A L_2 Discrepancy Learning Process with Applications to Outlier and Insider Detections With Large High-Dimensional Data

Faysal El Khettabi*

Abstract

In this paper, a discrepancy-based framework is first presented for outlier and insider detections purpose. Given any sequence of profiles, a local discrepancy first identifies regions where the profiles are clumped or scarce then a global L_2 discrepancy summarizes the overall distribution patterns of the data into one real value. A L_2 discrepancy learning process is formulated to rank each profile in the sequence on the basis of optimizing the L_2 discrepancy value. This L_2 discrepancy learning process allows an access to many levels of information about outliers and insiders in the data. Experimental results are given to demonstrate the application of the L_2 discrepancy learning process with different features data sets showing that the algorithm efficiently detects the outliers and insiders in the data.

1 Introduction and motivation

Data mining can be defined as an information extraction activity whose goal is to discover hidden facts contained in a data set; see a review in [1]. Procedures and algorithms designed to analyze the sequence of profiles[1] are called data mining method, using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results.

Outlier and insider detections are an outstanding data mining task that has a lot of practical applications in many different domains. For instance, outlier detection can be defined as follows: Given a set of data points or objects find subsequent profiles that are considerably dissimilar, exceptional or inconsistent with respect to the remaining data, see [2] for a complete review.

In the present work, we are concerned by a new outlier and insider definitions based on the mathematical framework of discrepancy. Essentially, discrepancy gives a global indication about the non-uniformity of the distribution of a sequence of profiles in s-dimensional hypercube. Under this framework, the local discrepancy identifies the sparse and clumped regions among all multiresolution grid cells, thereby L_2 discrepancy summarizes the overall distribution patterns of the data into one real value.

Thus we derive a discrepancy learning process that ranks each profile on the basis of optimizing the L_2 discrepancy relaying only on the sequence profiles without assuming any specific model form in the data set or using any external parameters. This learning process recovers the visibility of the data i.e. where the data is considerably dissimilar(outliers) and more similar(insiders). Nevertheless the learning process allows an access to many quantitative levels of information to track the most outliers and insiders.

2 Methods and approach

We are given a finite data sequence with N profiles with a finite set of s variables describing properties of a profile; the context will always show what is meant. We are concerned with quantitative values for variables. Mathematically, the values for variables are in $I^s = \prod_{i=1}^{s}[a_i, b_i] \subset \mathbf{R}^s$. We scale and translate each interval $[a_i, b_i]$ to the interval $[0, 1]$. We can assume $I^s = [0, 1)^s$ without lost of generality. The sequence

*Laboratory for Threat Material Detection, University of New Brunswick, Fredericton, New Brunswick, E3B 5A3, Canada.

[1] In the sequel, we use the expressions "data set", "sequence of points" and "sequence of profiles" with the same meaning.

of profiles is $X = \{x_n\}_{1 \leq n \leq N}$ where each profile is an s-dimensional vector $x_n = (x_{n,1}, x_{n,2} \ldots, x_{n,s}) \in I^s$.

Our deterministic framework to profiles analysis is based on the mathematical foundation of discrepancy, the discrepancy will measure how evenly the profiles are scattered in space, I^s. In the next subsection 2.1, we recall from [5] the basic concepts of the mathematical formulation of discrepancy.

2.1 Mathematical formulation of discrepancy

Let λ_s denote the s-dimensional Lebesgue measure. The Lebesgue measure is the standard way of assigning a volume to subsets of Euclidean space, for instance I^s. It is used throughout real analysis, in particular to define Lebesgue integration. Sets which can be assigned a volume are called Lebesgue measurable; the volume or measure of the Lebesgue measurable set $J \subset I^s$ is denoted by $\lambda_s(J)$. We note by χ_J the characteristic function of a set J:

$$\chi_J(x) = \begin{cases} 1 & \text{if} \quad x \in J, \\ 0 & \text{if} \quad x \notin J. \end{cases} \tag{1}$$

The function χ_J has the value 1 on the set J, and is zero elsewhere. The set of discontinuous points is the boundary of the set J.

