BMC Bioinformatics

BioMed Central

Research

Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery

Henry Han

Address: Department of Mathematics and Bioinformatics, Eastern Michigan University, Ypsilanti MI 48109, USA

E-mail: henry.ahan@gmail.com

from The Eighth Asia Pacific Bioinformatics Conference (APBC 2010)
Bangalore, India 18-21 January 2010

Published: 18 January 2010

BMC Bioinformatics 2010, **11**(Suppl 1):S1 doi: 10.1186/1471-2105-11-S1-S1

This article is available from: http://www.biomedcentral.com/1471-2105/11/S1/S1

Abstract

Background: As a novel cancer diagnostic paradigm, mass spectroscopic serum proteomic pattern diagnostics was reported superior to the conventional serologic cancer biomarkers. However, its clinical use is not fully validated yet. An important factor to prevent this young technology to become a mainstream cancer diagnostic paradigm is that robustly identifying cancer molecular patterns from high-dimensional protein expression data is still a challenge in machine learning and oncology research. As a well-established dimension reduction technique, PCA is widely integrated in pattern recognition analysis to discover cancer molecular patterns. However, its global feature selection mechanism prevents it from capturing local features. This may lead to difficulty in achieving high-performance proteomic pattern discovery, because only features interpreting global data behavior are used to train a learning machine.

Methods: In this study, we develop a nonnegative principal component analysis algorithm and present a nonnegative principal component analysis based support vector machine algorithm with sparse coding to conduct a high-performance proteomic pattern classification. Moreover, we also propose a nonnegative principal component analysis based filter-wrapper biomarker capturing algorithm for mass spectral serum profiles.

Results: We demonstrate the superiority of the proposed algorithm by comparison with six peer algorithms on four benchmark datasets. Moreover, we illustrate that nonnegative principal component analysis can be effectively used to capture meaningful biomarkers.

Conclusion: Our analysis suggests that nonnegative principal component analysis effectively conduct local feature selection for mass spectral profiles and contribute to improving sensitivities and specificities in the following classification, and meaningful biomarker discovery.

Background

With the rapid advances in proteomics, mass spectroscopic serum proteomic pattern diagnostics has been appearing as a revolutionary cancer diagnostic paradigm. However, this technology still remains as an important field in clinical research study rather than a clinical routine testing [1]. There are many issues to be resolved to realize the routine clinical testing. Asides from the issues like data reproducibility and quality control [2], one essential issue prevents it going beyond clinical research study sets is that there is no robust supervised learning algorithm to classify proteomic patterns with high sensitivities and specificities. Although there is an urgent need to predict cancer molecular patterns with high accuracies to support clinical decisions, it is still a challenge for oncologists and computational biologists to achieve a high-performance classification due to the special characteristics of mass spectral data.

The mass spectral data has large or even huge dimensionalities. It can be represented by a $n \times m$ matrix, each row of which represents the ion intensity values of all biological samples in investigation at a mass charge ratio (m/z); each column of which represents the ion intensity values of a single biological sample at different m/z values. Each raw data can be called a *pseudo-gene* since it is similar to a gene in a gene expression dataset. Generally, the total number of m/z ratios is in the order of 10^4~10^6 and the total number of biological samples is on the magnitude of hundreds, i.e., the number of variables is much greater than the number of biological samples. Although there are a large number of m/z ratios in a mass spectral profile, only a small number of them have meaningful contributions to the data variations.

Many feature selection algorithms are employed to reduce the protein expression data dimensions, remove noise, and extract meaningful features before further classification or clustering. These algorithms include two-sample t-tests, principal component analysis (PCA), independent component analysis (ICA), nonnegative matrix factorization (NMF) and their different variants [3-5]. PCA may be the most employed among them for its simplicity. It projects data in an orthogonal subspace generated by the eigenvectors of the data covariance matrix. The maximum variance direction-based subspace spanning guarantees the least information loss in the feature selection. However, as a holistic feature selection algorithm, PCA can only capture global features instead of local features [6]. The global and local features contribute to the global and local characteristics of data that are responsible for interpreting the global and local behavior of data respectively. The standard PCA by nature can not extract local features. This not only leads to difficulty in interpreting each principal component (PC) intuitively, but also causes some difficulty in achieving high-performance proteomic pattern discovery, because only the features interpreting global behavior of data are used to train a learning machine (e.g., a support vector machine (SVM) [7]). Since redundant global features may be involved in training, it will decrease the generalization of the learning machine and increase the risk of misclassifications or over-fitting. Moreover, the global data characteristics of a cancer or normal pattern are generally similar because they follow the same protein profiling mechanism. This can be easily verified by the direct visualization of mass spectral profiles. In other words, the local data characteristics play a key role in distinguishing cancer and normal proteomic patterns.

One reason for the holistic mechanism of PCA is that its data representation is not 'purely-additive'. The linear combination to calculate each PC contains both positive and negative weights. The positive and negative weights are likely to partially cancel each other in the linear combination. In fact, weights representing contributions from local features are more likely to be cancelled out because of their frequencies. This partial cancellation may directly lead to missing captures of local features for each loading vector. Another reason for the global nature of PCA is that it lacks some level sparse representation. Each loading vector receives contributions from all input variables in the linear combination. Changes in one variable will inevitably affect all loading vectors globally.

