Accepted Manuscript

Auto Insurance Fraud Detection Using Unsupervised Spectral Ranking for Anomaly

Ke Nian, Haofan Zhang, Aditya Tayal, Thomas Coleman, Yuying Li

PII: S2405-9188(16)30005-8

DOI: 10.1016/j.jfds.2016.03.001

Reference: JFDS 8

To appear in: *The Journal of Finance and Data Science*

Received Date: 29 February 2016

Accepted Date: 2 March 2016

Please cite this article as: Nian K, Zhang H, Tayal A, Coleman T, Li Y, Auto Insurance Fraud Detection Using Unsupervised Spectral Ranking for Anomaly, *The Journal of Finance and Data Science* (2016), doi: 10.1016/j.jfds.2016.03.001.

Auto Insurance Fraud Detection Using Unsupervised Spectral Ranking for Anomaly

Ke Nian[a,1], Haofan Zhang[a,1], Aditya Tayal[a,1], Thomas Coleman[b,2], Yuying Li[a,1,*]

[a] *Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1*
[b] *Combinatorics and Optimization, University of Waterloo, Waterloo, ON, Canada N2L 3G1*

Abstract

For many data mining problems, obtaining labels is costly and time consuming, if not practically infeasible. In addition, unlabeled data often includes categorical or ordinal features which, compared with numerical features, can present additional challenges. We propose a new unsupervised spectral ranking method for anomaly (SRA). We illustrate that the spectral optimization in SRA can be viewed as a relaxation of an unsupervised SVM problem. We demonstrate that the first non-principal eigenvector of a Laplacian matrix is linked to a bi-class classification strength measure which can be used to rank anomalies. Using the first non-principal eigenvector of the Laplacian matrix directly, the proposed SRA generates an anomaly ranking either with respect to the majority class or with respect to two main patterns. The choice of the ranking reference can be made based on whether the cardinality of the smaller class (positive or negative) is sufficiently large. Using an auto insurance claim data set but ignoring labels when generating ranking, we show that our proposed SRA significantly surpasses existing outlier-based fraud detection methods. Finally we demonstrate that, while proposed SRA yields good performance for a few similarity measures for the auto insurance claim data, notably ones based on the Hamming distance, choosing appropriate similarity measures for a fraud detection problem remains crucial.

*Corresponding author.
Email addresses: knian@uwaterloo.ca (Ke Nian), haofan.zhang@uwaterloo.ca (Haofan Zhang), amtayal@uwaterloo.ca (Aditya Tayal), tfcoleman@uwaterloo.ca (Thomas Coleman), yuying@uwaterloo.ca (Yuying Li)

[1] All authors acknowledge funding from the National Sciences and Engineering Research Council of Canada

[2] This author acknowledges funding from the Ophelia Lazaridis University Research Chair. The views expressed herein are solely from the authors.

Keywords: unsupervised learning, fraud detection, rare class ranking, similarity measure, kernels, spectral clustering, one-class svm

Auto Insurance Fraud Detection Using Unsupervised Spectral Ranking for Anomaly

1. Introduction

The main objective of the paper is to propose a method for fraud detection by detecting anomaly of interdependence relation, captured by a kernel similarity, among the feature variables. The method includes a new ranking scheme, with the top of the ranked list indicating the most suspicious case, based on spectral analysis of the Laplacian of the similarity kernel. To illustrate, the proposed method is applied to both synthetic data sets and an auto insurance application.

Fighting against insurance fraud is a challenging problem both technically and operationally. It is reported that approximately 21% $\sim$ 36% auto-insurance claims contain elements of suspected fraud but only less than 3% of the suspected fraud is prosecuted.[1,2] Traditionally, insurance fraud detection relies heavily on auditing and expert inspection. Since manually detecting fraud cases is costly and inefficient and fraud need to be detected prior to the claim payment, data mining analytics is increasingly recognized as a key in fighting against fraud. This is due to the fact that data mining and machine learning techniques have the potential to detect suspicious cases in a timely manner, and therefore potentially significantly reduce economic losses, both to the insurers and policy holders. Indeed there is great demand for effective predictive methods which maximize the true positive detection rate, minimize the false positive rate, and are able to quickly identify

new and emerging fraud schemes.

Fraud detection can be approached as an *anomaly ranking* problem. Anomaly detection encompasses a large collection of data mining problems such as disease detection, credit card fraud detection, and detection of any new pattern amongst the existing patterns. In addition, comparing to simply providing binary classifications, providing a ranking, which represents the degree of relative abnormality, is advanntageous in cost and benefit evaluation analysis, as well as in turning analytic analysis into action.

If anomaly can be treated as a rare class, many methods for supervised rare class ranking exist in the literature. RankSVM[3] can be applied to a bi-class rare class prediction problem. However, solving a nonlinear kernel RankSVM problem is computationally prohibitive for large data mining problems. Using SVM ranking loss function, a rare class based nonlinear kernel classification method, RankRC[4,5], is proposed.

Unfortunately it may not be feasible or desirable to use supervised anomaly ranking for fraud detection, since obtaining clearly fraudulent (and non-fraudulent) labels is very costly, if not impossible. Even if one ignores human investigative costs, it is quite common to find fraud investigators to differ in their claim assessments. This raises additional reliability issues in data (specifically in labels). In contrast, unsupervised learning has the advantages of being more economical and efficient application of knowledge discovery. Moreover, the need to detect fraudulent claims before payments are made and to quickly identify new fraud schemes essentially rule out supervised learning as a candidate solution to effective fraud detection in practice. Therefore, an unsupervised anomaly ranking is more appropriate and beneficial here.

Standard unsupervised anomaly detection methods include clustering analysis and outlier detection. Many outlier detection methods have been proposed in the literature. Examples include k-Nearest Neighbor (k-NN) outlier detection, one-class Support Vector Machine (OC-SVM) (including kernel-based), and density-based methods, e.g., Local Outlier Factor (LOF). The effectiveness of these methods has been investigated in numerous application domains, including network intrusion detection, credit card fraud detection, and abnormal activity detection in electronic commerce. However, many standard outlier detection methods, e.g., one-class SVM, are only suitable for detecting outliers with respect to a single global cluster, which we refer to as global outliers in this paper. An implicit assumption in this case is that normal cases are generated from one mechanism and abnormal cases are generated from other mechanisms. Density-based methods can be effective in detecting both global outliers and local outliers; but assumption here is that data density is the only discriminant for abnormality. Density-based methods fail when small dense clusters also constitute abnormality. In addition, density based methods, e.g., LOF, often require users to define a parameter which specifies a neighborhood to compare the density. Tuning these parameters can often be challenging.

Understandably, unsupervised learning is much more difficult than supervised learning, since learning targets are not available to guide the learning process. In practice, difficulty in unsupervised learning is further exacerbated by the additional challenge in identifying relevant features for unsupervised learning methods. Due to these challenges, the existing literature on auto insurance fraud detection typically formulates the problem as a supervised learning problem. [6,7]

The literature on *unsupervised* auto insurance fraud detection is extremely

sparse. To the best of our knowledge, it includes the self-organizing feature map method[8], and PRIDIT analysis[9,10,11]. PRIDIT is based on RIDIT scores and Principal Component Analysis. We note these studies[8,9,10,11] have been conducted using the single Personal Injury Protection (PIP) data set[9], which is provided by Automobile Insurance Bureau (AIB) from Massachusetts. This data set has been preprocessed by auditors and fraud detection inspectors. Specifically, data features are the red flags specified by domain experts, with attribute values for fraudulent instances in the data set typically smaller than that of the instances from the non-fraudulent class. This corresponds to a preliminary ranking of fraudulent suspiciousness. Thus it is reasonable to regard this as a partially supervised learning, which is susceptible to the issues associated with label unreliability in fraud detection. In addition, insurance claim data often consists of numerical, ordinal, categorical, and text data. Consequently it will be difficult to apply the self-organizing feature map method[8], and PRIDIT analysis[9,10,11] without first preprocessing from domain experts. Moreover, when claim data consists of many categorical attributes, it is reasonable to expect that data can form more than one major cluster. This makes any single class based outlier detection method less likely to be effective.

Based on spectral analysis, we propose a new unsupervised ranking method for anomaly detection. Specifically we consider both rare class ranking, in which anomaly is assessed with respect to a single majority class, as well as anomaly ranking which is assessed with respect to more than one major pattern. We note that the words, class and pattern, are used interchangeably in this paper.

Given a data set of input attributes, anomaly can be assessed with respect to marginal attribute distributions as well as attribute dependence relationships,

which can be either linear or nonlinear. Recently Gretton et al.[12] and Song et al.[13] have studied the use of kernels to measure nonlinear feature inter-dependence. In this paper we focus on detecting anomaly in the attribute dependence using similarity kernels, where the similarity constructed to capture dependence relation among input attributes is assumed to be given. We then use spectral analysis to compute the first bi-modal non-principal eigenvector to generate anomaly ranking. We also compute the second non-principal eigenvector to assist visualization. In addition, we apply several standard anomaly detection methods on auto insurance problem and compare the performance with our proposed ranking algorithms.

The main contributions of this paper include:

- We observe a connection between an unsupervised Support Vector Machine (SVM) optimization formulation and the spectral optimization. Specifically we demonstrate that spectral optimization based on the Laplacian matrix can be viewed as a relaxation of unsupervised SVM. Consequently the magnitudes of eigenvector components approximate the degree of support in the optimal bi-class separation function, which is used to yield the proposed anomaly ranking.

