

INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING
2019, ICRTAC 2019

Characterizing the Outlying Feature Set of Groups

Haiyang Xia^{a,b}, Huy Quan Vu^c, Jianlong Tan^d, Xiao Li^e, Gang Li^{e,c}^a Xi'an Shiyou University, Shaanxi 710065, China^b Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, China^c Deakin University, Geelong, VIC 3216, Australia^d Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China^e Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences

Abstract

Outlying feature set of groups is useful in many applications scenarios. However, most of existing literatures focused on characterizing outlying feature set of individuals rather than group level. A method that can identify outlying feature set of groups effectively from large scale dataset is not yet available. This paper aims to tackle this challenge by proposing a novel group outlying feature set identification algorithm, named GOFSI, which can identify the outlying feature set at the group level

automatically. The Experiments on both synthetic and real-world data sets confirmed the effectiveness of our method.

© 2019 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the scientific committee of the INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019.

Keywords: group outlying feature set detection, distribution similarity, earth mover's distance;

1. Main text

Outlying feature set characterization is the process of identifying feature set, on which the query data is the most distinct from the rest of the data under consideration, it has many real-world applications. For instances, a car buyer may be focus on the features that make a specific car distinct from others to support his purchase decision making. A manager may be interested in identifying factors, which make his brand different from others [11].

A major drawback in the existing technology is that their algorithms were designed to identify feature set that makes a particular object most distinct from others, but not for a group of objects. There many real-world applications where the practitioners or decision-makers are interested in outlying features of a specific group, such as

a group of web service [12] or hotels[13]. [7] formulated this task as *Group Outlying Aspects Mining* (GOAM), and proposed an algorithm with the same name to identify these outlying features for groups. However, the generated result of the GOAM algorithm contains two kind of outlying subspaces. Judgment from users is required to determine which feature(s) should be included in the final outlying feature set. This requirement is unsuitable for outlying feature set characterization that involves a large number of features. Besides, the efficiency of GOAM is largely affected by the settings of the user-specified threshold. An inappropriate threshold may decrease the effectiveness of this algorithm.

This paper aims to address the limitations of GOAM by proposing a new technique, called Group Outlying Feature Set Identification (GOFSI), which can identify the outlying feature set at the group level automatically. GOFSI has a major advantage over GOAM, as users are not required to specify thresholds, while the determination of the outlying feature set of a specific group are carried out automatically. Such improvement makes GOFSI applicable for the mining outlying feature set of groups that involve a large number of features.

2. Related work

Methods used in the outlying feature set characterizing problem can be categorized into feature selection-based techniques [1] and score-and-search based techniques [2]. Feature selection-based techniques solve outlying feature set characterization problem by treating the selection of feature set as a binary classification problem, the query data (positive class) versus rest of the data (negative class). One critical challenge in feature selection-based approaches is to balance the data in these two classes because the sample number in positive class is commonly smaller than that in the negative class. [1] tackled this issue by over-sampled the positive class. [5] used the samples extract from a Gaussian distribution centered at the target point to balance the data between these two classes. The feature set with the best classification performance will be reported as the final outlying feature set to interpret the distinctiveness of the query data. One main drawback of feature selection-based approaches are considered as the extension of the positive synthetic distributions [5] because the positive distribution always has the same variance in every dimension and is commonly determined by the distance of query data and its k th nearest neighbor in full feature set. This method expected to affect all feature set equally, but actually, some feature set may affected by this setting more than others. [6].

Score-and-search based approaches are another set of common techniques for the outlying feature set characterization [7]. Along this line of research, a function that can evaluate the outlying degree of the query object in each candidate feature set is required. The feature set with the highest outlying degree will be reported as the final output [8]. One pioneering work in the score-and-search based approach is from [10], in which they proposed an algorithm named HOS-Miner to evaluate the outlying degree of the target point. [4] proposed a kernel density based scoring function for the large-scale data processing task. Although these methods have praiseworthiness performance in solving the outlying feature set characterizing problems, all of these methods are not suitable for measuring the outlying degree of groups. To this end, [7] proposed an Earth Moving Distance based scoring function, and confirmed its effectiveness in measuring the outlying degree of group data. However, the GOAM algorithm proposed in this paper requires the user to determine which features should be included in the later iteration. This requirement is unsuitable for the outlying feature set characterizing in a large number of features.

