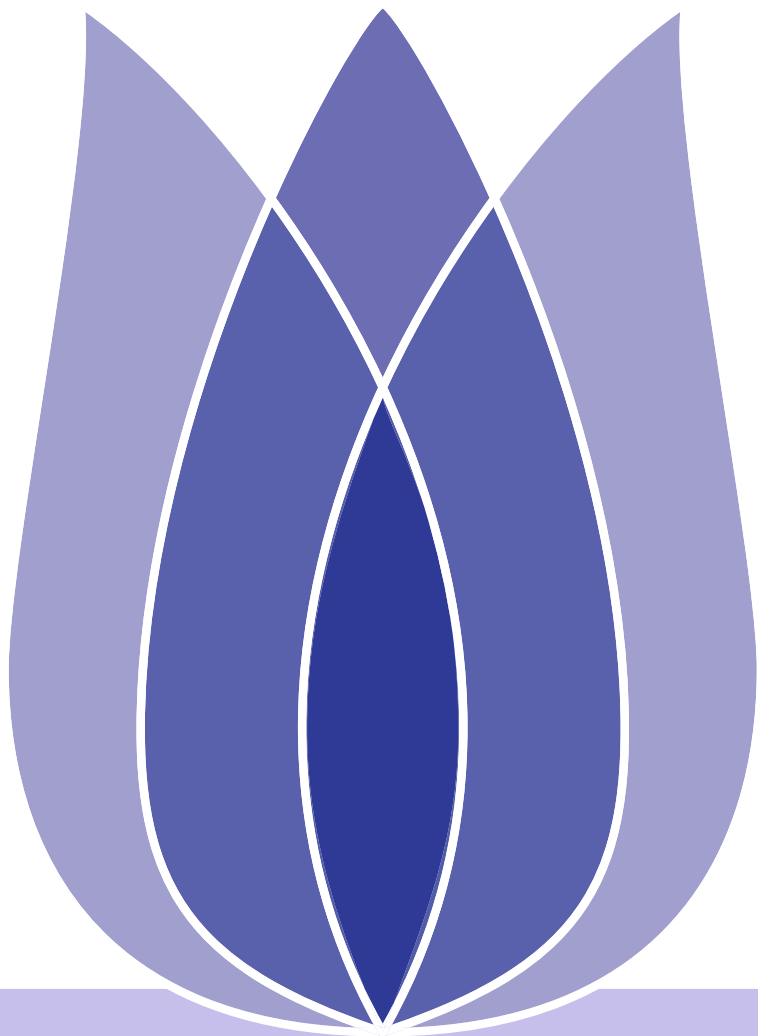


Air Pollution Prediction based on multicollinearity

Shanika Iroshi Nanayakkara
Deakin University

03/06/2022





Overview

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Evaluation Results](#)
- [Conclusion](#)

Problem Definition

- Air pollution predictors
- Group Outlying Aspects Mining

Related Work and Challenges

- Related Work - Outlying Aspects Mining
- Challenges (1)

GOAM Algorithm

- Step One - Group Feature Extraction
- Step Two - Outlying Degree Scoring
- Step Three - Outlying Aspects Identification

Evaluation Results

- Synthetic Dataset
- NBA Dataset

Conclusion



Problem Definition

Air pollution predictors
Group Outlying Aspects Mining

Related Work and Challenges

GOAM Algorithm

Evaluation Results

Conclusion

Problem Definition



- Problem Definition
- Air pollution predictors
- Group Outlying Aspects Mining
- Related Work and Challenges
- GOAM Algorithm
- Evaluation Results
- Conclusion

Defn

Outlying Aspects Mining aims to identify the outstanding features of the query object.

- A teacher may be interested in the **characteristics** that make **one student distinctive** from others.
- NBA coaches would prefer to find out the strengths and weaknesses of the player (a query object).

Player	3PT%	FTA	FT%	To
P_1	65	4	33	8
P_2	78	1	65	5
P_3	58	6	46	3
P_4	68	1.2	85	6.2
P_5	58	6.2	36	3.4



Linear Regression vs Rigid Regrsson

- Problem Definition
- Air pollution predictors
- Group Outlying Aspects Mining
- Related Work and Challenges
- GOAM Algorithm
- Evaluation Results
- Conclusion

Player	3PT%	FTA	FT%	To
P_1	65	4	33	8
P_2	78	1	65	5
P_3	58	6	46	3
P_4	68	1.2	85	6.2
P_5	58	6.2	36	3.4

Linear Regression

- Explain the distinctive aspects of the query object.
- The query object may (or may not) be an outlier.

Rigid Linear Regression

- Find out all unusual objects in the whole dataset.
- No explanation on how they are different.



Group Outlying Aspects Mining

- Problem Definition
- Air pollution predictors
- Group Outlying Aspects Mining
- Related Work and Challenges
- GOAM Algorithm
- Evaluation Results
- Conclusion

Defn

Group outlying aspects mining aims to identify the outstanding features of the group of query object.

