Protein Expression Molecular Pattern Discovery by Nonnegative Principal Component Analysis

Xiaoxu Han[1] and Joseph Scazzero[2]

[1] Department of Mathematics and Bioinformatics Program
[2] Department of Accounting and Finance
Eastern Michigan University, Ypsilanti MI 48197, USA
{xiaoxu.han,jscazzero}@emich.edu

Abstract. Identifying cancer molecular patterns robustly from large dimensional protein expression data not only has significant impacts on clinical ontology, but also presents a challenge for statistical learning. Principal component analysis (PCA) is a widely used feature selection algorithm and generally integrated with classic classification algorithms to conduct cancer molecular pattern discovery. However, its holistic mechanism prevents local data characteristics capture in feature selection. This may lead to the increase of misclassification rates and affect robustness of cancer molecular diagnostics. In this study, we develop a nonnegative principal component analysis (NPCA) algorithm and propose a NPCA-based SVM algorithm with sparse coding in the cancer molecular pattern analysis of proteomics data. We report leading classification results from this novel algorithm in predicting cancer molecular patterns of three benchmark proteomics datasets, under 100 trials of 50% hold-out and leave one out cross validations, by directly comparing its performances with those of the PCA-SVM, NMF-SVM, SVM, k-NN and PCA-LDA classification algorithms with respect to classification rates, sensitivities and specificities. Our algorithm also overcomes the overfitting problem in the SVM and PCA-SVM classifications and provides exceptional sensitivities and specificities.

Keywords: Nonnegative principle component analysis, sparse coding, support vector machine (SVM).

1 Introduction

Molecular diagnostics has been challenging traditional cancer diagnostics in oncology by generating gene/protein expression data from a patient's tissue, serum or plasma samples through the DNA and protein array technologies. In clinical oncology, the gene/protein expressions are molecular patterns of cancers, reflecting gene/protein activity patterns in different types of cancerous or precancerous cells. However, robustly classifying cancer molecular patterns to support clinical decision making in early cancer diagnostics is still a challenge because of the special characteristics of gene and protein expression data. In this study, we focus on the mass spectrometry based protein expression data (MS data).

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 388–399, 2008.

Similar to general gene expression data, MS data can have large or even huge dimensionalities. It can be represented by a n×m matrix, each row of which represents the intensity values of a measured data point at a mass charge ratio (m/z) across different biological samples; each column of which represents the intensity values of all measured data points at different m/z values in a sample. Generally, the total number of measured data points is in the order of $10^5 \sim 10^6$ and the total number of biological samples is in 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 variables in these data, only a small set of variables have meaningful contributions to the data variations. Actually, these high-dimensional data are not noise-free. This is because the raw data contains systematic noise and the preprocessing algorithms can not remove it completely.

1.1 Principal Component Analysis Is a Holistic Feature Selection Algorithm

Many feature selection algorithms are employed to reduce protein expression data dimensions and decrease data noise before further classification or clustering [1,2]. Principal component analysis (PCA) is a commonly used approach among them [3,4]. It projects data in an orthogonal subspace generated by the eigenvectors of the data covariance or correlation matrix. The data representation in the subspace is uncorrelated and 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 the global characteristics of data instead of local characteristics of data. This leads to difficulty in interpreting each principal component (PC) intuitively, because each PC contains some levels of global characteristics of data. In the cancer pattern analysis of proteomics data, the holistic mechanism will prevent the following supervised/unsupervised learning algorithm from capturing the local behaviors of proteomics data in the clustering/classification. This would lead to the increase of misclassification rates and finally affect the robustness of the cancer molecular diagnostics.

One main reason for the holistic mechanism of the PCA is that data representation in the classic PCA is not '*purely additive*', i.e. the linear combination in the PCA contains both positive and negative weights and each PC consists of both negative and positive entries. The positive and negative weights are likely to cancel each other partially in the data representation. In fact, it is more likely that weights contributing from local features are partially cancelled out because of their frequencies. This directly leads to the holistic feature selection characteristics in the PCA.

