An Efficient Method for Outlying Aspect Mining Based on Genetic Algorithm

Zihao Chen, Lei Duan(✉), and Xinye Wang

School of Computer Science, Sichuan University, Chengdu, China
{chenzihao,wangxinye}@stu.scu.edu.cn, leiduan@scu.edu.cn

Abstract. Outlying aspect mining (OAM) aims to identify a feature subspace in which a given query object is dramatically distinctive from the rest data. The identified features can assist the formulation and optimization of decisions. Score-and-search methods are widely used in outlying aspect mining. However, limited by scoring instability and search inefficiency, studies using this strategy are unable to be comprehensive and accurate for mining outlying aspects. In this paper, it proposes a novel OAM method based on genetic algorithm, named OSIER, which can be applied in mining outlying aspects from multi-dimensional spaces. OSIER improves the search efficiency by analyzing the correlations between dimensions. By combining the genetic algorithm with the traditional beam search strategy, OSIER effectively improves the diversity of the searched aspects. As a result, the execution time for candidate outlying aspects search is controlled in an acceptable range. Experiments show that OSIER outperforms the benchmark methods in terms of effectiveness on the OAM task. Besides, OSIER is capable of providing valuable outlying aspect mining results for various types of datasets.

Keywords: Outlying aspect mining · Kernel density estimation · Genetic algorithm

1 Introduction

Outlying aspect mining aims to discover an aspect (i.e., a set of features or attributes) in which a query point has the most significant outlyingness. In many real-life application scenarios, it can provide interpretable information and decision support for downstream tasks [13]. For example, a recruitment team somehow highly interested in identifying what are the most outstanding merits or shortcomings of a particular candidate compared to others.

It is worth noting that the distribution of query points in different spaces varies significantly. As Fig. 1(a) shows, all the data samples are scattered in a 3-dimensional space. The outlyingness of the query point (red triangle) is not significant in the full space. After projecting the data into various 2-dimensional subspaces, the red triangle is more distinguishable from the other points (blue

W. Chen et al. (Eds.): ADMA 2022, LNAI 13725, pp. 337–351, 2022.
https://doi.org/10.1007/978-3-031-22064-7_25

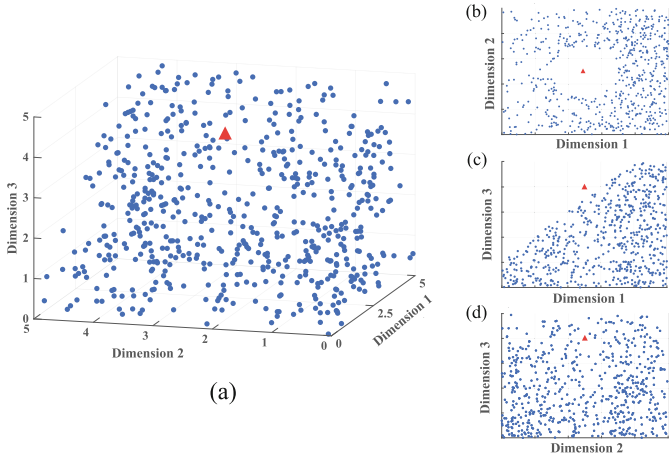

Fig. 1. An example of data distribution in different spaces. (a) Data distribution in full space; (b)(c)(d) Data distribution in three projection subspaces.

dots) in Fig. 1(b) and Fig. 1(c). The space that can exhibit significant differences between the query point and others is called the outlying aspect.

Technically, it is not feasible to enumerate all aspects due to the number of subspaces grows exponentially with the increasing data dimensionality. To find the optimal outlying aspect efficiently, we consider the following two challenges:

- **(C1) How to search aspects efficiently and comprehensively?** To reduce the enormous computational cost of enumerating aspects, it should select representative aspects in the search process. But the distributions of different spaces are not regular, making it challenging to find all representative aspects.
- **(C2) How to measure the outlyingness of different aspects impartially?** A scoring function needs to be designed to quantify the outlyingness of a query point in different aspects. However, calculating the outlyingness in different dimensional aspects may lead to biased results.

Current approaches have limitations in addressing the above challenges. Concerning the search strategy **(C1)**, the most advanced and general approach is the beam algorithm [10] using heuristic rules. This search strategy has an assumption that if a point scores high in an l-dimensional aspect, it generally generated by adding a dimension in a well-behaved $(l-1)$-dimensional aspect. However, Fig. 1(b) and Fig. 1(c) show extreme cases that disprove this assumption. Higher-dimensional aspects with extreme distribution are not searchable. As for the outlyingness scoring function **(C2)**, Zhang *et al.* [17] used the distance from surrounding points to the query point as a metric, leading to results biased towards higher-dimensional aspects. The subsequent methods chose density as the evaluation criterion, but they still have shortcomings. The density ranking [5] loses

the absolute degree of deviation, Z-score normalization [10] tends to aspects with high variance, and the method of generating hypersphere simulation densities by random sampling [12] produces a large instability.