- Doctors desire to identify the merits & demerits between a group of cancer patients and normal people.
- NBA coaches are passionate about exploring the obvious advantages & disadvantages of the team.



Figure 1: Medical



Figure 2: NBA-Team

Problem Formalization

Problem Definition
Air pollution predictors
Group Outlying Aspects Mining
Related Work and Challenges
GOAM Algorithm
Evaluation Results
Conclusion

Defn

Group outlying aspects mining aims to identify the top-k group outlying subspace $s \subseteq F$ in which the query group G_q is distinctive with other groups.

- $G = \{G_q, G_2, G_3, \dots, G_n\} \Leftrightarrow$ a set of groups.
- $G_q \Leftrightarrow$ the query group.
- Other groups \Leftrightarrow comparison groups.
- Each object in the group has d features $F = \{f_1, f_2, \dots, f_d\}$.



Term Definition

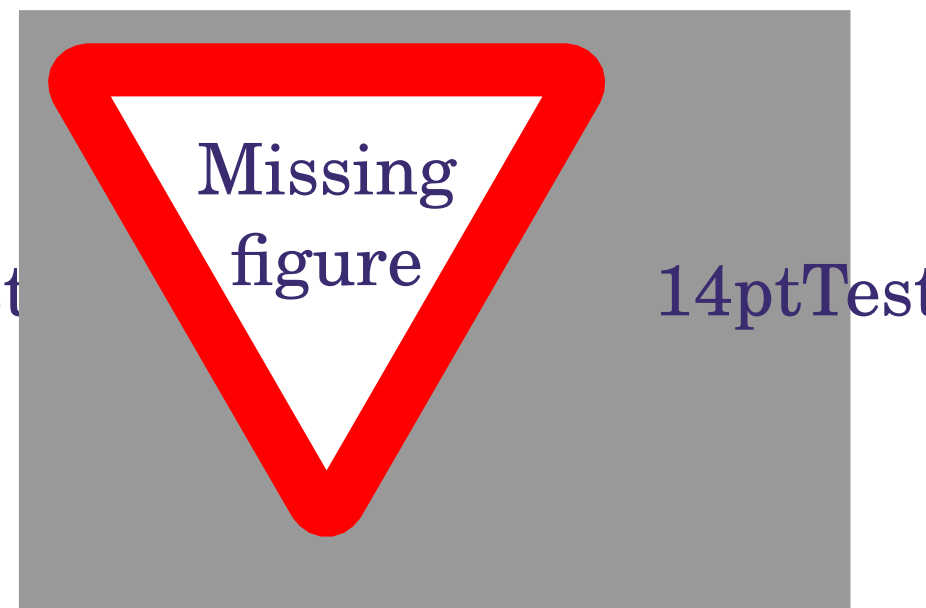
- Problem Definition
- Air pollution predictors
- Group Outlying Aspects Mining
- Related Work and Challenges
- GOAM Algorithm
- Evaluation Results
- Conclusion

■ Top-k group outlying subspaces

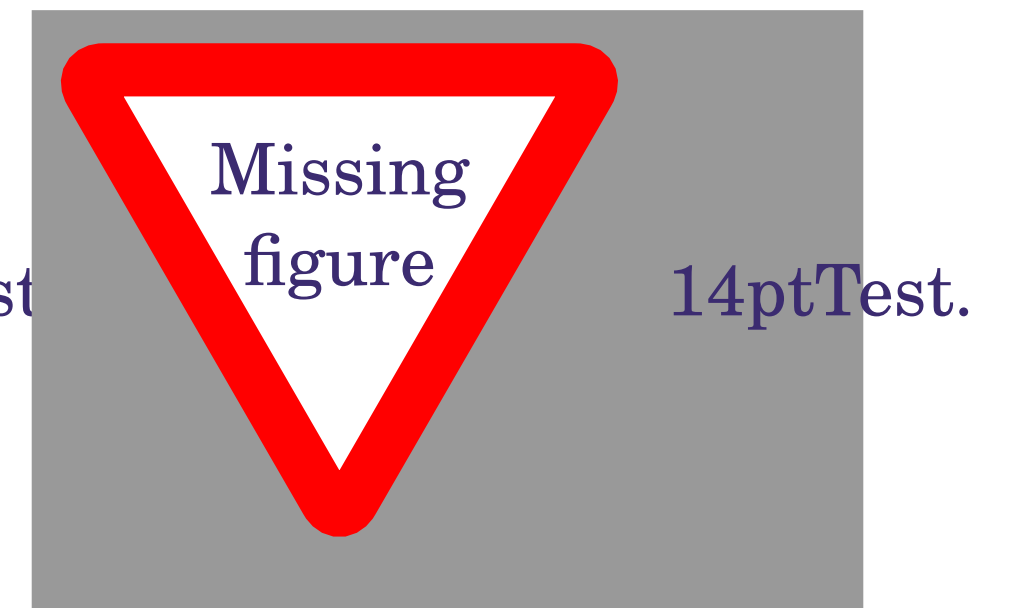
- ◆ $\rho_s(\cdot) \Rightarrow$ outlying scoring function.
- ◆ $\rho_s(\cdot)$ quantifies the outlying degree of the query group G_q in the subspace s .
- ◆ Order by DESC using scoring function $\rho(\cdot)$ to identify top K group outlying subspaces.



