APPUNTI LEGATI AI RISULTATI DEI TEST EFFETTUATI

Chiamerò così i dataset:

* 1 – 1043 record
* 2 – 2235 record
* 3 – 3129 record
* 4 – 5216 record

Considerazioni GENERALI:

* i risultati non arrivano mai oltre il 4 livello di albero anche con il 4.
* Larceny essendo l’evento più frequente è sempre in cima ai risultati
* Probabilmente per comprendere meglio queste associazioni serve un esperto di dominio
* non considerare i risultati in cima (Top) in quanto sono di lunghezza 2 e legati all’alta frequenza degli stessi connessi a eventi più sparsi di bassa frequenza

USARE vicinato grande genera risultati con alti pi ma tutti vicini come valore (difficile distinguere quali sono importanti e quali no) fare vicinato troppo piccolo genera risultati con bassi pi e probabilmente più spaziati nel range 0 – 1. Credo sia meglio una via di mezzo tendente al vicinato piccolo in quanto più i risultati si differenziano più si notano le differenze.

NON so a che quota di pi iniziano i risultati significativi, è molto relativo

Il **primo** dataset un po’ in tutti i casi non ha un teta finale > 0.25 e le sequenze con alto PI si fermano al secondo livello raramente si trovano sequenze lunghe 3 importanti.

Nell’analisi non aiuta a comprendere le potenzialità dell’algoritmo e i risultati però potrei usarlo per capire come e quanto scala nel tempo

Il **secondo** dataset ha dei teta finali interessanti quando il vicinato è sufficientemente ampio, 3 di raggio e 120 tempo.

Il più significativo r = 3 e t = 168 (0.75), però in quest’ultimo caso ho moltissimi risultati vicini per cui non so quanto potrei ricavarne, notare sempre che le sequenze con pi più elevato sono quelle lunghe 2, la prima lunga 3 è l’ottava.

Il **terzo** dataset è probabilmente quello più equilibrato, i teta finali dei test hanno risultati medi (attorno al 0.5) per vicinati di grandezza media (rispetto agli altri). teta bassi per vicinati piccoli e teta alti per vicinati grandi.

Nel caso r = 1.5, t = 168: sono più spaziati i risultati ma non superano l’0.85 e le sequenze di lunghezza 3 sono sotto 0.68.

Il **quarto** dataset è quello con valori di teta finali maggiori, quindi trova tendenzialmente sequenze di alto valore di pi in quanto vi sono più record nell’analisi.

SKYPE prof:

La variabilità dei record è per capire quanto scala rispetto ai parametri presi singolarmente, come scala rispetto al raggio, rispetto a tempo, rispetto alla lunghezza delle sequenze.

Qualità delle regole: basket problem per distinguere sequenze utili o meno utili (le trovo online in R) cerca in letteratura

Dove si apprezza la differenza di scala, dire che il numero di record è molto rilevante.

Misure di valutazione collegare su moodle di data e text mining vedere lezione sul tema regole di associazione sono già implementate in R. Quanto una regola è interessante, interestingness, concetti di regola massimale. Comprendere le coppie più significative in base a quello che viene dopo nella sequenza. Viene fatto dopo le regole trovate. Faccio questo lavoro sul test più significativo, devo stabilire io qual è il più significativo e perchè.

Scelgo il gruppo di rules da valutare e cerco di fare la seguente cosa:

* Lhs = Csv con gli antecendenti delle regole
* Rhs = Csv con i conseguenti delle regole
* Support = Csv con i valore di PI

Tutto deve essere rigorosamente ordinato e coincidente con gli altri csv.

L’idea è quella di caricarli in R e creare così le rules – new(“rules”, lhs, rhs, support)

Dalle rules posso applicare la summary e la inspect per analizzare i risultati

Esempio da riapplicare:

set.seed(1234)

## Generate random data and coerce data to itemMatrix.

m <- matrix(runif(100000)>0.8, ncol=20)

dimnames(m) <- list(NULL, paste("item", c(1:20), sep=""))

i <- as(m, "itemMatrix")

## Get the number of elements (rows) in the itemMatrix.

length(i)

## Get first 5 elements (rows) of the itemMatrix as list. as(i[1:5], "list")

## Get first 5 elements (rows) of the itemMatrix as matrix. as(i[1:5], "matrix")

## Get first 5 elements (rows) of the itemMatrix as sparse ngCMatrix.

## Warning: for efficiency reasons, the ngCMatrix you get is transposed! as(i[1:5], "ngCMatrix")

## Get labels for the first 5 itemsets (first default and then with

## custom formating)

labels(i[1:5])

labels(i[1:5], itemSep = " + ", setStart = "", setEnd = "")

## create itemsets from itemMatrix

is <- new("itemsets", items = i[1:3])

inspect(is)

## create rules (rhs and lhs cannot share items so I use

## itemSetdiff here). Also assign (random) support.

rules <- new("rules", lhs=itemSetdiff(i[4:6],i[1:3]), rhs=i[1:3], quality = data.frame(support = runif(3)))

inspect(rules)

03/10/19

Last analisi

Lift è quasi sempre > 1 infatti significa che il nostro algoritmo estrae le sequenze di eventi significative non nelle quali i tipi di eventi sono stocasticamente indipendenti FATTO

Il supporto ha molto spesso valori bassi < 10% perché il dataset è suddiviso in 9 classi per cui è normale. (100/9) naturalmente qua influisce molto la frequenza dei tipi. Molto divesificata tra classi per dire il larceny ha un sacco di record in più rispetto alle altre. FATTO

La confidenza ci segnala quanto l’ultimo elemento della sequenza dipenda dalla sequenze di tipi antecedenti. Per questo motivo è molto rilevante nella nostra analisi, è anche correlato con pi

Le sequenze lunghe 2 sono poco significative alla nostra analisi in quanto ovviamente hanno i pi, confidenze e altri indici maggiorati rispetto alle sequenze più lunghe, e spiega perché. FATTO

Quindi mi concentro sulle sequenze lunghe 3 o 4 che hanno pi e confidenza alti.

Riass:

lift generale

pi e confidenza + importanti, assieme viene considerato il supporto

\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\

Presentazione

1. Titolo [fatto]
2. Paper - Dataset utilizzato [fatto]
3. Vicinato – setInstances [fatto]
4. PI e PR [fatto]
5. Algoritmo (threshold e numTop) [fatto]
6. Albero SPTree [fatto]
7. Generate/verify candidates [fatto]
8. Implementazione e dataset test [fatto]
9. Analisi tempi [fatto]
10. Analisi indici [fatto]
11. Conclusione e prospettive future [fatto]
12. Ringraziamenti [fatto]