Local Peculiarity Factor and Its Application in Outlier Detection

Jian Yang[1], Ning Zhong[1,2], Yiyu Yao[1,3], Jue Wang[4]

[1]International WIC Institute
Key Laboratory of Multimedia and Intelligent Software
Beijing University of Technology, Beijing 100124, China

[2]Department of Life Science and Informatics
Maebashi Institute of Technology, Maebashi-City 371-0816, Japan

[3]Department of Computer Science, University of Regina, Regina, Canada S4S 0A2

[4]Institute of Automation, Chinese Academy of Sciences, Beijing 100196, China

jianyang@bjut.edu.cn, zhong@maebashi-it.ac.jp
yyao@cs.uregina.ca, jue.wang@mail.ia.ac.cn

ABSTRACT

Peculiarity oriented mining (POM), aiming to discover peculiarity rules hidden in a dataset, is a new data mining method. In the past few years, many results and applications on POM have been reported. However, there is still a lack of theoretical analysis. In this paper, we prove that the peculiarity factor (PF), one of the most important concepts in POM, can accurately characterize the peculiarity of data with respect to the probability density function of a normal distribution, but is unsuitable for more general distributions. Thus, we propose the concept of local peculiarity factor (LPF). It is proved that the LPF has the same ability as the PF for a normal distribution and is the so-called ϵ-sensitive peculiarity description for general distributions. To demonstrate the effectiveness of the LPF, we apply it to outlier detection problems and give a new outlier detection algorithm called LPF-Outlier. Experimental results show that LPF-Outlier is an effective outlier detection algorithm.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; I.2.6 [**Learning**]: Knowledge acquisition

General Terms

Algorithms, Theory

Keywords

data mining, peculiarity factor, local peculiarity factor, ϵ-sensitive peculiarity description, outlier detection

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.
Copyright 2008 ACM 978-1-60558-193-4/08/08 ...$5.00.

1. INTRODUCTION

A fundamental task of data mining is to discover useful knowledge hidden in a dataset. The knowledge may be expressed in many forms, such as association rules, classification rules, exception rules, clusters, frequent patterns, trends and so on [13, 15]. Zhong *et al.* considered a new type of knowledge, called peculiarity rules, which may be hidden in a small subset of records in a large dataset [23]. The associated peculiarity oriented mining (POM) aims to find peculiarity rules by focusing on a small subset of interesting data called peculiar data. Peculiar data have two essential properties, one is that they represent cases described by a relatively small number of objects, and the other is that they are very different from other objects in the dataset [15].

The necessity and importance of studying peculiar data and peculiarity rules can be seen from many real world problems. In many situations, datasets are collected with data polluted by noises from hardware systems, human factors or other malfunctions. Detecting the noises from polluted data is an important preprocessing step. This may be done by considering noises as peculiar data. It may also happen that only a small sub-dataset consisting of data very different from others is interesting and important to a particular application. For example, in the stock market supervising only a small subset of abnormal data is interesting and important to the organization of securities regulatory. These abnormal data are peculiar data in the dataset. In fact, peculiar data have been interpreted in many terms, such as outliers, data points with very small probability density values and so on. POM is therefore closely related to other data mining or machine learning approaches, including the outlier detection and the probability density estimation.

Studies of POM have produced useful theoretical results and applications. Zhong *et al.* used the RVER (reverse variant entity-relationship) model to represent peculiar data and conceptual relationships among peculiar data discovered from multi-database [23]. The model has been applied to mine peculiarity rules in many databases, such as Japan-survey, amino-acid, weather, super-market, hepatitis, fMRI brain images and EEG brain waves [16, 21, 22]. The results show that POM is an effective data mining method. A preliminary analysis of the peculiarity factor (PF), one of the most important concepts in POM, was given in [20], which demonstrates that the PF indeed reflects our intuitive understanding of the two properties of peculiar data. The concept and a framework of relational POM were proposed in [21], which combines results from relational mining and POM at the record level. Techniques from inductive logic programming for relational inductive learning [18] were used in this framework. A more systematic and detailed study of relational POM based on the record PF was given in [15]. Particularly, two basic tasks of relational POM, namely, description and explanation, were discussed. A prototype system was implemented to find peculiar records and first-order peculiarity rules from multiple relations. The analysis of a database from China Statistics Yearbook by using the system has produced interesting peculiarity rules [15].

When compared with applications of POM, there is still a lack of theoretical and quantitative analysis of the sensitivity, reasonability, and applicability of the PF. In this paper, we present a theoretical analysis of the PF. We prove that, from the viewpoint of low frequency, the PF can characterize accurately the peculiar data of a normal distribution, but is incapable of describing the peculiar data of general distributions. To resolve this difficulty, we introduce the concept of local peculiarity factor (LPF). It is shown that the LPF has the same ability as the PF in describing the peculiar data of a normal distribution. In addition, it is the so-called ϵ-sensitive peculiarity description with respect to a more general distribution. We apply LPF to outlier detection problems and give an outlier detection algorithm called LPF-Outlier. Experimental results on a synthetic dataset show that the LPF can characterize the peculiar data of a general distribution more accurately than the PF. Experimental results on several real-world datasets demonstrate that the proposed LPF-Outlier is an effective outlier detection method.

The rest of the paper is organized as follows. In Section 2, we first analyze the original peculiarity factor and then propose the concept of local peculiarity factor. Then we prove that the local peculiarity factor has two desirable properties. Finally we propose the LPF-Outlier algorithm. In Section 3, we report the results from a set of extensive experiments by using both synthetic and real-world datasets. Conclusions are given in Section 4. Due to the page limit, we only provide the proof of one theorem in the Appendix.

2. LOCAL PECULIARITY FACTOR

By extending the concept of peculiarity factor (PF), we introduce the notion of local peculiarity factor (LPF). For convenience, we first give some explanations and notations. In the following discussions, unless noted, we only discuss the PF and the LPF for a one-dimensional distribution. We mainly consider the attribute PF and the attribute LPF, since the record PF and the record LPF can be easily obtained from attribute ones. We suppose that $p(x)$ is the probability density function (PDF) of a one-dimensional continuous random variable, and $D : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a continuous and shift-invariant conceptual distance on $\mathbb{R}$, that is, $D(x_1, x_2)$ can be represented by a function of $|x_1 - x_2|$, $x_1, x_2 \in \mathbb{R}$. We denote $\phi(|x_1 - x_2|) = D(x_1, x_2)$ and further suppose that ϕ is strictly increasing. Obviously, $D(x_1, x_2) = \phi(|x_1 - x_2|) = |x_1 - x_2|^\alpha$ with $\alpha > 0$ are such conceptual distances.

