BMC Bioinformatics

BioMed Central

Proceedings

Open Access

A kernel-based approach for detecting outliers of high-dimensional biological data

Jung Hun Oh and Jean Gao*

Address: Department of Computer Science and Engineering, The University of Texas, Arlington, Texas, USA

Email: Jung Hun Oh - jung.oh@uta.edu; Jean Gao* - gao@uta.edu

* Corresponding author

from IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2008
Philadelphia, PA, USA. 3–5 November 2008

Published: 29 April 2009

BMC Bioinformatics 2009, **10**(Suppl 4):S7 doi:10.1186/1471-2105-10-S4-S7

This article is available from: http://www.biomedcentral.com/1471-2105/10/S4/S7

Abstract

Background: In many cases biomedical data sets contain outliers that make it difficult to achieve reliable knowledge discovery. Data analysis without removing outliers could lead to wrong results and provide misleading information.

Results: We propose a new outlier detection method based on Kullback-Leibler (KL) divergence. The original concept of KL divergence was designed as a measure of distance between two distributions. Stemming from that, we extend it to biological sample outlier detection by forming sample sets composed of nearest neighbors. KL divergence is defined between two sample sets with and without the test sample. To handle the non-linearity of sample distribution, original data is mapped into a higher feature space. We address the singularity problem due to small sample size during KL divergence calculation. Kernel functions are applied to avoid direct use of mapping functions. The performance of the proposed method is demonstrated on a synthetic data set, two public microarray data sets, and a mass spectrometry data set for liver cancer study. Comparative studies with Mahalanobis distance based method and one-class support vector machine (SVM) are reported showing that the proposed method performs better in finding outliers.

Conclusion: Our idea was derived from Markov blanket algorithm that is a feature selection method based on KL divergence. That is, while Markov blanket algorithm removes redundant and irrelevant features, our proposed method detects outliers. Compared to other algorithms, our proposed method shows better or comparable performance for small sample and high-dimensional biological data. This indicates that the proposed method can be used to detect outliers in biological data sets.

Background

Outlier detection is an active research area that has many applications such as network intrusion detection [1], fraud detection [2] and biomedical data analysis [3]. In particular, outliers caused from instrument error or human error in the biomedical data analysis such as biomarker selection and disease diagnosis could deeply degrade the performance of the data analysis. Therefore, prior to the analysis, during preprocessing it is imperative to remove outliers to prevent wrong results. To detect such

anomalous observations from normal ones, data mining techniques are widely used.

Outlier detection has been studied by researchers using a diversity of approaches. Statistical methods often view objects that are located relatively far from the center of the data distribution as outliers. Several distance measures were implemented. The Mahalanobis distance is the most commonly used multivariate outlier criterion. Based on Akaike's Information Criterion (AIC), Kadota *et al*. developed a method for detecting outliers, which is free from a significance level [4]. Knorr and Ng introduced a distance-based approach in which outliers are those objects for which there are less than k points within a given threshold in the input data set [5,6]. Angiulli *et al*. proposed a distance-based outlier detection method which finds the top outliers and provides a subset of the data set, called outlier detection solving set, that can be used to predict if new unseen objects are outliers [7]. Distance-based strategies are advantageous since model learning is not required. As an alternative, clustering algorithms can be used for outlier detection in which objects that do not belong to any cluster are regarded as outliers. Wang and Chiang proposed an effective cluster validity measure with outlier detection and cluster merging strategies for support vector clustering (SVC) [8]. The validity measure is capable of finding suitable values for the kernel parameter and soft margin constant. Based on these parameters, SVC algorithm can identify the ideal cluster number and increase robustness to outliers and noises. Schölkopf proposed a method of adapting support vector machine (SVM) to one-class classification problems [9]. Manevitz and Yousef presented two versions using the one-class SVM, both of which can identify outliers: Schölkopf's method and their proposed suggestion [10]. In such methods, after mapping the original samples into a feature space using an appropriate kernel function, the origin is referred to as the second class. In the feature space, samples close to the origin or lying on the standard subspace such as axes are regarded as outliers. Bandyopadhyay and Santra applied a genetic algorithm to the outlier detection problem in a lower dimensional space of a given data set, dividing these spaces into grids and efficiently computing the sparsity factor of the grid [11]. Aggarwal and Yu studied the problem of outlier detection for high-dimensional data, which works by finding lower dimensional projections [12]. Malossini *et al*. proposed two methods for detecting potential labeling errors: Classification-stability algorithm (CL-stability) and Leave-One-Out-Error-sensitivity algorithm (LOOE-sensitivity) [13]. In CL-stability, the stability of the classification of a sample is evaluated with a small perturbation of the other samples. LOOE-sensitivity was derived from the fact that if a sample is mislabeled, flipping the label of the sample should improve the prediction power.

In this paper, we propose a new outlier detection method based on KL divergence [14]. Due to the possible non-linearity of data structure, we deal with this problem in a higher feature space rather than the original space. Several issues arise after data mapping such as singularity because of small sample size versus high feature dimension. We address the computational issues and show the effectiveness of the proposed approach, KL divergence for outlier detection (KLOD).