A bounded set J is called **Jordan measurable** if its boundary, δJ is of measure zero, $\lambda_s(\delta J) = 0$. In this case, the Lebesgue integral of its characteristic function, χ_J, exists, since the measure of the set of discontinuous points of the function, χ_J, is zero. The value of this integral is called the content of J or its (s-dimensional) volume,

$$\lambda_s(J) = \int_{I^s} \chi_J(x) dx. \tag{2}$$

An infinite sequence of profiles, $\left\{x_\ell\right\}$, is called uniformly distributed or scattered in I^s if, in the limit, the number of profiles x_n falling in any given subsets J of I^s is proportional to its volume. Mathematically, $\left\{x_\ell\right\}$ is uniformly distributed or scattered in I^s if for all **Jordan measurable** subsets J of I^s

$$\lim_{N \to \infty} \frac{1}{N} \sum_{\ell=1}^{N} \chi_J(x_\ell) = \lambda_s(J), \tag{3}$$

holds.

In practice, a sequence of profiles, X, have a finite number of profiles, it is necessary to define some measure of uniformity for finite sequence of profiles. Such a quantity is know as discrepancy. For a subinterval J of I^s the *local discrepancy* is defined by

$$D_N(J, X) = \frac{1}{N} \sum_{1 \leq \ell \leq N} \chi_J(x_\ell) - \lambda_s(J). \tag{4}$$

Thus the more uniformly space filling the sequence of profiles is, the smaller its *local discrepancy*. The *local discrepancy* allows us to observe locally how the profiles are scattered in the domain I^s. The positive value of the *local discrepancy* indicates a high number of profiles share common characteristics in J. Alternatively, the *local discrepancy* is also a sparsity indicator of profiles as a negative value indicates that the presence of the profiles is considerably lower in J. Therefore the mathematical formulation of the *local discrepancy* handles the similarity and the sparsity which are an useful asset to measure globally how a sequence of profiles is scattered in the domain I^s.

2.2 L_2 discrepancy for profiles analysis

A scientific goal in data mining is to find groups of profiles which are significantly correlated with each other. In addition to identifying outlier profiles showing an exceptional behaviors. Then a natural question " How should we select a *global discrepancy* framework to derive this knowledge?"

By restricting the subinterval J to a certain class of sets and taking a norm of $D_N(J, X)$ over this class, various kinds of discrepancy can be defined as quantitative measures of the uniformity of sequence X. Note that the important sets for describing what is called the topological structure of the Euclidean space I^s are the open subsets $J = \prod_{i=1}^{s}]a_i, b_i[$ with

$\mathbf{a} = (a_1, a_2 \ldots, a_s)$ and $\mathbf{b} = (b_1, b_2 \ldots, b_s)$ are in I^s. Therefore, the L_2 integration of $D_N(J, X)$ over this subrectangles J enable us to give a topological characterization of the notion of complexities and features of the sequence of profiles which is implicitly framed in terms of the Euclidean distance, i.e. explicitly formulated via the metric topology based on the Euclidean metric. The mathematical definition of L_2 discrepancy is

$$\mathbf{T}_N(X) = \left[\int_{(\mathbf{a},\mathbf{b}) \in I^{2s}, a_i < b_i} \Big(D_N(J, X) \Big)^2 d\mathbf{a} d\mathbf{b} \right]^{\frac{1}{2}}, \tag{5}$$

This L_2 discrepancy uses all multiresolution subrectangles of I^s, hence gives an indication without loss of informational content of the sequence's features. Each subrectangle summarizes the information of a group profiles that map into it by a value computed via the equation 4. Thus by summing over all multiresolution subrectangles, L_2 discrepancy summarizes the overall distribution patterns of the data into one real value which can be computed directly from the analytical formulas of L_2 discrepancy:

$$\Big(\mathbf{T}_N(X) \Big)^2 = A + B + 12^{-s}, \tag{6}$$

where

$$A = \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \prod_{i=1}^{s} \Big(1 - \max(x_{n,i}, x_{m,i}) \Big) \min(x_{n,i}, x_{m,i}),$$

and

$$B = \frac{2^{1-s}}{N} \sum_{n=1}^{N} \prod_{i=1}^{s} x_{n,i}(1 - x_{n,i}).$$

A detailed derivation of this analytical formulas of L_2 discrepancy is in the paper [4].