2.1 Peculiarity Factor

A central notion of peculiarity oriented mining (POM) is the peculiarity factor (PF) introduced by Zhong et al. [15, 24]. In particular, two levels of peculiarity can be identified, representing attribute peculiarity and record peculiarity.

Definition 1. ([15, 24]) *Suppose that $\{Z_1, Z_2, \cdots, Z_n\}$ is a sample set with n points and each point $Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{im})$ is described by attributes $A_1, A_2, \cdots, A_m$. Then for the dataset, the attribute PF of the attribute value Z_{ij} is defined by:*

$$PF(Z_{ij}) = \sum_{l=1}^{n} D(Z_{ij}, Z_{lj}) = \sum_{l=1}^{n} |Z_{ij} - Z_{lj}|^\alpha, \tag{1}$$

where α is a parameter. And for the dataset, the record PF of the point Z_i is defined by:

$$PF(Z_i) = \sum_{l=1}^{n} \sqrt{\sum_{j=1}^{m} \beta_j \times (PF(Z_{ij}) - PF(Z_{lj}))^2}, \tag{2}$$

where β_j is the weight of attribute A_j, and $PF(Z_{ij})$ is given by Eq. (1).

The peculiarity factor is determined by the parameter α, which may be adjusted by users and $\alpha = 0.5$ is used as default. For the record PF, we also need the weights β_j's for attributes, which are given by users and $\beta_j = 1$ is used as default.

In general, one can define the record PF as follows:

$$PF(Z_i) = \sum_{l=1}^{n} \left(\sum_{j=1}^{m} \beta_j \times (PF(Z_{ij}) - PF(Z_{lj}))^p \right)^{\frac{1}{p}}.$$

That is, we use general p-norm instead of Euclidean distance in Eq. (2). This is more consistent with the attribute PF. But if the attribute PF can accurately describe the peculiarity of the data on each attribute, we can simply define the record PF of a point by the weighted sum of the attribute PF values on all attributes, that is,

$$PF(Z_i) = \sum_{j=1}^{m} \beta_j \times PF(Z_{ij}). \tag{3}$$

Meanwhile the record PF given by Eq. (3) has a good property. That is, for a one-dimensional data point, its attribute PF is same to its record PF.

In the following discussions of this paper, we use Eq. (3) to get the record level PF from the attribute level PF and briefly refer to the record PF as PF.

In order to gain more insights about the attribute PF, we can analyze it from the viewpoint of statistics. Thus we first generalize the definition of the PF to the case of a one-dimensional continuous distribution.

Definition 2. *Suppose that P is a one-dimensional distribution with continuous PDF $p(x)$, then for the distribution, the PF of any point x_0 sampled from P is defined by:*

$$LPF(x_0) = \int_{-\infty}^{+\infty} D(x_0, x) p(x) dx.$$

The idea hidden in Definition 2 is same to that in Definition 1, and we only use the one-dimensional continuous distribution and the integral to replace the dataset and the sum in Definition 1, respectively. Thus we can use the PF given by Definition 2 to analyze the property of the attribute PF from the viewpoint of statistics. For a normal distribution, we have the following important theorem.

Theorem 1. *For a data point sampled from a one-dimensional normal distribution, its PF value strictly increases with its conceptual distance to the mean of the distribution.*

From Theorem 1, we know that for a normal distribution and a given proper conceptual distance, data points further away from the mean have larger PF values. Such points have lower probabilities and should be considered to be more peculiar than other data points

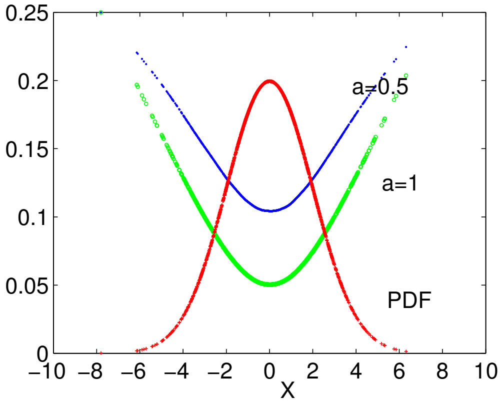

Figure 1: The PDF values and PF values ($\alpha = 1$ or $\alpha = 0.5$) of data sampled from the normal distribution $\mathbf{N}(0, 4)$.

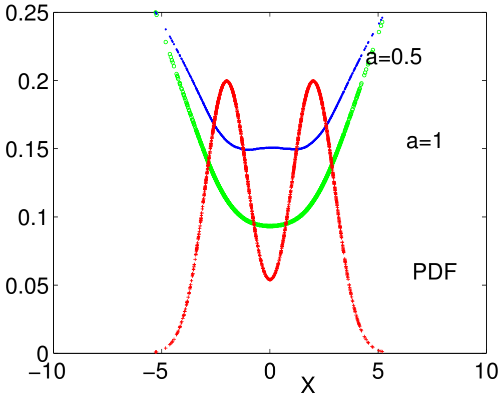

Figure 2: The PDF values and PF values ($\alpha = 1$ or $\alpha = 0.5$) of data sampled from the distribution $\frac{1}{2}(\mathbf{N}(2, 1) + \mathbf{N}(-2, 1))$.

closer to the mean. It follows that the PF correctly reflects our intuitive interpretation of peculiarity for a normal distribution. That is, for a normal distribution, the PF can indicate the location of a point in the distribution, which is closely related to the frequency-based interpretation of peculiarity. Furthermore, from Theorem 1 we know that the PF can indicate the location of any point in a normal distribution. Thus the PF can give a description of the PDF from the viewpoint of data analysis. The description is important and valuable, since in most practical problems there are only some data points sampled from an unknown distribution and the estimation of the PDF is very important but quite difficult.

We demonstrate the ideas of Theorem 1 by a concrete example. We sampled 2000 data points from the normal distribution $\mathrm{N}(0, 4)$ and calculated their PDF values and PF values. Since we only want to illustrate the relation between PDF and PF values, which is determined only by the relative magnitude of these values, we compressed these PF values into (0,0.25] and plotted PDF values and PF values (with $\alpha = 1$ or $\alpha = 0.5$ in Eq. (1)) in one figure, as shown in Fig. 1. We can see that points with smaller PDF values indeed have larger PF values.

In general, points sampled from a non-normal distribution may not have the property given in Theorem 1. This can be illustrated by another example.