- We demonstrate the unsupervised SRA for anomaly detection with respect to a single majority class as well as multiple clusters using a few synthetic data sets.

- Using the real auto insurance claim data[6], we evaluate effectiveness of the unsupervised SRA for detecting anomaly with respect to multiple major patterns. We emphasize that here the anomaly ranking is generated without using labels. We show that the proposed SRA performs significantly better

than existing methods. In addition we demonstrate that, for this data set, SRA achieves performance close to the in-sample training performance of a random forest supervised classification, which can be considered a performance upper bound.

Presentation of the paper is organized as follows. In §2, we summarize the basic idea of spectral analysis and describe the interpretation of the eigenvector components which motivates our ranking algorithm. In §3, we propose a new method of unsupervised spectral ranking for anomaly (SRA). In §4, we review and justify similarity measures for categorical attributes which are used in our computational investigation. We present and discuss the results from the proposed ranking method for the auto insurance claim data in §5. Concluding remarks are given in §6

2. Spectral Analysis and Clustering

Spectral clustering[14] has become a widely used clustering technique, often outperforming traditional clustering techniques such as k-means and hierarchical clustering. Before proposing our ranking method, we first briefly review the spectral clustering technique.

The main objective of clustering is to partition data into groups so that similarity between different groups is minimized. Hence similarity based clustering can be modeled as a graph cut problem. An undirected graph $G = (V, E)$ is used to represent the data set and pairwise similarity, with vertices $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ corresponding to data instances and an adjacency matrix $W = (W_{ij})$ specifying pairwise similarities, where $W_{ij} \geq 0$ is the similarity between $\mathbf{v}_i$ and $\mathbf{v}_j$.

Let $\mathbf{d}$ be the degree vector of vertices, with $\mathbf{d}_i = \sum_j W_{ij}$, and D be the diagonal

matrix with $\mathbf{d}$ on the diagonal. From the degree matrix D and the weighted adjacency matrix W, a Laplacian matrix, which plays an important role in spectral clustering computation, is defined. Although there are variations in the definition of Laplacian, the relevant definition to our discussion in this paper is the symmetric normalized Laplacian[15] $L = I - D^{-1/2} W D^{-1/2}$.

The key idea of a spectral clustering algorithm is to determine clustering membership of data instances by applying a simple clustering technique, e.g., k-means, to a small subset of eigenvectors of a graph Laplacian matrix.[16,15]

Assume that the eigenvalues of the Laplacian L are

$$\lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$$

and $\mathbf{g}_k^*$ is an eigenvector associated with λ_k, $k = 0, \cdots, n-1$. Let $\mathbf{e}$ be a n-by-1 vector of ones. Since $LD^{\frac{1}{2}}\mathbf{e} = 0$ and L is positive semidefinite, $\mathbf{g}_0^* = D^{\frac{1}{2}}\mathbf{e}$ is the principal eigenvector associated with the minimum eigenvalue $\lambda_0 = 0$. Therefore, the first non-principal eigenvector solves

$$\begin{aligned} & \min_{\mathbf{g} \in \Re^n} \mathbf{g}^T L \mathbf{g} \\ \text{subject to} \quad & \mathbf{g}^T \mathbf{g}_0^* = 0 \\ & \mathbf{g}^T \mathbf{g} = \upsilon \end{aligned} \tag{1}$$

and the kth non-principal eigenvector solves

$$\begin{aligned} & \min_{\mathbf{g} \in \Re^n} \mathbf{g}^T L \mathbf{g} \\ \text{subject to} \quad & \mathbf{g}^T \mathbf{g}_i^* = 0, \quad i = 0, \cdots, k-1 \\ & \mathbf{g}^T \mathbf{g} = \upsilon \end{aligned} \tag{2}$$

where $\upsilon = \sum_{i=1}^{n} \mathbf{d}_i$.

From $\mathbf{g}_0^* = D^{\frac{1}{2}}\mathbf{e}$, each non-principal eigenvector $\mathbf{g}_i^*$ satisfies

$$(D^{\frac{1}{2}}\mathbf{e})^T \mathbf{g}_i^* = 0.$$

Thus each eigenvector $\mathbf{g}_i^*$ always contains both positive and negative components, which we can group into a non-negative class $\mathcal{C}_+ = \{j : (\mathbf{g}_i^*)_j \geq 0\}$ and a negative class $\mathcal{C}_- = \{j : (\mathbf{g}_i^*)_j < 0\}$.

A typical motivation of spectral clustering is that the spectral optimization problem (1) can be regarded as a relaxation of an integer programming problem modeling a normalized graph 2-cut problem. We refer an interested reader to[14] for a more detailed discussion. A spectral clustering based on m eigenvectors can be regarded as a m-step iterative bi-clustering method, in which each successive iteration looks for a bi-clustering in the space orthogonal to the first $(m-1)$ bi-clustering eigenvector space. To determine clustering membership from the eigenvectors of a Laplacian matrix, another clustering method, e.g., k-means, is subsequently applied to the eigenvectors. Clustering methods such as k-means typically require the number of clusters to be specified *a prior*.

Instead of determining cluster membership, our goal in this paper is to develop an anomaly ranking method. To motivate, we first investigate information which is potentially present in the eigenvector components. Specifically we show next that optimization problem in (1) can be considered a relaxation to the unsupervised SVM.

Assume that we are given the data instances without labels. For unsupervised SVM, the goal is to find the optimal label assignment such that, the optimal hyperplane from supervised kernel SVM, using the assigned labels, has the maximal

margin.

Formally, given a feature mapping $\phi(\mathbf{x})$,we want to solve the following nested minimization problem:

$$\min_{y_i\in\{\pm 1\}} \left\{ \min_{w,\xi,b,y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b)\geq 1-\xi_i,\xi_i\geq 0} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \right\} \tag{3}$$

Due to the integer constraints on y_i, (3) is a NP-hard problem.[17]

Let $K = (K_{ij})$, $K_{ij} = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j), i, j = 1, \cdots, m$, be the kernel matrix and $Y = \mathrm{diag}(y)$. The inner convex optimization problem satisfies strong duality and has the equivalent dual formulation

$$\max_{\substack{0\leq\alpha_i\leq C\\ \mathbf{y}^T\boldsymbol{\alpha}=0}} -\frac{1}{2}\boldsymbol{\alpha}^T YKY\boldsymbol{\alpha} + \mathbf{e}^T\boldsymbol{\alpha} \tag{4}$$

Replacing the inner minimization by its dual problem, we have the following equivalent minmax problem

$$\min_{y_i\in\{\pm 1\}} \max_{\substack{0\leq\alpha_i\leq C\\ \mathbf{y}^T\boldsymbol{\alpha}=0}} -\frac{1}{2}\boldsymbol{\alpha}^T YKY\boldsymbol{\alpha} + \mathbf{e}^T\boldsymbol{\alpha} \tag{5}$$

Let $\mathbf{z} \in \mathbb{R}^m$ where

$$z_i = \alpha_i \cdot y_i, \quad i = 1, \cdots, m$$

It immediately follows that

$$\boldsymbol{\alpha}^T YKY\boldsymbol{\alpha} = \mathbf{z}^T K\mathbf{z}, \quad \text{and} \quad \mathbf{y}^T\boldsymbol{\alpha} = \mathbf{e}^T\mathbf{z}$$

Moreover, for any $\alpha_i \neq 0$, we have

$$y_i = \mathrm{sign}(z_i), \quad i = 1, \cdots, m \tag{6}$$

Consequently

$$\mathbf{e}^T \alpha = \mathbf{e}^T |\mathbf{z}|$$

Therefore, given y, solution to (4) is one of local maximizers of

$$\begin{aligned} \max_{\mathbf{z}} \quad & \mathbf{e}^T |\mathbf{z}| - \frac{1}{2}\mathbf{z}^T K \mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T \mathbf{z} = 0, \\ & |\mathbf{z}| \leq C \end{aligned} \tag{7}$$

Note that the objective function in (7) is no longer concave and (7) has many local maximizers. Let

$$\mathcal{Z}^* = \left\{ \mathbf{z}^* : \mathbf{z}^* = \text{augmax}_{\substack{\mathbf{e}^T \mathbf{z}=0 \\ |\mathbf{z}| \leq C}} \mathbf{e}^T |\mathbf{z}| - \frac{1}{2}\mathbf{z}^T K \mathbf{z} \right\}$$

Now consider the following simpler problem

$$\begin{aligned} \min_{\mathbf{z}} \quad & -\frac{1}{2}\mathbf{z}^T K \mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T \mathbf{z} = 0, \\ & |\mathbf{z}| \leq C \end{aligned} \tag{8}$$

We note that (8) remains an NP-hard problem due to the concave objective function and rectangular constraints.[18] Assume K is positive definite in the space $\{\mathbf{z} : \mathbf{e}^T \mathbf{z} = 0\}$. Then all local minimizers of (8) are at the boundary of $|\mathbf{z}| \leq C$

Intuitively, each maximizer of (7) is created by moving from the origin into a quadrant in $\Re^n$, from the combined values of the first increasing term $\mathbf{e}^T |\mathbf{z}|$ and the second decreasing term $-\frac{1}{2}\mathbf{z}^T K \mathbf{z}$. For the unsupervised SVM, the main task is to assign the label $\mathbf{y}$ to determine the minimum of local maximums. In other