(a) Original Feature Spaces



(b) Group Outlying Spaces



(c) Another Subspaces





- Problem Definition
- Air pollution predictors
- Group Outlying Aspects Mining
- Related Work and Challenges
- GOAM Algorithm
- Evaluation Results
- Conclusion

- Trivial Outlying Features
 - ◆ One-dimension subspaces.
 - ◆ G_q 's outlying degree $\rho(\cdot) > \alpha$.

Table 1: $\alpha = 4$

Feature	Outlying Degree
$\{F_1\}$	4.351
$\{F_3, F_4\}$	4.024
$\{F_2, F_4\}$	2.318
$\{F_2\}$	2.002
$\{F_3\}$	1.028



- Problem Definition
- Air pollution predictors
- Group Outlying Aspects Mining
- Related Work and Challenges
- GOAM Algorithm
- Evaluation Results
- Conclusion

- Non-Trivial Outlying Subspaces
 - ◆ Multi-dimension subspaces.
 - ◆ G_q 's outlying degree $\rho(\cdot) > \alpha$.

Table 2: $\alpha = 4$

Feature	Outlying Degree
$\{F_1\}$	4.351
$\{F_3, F_4\}$	4.024
$\{F_2, F_4\}$	2.318
$\{F_2\}$	2.002
$\{F_3\}$	1.028



[Problem Definition](#)

[Related Work and Challenges](#)

[Related Work - Outlying Aspects](#)

[Mining](#)

[Challenges \(1\)](#)

[GOAM Algorithm](#)

[Evaluation Results](#)

[Conclusion](#)

Related Work and Challenges



Related Work - Outlying Aspects Mining

[Problem Definition](#)

[Related Work and Challenges](#)

[Related Work - Outlying Aspects Mining](#)

[Challenges \(1\)](#)

[GOAM Algorithm](#)

[Evaluation Results](#)

[Conclusion](#)

■ Existing Methods - Feature selection

- ◆ To distinguish two classes: the query point (positive) & rest of data (negative)

Disadvantages

- ◆ Positive and negative classes are **Not** balanced.
- ◆ **Not** quantify the outlying degree accurately.
- ◆ **Not** identify group outlying aspects.

Advantages

- ◆ Easy to operate.
- ◆ Resolve dimensionality bias.



TULIP

Team for Universal Learning and Intelligent Processing



Related Work - Outlying Aspects Mining

Problem Definition

Related Work and Challenges

Related Work - Outlying Aspects Mining

Challenges (1)

GOAM Algorithm

Evaluation Results

Conclusion

■ Existing Methods - Score-and-search

- ◆ Define an outlying score function.
- ◆ Search subspaces.

Disadvantages

- ◆ Dimensionality bias.
- ◆ Search efficiency is **Not** high (dataset is large).
- ◆ **Not** identify group outlying aspects.

Advantages

- ◆ Quantify the outlying degree correctly.
- ◆ High Comprehensibility.



TULIP

Team for Universal Learning and Intelligent Processing



- Problem Definition
- Related Work and Challenges
- Related Work - Outlying Aspects Mining
- Challenges (1)
- GOAM Algorithm
- Evaluation Results
- Conclusion

Group Outlying Aspects Mining

- Focus on differences between **groups**.
- **Multiple** points.

