Package ‘RWBP’

February 19, 2015

Type Package

Title Detects spatial outliers using a Random Walk on Bipartite Graph

Version 1.0

Date 2014-06-23

Author Sigal Shaked & Ben Nasi

Maintainer Sigal Shaked <shaksi@post.bgu.ac.il>

Description
a Bipartite graph and is constructed based on the spatial and/or non-spatial attributes of the spatial objects in the dataset. Secondly, RW techniques are utilized on the graphs to compute the outlierness for each point (the differences between spatial objects and their spatial neighbours). The top k objects with higher outlierness are recognized as outliers.

License GPL (>= 2)

Depends RANN, igraph, lsa, SnowballC

NeedsCompilation no

Repository CRAN

Date/Publication 2014-06-24 23:30:50

R topics documented:

`RWBP-package` *Random Walk on Bipartite Graph*

Description

Detects spatial outliers using Random Walk on Bipartite Graph technique

Details

Package:	RWBP
Type:	Package
Version:	1.0
Date:	2014-06-23
License:	GPL (>=2)

See the example below in order to use the package. important methods: predict.RWBP, RWBP.formula, RWBP

Author(s)

Sigal Shaked & Ben Nasi

Maintainer: Sigal Shaked <shaksi@post.bgu.ac.il>

References

Liu X., Lu C.T., Chen F.: Spatial outlier detection: Random walk based approaches. In: Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), San Jose, CA (2010).

Examples

```
#an example dataset:
trainSet <- cbind(
c(7.092073,7.092631,7.09263,7.093052,7.092876,7.092689,7.092515,7.092321,
7.092138,7.11455,7.11441,7.11408,7.11376,7.11338,7.11305,7.11277,7.1124,
7.11202,7.11161,7.11115,7.11068,7.11014,7.10963,7.1095,7.1089,7.10818,
7.10747,7.10674,7.116691,7.116142,7.115559,7.115007,7.114423,7.113838,
7.113272,7.112684,7.112067,7.111458,7.110869,7.110274,7.109696,7.109131,
7.109231,7.108546,7.10797,5.599215,5.597609,5.596588,5.595359,5.594478,5.593652),
c(50.77849,50.77859,50.7786,50.77878,50.77914,50.77952,50.77992,50.78035,
50.78081,53.8,53.7,53.6,53.5,54.2,55.3,55.2,56.6,57.6,57.7,58.8,59.4,59.7,
59,59.03,59.3,60.7,60.8,61.4,50.73922,50.73914,50.73905,50.73899,50.73889,
50.73881,50.73873,50.73865,50.73856,50.73847,50.73838,50.73831,50.73822,
50.73814,50.73937,50.73805,50.73798,43.2034,43.20338,43.20352,43.2037,43.20391,43.20409),
c(106.5,107.6,25,108.5,109.1,109.7,111.6,113.3,113.3,62.3,333.7,331.5,327.2,
325.5,324.8,323.5,322.3,320.3,319,317.8,316,315.1,315.3,12,312.4,311.3,310.8,
309.4,99.2,99.2,101.1,99.5,101.3,105.3,104.3,104.4,106.3,108.8,110.3,111.7,113.3,
112.1,5000,111.6,109.8,125.6,130,132.3,133.4,138,143.4),
c(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,1,0,0,0,0,0,0,0,0,0)
)

colnames(trainSet)<- c("lng","lat","alt","isOutlier")

#first to columns of the input data are assumed to be spatial coordinates,
#and the rest are non-spatial attributes according to which outliers will be extracted
myRW <- RWBP(as.data.frame(trainSet[,1:3]), clusters.iterations=6)

#predict classification:
testPrediction<-predict(myRW,3 )
```

```
#calculate accuracy:
sum(testPrediction$class==trainSet[,"isOutlier"])/nrow(trainSet)
#confusion table
table(testPrediction$class, trainSet[,"isOutlier"])

#other options:
myRW1 <- RWBP(isOutlier~lng+lat+alt, data=as.data.frame(trainSet))
#print model summary
print(myRW1)
#plot model graph
plot(myRW1)
#predict probabilities of each record to be an outlier:
predict(myRW1 , top_k=4,type="prob")
```

---

predict.RWBP *predict.RWBP*

Description

Predict spatial outliers according to a RWBP model

Usage

```
## S3 method for class 'RWBP'
predict(object, top_k = 3, type = "raw", ...)
```

Arguments

object	a RWBP object
top_k	the number of outliers to extract
type	"raw" returns classification results (0 for normal, 1 for outlier). "prob" returns probabilities for being outlier.
...	currently not in use

Value

Returns the input data frame/matrix with an additional column that contains the prediction results. The additional column is set according to the type parameter:

raw	"class" column is added
prob	"prob" column is added

Author(s)

Sigal Shaked & Ben Nasi

References

Liu X., Lu C.T., Chen F.: Spatial outlier detection: Random walk based approaches. In: Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), San Jose, CA (2010).