Imposing nonnegative constraints on the PCA can remove the likelihood of the partial cancellation and make data representation consists of only additive components. In addition, it also contributes to sparse data representation. In the context of feature selection, adding nonnegative constraints on the PCA can improve the data locality in feature selection and make the data latent structure explicit.

Adding nonnegativity on the PCA is also motivated by the cancer molecular pattern discovery itself, i.e., protein expression data generally are represented as positive or nonnegative matrices naturally or after simple preprocessing. It is reasonable to require their corresponding dimension reduction data to be positive or at least nonnegative to maintain data locality in order to catch more subtle or local behaviors in the following clustering or classification-based pattern discovery.

In this study, we present the nonnegative principal component analysis (NPCA) algorithm and demonstrate the superiority of the NPCA-based SVM classification algorithm (NPCA-SVM) with sparse coding, for three benchmark mass spectral serum datasets, by directly comparing it with five other similar classification algorithms, i.e., SVM, PCA-SVM, NMF-SVM, k-NN and PCA-LDA. This paper is organized as follows. Section 2 presents the nonnegative principal component analysis (NPCA) and NPCA-based SVM classification. Section 3 gives the experimental results of the NPCA-based SVM algorithm with sparse coding under 100 trials of 50% holdout cross validations for each dataset. It also compares the NPCA-SVM algorithm with the other five classification algorithms for the same training and test datasets. Finally, Section 4 concludes the paper.

2 Nonnegative PCA-Based Classification

Nonnegative PCA can be viewed as an extension of classic PCA by imposing PCA with nonnegativity constraints to capture data locality in the feature selection. Let $X = (x_1, x_2, \cdots x_n)$, $x_i \in \Re^d$, be a zero mean dataset, i.e., $\sum_{i=1}^{n} x_i = 0$. Then, the nonnegative PCA can be formulated as a constrained optimization problem to find maximum variance directions under nonnegative constraints as follows.

$$\begin{aligned} &\max\, J(U) = \frac{1}{2}\left\|U^T X\right\|_F^2, \quad s.t. \\ &U^T U = I,\ U \geq 0 \end{aligned} \tag{1}$$

where $U = [u_1, u_2, \cdots u_k]$, $k \leq d$, is a set of nonnegative PCs. The square Frobenius norm for a matrix A is defined as $\left\|A\right\|_F^2 = \sum_{i,j} a_{ij}^2 = trace(AA^T)$.

In fact, the rigorous orthonormal constraint under non-negativity is too strict for the practical cancer molecular pattern analysis, because it requires only one nonnegative entry in each column of U. The quadratic programming problem with the orthonormal-nonnegativity condition can be further relaxed as

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

where $\alpha \geq 0$ is a parameter to control the orthonormal degree of each column of U. After relaxation, matrix U is a near-orthonormal nonnegative matrix, i.e., $U^T U \sim I$. Computing the gradient of the objective function with respective to U , we have

$$U(t+1) = U(t) - \eta(t)\nabla_U J(t), \qquad U \geq 0 \tag{3}$$

where $\nabla_U J(U,\alpha) = (U^T X)X^T + 2\alpha(I - U^T U)U$ and $\eta(t)$ is the iteration step size in the t time level iteration. For convenience, we select the step size in the iteration as 1. In fact, this is equivalent to finding the local maximum of a function $f(u_{sl})$ under the conditions: $u_{sl} \geq 0$, $s = 1, 2 \cdots d; l = 1, 2 \cdots n$, in the scalar level.

