# New York City Taxi Trip Duration

Siyi Yu[1]

[1] Jilin University, China

## Introduction

Many real world applications call for one important function of identifying the set of features on which the interested object is most distinguished from others. Usually, this object is termed as the query object, and the set of features are referred to as the *subspaces* or *aspects*. Accordingly, this research problem is referred to as *outlying aspects mining*, which is different from *outlier detection*.

**Outlying Aspects Mining** aims to identify a subspace which makes the query object most outlying, rather than verifying whether it is an outlier or not. The task of *Outlying Aspects Mining* is to explain which aspects make the query object most different.

**Outlier Detection** aims to identify all possible outliers in the dataset, without explaining why or how they are different. Hence, the outlying aspects mining is also referred to *outlier interpretation* or *object explanation*.

In this paper, we extend the task of *outlying aspects mining* to the *group* level, formalize the research problem of *group outlying aspects mining*, and propose a novel algorithm named GOAM to solve the *group outlying aspects mining* problem.

## Group Outlying Aspects Mining

- It aims to *identify a subset of aspects (or subspace) which makes the query group, rather than the single object, obviously different*. What we are interested in the task of *group outlying aspects mining* is to explain which aspects make the query group distinctive different from the other groups.
- *Group Outlying Aspects Mining*, *Outlying Aspects Mining* and *Outlier Detection* are different with each other.


Group Outlying Aspects Mining


Outlying Aspects Mining


Outlier Detection

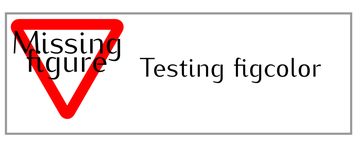
## GOAM Algorithm

We propose the *GOAM* algorithm to solve the research problem of *Group Outlying Aspects Mining*. The *GOAM* algorithm includes three major steps.


**Group Feature Extraction** Let $f_1$, $f_2$, $f_3$ represent three features of $G_q$. We count the frequency of each value for one feature. Then use the histogram to represent each feature. Similarly, we can extract other features for each group.


Histogram of $G_q$ on $f_1$


Histogram of $G_q$ on $f_2$




Histogram of $G_q$ on $f_3$

**Outlying Degree Scoring** In this step, we first calculate the *earth mover distance* (EMD) of one feature among different groups. The earth mover distance reflects the minimum mean distance between groups on one feature. So, we utilize the EMD to measure the difference between groups of each feature.

## GOAM Algorithm

Second, based on the *earth move distance*, we calculate the outlying degree.


where $G_q$ is the query group, $n$ is the number of compare groups, and $h_{k_s}$ is the histogram representation of $G_k$ in the subspace $s$.

**Outlying Aspects Identification** In this step, based on the value of outlying degree we will identify the group outlying aspects. If a feature's outlying degree is greater than a threshold, the more likely the feature is group outlying aspect. When the dimensionality of features is high, we adopt a stage-wise candidate subspace construction strategy to alleviate the exponential explosion.

## Experiment

**Synthetic Dataset** contains 10 groups and 8 features. Each group consists of 10 members, and each member has 8 features.

| Method | Truth Outlying Aspects | Identified Aspects | Accuracy |
|---|---|---|---|
| GOAM | $\{F_1\}$, $\{F_2F_4\}$ | $\{F_1\}$, $\{F_2F_4\}$ | 100% |
| Arithmetic Mean based OAM | $\{F_1\}$, $\{F_2F_4\}$ | $\{F_4\}$, $\{F_2\}$ | 0% |
| Median based OAM | $\{F_1\}$, $\{F_2F_4\}$ | $\{F_2\}$, $\{F_4\}$ | 0% |

It can be observed that the GOAM method can identify the trivial outlying features and non-trivial outlying subspaces correctly and is obvious from the table that the accuracy of GOAM is the best, which is (100%).

**NBA Dataset** was collected from Yahoo Sports website (`http://sports.yahoo.com.cn/nba`). The data include all teams from the six divisions, and each player in the team has 12 features.

| Teams | Trivial Outlying Aspects | NonTrivial Outlying Aspects |
|---|---|---|
| Cleveland Cavaliers | {3FA} | {FGA, FT%}, {FGA, FG%} |
| Orlando Magic | {Stl} | None |
| Milwaukee Bucks | {To}, {FTA} | {FGA, FTA}, {3FA, FTA} |
| New Orleans Pelicans | {FT%}, {FTA} | {FTA, Stl}, {FTA, To} |


New Orleans Pelicans on FT%


New Orleans Pelicans on FTA

`New Orleans Pelicans` has more players with lower {free throw percentage}, {free throws attempted}.

## Conclusion

**Problem Definition** Formalize the problem of Group Outlying Aspects Mining by extending outlying aspects mining.

**GOAM algorithm** Propose GOAM algorithm to solve the *Group Outlying Aspects Mining* problem.

**Strategies** Utilize the pruning strategies to reduce time complexity.

Acknowledgement

- International Cooperation Project (Y7Z0511101) of IIE, Chinese Academy of Sciences