See Also

RWBP,RWBP-package

Examples

```
#an example dataset:
trainSet <- cbind(
c(7.092073,7.092631,7.09263,7.093052,7.092876,7.092689,7.092515,7.092321,
7.092138,7.11455,7.11441,7.11408,7.11376,7.11338,7.11305,7.11277,7.1124,
7.11202,7.11161,7.11115,7.11068,7.11014,7.10963,7.1095,7.1089,7.10818,
7.10747,7.10674,7.116691,7.116142,7.115559,7.115007,7.114423,7.113838,
7.113272,7.112684,7.112067,7.111458,7.110869,7.110274,7.109696,7.109131,
7.109231,7.108546,7.10797,5.599215,5.597609,5.596588,5.595359,5.594478,5.593652),
c(50.77849,50.77859,50.7786,50.77878,50.77914,50.77952,50.77992,50.78035,
50.78081,53.8,53.7,53.6,53.5,54.2,55.3,55.2,56.6,57.6,57.7,58.8,59.4,59.7,
59,59.03,59.3,60.7,60.8,61.4,50.73922,50.73914,50.73905,50.73899,50.73889,
50.73881,50.73873,50.73865,50.73856,50.73847,50.73838,50.73831,50.73822,
50.73814,50.73937,50.73805,50.73798,43.2034,43.20338,43.20352,43.2037,43.20391,43.20409),
c(106.5,107.6,25,108.5,109.1,109.7,111.6,113.3,113.3,62.3,333.7,331.5,327.2,
325.5,324.8,323.5,322.3,320.3,319,317.8,316,315.1,315.3,12,312.4,311.3,310.8,
309.4,99.2,99.2,101.1,99.5,101.3,105.3,104.3,104.4,106.3,108.8,110.3,111.7,113.3,
112.1,5000,111.6,109.8,125.6,130,132.3,133.4,138,143.4),
c(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,1,0,0,0,0,0,0,0,0)
)

colnames(trainSet)<- c("lng","lat","alt","isOutlier")

#first to columns of the input data are assumed to be spatial coordinates,
#and the rest are non-spatial attributes according to which outliers will be extracted
myRW <- RWBP(as.data.frame(trainSet[,1:3]), clusters.iterations=6)

#predict classification:
testPrediction<-predict(myRW,3 )
#calculate accuracy:
sum(testPrediction$class==trainSet[,"isOutlier"])/nrow(trainSet)
#confusion table
table(testPrediction$class, trainSet[,"isOutlier"])

#other options:
myRW1 <- RWBP(isOutlier~lng+lat+alt, data=as.data.frame(trainSet))
#print model summary
print(myRW1)
#plot model graph
plot(myRW1)
#predict probabilities of each record to be an outlier:
```

```
predict(myRW1 , top_k=4,type="prob")
```

RWBP *Random Walk on Bipartite Graph*

Description

Performs an outlier detection on a given data frame/matrix.

Usage

```
RWBP(x,...,nn_k,min.clusters,clusters.iterations,
clusters.stepSize,alfa,dumping.factor)
## Default S3 method:
RWBP(x,...,nn_k=10,min.clusters=8,clusters.iterations=6,
clusters.stepSize=2,alfa=0.5,dumping.factor=0.9)
## S3 method for class 'formula'
RWBP(formula,data,...,nn_k=10,min.clusters=8,clusters.iterations=6,
clusters.stepSize=2,alfa=0.5,dumping.factor=0.9)
## S3 method for class 'RWBP'
print(x, ...)
## S3 method for class 'RWBP'
plot(x, ...)
```

Arguments

`formula`	a formula representation of the problem (the dependent variable (y) will be ignored, the first two x attributes have to be spatial coordinates and the rest are numeric attributes)
`data`	a data frame containing the data to be analysed (may contain additional columns).
`x`	a data frame containing the data to be analysed. the first two columns must be spatial coordinates and the other columns are non-spatial attributes on which we search for outliers
`nn_k`	neighbourhood size (for finding each objects k nearest neighbours)
`min.clusters`	the number of clusters in the first clustering process
`clusters.iterations`	the number of clustering process to be conducted
`clusters.stepSize`	increase the amount of clusters in the following clustering process by this size
`alfa`	helps to compute more accurate edge value (distance between object and cluster)
`dumping.factor`	dumping factor (the probability to return to the original node during each step along a random walk)
`...`	currently not in use

Details

A spatial outlier detection approach based on RW techniques. A Bipartite graph is constructed based on the spatial and/or non-spatial attributes of the spatial objects in the dataset. Secondly, RW techniques are utilized on the graphs to compute the outlierness for each point (the differences between spatial objects and their spatial neighbours). The top k objects with higher outlierness are recognized as outliers.