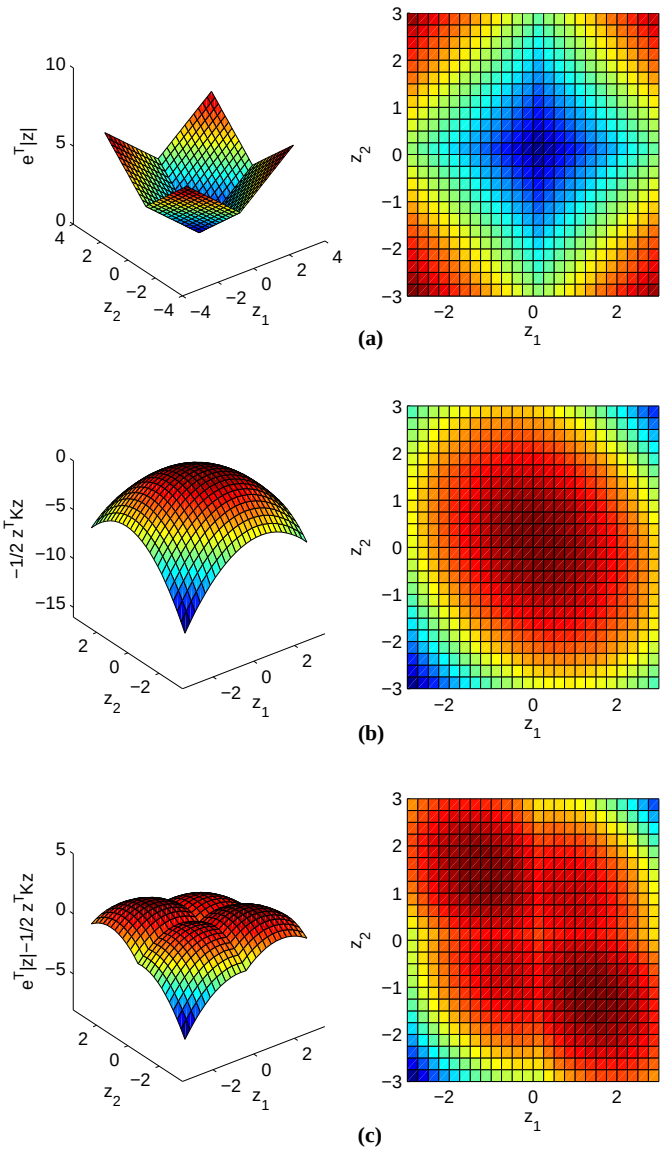


Figure 1: Graphical illustration of $\mathbf{e}^T|\mathbf{z}|$, $-\frac{1}{2}\mathbf{z}^T K\mathbf{z}$, and $\mathbf{e}^T|\mathbf{z}| - \frac{1}{2}\mathbf{z}^T K\mathbf{z}$

words, we want to determine the quadrant at which the global optimal solution locates. Assuming moving away from the origin in contours of $\mathbf{e}^T|\mathbf{z}|$, the optimal label assignments $\mathbf{y}$ for the minmax objective function (5) corresponds to the quadrant along which the objective function $-\frac{1}{2}\mathbf{z}^T K\mathbf{z}$ decreases the fastest. This suggests we can obtain a reasonable approximation to the optimal label assignment by computing the minimum of $-\frac{1}{2}\mathbf{z}^T K\mathbf{z}$ under the same constraints, i.e., (8) is a reasonable approximation to the unsupervised SVM (5). Figure 1 illustrates this motivation in the two dimensional case. Subplot (a), (b), and (c) graph possible shapes of functions $\mathbf{e}^T|\mathbf{z}|$, $-\frac{1}{2}\mathbf{z}^T K\mathbf{z}$, and $\mathbf{e}^T|\mathbf{z}| - \frac{1}{2}\mathbf{z}^T K\mathbf{z}$ respectively.

Next we show that the spectral optimization (1) is a reasonable approximation to (8) (and consequently to unsupervised SVM (5)). Denoting $\mathbf{z} = D^{\frac{1}{2}}\mathbf{g}$, $K = D^{-1}WD^{-1}$, and ignoring the constant υ in the objective function, spectral optimization (1) is equivalent to

$$\begin{aligned} \min_{\mathbf{z}\in\Re^m} \quad & -\mathbf{z}^T K\mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T\mathbf{z} = 0 \\ & \mathbf{z}^T D^{-1}\mathbf{z} = \upsilon. \end{aligned}$$

Assuming K is positive definite, the ellipsoidal equality constraint can be replaced by an inequality constraint

$$\begin{aligned} \min_{\mathbf{z}\in\Re^m} \quad & -\mathbf{z}^T K\mathbf{z} \\ \text{subject to} \quad & \mathbf{e}^T\mathbf{z} = 0 \\ & \mathbf{z}^T D^{-1}\mathbf{z} \leq \upsilon \end{aligned} \tag{9}$$

because the ellipsoidal constraint in (9) should be active at a solution. Assuming

$K = D^{-1}WD^{-1}$, $C = \upsilon \cdot \mathbf{d}^{\frac{1}{2}}$, and approximating the rectangular constraint by the ellipsoidal constraint, problem (9) becomes an approximation to the optimization problem (8). This suggests that the normalized spectral optimization problem (1) can be regarded as an approximation to the unsupervised SVM problem (5).

The optimal separating hypothesis from an unsupervised SVM has the form

$$f(\mathbf{x}) = \left(\sum_{j=1}^{m} y_j^* \alpha_j^* K(\mathbf{x}, \mathbf{x}_j) + b^* \right)$$

where the coefficient of the bi-class separating optimal decision function α_j^* represents a measure of support from the jth data point on the two class separation decision. Since the first non-principal eigenvector of the normalized spectral clustering $\mathbf{z}^*$ yields an approximation $|\mathbf{z}^*| \approx \alpha^*$ and $\textsc{sign}(\mathbf{z}^*) \approx y^*$, $|\mathbf{z}_j^*|$ also provides a measurement of the jth data point's support on the separation. However, because of the use of the ellipsoidal constraint rather than rectangular constraints and other approximations, the eigenvector components are mostly nonzero, yielding a continuous measure of support in this two clusters separation.

3. A New Spectral Ranking for Anomaly

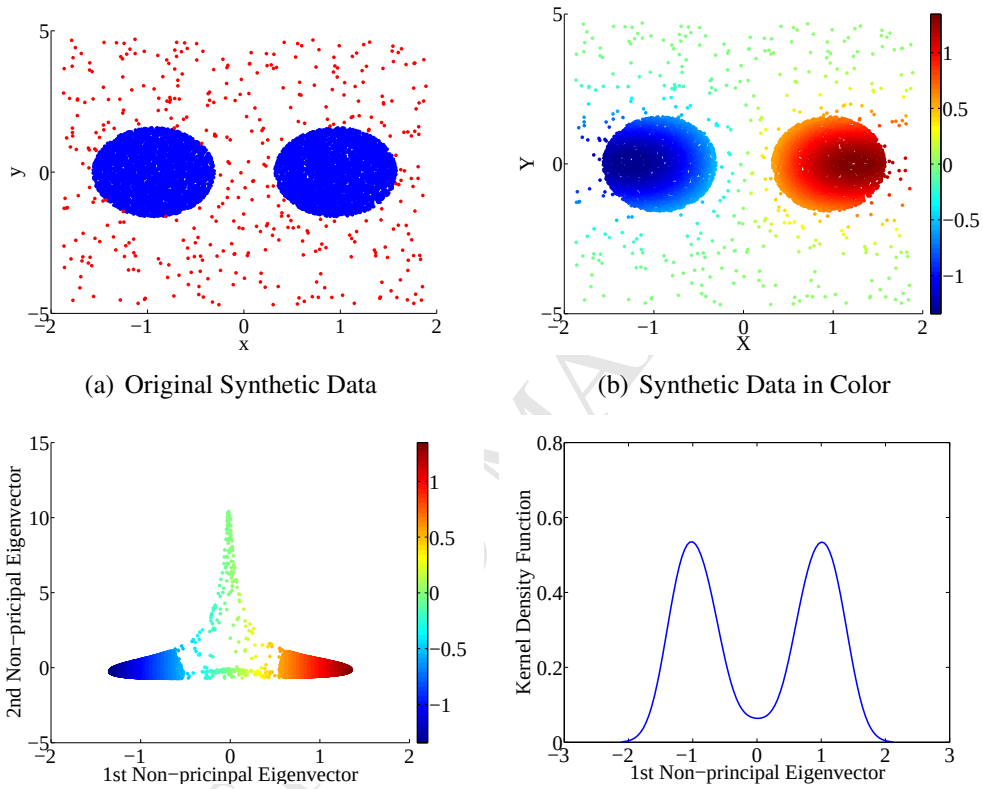
In the standard spectral analysis, k-means clustering is typically applied to non-principal eigenvectors to determine clustering memberships. Our discussion in §2 suggests that components of a non-principal eigenvector $\mathbf{z}^*$ have meaning beyond indicating cluster membership. In fact $|\mathbf{z}^*|$ provides a bi-class clustering strength measure in the optimal bi-class clustering in the high dimensional feature space according to the assumed similarity.