$$\max_{u_{sl} \geq 0} f(u_{sl}) = -\alpha u_{sl}^4 + c_2 u_{sl}^2 + c_1 u_{sl} + c_0 \tag{4}$$

where c_2 and c_1 are the coefficients of the u_{sl}^2 and u_{sl} ; c_0 is the sum of the constant items independent of u_{sl} . The local maximum finding of the equation (4) is actually a set of cubic polynomial nonnegative root finding. Computing the stationary points for the scalar function $f(u_{sl})$, we have a set of cubic function root finding problems: $p(u_{sl}) = df(u_{sl}) / du_{sl} = 0$ (see the appendix for details). The final U matrix is a set of nonnegative roots of the equation. By collecting the coefficients of u_{sl} and u_{sl}^2 , we have

$$c_2 = \frac{1}{2}\sum_{i=1}^{n} x_{si}^2 - \alpha \sum_{j=1, j\neq l}^{k} u_{sj}^2 - 2\alpha \sum_{t=1, t\neq s}^{d} u_{tl}^2 + 2\alpha \tag{5}$$

$$c_1 = \sum_{i=1}^{n} \sum_{t=1, t\neq s}^{d} x_{si} u_{tl} x_{ti} - 2\alpha \sum_{j=1, j\neq l}^{k} \sum_{t=1, t\neq s}^{d} u_{sj} u_{tl} u_{tj} \tag{6}$$

Actually, the constant term $c_0 = -k\alpha$ does not affect the entries of the matrix U. Only coefficients c_1 and c_2 are involved in the nonnegative root finding. The algorithm complexity of NPCA is $O(dkn \times N)$, where N is the total iteration number used in the algorithm. The detailed parameter derivations about equation 5 and 6 can be found in the appendix. Some authors also proposed a similar approach to solve the nonlinear optimization problem induced by a nonnegative sparse PCA [5]. However, their results lack technical soundness in the key parameter derivations.

2.1 NPCA-Based Classifications

The NPCA-based cancer molecular pattern classification employs the nonnegative principal component analysis (NPCA) to obtain a nonnegative representation of each sample in a low-dimensional, purely additive subspace spanned by the meta-variables first. A meta-variable is a linear combination of the intensity values of the measured data points for the MS data. The nonnegative representation for each sample is called a meta-sample, which is the prototype of the original sample with small dimensionalities. Then, a classification algorithm π_a , which is the SVM algorithm in this study, is applied to the meta-samples to gain classification information.

Theoretically, NPCA-based classification is rooted from a special nonnegative matrix factorization (NMF) [6] that we propose in this study: the nonnegative principal component induced NMF. We brief the principle of the NPCA-induced NMF as follows.

Let $X \in \Re^{d\times n}, d \ll n$, be a nonnegative matrix, which is a protein expression dataset with d number of samples for n number of measured points. Let $U \in \Re^{d\times d}$ be the nonnegative PCs, a near-orthogonal matrix for X before any further dimension selection. Projecting X^T into the purely additive subspace generated by U, we obtain the nonnegative projection $X^T U = P$. Alternatively, considering the PC matrix U is a near-orthogonal matrix, we can view it as an orthogonal matrix to decompose the data matrix, i.e., $X^T \sim PU^T$, where the nonnegative matrix P is equivalent to the basis matrix W and matrix U^T is equivalent to the feature matrix H in the classic NMF: $X \sim WH$. Similarly, the decomposition rank r in the NMF is the corresponding selected dimensionality in the nonnegative principal component analysis.

The NPCA-induced NMF can be also explained as follows. Each row of U is the corresponding meta-sample of each sample of X in the meta-variable space: $X_i^T \sim PU_i^T$. The meta-variable space is a subspace generated by columns of the basis matrix P, where each column/basis is a meta-variable. The meta-variable space is a purely additive space where each variable can be represented as the nonnegative linear combination of meta-variables as shown below.

$$X_i^T = \sum_{j=1}^{r} U_{ij}^T P_j, 1 \le r \le d \tag{7}$$

Based on the observation that proteomics data are nonnegative data or can be converted to corresponding nonnegative data easily, we have the NPCA-based SVM classification algorithm for proteomics data, i.e., starting from the NPCA-induced NMF for the protein expression dataset X, we input the corresponding normalized meta-samples $U = U / \|U\|_2$ to the SVM algorithm to conduct classification.