Value

Returns as RWBP object that contains several components:

- `data` — the data after removing records with empty fields
- `X` — a data frame containing the spatial attributes(first two columns) from the input data
- `Y` — a data frame containing the non-spatial attributes(all but the first two columns) from the input data
- `ID` — a vector with sequential numbers, used as an index
- `n` — number of valid records
- `n.orig` — number of records accepted in the input data
- `nn_k` — neighbourhood size for knn search
- `k` — clusters amount in the first clustering process
- `clusters.stepSize` — each next clustering process is increased by this size
- `h` — number of conducted clustering processes
- `alfa` — Helps to compute more accurate edge value (distance between object and cluster)
- `c` — Dumping factor (the probability to return to the original node during each step along a random walk)
- `nearest.indexes` — a matrix where each row contains a spatial object's nn_k nearest neighbours
- `clusteredData` — a data frame containing the results of all clustering process: an object, the cluster it belongs to and the distance between the two
- `igraph` — an igraph object built according to the connections between spatial objects and clusters
- `OutScore` — the outlierness scores of each record, sorted ascending by score, the first column is the index of the record and the second column is the given score
- `objects.similarity` — a matrix where each row holds the similarity between a spatial object and its nn_k neighbours

Note

First two columns must be spatial coordinates, and the rest of the columns must be numeric attributes. records with empty fields are removed from the input data.

Author(s)

Sigal Shaked & Ben Nasi

References

Liu X., Lu C.T., Chen F.: Spatial outlier detection: Random walk based approaches. In: Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS), San Jose, CA (2010).

See Also

predict.RWBP,RWBP-package

Examples

```
#an example dataset:
trainSet <- cbind(
c(7.092073,7.092631,7.09263,7.093052,7.092876,7.092689,7.092515,7.092321,
7.092138,7.11455,7.11441,7.11408,7.11376,7.11338,7.11305,7.11277,7.1124,
7.11202,7.11161,7.11115,7.11068,7.11014,7.10963,7.1095,7.1089,7.10818,
7.10747,7.10674,7.116691,7.116142,7.115559,7.115007,7.114423,7.113838,
7.113272,7.112684,7.112067,7.111458,7.110869,7.110274,7.109696,7.109131,
7.109231,7.108546,7.10797,5.599215,5.597609,5.596588,5.595359,5.594478,5.593652),
c(50.77849,50.77859,50.7786,50.77878,50.77914,50.77952,50.77992,50.78035,
50.78081,53.8,53.7,53.6,53.5,54.2,55.3,55.2,56.6,57.6,57.7,58.8,59.4,59.7,
59,59.03,59.3,60.7,60.8,61.4,50.73922,50.73914,50.73905,50.73899,50.73889,
50.73881,50.73873,50.73865,50.73856,50.73847,50.73838,50.73831,50.73822,
50.73814,50.73937,50.73805,50.73798,43.2034,43.20338,43.20352,43.2037,43.20391,43.20409),
c(106.5,107.6,25,108.5,109.1,109.7,111.6,113.3,113.3,62.3,333.7,331.5,327.2,
325.5,324.8,323.5,322.3,320.3,319,317.8,316,315.1,315.3,12,312.4,311.3,310.8,
309.4,99.2,99.2,101.1,99.5,101.3,105.3,104.3,104.4,106.3,108.8,110.3,111.7,113.3,
112.1,5000,111.6,109.8,125.6,130,132.3,133.4,138,143.4),
c(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,1,0,0,0,0,0,0,0,0)
)

colnames(trainSet)<- c("lng","lat","alt","isOutlier")

#first to columns of the input data are assumed to be spatial coordinates,
#and the rest are non-spatial attributes according to which outliers will be extracted
myRW <- RWBP(as.data.frame(trainSet[,1:3]), clusters.iterations=6)

#predict classification:
testPrediction<-predict(myRW,3 )
#calculate accuracy:
sum(testPrediction$class==trainSet[,"isOutlier"])/nrow(trainSet)
#confusion table
table(testPrediction$class, trainSet[,"isOutlier"])

#other options:
myRW1 <- RWBP(isOutlier~lng+lat+alt, data=as.data.frame(trainSet))
#print model summary
```

```
print(myRW1)
#plot model graph
plot(myRW1)
#predict probabilities of each record to be an outlier:
predict(myRW1 , top_k=4,type="prob")
```

Index