We now graphically illustrate the information in the component of the first non-principal eigenvector $\mathbf{z}^*$ and motivate how the information can be used to

rank anomaly. To aid visualization, we graph both components of the first and second non-principal eigenvectors, with each point in this 2-dimensional space corresponding to one original data instance. Figure 2 presents a visual illustration of the clustering strength information in the first non-principal eigenvector for a balanced two-cluster data set. We consider the Laplacian corresponding to the Gaussian kernel with bandwidth $\sigma = 1$ as the similarity. Subplot (a) graphs the original 2-D data while Subplot (c) graphs the components of the first and second non-principal eigenvectors. To see how the original data points correspond to points in the eigenvector space, we assign each data point in the dimension of the first non-principal eigenvector, $\mathbf{z}^* = D^{1/2}\mathbf{g}_1^*$, a unique color with the darker intensity corresponds to a larger magnitude in the first non-principal eigenvector component. The effect in the original data space is visualized in Subplot (b). The color of the data point in Subplot (b) is the same as the color of the corresponding point in the eigenvectors in Subplot (c). The colormap is shown at the right side of Subplot (b) and Subplot (c) in Figure 2. We also graph the probability density of the first non-principal eigenvector in Subplot (d), which clearly indicates presence of two clusters in this case.

From Figure 2, it can also be observed that the bi-class clustering strength $|\mathbf{z}^*|$ of global outliers, typically corresponding to the color green and yellow, is the smallest. In addition, data points which are closer to the other cluster, colored in light red and light blue, have approximately the medium bi-class clustering strength $|\mathbf{z}^*|$. This suggests that they offer less information in defining the clusters than that of the cluster cores, which are colored in dark red and dark blue.

In Figure 3 we consider a second synthetic example which includes major clusters in combination with small clusters. Subplots (a), (b), (c) and (d) are simi-



(a) Original Synthetic Data

(b) Synthetic Data in Color

(c) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

(d) Kernel Density Estimation

Figure 2: Visualizing Information in the First Non-principal Eigenvector: two balanced clusters with noise

larly generated as in Figure 2. From Figure 3, it can be observed that data points in two smaller clusters lie closer to the origin and are depicted in yellow and green. In addition, similar to Figure 2, we observe that global outliers mostly have the smallest clustering strength support and lie near the origin. In addition, data points which are closer to the other cluster offer less information in defining the clusters and consequently the corresponding clustering strength, $|\mathbf{z}^*|$, is smaller. We can also see, from the colormap, that the edge of the major clusters has relatively smaller clustering support strength ($|\mathbf{z}^*|$) than that of the core of the major clusters. Note that the probability density plot for the first non-principal eigenvector in Subplot (c) now has three modes, which indicates presence of additional small clusters.

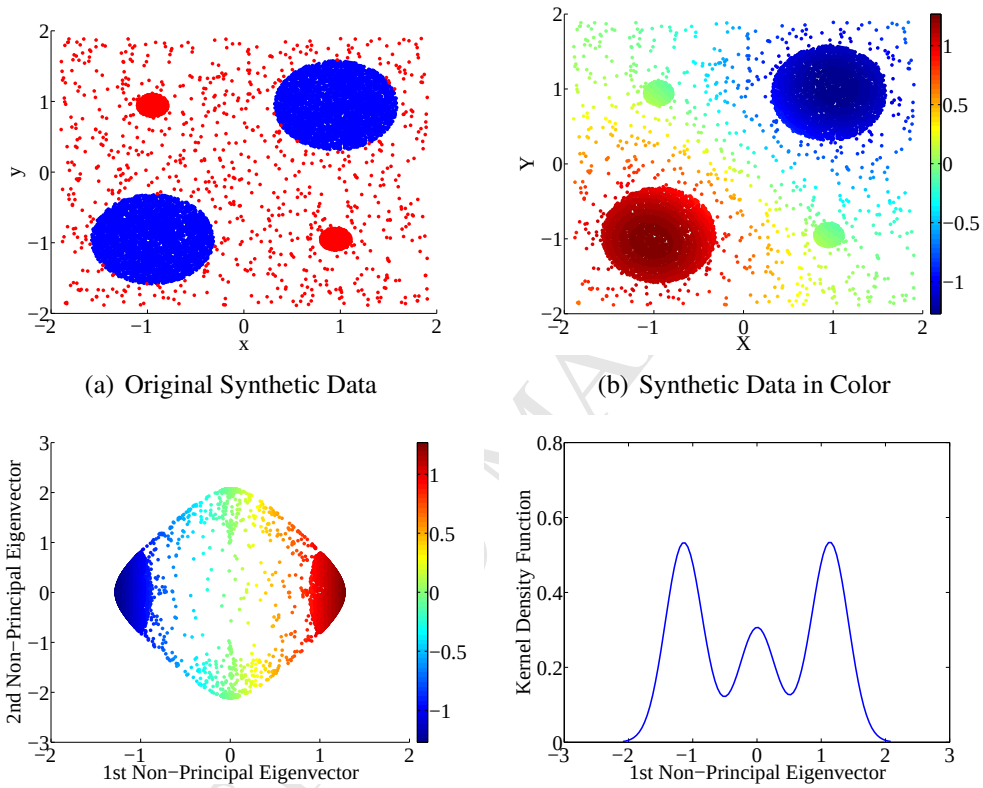
Figure 2 and 3 demonstrate that the first non-principal eigenvector indeed contains significant bi-class clustering support strength information, which can be used to produce ranking for anomaly, either global outliers or small anonymous patterns relative to major normal clusters. Using the cluster support strength information in the non-principal eigenvector, we now propose a new method of Spectral Ranking for Abnormality (SRA), which is summarized in Algorithm 1. We consider both the case when multiple major patterns exist and the case when there is only one major pattern for the normal class with smaller clusters representing abnormality (possibly also with global outliers). The proposed algorithm allows a choice of the reference in making the assessment of anomaly ranking. When ranking anomaly in referencing to multiple major classes, we define the anomaly score as $\mathbf{f}^* = \|\mathbf{z}^*\|_\infty - |\mathbf{z}^*|$. If the minority class does not have a sufficient mass, one can choose to assess anomaly likelihood with respect to a single majority class and ranking is generated suitably with this view. Otherwise, anomaly

is assessed with two main patterns. Specifically, let $\mathcal{C}_+ = \{i : (\mathbf{g}_1)_i^* \geq 0\}$ and $\mathcal{C}_- = \{i : (\mathbf{g}_1^*)_i < 0\}$ denote data instance index sets corresponding to non-negative and negative value in $\mathbf{g}_1^*$ respectively. Assume that an *a priori* upper bound for the anomaly ratio χ is given. If $\min\left\{\frac{|\mathcal{C}_+|}{n}, \frac{|\mathcal{C}_-|}{n}\right\} \geq \chi$, then neither the set $\mathcal{C}_+$ nor $\mathcal{C}_-$ is considered as an anomaly class and SRA outputs ranking $\mathbf{f}^* = \|\mathbf{z}^*\|_\infty - |\mathbf{z}^*|$ with respect to both the positive and negative majority classes. Otherwise, SRA outputs anomaly ranking $\mathbf{f}^*$ with respect to a single majority class, with $\mathbf{f}^*$ equal to either $\mathbf{z}^*$ or $-\mathbf{z}^*$ depending on which of the two classes has a smaller cardinality. If the distribution of component values of the first non-principal eigenvector is a relatively balanced bi-modal, $|\mathcal{C}_+| \approx |\mathcal{C}_-|$, SRA outputs anomaly ranking with respect to multiple patterns.

For the example depicted in Figure 3, the score $\mathbf{f}^*$ yields larger positive values for points closer to the origin, corresponding to global outliers or anonymous small clusters. The core of the major clusters have the smallest scores. The scores for the boundaries of the major clusters are in between the scores of the global outliers and the cores of the major clusters.

Figure 4 illustrates the case when there is only one major pattern for the normal class with anomaly expressed as small clusters and global outliers. For this example, the appropriate ranking score is $\mathbf{f}^* = \mathbf{z}^*$, since $\mathcal{C}_+$ is the minority class. The core of the major cluster has the smallest scores. In addition, global outliers (e.g., points in light red and yellow) have scores in between that of the cores of the major class and the minor class.