This analytical formulas of L_2 discrepancy gauges the irregularity of profiles distribution by using the entire set of profiles conforming by that the holistic nature of L_2 discrepancy.

2.3 L_2 Discrepancy learning process

The L_2 discrepancy value captures the global degree of isolation of the sequence of profiles in the domain, more higher is the value of L_2 discrepancy, more the sequence is scarce in some parts of the domain. For instance, if profiles are tightly coupled, they will tend to occupy a small region of space by consequence the $\mathbf{T}_N(X)$ will have a higher value. However, if object profiles are loosely coupled then their profiles tend to occupy a large region space. Thus the $\mathbf{T}_N(X)$ will be a small value.

In practice, however, one have to response the following question: How different is a specific profile x_n from other profiles? For a response we need to measure the importance of each profile in the data set and rank profile by the degree of similarity accordingly to a specific criteria.

A L_2 discrepancy learning process will rank each profile on the basis of optimizing the L_2 discrepancy, i.e. the profile is far-sighted to minimize the L_2 discrepancy value by removed it from the sequence, the profile only concerned to keep the same L_2 discrepancy value or finally the profile needs another point to be added in its immediate neighborhood to minimize the L_2 discrepancy:

DEFINITION **1** *A profile, x_n, is called insider if the L_2 discrepancy value will be minimized by removing, x_n, from the sequence of profiles.*

A profile, x_n,is called outlier if the L_2 discrepancy value will be maximized by removing, x_n, from the sequence of object profiles.

One of the key factors that will provide more information about the complexities and features in the profiles, is to define a weighted L_2 discrepancy that will offer a learning framework with many levels of information about the interdependence in the profiles.

We introduce the weighted L_2 discrepancy as:

$$\begin{aligned} (\mathbf{T}_N(X, \omega))^2 &= \Big| \frac{1}{(\sum_{k=1}^{N} \omega_k)^2} \sum_{n=1}^{N} \omega_n \sum_{m=1}^{N} \omega_m \mathbf{f}_{n,m} \\ &- \frac{2^{1-s}}{(\sum_{k=1}^{N} \omega_k)} \sum_{n=1}^{N} \omega_n \mathbf{g}_n + 12^{-s} \Big|, \end{aligned} \tag{7}$$

where the notations are, $x_n = (x_{n,1}, x_{n,2} \ldots, x_{n,s})$,

$$\mathbf{g}_n = \prod_{i=1}^{s}\Big(1 - x_{n,i}\Big)x_{n,i},$$

$$\mathbf{f}_{n,m} = \prod_{i=1}^{s}\Big(1 - \max(x_{n,i}, x_{m,i})\Big)\min(x_{n,i}, x_{m,i}),$$

and the weights $\omega = (\omega_k)_{1\leq k\leq N}$ are nonnegative reals.

Note that there is a neighborhood of $\omega_u = (\frac{1}{N})_{1\leq n\leq N}$, $\mathbf{O}_{\omega^u}$, such that the functions $\mathbf{T}_N(X,\omega)$ is differentiable at any point ω in $\mathbf{O}_{\omega_u}$. Thus, the partial derivatives of $\mathbf{T}_N(X,\omega)$:

$$S_n = \frac{\partial \mathbf{T}_N}{\partial \omega_n}(\omega_u), \quad 1 \leq n \leq N,$$

exist and are called the L_2 discrepancy sensitivities.

PROPOSITION **1** *The L_2 discrepancy sensitivity, S_n, is given by the following analytical formulas:*

$$S_n = \frac{\Big(\sigma_n - \sigma\Big) + 2^{-s}\Big(\gamma - \mathbf{g}_n\Big)}{\mathbf{T}_N(X)}, \tag{8}$$

where $\sigma = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbf{f}_{n,m}$, $\sigma_n = \frac{1}{N}\sum_{m=1}^{N}\mathbf{f}_{n,m}$ *and* $\gamma = \frac{1}{N}\sum_{m=1}^{N}\mathbf{g}_m$.