2.2 Sparse-Coding

To improve the generality of the NPCA-SVM classification algorithm, we conduct a sparse coding for the nonnegative PC matrix U. The sparseness of a nonnegative vector v with n tuples is a ratio between 0 and 1, which is defined in the equation (8) according to the relationship of two norms [7].

$$sparseness(v) = \frac{\sqrt{n} - \|v\|_1 / \|v\|_2}{\sqrt{n} - 1} \tag{8}$$

The sparse coding of the nonnegative PC matrix U finds the corresponding nonnegative vector satisfying the specified sparseness degree for each row $U_i^T, i = 1, 2 \cdots k$. In other words, for each row vector $x \ge 0$, the nearest vector $v \ge 0$ in the Euclidean sense is found that achieves a specified sparseness s. For convenience, we first normalize the nonnegative vector x such that $\|x\|_2 = 1$ before the sparse coding. Then, we project x into the hyperplane: $\sum v_i = \|x\|_1$ and compute the nonnegative intersection point with the hypersphere $\sum v_i^2 = 1$ under the condition $s = (\sqrt{n} - \|x\|_1) / (\sqrt{n} - 1)$ in real time to finish its sparse coding.

2.3 Cross Validations

Since different training sets will affect classification results of a classification algorithm, we conducted the NPCA-SVM classification under the *50% holdout* cross validation 100 times, i.e., 100 sets of training and test datasets are generated randomly for each cancer dataset in the classification, to evaluate the expected classification performances. To improve computing efficiency, the PC matrix *U* in the nonnegative principal component analysis (NPCA) is cached from the previous trial and used as the initial point to compute the next principal component matrix in the computation.

3 Experimental Results

Our experimental data consists of three mass spectral serum profiles: *Ovarian*, *Ovarian-qaqc* (quality assurance/quality control) and *Liver* [8,9]. These datasets include one low resolution dataset and two high resolution datasets. They are generated from the Surface-Enhanced Laser Desorption/Ionization Time-Of-Flight (SELDI-TOF) and SELDI-QaTOF (a hybrid quadrupole time-of-flight mass spectrometry) technologies respectively. The detailed information about the datasets is given in the Table 1.

Table 1. Three Mass Spectral Serum Profiles.

Dataset	Data type	#M/Z	#Samples
	SELDI-TOF		91 controls
Ovarian	Low resolution	15142	162 cancers
	SELDI-TOF		95 controls
Ovarian-qaqc	High resolution	15000	121 cancers
Liver	SELDI-QqTOF		181 controls
	High resolution	6710	176 cancers

3.1 Preprocessing and Basic Feature Selection

We conducted the basic preprocessing steps for each mass spectrometry dataset: spectrum calibration, baseline correction, smoothing, peak identification, intensity normalization, and peak alignments. In addition, we employed the two-sided t-test to conduct basic feature selection for the three proteomics datasets before classifications. After the basic feature selection, *3780*, *2000* and *3000* most significant features are selected for the 1st, 2nd and 3rd dataset respectively, before further classifications.

3.2 Classifications

We compared the classification results from the NPCA-SVM algorithm under the sparse coding (α=10, sparseness=0.20) with the PCA-SVM and SVM algorithm under

Table 2. Average classification performance of three algorithms

	Average Sensitivity	Average Specificity	Average Classifying rates
Ovarian			
npca-svm-linear	98.35±1.03	99.98±0.24	98.94±0.65
npca-svm-rbf	100.0±0.0	99.42±0.99	99.79±0.35
svm-linear	100.0±0.0	98.63±2.21	99.50±0.83
svm-rbf	100.0±0.0	0.0±0.0	64.13±2.88
pca-svm-linear	99.98±0.17	99.93±0.51	99.96±0.26
pca-svm-rbf	100.0±0.0	0.0±0.0	64.13±2.88
Ovarian-qaqc			
npca-svm-linear	98.01±1.94	99.27±0.90	98.70±0.89
npca-svm-rbf	98.11±2.25	99.57±0.82	98.91±0.98
svm-linear	96.16±3.52	96.97±2.19	96.57±1.99
svm-rbf	97.00±17.18	3.00±17.18	54.92±44.8
pca-svm-linear	97.14±2.16	97.94±1.57	97.12±1.17
pca-svm-rbf	3.20±17.22	96.80±17.22	54.95±44.7
Liver			
npca-svm-linear	97.68±1.71	94.40±2.22	96.02±1.35
npca-svm-rbf	98.35±1.67	96.20±2.01	97.25±1.30
svm-linear	92.57±3.84	91.04±3.76	91.78±2.27
svm-rbf	38.00±48.78	62.00±48.78	47.92±2.00
pca-svm-linear	90.96±3.69	89.57±3.56	90.21±1.99
pca-svm-rbf	38.00±48.78	62.00±48.78	47.92±2.00