Methods

Markov blanket

Markov blanket algorithm proposed by Koller and Sahami is a cross-entropy based technique to identify redundant and irrelevant features [15]. Let $\mathbf{F}$ be a full set of features and $\mathbf{M} \subseteq \mathbf{F}$ be a subset of features which does not contain feature F_i. Then, $\mathbf{M}$ is called a Markov blanket for F_i if F_i is conditionally independent of $\mathbf{F}$ - $\mathbf{M}$-$\{F_i\}$ given $\mathbf{M}$. Generally, the Markov blanket $\mathbf{M}_i$ of F_i is defined as a subset of features that consists of some features that have the highest Pearson correlation with F_i. To evaluate the closeness between F_i and its Markov blanket $\mathbf{M}_i$, the following expected cross-entropy Δ is estimated:

$$\Delta(F_i \mid \mathbf{M}_i) = \sum_{\mathbf{f}_{\mathbf{M}_i}, f_i} P(\mathbf{M}_i = \mathbf{f}_{\mathbf{M}_i}, F_i = f_i) \times D(P(c \mid \mathbf{M}_i = \mathbf{f}_{\mathbf{M}_i}, F_i = f_i) \,||\, P(c \mid \mathbf{M}_i = \mathbf{f}_{\mathbf{M}_i})),$$

where $\mathbf{f}_{\mathbf{M}i}$ and f_i are feature values to $\mathbf{M}_i$ and F_i, respectively, c is the class label, and $D(.||.)$ represents the cross-entropy (a.k.a. Kullback-Leibler divergence). For each feature, Δ value is computed and a feature with the smallest Δ value is eliminated from the whole feature set. With the remaining features, the procedure is repeated until a predefined number of features remains.

Kullback-Leibler (KL) divergence

KL divergence, widely used in information theory, is adopted in Markov blanket as a core component. As shown in Markov blanket, KL divergence represents a measure of the distance between two probability distributions [16], i.e., for two probability densities $p(\mathbf{x})$ and $q(\mathbf{x})$, the KL-divergence is defined as

$$D_{\mathrm{KL}}(p \,||\, q) = \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}.$$

Suppose that $\mathcal{N}(\mu, \Sigma)$ is a multivariate Gaussian distribution defined as

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \mu) \right),$$

where $\mathbf{x} \in \mathcal{R}^m$ and $|\Sigma|$ is the determinant of covariance matrix Σ. Given two different probability density

functions, $p(\mathbf{x}) = \mathcal{N}_1(\mu_1, \Sigma_1)$ and $q(\mathbf{x}) = \mathcal{N}_2(\mu_2, \Sigma_2)$, the KL divergence is defined as

$$D_{\mathrm{KL}}(\mathcal{N}_1||\mathcal{N}_2) = \frac{1}{2}\{(\mu_1-\mu_2)^{\mathrm{T}}\Sigma_2^{-1}(\mu_1-\mu_2) + \log\frac{|\Sigma_2|}{|\Sigma_1|} + \mathrm{tr}[\Sigma_1\Sigma_2^{-1} - \mathbf{I}_m]\}.$$

Concept of KL divergence for outlier detection (KLOD)

In Markov blanket, based on KL divergence, after calculating Δ value of Eq. (1) for each feature, a feature with the lowest Δ value is considered to be the most redundant. Using KL divergence, our new outlier detection method, called KLOD, employs similar strategy to the Markov blanket, i.e., while Markov blanket algorithm detects redundant and irrelevant features, our method identifies outliers. In KLOD, each sample $\mathbf{x}_i$ has a sample set that consists of t samples close to the $\mathbf{x}_i$. To calculate the distance between samples, Euclidean metric is used. More specifically, we define two sample sets, i.e., $\mathbf{S}_1$ and $\mathbf{S}_2$: $\mathbf{S}_2$ is a sample set close to $\mathbf{x}_i$ in Euclidean distance and the other set $\mathbf{S}_1$ consists of $\mathbf{x}_i$ and all samples in $\mathbf{S}_2$. The similarity, $D_{\mathbf{x}_i}(\mathbf{S}_1||\mathbf{S}_2)$, between $\mathbf{S}_1$ and $\mathbf{S}_2$ for each sample can be measured by using KL divergence, where $1 \le i \le n$ and n is the total number of samples in the data set. Intuitively, in our strategy, a sample $\mathbf{x}_i$ with the largest D is regarded as an outlier.

$$o = \arg\max_{1\le i\le n} D_{\mathbf{x}_i}$$

Given a data set with nonlinear data structure, if we model the linearity for the data set, it will cause our strategy to fail. Here, we focus on modeling the nonlinearity. Accordingly, with a mapping function φ, the original space is mapped into a higher dimensional feature space. Let $\mathbf{S}_1^{\Phi}$ and $\mathbf{S}_2^{\Phi}$ denote the two sample sets in the feature space in which we compute the similarity $D(\mathbf{S}_1^{\Phi}||\mathbf{S}_2^{\Phi})$ between $\mathbf{S}_1^{\Phi}$ and $\mathbf{S}_2^{\Phi}$. For each sample, its $D(\mathbf{S}_1^{\Phi}||\mathbf{S}_2^{\Phi})$ is calculated. A sample which has the largest $D(\mathbf{S}_1^{\Phi}||\mathbf{S}_2^{\Phi})$ is referred to as an outlier.

Please see an example in Figure 1. However, the calculation leads to several important issues to be considered, such as kernel trick, singularity problem, and calculation of KL divergence in the feature space. In the following sections, we will describe them.