In this paper, we propose a score-and-search method OSIER (short for outlying aspects mining based on genetic algorithm), which can effectively identify the candidate outlying aspects. OSIER makes improvements to overcome the shortcomings of existing methods from two perspectives. For **C1**, it retain the strategy of generating high-dimensional aspects from well-behaved aspects, while increasing the diversity of the searched aspect by replacing dimensions with a mutation operation. This search method has the probability of getting rid of the local optimum. For **C2**, it uses the strategy of stage comparison. It compares the absolute density of the query point in different aspects under the same dimension. Moreover, the optimal aspects under each dimension are compared with a normalized score. In this way, it retains the absolute density information of all data, avoiding biased results, and accelerates computational efficiency. The main contributions of this paper are summarized as follows:

- We propose a novel genetic algorithm-based method, named OSIER, to increase the diversity of searched aspects and generate more representative aspects efficiently.
- We use a simplified density estimation function in a multi-dimensional space and analyze an improved outlyingness measure guided by prior knowledge.
- We calculate the impact of each dimension on the outlying aspect by generating a maximum interval hyperplane, which can be used to guide the direction of evolution in genetic algorithms. Besides, we use a new comparison strategy preserving the absolute density of all data to avoid biased results.
- We demonstrate OSIER on multiple real-world and synthetic datasets. The experimental results show that OSIER has higher search efficiency and stability, which can be applied to some extreme cases.

The rest of the paper is organized as follows. The related work is reviewed in Sect. 2. Section 3 presents the details of OSIER. The experiments and results are provided in Sect. 4, and Sect. 5 shows the conclusion and future work.

2 Related Work

2.1 Outlying Aspect Mining

Assume that there exists a d-dimensional space $D = \{D_1, D_2, ..., D_d\}$ and a set of data points $X \in \mathbb{R}^d$. For a point $X_i \in X$, its feature under the full space is represented as $\{X_i.D_1, X_i.D_2, ..., X_i.D_d\}$.

We call the combination of multiple dimensions as aspects (a dimension can also be considered as an aspect), and the space composed of these dimensions is a subspace of the full space D. For an aspect $\mathcal{S} = \{D_{i_1}, D_{i_2}, ..., D_{i_{|S|}}\}$, we can define a measure of outlyingness scoring function $\rho_{\mathcal{S}}(q)$, which measures the outlyingness of query point q in the aspect $\mathcal{S}$. Based on the above definitions, we can formalize the *Outlying Aspect Mining* problem as follows:

Definition 1 (Outlying Aspect Mining). *Given a set of n instance X in d-dimensional space D, a query point $q \in X$, the outlying aspect mining is to identify the non-empty aspect $\mathcal{S} \subseteq D$ in which the query point q's outlying degree $\rho_{\mathcal{S}}(q)$ is larger than any other aspect.*

The existing OAM methods are mainly divided into two categories, feature selection methods [4,9] and score-and-search methods [5,10,17]. Feature selection methods transform the OAM problem into a classical feature selection classification problem. More specifically, the two classes are defined as the query points (positive class) and the rest of the data (negative class). Therefore, there is a data imbalance problem when the model is trained. Besides, the interpretability of these methods is poor. The score-and-search methods are more widely and deeply researched than feature selection methods.

The frame of score-and-search methods is to search each candidate aspect, calculate outlyingness for query points, and select the aspect with the highest outlyingness as the optimal result. Zhang *et al.* [17] proposed a metric based on the idea of kNN, called outlying degree. Duan *et al.* [5] applied kernel density estimation to multidimensional space by using a product of univariate Gaussian kernels. Meanwhile, they used a boundary pruning-based search strategy.

Vinh *et al.* [10] considered that the use of density ranking would lose important information about the degree of absolute deviation. Thus they designed a standard scoring function and proposed the concept of dimensionality unbiasedness for outlying aspect mining measures.

Definition 2 (Dimensionality Unbiasedness). *If a density scoring function $\rho_{\mathcal{S}}(\cdot)$ satisfies the formula:*

$$\frac{1}{n}\sum_{X_i \in X} \rho_{\mathcal{S}}(X_i) = const.w.r.t.|\mathcal{S}| \tag{1}$$

the function can be used to compare the outlyingness of query points in different dimensional spaces directly.

Dimensionality unbiasedness provides a desirable property for designing density scoring functions, thus avoiding bias due to different dimensions and making OAM more interpretable. Meanwhile, to prevent the problem of exploding the number of high-dimensional aspects, a beam search method was proposed to ensure that the number of search spaces is within a specific range by heuristic pruning. Wells *et al.* [15] analyzed the shortcomings of kernel density search and proposed SGrid density estimation instead, thus considerably speeding up the computational process. Samariya *et al.* [12] generated hyperspheres by random sampling to evaluate the outlyingness of query points to solve the complex problem of density computation in high-dimensional space.

It is worth noting that outlier detection and outlying aspect mining are different. Outlier detection aims to detect anomalous data that are exceptional with respect to the majority of objects in the databases. It can applied in various fields, such as disease detection [6], social media monitoring [19] and network

intrusion supervision [1]. However, outlier detection is difficult to provide a reasonable and intuitive explanation for the identified objects. The OAM task was proposed for discovering aspects where the query instance exhibits the most outlying characteristics.

