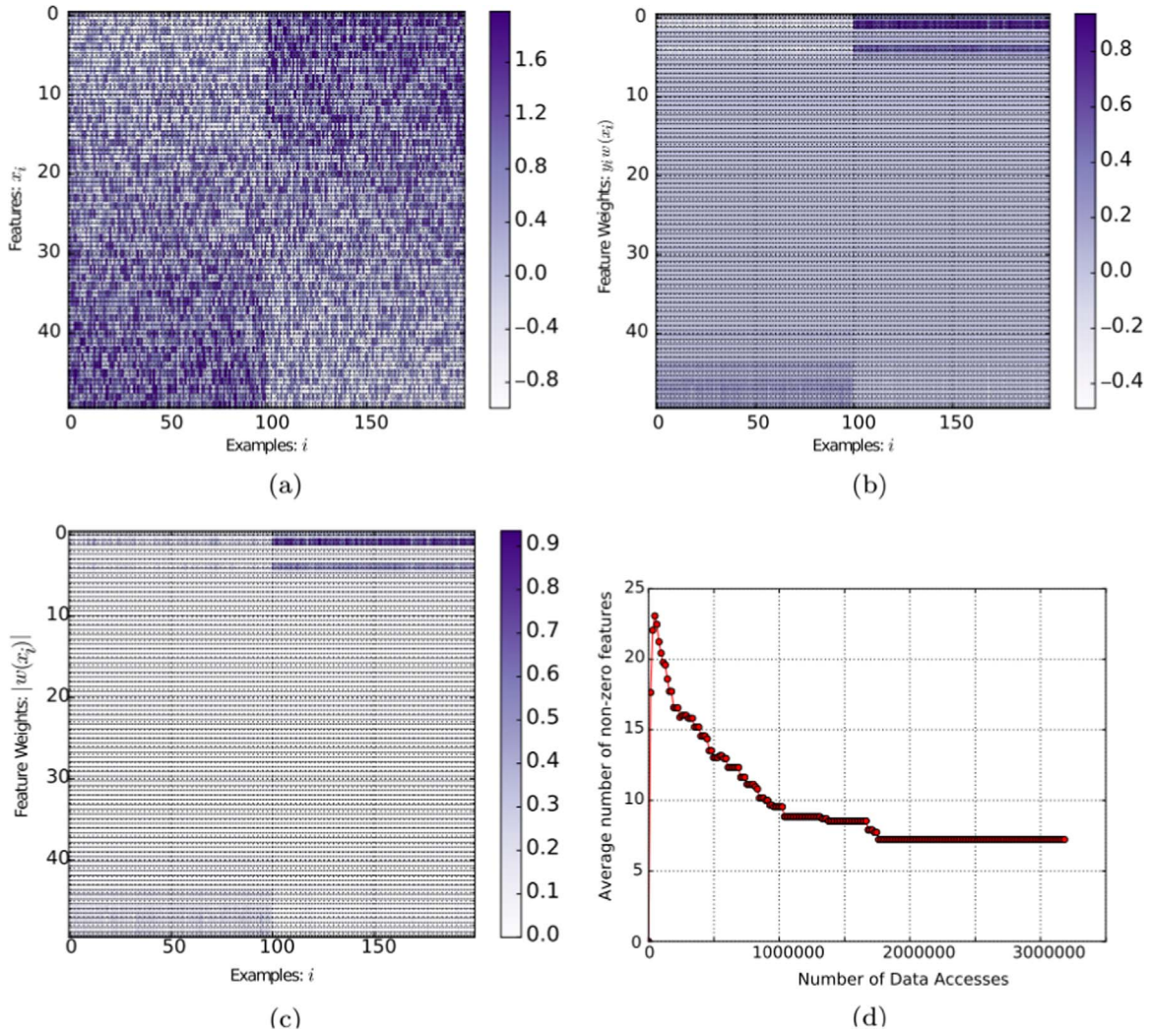

Fig. 5. Results of CAFÉ-Map with the linear interpolation data. (a) The linear interpolation data set with added noise. The first 100 examples belong to the negative class whereas the last 100 belong to the positive class. (b) Absolute values of the CAFÉ-Map's local weight vectors for each example. Notice that only a few features have high values. (c) CAFÉ-Map's local weight vector $\boldsymbol{w}(\boldsymbol{x}_i)=W\gamma(\boldsymbol{x}_i)$ multiplied by the training label y_i for all training examples. Notice the high correlation of $y_i\boldsymbol{w}(\boldsymbol{x}_i)$ with $\boldsymbol{x}_i$ and the ranking of the features. (d) The number of non-zero features averaged across all examples is plotted across the number of data accesses during training. Notice that the final classifier uses only a few features..