3. GOFSI Algorithm

3.1. GOFSI Problem Statement

This section presents and formally defines the problem of GOFSI, which help guide the subsequent algorithm development.

Definition 1 (GOFSI Problem). Let $G = \{G_0, G_1, G_2, \dots, G_n\}$ denote n groups, in which G_0 represents the query group, and $\{G_1, G_2, \dots, G_n\}$ are the contrast groups. Individuals in each group have d features $F = \{f_1, f_2, \dots, f_d\}$. Let $p(s)$ denote the scoring function that can measure the outlying degree of each candidate feature set. Then the

problem of GOFSI is to identify a feature set $s \subseteq F$ on which the query group G_q is most distinct with other groups, namely, with the highest outlying degree $\rho(s)$.

Definition 2 (Single Feature Set). Let $F = \{f_1, f_2, \dots, f_d\}$ represent d features of each member in every group F . Then the single feature set denotes feature set which contains only a single feature f_i , namely, $\{f_i\}, \{f_2\}, \dots, \{f_n\}$.

Definition 3 (Union Feature Set). The union feature set refers to the feature set, which contains multiple features f_i , such as $\{f_1, f_2\}, \{f_1, f_3\}, \{f_2, f_3\}, \{f_1, f_2, f_3\}$, etc.

3.2. Group Feature Representation

This stage aims to generate a representation that can best reflect the feature quality at the group level. One convenient approach is to use aggregation functions [3], such as *mean*, *median* or *mode*, which output a representative value for each feature. However, the use of a single value sacrifices the original feature distribution of the data, and is incapable of capturing the quality for groups of features. As such, we adopt an alternative approach based on probability distribution to accommodate this task. For simplicity, we use discrete probability distribution in this paper, however, the probability distribution for continuous variable can also be used. Let $A = \{o_1, o_2, \dots, o_n\}$ denotes a group of n objects with m features $\{f_1, f_2, \dots, f_m\}$. The values of all features are in the scale $[l, h]$. Suppose there are m discrete values in the scale $[l, h]$. For each feature f_i , we compute the frequency of discrete values and normalize them to represent probability distribution, $\{P^{f_i} = p_1^{f_i}, p_2^{f_i}, \dots, p_m^{f_i}\}$, the values of the elements $p_j \in P^{f_i}$ satisfy $\sum_{i=1}^m p_i^{f_i} = 1$.

3.3. Outlying Degree Computation

Once the feature values of group members are represented as a discrete probability distribution, *Earth Mover's Distance (EMD)* [7] is adopted to evaluate the outlying degree between groups on each candidate feature set. The EMD of two discrete probability distributions P_1 and P_2 can be measured by Eq. (1):

$$\text{EMD}(P_1, P_2) = \frac{\sum_{i=1}^m \sum_{j=1}^m d_{i,j} c_{i,j}}{\sum_{i=1}^m \sum_{j=1}^m c_{i,j}} \quad (1)$$

Where $d_{i,j}$ is the ground distance between the element i and j of probability distribution P_1 and P_2 respectively, which can be calculated by Euclidean distance, Manhattan distance or other distance. This work adopts the Euclidean distance due to its efficiency. $c_{i,j}$ represents the flow between the elements i and j of P_1 and P_2 respectively, which can be determined by solving a linear programming problem as Eq. (2):

$$\min \sum_{i=1}^m \sum_{j=1}^m c_{i,j} d_{i,j} \quad \left\{ \begin{array}{l} c_{i,j} \geq 0, 1 \leq i, j \leq m, \\ \sum_{i=1}^m c_{i,j} \leq p_i, 1 \leq i \leq m, \\ \sum_{j=1}^m c_{i,j} \leq q_j, 1 \leq j \leq m, \\ \sum_{i=1}^m \sum_{j=1}^m c_{i,j} = \min \left\{ \sum_{i=1}^m p_i, \sum_{j=1}^m q_j \right\} \end{array} \right. \quad (2)$$