linear and Gaussian kernels, for each proteomics dataset under the *same* 100 sets of training and test data (trials). The 100 trials of training/test data for each dataset are generated under the 50% holdout cross validations. The average classification rates, sensitivities and specificities and their corresponding standard deviations from each classification algorithm are given in the Table 2.

From the classification results, we can make the following observations. 1. It is clear that the PCA-SVM, SVM classification algorithms suffer from overfitting under a Gaussian ('*rbf*') kernel. This is due to the complementary results of the sensitivities and specificities for the three proteomics datasets. For instance, under a '*rbf*' kernel, the PCA-SVM and SVM classification for the *Ovarian* cancer dataset can only classify the positive (cancer) targets. Both of them have an average classification rate of 64.13%, which is approximately the ratio of the positive targets among the total samples: 162/253=64.03%. 2. There is no overfitting problem under a '*rbf*' kernel, for the NPCA-SVM algorithm with sparse coding. On the other hand, the NPCA-SVM has the best classification performance among all the algorithms for the three protein expression datasets. 3. Under a linear kernel, the PCA-SVM achieves slightly better or comparable results than the SVM for the two ovarian datasets. Similarly, the SVM classification also has slightly better average classification rates, sensitivities and specificities than the PCA-SVM for the *Liver* dataset. Thus, we can say that their classification performances for the experimental datasets are comparable. 4. The classification results of the NPCA-SVM have leading advantages for the three datasets, compared with those of the PCA-SVM and SVM classifications. Actually, the average specificities for the two ovarian cancer datasets reach *99%+* under the NPCA-SVM

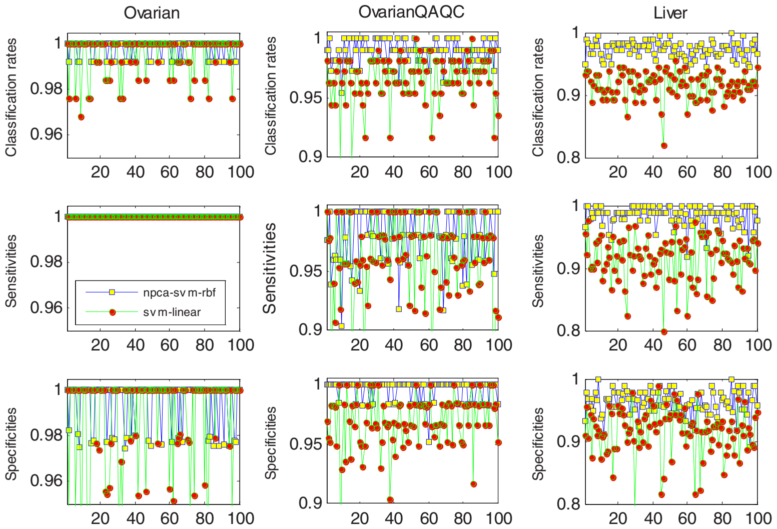

Fig. 1. Comparison on the SVM classification under a linear kernel and the NPCA-SVM classification under a '*rbf*' kernel. For the first dataset, NPCA-SVM ('rbf') has slightly better classification performance than the SVM ('linear'). For the 2nd and 3rd datasets, the NPCA-SVM ('rbf') classification has obvious leading advantages over the SVM ('linear') classification.

classification. The 99%+ specificity level is the population screening requirement ratio in general clinical diagnostics. Figure 1 shows the performances of the SVM algorithm under a linear kernel and the NPCA-SVM algorithm under a '*rbf*' kernel for three datasets.