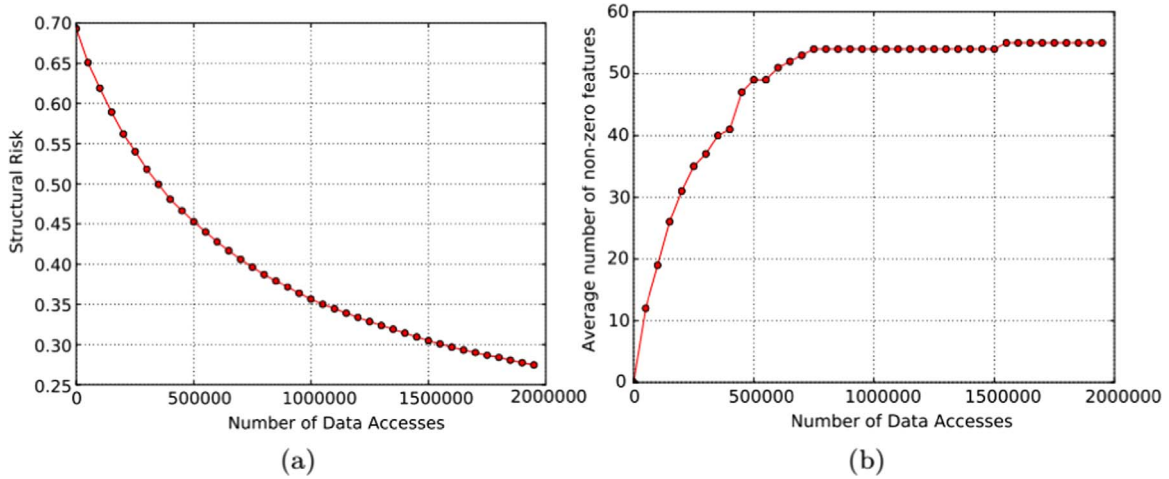

Fig. 6. Convergence of CAFÉ-Map with the arcene data set. (a) shows the structural risk minimization over different iterations and (b) shows the number of features with non-zeros values averaged across training examples.

obtained by other state of the art methods [54],[56–61]. Please note that, for this data set, the evaluation is performed on the validation data set after CAFÉ-Map is trained on the given training set. This follows the same protocol as used in the cited references. It is interesting to notice that, the number of active features is only 55 (0.55%) and even a smaller number (27) of local components are larger

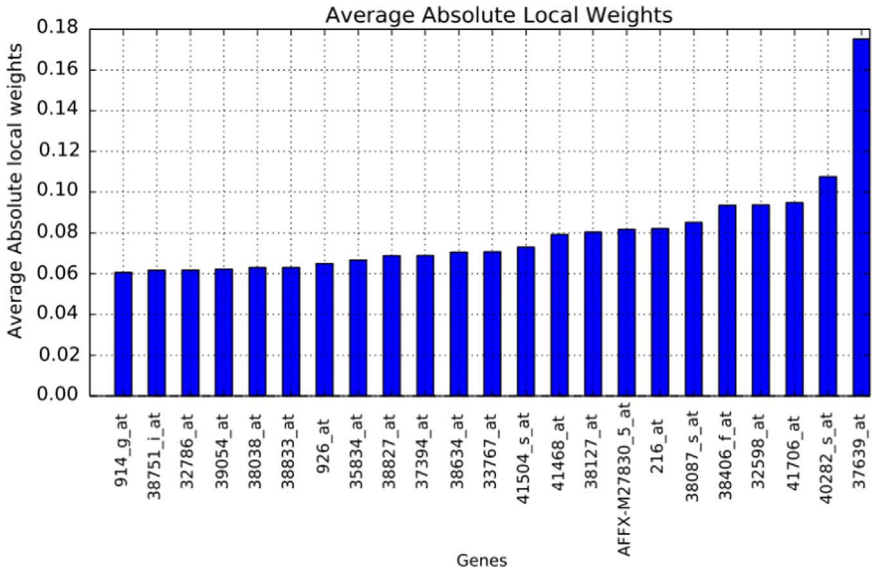

Fig. 7. Results of CAFÉ-Map with the prostate cancer data sets. The absolute value of average weights across all examples is shown for different genes.

Table 2
Identification of important genes for prostate cancer from CAFE-Map and their associated references.