Imposing nonnegativity constraints on PCA can remove the partial cancellations in the linear combinations and make data representation consist of only additive components, i.e., restrict all entries of the input data and each PC as nonnegative items. Adding nonnegativity on PCA is also motivated by proteomic pattern discovery itself. The mass spectral profiling data is generally represented as a positive matrix naturally. It is reasonable to require its corresponding dimension-reduction data to be positive or at least nonnegative to maintain data locality in the feature selection for the sake of pattern discovery. Furthermore, imposing nonnegativity constraints on PCA also leads to the sparse representation of loading vectors.

In this study, we present a nonnegative principal component analysis (NPCA) algorithm and propose a nonnegative principal component analysis based support vector machine algorithm (NPCA-SVM) for high-performance proteomic pattern discovery. We demonstrate its algorithm superiority by comparing it with six peer classification algorithms on four benchmark mass spectral serum datasets. In addition, we present an

effective biomarker discovery approach based on nonnegative principal component analysis.

This work is evolved from our previous naïve work on protein expression classification [8]. However, our current work has the following major advances/differences compared to the previous work. 1. A robust gradient learning scheme is developed for nonnegative principal component analysis and a complete nonnegative principal component analysis based support vector machine algorithm is proposed rigorously. 2. The optimal orthogonal parameter selection method is discussed and an empirical parameter choice approach is given. In addition, we also give a method to set the sparseness control parameter. 3. In addition to including previous three datasets and regenerating all simulation results, we include a new dataset: colorectal data in the experiment. Moreover, a new comparison algorithm: ICA-SVM is included in the simulation. 4. A nonnegative component principal analysis based filter-wrapper biomarker discovery by employing Bayesian t-test based filtering is proposed and its biomarker discovery results for the ovarian and colorectal data are analyzed and visualized. 5. The major global feature selection methods are presented and the two key concepts: global and local features are defined and their impacts in classifications are discussed. 6. We dropped all figures, and tables, and redundant results (e.g. over-fitting analysis about comparison algorithms) from the previous work.

Methods

Nonnegative principal component analysis is an extension of the classic PCA algorithm by imposing it with nonnegativity constraints to capture data locality in the feature selection. Let $X = (x_1, x_2, \ldots x_n)$, $x_i \in \mathbb{R}^d$ be a zero mean dataset, the nonnegative PCA can be formulated as a constrained optimization problem to find maximum variance directions under nonnegative constraints as follows,

$$\max_{U \geq 0} J(U, \alpha) = \frac{1}{2} \left\| U^T X \right\|_F^2 - \alpha \left\| I - U^T U \right\|_F^2 \qquad (1)$$

where $U = [u_1, u_2, \ldots u_k]$, $k \leq d$ is a set of nonnegative PCs. The square Frobenius norm for a matrix A is defined as $\|A\|_F^2 = \sum_{i,j} a_{ij}^2 = trace(AA^T)$. The penalty parameter $\alpha \geq 0$ controls the orthonormal degree of each loading vector. The principal component matrix U is a near-orthonormal nonnegative matrix, i.e., $U^T U \sim I$ Calculating the gradient of the objective function with respect to U, we have the learning scheme: $U(t+1) = U(t) - \eta(t)\nabla_U J(t) / ||\nabla_U J(t)||$, $U \geq 0$ where $\nabla_U J(U, \alpha) = (U^T X)X^T + 4\alpha(I - U^T U)U^T$ and $\eta(t)$ is the t time level iteration step size. We select $\eta(t) = 1$ in the implementation to avoid an expensive trust region search. This is equivalent to finding the local maximum of function $f(u_{sl}) = -\alpha u_{sl}^4 + c_2 u_{sl}^2 + c_1 u_{sl} + c_0$ under the constraints $u_{sl} \geq 0$ on the scalar level ($s = 1, 2 \ldots d$, $l = 1, 2 \ldots n$), where coefficients c_2, c_1 and c_0 are parameters to be determined in the local optimum finding. The final principal component matrix U is a collection of nonnegative roots of function $f(u_{sl})$. Calculating the stationary points and collecting the coefficients of u_{sl} and u_{sl}^2, we obtain the following coefficients c_2, c_1 (The coefficient $c_0 = -k\alpha$ has no contribution to the entries of the PC matrix).

$$c_2 = \frac{1}{2} \sum_{i=1}^{n} x_{si}^2 - \alpha \sum_{j=1, j \neq l}^{k} u_{sj}^2 - 2\alpha \sum_{t=1, t \neq s}^{d} u_{tl}^2 + 2\alpha \qquad (2)$$

$$c_1 = \sum_{i=1}^{n} \sum_{t=1, t \neq s}^{d} x_{si} u_{tl} x_{ti} - 2\alpha \sum_{j=1, j \neq l}^{k} \sum_{t=1, t \neq s}^{d} u_{sj} u_{tl} u_{tj} \qquad (3)$$

The nonnegative principal component analysis complexity is $O(dkn \times N)$, where N is the total iterations needed to meet the algorithm termination threshold $||\nabla_U J(t)|| \leq 10^{-4}$ in the implementation. Other authors also proposed a similar approach to solve a nonlinear optimization problem induced by a nonnegative sparse PCA [9], where two penalty parameters were employed to control the orthonormality and sparseness of the PC matrix. However, an additional sparseness control parameter will increase the risk of algorithmic convergence difficulty with the increasing of the parameter values [10].