One of the main advantages of the proposed SRA over existing anomaly ranking methods is that SRA can distinguish simultaneously small clusters and global outliers from majority patterns. The existing methods typically require a user to



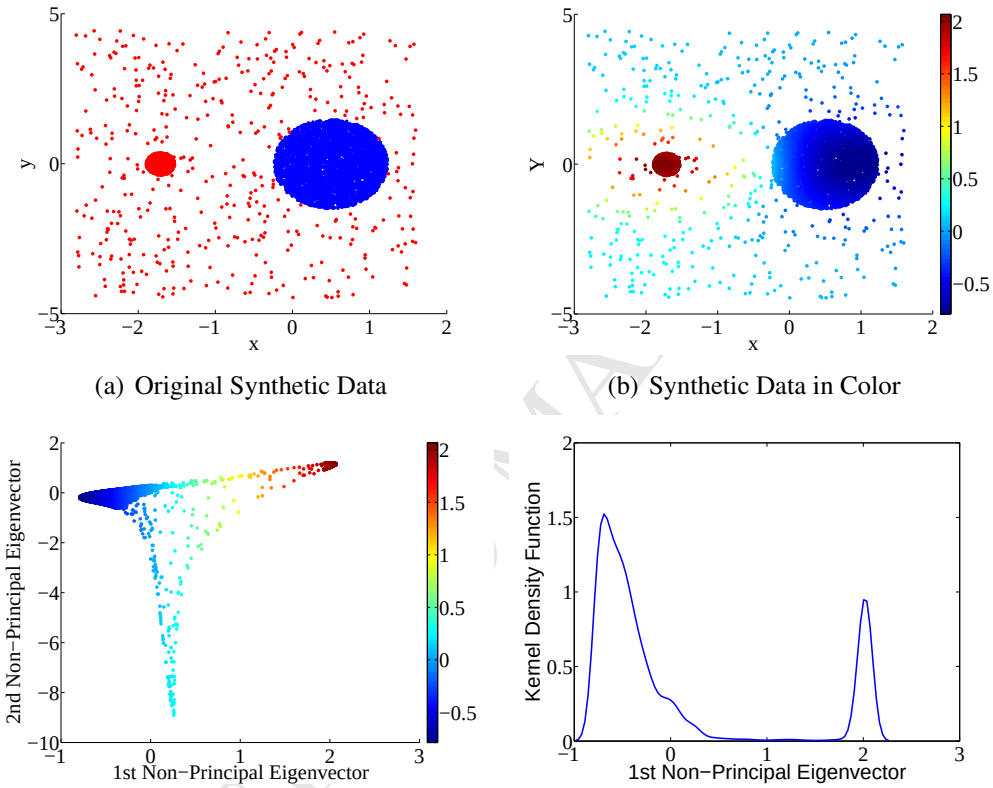
(a) Original Synthetic Data

(b) Synthetic Data in Color

(c) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

(d) Kernel Density Estimation

Figure 3: Visualizing Information in the First Non-principal Eigenvector: two major patterns and small clusters



(a) Original Synthetic Data

(b) Synthetic Data in Color

(c) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

(d) Kernel Density Estimation

Figure 4: Visualizing Information in the First Non-principal Eigenvector: one major pattern and anonymous small clusters

target one instead of both cases. In addition, our proposed SRA distinguishes edges of the main clusters from the core of the main clusters. Finally, SRA can be easily applied to both cases when there are multiple patterns for the normal class and when only one major pattern exists.

Algorithm 1: Spectral Ranking for Abnormality (SRA)

Input: W: an n-by-n similarity matrix W.
χ: a ratio upper bound on anomaly
Output: $\mathbf{f}^* \in \Re^n$: a ranking vector with a larger value representing more abnormal
mFLAG : a flag indicating the type of the ranking reference
begin
Form Laplacian $L = I - D^{-1/2} W D^{-1/2}$;
Compute $\mathbf{z}^* = D^{\frac{1}{2}} \mathbf{g}_1^*$ and $\mathbf{g}_1^*$ (the first non-principal eigenvector for L);
Let $\mathcal{C}_+ = \{i : \mathbf{z}_i^* \geq 0\}$ and $\mathcal{C}_- = \{i : \mathbf{z}_i^* < 0\}$;
if $\min\{\frac{|\mathcal{C}_+|}{n}, \frac{|\mathcal{C}_-|}{n}\} \geq \chi$ **then**
mFLAG = 1, $\mathbf{f}^* = \max(|\mathbf{z}^*|) - |\mathbf{z}^*|$, %ranking w.r.t. multiple patterns;
else if $|\mathcal{C}_+| > |\mathcal{C}_-|$ **then**
mFLAG = 0, $\mathbf{f}^* = -\mathbf{z}^*$, %ranking w.r.t. a single major pattern;
else
mFLAG = 0, $\mathbf{f}^* = \mathbf{z}^*$, %ranking w.r.t. a single major pattern ;
end
end

4. Similarity Measures for Categorical Data

In practice, representation of human activities and behavior often leads naturally to categorical and ordinal data. Since most existing data mining methods solely focus on numerical values, a common preprocessing technique for categorical data expands categorical attributes into a set of binary indicators and then use the Euclidean distance to measure the similarity. Unfortunately this simple

method often fails to capture relevant information of a data set, e.g., nominal value frequency distribution, and can potentially create distortion. Indeed we fail to obtain any meaningful results on the auto insurance data set considered in this paper, when categorical data is treated directly by performing binary expansion and applying existing clustering and outlier detection methods.

A more meaningful treatment for categorical data is to select a similarity measure to capture relationship between input categorical attributes. Indeed similarity measures have been studied for more than a century and hundreds of similarity measures exist.[19] These measures can be classified into two types, nominal value definition driven and nominal value distribution driven similarities. Next we briefly review the similarity measures which are used in this paper for the auto insurance fraud detection problem.

Nominal value definition driven similarity is defined directly from the specification of the attribute nominal values. For auto insurance fraud detection, matches and mismatches in nominal values of categorical attributes form an intuitive basis for comparing claim patterns. Hence we mainly focus on the simple overlapping similarity and its derived kernels. We assume that the data set comes from sampling of random n-dimensional categorical vector $\mathcal{D}$, with the ith attribute $\mathcal{D}_i$ having $|\mathcal{D}_i|$ distinct nominal values, $i = 1, 2, \cdots, n$.

Overlapping Similarity

Given two n-dimensional categorical attribute vector $\mathbf{x}$ and $\tilde{\mathbf{x}}$, overlapping similarity is given by

$$s^O(\mathbf{x}, \tilde{\mathbf{x}}) = 1 - d^H(\mathbf{x}, \tilde{\mathbf{x}})$$

where $d^H(\mathbf{x}, \tilde{\mathbf{x}})$ is the Hamming distance defined as the number of attributes that $\mathbf{x}$ and $\tilde{\mathbf{x}}$ do not match divided by the total number of attributes:

$$d^H(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{\sum_{i=1}^{n} \delta(\mathbf{x}_i, \tilde{\mathbf{x}}_i)}{n} \tag{10}$$

where

$$\delta(\mathbf{x}_i, \tilde{\mathbf{x}}_i) = \begin{cases} 1, & \mathbf{x}_i \neq \tilde{\mathbf{x}}_i \\ 0, & \mathbf{x}_i = \tilde{\mathbf{x}}_i \end{cases}$$

Overlapping similarity is in fact a valid kernel.[20] Replacing the Euclidean distance in a standard Gaussian kernel by the Hamming distance, we immediately obtain a Gaussian kernel below derived from the Hamming distance,

$$k^{GH}(\mathbf{x}, \tilde{\mathbf{x}}) = e^{-\frac{d^H(\mathbf{x}, \tilde{\mathbf{x}})}{2\sigma^2}} \tag{11}$$

where $\sigma > 0$ is a constant kernel width parameter.

Adaptive Gaussian Kernel

In the Gaussian Hamming Kernel (11), a single bandwidth σ is applied to every data instance. It has been argued that clustering performance can be improved using an adaptive bandwidth.[21] Considering the number of nominal values in each attribute, a weighted Hamming distance has also been proposed.[21]

$$d^{WH}(\mathbf{x}, \tilde{\mathbf{x}}) = \sum_{i=1}^{n} \frac{\delta(\mathbf{x}_i, \tilde{\mathbf{x}}_i)}{|\mathcal{D}_i|}.$$

Corresponding to the weighted Hamming distance, one can consider an adaptive kernel

$$k^{WH}(\mathbf{x}, \tilde{\mathbf{x}}) = e^{-\frac{d^{WH}(\mathbf{x}, \tilde{\mathbf{x}})}{\sigma(\mathbf{x}, \tilde{\mathbf{x}})}} \tag{12}$$

where $\sigma(\mathbf{x}, \tilde{\mathbf{x}})$ is a data driven adaptive bandwidth determined by a fixed number of nearest neighbors of the data instance $\mathbf{x}$ and $\tilde{\mathbf{x}}$.

Hamming Distance Kernel

For categorical data, Hamming distance kernel[22] has been proposed, which can be viewed as a variant of the string kernel[23].

Let $\mathcal{D}^n$ be the cross product over all n input attribute domains. The Hamming distance kernel is defined by implicitly considering the $|\mathcal{D}^n|$-dimensional attribute space in which each possible nominal value combination represents one dimension. For each $\mathbf{u} \in \mathcal{D}^n$, representing one dimension in the $|\mathcal{D}^n|$-dimensional kernel attribute space, an explicit mapping $\theta_{\mathbf{u}}(\mathbf{x})$ is defined to map a data instance $\mathbf{x}$ onto this dimension. For any given original input attributes $\mathbf{x}$ and $\tilde{\mathbf{x}}$, the kernel function $k(\mathbf{x}, \tilde{\mathbf{x}})$ equals the inner product of the mapped attributes $\{\theta_{\mathbf{u}}(\mathbf{x}), \mathbf{u} \in \mathcal{D}^n\}$ and $\{\theta_{\mathbf{u}}(\tilde{\mathbf{x}}), \mathbf{u} \in \mathcal{D}^n\}$. More specifically, let $\mathcal{D}_i$ be the domain of the ith attribute. For each $\mathbf{u} \in \mathcal{D}^n$, given a categorical instance $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i \in \mathcal{D}_i$, we define an explicit mapping:

$$\theta_{\mathbf{u}}(\mathbf{x}) = \lambda^{d^H(\mathbf{u},\mathbf{x})} \tag{13}$$

where $\lambda \in (0, 1)$ is a damping parameter and $d^H(\cdot)$ is the Hamming distance (10). Note that $\theta_{\mathbf{u}}(\mathbf{x})$ is only one dimension of the kernel feature space. Thus the Hamming distance kernel between instances $\mathbf{x}$ and $\tilde{\mathbf{x}}$ is:

$$k^H(\mathbf{x}, \tilde{\mathbf{x}}) = \sum_{\mathbf{u} \in \mathcal{D}^n} \theta_{\mathbf{u}}(\mathbf{x}) \theta_u(\tilde{\mathbf{x}}) \tag{14}$$

Directly computing the Hamming distance kernel has an exponential computational complexity.[22] Fortunately a dynamic programming technique can be applied, which allows this kernel to be computed efficiently following a recursive

procedure.[22]

While evaluating similarity between a pair of instance $\mathbf{x}$ and $\tilde{\mathbf{x}}$, the Hamming distance directly compares nominal values of $\mathbf{x}$ and $\tilde{\mathbf{x}}$ to determine the number of different values. In contrast, a Hamming distance kernel assesses the similarity between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ in referencing to all possible nominal value combinations of the categorial attribute. Consequently, a Hamming distance kernel can capture more information than the Hamming distance based on the simple overlapping similarity.