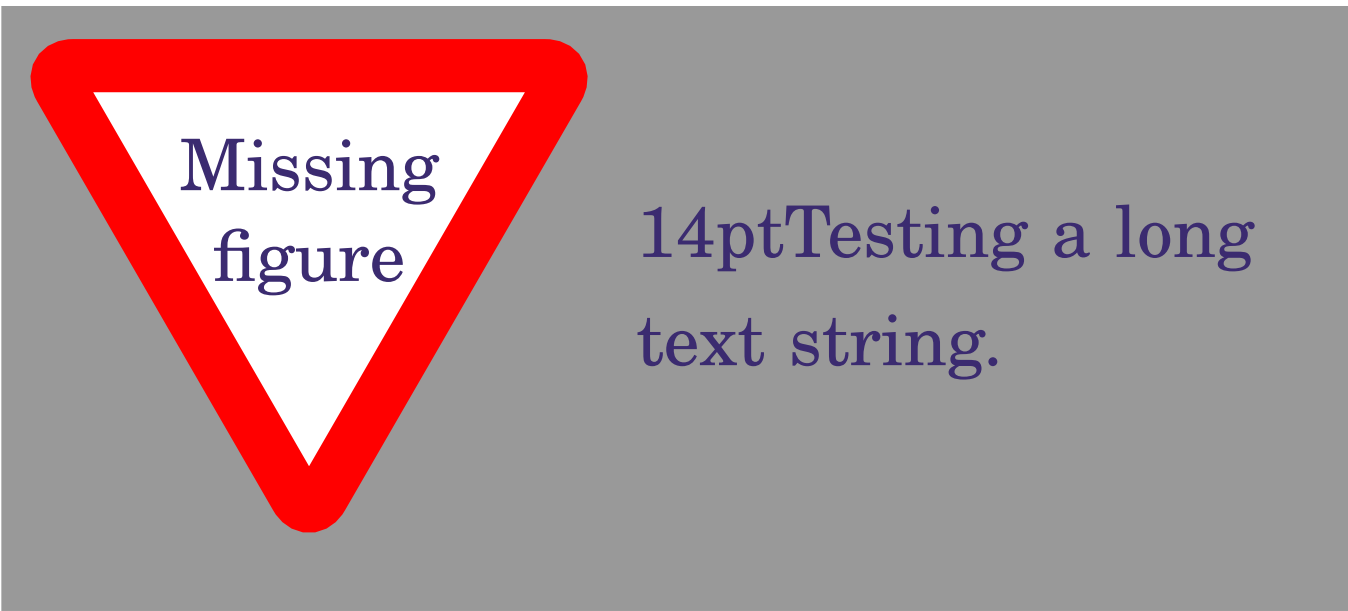


Figure 3: Group Outlying Aspects Target

Outlying Aspects Mining

- Concentrates on differences between **objects**.
- **One** point.

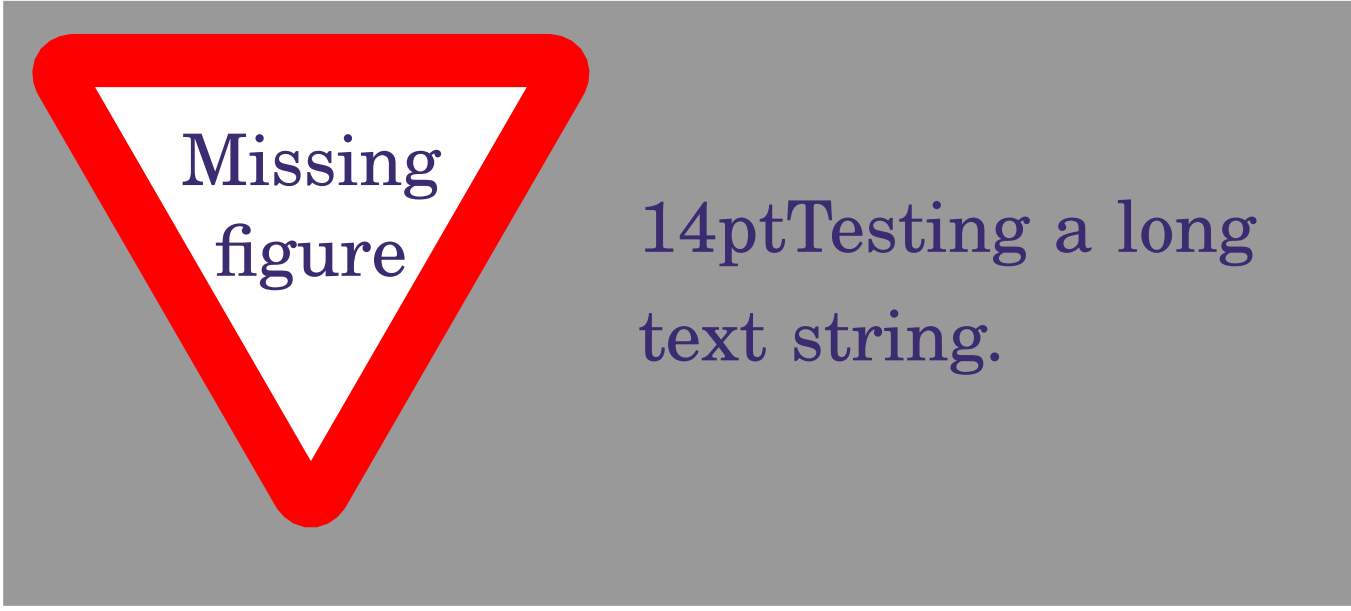


Figure 4: Outlying Aspects Target



Challenges (1)

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [Related Work - Outlying Aspects](#)
- [Mining](#)
- [Challenges \(1\)](#)**
- [GOAM Algorithm](#)
- [Evaluation Results](#)
- [Conclusion](#)

- How to **represent** the group features.
 - ◆ Can be affected by outlier values.
 - ◆ Can **Not** reflect the overall distribution of group features.



Challenges (2)

[Problem Definition](#)

[Related Work and Challenges](#)

[Related Work - Outlying Aspects](#)

[Mining](#)

[Challenges \(1\)](#)

[GOAM Algorithm](#)

[Evaluation Results](#)

[Conclusion](#)

- How to **evaluate** the outlying degree in different aspects.
 - ◆ Need design a scoring function when necessary.
 - ◆ Adopting an appropriate scoring function (without dimension bias) remains a problem.