We propose a nonnegative principal component analysis based classification algorithm to achieve the high-performance proteomic pattern prediction. The algorithm employs nonnegative principal component analysis to obtain the nonnegative representation of each sample in a low-dimensional, purely-additive subspace spanned by meta-variables. A meta-variable is a linear combination of the intensity values of the pseudo-genes in a mass spectral profile. The nonnegative representation for each sample is denoted as a meta-sample, which is the locality-preserved prototype of the original biological sample with low dimensionalities. Then, a classification algorithm, which is chosen as a support vector machine algorithm (SVM) [7] in this study, is applied to the meta-samples to gain classification information. Given a protein expression training dataset consisting of d biological samples across n pseudo-genes and their label information: $\{x_i, c_i\}_{i=1}^d$, where $X = [x_1, x_2 \ldots x_d]^T$ $x_i \in \mathbb{R}^n$ and $c = [c_1, c_2 \ldots c_d]^T$, $c_i \in \{-1, 1\}$, the NPCA-SVM algorithm finds the meta-samples

$U = [u_1, u_2...u_d]^T$, $U \in \mathbb{R}^{d\times k}$, $k \le d \ll n$, by the described steepest descent method. Then, an optimal separating hyperplane O_h: $w^T u + b = 0$ in $\mathbb{R}^d$ is computed to attain the maximum margin between the '-1' and '1' types of the meta-samples. This is equivalent to solving the following quadratic programming problem in $\mathbb{R}^d$,

$$\begin{aligned} &\min_{w,\xi,b} \frac{1}{2} ||\, w \,||^2 + C\sum_{i=1}^{d} \xi_i \\ &s.t.\ c_i(w^T u_i + b) \ge 1 - \xi_i, i = 1, 2\cdots d, \\ &\quad \xi_i \ge 0 \end{aligned} \tag{4}$$

Given an unknown type sample $x' \in \mathbb{R}^n$, the NPCA-SVM learning machine employs the following decision rule to determine its class type: $f(x') = sign(\sum_{i=1}^{d} \alpha_i c_i k(u_i \bullet u') + b)$, where u_i, $u' \in \mathbb{R}^d$ are the meta-samples of samples x_i, x' computed from nonnegative principal component analysis respectively. The vector $\alpha = [\alpha_1, \alpha_2... \alpha_d] \ge 0$ is the solution of the dual problem of the QP in Eq. (4) and $k(u_i \bullet u')$ is a kernel function for the support vector machine, which maps these meta-samples into a same-dimensional or high-dimensional feature space. We only focus on the linear and *'rbf'* kernels for their popularity [7].

We employ a sparse-coding approach to improve the sparseness for each meta-sample. The sparseness of a nonnegative vector $v = [v_1, v_2...v_n]^T$, $v_i \ge 0$, $i = 1, 2...n$, is defined as a ratio between 0 and 1: $\delta_v = (\sqrt{n} - \|v\|_1 / \|v\|_2)/(\sqrt{n} - 1)$ according to the relationship of two norms [6]. A large sparseness δ_v indicates less number of positive entries in the vector v Extreme cases $\delta_v = 1$ or $\delta_v = 0$ indicate that there is only one entry or all entries are equal in v respectively. The sparse coding of a meta-sample $u_i^T \in \mathbb{R}^{1\times k}$, $i = 1, 2...d$, $k \le d \ll n$ seeks to find a nonnegative vector $v \in \mathbb{R}^{1\times k}$ such that $\|v\|_1 = \|u_i^T\|_1$, $\|v\|_2 = \|u_i^T\|_2$, and δ_v achieving a specified sparseness value. In other words, for each loading vector u_i^T in the nonnegative PC matrix, the nearest nonnegative vector v on behalf of L_1 and L_2 distances is found to achieve a specified sparseness δ_v. It is equivalent to calculating the nonnegative intersection point between a hyperplane $\pi_1 : \sum_{j=1}^{k} v_j = \sum_{j=1}^{k} u_{ij}$ and a hypersphere $\pi_2 : \sum_{j=1}^{k} v_j = \sum_{j=1}^{k} u_{ij}$ such that the sparseness degree $\delta_v = (\sqrt{k} - \|u_i^T\|_1)/(\sqrt{k} - 1)$. Since the traditional approach to find the optimal α is computationally expensive [10], we select $\alpha \propto d$ in practice because of $||U^T U|| = d^2$ in the extreme case where U is the identity matrix, if there is no further sparse coding applied to loading vectors. Otherwise, we select $\alpha \propto \sqrt{d}$. Also because data sparseness is a by-product of the non-negativity constraints in Eq. (1), we usually select the sparseness degree for each nonnegative principal component as $\delta_v \le 0.5$.

We implement the NPCA-SVM algorithm under the 100 trials of 50% holdout cross validations (HOCV), i.e., 100 sets of training and testing data are generated randomly for each dataset. The final classification rate, sensitivity and specificity are the average values of these measures among the 100 trials of classifications. To improve computing efficiency, the PC matrix U in the nonnegative principal component analysis is cached from the previous trial and used as the initial point to compute the next principal component matrix in the computation.