Kernel function

Suppose that $\{\mathbf{x}_1, \mathbf{x}_2, \cup \mathbf{x}_n\}$ are the given samples in the original space. After mapping the samples into a higher feature space by a nonlinear mapping function φ, the samples in the feature space are observed as $\Phi_{mxn} = [\varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2), \cup, \varphi(\mathbf{x}_n)]$ where m is the number of features. Denote $\mathbf{K}$ as follows:

$$\mathbf{K} = \Phi^{\mathrm{T}}\Phi.$$

The calculation can be performed using kernel trick, i.e., the ijth element, $\varphi(\mathbf{x}_i)^{\mathrm{T}}\varphi(\mathbf{x}_j)$, of the $\mathbf{K}$ matrix can be computed as a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$. In literatures, the polynomial kernel and the Gaussian kernel are the most widely used kernel functions. In this study, the Gaussian kernel function is used:

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\right),$$

where σ controls the kernel width. Similar to Eq. (6), we define $\mathbf{K}_{ij}$ as follows:

$$\mathbf{K}_{ij} = \Phi_i^{\mathrm{T}}\Phi_j,$$

where if $i \neq j$, Φ_i and Φ_j are different sample sets in the feature space; if $i = j$, $\mathbf{K}_{ij}$ is equivalent to the definition of $\mathbf{K}$. Indeed, the feature space and the mapping function may not be explicitly known. However, once the kernel function is known, we can easily deal with the nonlinear mapping problem by replacing the mapping functions by the kernel functions.

KL divergence equation is composed of mean and covariance components. The mean and the covariance matrix in the feature space are estimated as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i) = \mathbf{\Phi s},$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\phi(\mathbf{x}_i)-\mu)(\phi(\mathbf{x}_i)-\mu)^{\mathrm{T}} = \mathbf{\Phi J J}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}},$$

where $\mathbf{s}_{n\times 1} = \frac{1}{n}\vec{\mathbf{1}}^{\mathrm{T}}$, $\mathbf{J} = \frac{1}{\sqrt{n}}(\mathbf{I}_n - \mathbf{s}\vec{\mathbf{1}})$ and $\vec{\mathbf{1}} = [1, 1, \cup, 1]$. Then, an $m \times n$ matrix $\mathbf{W}$ is denoted as

$$\mathbf{W} = \mathbf{\Phi J} = \frac{\mathbf{1}}{\sqrt{n}}[(\phi(\mathbf{x}_1)-\mu), \cdots, (\phi(\mathbf{x}_n)-\mu)].$$

Singularity problem

The covariance matrix in Eq. (10) is rank-deficient due to the small number of samples against the number of features. This problem, called singularity problem, makes it impossible to calculate the inverse of the covariance matrix. To overcome the problem, several methods have