EXAMPLE 1. *Let's consider a mixed normal distribution with PDF $p(x)$, $p(x) = \frac{1}{2}(N(2,1) + N(-2,1)) = \frac{1}{2\sqrt{2\pi}}(e^{-\frac{(x-2)^2}{2}} + e^{-\frac{(x+2)^2}{2}})$. For the distribution, there exist data points which have both smaller PDF values and smaller PF values. That is, the PF does not necessarily have higher values for lower frequent data points, and hence does not reflect our intuitive interpretation of peculiarity.*

To illustrate the inability of the PF in characterizing peculiar data of the distribution $\frac{1}{2}(\mathrm{N}(2, 1) + \mathrm{N}(-2, 1))$, we sampled 2000 data points and plotted their PDF values and PF values in Fig. 2, where PF values were calculated with $\alpha = 1$ or $\alpha = 0.5$ in Eq. (1), and compressed into (0,0.25]. We can see that points close to the origin have smaller PDF values and smaller PF values, and points with their absolute values larger than 4 have smaller PDF values and larger PF values. Hence the PF cannot describe the frequency based interpretation of peculiarity.

From Fig. 2, we observe that for points near the origin most sample data are near to them, but for points with absolute values larger than 4 almost half sample data are far from them. Hence the PF value of a point with conceptual distances to all the other sample data cannot describe its location in the distribution accurately. This is the reason for the inability of the PF in characterizing the PDF of general distributions. Intuitively the PF value with conceptual distances between the point and its near neighbors can indicate its location in the distribution. Justly we review the result of Theorem 1 for a unimodal non-normal distribution. We know that a point sampled from the distribution does not necessarily have the property of Theorem 1. But a bit modification on the PF can make it have the property. We divide the distribution into two regions, the left region and the right region of the peak value. If we consider the PF only in one of the regions, that is, the PF value of a point sample from the left region is defined only in the left region and the PF of a point sample from the right region is defined only in the right region, one can prove that Theorem 1 is still true for each region. This shows that by defining the PF locally it can have more good properties. These motivated us to improve the definition of the PF in a local manner.

2.2 Local Peculiarity Factor

Consider first the following theorem.

THEOREM 2. *Suppose that $c_0 \in (0, 1]$ is a constant and $p(x)$ is a continuous PDF of a one-dimensional distribution. For any x_0 with $p(x_0) > 0$, the sum of its conceptual distances to all points belonging to an interval with its probability equal to c_0 is given by:*

$$\begin{aligned} F_{(x_0,c_0)}(a, b) &= \int_{x_0-a}^{x_0+b} D(x_0, x)p(x)dx \\ s.t. &\qquad \int_{x_0-a}^{x_0+b} p(x)dx = c_0. \end{aligned}$$

Then the function $F_{(x_0,c_0)}$ reaches its minimum value at $a = b$, that is,

$$\min_{a,b} F_{(x_0,c_0)}(a, b) = \int_{x_0-a}^{x_0+a} D(x_0, x)p(x)dx.$$

From Theorem 2 it can be seen that for a point of a distribution and a given parameter c_0, the interval minimizing the sum of the weighted conceptual distances from the point to all its near neighbor points belonging to an interval with probability c_0 is centered at the point. The point x_0 and the constant c_0 can determine uniquely the interval. This result enables us to introduce the concept of local peculiarity factor (LPF) for a one-dimensional continuous distribution.

DEFINITION 3. *Suppose that P is a one-dimensional distribution with continuous PDF $p(x)$ and $c_0 \in (0,1]$ is a constant, then the attribute LPF of a point x_0 sampled from P is defined by:*

$$LPF(x_0) = \int_{x_0-a}^{x_0+a} D(x_0,x)p(x)dx,$$

where a satisfies

$$\int_{x_0-a}^{x_0+a} p(x)dx = c_0. \tag{4}$$

The constant c_0 is the probability of all the data points selected to describe the peculiarity of a point, which introduces a local manner to improve the definition of the PF. Definition 3 is a local version of Definition 2, and if $c_0 = 1$, we have $a = +\infty$. In the case, for a one-dimensional continuous distribution the attribute LPF is equivalent to the attribute PF.

For a one-dimensional normal distribution, the attribute LPF has the same property as the PF.

THEOREM 3. *For any $c_0 \in (0,1]$, the attribute LPF value of a point sampled from a one-dimensional normal distribution strictly increases with its conceptual distance to the mean of the distribution.*

From Theorem 3, we know that for any $c_0 \in (0,1]$, the attribute LPF value of a point sampled from a normal distribution can indicates its position in the distribution, namely, a point with a smaller PDF value has a larger attribute LPF value and is farther to the mean point. For the distribution, the attribute LPF with any $c_0 \in (0,1]$ has the same ability in describing the peculiar data as the PF. For more general distributions, by selecting a proper c_0, the attribute LPF can describe their PDF at any accuracy. This requires us to introduce first the notion of "description at any accuracy".

DEFINITION 4. *Suppose that P is a one-dimensional distribution with continuous PDF $p(x)$, f is a description of P and $\epsilon > 0$ is a given constant. If for all x_1, x_2 with $p(x_1) - p(x_2) > \epsilon$, the inequality $f(x_1) < f(x_2)$ holds, f is called an ϵ-sensitive peculiarity description of P.*

The parameter ϵ can be adjusted by users according to the requirement of the problem or their own prior knowledge on the problem. For different problems or different aspects of applications, we are always required to describe one distribution with different precisions, which corresponds to different requirements on sampling from the distribution. This idea is captured by introducing a parameter ϵ into the definition. The smaller the ϵ value, the more accurate the description, which implies more sample data needed. The effect of all data points with the PDF value smaller than ϵ is ignored in an ϵ-sensitive peculiarity description.

With the definition of ϵ-sensitive peculiarity description, the attribute LPF has the following property.

THEOREM 4. *For a one-dimensional distribution P with continuous PDF $p(x)$ and any given $\epsilon > 0$, there exists a constant $c_0 \in (0,1]$ such that for any $c \in (0, c_0)$ the attribute LPF is an ϵ-sensitive peculiarity description of P.*

From Theorem 4 we know that the attribute LPF can describe the PDF of a distribution at any accuracy, which is determined by the parameter ϵ. Generally speaking, for a given distribution the smaller the parameter ϵ, the smaller the constant c_0, and hence the more accurate the description.