Results

Four serum proteomic datasets: ovarian, ovarian-qaqc (quality assurance/quality control), liver and colorectal are included in this study [11-13], which are generated from three different profiling technologies. Table 1 provides the detailed information about the datasets.

We conducted the following preprocessing for each dataset: baseline correction, smoothing, normalization, peak identification and peak calibration by using Matlab bioinformatics toolbox 3.3. In addition, we applied the standard two-sample t-test to select 3780, 2500, 3000 and 1000 most significant pseudo-genes for the ovarian, ovarian-qaqc, liver, and colorectal data respectively before further classifications. The goal of this basic feature selection is to select approximately $10 \times d$ most significant features for each input dataset $X \in \mathbb{R}^{d\times n}$ before classification. We compared the nonnegative principal component analysis based support vector machine algorithm with the six peers: k-NN, SVM, PCA-SVM, NMF-SVM, ICA-SVM and PCA-LDA algorithms in terms of average classification rates, sensitivities, and specificities under 100 trials of 50% HOCV. Detailed information about the algorithms: LDA, NMF and ICA algorithms can be found in [14,6,5]. In the NPCA-SVM algorithm, we set the orthonormal control $\alpha = 10$, the sparseness for each loading vector $\delta_v = 0.20$, and $k = d - 1$ in the NPCA feature selection due to $n \ll d$.

We showed the average performance of the seven algorithms in terms of average classification rates, sensitivities, specificities, and their corresponding standard deviations in Table 2. We did not include performance of the SVM, PCA-SVM, ICA-SVM and NMF-SVM algorithms under the *'rbf'* kernel, because the first three encountered over-fitting and the last had

Table 1: Four mass spectral serum profiles

Dataset	Technology	#m/z	#Samples
Ovarian	SELDI-TOF low resolution	15142	91 controls + 162 cancers
Ovarian-qaqc	SELDI-TOF high resolution	15000	95 controls + 121 cancers
Liver	SELDI-QqTOF high resolution	6107	176 controls + 181 cancers
Colorectal	MADLI-TOF high resolution	16331	48 controls + 64 cancers

Table 2: Comparisons of the seven algorithms

	Average Classifying rate (%)	Average Sensitivity (%)	Average Specificity (%)
Ovarian			
npca-svm-linear	98.94 ± 00.65	98.35 ± 01.03	99.98 ± 00.24
npca-svm-rbf	99.79 ± 00.35	100.0 ± 00.00	99.42 ± 00.99
svm-linear	99.50 ± 00.83	100.0 ± 00.00	98.63 ± 02.21
pca-svm-linear	99.96 ± 00.26	99.98 ± 00.17	99.93 ± 00.51
nmf-svm-linear	97.41 ± 00.94	99.91 ± 00.31	92.92 ± 02.50
knn	96.53 ± 01.57	99.28 ± 01.34	91.67 ± 03.67
pca-lda	99.67 ± 00.87	99.93 ± 00.38	99.21 ± 02.00
ica-svm-linear	99.99 ± 00.08	99.99 ± 00.12	100.0 ± 00.00
Ovarian-qaqc			
npca-svm-linear	98.70 ± 00.89	98.01 ± 01.94	99.27 ± 00.90
npca-svm-rbf	98.91 ± 00.98	98.11 ± 02.25	99.57 ± 00.82
svm-linear	96.57 ± 01.99	96.16 ± 03.52	96.97 ± 02.19
pca-svm-linear	97.12 ± 01.17	97.14 ± 02.16	97.94 ± 01.57
nmf-svm-linear	88.69 ± 03.47	92.02 ± 05.01	86.24 ± 05.67
knn	90.87 ± 02.92	89.99 ± 04.68	91.82 ± 04.43
pca-lda	97.69 ± 00.65	98.81 ± 01.68	96.99 ± 00.03
ica-svm-linear	97.56 ± 01.45	97.80 ± 02.46	97.41 ± 01.77
Liver			
npca-svm-linear	96.02 ± 01.35	97.68 ± 01.71	94.40 ± 02.22
npca-svm-rbf	97.25 ± 01.30	98.35 ± 01.67	96.20 ± 02.01
svm-linear	91.78 ± 02.27	92.57 ± 03.84	91.04 ± 03.76
pca-svm-linear	90.21 ± 01.99	90.96 ± 03.69	89.57 ± 03.56
nmf-svm-linear	77.76 ± 02.48	84.58 ± 05.14	71.30 ± 05.12
knn	76.48 ± 02.20	72.27 ± 04.60	80.80 ± 04.57
pca-lda	90.08 ± 02.13	91.39 ± 03.53	88.87 ± 03.95
ica-svm-linear	86.61 ± 02.87	87.78 ± 04.55	86.50 ± 04.86
Colorectal			
npca-svm-linear	98.14 ± 01.27	97.93 ± 02.32	98.35 ± 02.00
npca-svm-rbf	97.15 ± 01.07	95.81 ± 02.78	98.18 ± 02.22
svm-linear	96.55 ± 01.87	94.35 ± 03.47	98.26 ± 02.16
pca-svm-linear	93.21 ± 03.38	92.59 ± 04.68	93.89 ± 05.56
nmf-svm-linear	94.73 ± 03.09	92.71 ± 06.14	96.49 ± 03.45
knn	95.05 ± 03.17	96.17 ± 02.91	94.28 ± 05.33
pca-lda	94.05 ± 02.78	94.16 ± 03.74	94.01 ± 04.12
ica-svm-linear	96.04 ± 02.02	94.38 ± 03.66	97.39 ± 02.97