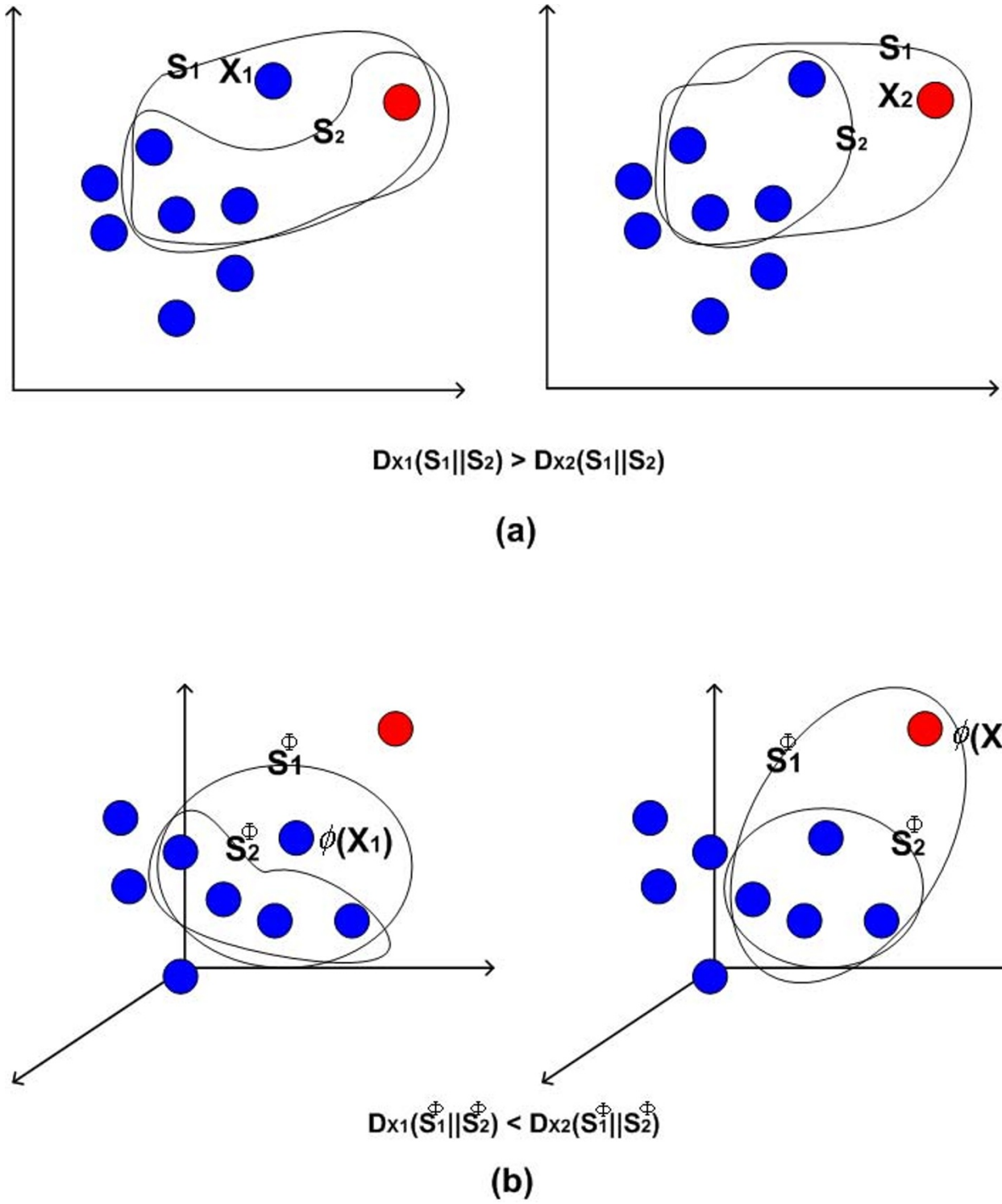

Figure 1
Outlier detection in a high feature space. Suppose that the red dot is a real outlier which is the farthest one from the majority of data. (a) in the original space, $\mathbf{x}_1$ is regarded as an outlier. (b) in the higher feature space, $\mathbf{x}_2$ is correctly detected as an outlier.

been proposed. In this study, we make use of a simple regularized approximation in which some positive constant values are added to the diagonal elements of the covariance matrix [17]. Therefore, the modified covariance matrix is of full rank, hence nonsingular. Let **C** denote

$$\begin{aligned}\mathbf{C} &= \mathbf{\Phi J J}^{\mathbf{T}}\mathbf{\Phi}^{\mathbf{T}} + \rho\mathbf{I}_m,\\ &= \mathbf{W}\mathbf{W}^{\mathbf{T}} + \rho\mathbf{I}_m,\\ &= \mathbf{\Phi R \Phi}^{\mathbf{T}} + \rho\mathbf{I}_m,\end{aligned}$$

where $\mathbf{R} = \mathbf{J}\mathbf{J}^{\mathbf{T}}$, $\rho > 0$, and $\mathbf{I}_m$ is an identity matrix. In this study, $\rho = 1$ is used. Then, the inversion of the matrix **C** can be computed by using *Woodbury formula*:

$$\begin{aligned}\mathbf{C}^{-1} &= (\rho\mathbf{I}_m + \mathbf{\Phi J J}^{\mathbf{T}}\mathbf{\Phi}^{\mathbf{T}})^{-1},\\ &= (\rho\mathbf{I}_m + \mathbf{W}\mathbf{W}^{\mathbf{T}})^{-1},\\ &= \rho^{-1}(\mathbf{I}_m - \rho^{-1}\mathbf{W}(\mathbf{I}_n + \rho^{-1}\mathbf{W}^{\mathbf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathbf{T}}),\\ &= \rho^{-1}(\mathbf{I}_m - \mathbf{W}(\rho\mathbf{I}_n + \mathbf{W}^{\mathbf{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathbf{T}}),\\ &= \rho^{-1}(\mathbf{I}_m - \mathbf{\Phi J M}^{-1}\mathbf{J}^{\mathbf{T}}\mathbf{\Phi}^{\mathbf{T}}),\\ &= \rho^{-1}(\mathbf{I}_m - \mathbf{\Phi B \Phi}^{\mathbf{T}}),\end{aligned}$$

where $\mathbf{B} = \mathbf{J}\mathbf{M}^{-1}\mathbf{J}^{\mathrm{T}}$ and $\mathbf{M} = \rho\mathbf{I}_n + \mathbf{W}^{\mathrm{T}}\mathbf{W} = \rho\,\mathbf{I}_n + \mathbf{J}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi J} = \rho\mathbf{I}_n + \mathbf{J}^{\mathrm{T}}\mathbf{K J}$.

Definition (Woodbury formula): Let **A** be a square $r \times r$ invertible matrix, where **U** and **V** are two $r \times k$ matrices with $k \le r$. Assume that a $k \times k$ matrix $\Sigma = \mathbf{I}_k + \beta\,\mathbf{V}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{U}$, in which $\mathbf{I}_k$ denotes a $k \times k$ identity matrix and β is an arbitrary scalar, is invertible. Then

$$(\mathbf{A} + \beta\,\mathbf{U}\mathbf{V}^{\mathrm{T}})^{-1} = \mathbf{A}^{-1} - \beta\,\mathbf{A}^{-1}\mathbf{U}\Sigma^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{A}^{-1}.$$

Calculation of KL divergence

Suppose that $\mathbf{S}_1^{\Phi}$ and $\mathbf{S}_2^{\Phi}$ are two sample sets in the feature space as mentioned in section. We know that the covariance matrices for both sets are singular. Let $\mathbf{C}_1$ and $\mathbf{C}_2$ denote the approximated covariance matrices for $\mathbf{S}_1^{\Phi}$ and $\mathbf{S}_2^{\Phi}$, respectively, where the size of $\mathbf{S}_1^{\Phi}$ is one larger than that of $\mathbf{S}_2^{\Phi}$. Also, let μ_1 and μ_2 be mean matrices for $\mathbf{S}_1^{\Phi}$ and $\mathbf{S}_2^{\Phi}$, respectively. Therefore, KL divergence for $\mathbf{S}_1^{\Phi}$ and $\mathbf{S}_2^{\Phi}$ is expressed as follows:

$$2D_{\mathrm{KL}}(\mathcal{N}_1||\mathcal{N}_2) = (\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{C}_2^{-1}(\mu_1 - \mu_2) + \log\frac{|\mathbf{C_2}|}{|\mathbf{C_1}|} + \mathbf{tr}\,|\,\mathbf{C}_1\mathbf{C}_2^{-1}] - m.$$

The KL divergence above is composed of three terms, i.e.,

$$\begin{cases}(\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{C}_2^{-1}(\mu_1 - \mu_2)\\ \log\dfrac{|\,\mathbf{C}_2\,|}{|\,\mathbf{C}_1\,|}\\ \mathbf{tr}[\mathbf{C}_1\mathbf{C}_2^{-1}].\end{cases}$$

It should be noted that as shown in Eq. (9), Eq. (12) and Eq. (13), μ_i, $\mathbf{C}_i$ and $\mathbf{C}_i^{-1}$ (i = 1 or 2) have mapping functions rather than kernel functions.