TULIP

Team for Universal Learning and Intelligent Processing



Challenges (3)

[Problem Definition](#)

[Related Work and Challenges](#)

[Related Work - Outlying Aspects](#)

[Mining](#)

[Challenges \(1\)](#)

[GOAM Algorithm](#)

[Evaluation Results](#)

[Conclusion](#)

- How to **improve** the efficiency.
 - ◆ When the dimension of the **data is high**, the candidate subspace grows exponentially.
 - ◆ It will easily go beyond the limits of the computation resources.



TULIP

Team for Universal Learning and Intelligent Processing



[Problem Definition](#)

[Related Work and Challenges](#)

[GOAM Algorithm](#)

[Step One - Group Feature Extraction](#)

[Step Two - Outlying Degree Scoring](#)

[Step Three - Outlying Aspects
Identification](#)

[Evaluation Results](#)

[Conclusion](#)

GOAM Algorithm



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
 - Step One - Group Feature Extraction
 - Step Two - Outlying Degree Scoring
 - Step Three - Outlying Aspects Identification
- [Evaluation Results](#)
- [Conclusion](#)

Framework of GOAM algorithm:

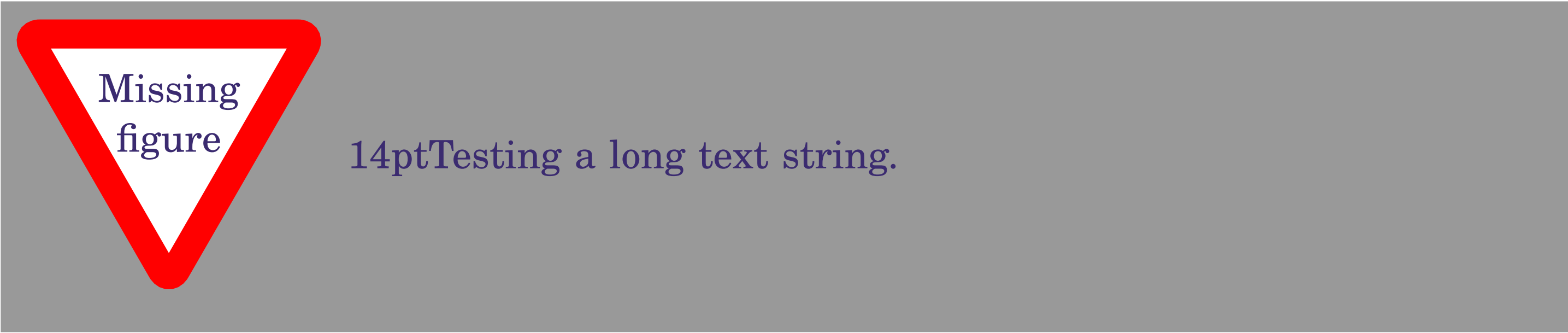


Figure 5: Framework of GOAM Algorithm



Step One - Group Feature Extraction

- Problem Definition
- Related Work and Challenges
- GOAM Algorithm
- Step One - Group Feature Extraction**
- Step Two - Outlying Degree Scoring
- Step Three - Outlying Aspects Identification
- Evaluation Results
- Conclusion

■ Suppose f_1, f_2, f_3 are three features of G_q .