3.3 Compare Classification Results with Those of Other Algorithms

We also compare the classification performance of the NPCA-SVM algorithm with those of three other classification algorithms: k-NN, PCA-LDA and NMF-SVM. For each dataset, we still use the previous 100 trials of training/test datasets generated under the 50% holdout cross validations in the classifications.

The k-NN and PCA-LDA are widely used algorithms in proteomics data classifications. The k-NN is a simple Bayesian inference method. It determines the class type of a sample based on the class belonging to its nearest neighbors, which are measured by correlation, Euclidean or other distances. The PCA-LDA conducts the PCA processing for the training samples and projects the test samples in the subspace spanned by the principal components of the training data. Then, linear discriminant analysis (LDA) is used to classify the projections of the test data, which is equivalent to solving a generalized eigenvalue problem [10].

The NMF-SVM algorithm is similar to the NPCA-SVM classification algorithm. It conducts the SVM classification for the meta-samples of a proteomics dataset, which are the columns of the feature matrix H in the NMF. We briefly describe the

NMF-SVM algorithm as follows. The NMF-SVM classification decomposes the nonnegative protein expression data $X \in \Re^{n\times m}$ into the product of two nonnegative matrices: $X \sim WH$, under a rank r with the least reconstruction error in the Euclidean sense. The matrix $W \in \Re^{n\times r}$ is termed a basis matrix. Its column space sets up a new coordinate system for *X*. The matrix $H \in \Re^{r\times m}$ is called a feature matrix. It stores the new coordinate values for each variable of *X* in the new space. Then, the SVM algorithm is employed to classify the corresponding meta-sample of each sample in the protein expression matrix *X***.** Each meta-sample is just the corresponding column in the feature matrix *H*.

In the k-NN, the distance measures are chosen as the correlation and Euclidean distances. The number of nearest neighbors for each test sample is selected from *2* to *7*; In the NMF-SVM, the matrix decomposition rank in the NMF is selected from *2* to *18* for each dataset under the linear and Gaussian kernel. The final average classification rate for each dataset under the k-NN and NMF-SVM is selected as the best average classification rate of the 100 trials of training and test data among all cases. Table 3 shows the expected classification rates, sensitivities and specificities of the three algorithms and corresponding standard deviations for each of the three datasets, under the 100 trials of training/test datasets generated from 50% holdout cross validations.

Actually, we have found the k-NN algorithm achieves better classification performances under the correlation distance than the Euclidean distance. The NMF-SVM algorithm achieves better classification performances under the correlation distance than the Euclidean distance. For the three protein expression datasets, the NMF-SVM and k-NN classification results are comparable. However, it is obvious that the PCA-LDA algorithm achieves the best performances among the three algorithms.

Table 3. Average classification performances of the NMF-SVM, k-NN, PCA-LDA algorithms

	Average Sensitivity	Average Specificity	Average Classifying rates
Ovarian			
nmf-svm-linear	99.91±0.31	92.92±2.50	97.41±0.94
nmf-svm-rbf	96.27±3.35	90.83±4.48	94.29±2.72
knn-correlation	99.28±1.34	91.67±3.67	96.53±1.57
knn-euclidean	99.58±0.76	90.77±3.19	96.41±1.29
pca-lda	99.93±0.38	99.21±2.00	99.67±0.87
Ovarian-qaqc			
nmf-svm-linear	92.02±5.01	86.24±5.67	88.69±3.47
nmf-svm-rbf	76.18±9.12	78.57±6.38	77.30±3.67
knn-correlation	89.99±4.68	91.82±4.43	90.87±2.92
knn-euclidean	82.03±6.86	87.71±5.86	85.03±3.71
pca-lda	98.81±1.68	96.99±0.03	97.69±0.65
Liver			
nmf-svm-linear	84.58±5.14	71.30±5.12	77.76±2.48
nmf-svm-rbf	80.69±6.01	69.21±5.57	74.79±2.25
knn-correlation	72.27±4.60	80.80±4.57	76.48±2.20
knn-euclidean	77.04±5.81	75.38±5.33	76.11±2.51
pca-lda	91.39±5.81	88.87±3.95	90.08±2.13