DISC Similarity

To illustrate the effect of the similarity measure on the performance of a ranking method, we consider in this paper DISC (Data-Intensive Similarity Measure for Categorical Data)[24], which is an example of distribution driven similarity measure for categorical data. Distribution driven similarity is determined from distributions of occurring feature values. This type of similarity measures may adjust similarity for rare nominal values, which introduces additional information in representing the relationship between data instances. This can potentially detect different patterns in comparison to simpler feature definition driven similarity measures.

5. Performance Evaluation for SRA

To evaluate effectiveness of the proposed SRA for anomaly detection, we examine the Receiver Operating Characteristic (ROC)[25,26,27] to obtain a class skew independent performance measure. We regard the anomaly cases as the positive class and the normal cases as the negative class. Assume that n_+ and n_- denote the total number of instances in the positive and negative class respectively. The

ROC graph is obtained by plotting the true positive rate (number of true positives divided by n_+) against the false positive rate (number of false positives divided by n_-), as the threshold level is varied. The true positive rate (also known as sensitivity) is one evaluation criterion and the false positive rate (or one minus specificity) is the second evaluation criterion.

The ROC curves (or true-positive and false-positive frontiers) depict the trade-off between the two criteria, benefits (true positive) and costs (false positives), for different choices of the threshold. Thus it does not depend on *a priori* knowledge to combine the two objectives into one. A ROC curve that dominates another provides a better solution at any cost point and it corresponds to a higher area under the curve (AUC). The AUC yields the probability that the generated ranking places a positive class sample above a negative class sample, when the positive sample is randomly drawn from the positive class and the negative class sample is drawn randomly from the random class respectively.[28] Thus, the ROC curves and AUC can be used as a criteria to measure how well an algorithm performs on certain data sets. Subsequently, we will use AUC and ROC curves as performance evaluation measures to compare different methods.

Synthetic Examples

To illustrate, we first present, in Subplot (a) in Figure 5, performance of SRA on the synthetic data sets depicted in Figure 2–4. For the synthetic data set 1 & 2, there are multiple main patterns which can be deduced from density plots of the first non-principal eigenvector; consequently mFlag is set to 1. For the synthetic dataset 3, mFlag is set to zero to reflect the observation that the positive class does not have sufficient instances. Here the performance is assessed regarding the small cluster and outliers as constituting anomaly. We observe very high AUCs

0.95, 0.99, 0.99, respectively for the synthetic data set 1, 2 & 3.

Auto Fraud Detection

To further evaluate effectiveness of the proposed SRA, we now apply it to a fraud detection problem from auto insurance claim data, which has been used for the supervised detection[6]. This is the only publicly available auto insurance fraud detection data that we can find from the academic literature. This data set, which is provided by Angoss KnowledgeSeeker Software, consists of 15420 claim instances from January 1994 to December 1996. The data set consists of 6% fraudulent labels and 94% legitimate labels, with an average of 430 claims per month. In addition, the data set has 6 ordinal features and 25 categorical attributes. Feature examples include base policy, fault, vehicle category, vehicle price (6 nominal values), month of accidents, make of the car, accidental area, holiday, and sex. Intuitively, anomaly in this case should be assessed with respect to nominal value combinations. Consequently the Hamming distance and Hamming distance based kernels are reasonable similarities to use.

This data set is used to assess achievable prediction quality using supervised learning.[6] Since labels for auto insurance claims are generally not available at the detection time, here we apply the proposed *unsupervised* SRA to this claim data set. In other words,ranking is generated without using labels and the labels are used only for performance evaluation.

When performing unsupervised fraud detection on this data, we recall two major challenges which have been briefly mentioned in previous sections. Firstly, most of the features in this dataset are categorical or ordinal. Secondly, unlike common anomaly detection problems, the claim data forms multiple patterns. Consequently a single cluster based global outlier detection method generally pro-

duces unsatisfactory results.

For comparison, we include performance of two popular unsupervised outlier detection methods, one-class SVM (OC-SVM) and Local Outlier Factor (LOF). The implementation of OC-SVM comes from LIBSVM library.[29] The implementation of LOF comes from Ddtools library.[30] In addition, we train *supervised* Random Forest (RF) on the full dataset, since the supervised training accuracy of RF can then be used as an upper bound for evaluations of unsupervised learning methods. For the RF computational results, the number of trees built is 500 and the number of features used for building each tree is 6. The resulting votes are used as the decision value. The implementation of RF comes from the Treebagger class in Matlab software.

We note that, the only parameter required by SRA is the upper bound on the anomaly rate, which we believe that it is reasonable to expect a crude approximation at least in practice. In contrast, LOF requires a parameter for the number of neighborhood n_{LOF} and OC-SVM requires an additional width parameter μ_{SVM}. Unfortunately it is far more difficult to determine appropriate values for these parameters. This can present more challenges for unsupervised learning since there is no mechanism to tune their values. For the auto insurance data set considered here, since $|\mathcal{C}_+| \approx |\mathcal{C}_-|$ for all the similarities we have experimented with, subsequently we always report ranking with regard to multiple patterns, i.e., mFLAG=1. Consequently the choice of χ is practically irrelevant in this case.

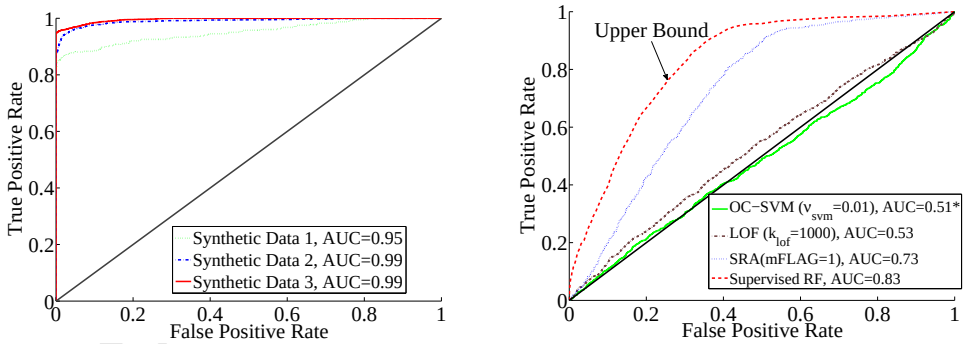
Comparisons with LOF and OC-SVM Using Overlapping Similarity

Subplot (b) in Figure 5 presents ROC curves for SRA, LOF and OC-SVM, using overlapping similarity. In addition, ROC of the supervised RF is also included as a benchmark. For LOF and OC-SVM however, parameter choices are

Table 1: Summary of the AUCs for the Auto Insurance Fraud Detection Data Set

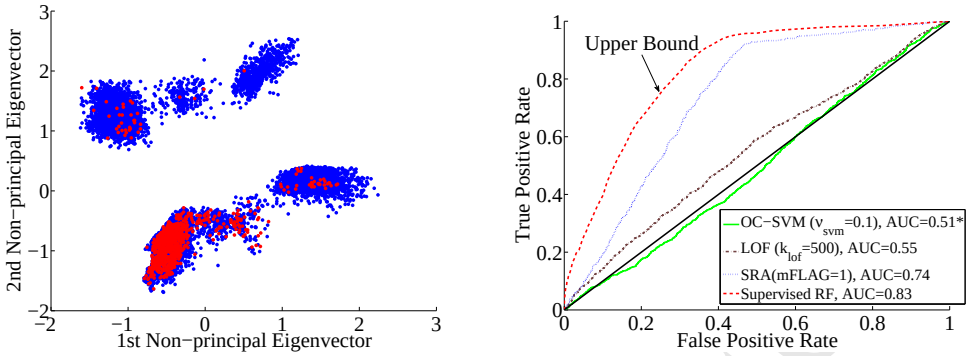
Automobile Fraud Detection Data Set										
Method			OS	AGK				HDK		DISC
				β				λ		
				10	100	1000	3000	0.5	0.8	
LOF	n_{lof}	10	0.53	0.5	0.52	0.58	0.64	0.52	0.53	0.55
		100	0.51	0.51*	0.54	0.58	0.67	0.51	0.52	0.57
		500	0.53	0.52*	0.55	0.59	0.68	0.51	0.51	0.57
		1000	0.53	0.52*	0.53	0.59	0.69	0.5	0.5	0.56
		3000	0.5	0.58*	0.55	0.58	0.69	0.54*	0.55*	0.53
OC-SVM	μ_{svm}	0.01	0.51*	0.53*	0.51*	0.54	0.59	0.51*	0.52*	0.53*
		0.05	0.51*	0.53*	0.51*	0.55	0.59	0.52*	0.53*	0.52*
		0.1	0.51*	0.54*	0.51*	0.55	0.59	0.53*	0.54*	0.56*
SRA	mFLAG	1	**0.73**	**0.74**	**0.74**	**0.66**	**0.74**	**0.74**	**0.74**	**0.66**

For entries marked by *, AUC reported is one minus the actual AUC (≤ 0.5).
OS:Overlapping Similarity, AGK: Adaptive Gaussian Kernel
HDK:Hamming Distance Kernel, DISC: DISC Similarity



(a) (b) ROC Curves

Figure 5: Subplot (a) Displays ROC and AUC of SRA for Synthetic Examples. Subplot (b) Compares SRA with LOF and OC-SVM on Auto Insurance Data Set Based on the Overlapping Similarity. SRA clearly dominates LOF and OC-SVM.