Suppose a group A has a s number of compared groups B , whose probability distribution on feature f_i is $\{P_A^{f_i}, P_B^{f_i}, \dots, P_s^{f_i}\}$. The outlying degree of these two group on the feature f_i can be measured by Eq.(3):

$$Diff(A, B)^{f_i} = \sum_{i=1}^s EMD(P_A^{f_i}, P_B^{f_i}) \quad (3)$$

Note that the above description focuses on the single feature set. To evaluate the outlying degree for union feature sets, the values of union feature sets for groups should be transformed to joint distributions of individual features. Suppose $D(D \in F; |D| \leq s)$ is a union feature set, then its joint probability distribution of individual features can be represented as $\{P^D = P_1^D, \dots, P_m^D\}$ and satisfies the constrain $\{\sum_{i=1}^m p_i^D = 1\}$. The outlying degree between union feature sets can be measured with joint probability distribution following Eq. (1).

3.4. Outlying Feature Set Identification

We noted that the outlying degree tends to increase along with the dimension of feature set because adding more feature(s) to feature set would amplify the distinctiveness. However, in the real-world applications, feature sets with have large variances in individual features are not desirable for interpretation purpose. In some cases, the identified feature set with the highest outlying degree can contain both high and low features, such feature set are not useful to support decision making in practices. Therefore, we propose an approach to identify valuable outlying feature set, by taking the inner difference between individual features in the union feature set S into consideration. The inner difference between individual features in feature set D concerning group A is computed as Eq. (4):

$$Inner(A)^D = \sum_{i=1}^{|D|} \sum_{j=2}^{|D|} EMD(P^{f_i}, P^{f_j}), (i \neq j) \quad (4)$$

The adjusted outlying degree of group A and group B on a feature set D is computed as Eq. (5):

$$AdjDiff(A, B)^D = Diff(A, B)^D - Inner(A, B)^D \quad (5)$$

Table 1. Synthetic Dataset

	Group 1				Group 2				Group 3				Group4				Group 5			
	f_1	f_2	f_3	f_4	f_1	f_2	f_3	f_4	f_1	f_2	f_3	f_4	f_1	f_2	f_3	f_4	f_1	f_2	f_3	f_4
Member 1	4	5	2	2	4	1	2	3	4	4	5	3	4	4	3	5	4	4	1	2
Member 2	2	4	4	4	2	1	4	5	2	1	5	3	2	2	3	5	2	3	1	2
Member 3	1	5	3	3	1	1	3	4	1	3	5	4	1	2	4	5	1	3	1	2
Member 4	3	4	1	3	3	1	4	4	3	3	5	2	3	3	2	5	3	3	2	2
Member 5	4	5	3	3	4	2	3	5	4	1	5	3	4	1	3	5	4	2	2	1
Member 6	4	5	4	3	4	1	4	2	4	2	4	3	4	2	3	4	4	2	2	1
Member 7	5	5	3	3	5	1	3	4	5	2	4	3	5	2	3	4	5	3	2	1
Member 8	3	5	3	3	3	2	2	2	3	2	4	3	3	2	3	4	3	2	2	2
Member 9	2	5	3	3	2	1	3	3	2	2	4	3	2	2	3	4	2	2	2	1
Member 10	1	4	3	3	1	1	3	3	1	2	4	3	1	2	3	4	1	2	2	1

4. Experiment and Analysis

4.1. Evaluation on Synthetic Data Set

We first experiment with a synthetic dataset to examine the performance of the proposed algorithm. Our data set contains 5 groups, each with 10 members. The details of this data set are shown in Table 1. Intuitively, the most distinctive feature of *Group 1* is f_2 , because members of *Group 1* tends to have higher values on this feature than all other groups. Similarly, the most distinctive feature of *Group 3* and *Group 4* are f_3 and f_4 respectively. Besides, f_2 is also the most distinct features of *Group 2* is, because it tends to have lower values on this feature than most of the other groups. The most distinctive features of *Group 5* is the union feature set $\{f_2, f_4\}$, because it has lower values

Table 2. Outlying degree of each group on every feature set.