Proof

By observing that

$$\begin{aligned} A(\omega) &= \frac{1}{(\sum_{k=1}^{N}\omega_k)^2}\sum_{k=1}^{N}\omega_k\sum_{m=1}^{N}\omega_m\mathbf{f}_{n,m} \\ &= \frac{\omega_n^2\mathbf{f}_{n,n}}{(\sum_{k=1}^{N}\omega_k)^2} \\ &+ \frac{2\omega_n}{(\sum_{k=1}^{N}\omega_k)^2}\sum_{m\neq n}^{N}\omega_m\mathbf{f}_{n,m} \\ &+ \frac{1}{(\sum_{k=1}^{N}\omega_k)^2}\sum_{r\neq n}^{N}\omega_r\sum_{m\neq n}^{N}\omega_m\mathbf{f}_{r,m}, \end{aligned}$$

and

$$\begin{aligned} B(\omega) &= \frac{1}{(\sum_{k=1}^{N}\omega_k)}\sum_{m=1}^{N}\omega_m\mathbf{g}_m \\ &= \frac{\omega_n\mathbf{g}_n}{\Big(\sum_{k=1}^{N}\omega_k\Big)} \\ &+ \frac{1}{\Big(\sum_{k=1}^{N}\omega_k\Big)}\sum_{m\neq n}^{N}\omega_m\mathbf{g}_m. \end{aligned}$$

The partial derivatives of $A(\omega)$ and $B(\omega)$ with respect to ω_n are

$$\begin{aligned} \frac{\partial A}{\partial \omega_n}(\omega) &= \frac{2\omega_n\mathbf{f}_{n,n}}{(\sum_{k=1}^{N}\omega_k)^3}\Big[\sum_{m\neq n}^{N}\omega_m\Big] \\ &+ \frac{2\Big(\sum_{m\neq n}^{N}\omega_m\mathbf{f}_{n,m}\Big)}{(\sum_{k=1}^{N}\omega_k)^3}\Big[\sum_{r\neq n}^{N}\omega_r\Big] \\ &- \frac{2\sum_{r\neq n}^{N}\omega_r\sum_{m\neq n}^{N}\omega_m\mathbf{f}_{r,m}}{(\sum_{k=1}^{N}\omega_k)^3}, \end{aligned}$$

and

$$\frac{\partial B}{\partial \omega_n}(\omega) = \frac{\mathbf{g}_n}{(\sum_{k=1}^{N}\omega_k)^2}\Big[\sum_{m\neq n}^{N}\omega_m\Big]$$

$$- \quad \frac{\sum_{m\neq n}^{N} \omega_m \mathbf{g}_m}{(\sum_{k=1}^{N} \omega_k)^2}.$$

After evaluating $\frac{\partial A}{\partial \omega_n}$ and $\frac{\partial B}{\partial \omega_n}$ with all ω_k set to $\frac{1}{N}$, we obtain

$$\frac{\partial A}{\partial \omega_n}(\omega^u) \quad = \quad 2\Big[\frac{1}{N}\sum_{m=1}^{N} \mathbf{f}_{n,m} - \frac{1}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N} \mathbf{f}_{k,m}\Big],$$

and

$$\frac{\partial B}{\partial \omega_n}(\omega^u) \quad = \quad \mathbf{g}_n - \frac{1}{N}\sum_{m=1}^{N} \mathbf{g}_m.$$

Thus, the partial derivative S_n is given by the analytical formulas:

$$S_n = \frac{\Big(\sigma_n - \sigma\Big) + 2^{-s}\Big(\gamma - \mathbf{g}_n\Big)}{\mathbf{T}_N(X)}, \tag{9}$$

where $\sigma = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} \mathbf{f}_{n,m}$, $\sigma_n = \frac{1}{N}\sum_{m=1}^{N} \mathbf{f}_{n,m}$ and $\gamma = \frac{1}{N}\sum_{m=1}^{N} \mathbf{g}_m$.