lower performance under the '*rbf*' kernel than the linear kernel. We had the following observations from these results. 1) The NPCA-SVM algorithm achieved obviously leading advantages over the others. Its average specificities for the two ovarian cancer datasets reached 99%+ that was the population screening requirement ratio in the clinical diagnostics. It also achieved 98.35% average specificity for the colorectal data and 98.35% average sensitivity for the liver data. It was the only algorithm among the seven algorithms that achieved consistently leading performances for all datasets. 2) There was no over-fitting associated with the NPCA-SVM algorithm under the '*rbf*' kernel. Alternatively, it achieved exceptional sensitivities and specificities under this kernel. 3) The conventional feature selection algorithms PCA, NMF and ICA generally did not contribute to the improvements of SVM classifications.

Figure 1 compares the average classification rates, sensitivities, specificities and negative target prediction ratios of the NPCA-SVM algorithms with those of the other four algorithms: ICA-SVM, PCA-SVM, SVM and PCA-LDA. It was obvious that the NPCA-SVM algorithm with sparse coding under the '*rbf*' and '*linear*' kernels demonstrated superior or comparable performance compared with the other four algorithms.

Biomarker discovery by nonnegative principal component analysis

In this section, we presented a nonnegative principal component analysis based filter-wrapper biomarker capturing algorithm. The Bayesian two-sample t-test [15] and nonnegative principal component analysis functioned as filters and a SVM classifier worked as a wrapper in this algorithm. Unlike other peak-selection based biomarker capturing methods [12,13], our algorithm could identify which pseudo-genes were more effective in predicting cancer patterns. The NPCA-based biomarker discovery algorithm can be described as follows. For an input mass spectral data $X \in \mathbb{R}^{n \times m}$ with m pseudo-genes and n biological samples, we first filter a potential biomarker set S_b by conducting the two-sample Bayesian t-test, which is a novel approach to evaluate each pseudo-gene according to their differentially expressed levels. The potential biomarker set S_b consists of significantly differentially-expressed pseudo-genes. For each dataset, we select at least the top 1% pseudo-genes with the smallest Bayesian factors, i.e., $|S_b| = \lceil m \times 0.01 \rceil$ to construct S_b. Then, nonnegative principal component analysis (NPCA) is employed to decompose the input data: $X^T \sim PU^T$ For each pseudo-gene, a coefficient τ is used to rank its contribution to all PCs. For example, the coefficient for the i^{th} pseudo-gene is calculated as the weighted sum of the i^{th} row in the nonnegative P matrix: $\tau_i = \log \sum_{j=1}^{\#PC} w_j P_{ij}$, where