Probe (Feature)	Gene name	Absolute weight	References
37639_at	HPN	0.18	[71]
40282_s_at	CFD	0.11	[72]
41706_at	AMACR	0.09	[73]
32598_at	NELL2	0.09	[74]
38406_f_at	PTGDS	0.09	[75]
38087_s_at	S100A4	0.09	[76, p. 4]
216_at	PTGDS	0.08	[75]
AFFX-M27830_5_at	Control Probe	0.08	[62]
38127_at	SDC1	0.08	[77]
41468_at	TRGV9	0.08	[78]
41504_s_at	MAF	0.07	[79]
33767_at	NEFH	0.07	[80]
38634_at	RBP1	0.07	[81, p. 1]
37394_at	C7	0.07	[82]
38827_at	AGR2	0.07	[83, p. 2]
35834_at	AZGP1	0.07	[84]
926_at	MT1G	0.06	[85]
38833_at	HLA-DPA1	0.06	[86]
38038_at	LUM	0.06	[87]
39054_at	GSTM4	0.06	[88]
32786_at	JUN-B	0.06	[89]
38751_i_at	ATP5I	0.06	[90]
914_g_at	ERG	0.06	[91]

than $\frac{\lambda}{\beta}$. This clearly illustrates the effectiveness of CAFÉ-Map in feature selection. Fig. 6 shows the convergence characteristics of CAFÉ-Map for this data set in terms of the structural risk P(W) defined in Eq. (3) and the number of active features.

For the three microarray data sets, the cross-validation performance of CAFÉ-Map is better than existing approaches as shown in Table 1. For comparison, we present the best results among a number of different classifiers given in the work by Glaab et al. (see table 4 and table 5 in their work [62]). It is interesting to note that the proposed scheme performs better than earlier approaches over all these data sets with exactly the same evaluation protocol. The number of active features obtained is very small for all these data sets relative to the number of original features. The number of features with absolute values greater than $\frac{\lambda}{\beta}$ is even smaller: 5, 41 and 2 for the Prostate, Lymphoma and Breast cancer data sets, respectively.

Table 1 also gives the average run times across multiple cross-validation runs for different data sets on a Dell Core-i5 laptop with 4 GB RAM. It can clearly be seen that CAFÉ-Map offers very high speed of convergence over these data sets.

3.3. Comparison with Locally Linear SVM

The formulation of CAFÉ-Map is similar to that of the locally linear SVM (LLSVM) proposed by Ladicky and Torr. Both CAFÉ-Map and LLSVM are locally linear classification methods that use a context aware weight function. However, the major difference between these techniques is the choice of the regularization function in CAFÉ-Map. CAFÉ-Map uses an L1 regularization function over the weight matrix which enforces sparsity. LLSVM, on the other hand, uses L2 regularization and a stochastic gradient descent based optimization algorithm. As a consequence, CAFÉ-Map can be expected to produce a smaller number of active features in comparison to LLSVM. We tested this hypothesis over the prostate cancer data set by applying LLSVM. The best cross-validation AUC score for the LLSVM is 94.3 with 480 active features. In comparison, CAFÉ-Map gives an AUC score of 98.0 and only 38 active features. This clearly shows the effectiveness of using CAFÉ-Map in comparison to LLSVM and similar approaches.

3.4. Analysis of prostate cancer features

In order to see if the features selected by CAFÉ-Map are meaningful or not, we mined the literature for relevance of top scoring genes in the prostate data set to prostate cancer. For this purpose, we ranked genes by the average of the absolute value of local weights across all examples in the prostate data set after training through CAFÉ-Map. Fig. 7 plots the weight values for top ranked genes. We found references in the literature for all genes with absolute weight values from CAFÉ-Map higher than 0.06. Table 2 lists all such microarray probes and their gene identifiers together with associated literature references indicating their relevance to prostate cancer. For example, Hepsin (HPN), the top scoring gene selected by CAFÉ-Map, is known to be overexpressed consistently in prostate cancer cases [71].

A significance of CAFÉ-Map is its unique ability to analyze the impact of different features at the individual instance level which can be very useful in interpreting why an example is being classified in a certain way. In order to demonstrate this, we clustered the positive and negative examples in the data set based on the 40 top ranked components of their local weight vectors $\boldsymbol{w}(\boldsymbol{x}_i)=W\boldsymbol{\gamma}(\boldsymbol{x}_i)$. The clustering was done using the K-means algorithm. The results of this clustering are shown in the figure below. The figure shows the local weight values