$$f_1: \{x_1, x_2, x_3, x_4, x_5, x_2, x_3, x_4, x_1, x_2\}$$

$$f_2: \{y_2, y_2, y_1, y_2, y_3, y_3, y_5, y_4, y_4, y_2\}$$

$$f_3: \{z_1, z_4, z_2, z_4, z_5, z_3, z_1, z_2, z_4, z_2\}$$

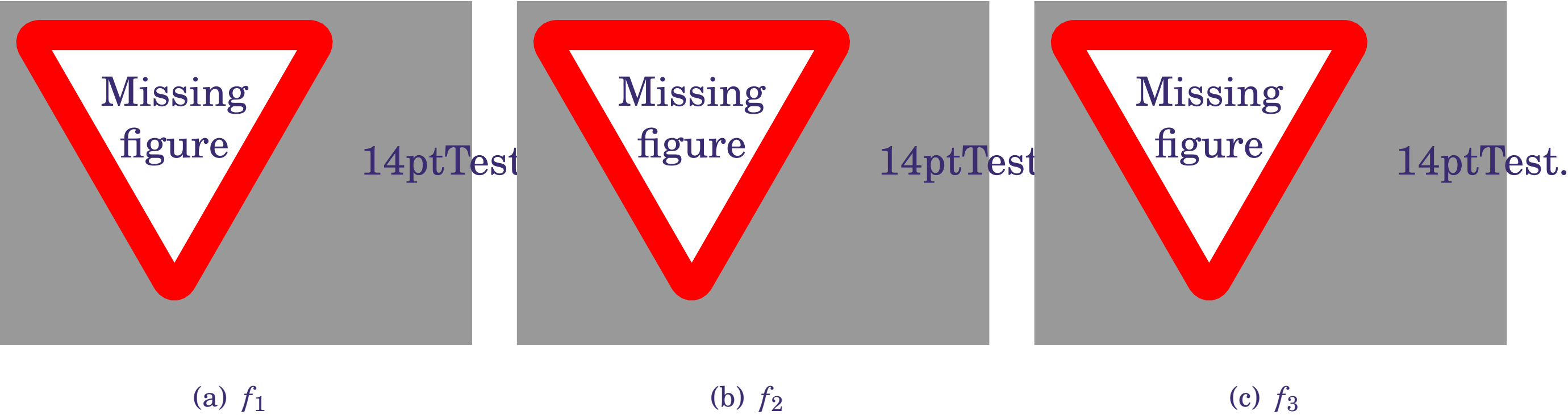


Figure 6: Histogram of G_q on three features



Step Two - Outlying Degree Scoring

- Problem Definition
- Related Work and Challenges
- GOAM Algorithm
- Step One - Group Feature Extraction
- Step Two - Outlying Degree Scoring
- Step Three - Outlying Aspects Identification
- Evaluation Results
- Conclusion

- Calculate Earth Mover Distance
 - ◆ Represent one feature among different groups
 - ◆ Purpose: calculate the minimum mean distance

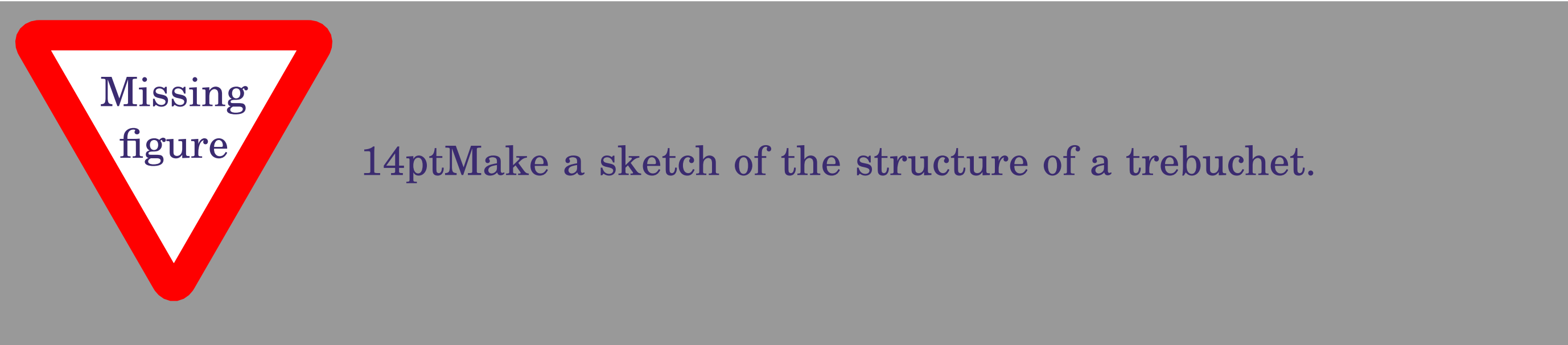


Figure 7: EMD of one feature



Step Two - Outlying Degree Scoring

Problem Definition
Related Work and Challenges
GOAM Algorithm
Step One - Group Feature Extraction
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
Conclusion

■ Calculate the outlying degree

$$OD(G_q) = \sum_1^n EDM(h_{q_s}, h_{k_s})$$

- ◆ $n \Leftrightarrow$ the number of contrast groups.
- ◆ $h_{k_s} \Leftrightarrow$ the histogram representation of G_k in the subspace s .





Step Three - Outlying Aspects Identification

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Step One - Group Feature Extraction](#)
- [Step Two - Outlying Degree Scoring](#)
- [Step Three - Outlying Aspects Identification](#)
- [Evaluation Results](#)
- [Conclusion](#)

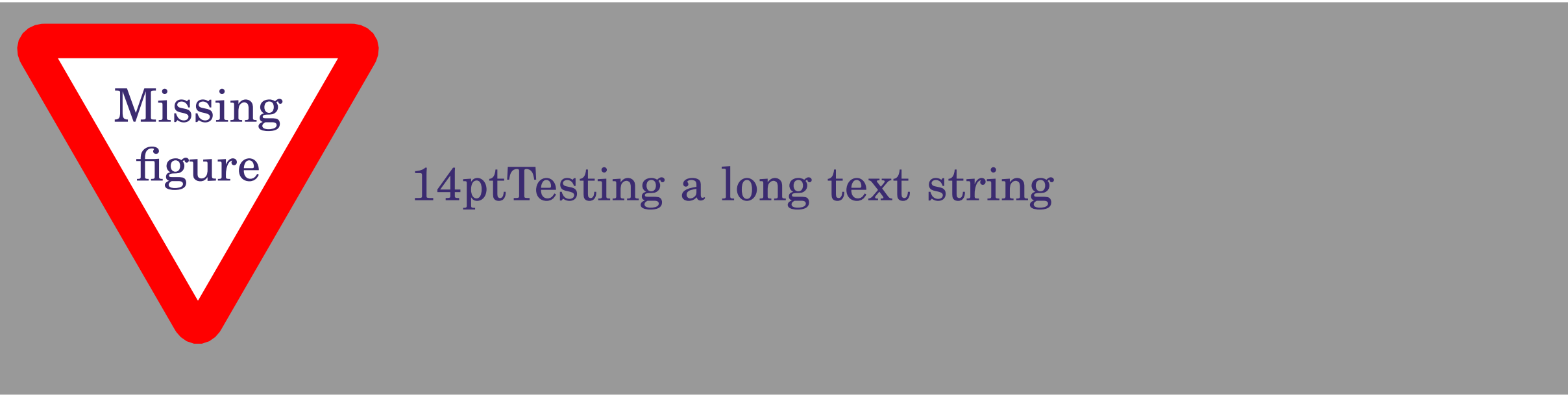
- Identify group outlying aspects mining based on the value of outlying degree.
- The greater the outlying degree is, the more likely it is group outlying aspect.