Here, we will show how each term can be expressed by kernel functions instead of mapping functions. The first term consists of four sub-terms,

$$(\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{C}_2^{-1}(\mu_1 - \mu_2) = \mu_1^{\mathrm{T}}\mathbf{C}_2^{-1}\mu_1 + \mu_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mu_2 - \mu_1^{\mathrm{T}}\mathbf{C}_2^{-1}\mu_2 - \mu_2^{\mathrm{T}}\mathbf{C}_2^{-1}\mu_1.$$

Substituting Eq. (9) and Eq. (13) into each sub-term $\mu_i^{\mathrm{T}}\mathbf{C}_j^{-1}\mu_k$, we have

$$\begin{aligned}\mu_i^{\mathrm{T}}\mathbf{C}_j^{-1}\mu_k &= \mathbf{s}_i^{\mathrm{T}}\mathbf{\Phi}_i^{\mathrm{T}}\rho^{-1}(\mathbf{I}_m - \mathbf{\Phi}_j\mathbf{B}_j\mathbf{\Phi}_j^{\mathrm{T}})\mathbf{\Phi}_k\mathbf{s}_k,\\ &= \rho^{-1}(\mathbf{s}_i^{\mathrm{T}}\mathbf{K}_{ik}\mathbf{s}_k - \mathbf{s}_i^{\mathrm{T}}\mathbf{K}_{ij}\mathbf{B}_j\mathbf{K}_{jk}\mathbf{s}_k),\\ &= \rho^{-1}\theta_{ijk}.\end{aligned}$$

As a result of the effort, all mapping functions in the first term are replaced with kernel functions. Before dealing with the second term, we want to introduce the following three properties of determinant that are essential in the calculation of the second term.

Properties of determinant

(a) If **A** is an r-by-r matrix, $\mathbf{det}|d\mathbf{A}| = \mathbf{det}|d\mathbf{I}_r\mathbf{A}| = d^r\mathbf{det}|\mathbf{A}|$.

(b) If **A** and **B** are k-by-r matrices, $\mathbf{det}|\mathbf{I}_k + \mathbf{A}\mathbf{B}^{\mathrm{T}}| = \mathbf{det}|\mathbf{I}_r + \mathbf{B}^{\mathrm{T}}\mathbf{A}|$.

(c) If **A** is invertible, $\mathbf{det}|\mathbf{A}^{-1}| = \mathbf{1}/\mathbf{det}|\mathbf{A}|$.

In the second term, we should compute the determinant of **C**($\mathbf{C}_1$ or $\mathbf{C}_2$). Instead of directly calculating the determinant of **C**, we try to obtain it through the determinant of $\mathbf{C}^{-1}$. That is,

$$\begin{aligned}|\,\mathbf{C}^{-1}\,| &= |\,\rho^{-1}(\mathbf{I}_m - \mathbf{\Phi B \Phi}^{\mathbf{T}})\,|,\\ &= \rho^{-m}\,|\,\mathbf{I}_m - \mathbf{\Phi B \Phi}^{\mathbf{T}}\,|, \quad \text{by property (a)}\\ &= \rho^{-m}\,|\,\mathbf{I}_m - \mathbf{Q}\mathbf{\Phi}^{\mathbf{T}}\,|,\\ &= \rho^{-m}\,|\,\mathbf{I}_n - \mathbf{\Phi}^{\mathbf{T}}\mathbf{Q}\,|, \quad \text{by property (b)}\\ &= \rho^{-m}\,|\,\mathbf{I}_n - \mathbf{\Phi}^{\mathbf{T}}\mathbf{\Phi B}\,|,\\ &= \rho^{-m}\,|\,\mathbf{I}_n - \mathbf{K B}\,|,\end{aligned}$$

where $\mathbf{Q} = \Phi\mathbf{B}$. Here, by property (c), we can easily calculate $|\mathbf{C}|$, i.e.,

$$|\mathbf{C}| = \frac{\mathbf{1}}{|\mathbf{C^{-1}}|} = \frac{\rho^m}{|\mathbf{I}_n - \mathbf{KB}|}.$$

By taking logarithm of $|\mathbf{C}|$, we have

$$\log|\mathbf{C}| = \log\frac{\rho^m}{|\mathbf{I}_n - \mathbf{KB}|} = m\log\rho - \log|\mathbf{I}_n - \mathbf{KB}|.$$

Note that the size of $\mathbf{S}_1^{\Phi}$ is one larger than that of $\mathbf{S}_2^{\Phi}$. If the size of $\mathbf{S}_2^{\Phi}$ is k, the size of $\mathbf{S}_1^{\Phi}$ becomes $k + 1$.