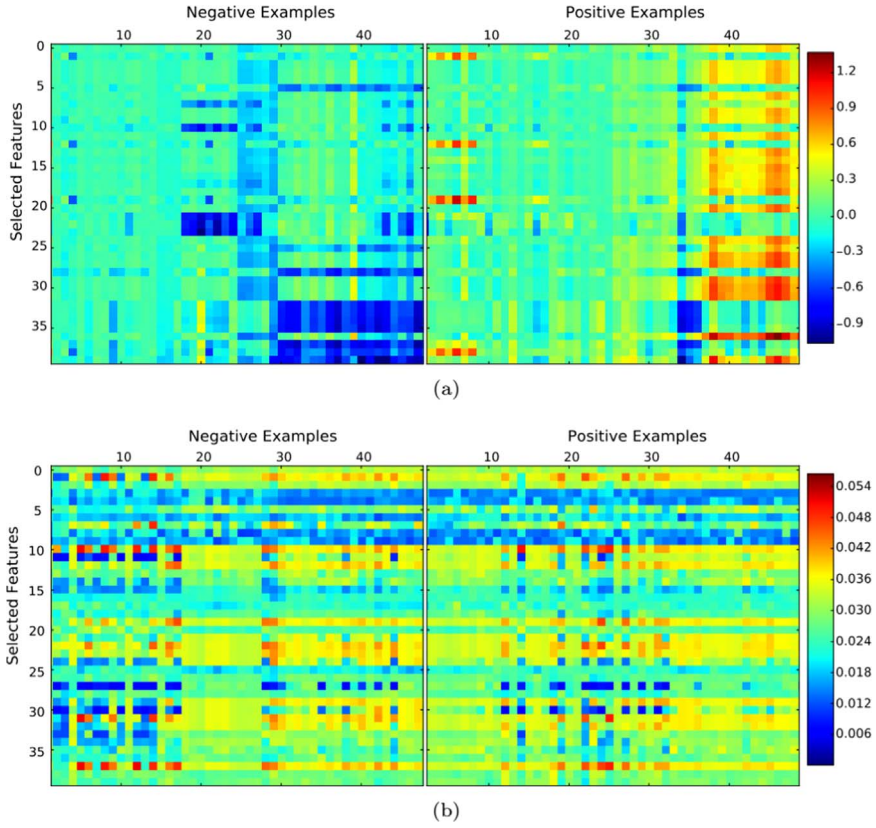

Fig. 8. Clustering analysis of local weights from CAFÉ-Map. (a) Top ranked Local weights and (b) Original Features of each example.

of each example as well as the original features. Please note that this clustering is not based on the original features. The components of the local weight vector or features are indexed along the vertical axis of the heatmap based on their rank and the examples are indexed based on their cluster membership. This clustering reveals an interesting structure in the data. Examples within the same cluster have similar local weights which correspond to similar expression patterns. For instance, the local weight vectors for examples 1–8 of the positive class are very different from examples 38–52 even though both of them belong to the positive class. Unlike other positive class instances, examples 33–37 have large negative local weights for certain features. This reveals that there are large differences between the expression profiles of these examples. It must be noted that such strong clustering is not visible in the heatmap of the original features shown in Fig. 8. A similar structure is visible in the local weight values of the negative class. This figure also shows that the relative importance of features varies across examples based on their local context. It can be postulated that such differences are be a consequence of differences in age, gender, disease progression, etc. Unfortunately, the prostate cancer data set does not provide sufficient information to investigate the source of these differences. However, it clearly illustrates the primary idea behind CAFÉ-Map and its usefulness in analyzing similar data sets.

4. Discussion and conclusions

CAFÉ-Map is a locally linear classifier with built-in feature ranking capabilities. It allows the user to estimate the relative importance of different features for individual examples or in different regions of the feature space. Our comparative analysis reveals that CAFÉ-Map compares very well with state of the art feature analysis algorithms and is particularly well suited to biomedical data. For the real world data sets such as Arcene, Prostate, Lymphoma and Breast cancer, the results of CAFÉ-Map are comparable to or better than published feature ranking and selection methods. Specifically, for the Arcene dataset, our comparison with SVM, filter based, game theoretic and random forest methods show that CAFÉ-Map achieves better classification accuracy than these methods [56–61]. Similarly, for the prostate cancer data set, we have compared CAFÉ-Map with rule based, dimensionality reduction, Gaussian Process Modeling and Genetic Algorithm techniques [62–67]. CAFÉ-Map achieves the best classification performance among all these methods as well. This also holds true for the Lymphoma data set which was used to compare CAFÉ-Map with rule based, genetic algorithm, ensemble or hybrid feature selection and Ontology based frameworks [16], [62], [65], [67–70]. For the breast cancer data set, CAFÉ-Map performs significantly better than the published results in the work by Glaab et al. [62]. We have also compared CAFÉ-Map with the locally linear SVM by Ladicky and Torr and found that CAFÉ-Map offers significantly better performance in terms of feature selection and classification accuracy. CAFÉ-Map is computationally efficient and cross-validation results for multiple runs over the 5 real world data sets took no longer than 10 min on a typical laptop computer. However, the biggest advantage of CAFÉ-Map is its ability to generate information about the importance of different features at the level of individual examples. We demonstrated this ability of CAFÉ-Map over the prostate cancer data set, where the genes selected by CAFÉ-Map are known to have significant relation to prostate cancer.

Conflict of Interest

There is no conflict of interest.
Dr Muhammad Arif.

Acknowledgements

The authors would like to thank the Deanship of Scientific Research at Umm Al-Qura University (project # 43408032) for the financial support.

References

[1] A. Perez-Diez, A. Morgun, N. Shulzhenko, Microarrays for cancer diagnosis and classification, Adv. Exp. Med. Biol. 593 (2007) 74–85.

[2] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer [Jan.]Nature 415 (6871) (2002) 530–536.