2.2 Genetic Algorithm

In genetic algorithms, each individual consists of a gene string that represents a feasible solution to that problem. Fitness is the metric used to evaluate the individual, and the fitness function is usually determined based on the objective function. The selection operation selects a parent based on fitness and inherits its genes to the next generation of individuals. The selected parent undergoes a crossover operation with a certain probability to produce the next individual. After many generations, the genetic algorithm jumps out of the loop with a defined threshold or number of iterations and obtains a better quality solution. Genetic algorithm has been used to solve a large variety of problems efficiently, including classification [3], credit risk assessment [8] and time-series analysis [11].

Zhu *et al.* [18] used genetic algorithms in the outlier detection problem. They used cell-based segmentation techniques, which resulted in a high outlyingness computation cost in a high-dimensional space. Zhang *et al.* [16] devised a method that does not depend on the upper and lower bound closure properties. Similar to [17], they chose distance as outlyingness, which is used to guide the evolution.

The genetic algorithm is an efficient optimization algorithm for intelligent global search, which is simple and robust. Thus, we can use these characteristics to discover outlying aspects efficiently.

3 Design of OSIER

In this section, we discuss the details of OSIER. It takes a query point q together with a dataset X of n points $\{X_1, ..., X_n\}$ as input, $X_i \in \mathbb{R}^d$, and outputs an aspect in which the given point has the highest outlyingness.

3.1 Outlying Scoring Function

In the choice of scoring function, we use a simplified version of the multidimensional density estimation function:

$$\rho_{\mathcal{S}}(q) = \frac{1}{nh^{|\mathcal{S}|}} \sum_{i=1}^{n} K\left(\frac{\|q - X_i\|_p}{h}\right) \tag{2}$$

where p denotes norm. In the absence of any prior knowledge, we choose the Euclidean norm ($p = 2$), adopt the Gaussian kernel as kernel function $K(\cdot)$, and calculate the bandwidth h follows Silverman's rule of thumb [14] is more general.

This default density estimation parameter can be improved by prior knowledge. One type of prior knowledge derived from the data description is the

bound of each dimension. Suppose the dataset is restricted in a dimension to a range of values. It is not reasonable to have a density distribution for points outside the range. For example, age is a non-negative number. It is unreasonable to produce a probability distribution in the space where age is negative by the density estimation function. We use a reflection strategy to solve this problem. If the data has a minimum value boundary b in dimension D_i, for query point $q = \{q.D_1, ..., q.D_i, ..., q.D_{|\mathcal{S}|}\}$, we set the symmetry point $q_{sym} = \{q.D_1, ..., 2b - q.D_i, ..., q.D_{|\mathcal{S}|}\}$. The optimized scoring function is:

$$\rho'_{\mathcal{S}}(q) = \rho_{\mathcal{S}}(q) + \rho_{\mathcal{S}}(q_{sym}) \tag{3}$$

For the case where there are boundaries on both sides, we only consider the first reflection point because the appropriate bandwidth ensures that the density distribution after multiple reflections is equal to 0 or infinitely close to 0.

In addition, if a dataset is composed of multiple datasets, resulting in far from normally distributed data, Silverman's rule of thumb [14] will result in a poor density estimate. We prefer to use an improved sheather algorithm [2] which can achieve better results when the dataset is distributed in multiple dense regions.

3.2 Dimensions Correlation Analysis

Before formally calculating the query points' outlyingness, each dimension of the dataset needs to be analyzed. Different dimensions provide different contributions to the generation of outlying aspects. Thus the analysis of a single dimension helps to guide subsequent search process.

To make the outlying aspects result more credible, we perform a deeper analysis of the density estimation method. For a query point, if the distribution of the projection on a dimension is remote and the probability density is minimal, the dimension can have a large impact on aspect generation. This kind of dimensions is called trivial outlying dimension, which is defined as follows:

Definition 3 (Trivial Outlying Dimension). *Given a query point q, an outlyingness scoring function $\rho(\cdot)$ and a threshold ϵ, a dimension D_i is called trivial outlying dimension if $\rho_{D_i}(q) \leq \epsilon$.*

An intuitive fact is that when a trivial outlying dimension is coupled with another dimension, the generated aspects may still have a good outlyingness score for the query point. Therefore, the trivial outlying dimensions should be pre-processed to reduce their impact on the search results. For a query point, if a trivial outlying dimension exists, aspects that may cause superior outlyingness will be replaced by different combinations of that trivial outlying dimension with other dimensions.

Example 1. Table 1 shows the shooting statistics for Los Angeles Lakers. We employ OAMiner [5] to figure out outlying aspects of *LeBron James*. The result shows that the top-5 outlying aspects are $\{3PM\}$, $\{3PM, 3PA\}$, $\{3PM, FGM\}$, $\{3PM, FTM\}$ and $\{3PM, 3PA, FGM\}$. When we ignore the effects of feature $3PM$, outlying aspect $\{FG\%, FT\%, 2P\%\}$ is revealed, which can offer more hidden information.