Pseudo code

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
 - [Step One - Group Feature Extraction](#)
 - [Step Two - Outlying Degree Scoring](#)
 - [Step Three - Outlying Aspects Identification](#)
- [Evaluation Results](#)
- [Conclusion](#)

■ Pseudo code of GOAM algorithm





Illustration

Problem Definition
Related Work and Challenges
GOAM Algorithm
Step One - Group Feature Extraction
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
Conclusion

Table 3: Original Dataset

G_1	F_1	F_2	F_3	F_4	G_2	F_1	F_2	F_3	F_4
	10	8	9	8		7	7	6	6
	9	9	7	9		8	9	9	8
	8	10	8	8		6	7	8	9
	8	8	6	7		7	7	7	8
	9	9	9	8		8	6	6	7
G_3	F_1	F_2	F_3	F_4	G_4	F_1	F_2	F_3	F_4
	8	10	8	8		9	8	8	8
	9	9	7	9		7	7	7	9
	10	9	10	7		8	6	6	8
	9	10	8	6		9	8	8	7
	9	9	7	9		8	7	9	8



Illustration

- Problem Definition
- Related Work and Challenges
- GOAM Algorithm
 - Step One - Group Feature Extraction
 - Step Two - Outlying Degree Scoring
 - Step Three - Outlying Aspects Identification
- Evaluation Results
- Conclusion

Table 4: outlying degree of each possible subspaces

Feature	Outlying Degree	Feature	Outlying Degree
$\{F_1\}$	4.351	$\{F_2, F_3\}$	4.023
$\{F_2\}$	2.012	$\{F_3, F_4\}$	4.324
$\{F_3\}$	1.392	$\{F_2, F_4\}$	2.018
$\{F_4\}$	2.207	$\{F_2, F_3, F_4\}$	2.012

■ Search process:

$OD(\{F_1\}) > \alpha$, save to T_1 .
 $OD(\{F_2\}) < \alpha$, save to C_1 .
 $OD(\{F_3\}) < \alpha$, save to C_2 .
 $OD(\{F_4\}) < \alpha$, save to C_3 .

$OD(\{F_2, F_3\}) > \alpha$, save to N_1 .
 $OD(\{F_3, F_4\}) > \alpha$, save to N_2 .
 $OD(\{F_2, F_4\}) < \alpha$, remove.
 $OD(\{F_2, F_3, F_4\}) < \alpha$, remove.



Strengths of GOAM Algorithm

Problem Definition
Related Work and Challenges
GOAM Algorithm
Step One - Group Feature Extraction
Step Two - Outlying Degree Scoring
Step Three - Outlying Aspects Identification
Evaluation Results
Conclusion

- **Reduction of Complexity**
 - ◆ Bottom-up search strategy.
 - ◆ Reduce the size of candidate subspaces.
- **Efficiency**
 - ◆ Before: $O(2^d)$
Now: $O(d * n^2)$





[Problem Definition](#)

[Related Work and Challenges](#)

[GOAM Algorithm](#)

Evaluation Results

[Synthetic Dataset](#)

[NBA Dataset](#)

[Conclusion](#)

Evaluation Results



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Evaluation Results](#)
 - [Synthetic Dataset](#)
 - [NBA Dataset](#)
- [Conclusion](#)

- $Accuracy = \frac{P}{T}$
 - P: Identified outlying aspects
 - T: Real outlying aspects



- Problem Definition
- Related Work and Challenges
- GOAM Algorithm
- Evaluation Results
- Synthetic Dataset
- NBA Dataset
- Conclusion