Those are the first partial derivatives of $\mathbf{T}_N(X, \omega)$, with respect to the weight, ω_n, and evaluated at $\omega_u = (\frac{1}{N})_{1\leq n\leq N}$, i.e. weights of all profiles are identically initialized.

Higher order sensitivities may be obtained in a similar fashion, however, for our definition of outlier and insider, the first sensitivities are the most important ones. Let h a very small real, we note by $\mathbf{w}_h^n = (\omega_i)_{1\leq i\leq N}$ when $\omega_n = \frac{1}{N} + h$ and $\omega_j = \frac{1}{N}$ if $j \neq n$. By using the law of the mean, we have

$$\mathbf{T}_N(X, \mathbf{w}_h^n) - \mathbf{T}_N(X) \approx hS_n. \tag{10}$$

We assume that N is very large, and we choose $h = -\frac{1}{N}$. Therefore, these first sensitivities can be used as a basis for monitoring the variation of L_2 discrepancy value:

PROPOSITION **2** *When N is very large, we have*

- *If $S_n > 0$, the profile, x_n, is an insider as the L_2 discrepancy value is minimized by removing, x_n, from the sequence of profiles.*
- *If $S_n < 0$, the profile, x_n,is an outlier as the L_2 discrepancy value is maximized by removing, x_n, from the sequence of profiles.*

The sign of the L_2 discrepancy sensitivity can be regarded as the oracle on spatial visibility for each profile.

- If $S_n < 0$, the point x_n is in a region quasi-empty of sequence points, i.e. add more points in its immediate neighborhood will minimize L_2 discrepancy.
- If $S_n > 0$, the point x_n is in a region where the sequence is clumped, i.e. L_2 discrepancy will decrease if the point x_n is removed.
- If $S_n = 0$, the point x_n is in a region where the sequence is uniformly distributed.

Nevertheless, the L_2 discrepancy sensitivity magnitudes can be used to define:

- Greatest rate of *disimilarity* (GRDS),

$$GRD = (\frac{1}{N_{dis}} \sum_{n,S(x_n)<0} S(x_n)^2)^{1/2}, \tag{11}$$

- Greatest rate of *similarity* (GRS),

$$GRS = (\frac{1}{N_{si}} \sum_{n,S(x_n)>0} S(x_n)^2)^{1/2}. \tag{12}$$

- Ratio between *disimilarity* and *similarity* is defined by

$$R = \frac{GRDS}{GRS}.$$

The parameters GRS and GRDS with the L_2 discrepancy value can be considered as the fingerprint of the profiles distribution. The ratio R will be used as measure of autocorrelation between the profiles. For instance, when some profiles are considerably dissimilar or inconsistent with respect to the remaining data, the value of R will be high.

3 Effectiveness analysis

In the previous section, two explicit formulas (6) and (8) available for straightforward computation of the L_2 discrepancy and the discrepancy learning process, the cost will require $O(N^2 \times s)$ operations. In this section, we conduct a effectiveness analysis, meaning the effectiveness of the discrepancy learning process for high-dimensional data, as is always the case in data mining.

The L_2 discrepancy remains valid as a measure of the uniformity only if the number of points, N, has to grow exponentially with the dimension, s. Mathematically speaking,

PROPOSITION **3** *For any sequence of profiles* $X = \{x_i\}_{1 \leq i \leq N}$ *in* I^s *with* $N \leq 2^s$*, we have*

$$12^{-s} \leq \mathbf{T}_N^2(X). \tag{13}$$

Proof

By using the definition of the *local discrepancy* as formulated in equation (4), we have

$$\frac{1}{N}\mathbf{E}(N\lambda_s(J)) \leq \Big|D_N(J,X)\Big|, \tag{14}$$

where $\mathbf{E}(r)$ is the distance between the real number r to the nearest integer. Thus

$$\frac{1}{N^2}\left[\int_{(\mathbf{a},\mathbf{b})\in I^{2s}, a_i<b_i} \Big(\mathbf{E}(N\lambda_s(J))\Big)^2 d\mathbf{a}d\mathbf{b}\right] \leq \mathbf{T}_N^2(X), \tag{15}$$

where $\lambda_s(J) = \prod_{i=1}^{s}\Big(b_i - a_i\Big)$.