Feature Set	Group 1 vs. C	Group 2 vs. C	Group 3 vs. C	Group 4 vs. C	Group 5 vs. C
$\{f_1\}$	0.00	0.00	0.00	0.00	0.00
$\{f_2\}$	<u>10.65</u>	<u>7.12</u>	4.26	4.24	4.48
$\{f_3\}$	3.90	3.88	<u>8.34</u>	4.38	6.34
$\{f_4\}$	3.77	4.42	3.77	<u>7.02</u>	8.19
$\{f_1, f_2\}$	8.76	5.22	3.41	3.18	3.55
$\{f_1, f_3\}$	3.44	3.28	6.56	3.63	5.14
$\{f_1, f_4\}$	3.14	3.76	3.14	5.36	6.61
$\{f_2, f_3\}$	9.86	6.65	7.70	6.75	7.33
$\{f_2, f_4\}$	9.99	6.51	6.06	6.74	8.56
$\{f_3, f_4\}$	5.93	5.94	7.81	6.74	<u>10.64</u>
$\{f_1, f_2, f_3\}$	7.58	4.57	5.38	5.16	5.59
$\{f_1, f_2, f_4\}$	7.33	4.45	4.57	4.40	6.46
$\{f_1, f_3, f_4\}$	4.75	5.21	5.42	4.45	8.27
$\{f_2, f_3, f_4\}$	8.94	5.76	6.46	6.23	10.03
$\{f_1, f_2, f_3, f_4\}$	5.69	3.12	3.26	2.97	6.83

on both of these features. Then we applied the GOFSI algorithm to this data set, and the adjusted outlying degrees of all feature sets are computed, as shown in Table 2. The optimal feature set concerning each targeted group is highlighted by underline, which is consistent with the above reasoning.

We also apply the algorithm, proposed in [7], to the data set for comparison. Table 3 shows the identified feature set of GOAM for each group, the GOAM generated some individual feature sets and union feature sets for user's manual inspection. Although, GOAM correctly identified the outlying feature set of g_1, g_2, g_3 and g_4 it failed to identify the outlying feature set of g_5 . The outlying feature set of g_5 should be $\{f_3, f_4\}$ rather than $\{f_2, f_3\}$. This is probably because GOAM removed individual features with high outlying degrees to ensure the distribution similarities of members in the union feature set. This influent the outlying degrees of union feature set, which contains individual features with high outlying degrees.

4.2. Evaluation on Real-World Data Set

We further validate the capability of GOFSI on a real-world data set collected from Booking.com. We collected the rating data of 10 major hotel brands in Singapore in February 2019. Each hotel has

Table 3: The identified outlying feature set of GOAM.

Groups	Individual feature set	Union feature set
Group 1	$\{f_2\}(10.65)$	$\{f_2, f_3\}(14.92)$
Group 2	$\{f_2\}(7.12)$	$\{f_2, f_3\}(14.57)$
Group 3	$\{f_3\}(8.34)$	$\{f_2, f_3\}(14.81)$
Group 4	$\{f_4\}(7.02)$	$\{f_2, f_3\}(15.71)$
Group 5	$\{f_4\}(8.19)$	$\{f_2, f_3\}(18.12)$

rating data on seven features, since we adopted discrete probability distribution for GOFSI, the ratings are rounded the nearest integer.

We use each hotel brand as the query group and the other groups as the compared groups. To ensure the comprehensibility of discovered outlying feature set, we only evaluate the outlying degree up to three features according to settings suggested in [7]. The outlying degrees of all those feature sets were computed, as shown in Figure 1. The horizontal axis denotesthefeature set index, and the vertical axis denotes the adjusted outlying degree

of the corresponding feature set. The red diamond represents those outlying feature set with the highest outlying degree for each hotel brand. From this figure, we can see that the outlying degree of the feature sets tends to increase as the number of features in each set increase. Details about the identified outlying feature set of each group are shown in Table 4.