(a) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

(b) ROC Curves

Figure 6: Comparisons Based on an Adaptive Gaussian Kernel with $\beta = 100$

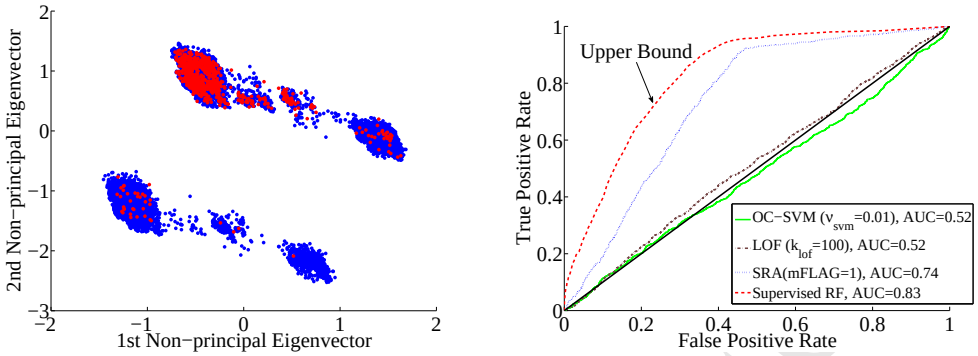
important and we report the optimal AUCs among different parameters we have experimented with. It can be observed that the ROC from SRA clearly dominates those from LOF and OC-SVM using overlapping similarity.

Table 1 summarizes the AUCs from different unsupervised methods using different similarities. Here we use β to denote the neighborhood size parameter of the adaptive Gaussian kernel in (12). The results indicate that SRA outperforms LOF and OC-SVM significantly on all similarities considered here.

Next we present more detailed discussions on the comparison in terms of ROCs for different kernels and different methods. Unless otherwise noted, we always include performance of the supervised RF to show the upper bound.

Using Adaptive Gaussian Kernel with Weighted Hamming Distance.

Figure 6 shows ROC curves for SRA, LOF, OC-SVM achieved with the Adaptive Gaussian Kernel with the weighted Hamming distance and neighborhood size $\beta = 100$. We observe that AUC for each of SRA, LOF and OC-SVM is improved. We conjecture that the improvement comes from the facts that the weighted Ham-



(a) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

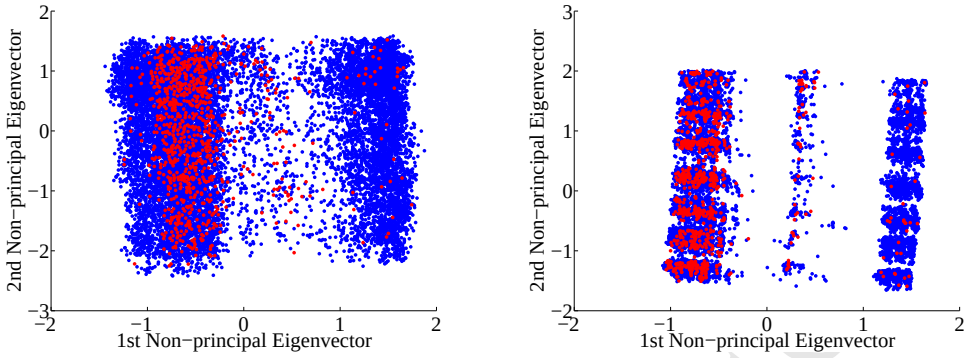
(b) ROC Curves

Figure 7: Comparisons Based on a Hamming Distance Kernel with $\lambda = 0.8$

ming distance is a better distance measures than Hamming distance since information on the number of distinct value in each feature is also included. From Figure 6 (a), we also observe that, using the adaptive Gaussian kernel, clusters are also more distinct comparing to the overlapping kernel. Note that in Figure 6 (a), Figure 7 (a), and Figure 8, red points are the corresponding points of fraudulent cases in the space of first and second non-principal eigenvectors.

Using the Hamming Distance Kernel

The Hamming distance kernel is defined based on the overlapping similarity measure. However, as discussed previously, a Hamming distance kernel can capture more information than the simple Hamming distance (overlapping similarity). Recall that, while Hamming distance directly compare nominal values of a pair of data instances to determine the number of different values, a Hamming distance kernel assesses the similarity between the pair in referencing to all possible nominal value combinations of categorial attributes. From Figure 7 for which the Hamming Distance Kernel is used, we observe a more complex cluster structure



(a) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

(b) Visualization of $\mathbf{z}_1^* = D^{1/2}\mathbf{g}_1^*$ and $\mathbf{z}_2^* = D^{1/2}\mathbf{g}_2^*$

Figure 8: Overlapping Similarity and DISC Similarity:using DISC

, even though SRA achieves a similar 0.74 AUC. Although SRA achieves similar AUCs when λ value is changed, different clustering structures are observed.

Using DISC Similarity

We have also experimented with a few distribution driven similarity measures, e.g., Lin similarity[19]; the results are consistently worse (less than 0.6 AUC) in comparison to overlapping, adaptive Gaussian and Hamming distance kernel. The only exception is that of the DISC similarity measure, for which AUC = 0.52 for OC-SVM (μ_{svm} = 0.05), AUC= 0.56 for LOF (n_{lof} = 1000), AUC= 0.66 for SRA (mFLAG=1). Once again, performance of SRA significantly dominates LOF and OC-SVM. Figure 8 compares cluster structure revealed using overlapping similarity and DISC similarity respectively. We observe here that different cluster structures are revealed when different similarity measures are used.

Understanding Cluster Structure: Further Validation of SRA Ranking

To further justify SRA ranking generated, we further investigate the cluster structure revealed. Specifically we analyze significant features based on which

highly ranked claims are different from the majority. If deviation from the majority is based on features which are likely to lead to fraud, this provides support for the generated ranking. As an example, we consider clusters identified using Hamming kernel with $\lambda = 0.8$.

Figure 9 is identical to Subplot (a) in Figure 7, except that clusters formed in the space of the first and second non-principal eigenvectors are explicitly labeled. We report the fraud ratio of each cluster in Table 2. It can be seen from the plot and the table that 92% of the fraudulent cases actually reside in the clusters 4, 6 and 7. These are also the clusters that have relatively high anomaly score from SRA. Hence we further analyze each cluster, especially the one with the highest fraud ratio.

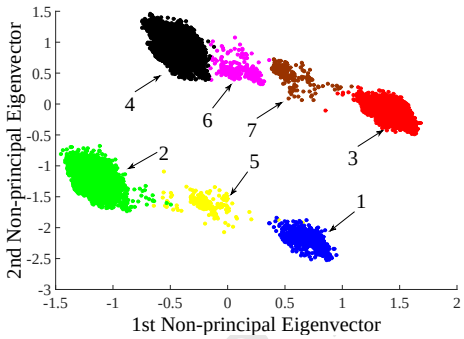


Figure 9: Clusters with Labels

Cluster	n	n_-	n_+	fraud %
1	723	722	1	0.14%
2	3261	3228	33	1.01%
3	4266	4231	35	0.82%
4	6423	5622	761	11.85%
5	206	203	3	1.46%
6	340	294	46	13.53%
7	201	157	44	21.89%

Table 2: Cluster Summary Information

To gain additional insight, we build a standard CART (classification and regression tree) to determine the decision rule for each cluster, with the pruning rule that the number of data points at a leaf node is no less than 15. The training labels for CART are the cluster labels we manually identify from the first and second non-principal eigenvectors and the training data is the whole data set. Figure 10 presents the computed CART tree.