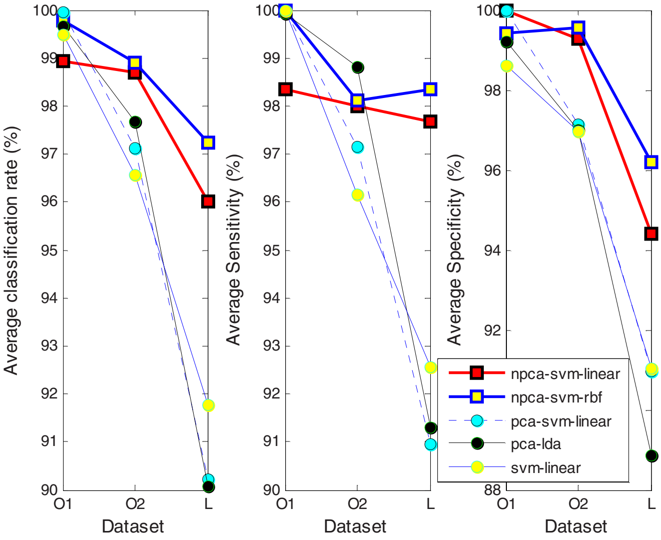

Fig. 2. Comparison on the classification performances of four algorithms for three proteomics datasets: 'O1' (*Ovarian*), 'O2' (*Ovarian-qaqc*), 'L' (*Liver*). The NPCA-SVM has the best performances among the four algorithms with respect to average classification rates, sensitivities and specificities for all three datasets, though the PCA-LDA and SVM algorithms both achieve comparable classification performances for the first ovarian dataset.

Figure 2 compares the classification performances of the NPCA-SVM algorithm with sparse coding to those of the PCA-LDA, PCA-SVM and SVM with respect to average classification rates, sensitivities and specificities. For all three proteomics datasets, it is obvious that the NPCA-SVM algorithm with sparse coding under the '*rbf*' and '*linear*' kernel has generally achieved the best or second-best classification results among all these algorithms respectively. The NPCA-SVM algorithm with sparse coding also gives the same leading results under the leave-one-out cross validation (*LOOCV)* according to our experimental results.

4 Conclusion and Discussions

In this study, we develop a novel feature selection algorithm: nonnegative principal component analysis (NPCA) and propose the NPCA-SVM algorithm under sparse coding for the cancer molecular pattern discovery of protein expression data. We also demonstrate the superiority of this novel algorithm over the NMF/PCA-SVM, SVM, k-NN and PCA-LDA classification algorithms for three benchmark proteomics datasets. Our algorithm also overcomes the overfitting problem of the SVM and PCA-SVM classifications under a Gaussian kernel.

With nonnegative principal component analysis, we can develop a family of NPCA-based statistical learning algorithms by applying NPCA as a feature selection algorithm before a classification or clustering algorithm, e.g., NPCA-based Fisher discriminant analysis (NPCA-FDA), NPCA-based K-means or hierarchical clustering. In future work, we plan to investigate the NPCA-based classifications, such as NPCA-FDA, NPCA-SVM in the protein folding, gene, microRNA profiles data classification and biomarker discovery.