As given in Theorem 3 and Theorem 4, the attribute LPF for a one-dimensional distribution has better statistical properties than the PF. Thus we can improve the definitions of the attribute PF and the record PF for a dataset by introducing the idea of locality. For the case of a dataset, the constant c_0 in Eq. (4) is replaced by a fixed number of neighbor points and the integral is replaced by a sum.

DEFINITION 5. *Suppose that $\{Z_1, Z_2, \cdots, Z_n\}$ is a sample set with n points and each point $Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{im})$ is described by attributes $A_1, A_2, \cdots, A_m$. Then for the dataset, the attribute LPF of the attribute value Z_{ij} is defined by:*

$$LPF(Z_{ij}) = \sum_{Z_{lj} \in N_k(Z_{ij})} D(Z_{ij}, Z_{lj}), \tag{5}$$

where α is a parameter and $N_k(Z_{ij})$ is the set of k near neighbors of Z_{ij} in the set $\{Z_{1j}, Z_{2j}, \cdots, Z_{nj}\}$, that is, $N_k(Z_{ij})$ consists of k near neighbors of Z_i on attribute A_j. And the record LPF of the point Z_i is defined by:

$$LPF(Z_i) = \sum_{j=1}^{m} \beta_j LPF(Z_{ij}), \tag{6}$$

where β_j is the weight of attribute A_j, and $LPF(Z_{ij})$ is given by Eq. (5).

The above definition gives a large class of LPF measures determined by the weights of attributes. Those weights value can be adjusted by users to derive a particular LPF measure, and $\beta_j = 1$ is used as default.

The introduction of the LPF enables us to identify peculiar data at both the attribute and the record level. As shown in the next section, they can easily be incorporated into a peculiar data identification algorithm. Similar to the PF, we briefly refer to the record LPF as LPF. For a one-dimensional data point its attribute LPF is also same to its record LPF.

2.3 A LPF based Outlier Detection Algorithm

We can design many LPF based algorithms for many data mining problems. As an application of peculiar data identification, we use it to solve outlier detection problems. Outliers are a kind of peculiar data and outlier detection is a very important problem in data mining aiming to find outliers. There are already many studies on the problem [1, 4, 7, 8, 9, 10, 11, 17, 19]. We give an outlier detection algorithm called LPF-Outlier in Algorithm 1. Its computational complexity is $O((k + \log n)mn)$, which is acceptable for many real world problems.

Algorithm 1 The LPF-Outlier algorithm for outlier detection

```
Input dataset $\{Z_1, Z_2, \cdots, Z_n\}$, the near neighbor parameter $k$, the parameter $q$ determining the number of outliers.
Output $q$ outliers.
for $i = 1$ to $n$ do
    calculate the LPF value of $Z_i$, $LPF(Z_i)$, by Eq. (5) and Eq. (6).
end for
Sort these $LPF(Z_i)$s in descending order and mark sample data corresponding to the former $q$ ones as outliers.
```

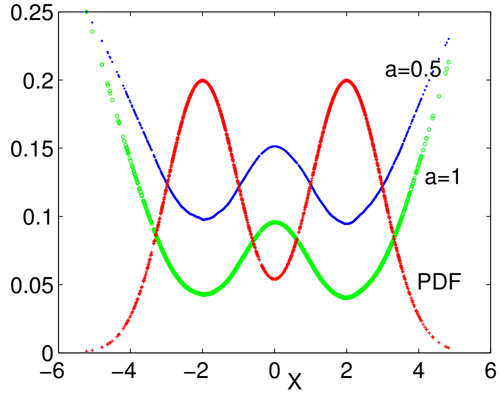

Figure 3: The PDF values and LPF values ($k = 700$, and $\alpha = 1$ or $\alpha = 0.5$) of data sampled from the distribution $\frac{1}{2}(\mathbf{N}(2, 1) + \mathbf{N}(-2, 1))$.

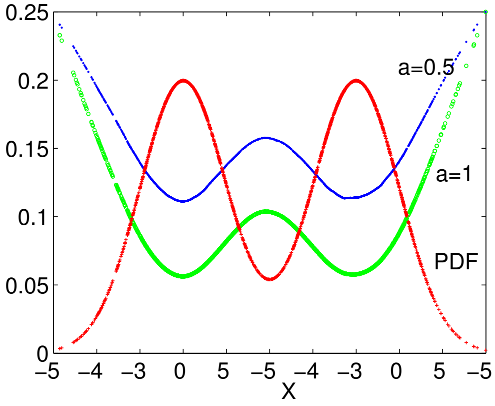

Figure 4: The PDF values and LPF values ($k = 900$, and $\alpha = 1$ or $\alpha = 0.5$) of data sampled from the distribution $\frac{1}{2}(\mathbf{N}(2, 1) + \mathbf{N}(-2, 1))$.

3. EXPERIMENTS

Unless noted, in the following experiments we select $D(Z_{ij}, Z_{lj}) = |Z_{ij} - Z_{ij}|^{\alpha}$, $\alpha = 1$ and $\beta_j = 1$ to calculate PF values and LPF values.

3.1 Experiments on a Synthetic Dataset

For a dataset with its generating PDF known, we intend to illustrate that a data point with a smaller PDF value has a large LPF value, and that the LPF can describe the PDF more accurately than the PF. We sampled 2000 data points from the distribution $\frac{1}{2}(\mathrm{N}(2, 1) + \mathrm{N}(-2, 1))$ and calculated their PDF values and LPF values to demonstrate that the LPF can describe the PDF accurately. We selected $\alpha = 1$ or $\alpha = 0.5$ in Eq. (6) to calculate LPF values and compressed them into the range (0, 0.25]. The PDF values and LPF values with $k = 700$ and $k = 900$ are plotted in Fig. 3 and Fig. 4, respectively. It can be seen that for this distribution, data points with smaller PDF values have larger LPF values. Hence LPF can really describe the peculiarity of low frequency data points. By comparing Fig. 2 to Fig. 3 and Fig. 4, we can see that the LPF can more accurately characterize the PDF of the distribution than the PF.

To more accurately demonstrate the fact that data points with smaller PDF values have larger LPF values, we sort these PDF values in ascending order and LPF values in descending order, and compare the two ordered sequences. A usual measure between two rankings in learning to ranking problems of machine learning is Kendall's Tau distance [2]. We give its definition for ordered PDF values and LPF values as follows.