Now we have the second term composed of kernel functions:

$$\log\frac{|\mathbf{C_2}|}{|\mathbf{C_1}|} = \log|\mathbf{C}_2| - \log|\mathbf{C}_1|,$$
$$= \log|\mathbf{I}_{k+1} - \mathbf{K}_{11}\mathbf{B}_1| - \log|\mathbf{I}_k - \mathbf{K}_{22}\mathbf{B}_2|.$$

The third term can be replaced with kernel functions using properties of trace:

$$\begin{aligned}\mathbf{tr}[\mathbf{C}_1\mathbf{C}_2^{-1}] &= \mathbf{tr}[(\Phi_1\mathbf{R}_1\Phi_1^{\mathrm{T}} + \rho\mathbf{I}_m)\rho^{-1}(\mathbf{I}_m - \Phi_2\mathbf{B}_2\Phi_2^{\mathrm{T}})],\\ &= \rho^{-1}\mathbf{tr}[\Phi_1\mathbf{R}_1\Phi_1^{\mathrm{T}}] - \rho^{-1}\mathbf{tr}[\Phi_1\mathbf{R}_1\Phi_1^{\mathrm{T}}\Phi_2\mathbf{B}_2\Phi_2^{\mathrm{T}}] + m - \mathbf{tr}[\Phi_2\mathbf{B}_2\Phi_2^{\mathrm{T}}],\\ &= \rho^{-1}\mathbf{tr}[\mathbf{R}_1\mathbf{K}_{11}] - \rho^{-1}\mathbf{tr}[\mathbf{R}_1\mathbf{K}_{12}\mathbf{B}_2\mathbf{K}_{21}] + m - \mathbf{tr}[\mathbf{B}_2\mathbf{K}_{22}].\end{aligned}$$

Successfully, we substitute all mapping functions in the three terms of KL divergence by kernel functions so that we can calculate KL divergence between two sample sets in the feature space.

Results and discussion

To evaluate the performance of KLOD method, we performed several experiments using a synthetic data, two gene expression data sets, and a high-resolution mass spectrometry data. To obtain unbiased results, all experiments were repeated 30 times with 10-fold cross validation (CV) and the performance was averaged. The performance of KLOD was compared with one-class SVM and Mahalanobis distance based outlier detection methods. Given n samples, the Mahalanobis distance for each multivariate sample $\mathbf{x}_i$ is as follows:

$$D_i = \sqrt{(\mathbf{x}_i - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}_i - \mu)}$$

where Σ and μ are the sample covariance matrix and sample mean vector, respectively. Samples with a large Mahalanobis distance are regarded as outliers.

Results on synthetic data

First, using a synthetic data, we evaluated KLOD to see the ability in detecting outliers. The synthetic data consists of 100 samples, denoted as **N**, each of which has 100 features generated from a mixture of Gaussian $\mathcal{N}(0, \mathbf{I})$. In addition, two sample sets called quasi-outlier set **Q** and perfect outlier set **P** were produced, each of which has 10 samples with 100 features, which were generated from a mixture of Gaussian $\mathcal{N}(0, \mathbf{I})$ and $\mathcal{N}(2, \mathbf{I})$, respectively. It is noted that **Q** was created from the same distribution as **N**. Here, we corrupted **Q** by changing the values in some features. To do so, some features from each sample in **P** were randomly selected. The values of the selected features replaced those of features randomly selected from the corresponding sample in **Q**. Finally, we merged **N** and **Q**, which were used as a synthetic data. Figure 2 illustrates an example of generating the synthetic data. In this experiment, we tested KLOD changing the number of corrupted features from 10 to 30 increasing by 2 and the size of a set, denoted as t, that consists of close samples of each sample from 5 to 20 increasing by 5. With the synthetic data, we measured how accurate our method is in identifying outliers in a way that the number of real outliers is counted out of the first 10 samples detected by KLOD.

Figure 3 shows the experimental results. When the number of noisy features increases, the accuracy shows a tendency to increase as well. It should be noted that for all set sizes, when the number of noisy features is 18, an accuracy of over 90% was obtained. Particularly, for t = 10, 15 and 20, when the number of noisy features is 30, an accuracy of 100% was achieved.

Performance evaluation after outlier removal

Before introducing the outlier removal for real biomedical data, we first introduce the performance evaluation metric we will use which is PCA (principal component analysis) + LDA (linear discriminant analysis). LDA maps the data into a very low dimensionality of c -1, where c is the number of classes. In the reduced space, a simple matching procedure is used for classification. However, in order to guarantee a non-degenerate result from LDA, before the LDA task, the dimensionality of the data must be reduced to at most n - c where n is the number of samples. Principal component analysis (PCA) is often used in the analysis of high dimensional data set. PCA performs a transformation of the original space into a lower dimensional space with little or no information loss while maximally preserving variance.

Lilien *et al*. used the PCA+LDA method in the analysis of mass spectrometry data sets [18]. In this framework, the PCA dimensionality-reduced samples are projected by LDA onto a hyperplane in the way of maximizing the between-class variance and minimizing the within-class