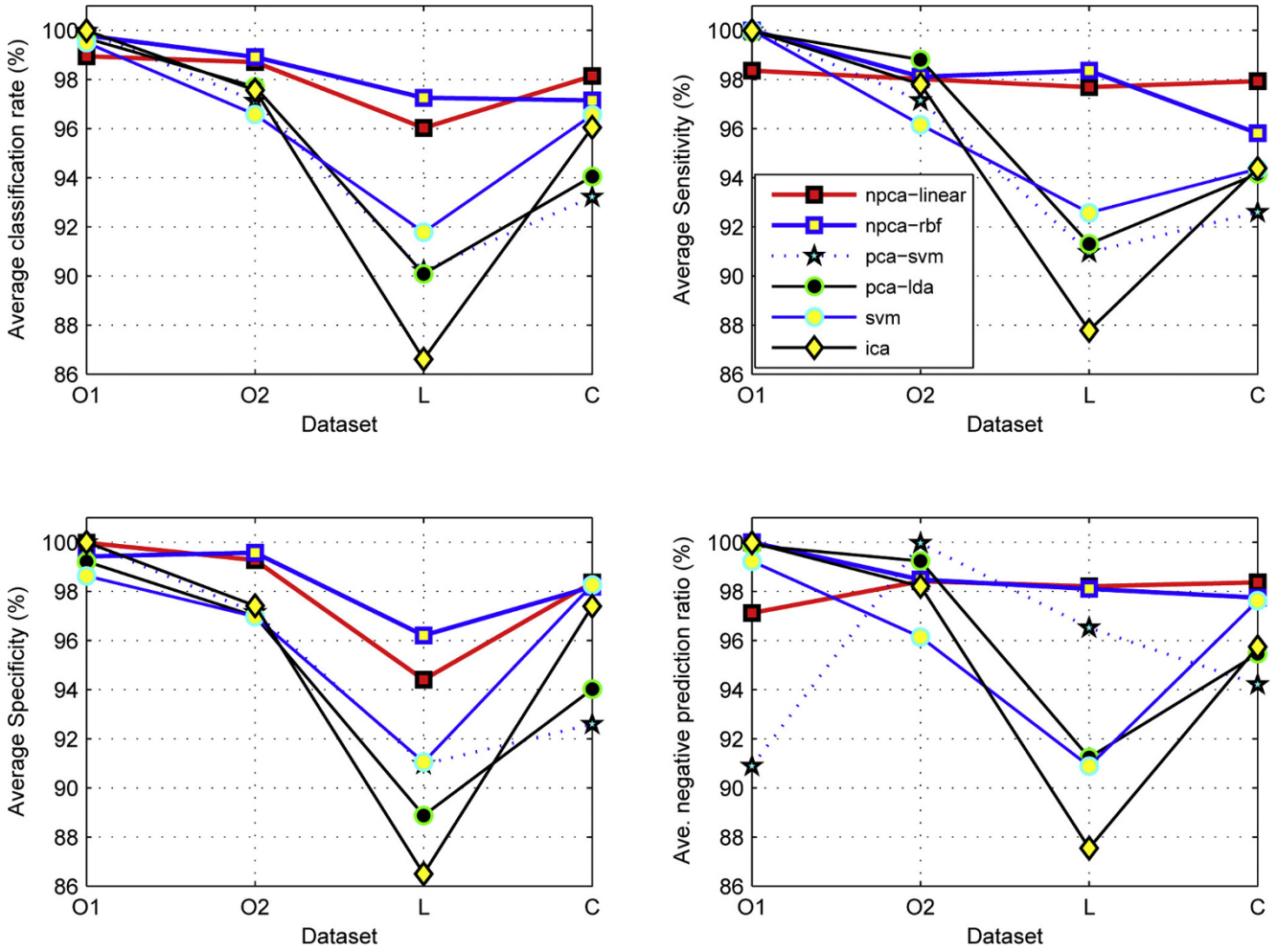

Figure 1

Comparison on the five algorithm performance. Comparison on the five algorithm performance on four datasets: 'O1' (*ovarian*), 'O2' (*ovarian-qaqc*), 'L' (*liver*), and 'C' (*colorectal*). The NPCA-SVM algorithm demonstrated leading performance over the other four algorithms.

$w_j = \lambda_j / \sum_{i=1}^{\#PC} \lambda_i$ is the ratio of variance explained in the j^{th} PC among the total data variance. A large coefficient value of a pseudo-gene indicates it has significant contributions to the PCs.

Each pseudo-gene in S_b is used to train a SVM classifier under the leave-one-out cross validation (LOOCV). The first biomarker g_1 is selected as the pseudo-gene with the highest accuracy. If there is more than one candidate, the pseudo-gene with the largest coefficient in NPCA-ranking will be selected. The potential biomarker set is updated by removing the selected biomarker, i.e, $S_b = S_b - \{g_1\}$. The second biomarker g_2 is selected from the current S_b such that the SVM classifier reaches its maximum classification rate for the combination of g_1 and g_2. If there is more than one candidate, the pseudo-gene with the largest coefficient in the NPCA-ranking will be chosen as g_2. Similarly, S_b is updated as $S_b = S_b - \{g_2\}$. Such a proceeding continues until the SVM classifier achieves the maximum classification accuracy with the fewest biomarkers.

We applied the nonnegative principal component analysis based biomarker capturing algorithm to the colorectal dataset. The potential biomarker set S_b was initialized by 200 pseudo-genes with the smallest Bayes factors. The alpha value in NPCA was set as α = 10 to maintain consistency with the previous classification setting. Table 3 shows the information about three biomarkers discovered for the colorectal data. The total SVM accuracy under the three biomarkers was 98.21% and the corresponding sensitivity and specificity were 95.83% and 100% respectively, which was better than

Table 3: Biomarkers captured for the colorectal data

m/z	Bayes factor	npca-coefficient	SVM ratio (%)
969.1849	7.7881e-031	-1.1205	0.9643
997.5336	1.4236e-026	-1.1571	0.9018
1016.389	7.6644e-013	1.2773	0.8152

the biomarker discovery results obtained in [13]. It was interesting that these biomarkers were not peaks with very large intensity values. The similar results can also be obtained by running the biomarker capture algorithms under the *'rbf'* kernel. The final SVM accuracy also reached 98.21% with three biomarkers at 970.0379, 973.1689 and 997.5336 Da. Interestingly, the biomarkers from different kernels not only shared a same pseudo-gene at 997.5336 Da, but also demonstrated a spatial coherence, i.e., they were neighbors close or very close to each other among 16331 m/z ratios in the data. It indicated that m/z ratios in the downstream interval 960-1030 Da may be more sensitive in discovering cancer patterns than others. Figure 2 visualizes all samples of the colorectal data by using the three biomarkers found under the linear kernel. It is clear that the 112 samples are partitioned into two groups: 64 cancers and 48 controls, and the two types of samples showed significantly different mean and variance values.