Definition 6. ([2]) *Suppose that $\mathcal{Z} = \{Z_1, Z_2, \cdots, Z_n\}$ is a dataset, $p(\mathcal{Z})$ and $LPF(\mathcal{Z})$ are the set of PDF values and LPF values of Z_is. Sort $p(\mathcal{Z})$ in ascending order and denote $I(i)$ the index of the ith small element of the sorted $p(\mathcal{Z})$, that is, $p(Z_{I(1)}) \leq p(Z_{I(2)}) \leq \cdots \leq p(Z_{I(n)})$. Then the Kendall's Tau distance between the sorted $LPF(\mathcal{Z})$ and $p(\mathcal{Z})$ is defined by*

$$\tau(LPF(\mathcal{Z}), p(\mathcal{Z})) = \frac{2}{n(n-1)} \sum_{i<l} S((p(Z_{I(i)}) - p(Z_{I(l)})) (LPF(Z_{I(l)}) - LPF(Z_{I(i)}))), \tag{7}$$

where S is the sign function.

The measure $\tau(LPF(\mathcal{Z}), p(\mathcal{Z}))$ is a function of $LPF(\mathcal{Z})$ and $p(\mathcal{Z})$ with range [-1, 1]. If the index sequence of the sorted $p(X)$ (ascending order) and the index sequence of the sorted $LPF(\mathcal{Z})$ (descending order) are completely different, we have $\tau(LPF(\mathcal{Z}), p(\mathcal{Z})) = -1$, and if the two index sequences are completely same, we have $\tau(LPF(\mathcal{Z}), p(\mathcal{Z})) = 1$. For general cases, it can be seen from Eq. (7) that the more similar the two sequences are, the larger $\tau(LPF(\mathcal{Z}), p(\mathcal{Z}))$ is. Thus the LPF can describe the PDF more accurately.

We use $\tau(LPF(\mathcal{Z}), p(\mathcal{Z}))$ to quantitatively analyze the ability of the LPF in describing PDF and to intuitively demonstrate the impact of the parameter k on the performance of the LPF, which can help us to select a better parameter value in calculating LPF values. We sampled 2000 data points from the distribution $\frac{1}{2}(\mathrm{N}(2, 1) + \mathrm{N}(-2, 1))$, calculated their PDF values and LPF values as plotted in Fig. 3 and Fig. 4, and calculated the Kendall's Tau distance between the two sets of ordered values. We did 50 random experiments for each k, and the means and standard deviations of these distances are plotted in Fig. 5. It can be seen that for the distribution, the LPF accurately characterizes the PDF when k is in the range [500, 1000]. When k is increased further, the ability of the LPF drops. The description is worst with k equal to the size of the dataset, at which the LPF is equivalent to the PF. That is to say, for the distribution, the LPF can always characterize the PDF more accurately than the PF.

Fig. 5 also shows that the performance of the LPF is reasonably good and stable with k in a large range. This makes the selection of k easy. For a dataset with its generating PDF unknown, we roughly estimate the number of the peaks in the PDF by our prior knowledge. If the number is big, we select a small parameter k, and if the number is small, we select a large k. The parameter k should be the size of the dataset while there is only one peak in the PDF, in which case the LPF is same as the PF.

3.2 Experiments on Real Life Datasets

The LPF-Outlier algorithm has been implemented on four UCI datasets, a KDD dataset and a Mammography dataset, and its results have been compared with that of some other outlier detection methods. These experiments are divided into two groups. The first group is on two UCI datasets, Lymphography dataset and Wisconsin breast cancer dataset [14], and compared outlier detection algorithms includes KNN [17], RNN [7], FindCBLOF [8] and FindFPOF [9]. The second group is on two other UCI datasets, Ann-

Table 1: Detected rare classes in Lymphography dataset.

$Top\ ratio$	Number of rare classes included ($coverage$)			
(number of records)	LPF-Outlier	FindFPOF	FindCBLOF	KNN
5% (7)	**5 (83%)**	**5 (83%)**	4 (67%)	1 (17%)
7% (10)	**6 (100%)**	5 (83%)	4 (67%)	1 (17%)
10% (15)	**6 (100%)**	5 (83%)	4 (67%)	1 (17%)
11% (16)	**6 (100%)**	**6 (100%)**	4 (67%)	1 (17%)

Table 2: Detected malignant records in Wisconsin breast cancer dataset.

$Top\ ratio$	Number of malignant included ($coverage$)				
(number of records)	LPF-Outlier	FindCBLOF	FindFPOF	RNN	KNN
0.83% (4)	**4 (10.26%)**	**4 (10.26%)**	3 (7.69%)	3 (7.69%)	**4 (10.26%)**
1.66% (8)	**8 (20.51%)**	7 (17.95%)	7 (17.95%)	7 (17.95%)	4 (10.26%)
3.31% (16)	**16 (41.03%)**	14 (35.90%)	14 (35.90%)	11 (28.21%)	9 (23.80%)
4.97% (24)	**24 (61.54%)**	21 (53.85%)	21 (53.85%)	18 (46.15%)	15 (38.46%)
6.63% (32)	**30 (76.92%)**	27 (69.23%)	28 (71.79%)	25 (64.10%)	20 (51.28%)
8.28% (40)	**35 (89.74%)**	32 (82.05%)	31 (79.49%)	30 (76.92%)	21 (53.83%)
9.94% (48)	**39 (100.00%)**	35 (89.74%)	35 (89.74%)	35 (89.74%)	26 (66.67%)
11.59% (56)	**39 (100.00%)**	38 (97.44%)	**39 (100.00%)**	36 (92.31%)	28 (71.79%)

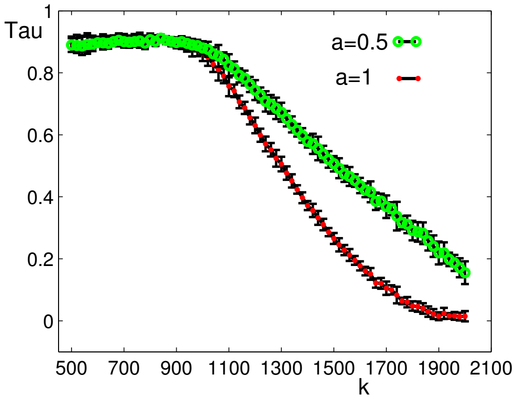

Figure 5: The means and covariances of the Kendall's Tau distances between ordered PDF values and ordered LPF values with different k, where data points were sampled from the distribution $\frac{1}{2}(\mathbf{N}(2,1) + \mathbf{N}(-2,1))$, and the parameter α in the LPF is $\alpha = 1$ or $\alpha = 0.5$.