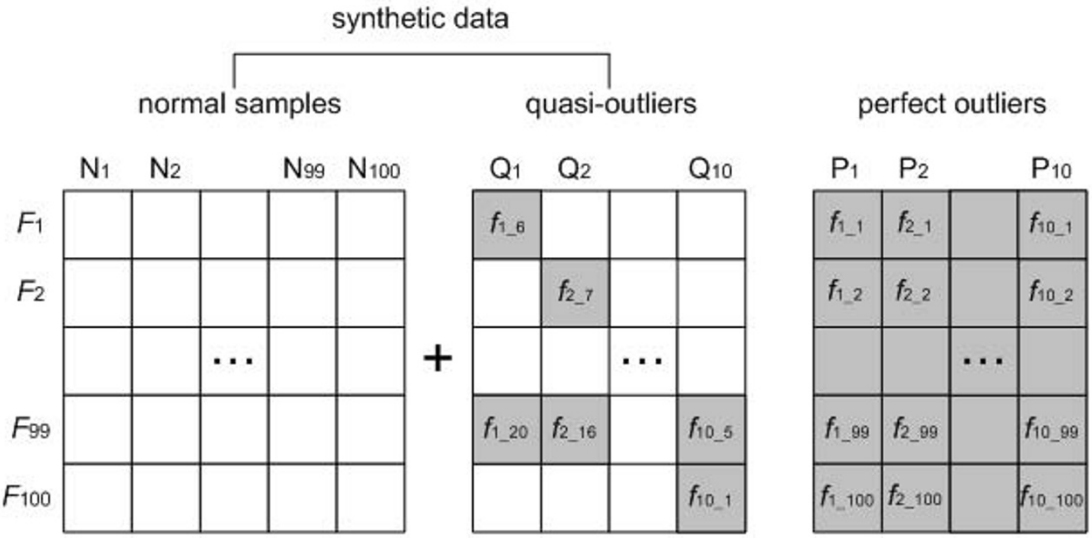

Figure 2
Generation of a synthetic data. This example shows a way used in this study to generate a synthetic data.

variance of the projected samples. To evaluate the performance after outlier removal in our experiments, we employed the PCA+LDA strategy.

Results on gene expression data sets

In this study, two public microarray data sets were used.

• The leukemia data set covers two types of acute leukemia: 47 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples with 7,129 genes. The data set is publicly available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi/[19].

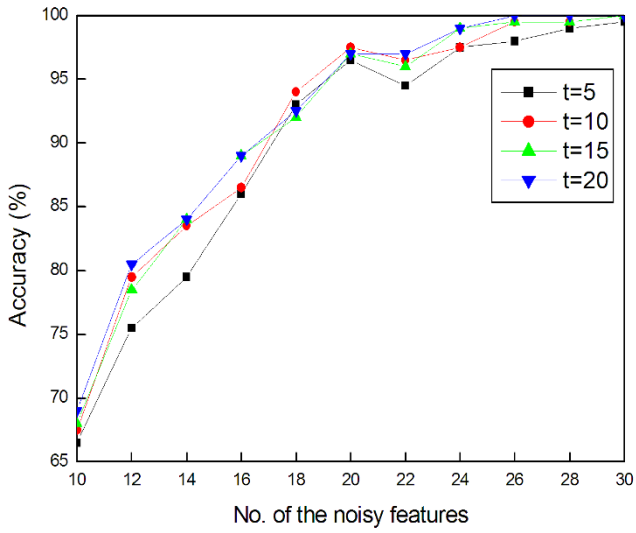

Figure 3
Accuracy of detecting outliers on a synthetic data. The data consists of 100 normal samples and 10 outliers, each having 100 features.

• The colon data set contains 40 tumor and 22 normal colon tissues with 2,000 genes. The data set is available at http://microarray.princeton.edu/oncology/[20].

In experiments with the two microarray data sets, specificity, sensitivity, and accuracy were measured using PCA+LDA classification strategy after removing outliers detected by KLOD with t = 10, Mahalanobis distance based method, and one-class SVM. We define the specificity as the ratio of correctly classified negatives to the actual number of negatives. For leukemia and colon microarray data sets, negatives are ALL and normal samples, respectively. For KLOD and Mahalanobis distance based method, the performance was measured after removing a sample having the largest distance from each class at each iteration. If the prediction rate (specificity or sensitivity) decreases more than a threshold γ compared to the prediction rate before the outlier removal, we stop the outlier detection in the corresponding class. In this study, we used γ = 0.5%. In contrast, for one-class SVM, after excluding all samples regarded as outliers in each class, the performance was assessed.

Table 1 shows the experimental results obtained using leukemia and colon microarray data sets. For the leukemia data set, KLOD achieved the best accuracy with 9 outliers (2 ALL and 7 AML samples).

Table 1: Performance after outlier detection in leukemia and colon data sets.

Data set	Measurements	Without outlier removal	After outlier removal		
			KLOD	Mahalanobis	One-class SVM
Leukemia	Specificity (%)	96.17	99.00	97.37	100
	Sensitivity (%)	95.60	99.44	100	95.24
	Accuracy (%)	95.97	99.13	98.28	98.33
	No. of the outliers	ALL	2	9	8
		AML	7	5	4
Colon	Specificity (%)	82.50	85.95	83.25	85.26
	Sensitivity (%)	88.25	94.43	85.90	94.17
	Accuracy (%)	86.21	91.25	85.00	91.09
	No. of the outliers	normal	1	2	3
		tumor	5	1	4

Mahalanobis distance based method and one-class SVM found 14 and 12 outliers, respectively. For the colon data set, KLOD found 6 outliers (1 normal and 5 tumor samples) with 84.95% specificity, 94.43% sensitivity, and 91.25% accuracy. It should be noted that the performance of Mahalanobis distance based method was degraded in terms of sensitivity and accuracy compared to the performance obtained using all samples without outlier removal, suggesting that outliers detected by Mahalanobis distance based method are unlikely to be real ones.