Similarly, we applied this algorithm to the ovarian data and obtained 100% predication accuracy (sensitivity: 100%, specificity: 100%) from four biomarkers at m/z ratios: (0.452124, 0.000096, 0.530561, 1.276201) under the linear kernel. Moreover, The SVM classifier also achieved 99.60% accuracy, 100% sensitivity, and 98.90% specificity under the *'rbf'* kernel from three biomarkers at m/z ratios: (0.464762, 0.000096, 0.517053). Also similar to the previous case, the biomarkers discovered under different kernels illustrated spatial proximity and shared same pseudo-genes. Figure 3 visualizes all 253 samples by using the three biomarkers obtained from the *'rbf'* kernel. It was also obviously that cancer and control samples were separated clearly by the three biomarkers.

Discussion

Although nonnegative principal component analysis has overcome the global nature of the standard PCA algorithm, and contributed to the high-performance proteomic pattern prediction and effective biomarker capture, it is an expensive algorithm with a high complexity $O(d^2n \times N)$ compared to the classic PCA algorithm $O(d^3)$ for an input data $X \in \mathbb{R}^{d \times n}$, $d \ll n$. It may require some basic feature selection preprocessing such as the two-sample t-test to avoid a large computing burden for a high-dimensional dataset. On the other hand, since the final PC matrix in nonnegative principal component analysis is computed through a fixed instead of an optimal step size in the iteration, it may miss some local optimal solutions and lead to potential convergence problems. In the following work, we plan to improve nonnegative principal component analysis

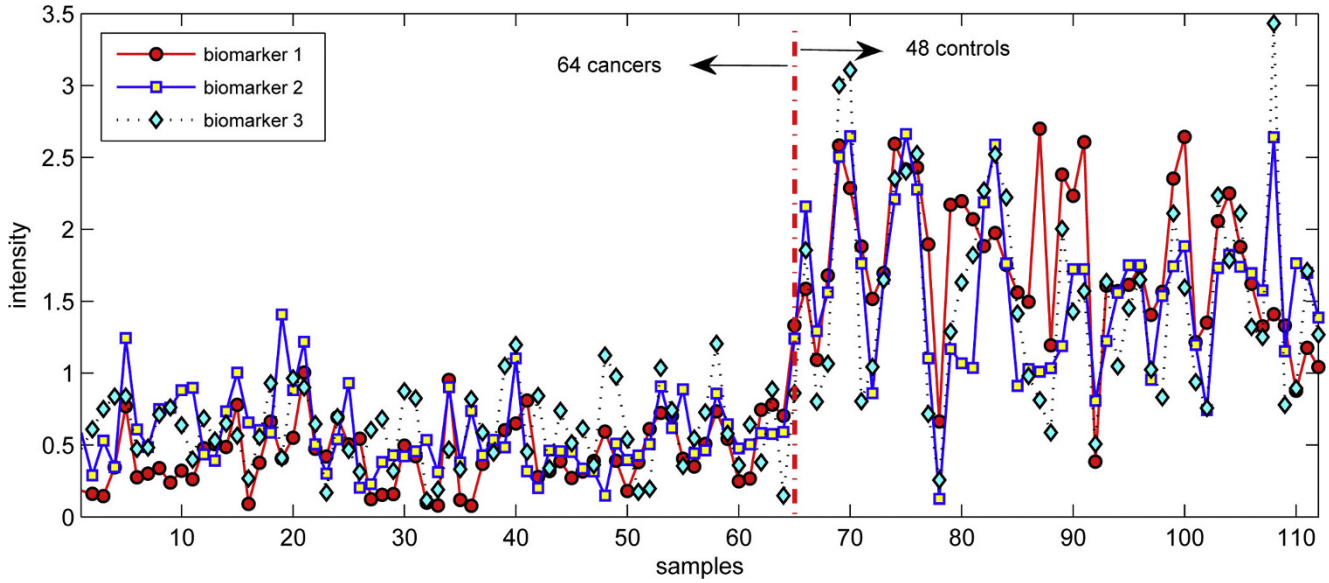

Figure 2
Visualization of the colorectal samples by using three biomarkers. The 48 control and 64 cancer samples are visualized by using the three biomarkers. Two types of samples demonstrated significantly different means and variations.

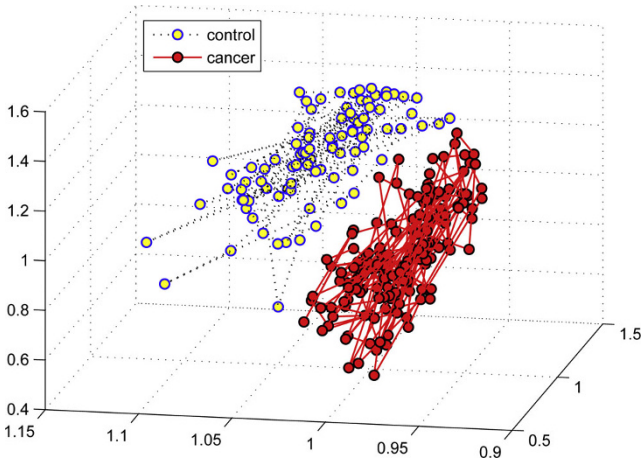

Figure 3
Visualization of the ovarian samples by using three biomarkers. The 253 ovarian samples are visualized by using the three biomarkers. The 91 control and 162 cancer samples are separated into two disjoint clusters.