Table 4: Discovered outlying feature set of each group

Hotel Brand Name	Outlying degree	Outlying Feature Set
Aqueen	11.87	{comfort,facilities,wifi}
Far East Hospitality	13.64	{comfort,location,value}
Fragrance Hotel	14.18	{cleanliness,comfort,wifi}
Holiday InnExpress	17.12	{cleanliness,comfort,wifi}
Hotel 81	15.90	{cleanliness,comfort,facilities}
Millennium Hotels	11.57	{comfort,location,wifi}
Park Hotel Group	12.57	{cleanliness,comfort,wifi}
RedDoorz	14.24	{cleanliness,comfort,facilities}
Worldhotels	16.32	{comfort,location,wifi}
ZEN Rooms	15.42	{cleanliness,comfort,wifi}

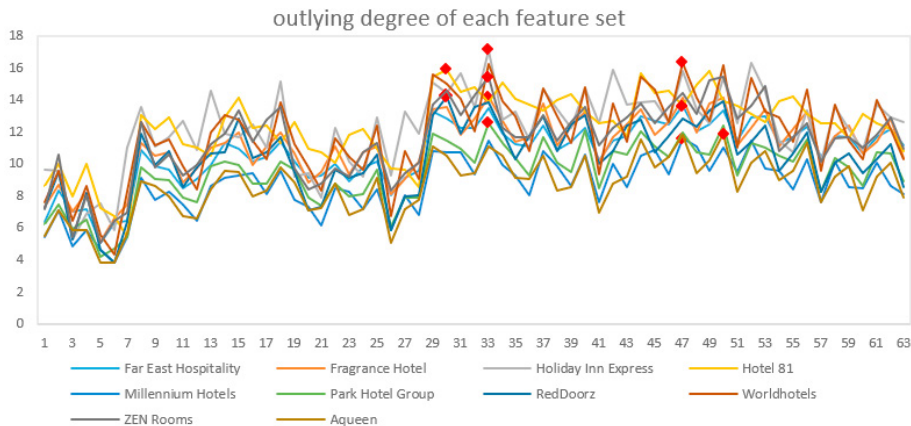


Figure 1: Outlying degrees of each feature Set.

We examined the rating distributions of each hotel brand to determine whether the distinctiveness is a competitive advantage or disadvantage. Taking Aqueen as an example, we found this brand is likely to have high ratings (8, 8, 8) on all three features {comfort, facilities, wifi} than other brands. We can conclude that the competitive advantage of Aqueen hotel brand is to have high ratings on comfort, facilities, and wifi features. As such, the managers of Aqueen can emphasize this distinctiveness in their marketing or promotional materials to attract consumers focus on both comfort, facilities and wifi. For Far East Hospitality brand, {comfort, location, value} are its competitive advantages, because its member hotels are more likely to have high ratings on these features (9, 9, 8 or 8, 9, 8) than other brands. The words of “high value”, “superior location” and “high comfort” can appear on the advertising board of Far East Hospitality to attract potential customers.

On the other hand, the brands Hotel 81 and RedDoorz are distinct from other brands based on the feature sets {cleanliness, comfort, facilities}. Hotels in these brands are likely to have lower rating on these three features (7; 7; 7), which reflect their competitive disadvantages. Hotel managers of Hotel 81 and RedDoorz should develop strategies to improve the user’s satisfaction on such features to eliminate the disadvantages. Similar competitive disadvantages were found for the Fragrance Hotel and ZEN rooms brands, whose member hotels are more likely to have lower ratings on {cleanliness, comfort, wifi} than other brands.