References

1. Han, X.: Cancer molecular pattern discovery by subspace consensus kernel classification, Computational Systems Bioinformatics, In: Proceedings of the Conference CSB 2007, vol. 6, pp. 55–65 (2007)
2. Hauskrecht, H., et al.: Feature Selection for Classification of SELDI-TOF-MS Proteomic Profiles. Applied Bioinformatics 4(4), 227–246 (2005)
3. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. Journal of Computational and Graphical Statistics 15(2), 262–286 (2006)
4. D'Aspremont, A., Ghaout, L., Jordan, M., Lanckriet, G.: A direct formulation for sparse PCA using Semidefinite Programming. SIAM Review 49(3), 434–448 (2007)
5. Zass, R. and Shashua, A.: Nonnegative sparse PCA, Neural Information and Processing Systems (NIPS) (2006)
6. Lee, D.D., Sebastian Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
7. Hoyer, P.O.: Hoyer: Non-negativematrix factorization with sparseness constraints. Journal of Machine Learning Research 5, 1457–1469 (2004)
8. National Center Institute Center for Cancer Research Clinical Proteomics Program, http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp
9. Ressom, H., Varghese, R., Saha, D., Orvisky, R., et al.: Analysis of mass spectral serum profiles for biomarker selection. Bioinformatics 21(21), 4039–4045 (2005)
10. Lilien, R., Farid, H.: Probabilistic Disease Classification of Expression-dependent Proteomic Data from Mass Spectrometry of Human Serum. Journal of Computational Biology 10(6), 925–946 (2003)

Appendix: Nonnegative Principal Component Analysis Parameter Derivations

In this section, we give the detailed parameter derivation for equations (5) and (6). Computing the stationary points for the objective function $f(u_{sl})$ in equation (4), we have a cubic root finding problem. The final U matrix consists of a set of nonnegative roots of equation (9).

$$p(u_{sl}) = df(u_{sl}) / du_{sl} = -4\alpha u_{sl}^3 + 2c_2 u_{sl} + c_1 = 0 \tag{9}$$

We derive coefficients c_2, c_1 in equation (9) as follows. For convenience, we rewrite the terms in equation (2) as $\left\| I - U^T U \right\|_F^2 = L_1 + L_2$, where L_1 and L_2 represent the contributions from the diagonal elements and non-diagonal elements of the matrix $I-U^TU$ to its Frobenius norm respectively, where $L_1=\sum_{l=1}^{k}(1-u_l^Tu_l)^2$ and $L_2=\sum_{l=1}^{k}\sum_{j=1,j\neq l}^{k}(u_l^Tu_j)^2$. Similarly, we also set $L_3=\left\|U^TX\right\|_F^2=\sum_{l=1}^{k}\sum_{j=1}^{n}(u_l^Tx_j)^2$ and compute the parameters c_2 , c_1 by checking the coefficients of u_{sl} and u_{sl}^2 in equation (2). From the equation, we have following results:

$$L_1=k-2\sum_{l=1}^{k}\sum_{s=1}^{d}u_{sl}^2+\sum_{l=1}^{k}\sum_{s=1}^{d}u_{sl}^4+2\sum_{l=1}^{k}\sum_{s=1}^{d}\sum_{t=1,t\neq s}^{d}u_{sl}^2u_{tl}^2 \tag{10}$$

$$L_2=\sum_{s=1}^{d}\sum_{l=1}^{k}\sum_{j=1,l\neq j}^{k}u_{sl}^2u_{sj}^2+2\sum_{s=1}^{d}\sum_{l=1}^{k}\sum_{j=1,l\neq j}^{k}\sum_{t=1,t\neq s}^{d}u_{sl}u_{sj}u_{tl}u_{tj} \tag{11}$$

$$L_3=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{s=1}^{d}u_{sl}^2x_{si}^2+2\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{s=1}^{d}\sum_{t=1,t\neq s}^{d}u_{sl}x_{si}u_{tl}x_{ti} \tag{12}$$

By substituting for the coefficients of u_{sl} and u_{sl}^2 , we have

$$c_2=\frac{1}{2}\sum_{i=1}^{n}x_{si}^2-\alpha\sum_{j=1,j\neq l}^{k}u_{sj}^2-2\alpha\sum_{t=1,t\neq s}^{d}u_{tl}^2+2\alpha \tag{13}$$

$$c_1=\sum_{i=1}^{n}\sum_{t=1,t\neq s}^{d}x_{si}u_{tl}x_{ti}-2\alpha\sum_{j=1,j\neq l}^{k}\sum_{t=1,t\neq s}^{d}u_{sj}u_{tl}u_{tj} \tag{14}$$