(NPCA) in the following aspects. (1) We plan to employ the wavelet based multi-resolution approach to overcome the high algorithm complexity in NPCA. A wavelet transform is first employed to decompose an input data into a multi-resolution form. The nonnegative principal component analysis (NPCA) is then employed to extract the local data features from the fine level wavelet transform coefficients, which are relatively low dimensional data compared with the input protein expression data. (2) We will employ a projected-gradient algorithm [10] with a dynamic step size to improve the nonnegative principal component analysis algorithm convergence. As a local feature selection algorithm, nonnegative principal component analysis can be integrated with other state-of-the-art classification and clustering algorithms to develop a family of statistical learning algorithms. For instance, we are interested in combining it with the linear programming SVM algorithm [7] to further explore its potentials in proteomic data pattern prediction. Moreover, we will continue to investigate the applications of the NPCA-SVM algorithms in SNP, CGH array data analysis, and other related topics in future work, in addition to integrating the sparse-coding in our previous NPCA-SVM algorithm developed for gene expression profiles [16].

Conclusion

In this work, we developed a novel feature selection algorithm, nonnegative principal component analysis, and proposed the nonnegative principal component analysis based support vector machine algorithm with sparse coding for high performance proteomic pattern discovery. We demonstrated the superiority of this algorithm by comparing it with other six peer algorithms on four proteomic datasets. In addition, we have designed a NPCA-based filter-wrapper biomarker capturing algorithm and applied it to effectively capture meaningful biomarkers for the colorectal and ovarian data. Our analysis suggests that nonnegative principal component analysis has advantages over the conventional feature selection algorithm such as PCA, ICA, and NMF in local feature selections. Although its algorithmic complexity is higher than that of widely used PCA algorithm, its nature of local feature selection contributes to the high-performance serum proteomic pattern classification and meaningful biomarker discovery.

Competing interests

The author declares that they have no competing interests.

Authors' contributions

HEY did all the work for this paper.

Acknowledgements

This work was partially supported by faculty research fellowship at Eastern Michigan University. The author also thanks Dr. Laxmi Parida and Dr. Asif Javed for their help in improving this manuscript.

This article has been published as part of *BMC Bioinformatics* Volume 11 Supplement 1, 2010: Selected articles from the Eighth Asia-Pacific Bioinformatics Conference (APBC 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/11?issue=S1.

References

1. Petricoin E and Liotta A: **SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer.** *Curr Opin Biotechnol* 2004, **15:**24–30.
2. Coombes KR, Morris JS, Hu J, Edmonson SR and Baggerly KA: **Serum proteomics profiling - a young technology begins to mature.** *Nat Biotechnol* 2005, **23:**291–292.
3. Hauskrecht M, Pelikan R, Malehorn DE, Bigbee WL, Lotze MT, Zeh HJ, Whitcomb DC and Lyons-Weiler J: **Feature Selection for Classification of SELDI-TOF-MS Proteomic Profiles.** *Applied Bioinformatics* 2005, **4(4):**227–246.
4. Yu JS, Ongarello S, Fiedler R, Chen XW, Toffolo G, Cobelli C and Trajanoski Z: **Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data.** *Bioinformatics* 2005, **21(10):**2200–2209.
5. Mantini D, Petrucci F, Del Boccio P, Pieragostino D, Di Nicola M, Lugaresi A, Federici G, Sacchetta P, Di Ilio C and Urbani A: **Independent component analysis for the extraction of reliable protein signal profiles from maldi-tof mass spectra.** *Bioinformatics* 2008, **24(1):**63–70.
6. Hoyer P: **Non-negativematrix factorization with sparseness constraints.** *Journal of Machine Learning Research* 2004, **5:**1457–1469.
7. Vapnik V: *Statistical Learning Theory* John Wiley, New York; 1998.
8. Han X and Scazzero J: **Protein Expression Molecular Pattern Discovery by Nonnegative Principal Component Analysis.** *PRIB* 2008, **5265:**388–399.
9. Zass R and Shashua A: **Nonnegative sparse PCA.** *Neural Information Processing Systems* 2007.
10. Nocedal J and Wright S: *Numerical Optimization* Springer, New York; 1999.
11. **NCI clinical proteomics program.** http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp.

12. Ressom HW, Varghese RS, Abdel-Hamid M, Eissa SA, Saha D, Goldman L, Petricoin EF, Conrads TP, Veenstra TD, Loffredo CA and Goldman R: **Analysis of mass spectral serum profiles for biomarker selection.** *Bioinformatics* 2005, **21(21):**4039–4045.
13. Alexandrov T, Decker J, Mertens B, Deelder AM, Tollenaar RA, Maass P and Thiele H: **Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation.** *Bioinformatics* 2009, **25(5):**643–649.
14. Lilien R and Farid H: **Probabilistic Disease Classification of Expression-dependent Proteomic Data from Mass Spectrometry of Human Serum.** *Journal of Computational Biology* 2003, **10(6):**925–946.
15. Gonen M, Johnson W, Lu Y and Westfall P: **The Bayesian Two-Sample t Test.** *The American Statistician* 2005, **59(3):**252–257.
16. Han X: **Nonnegative Principal Component Analysis for Cancer Molecular Pattern Discovery, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.** *IEEE computer society Digital Library* 2009.