[3] C.M. Perou, T. Sørlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lønning, A.-L. Børresen-Dale, P.O. Brown, D. Botstein, Molecular portraits of human breast tumours [Aug.]Nature 406 (6797) (2000) 747–752.

[4] M.L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A.F. Wright, J.F. Wilson, F. Agakov, P. Navarro, C.S. Haley, Application of high-dimensional feature selection: evaluation for genomic prediction in man [May]Sci. Rep. 5 (2015) 10312.

[5] D.-E. van der Merwe, K. Oikonomopoulou, J. Marshall, E.P. Diamandis, Mass spectrometry: uncovering the cancer proteome for diagnostics, Adv. Cancer Res. 96 (2007) 23–50.

[6] K.D. Rodland, Proteomics and cancer diagnosis: the potential of mass spectrometry [Jul]Clin. Biochem. 37 (7) (2004) 579–583.

[7] E.P. Diamandis, Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. Opportunities and potential limitations, Mol. Cell. Proteom. 3 (2004) 367–378.

[8] C. Kumar, M. Mann, Bioinformatics analysis of mass spectrometry-based proteomics data sets [Jun]FEBS Lett. 583 (11) (2009) 1703–1712.

[9] J.L. Semmlow, B. Griffel, Biosignal and Medical Image Processing, Third ed., CRC Press, Boca Raton, FL, USA, 2014.

[10] I. Bankman, Handbook of Medical Image Processing and Analysis, Academic Press, USA, 2008.

[11] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Prog Artif Intell 5 (65) (2016). http://dx.doi.org/10.1007/s13748-015-0080-y.

[12] K.R. Foster, R. Koprowski, J.D. Skufca, Machine learning, medical diagnosis, and biomedical engineering research - commentary, Biomed. Eng. OnLine 13 (2014) 94.

[13] P. Sajda, Machine Learning for Detection and Diagnosis of Disease, Annu. Rev. Biomed. Eng. 8 (1) (2006) 537–565.

[14] X. Li, R. Xu, High-Dimensional Data Analysis in Cancer Research, Springer Science & Business Media, NY, USA, 2008.

[15] R. Clarke, H.W. Ressom, A. Wang, J. Xuan, M.C. Liu, E.A. Gehan, Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data [Jan.]Nat. Rev. Cancer 8 (1) (2008) 37–49.

[16] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C.T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning [Jan.]Nat. Med. 8 (1) (2002) 68–74.

[17] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior [Mar.]Cancer Cell 1 (2) (2002) 203–209.

[18] A. Naderi, A.E. Teschendorff, N.L. Barbosa-Morais, S.E. Pinder, A.R. Green, D.G. Powe, J.F.R. Robertson, S. Aparicio, I.O. Ellis, J.D. Brenton, C. Caldas, A gene-expression signature to predict survival in breast cancer across independent data sets [Aug.]Oncogene 26 (10) (2006) 1507–1516.

[19] S.F. Chin, A.E. Teschendorff, J.C. Marioni, Y. Wang, N.L. Barbosa-Morais, N.P. Thorne, J.L. Costa, S.E. Pinder, M.A. van de Wiel, A.R. Green, I.O. Ellis, P.L. Porter, S. Tavaré, J.D. Brenton, B. Ylstra, C. Caldas, High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer, Genome Biol. 8 (10) (2007) R215.

[20] H. Zhang, E.A. Rakha, G.R. Ball, I. Spiteri, M. Aleskandarany, E.C. Paish, D.G. Powe, R.D. Macmillan, C. Caldas, I.O. Ellis, A.R. Green, The proteins FABP7 and OATP2 are associated with the basal phenotype and patient outcome in human breast cancer [May]Breast Cancer Res. Treat. 121 (1) (2010) 41–51.

[21] H.O. Habashy, D.G. Powe, E. Glaab, G. Ball, I. Spiteri, N. Krasnogor, J.M. Garibaldi, E.A. Rakha, A.R. Green, C. Caldas, I.O. Ellis, RERG (Ras-like, oestrogen-regulated, growth-inhibitor) expression in breast cancer: a marker of ER-positive luminal-like subtype [Jul.]Breast Cancer Res. Treat. 128 (2) (2011) 315–326.

[22] A. Ben-Hur, C.S. Ong, S. Sonnenburg, B. Schölkopf, G. Rätsch, Support vector machines and kernels for computational biology [Oct.]PLoS Comput. Biol. 4 (10) (2008) e1000173.

[23] A. Vellido, J. Martín-guerrero, P. J. G. Lisboa, Making machine learning models interpretable, in: In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, 2012.

[24] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, An Interpretable Stroke Prediction Model using Rules and Bayesian Analysis, Working Paper, Nov. 2013.

[25] D. Hofmann, F.-M. Schleif, B. Paaßen, B. Hammer, Learning interpretable kernelized prototype-based models [Oct.]Neurocomputing 141 (2014) 84–96.