Table 1. Shooting statistics of six active players in Los Angeles Lakers

Name	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	2PM	2PA	2P%
LeBron James	11.1	21.3	52.3	2.9	8.0	36.2	4.6	6.0	76.6	8.3	13.4	61.8
Anthony Davis	9.2	17.2	53.7	0.3	1.8	18.2	4.4	6.1	70.9	8.9	15.4	57.8
Russell Westbrook	6.8	15.7	43.3	0.9	3.3	27.7	3.4	5.1	67.0	5.9	12.4	47.4
Stanley Johnson	2.3	4.8	47.0	0.6	2.0	31.9	0.9	1.2	70.7	1.6	2.8	57.9
Austin Reaves	2.3	4.8	47.6	0.8	2.6	31.8	1.3	1.5	83.8	1.4	2.1	67.3
Dwight Howard	1.9	3.1	60.9	0.1	0.2	66.7	1.3	2.0	62.9	1.8	3.0	60.6

Afterwards, we need to evaluate the correlation between dimensions and analyze the contribution of each dimension to the degree of query point outliers, represented as a set of weights $\mathbf{w} = \{w_1, w_2, ..., w_d\}$. We construct a hyperplane on the full space so that the query points are as separated as possible from other points to calculate w. Since the query point and the rest of the sample points are two classes of samples with extreme imbalance, we choose the one-class SVM method and set the query point as the origin. Solving this maximum geometric margin hyperplane is essentially a convex optimization problem:

$$\begin{aligned} &\min \frac{1}{2}||\mathbf{w}||^2 - \tau + \frac{1}{\upsilon n}\sum_{i=1}^{n}\xi_i \\ &s.t.\ \mathbf{w}^T X_i \geq \tau - \xi_i, \quad \xi_i \geq 0 \end{aligned} \tag{4}$$

where X_i denotes spatial vectors of the i-th point, τ denotes the hyperplane bias, ξ_i denotes the relaxation variable of the i-th point and υ denotes a trade-off parameter. We solve the quadratic programming problem by Lagrange Multiplier Method:

$$\mathcal{L}(\mathbf{w}, \tau, \xi, \alpha, \beta) = \frac{1}{2}||\mathbf{w}||^2 - \tau + \frac{1}{\upsilon n}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(\mathbf{w}^T X_i - \tau + \xi_i) - \sum_{i=1}^{n}\beta_i\xi_i \tag{5}$$

In order to obtain a specific form for solving the dual problem, let the partial derivative of $\mathcal{L}(\mathbf{w}, \tau, \xi, \alpha, \beta)$ with respect to W, τ, and ξ equal to zero. We can obtain the following three conditions:

$$\mathbf{w} - \sum_{i=1}^{n}\alpha_i X_i = 0 \tag{6}$$

$$\sum_{i=1}^{n}\alpha_i - 1 = 0 \tag{7}$$

$$0 \leq \alpha_i \leq \frac{1}{\upsilon n} \tag{8}$$

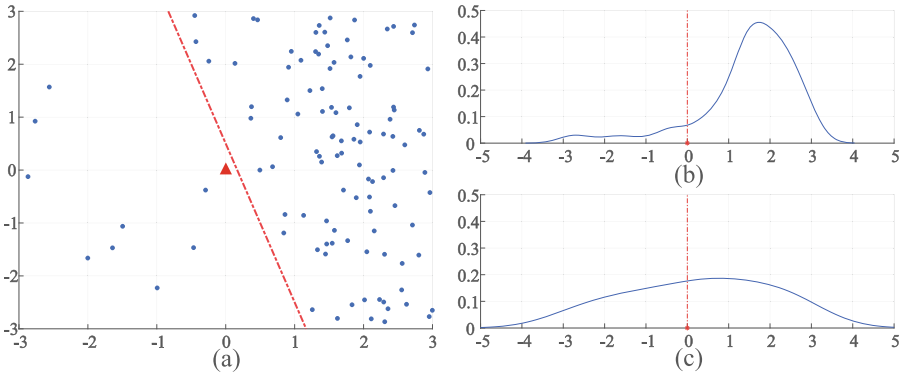

Fig. 2. Significance of the weights generated by hyperplane.

Equation 6 can be solved for $\mathbf{w}$, where $X = \{X_1, X_2, ..., X_d\}$ is known, and the optimal solution for α can be obtained by substituting Eq. (6)-(8) into Eq. 5:

$$\min \& \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i {X_i}^T X_j \alpha_j \\ s.t. \sum_{i=1}^{n} \alpha_i = 1, \quad 0 \leq \alpha_i \leq \frac{1}{\upsilon n} \tag{9}$$

In reality, it is reasonable to expect that if this hyperplane is more perpendicular to a dimension, the greater the contribution of this dimension in the classification. As Fig. 2(a) shows, red triangle indicates the query, and the red dashed line indicates the generated hyperplane. The hyperplane corresponds to a weight vector of $\mathbf{w} = [3, 1]$, indicating that the dimension corresponding to the x-axis has a greater influence on the query point becoming an outlier. Figure 2(b) and Fig. 2(c) denote the density distributions of the data after projection on the x-axis and y-axis, respectively, which also justify the analysis.

3.3 Outlying Aspect Generation

The strategy of searching candidate aspects is the core problem of computing the outlying aspects in a high-dimensional data set. When the dimensionality is large enough, the computation of traversing every aspect brings an unbearable computational cost. This cost is exponentially related to the number of dimensions. Taking OAMiner [5] as an example, it takes over 24 h on a dataset with 30 dimensions and 10,000 points, which is impracticable for many real-world high-dimensional datasets. OSIER uses a genetic algorithm, which includes recombination, mutation, and selection operations to search for representative aspects efficiently. The search strategy is given in Algorithm 1.