■ Synthetic Dataset and Ground Truth

Table 5: Synthetic Dataset and Ground Truth

Query group	F₁	F₂	<i>F₃</i>	F₄	<i>F₅</i>	<i>F₆</i>	<i>F₇</i>	<i>F₈</i>
<i>i₁</i>	10	8	9	7	7	6	6	8
<i>i₂</i>	9	9	7	8	9	9	8	9
<i>i₃</i>	8	10	8	9	6	8	7	8
<i>i₄</i>	8	8	6	7	8	8	6	7
<i>i₅</i>	9	9	9	7	7	7	8	8
<i>i₆</i>	8	10	8	8	6	6	8	7
<i>i₇</i>	9	9	7	9	8	8	8	7
<i>i₈</i>	10	9	10	7	7	7	7	7
<i>i₉</i>	9	10	8	8	7	6	7	7
<i>i₁₀</i>	9	9	7	7	7	8	8	8



Synthetic Dataset Results

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Evaluation Results](#)
- [Synthetic Dataset](#)
- [NBA Dataset](#)
- [Conclusion](#)

Table 6: The experiment result on synthetic dataset

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%



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Evaluation Results](#)
- [Synthetic Dataset](#)
- [NBA Dataset](#)
- [Conclusion](#)

Data Collection

Source

Yahoo Sports website (<http://sports.yahoo.com.cn/nba>)

Data

- Extract NBA teams’ data until March 30, 2018;
- 6 divisions;
- 12 features (eg: *Point Scored*).



- Problem Definition
- Related Work and Challenges
- GOAM Algorithm
- Evaluation Results
- Synthetic Dataset
- NBA Dataset
- Conclusion

The detail features are as follows:

Table 7: Collected data of Brooklyn Nets Team

Pts	FGA	FG%	3FA	3PT%	FTA	FT%	Reb	Ass	To	Stl	Blk
18	12	42	2.00	50	7.00	100	0	4	3	0	0
15.7	14.07	41	5.45	32	3.05	75	3.98	5.1	2.98	0.69	0.36
14.5	11.1	47	0.82	26	4.87	78	6.82	2.4	1.74	0.92	0.66
13.5	10.8	42	5.37	37	3.38	77	6.66	2	1.38	0.83	0.42
12.7	10.59	39	5.36	33	3.37	82	3.24	6.6	1.56	0.89	0.31
12.6	10.93	40	6.94	37	1.70	84	4.27	1.5	1.06	0.61	0.44
12.2	10.39	44	3.42	35	2.70	72	3.79	4.1	2.15	1.12	0.32
10.6	7.85	49	4.51	41	1.35	83	3.34	1.6	1.15	0.45	0.24



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Evaluation Results](#)
- [Synthetic Dataset](#)
- [NBA Dataset](#)
- [Conclusion](#)

■ Data Preprocess

Table 8: The bins that used to discrete data of each feature

Labels	Pts	FGA	FG%	3FA	3PT%	FTA
low	[0,5]	[0,4]	[0,0.35]	[0,1.0]	[0,0.2]	[0,1.0]
medium	(5,10]	(4,7]	(0.35,0.45]	(1.0,2.5]	(0.2,0.3]	(1.0,1.5]
high	(10,15]	(7,10]	(0.45,0.5]	(2.5,3.5]	(0.3,0.35]	(1.5,2.5]
very high	(15,+∞]	(10,+∞]	(0.5,1]	(3.5,+∞]	(0.35,1]	(2.5,+∞]
Labels	FT%	Reb	Ass	To	Stl	Blk
low	[0,0.6]	[0,2.0]	[0,1.0]	[0,0.6]	[0,0.2]	[0,0.25]
medium	(0.6,0.65]	(2,5]	(1,2]	(0.6,0.9]	(0.2,0.5]	(0.25,0.5]
high	(0.65,0.75]	(5,6]	(2,4]	(0.9,1.7]	(0.6,0.75]	(0.5,0.7]
very high	(0.75,1]	(6,+∞]	(4,+∞]	(1.7,+∞]	(0.75,+∞]	(0.7,+∞]



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Evaluation Results](#)
- [Synthetic Dataset](#)
- [NBA Dataset](#)
- [Conclusion](#)

Table 9: The identified outlying aspects of groups

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}
Golden State Warriors	{FG%}	{FT%, Blk}, {FGA, 3PT%, FTA}
Utah Jazz	{Blk}	{3FA, 3PT%}
New Orleans Pelicans	{FT%}, {FTA}	{FTA, Stl}, {FTA, To}



- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Evaluation Results](#)
- [Conclusion](#)

Conclusion



Conclusion

[Problem Definition](#)

[Related Work and Challenges](#)

[GOAM Algorithm](#)

[Evaluation Results](#)

[Conclusion](#)

- Formalize the problem of *Group Outlying Aspects Mining* by extending outlying aspects mining;
- Propose a novel method **GOAM algorithm** to solve the *Group Outlying Aspects Mining* problem;
- Utilize the pruning strategies to reduce time complexity.



TULIP

Team for Universal Learning and Intelligent Processing



Questions?

- [Problem Definition](#)
- [Related Work and Challenges](#)
- [GOAM Algorithm](#)
- [Evaluation Results](#)
- [Conclusion](#)



Contact Information

Associate Professor Gang Li
School of Information Technology
Deakin University, Australia



GANGLI@TULIP.ORG.AU



TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING