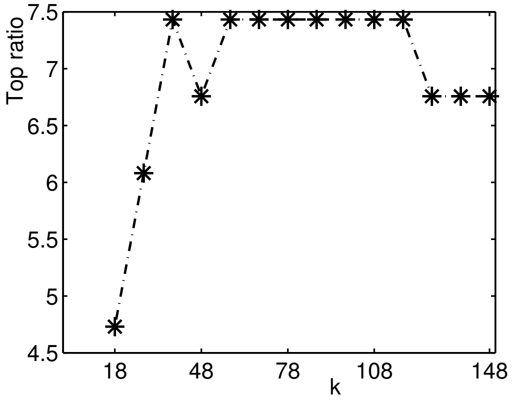

Figure 6: The ratio of the smallest number of top ranked records with all rare classes being detected to the total number of the records for LPF-Outlier with different ks.

thyroid dataset and Shuttle dataset, KDD dataset and Mammography dataset. On these datasets we compare LPF-outlier with methods including LOF [3, 4], Feature Bagging [11], Bagging, Boosting and Active-Outlier [1].

3.2.1 Experiments on the First Group

The Lymphography dataset consists of 148 instances with labels 1, 2, 3 or 4, and each record is described by 18 continuous or discrete attributes. Only 6 (4.1%) of the records are labeled with 1 or 4, and 142 (95.9%) are labeled with 2 or 3. We take instances with label 1 or 4 as outliers, and with label 2 or 3 as normal instances. The data was preprocessed by the method in [9, 12], and the experimental results in [12] and [9] are directly cited for comparison.

The Wisconsin breast cancer dataset consists of 458 instances with label benign and 241 instances with label malignant, and each record is described by 9 attributes. We followed the experimental technique in [7] and [9] by removing some of the malignant records to form a very unbalanced distribution, and the resultant dataset has 39 (8%) malignant and 444 (92%) benign records. The dataset can be used to evaluate the performance of an outlier detection algorithm by its ability in detecting the malignant records.

The results of LPF-Outlier, FindCBLOF, FindFPOF and KNN on Lymphography dataset are given in Table 1, where $Top\ ratio$ is the ratio of the number of detected records to the total data, $coverage$ is the ratio of the number of detected outliers to the total outliers, and the results of LPF-Outlier were calculated with $k = 30$ in Eq. (5). We can see that LPF-Outlier and FindFPOF are both the best ones at $Top\ \ ratio = 5\%$, and at all other $Top\ \ ratio$'s LPF-Outlier can detect outliers more effectively than FindCBLOF, FindFPOF and KNN. The results of LPF-Outlier, FindCBLOF, FindFPOF, RNN and KNN on the Wisconsin breast cancer dataset are given in Table 2, where the results of LPF-Outlier were calculated with $k = 200$ in Eq. (5). It can be seen that LPF-Outlier, FindCBLOF and KNN are all the best ones at the $Top\ ratio = 0.83\%$,

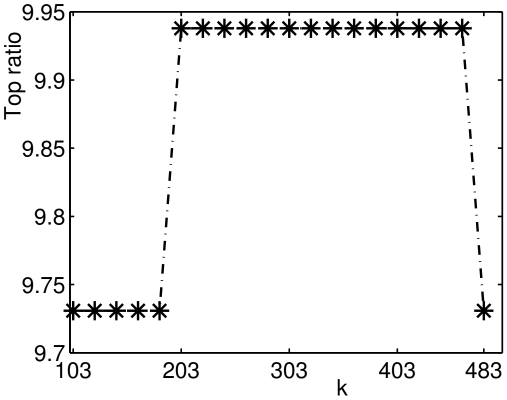

Figure 7: The ratio of the smallest number of top ranked records with all malignant being detected to the total number of the records for LPF-Outlier with different ks.

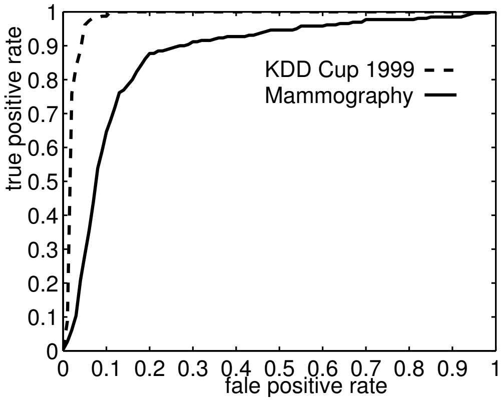

Figure 8: ROC curves for LPF-Outlier for the KDD dataset ($k = 13000$) and the Mammography dataset ($k = 11183$).

and at all other $Top\ \ ratio$'s LPF-Outlier can detect outliers more effectively than FindCBLOF, FindFPOF, RNN and KNN.

To investigate the impact of the parameter k on the performance of LPF-Outlier, the values of $Top\ \ ratio$ with different ks and $coverage = 100\%$ for the Lymphography dataset and the Wisconsin breast cancer dataset are plotted in Fig. 6 and Fig. 7, respectively. Fig. 6 shows that the performance of LPF-Outlier is comparatively stable while k varying and LPF-Outlier with $k \leq 28$ performs better than the PF based outlier detection algorithm (LPF-Outlier with $k = 148$). Fig. 7 shows that the performance of LPF-Outlier is also comparatively stable and while $k \leq 183$ it performs better than the outlier detection algorithm based on the PF (LPF-Outlier with $k = 483$).

3.2.2 Experiments on the Second Group

For the four datasets of the second group, the set-up for our experiments followed the ones in [1, 5, 11], and our results were directly compared with the results reported in these papers. In particular, in order to preprocess discrete attributes, we use the concept of inverse document frequency (IDF), where each value of categorical attribute is represented with the inverse frequency of its appearance in the dataset, and the values of each attribute are normalized to [0,1]. IDF is already used in outlier detection problems [6].

The Mammography dataset[1] consists of 10923 records with label 1 and 260 records with label 2, and each record is characterized by 6 continuous attributes. We take records with label 2 as outliers and others as normal records.

KDD Cup 1999 was prepared for network intrusion detection and is available from the UCI KDD Archive (http://kdd.ics.uci.edu/). There are six sub-datasets at the address and we selected the test data with corrected labels (corrected.gz) to extract the KDD dataset. Each data in the dataset is specified by 41 attributes (34 continuous and 7 categorical) and a label describing attack type, where all labels except "normal" indicate one of the four classes of attack, i.e., U2R, DOS, R2L and Probe. The KDD dataset consists of the normal records (60593) and the records of U2R (228), which is the smallest class of intrusion.

The Shuttle dataset consists of 14500 instances with label 1, 2, 3, 4, 5, 6 or 7, and each record has 9 numerical attributes. We created five data sets by selecting classes 2, 3, 5, 6 and 7 to be detected as outliers compared to the biggest remaining class 1.

The Ann-thyroid dataset consists of 3428 instances with label 1, 2 or 3 and each record is characterized by 15 binary attributes and 6 continuous attributes. We take the records of class 1 and class 2 as outliers versus the class 3 as the normal (majority) class, and two datasets, Ann-thyroid 1 and Ann-thyroid 2, were created respectively.

Before giving the experimental results on above datasets, we first introduce the concept of receive operating characteristic curve (ROC) and area under the curve (AUC), which have been used to evaluate the performance of outlier detection algorithms[1, 11]. Suppose that the dataset consists of actual outliers and actual normal records, predicted outliers include true positives (TP) and false positives (FP), and predicted normal records include false negatives (FN) and true negatives (TN). Then the detection rate and the false alarm rate are defined by TP/(TP+FN) and FP/(FP+TN), respectively. The ROC curve is the curve plotted with the detection rate as x-coordinates and the false alarm rate as y-coordinates. The AUC is the area under the ROC curve. The ideal ROC curve has 0% false alarm rate and 100% detection rate. However, the ideal ROC curve can only be approximated in practice and AUC quantitatively evaluate the approximation. Outlier detection algorithms with AUC closer to 1 have ROC closer to the ideal ROC, and are better algorithms.

The ROC curves for LPF-Outlier on the KDD dataset and the Mammography dataset are plotted in Fig. 8, where LPF values in the LPF-Outlier were calculated with $k = 13000$ and $k = 11183$, respectively. The AUC achieved by LPF-Outlier with different k's for the KDD dataset and the Mammography dataset are given in Fig 9.and Fig. 10, where$D(Z_{ij}, Z_{lj}) = |Z_{ij} - Z_{lj}|$ or $D(Z_{ij}, Z_{lj}) = |Z_{ij} - Z_{lj}|^{\frac{1}{2}}$ was selected for Fig. 9. The AUC achieved by LPF-Outlier, Active-Outlier, Bagging, Boosting, LOF and Feature Bagging for all the four datasets are given in Table 3. The results of LPF-Outlier given in Table 3 are the best ones with k shown in Fig. 9 and Fig. 10 for the KDD dataset and the Mammography dataset, is the mean of the best ones obtained on its five sub-datasets for the Shuttle dataset, and are chosen using the best parameter choices for Ann-thyroid 1 and Ann-thyroid 2. The AUC for other algorithms are directly copied from [1]. From Fig. 9, Fig. 10 and Table 3, we can see that for the KDD dataset and the Mammography dataset, LPF-Outlier with any k can detect outliers more ef-

[1] We would like to thank Professor Nitesh V. Chawla for providing us this dataset.

Table 3: The AUC achieved by LPF-Outlier, Active-Outlier, Bagging, Boosting, LOF and Feature Bagging for the four datasets of the second group.

Data sets	LPF-Outlier	Active-Outlier	Bagging	Boosting	LOF	Feature Bagging
Mammography	**0.87**	0.81(±0.03)	0.74(±0.07)	0.56(±0.02)	0.64(±0.1)	0.80(±0.1)
KDD dataset	**0.98**	0.94(±0.04)	0.61(±0.25)	0.51(0.004)	0.61(±0.1)	0.74(±0.1)
Shuttle (mean)	0.992	**0.999(±0.0006)**	0.985(±0.031)	0.784(±0.13)	0.852	0.839
Ann-thyroid 1	0.97	0.97(±0.01)	**0.98(±0.01)**	0.64(±0.08)	0.869	0.869
Ann-thyroid 2	0.73	0.89(±0.11)	**0.96(±0.02)**	0.54(±0.01)	0.761	0.769

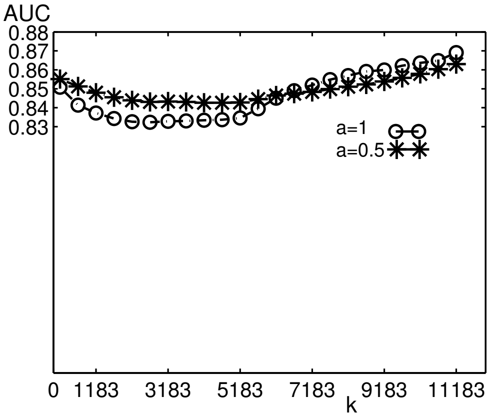

Figure 9: The AUC achieved by LPF-Outlier with different ks for the Mammography dataset.

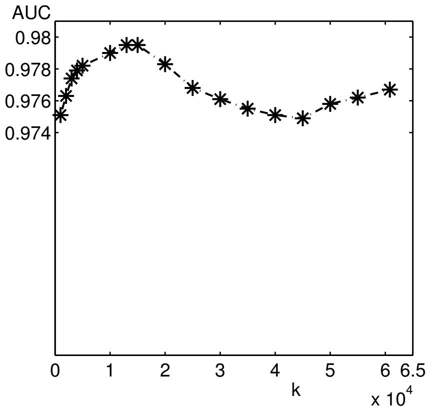

Figure 10: The AUC achieved by LPF-Outlier with different ks for the KDD dataset.

fectively than Active-Outlier, Bagging, Boosting, LOF and Feature Bagging. For the Shuttle dataset and the Ann-thyroid 1 dataset, LPF-Outlier performs quite well. However, for the Ann-thyroid 2 dataset, LPF-Outlier performs worse than four of the other five methods. The unfavorable performance may be explained by the fact that there are too many discrete attributes comparing to continuous attributes and the IDF may be ineffective in preprocessing such a dataset. In summary, these experimental results show that LPF-Outlier is effective for outlier detection.

4. CONCLUSIONS

In this paper, we have studied the PF from both theoretical analysis and experimental evaluation. We have proved that the PF can describe the PDF of a normal distribution accurately and is unsuitable for characterizing the PDF of general distributions. The concept of LPF was proposed to solve the difficulty. The LPF was proved to have the same power as the PF in characterizing the PDF of a normal distribution and be the so-called ϵ-sensitive peculiarity description for general distributions. Based on the LPF, an outlier detection algorithm, LPF-Outlier, was proposed for outlier detection problems. Experiments on a synthetic dataset have shown that the LPF can characterize the PDF of a distribution more accurately than the PF and experiments on several real life datasets have demonstrated that LPF-Outlier is a novel and quite effective outlier detection method.

Acknowledgments

The first author would like to thank Fuxin Li for his helpful discussions. This work was partially supported by the National Grand Fundamental Research 973 Program of China (No. 2004CB318103), the National Science Foundation of China (No. 60575001 and No. 60673015), the China Postdoctoral Science Foundation (No. 20070 420016), and the grant-in-aid for scientific research (No. 18300053) from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

5. REFERENCES

[1] N. Abe, B. Zadrozny, and J. Langford. Outlier detection by active learning. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 504–509, 2006.

[2] N. L. Bhamidipati and S. K. Pal. Comparing rank-inducing scoring systems. *Proceedings of the 18th International Conference on Pattern Recognition*, pages 300–303, 2006.

[3] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[4] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. *Proceedings of the 6th ACM SIGMOD International Conference on Management of Data*, pages 93–104, 2000.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[6] L. Ertoz. *Similarity Measures*. Ph.D. Dissertation, University of Minnesota, 2005.

[7] S. Harkins, H. He, G. J. Willams, and R. A. Baster. Outlier detection using replicator neural networks. *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, pages 170–180, 2002.

[8] Z. Y. He, X. F. Xu, and S. C. Deng. Discovering cluster based local outliers. *Pattern Recognition Letters*, 24:164–1650, 2003.

[9] Z. Y. He, X. F. Xu, Z. X. Huang, and S. C. Deng. A frequent pattern discovery method for outlier detection. *Proceedings of the 5th International Conference on Web-Age Information Management*, pages 726–732, 2004.

[10] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. *Proceedings of the 12th International Conference on Very Large Data Bases*, pages 392–403, 1998.

[11] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 157–166, 2005.

[12] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 80–86, 1998.

[13] K. Mcgarry. A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20:39–61, 2005.

[14] G. Merz and P. Murphy. Uci repository of machine learning databases. *http://www.ics.uci.edu/mlearn/MLRepository.html*, 1996.

[15] M. Ohshima, N. Zhong, Y. Y. Yao, and C. Liu. Relational peculiarity oriented mining. *Data Mining and Knowledge Discovery*, 15:249–273, 2007.

[16] M. Ohshima, N. Zhong, Y. Y. Yao, and S. Murata. Peculiarity oriented analysis in multi-people tracking images. *Advances in Knowledge Discovery and Data Mining*, pages 508–518, 2004.

[17] S. Ramaswamy, R. Rastogi, and S. Kyuseok. Efficient algorithms for mining outliers from large data sets. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 427–438, 2000.

[18] D. Saso and L. Nada. An introduction to inductive logic programming. *Relational Data Mining*, pages 48–73, 2001.

[19] Y. Y. Yao, F. Y. Wang, J. Wang, and D. D. Zeng. Rule + exception strategies for security information analysis. *IEEE Intelligent Systems*, 20:52–57, 2005.

[20] Y. Y. Yao and N. Zhong. An analysis of peculiarity oriented data mining. *Proceedings of the 2002 IEEE International Conference on Data Mining Workshop on the Foundation of Data Mining and Discovery*, pages 185–188, 2002.

[21] N. Zhong, C. Liu, Y. Y. Yao, M. Ohshima, M. X. Huang, and J. J. Huang. Relational peculiarity oriented data mining. *Proceedings of the 2004 IEEE International Conference on Data Mining*, pages 575–578, 2004.

[22] N. Zhong, M. Ohshima, and S. Ohsuga. Peculiarity oriented mining and its application for knowledge discovery in amino-acid data. *Advances in Knowledge Discovery and Data Mining*, pages 260–269, 2001.

[23] N. Zhong, Y. Yao, and M. Ohshima. Peculiarity oriented multi-database mining. *IEEE Transactions on Knowledge and Data Engineering*, 15:952–960, 2003.

[24] N. Zhong, Y. Y. Yao, M. Ohshima, and S. Ohsuga. Interestingness, peculiarity, and multi-database mining. *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 566–573, 2001.

APPENDIX

The Proof of Theorem 1: Without loss of generality, we may suppose that the mean of the distribution locates at the origin, that is, $p(x) = p(-x)$ for all $x \in \mathbb{R}^+$. We only prove that $PF(x)$ strictly increases with nonnegative x and a similar deduction can be done for $x < 0$. Furthermore, we suppose that $0 \leq x_1 < x_2$ and will prove that $PF(x_1) < PF(x_2)$.

Denote $H(x) = \phi(|x_2 - x|) - \phi(|x - x_1|)$. Since $p(x)$ is positive and decreasing, and $H(x) = \phi(x_2 - x) - \phi(x - x_1)$ is integrable on $[x_1, x_2]$, there exists a $\xi \in [x_1, x_2]$ such that

$$\int_{x_1}^{x_2} H(x)p(x)dx = p(x_1) \int_{x_1}^{\xi} H(x)dx.$$

If $\xi \leq \frac{x_2+x_1}{2}$, we have

$$\int_{x_1}^{\xi} H(x)dx \geq 0.$$

If $\xi > \frac{x_2+x_1}{2}$, we have

$$\int_{x_1}^{\xi} H(x)dx \geq \int_{x_1}^{x_2} H(x)dx = 0.$$

Hence we can get

$$\int_{x_1}^{x_2} H(x)p(x)dx \geq 0.$$

Since

$$\int_{0}^{x_1} H(x)dx = \int_{0}^{x_1} H(x_1 - x)dx,$$

the following inequalities hold

$$\int_{0}^{x_1} H(x)p(x)dx > \int_{0}^{x_1} H(x_1 - x)p(x_2 + x) > 0.$$

Therefore

$$\begin{aligned}
& \int_{-\infty}^{x_1} H(x)p(x)dx + \int_{x_2}^{+\infty} H(x)p(x)dx \\
= & \int_{0}^{+\infty} H(-x)p(x)dx - \int_{0}^{+\infty} H(-x)p(x_1 + x_2 + x)dx \\
& - \int_{0}^{x_1} H(x_1 - x)p(x_2 + x)dx + \int_{0}^{x_1} H(x)p(x)dx \\
> & \ 0.
\end{aligned}$$

Then we get

$$\begin{aligned}
& PF(x_2) - PF(x_1) \\
= & \int_{-\infty}^{x_1} H(x)p(x)dx + \int_{x_1}^{x_2} H(x)p(x)dx + \int_{x_2}^{+\infty} H(x)p(x)dx \\
> & \ 0.
\end{aligned}$$

That is to say, $PF(x)$ strictly increases with x for $x \geq 0$. ■