Let $0 < v = \max\limits_{1\leq i\leq s}\Big(b_i - a_i\Big) < 1$, then $N\lambda_s(J) \leq Nv^s$. Thus if the dimension s is high and the number of profiles N is relatively small such that $Nv^s < 1$, then $\mathbf{E}(N\lambda_s(J)) = N\lambda_s(J)$ accordingly to the definition of $\mathbf{E}(.)$, we have

$$12^{-s} = \prod_{i=1}^{s}\left[\int_{(a_i,b_i)\in I^2, a_i<b_i} \Big(b_i - a_i\Big)^2 da_i db_i\right] \leq \mathbf{T}_N^2(X). \tag{16}$$

Table 1: Smallest value of N for which the bound (16) is not valid. The sequence X was generated as a random sequence

s	6	8	10	12	14
N	70	310	1,100	4,500	17,000
$\frac{12^{-s}}{(\mathbf{T}_N^2(X))}$	1.0700	1.0500	1.1600	1.09860	1.0355
$\frac{2^s-1}{N}$	1.1100	1.2100	1.0700	1.0989	1.0376

Knowing that the root mean square expectation of $\mathbf{T}_N$ for a random sequence X in I^s is given by, see [4],

$$\mu(\mathbf{T}_N^2(X)) = \frac{12^{-s}}{N}(2^s - 1).$$

Using the inequality 16, thus,

$$N < 2^s.$$

The condition "if the dimension s is high and the number of profiles N is relatively small" is mathematically equivalent to $N < 2^s$ as the average value of v is $\frac{1}{2}$.

The table 1 shows the smallest values N for which the bound (16) is not valid when the sequence X is generated in a random fashion using different dimensions, s. When the dimension, s, is high, we conclude that the bound (16) is not valid when the number of points N is greater than $2^s - 1$ and only beyond this number, the L_2 discrepancy remains valid as a measure of the uniformity.

4 Numerical illustrations

We implemented the algorithm using the Fortran 90 programming language for computation and C programming language for graphics using OpenGl on IBM RS/6000 cluster. We used a 32 bit floating-point type to represent the attributes of the points and the computed values are rounded to four significant digits. In all the coming experiments, we compute first σ and γ and they are used to compute the L_2 discrepancy and L_2 discrepancy learning process(DLP). This appears to work well enough to minimize the execution time.

Table 2: Computed values of DLP parameters for random data and Gaussian data.

Data	Random	Gaussian
$T_N^2(X)$	2.0107E-6	0.0104
$GRDS$	0.0242	0.0553
GRS	0.0489	0.0363
R	0.4948	1.5234

Table 3: Computed values of DLP parameters for 3D Sobol sequence and 2D projections.

Data	$(28, 29)$	$(28, 30)$	$(29, 30)$	$(28, 29, 30)$
$T_N^2(X)$	1.6470E-7	4.5357E-7	2.7389E-5	2.3470E-6
$GRDS$	2.0984E-3	9.0162E-3	3.3140E-3	1.233E-3
GRS	2.8116E-3	0.0160	3.6074E-3	1.597E-3
R	0.7400	0.900	0.914	0.770

4.1 Outlier and insider detections in random and Gaussian data sets

For testing, we used two families of synthetic data sets with $s = 2$. The first data set have 10^4 profiles, randomly generated, $(r_1, r_2) \in (0, 1)^2$. The second data set is a transformation of the first data set into a Gaussian distribution $x = \sqrt{-2\log(r_1)}\cos(2\pi r_2)$ and $y = \sqrt{-2\log(r_1)}\sin(2\pi r_2)$ and scaled to $[0, 1)^2$. Figures 1 (a) and (b) show the two dimensional random and Gaussian data sets where color is encoded from blue (low values) to red (high values), the black color is the background of the graphics and is not related to the values generated by the DLP. The values of DLP parameters are listed in table 2. Gaussian data set has the higher discrepancy and GRS values, this is due to the fact that the Gaussian data set becomes more sparse and clumped than the random data set. The higher GRDS value is due to the fact that the Gaussian data set has more isolated points than the random data set. The higher value of R is reflecting the fact that the Gaussian data set has profiles with exceptional spatial distribution with respect to the remaining data.