Results on mass spectrometry data

To evaluate the effectiveness of KLOD, we also used a public mass spectrometry data for liver cancer study that consists of 201 spectra containing hepatocellular carcinoma (HCC) (78), cirrhosis (51), and health (72) [3]. From http://microarray.georgetown.edu/ressomlab/, we downloaded the binned spectra that have 23,846 peaks for each spectrum. To test outlier detection methods, only cirrhosis and HCC spectra were used as in [3]. By using t-test with the significance level of 0.05 in cirrhosis and HCC spectra, we selected 10,682 peaks. That is, the top 10,682 peaks selected by t-test with cirrhosis and HCC spectra were used in outlier detection methods. The same way as performed with the microarray data sets was employed. Here cirrhosis samples are regarded as negatives. As shown in Table 2, KLOD obtained slightly higher performance with the smallest number of outliers than Mahalanobis distance based method and one-class SVM. From the results in experiments using mass spectrometry and microarray data sets, it seems that one-class SVM detects more outliers than KLOD and Mahalanobis distance based method.

Conclusion

We proposed a new outlier detection method based on KL divergence called KLOD. Our idea was derived from Markov blanket algorithm where redundant and irrelevant features are removed based on KL divergence. We tackled the outlier detection problem in a higher feature space after mapping the original data. The mapping leads to several issues. In particular, we showed how to calculate KL divergence in the higher feature space by using the properties of determinant and trace of matrix. To asses the usefulness of KLOD, we used a synthetic data and real life data sets. Compared to Mahalanobis distance based method and one-class SVM, KLOD achieved higher or comparable performance.

Table 2: Performance after outlier detection in liver cancer mass spectrometry data.

Measurements	Without outlier removal	After outlier removal		
		KLOD	Mahalanobis	One-class SVM
Specificity (%)	93.63	94.69	94.29	94.35
Sensitivity (%)	92.82	93.95	93.51	93.89
Accuracy (%)	93.14	94.23	93.82	94.07
No. of the outliers	Cirrhosis	3	2	5
	HCC	2	4	6

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JHO performed data analysis and wrote the manuscript. JG supervised the project and edited the paper.

Acknowledgements

This work was supported in part by NSF under grants IIS-0612152 and IIS-0612214.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 4, 2009: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2008. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S4.

References

1. Lee W, Stolfo S, Mok K: **Mining audit data to build intrusion detection models.** *Proc Int Conf Knowledge Discovery and Data Mining (KDD 1998)* 1998:66-72.
2. Fawcett T, Provost F: **Adaptive fraud detection.** *Data Mining and Knowledge Discovery* 1997, **1:**291-316.
3. Ressom H, Varghese R, Drake S, Hortin G, Abdel-Hamid M, *et al.*: **Peak selection from MALDI-TOF mass spectra using ant colony optimization.** *Bioinformatics* 2007, **23:**619-626.
4. Kadota K, Tominaga D, Akiyama Y, Takahashi K: **Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification.** *Chem-Bio Informatics Journal* 2003, **3:**30-45.
5. Knorr E, Ng R: **Algorithms for mining distance-based outliers in large datasets.** *Proc Int Conf Very Large Databases (VLDB 1998)* 1998:392-403.
6. Knorr E, Ng R, Tucakov V: **Distance-based outlier: algorithms and applications.** *Proc Int Conf Very Large Databases (VLDB 2000)* 2000:237-253.
7. Angiulli F, Basta S, Pizzuti C: **Distance-based detection and prediction of outliers.** *IEEE Trans on Knowledge and Data Engineering* 2006, **18:**145-160.
8. Wang JS, Chiang JC: **A cluster validity measure with outlier detection for support vector clustering.** *IEEE Trans on Systems, Man, and Cybernetics, Part B* 2008, **38:**78-89.
9. Schölkopf B, Platt J, Shawe-Taylor J, Smola A, Williamson R: **Estimating the support of a high-dimensional distribution.** *Neural Computation* 2001, **13:**1443-1471.
10. Manevitz L, Yousef M: **One-class SVMs for document classification.** *Journal of Machine Learning Research* 2001, **2:**139-154.
11. Bandyopadhyay S, Santra S: **A genetic approach for efficient outlier detection in projected space.** *Pattern Recognition* 2008, **41:**1338-1349.
12. Aggarwal C, Yu P: **Outlier detection for high dimensional data.** *Proc ACM SIGMOD* 2001:37-46.
13. Malossini A, Blanzieri E, Ng R: **Detecting potential labeling errors in microarrays by data perturbation.** *Bioinformatics* 2006, **22:**2114-2121.
14. Oh J, Gao J, Rosenblatt K: **Biological data outlier detection based on Kullback-Leibler divergence.** *Proc IEEE Int Conf on Bioinformatics and Biomedicine (BIBM 2008)* 2008:249-254.
15. Koller D, Sahami M: **Toward optimal feature selection.** *Proc Int Conf on Machine Learnin* 1996.
16. Tumminello M, Lillo F, Mantegna R: **Kullback-Leibler distance as a measure of the information filtered from multivariate data.** *Physical Review E* 2007, **76:**256-67.
17. Zhou S, Chellappa R: **From sample similarity to ensemble similarity: probabilistic distance measures in reproducing kernel Hilbert space.** *IEEE Trans on Pattern Analysis and Machine Intelligence* 2006, **28:**917-929.
18. Lilien R, Farid H, Donald B: **Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum.** *Journal of Computational Biology* 2003, **10:**925-946.
19. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
20. Alon U, Barkai N, Notterman D, Gish K, Ybarra S, *et al.*: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci U S A* 1999, **96:**6745-6750.