[26] C. Otte, Safe and interpretable machine learning: a methodological review,Intelligent Data Analysis, Springer, Berlin Heidelberg, Germany, 2013, pp. 111–122.

[27] C. Rudin, Algorithms for Interpretable Machine Learning, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2014, pp. 1519–1519.

[28] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics [Oct.]Bioinformatics 23 (19) (2007) 2507–2517.

[29] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection [Mar.]J. Mach. Learn Res. 3 (2003) 1157–1182.

[30] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, [Oct.]Inf. Sci. 282 (2014) 111–135.

[31] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, A review of feature selection methods on synthetic data,, Knowl. Inf. Syst. 34 (3) (. 2012) 483–519.

[32] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Springer Science & Business Media, NY, USA, 2012.

[33] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, Feature selection for SVMs, Adv. Neural Inf. Process. Syst. 13 (2000) 668–674.

[34] L. C. Molina, L. Belanche, A. Nebot, Feature selection algorithms: a survey and experimental evaluation, in: Proceedings of the 2002 IEEE International Conference on Data Mining, 2002. ICDM 2003, pp. 306–313.

[35] Huan Liu, Hiroshi Motoda, Rudy Setiono, Zheng Zhao, Feature selection: an ever evolving frontier in data mining, J. Mach. Learn. Res. 10 (2010) 4–13.

[36] V. Vapnik, The Nature of Statistical Learning Theory, Springer Science & Business Media, 2013.

[37] Chang, Feature ranking using linear SVM, in: Proceedings if the JMLR Workshop and Conference: Causation and Prediction Challenge, 2008, pp. 53–64.

[38] V. Kecman, J. P. Brooks, Locally linear support vector machines and other local models, in: Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), 2010, pp. 1–6.

[39] L. Ladicky, P. Torr, Locally Linear Support Vector Machines, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), New York, NY, USA, 2011, pp. 985–992.

[40] M. Fornoni, B. Caputo, and F. Orabona, Multiclass Latent Locally Linear Support Vector Machines, presented at the Asian Conference on Machine Learning, 2013, pp. 229–244.

[41] M. Fornoni, B. Caputo, F. Orabona, Multiclass Latent Locally Linear Support Vector Machines, presented at the Asian Conference on Machine Learning, 2013, pp. 229–244.

[42] J. Chen, Y. Liu, Locally linear embedding: a survey, [Jan.]Artif. Intell. Rev. 36 (1) (2011) 29–48.

[43] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, ArXiv12065538 Cs, Jun. 201

[44] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, 2011, pp. 76.1–76.12.

[45] B.-D. Liu, Y.-X. Wang, Y.-J. Zhang, B. Shen, Learning dictionary on manifolds for image classification [Jul.]Pattern Recognit. 46 (7) (2013) 1879–1890.

[46] Y. Huang, Z. Wu, L. Wang, T. Tan, Feature coding in image classification: a comprehensive study, [Mar.]IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 493–506.

[47] A. Coates, A. Y. Ng, Learning Feature Representations with K-Means, in Neural Networks: Tricks of the Trade, G. Montavon, G. B. Orr, K.-R. Müller, Eds. Springer Berlin Heidelberg, 2012, pp. 561–580.

[48] S.T. Roweis, L.K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding [Dec.]Science 290 (5500) (2000) 2323–2326.

[49] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, Kernel Codebooks for Scene Categorization, in Computer Vision – ECCV 2008, D. Forsyth, P. Torr, and A. Zisserman, Eds. Springer Berlin Heidelberg, 2008, pp. 696–709.

[50] S. Gao, I. W. H. Tsang, L. T. Chia, P. Zhao, Local features are not lonely - Laplacian sparse coding for image classification, in: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 3555–3561.

[51] T.Z. Kai Yu, Yihong Gong, Nonlinear Learning using Local Coordinate Coding, Adv. Neural Inf. Process. Systems NIPS (2009) 2223–2231.

[52] S. Shalev-Shwartz, A. Tewari, Stochastic Methods for L1-regularized Loss Minimization [Jul.]J. Mach. Learn. Res. 12 (2011) 1865–1892.

[53] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, K. Muller, Fisher discriminant analysis with kernels, in: Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX, 1999, pp. 41–48.

[54] I. Guyon, J. Li, T. Mader, P.A. Pletscher, G. Schneider, M. Uhr, Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark [Sep.]Pattern Recognit. Lett. 28 (12) (2007) 1438–1444.

[55] E. Glaab, J. Bacardit, J.M. Garibaldi, N. Krasnogor, Using Rule-Based Machine Learning for Candidate Disease Gene Prioritization and Sample Classification of Cancer Gene Expression Data [Jul.]PLoS ONE 7 (7) (2012) e39932.

