Bike Sharing Demand Prediction

Dong Zhu
Deakin University, Australia

Introduction

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.
The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

Dataset Description

- The competition provides hourly rental data spanning two years. For this competition, the training set is comprised of the first 19 days of each month, while the test set is the 20th to the end of the month. You must predict the total count of bikes rented during each hour covered by the test set, using only information available prior to the rental period.
- Data Fields

datetime: hourly date + timestamp
season: 1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday: whether the day is considered a holiday
workingday: whether the day is neither a weekend nor holiday
weather: 1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp: temperature in Celsius
atemp: "feels like" temperature in Celsius
humidity: relative humidity
windspeed: wind speed
casual: number of non-registered user rentals initiated
registered: number of registered user rentals initiated
count: number of total rentals

GOAM Algorithm

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

Missing figure Testing figcolor

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.

Missing figure Testing figcolor

Histogram of G_q on f_1

Missing figure Testing figcolor

Histogram of G_q on f_2

Missing figure Testing figcolor

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.

Missing figure Testing figcolor

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}

Missing figure Testing figcolor

New Orleans Pelicans on FT%

Missing figure Testing figcolor

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