Procedure *Recombination*$(\cdot)$ generates new individuals by reorganizing parts of the structure of multiple parent individuals in Step 4. If an aspect performs

Algorithm 1. Pseudocode of OSIER

Input: a d-dimensional dataset X, a query point q, population P, mutation rate α.
Output: the outlying aspect of q.

1: Initialize the candidate dimension set $CSet = \{D_i | \rho_{D_i}(q) \leq \epsilon\}$ (optional, The full set can be used directly without considering trivial outlying dimension)
2: $C_1 \leftarrow CSet$
3: Best-scored aspect set $BS \leftarrow \{\arg\min_{D_i \in CSet} \rho_{D_i}(q)\}$
4: **for** $l \leftarrow 2$ *to* $|Cset|$ **do**
5: $\quad RC_l \leftarrow Recombination(C_{l-1}, P)$
6: $\quad MC_l \leftarrow Mutation(RC_l, P, \alpha)$
7: $\quad$ **for** each candidate aspect $\mathcal{S}$ in $RC_l \cup MC_l$ **do**
8: $\quad\quad$ **if** $\mathcal{S}$ has not been considered **then**
9: $\quad\quad\quad$ **if** $|C_l| < P$ **then**
10: $\quad\quad\quad\quad C_l \leftarrow C_l \cup \{\mathcal{S}\}$
11: $\quad\quad\quad$ **else if** $\rho_{\mathcal{S}}(q) < Max(\{\rho_{D_i}(q) | D_i \in C_l\})$ **then**
12: $\quad\quad\quad\quad$ replace the worst aspect in C_l by $\mathcal{S}$
13: $\quad\quad\quad$ **end if**
14: $\quad\quad$ **end if**
15: $\quad$ **end for**
16: $\quad BS \leftarrow BS \cup \{\arg\min_{A_i \in C_l} \rho_{A_i}(q)\}$
17: **end for**
18: $BS \leftarrow Normalization(BS)$
19: **return** $bestAspect \leftarrow \arg\min_{A_i \in BS} \rho_{A_i}(q)$

better in the paternal generation, the dimensions that make up this aspect are more likely to participate in the generation of new individuals. Recombination can discover most of the high-scored aspects by the heuristic search strategy.

Procedure $Mutation(\cdot)$ in Step 5 can handle extreme cases (e.g. Figure 1(b) and Fig. 1(c)). OSIER use sampling to calculate the probability of each dimension participating in the mutation, which is calculated as:

$$pro(D_i) = \frac{W_{D_i}}{\sum_{D_j \in H} W_{D_j}} \tag{10}$$

where H is the set of dimensions not involved in the recombination, W_{D_i} indicates the weight of the dimension D_i on the outlyingness in full space, which is mentioned before. A dimension with a larger weight will have a larger opportunity to be selected for next generation to reproduce with modification. OSIER uses bit-wise mutation which randomly replacing one of the dimensions that make up an individual.

Moreover, in order to ensure a fair comparison between different dimensional aspects, the outlyingness score needs to be normalized after calculation. A well-known normalization method in the OAM task is Z-score [10]:

$$Z(\rho_{\mathcal{S}}(q)) \triangleq \frac{\rho_{\mathcal{S}}(q) - \mu_{\rho_{\mathcal{S}}}}{\sigma_{\rho_{\mathcal{S}}}} \tag{11}$$

Table 2. Characteristics of the datasets

Data set	# objects(n)	# attributes(d)
Synthetic datasets	1000	10–100
Seed	210	7
Music emotion	400	50
Climate model	540	18
KSD	2856	71

where μ_{ρ_S} is the mean of the density of all points in the aspect $\mathcal{S}$, and σ_{ρ_S} is the standard deviation. The score obtained by this transformation satisfies the dimensional unbiasedness requirement of Eq. 1. However, the computational cost of normalizing each searched aspect is still large, so we consider a staged comparison. When generating the aspects in each dimensionality, the optimal aspect is obtained by comparing the original density evaluation score (Steps 6–18). After obtaining the optimal aspects under each dimension, the outliers are normalized (Step 19) and compared (Step 20) in these aspects. Compared to OAMiner [5] and Density Z-score [10], the overall search complexity is reduced from $O(n^2 d \cdot Wd)$ to $O(nd \cdot Wd + n^2 d \cdot d)$, where n and d are the size and dimensionality of data set, W is the average search width of each dimension.

4 Experiments and Result Discussion

4.1 Experimental Setting

Datasets. We use four real-world datasets from the UCI machine learning repository[1]. We also use the synthetic datasets provided by Keller *et al.* [7]. Table 2 shows the characteristics of these datasets. The attribute values in the dataset are all real numbers. For the sake of the subsequent description, we use incremental subscripts to record the attributes of the dataset.

Baselines. Four OAM methods are selected as baselines to demonstrate the efficiency and stability of OSIER, including kernel density rank (OAMiner) [5], Z-score normalized Kernel density (ZKDE) [10], sGrid [15] and SINNE [12]. We have made a brief summary in Table 3 for the outlyingness calculation used by each method, where ψ denotes the size of the random subsample, t denotes the number of ensemble models, and *Individual Complexity* represents the complexity of computing outlyingness in one aspect for a query point. There are three additional notes: (1) Although OAMiner introduces a boundary method to perform pruning operations, the amount of search space is uncertain and usually much larger than other methods. The exact number depends on the true distribution of the data. (2) SINNE is a sampling-based method. Given a query

[1] http://archive.ics.uci.edu/ml/index.php.

point and an aspect, the results will vary each time. In particular, the use of a heuristic search approach can also lead to a volatile search of the aspects. (3) The stability of OSIER depends on the mutation rate, which is usually small. It works by accepting the probability of generating an aspect that is worse than the current one, so it is possible to jump out of the local optimal solution. If there is no extreme case, the searched aspects are stable for the same dataset.

Table 3. Summary the characteristics of baselines

Methods	Individual complexity	Complexity	Interpretability	Stability
OAMiner	$O(n^2d)$	–	High	High
ZKDE	$O(n^2d)$	$O(n^2d \cdot Wd)$	High	High
Sgrid	$O(n^2d/\omega)$	$O(n^2d/\omega \cdot Wd)$	Medium	High
SINNE	$O(\psi t^2 d)$	$O(\psi t^2 d \cdot Wd)$	Low	Low
OSIER	$O(nd)$	$O(nd \cdot Wd + nd^2)$	High	Medium

Experimental Setup. We use the parameters mentioned in Sect. 3.1. For the aforementioned methods, we apply the default parameters. For the scoring function, OAMiner and ZKDE use the Gaussian kernel. We set ψ to 8 and t to 100 in SINNE. For the search strategy, epsilon neighborhood range of OAMiner is set to 1. Search width W and maximum dimensionality of searched aspects d_{max} in beam search are set to 120 and 3, respectively.

Experiments were conducted on a PC with four Intel Xeon E5-2698 CPUs, four GeForce RTX 2080 Ti GPUs and 512 GB memory, running Ubuntu 20.04. The algorithms were implemented in Java and compiled by Java version 13.

4.2 Effectiveness

We use four real-world datasets from different domains to demonstrate the effectiveness of OSIER in the OAM problem. Since we do not have a standard to measure the quality of the found aspects, we display the visualization of outlier points in the aspects as shown in Fig. 3. We choose query points with aspects of three dimensions as examples for better visualisation. The red triangles indicate the query points which exhibit good outlyingness in each obtained subspaces. Visualization results show that OSIER can obtain well-behaved aspects.

We also compare our method with the baseline methods on 10, 20, 30, 40, 50, 75, and 100-dimensional synthetic datasets, respectively. The original datasets provide with 19–136 outliers and the ground-truth of their outlying aspects. For the fairness of the experimental criteria, we augmented the number of outliers. For example, there is an outlier X_1 and a normal point X_2 in a d-dimensional data set, and the outlying aspect of X_1 is $\{D_1, D_2\}$. By replacing the dimensions which are not in outlying aspect of X_1, we can generate a new outlier $X_3 = \{X_1.D_1, X_1.D_2, X_2.D_3, ...X_2.D_d\}$. Since OAMiner is committed to finding aspects ranked 1, it may find more than one aspect. We consider that the correct aspect has been found if the result contains the ground truth.

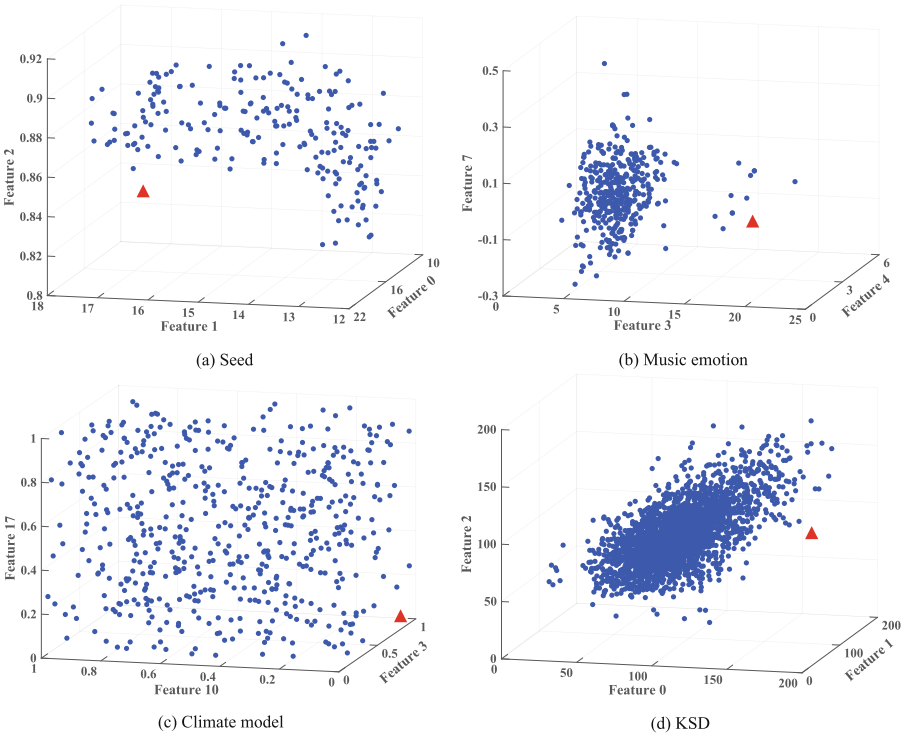

Fig. 3. Visualization results on four real data sets.