4.2 Insider and outlier detections in lower dimensional projections

The essential idea behind this experiments is to show that the value L_2 discrepancy informs about low density or sparsity of profiles. The sparsity of data is indicated by a high discrepancy value. This test uses 3D orthogonal projection $(28, 29, 30)$ of Sobol sequence in $[0, 1)^{30}$ given in [6].

Figure 2 shows a 2D projections visualization. The top, the colormap is coded using values of the discrepancy learning process computed over 3D sequence,$(28, 29, 30)$. But below, the colormap is coded using values of the DLP computed for each 2D projection.

The values of the L_2 discrepancy are listed in table 3 with N = 4096 points. An interesting observation is that the 2D projection $(29, 30)$ has the higher discrepancy and that indicates the sparsity of profiles and its visualization shows outliers(points with blue color) as points in a region of low density. However, the visualization of $(28, 29)$ and $(28, 30)$ projections show that outlier profiles in 3D are not directly comparable to one in 2D.

In general, the lower dimensional projections with low discrepancy are not suitable to detect sparsity as the full feature descriptions of the sequence in high dimensional often do not exist in the lower dimensional projections with low discrepancy.

4.3 Outlier detection in falsified data

We tested our outlier detection technique on the following real data set: ColorMoments ($s = 9, N = 68,040$). This data set represents a collections of real images. ColorMoments are image features extracted from a Corel image collection[2]. We have deliberately falsified the first 1000 points of the data set by introducing 17 points among them with considerably dissimilar attributes. We plot the data as (n, S_n) where n is the indice of the profile x_n and S_n is the value of of the DLP for x_n. Figure 3 shows the result-

[2]See http://kdd.ics.uci.edu/databases/CorelFeatures/CorelFeature.html for more information

Table 4: Computed values of DLP parameters for actual and falsified data sets.

Data	Actual	Falsified
$T_N^2(X)$	0.2175E-06	0.9684E-09
$GRDS$	0.2582E-03	0.3047E-05
GRS	0.2658E-03	0.1592E-05
R	0.9714	1.9139

ing visualization for the falsified data. Table 4 lists all L_2 discrepancy parameters. The DLP, $S(x_n)$, reports information accordingly to the change in the real distribution, i.e. maintain an excellent detection rate, and the ratio between GRDS and GRDS, $R = \frac{GRDS}{GRS}$, confirms the changes in the actual data.

5 Conclusion

Outlier mining is a new area of research, especially for computational mathematics. It is becoming an important activity for many companies, especially in scientific measurement. We presented new definition for outlier and insider mining using a deterministic and a holistic framework based on the well established theory of discrepancy. Beginning with a mathematical formulation of the *local discrepancy* that handles the similarity and the sparsity. and using a mathematical definition of L_2 discrepancy as a global measure of the discrepancy of the sequence of profiles.

A L_2 discrepancy learning process ranks each point on the basis of optimizing the L_2 discrepancy value. This L_2 discrepancy learning process allows an access to many levels of information about outliers and insiders in the data.

Experimental results showed that the L_2 discrepancy learning process captured the actual features of the data via the parameters L_2 discrepancy value, the greatest rate of *disimilarity*, GRDS, and the greatest rate of *similarity*, GRS. The ratio between GRDS and GRDS, $R = \frac{GRDS}{GRS}$, can be used as an indicator of the changes in the data set.

In future work we will further investigate the usage of this L_2 discrepancy learning process in real cases. Also we look to study a parallel implementation as the algorithm lends itself to parallelism and can map efficiently onto parallel computers. Thus it fit the requirement of high-performance data mining that, refers to developing efficient parallel algorithms for data-mining techniques.

Acknowledgments

The bulk of the reported calculations were performed on the IBM RS/6000 cluster of the Advanced Computational Research Laboratory of the University of New Brunswick.

References

[1] U. Fayyad, *Mining Databases: Towards Algorithms for Knowledge Discovery*, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 22 (1998), p. 39–48.