[56] Y.-W. Chen, C.-J. Lin, Combining SVMs with Various Feature Selection Strategies, in Feature Extraction, I. Guyon, M. Nikravesh, S. Gunn, L. A. Zadeh, Eds. Springer Berlin Heidelberg, 2006, pp. 315–324.

[57] T.N. Lal, O. Chapelle, B. Schölkopf, Combining a Filter Method with SVMs,, in: I. Guyon, M. Nikravesh, S. Gunn, L.A. Zadeh (Eds.), Feature Extraction, Springer, Berlin Heidelberg, Germany, 2006, pp. 439–445.

[58] R. Gaudel, M. Sebag, Feature Selection as a One-Player Game, presented at in: Proceedings of the International Conference on Machine Learning (ICML), Haifa, Israel, 2010, pp. 359–366.

[59] C. Vens, F. Costa, Random Forest Based Feature Induction, in: Proceedings of the 2011 IEEE 11th International Conference on Data Mining, 2011, pp. 744–753.

[60] S. Cohen, G. Dror, E. Ruppin, Feature selection via coalitional game theory [Jul.] Neural Comput. 19 (7) (2007) 1939–1961.

[61] M. Seo, S. Oh, CBFS: high Performance feature selection algorithm based on

feature clearness [Jul.]PLOS ONE 7 (7) (2012) e40419.
[62] E. Glaab, J. Bacardit, J.M. Garibaldi, N. Krasnogo, Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data [Jul.]PLOS ONE 7 (7) (2012) e39932.
[63] L. Shen, E.C. Tan, Dimension reduction-based penalized logistic regression for cancer classification using microarray data [Apr.]IEEEACM Trans. Comput Biol. Bioinforma. 2 (2) (2005) 166–175.
[64] T. K. Paul, H. Iba, Extraction of Informative Genes from Microarray Data, presented at in: proceedings of the Genetic and Evolutionary Computation Conference (GECCO), Washington, DC, USA, 2005.
[65] L.F.A. Wessels, M.J.T. Reinders, A.A.M. Hart, C.J. Veenman, H. Dai, Y.D. He, L.J. van't Veer, A protocol for building and evaluating predictors of disease state based on microarray data [Oct.]Bioinforma. Oxf. Engl. 21 (19) (2005) 3755–3762.
[66] W. Chu, Z. Ghahramani, F. Falciani, D.L. Wild, Biomarker discovery in microarray gene expression data with Gaussian processes [Aug.]Bioinforma. Oxf. Engl. 21 (16) (2005) 3385–3393.
[67] M. Lecocke, K. Hess, An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data, Cancer Inform. 2 (2006) 313–327.
[68] J. Liu, H.-B. Zhou, Tumor classification based on gene microarray data and hybrid learning method, in: Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, 2003, vol. 4, pp. 2275–2280.
[69] L. Goh, Q. Song, N. Kasabov, A Novel Feature Selection Method to Improve Classification of Gene Expression Data, presented at In: Proceedings of the Second Conference on Asia-Pacific bioinformatics, New Zealand, vol. 29, pp. 161–166.
[70] Y. Hu and N. Kasabov, Ontology-Based Framework for Personalized Diagnosis and Prognosis of Cancer Based on Gene Expression Data, in Neural Information Processing, M. Ishikawa, K. Doya, H. Miyamoto, and T. Yamakawa, Eds. Springer Berlin Heidelberg, 2007, pp. 846–855.
[71] S.K. Holt, E.M. Kwon, D.W. Lin, E.A. Ostrander, J.L. Stanford, Association of hepsin gene variants with prostate cancer risk and prognosis [Jun.]Prostate 70 (9) (2010) 1012–1019.
[72] W. Tan, L. Wang, Q. Ma, M. Qi, N. Lu, L. Zhang, B. Han, Adiponectin as a potential tumor suppressor inhibiting epithelial-to-mesenchymal transition but frequently silenced in prostate cancer by promoter methylation [Aug.]Prostate 75 (11) (2015) 1197–1205.
[73] V. Ananthanarayanan, R.J. Deaton, X.J. Yang, M.R. Pins, P.H. Gann, Alpha-methylacyl-CoA racemase (AMACR) expression in normal prostatic glands and high-grade prostatic intraepithelial neoplasia (HGPIN): association with diagnosis of prostate cancer [Jun.]Prostate 63 (4) (2005) 341–346.
[74] U.S. Shah, R.H. Getzenberg, Fingerprinting the diseased prostate: associations between BPH and prostate cancer [Jan.]J. Cell. Biochem. 91 (1) (2004) 161–169.