We discover the following rules from Figure 10 for the cluster 4, 6 and 7, for

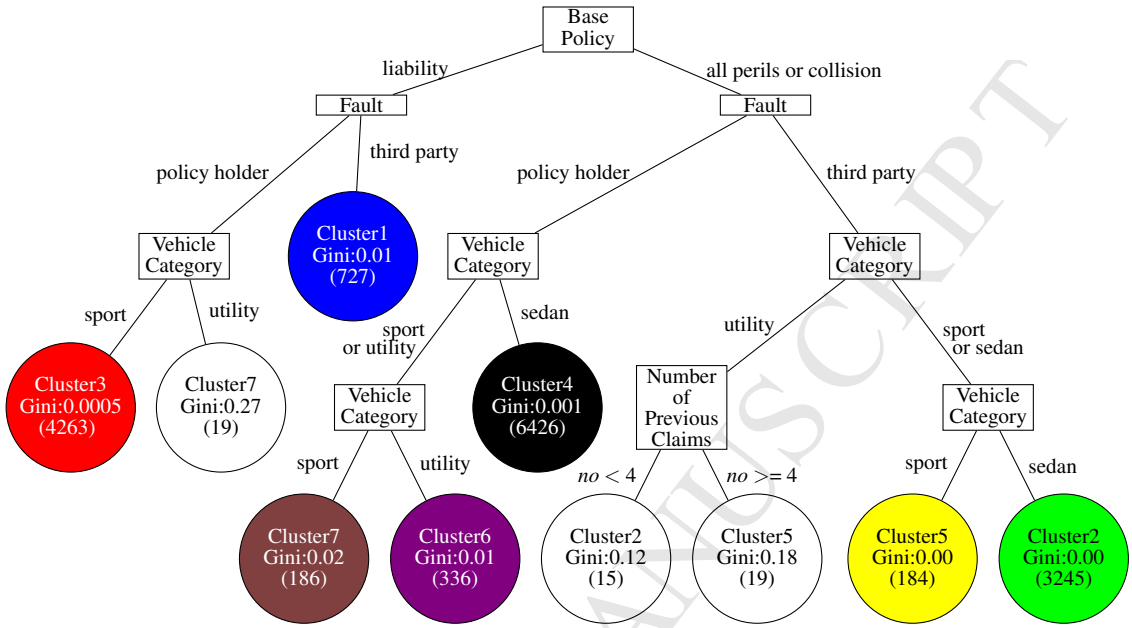


Figure 10: Decision Tree: rules leading to the colored clusters

which SRA has assigned high ranks:

- If the insurance policy is for *collision* or *all perils* and it is the policy holder who causes the accident (policy holder at fault), the corresponding claim belongs to the clusters with high fraud ratio (cluster 4, 6, & 7).
- If the insurance policy is for *liability* and/or and it is **not** the policy holder who causes the accident (third party at fault), the corresponding claim belongs to the other clusters (cluster 1, 2, 3, and 5).
- Following the first rule, if the policy holder drives *sports car*, the corresponding claim belongs to cluster 7. If the policy holder drive *utility car*, the corresponding claim belongs to cluster 6. Otherwise, the corresponding claim belongs to cluster 4.

Indeed these rules identify cases with reasonable suspiciousness. In addition, from training *supervised* random forest in the previous section, we have discovered that the three most important features in classify fraudulent cases against legal cases are **base policy**, **car types** and **fault**. These three features are actually the features used in defining the clusters. This analysis supports that the ranking from SRA is meaningful and reasonable.

6. Concluding Remarks

In this paper, we propose a spectral ranking method for anomaly detection. We observe that the spectral optimization problem can be interpreted as an approximation to an unsupervised support vector machine and a non-principal eigenvector can be used to derive a ranking vector directly. Based on this information in an eigenvector, a data instance is more likely to be an anomaly if its magnitude is smaller, when both positive and negative classes cannot be ruled as abnormal based on instance count percentages. Furthermore, we allow a choice of the reference in the assessment of anomaly ranking. If the minority class does not have a sufficiently large count percentage, one can choose to assess anomaly likelihood with respect to a single majority class and ranking is generated suitably with this view. Otherwise, anomaly is assessed with two main patterns.

As an illustration of the proposed SRA, we consider the challenging auto fraud insurance detection problem based on a real claim data set. Since obtaining labels is time consuming, costly, and error prone in practice, we model the problem as unsupervised learning and ignore labels when generating ranking using SRA, even though fraud labels are available for this particular data set. Since data attributes are categorical, we assess anomaly in nominal value combinations which lead to suspiciousness of the claim. We choose the Hamming distance and Hamming

distance based kernels in generating spectral ranking for this data set. SRA yields an impressive 0.74 AUC, particularly considering that the supervised RF generates 0.83 AUC, which can be regarded as an upper bound.

Finally we note that, although we focus in this paper on applying the proposed method for the auto insurance fraud detection, it can be applied to other anomaly detection problems as well. We recognize however that appropriate choice of a similarity measure may depend on the specific application context.

References

1. Tennyson Sharon, Salsas-Forn Pau. Claims auditing in automobile insurance: fraud detection and deterrence objectives *Journal of Risk and Insurance.* 2002;69:289–308.

2. Derrig Richard A. Insurance fraud *Journal of Risk and Insurance.* 2002;69:271–287.

3. Herbrich Ralf, Graepel Thore, Obermayer Klaus. *Large Margin Rank Boundaries for Ordinal Regression*. MIT Press 2000.

4. Tayal A. D., Coleman T. F., Li Y.. RankRC: Large-scale Nonlinear Rare Class Ranking *IEEE Transactions on Knowkedge and Data Engineering.* 2015;27:3347-3359.

5. Tayal A. D., Coleman T. F., Li Y.. Bounding the Difference Between RankRC and RankSVM and an Application to Multi-Level Rare Class Kernel Ranking . 2015;submitted.

6. Phua Clifton, Alahakoon Damminda, Lee Vincent. Minority report in fraud detection: classification of skewed data *ACM SIGKDD Explorations Newsletter.* 2004;6:50–59.

7. Phua Clifton, Lee Vincent, Smith Kate, Gayler Ross. A comprehensive survey of data mining-based fraud detection research *arXiv preprint arXiv:1009.6119.* 2010.

8. Brockett Patrick L, Xia Xiaohua, Derrig Richard A. Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud *Journal of Risk and Insurance.* 1998:245–274.

9. Brockett Patrick L, Derrig Richard A, Golden Linda L, Levine Arnold, Alpert Mark. Fraud classification using principal component analysis of RIDITs *Journal of Risk and Insurance.* 2002;69:341–371.

10. Ai Jing, Brockett Patrick L, Golden Linda L, Guillén Montserrat. A robust unsupervised method for fraud rate estimation *Journal of Risk and Insurance.* 2012.

11. Ai Jing, Brockett Patrick L, Golden Linda L. Assessing Consumer Fraud Risk in Insurance Claims: An Unsupervised Learning Technique Using Discrete and Continuous Predictor Variables *North American Actuarial Journal.* 2009;13:438–458.

12. Gretton A., Bousquet O., Smola A. J., Schölkopf B.. Measuring statistical dependence with Hilbert-Schmidt norms in *Proceedings of the International Conference on Algorithmic Learning Theory* (Jain S., Simon H. U., Tomita E.. , eds.):63–77Springer-Verlag 2005.

13. Song Le, Smola Alex, Gretton Author, Bedo J., Borgwardt K.. Feature Selection via Dependence Maximization *Journal of Machine Learning Research.* 2012;13:1393–1434.

14. Von Luxburg Ulrike. A tutorial on spectral clustering *Statistics and computing.* 2007;17:395–416.

15. Ng Andrew Y, Jordan Michael I, Weiss Yair, others . On spectral clustering: Analysis and an algorithm *Advances in neural information processing systems.* 2002;2:849–856.

16. Shi Jianbo, Malik Jitendra. Normalized cuts and image segmentation *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* 2000;22:888–905.

17. Karnin Zohar, Liberty Edo, Lovett Shachar, Schwartz Roy, Weinstein Omri. Unsupervised SVMs: On the complexity of the furthest hyperplane problem *Journal of Machine Learning Research.* 2012;1:18.

18. Horst Reiner, Pardalos Panos M. *Handbook of global optimization*;2. Springer Science & Business Media 2013.

19. Boriah Shyam, Chandola Varun, Kumar Vipin. Similarity Measures for Categorical Data: A Comparative Evaluation in *SDM*:243-254SIAM 2008.

20. Gärtner Thomas, Le Quoc V., Smola Alex J.. A short tour of kernel methods for graphs tech. rep.NICTA, AustraliaCanberra 2006.

21. David Gil, Averbuch Amir. SpectralCAT: Categorical spectral clustering of numerical and nominal data *Pattern Recognition.* 2012;45:416–433.

22. Couto Julia. Kernel k-means for categorical data in *Advances in Intelligent Data Analysis VI*:46–56Springer 2005.

23. Lodhi Huma, Saunders Craig, Shawe-Taylor John, Cristianini Nello, Watkins Chris. Text classification using string kernels *The Journal of Machine Learning Research.* 2002;2:419–444.

24. Desai Aditya, Singh Himanshu, Pudi Vikram. DISC: data-intensive similarity measure for categorical data in *Advances in Knowledge Discovery and Data Mining*:469–481Springer 2011.

25. Provost Foster, Fawcett Tom, Kohavi Ron. The Case Against Accuracy Estimation for Comparing Induction Algorithms in *In Proceedings of the Fifteenth International Conference on Machine Learning*:445–453 1997.

26. Bradley Andrew P.. The use of the area under the ROC curve in the evaluation of machine learning algorithms *Pattern Recognition.* 1997;30:1145–1159.

27. Metz C. E.. Basic principles of ROC analysis. *Seminars in nuclear medicine.* 1978;8:283–298.

28. DeLong Elizabeth R., DeLong David M., Clarke-Pearson Daniel L.. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach *Biometrics.* 1988;44:837–845.

29. Chang Chih-Chung, Lin Chih-Jen. LIBSVM: a library for support vector machines *ACM Transactions on Intelligent Systems and Technology (TIST).* 2011;2:27.

30. Tax D.M.J.. DDtools, the Data Description Toolbox for Matlab 2015. version 2.1.2.