5. Discussion and Conclusions

Many practical applications need to identify a set of features that make a particular group most distinct from others. If practitioners can identify such features, beneficial operation strategies for their business will be easy making. For instance, based on patterns discovered in our experiment, the hotel managers in Singapore can explicitly identify the competitive pros and cons of their hotel brands. Then a target-oriented advertisement strategy can be developed to attract the potential customers. Previous works along this line of research mainly focused on identifying the outlying feature set at the individual level, rather than group level, although [7] formalized this problem, the proposed GOAM algorithm has several limitations which make it inefficient for application with large scale data set with many features under consideration.

Aiming to address this research gap, we proposed a novel outlying feature set characterizing algorithm named GOFSI, which can determinate the outlying feature set of query group automatically. No manual intervention is required during the outlying feature set characterizing procedure, which allow users to characterize the outlying feature set of groups with many features. The experiment results confirmed that GOFSI has better precision than the prior technique (GOAM). GOFSI does not require any threshold settings, which ensure for its robust and stable performance. A shortcoming of GOFSI is that it is currently designed for the discrete probability distribution, where the dimension of probability distribution vector can be high when the scale of feature rating is large. Future works can address this issue by extending GOFSI to work with the continuous probability distribution.

Acknowledgements

This work was supported by Xinjiang research fund for Chinese Academy of Sciences and Guangxi Key Laboratory of Trusted Software (No KX201528).

References

- [1] Dang, X.H., Micenkov'a, B., Assent, I., Ng, R.T., 2013. Local Outlier Detection with Interpretation. Springer Berlin Heidelberg.
- [2] Duan, L., Tang, G., Pei, J., Campbell, A., Campbell, A., Tang, C., 2015. Mining outlying aspects on numeric data. *Data Mining and Knowledge Discovery* 29, 1116–1151.
- [3] Han, J., Xia, H., 2017. A weighted non-monotonic averaging image reduction algorithm, in: *International Conference on Knowledge Science, Engineering and Management*, Springer. pp. 458–465.
- [4] Li, Q., Niu, W., Li, G., Cao, Y., Tan, J., Guo, L., 2015. Lingo: Linearized grassmannian optimization for nuclear norm minimization, in: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, ACM. pp. 801–809.
- [5] Vinh, N.X., Chan, J., Bailey, J., 2014. Reconsidering mutual information based feature selection: A statistical significance view, in: *IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*, pp. 1–10.
- [6] Vinh, N.X., Chan, J., Romano, S., Bailey, J., Leckie, C., Ramamohanarao, K., Pei, J., 2016. Discovering outlying aspects in large datasets. *Data Mining & Knowledge Discovery* 30, 1520–1555.
- [7] Wang, S., Xia, H., Li, G., Tan, J., 2018. Group outlying aspects mining, in: Liu, W., Giunchiglia, F., Yang, B. (Eds.), *Knowledge Science, Engineering and Management*, Springer International Publishing, Cham. pp. 200–212.
- [8] Xia, H., Vu, H.Q., Lan, Q., Law, R., Li, G., 2019. Identifying hotel competitiveness based on hotel feature ratings. *Journal of Hospitality Marketing & Management* 28, 81–100.
- [9] Xu, H., Wang, Y., Cheng, L., Wang, Y., Ma, X., 2018. Exploring a high-quality outlying feature value set for noise-resilient outlier detection in categorical data, in: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ACM. pp. 17–26.
- [10] Zhang, J., Lou, M., Ling, T.W., Wang, H., 2004. Hos-miner: A system for detecting outlying subspaces of high-dimensional data, in: *Thirtieth International Conference on Very Large Data Bases*, pp. 1265–1268.
- [11] Vu, H. Q., Li, G., Law, R., & Zhang, Y. (2018). Tourist activity analysis by leveraging mobile social media data. *Journal of travel research*, 57(7), 883-898.
- [12] Niu, W., Lei, J., Tong, E., Li, G., Chang, L., Shi, Z., & Ci, S. (2014). Context-aware service ranking in wireless sensor networks. *Journal of network and systems management*, 22(1), 50-74.
- [13] Xia, H., Vu, H. Q., Law, R., & Li, G. (2019). Evaluation of hotel brand competitiveness based on hotel features ratings. *International Journal of Hospitality Management*, 102366.