The results are shown in Table 4. Experiment shows that OSIER achieves the state-of-the-art performance on all datasets of different dimensions. There is little difference in the effectiveness of each method in the low-dimensional space. As the number of dimensions rises, the accuracy of OSIER improves more significantly. The trend shows that the search strategy plays a greater role in high-dimensional space. Besides, the heuristic rule pruning strategy (OSIER and ZKDE) can search for more representative aspects than the boundary pruning (OAMiner). ZKDE performs better than Sgrid and SINNE, which indicates that using partial data makes the searched aspects unstable.

Figure 4 shows the efficiency test on the synthetic datasets with varying number of dimensions d and data size n. The base OAMiner method [5] is chosen as the baseline method, which can reduce the impact caused by the different functions of calculating the outlyingness. We select several query points with a result subspace of no more than 3 dimensions to calculate the average running time. Experiment shows that the search efficiency of our method is much faster. Moreover, the execution time rises slower than the baseline as the dimensionality and size of the dataset expand.

Table 4. Overall Performances of Comparison with Baselines

Method	*syn_10D*	*syn_20D*	*syn_30D*	*syn_40D*	*syn_50D*	*syn_75D*	*syn_100D*
OAMiner	0.893	0.664	0.463	0.352	0.320	0.308	0.192
ZKDE	0.953	0.808	0.839	0.631	0.601	0.635	0.573
Sgrid	0.942	0.629	0.574	0.556	0.524	0.508	0.426
SINNE	0.879	0.709	0.558	0.640	0.609	0.648	0.595
OSIER	**0.967**	**0.856**	**0.843**	**0.770**	**0.694**	**0.671**	**0.654**

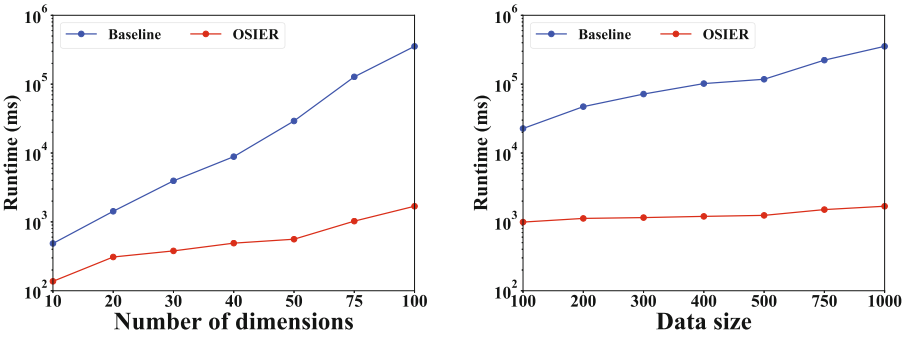

Fig. 4. Efficiency test w.r.t the number of dimensions d and data size n.

4.3 Parameter Analysis

For ease of understanding, we select the data under two dimensions of the synthetic datasets and visualize the results by contour lines.

The choice of norm comes in to play when $d \geq 2$. In the previous norm studies, the commonly used norms are $p = 1$ (Manhattan distance), $p = 2$ (Euclidean norm), and $p = \infty$ (Maximum norm). As shown in Fig. 5(a), the value of p has a tiny effect on the density distribution in a dense region. As the number of data points increases, the choice of p is less important. We recommend using the 2-norm for its stronger symmetry.

Figure 5(b) shows that the value of bandwidth cannot be static. A reasonable bandwidth should depend on the distribution of the data. A small bandwidth will result in relatively independent density estimates (e.g. $N = 100$, $h = 0.1$), and a large bandwidth will result in a dispersed density distribution, which makes it difficult to reflect the differences. Silverman's rule of thumb [14] is a variance-based bandwidth selection method. By this method, the value of h is taken closer to 0.25 when $N = 100$, and it can be observed that $h = 0.25$ is more representative compared with the others in visualization results. bandwidth essentially scales the kernel density estimation function in different dimensions to obtain better experimental results.

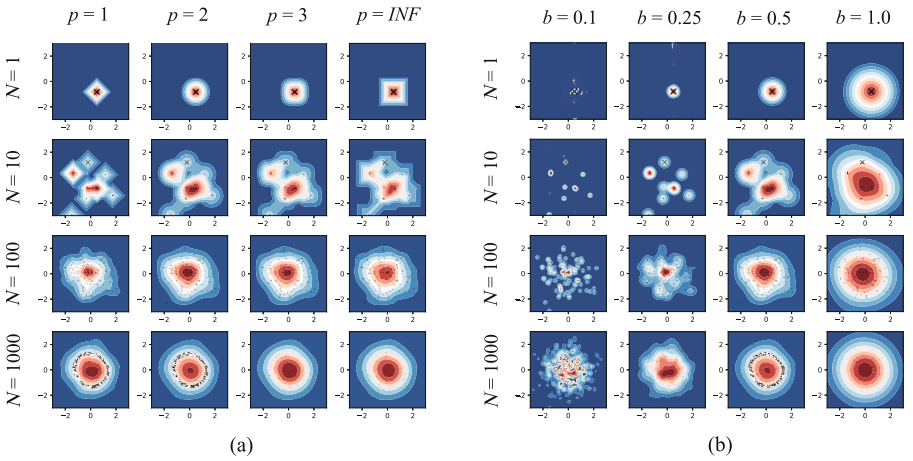

Fig. 5. Influence of different norm p and bandwidth b.