[75] D. Wang, R.N. DuBois, Prostaglandins and cancer [Jan.]Gut 55 (1) (2006) 115–122.
[76] K. Boye, G.M. Mælandsmo, S100A4 and Metastasis [Feb.]Am. J. Pathol. 176 (2) (2010) 528–535.
[77] J. Kiviniemi, M. Kallajoki, I. Kujala, M.-T. Matikainen, K. Alanen, M. Jalkanen, M. Salmivirta, Altered expression of syndecan-1 in prostate cancer [Feb.]APMIS Acta Pathol. Microbiol. Immunol. Scand. 112 (2) (2004) 89–97.
[78] K. Rostad, M. Mannelqvist, O.J. Halvorsen, A.M. Oyan, T.H. Bø, L. Stordrange, S. Olsen, S.A. Haukaas, B. Lin, L. Hood, I. Jonassen, L.A. Akslen, K.-H. Kalland, ERG upregulation and related ETS transcription factors in prostate cancer [Jan.] Int. J. Oncol. 30 (1) (2007) 19–32.
[79] P.S. Nelson, N. Clegg, H. Arnold, C. Ferguson, M. Bonham, J. White, L. Hood, B. Lin, The program of androgen-responsive genes in neoplastic prostate epithelium [Sep.]Proc. Natl. Acad. Sci. 99 (18) (2002) 11890–11895.
[80] N. Dubrowinskaja, K. Gebauer, I. Peters, J. Hennenlotter, M. Abbas, R. Scherer, H. Tezval, A.S. Merseburger, A. Stenzl, V. Grünwald, M.A. Kuczyk, J. Serth, Neurofilament Heavy polypeptide CpG island methylation associates with prognosis of renal cell carcinoma and prediction of antivascular endothelial growth factor therapy response [Apr.]Cancer Med 3 (2) (2014) 300–309.
[81] C. Jerónimo, R. Henrique, J. Oliveira, F. Lobo, I. Pais, M.R. Teixeira, C. Lopes, Aberrant cellular retinol binding protein 1 (CRBP1) gene expression and promoter methylation in prostate cancer [Aug.]J. Clin. Pathol. 57 (8) (2004) 872–876.
[82] J.S. Shoemaker, S.M. Lin, Methods of Microarray Data Analysis I, Springer Science & Business Media, 2006.
[83] H. Bu, S. Bormann, G. Schäfer, W. Horninger, P. Massoner, A. Neeb, V.-K. Lakshmanan, D. Maddalo, A. Nestl, H. Sültmann, A.C.B. Cato, H. Klocker, The anterior gradient 2 (AGR2) gene is overexpressed in prostate cancer and may be useful as a urine sediment marker for prostate cancer detection [May]Prostate 71 (6) (2011) 575–587.
[84] J. Lapointe, C. Li, J.P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A.M. DeMarzo, R. Tibshirani, D. Botstein, P.O. Brown, J.D. Brooks, J.R. Pollack, Gene expression profiling identifies clinically relevant subtypes of prostate cancer [Jan.]Proc. Natl. Acad. Sci. USA. 101 (3) (2004) 811–816.
[85] R. Henrique, C. Jerónimo, M.O. Hoque, S. Nomoto, A.L. Carvalho, V.L. Costa, J. Oliveira, M.R. Teixeira, C. Lopes, D. Sidransky, MT1G Hypermethylation is associated with higher tumor stage in prostate cancer [May]Cancer Epidemiol. Biomark. Prev. 14 (5) (2005) 1274–1278.
[86] J.L. Gregg, K.E. Brown, E.M. Mintz, H. Piontkivska, G.C. Fraizer, Analysis of gene expression in prostate cancer epithelial and interstitial stromal cells using laser capture microdissection, BMC Cancer 10 (2010) 165.
[87] V.J. Coulson-Thomas, Y.M. Coulson-Thomas, T.F. Gesteira, C.A.A. de Paula, C.R.W. Carneiro, V. Ortiz, L. Toma, W.K. Kao, H.B. Nader, Lumican expression, localization and antitumor activity in prostate cancer [Apr.]Exp. Cell Res. 319 (7) (2013) 967–981.
[88] M.J. Monument, K.M. Johnson, A.H. Grossmann, J.D. Schiffman, R.L. Randall, S.L. Lessnick, Microsatellites with Macro-Influence in Ewing Sarcoma [Jul.]Genes 3 (3) (2012) 444–460.
[89] M.K. Thomsen, L. Bakiri, S.C. Hasenfuss, H. Wu, M. Morente, E.F. Wagner, Loss of JUNB/AP-1 promotes invasive prostate cancer [Apr.]Cell Death Differ. 22 (4) (2015) 574–582.
[90] Q.Y. Ning, J.Z. Wu, N. Zang, J. Liang, Y.L. Hu, Z.N. Mo, Key pathways involved in prostate cancer based on gene set enrichment analysis and meta analysis, Genet. Mol. Res. GMR 10 (4) (2011) 3856–3887.
[91] M. Taris, J. Irani, P. Blanchet, L. Multigner, X. Cathelineau, G. Fromont, ERG expression in prostate cancer: the prognostic paradox [Nov.]Prostate 74 (15) (2014) 1481–1487.