Data Mining - The Diary

Rodion "rodde" Efremov

March 24, 2015

Introduction

This document is my learning diary written on behalf of Data Mining course led at spring term 2015 at University of Helsinki.

1 Week 1

The **support count** $\sigma(X)$ of an item set X is the amount of transactions containing X ($X \subset t_i$). Basically, we were computing support counts for various itemsets with the exception of applying additional constraints to the queries (such as particular grade range).

The **support** of an item set X is $\sigma(X)/N$, where N is the amount of all transactions. Support of X may be thought of as a classical probability of a random transaction containing X.

An **association rule** is an implication of the form $X \to Y$, where X and Y are itemsets having no items in common. The interpretation of an association rule is that if a transaction contains X, it "tends" to contain Y as well. Note that "tends" depends on parameters we specify to a data mining system. **Support** of an association rule $X \to Y$ is

$$s(X \to Y) = \frac{\sigma(X \cup Y)}{N}.$$

Support of the rule R may be thought of as a classical probability of R appearing in a random transaction. **Rule confidence** gives the probability of Y appearing in the same transactions with set X and is defined as

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

1.1 Reflection

Getting the data from a file to internal representation was pretty challenging: the data seems a little bit "dirty" and I am sure there is room for improvement. What comes to accessing data, I have made an effort to make sure that it runs fast. Basically I have three model classes:

`Course` holds the course name, the course code, grading mode and the amount of credits awarded,

`Student` holds only a unique student ID and enrollment year,

`CourseAttendanceEntry` holds a course C, a student S, the year and month S attended C, and the grade S received. Basically, these entries implement a many-to-many relationship between courses and students.

2 Week 2

Task 5

The supports are as follow:

E	0.684
O	0.632
P	0.526
W	0.158
EO	0.474
EP	0.316
EW	0.053
OP	0.263
OW	0.053
PW	0.105
EOP	0.221
EOW	0.053
EPW	0
OPW	0
EOPW	0

The only observation that I was able to come up with is that if $s(X)$ is support of an itemset X, then

$$s(X) \leq \min_{A \subsetneq X} s(A).$$

Task 10

We have around 23 million (N) different paperback books and we want to generate all 10-combinations of those. Suppose we are given an index tuple $t = (t_1, t_2, \ldots, t_{10}) = (1, 2, \ldots, 10)$. Next generate a combination of books indexed by t and increment t_{10}. When $t_{10} = N + 1$, increment t_9 and set $t_{10} = t_9 + 1$. After $t_9 = N - 1$ (and thus $t_{10} = N$) has been generated, increase t_8 and set $t_9 = t_8 + 1, t_{10} = t_8 + 2$. Continue this routine until $t_1 = N - 9, t_2 = N - 8, \ldots, t_9 = N - 1, t_{10} = N$.

Task 15

In this task we are supposed to measure time of generating k-combinations of courses for $k \in \{2, 3, 5\}$. The results are summarized in the following table:

k	t
2	4 ms
3	40 ms
5	291 ms

Increasing k from 2 to 3 increases the running time by a factor of 10; increasing k from 3 to 5 increases the running time by a factor of 7,3. Since $n = 213$,

$$\begin{aligned}\binom{n}{3}\binom{n}{2}^{-1} &= \frac{n!2!(n-2)!}{n!3!(n-3)!)} \\ &= \frac{(n-2)!}{3(n-3)!} \\ &= \frac{n-2}{3} \\ &\approx 70,\end{aligned}$$

and

$$\begin{aligned}\binom{n}{5}\binom{n}{3}^{-1} &= \frac{n!3!(n-3)!}{n!5!(n-5)!)} \\ &= \frac{(n-3)!}{20(n-5)!} \\ &= \frac{(n-4)(n-3)}{20} \\ &\approx 2100,\end{aligned}$$

which does not quite go hand in hand with the measurements.

Task 19

The objective of this task is to compare brute-force and Apriori algorithms for frequent itemset generation.

k	support	Brute-force (ms)	Apriori (ms)
2	0.3	379	154
3	0.175	9389	774
4	0.1	N/A	1845
5	0.1	N/A	1637

After Arto's counsel, I was able to speedup generation of 3-combinations by a factor of 20, but I was not able to make 4-combination generation feasible.

Task 21

The largest size of itemsets with support at least 0.05 seems to be 11. I got 19 of such itemsets; one of them is

- TVT-ajokortti
- Ohjelmoinnin perusteet

- Opiskelutekniikka
- Tietokantojen perusteet
- Ohjelmoinnin jatkokurssi
- Tietoliikenteen perusteet
- Tietorakenteet ja algoritmit
- Johdatus tietojenkäsittelytieteeseen
- Tietokone työvälineenä
- Ohjelmistotekniikan menetelmät
- Aineopintojen harjoitustyö: Tietokantasovellus