5 Discussion and Conclusion

In this paper, we study the outlying aspect mining problem and propose OSIER, which address the shortcomings of existing methods effectively and provide more interpretable and credible results. We analyze the application of kernel density estimation methods to outlying aspect mining and design an adaptive scoring function. In addition, we improve the commonly used aspects search strategy. We introduce the idea of a genetic algorithm to obtain the fitness of individuals in the process of genetic inheritance by analyzing the correlation among dimensions. Also, the mutation operation in the genetic algorithm can handle some extreme cases during the search process, thus avoiding getting trapped in a local optimum. Experimental results on several real and synthetic datasets demonstrate the effectiveness of the proposed method in outlying aspect mining.

Our future work will focus on applying OAM to hybrid or time-series data. We plan to design a rational outlying aspect structure for interpretable results.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (61972268), the Sichuan Science and Technology Program (2020YFG0034), and the Med-X Center for Informatics funding project of SCU (YGJC001).

References

1. Beulah, J.R., Punithavathani, D.S.: An efficient mixed attribute outlier detection method for identifying network intrusions. Int. J. Inf. Secur. Priv. **14**(3), 115–133 (2020)
2. Botev, Z.I., Grotowski, J.F., Kroese, D.P.: Kernel density estimation via diffusion. Ann. Stat. **38**(5), 2916–2957 (2010)

3. Carvalho, E.D., Silva, R.R.V., Araújo, F.H.D., de A. L. Rabelo, R., de Carvalho Filho, A.O.: An approach to the classification of COVID-19 based on CT scans using convolutional features and genetic algorithms. Comput. Biol. Med. **136**, 104744 (2021)
4. Dang, X., Assent, I., Ng, R.T., Zimek, A., Schubert, E.: Discriminative features for identifying and interpreting outliers. In: ICDE, pp. 88–99 (2014)
5. Duan, L., Tang, G., Pei, J., Bailey, J., Campbell, A., Tang, C.: Mining outlying aspects on numeric data. Data Min. Knowl. Disc. **29**(5), 1116–1151 (2015). https://doi.org/10.1007/s10618-014-0398-2
6. Jenkinson, W.G., Li, Y.I., Basu, S., Cousin, M.A., Oliver, G.R., Klee, E.W.: Leafcuttermd: an algorithm for outlier splicing detection in rare diseases. Bioinform. **36**(17), 4609–4615 (2020)
7. Keller, F., Müller, E., Böhm, K.: HICS: high contrast subspaces for density-based outlier ranking. In: ICDE, pp. 1037–1048 (2012)
8. Lappas, P.Z., Yannacopoulos, A.N.: A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Appl. Soft Comput. **107**, 107391 (2021)
9. Micenková, B., Ng, R.T., Dang, X., Assent, I.: Explaining outliers by subspace separability. In: ICDM, pp. 518–527 (2013)
10. Vinh, N.X., et al.: Discovering outlying aspects in large datasets. Data Min. Knowl. Disc. **30**(6), 1520–1555 (2016). https://doi.org/10.1007/s10618-016-0453-2
11. do Prado Ribeiro, K., Fontes, C.H., de Melo, G.J.A.: Genetic algorithm-based fuzzy clustering applied to multivariate time series. Evol. Intell. **14**(4), 1547–1563 (2021)
12. Samariya, D., Aryal, S., Ting, K.M., Ma, J.: A new effective and efficient measure for outlying aspect mining. In: WISE, pp. 463–474 (2020)
13. Samariya, D., Ma, J.: Mining outlying aspects on healthcare data. In: HIS, pp. 160–170 (2021)
14. Silverman, B.W.: Density estimation for statistics and data analysis (1986)
15. Wells, J.R., Ting, K.M.: A new simple and efficient density estimator that enables fast systematic search. Pattern Recognit. Lett. **122**, 92–98 (2019)
16. Zhang, J., Gao, Q., Wang, H.H.: A novel method for detecting outlying subspaces in high-dimensional databases using genetic algorithm. In: ICDM, pp. 731–740 (2006)
17. Zhang, J., Lou, M., Ling, T.W., Wang, H.H.: HOS-Miner: a system for detecting outlying subspaces of high-dimensional data. In: VLDB, pp. 1265–1268 (2004)
18. Zhu, C., Kitagawa, H., Faloutsos, C.: Example-based robust outlier detection in high dimensional datasets. In: ICDM, pp. 829–832 (2005)
19. Zrira, N., Mekouar, S., Bouyakhf, E.: A novel approach for graph-based global outlier detection in social networks. Int. J. Secur. Networks **13**(2), 108–128 (2018)