[2] J. Han and M. Kamber, *Data Mining, Concepts and Technique*, Morgan Kaufmann, San Francisco, 2001.

[3] F. E. Khettabi,*Numerical Methods For Boltzmann Equation*, Ph.D. thesis, Department of Mathematics,University of Savoie, Chambéry, France, 1998.

[4] W. J. Morokoff and R. E. Caflisch, *Quasi-random sequences and their discrepancies*, SIAM J. Sci. Comput.,6 (1994), p. 1251–1279.

[5] H. Niederreiter,*Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia, 1992.

[6] W. H. Press and S. A. Teukolsky, *Quasi-(that is,Sub-) Random Numbers* Computers in Physics, 6 (1989), p. 76–79.

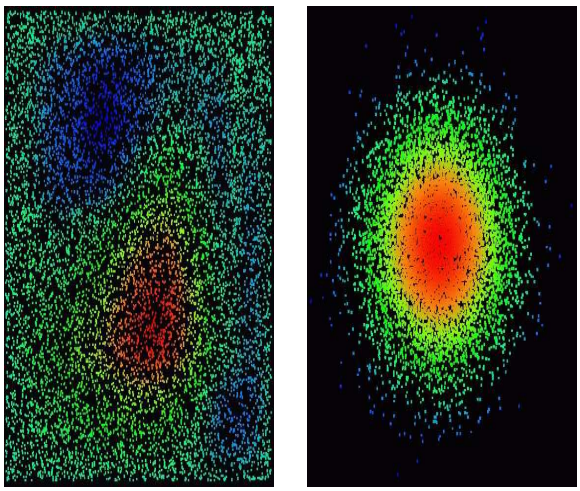

Figure 1: Outliers in a random distribution data and synthetic Gaussian data

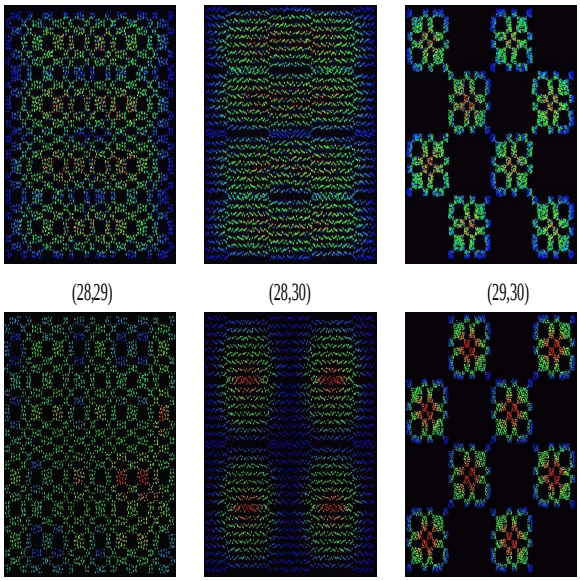

Figure 2: Comparaison of outliers in 3D Sobol sequence and in lower dimensional projections.

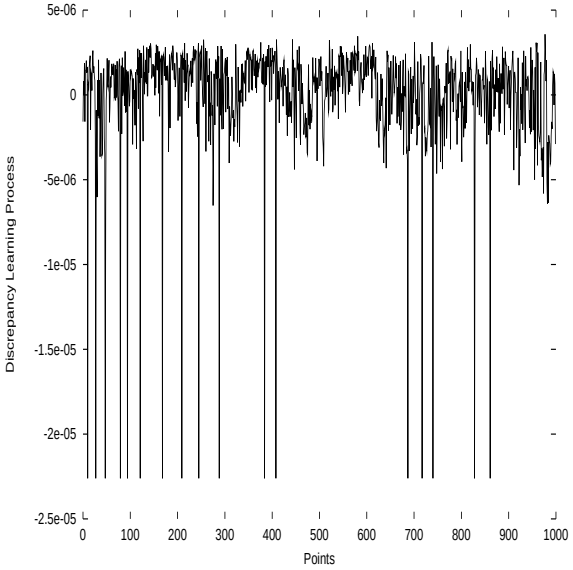

Figure 3: Detection of deliberately falsified data