COMP4433 Data Mining and Data Warehousing

A Second Thought of Support and Confidence

- Among 5000 students
  - 3000 play basketball (60%)
  - 3750 eat cereal (75%)
  - 2000 both play basket ball and eat cereal (40%)

	Basketball	Not Basketball	Sum(row)
Cereal	2000	1750	3750
Not Cereal	1000	250	1250
Sum(column)	3000	2000	5000

Two rules of concern:
R1: *play basketball* $\Rightarrow$ *eat cereal* [40%, 66.7%]
R2: *play basketball* $\Rightarrow$ *not eat cereal* [20%, 33.3%]

Basically, we may think of a database like this one for which the statistics above can be collected.

Student ID	Student Name	Address	Play Basketball	Eat Cereal	Attribute x1 …
0001	David Chan	…	Yes	No	
0002	Stephen Chung	…	Yes	Yes	
	…				
3000	May Yeung	…	Yes	No	
3001	Aaron Chin	…	No	Yes	
3002	John Mok	…	No	Yes	
	…	…			
5000	Eric Young		No	No	

Which recorded the information from a survey of 5000 PolyU UGstudents, like:

Life Style Survey

…

Q.1 Do you play basketball? □ Yes □ No

Q.2 Do you eat cereal? □ Yes □ No

…

Now, let's think about a promotion campaign for 5000 CityU UG students. The problem is that we only have a budget to send promotion pamphlet of a new brand of cereal to 3000 CityU UG students only. For the PolyU database above, we have 3750 students "eat cereal". We can simply select any 3000 students from there to send out the pamphlets. However, for CityU's DB, we have the following:

Student ID	Student Name	Address	Play Basketball	Eat Cereal	Attribute x1 …
5001	…	…	Yes	Unknown	
5002	…	…	Yes	Unknown	
8000	…	…	Yes	Unknown	
8001	…	…	No	Unknown	
8002	…	…	No	Unknown	
10000	…	…	No	Unknown	

where the "eat cereal" attribute values are not known. Assume that this database follows the statistics of PolyU, which 3000 students should be sent the pamphlets?

Solution 0 (without using data mining result):
Randomly select 3000 students to send the pamphlets. The response rate should be 75% (cf. 3750/5000), i.e. 2250 returns will be received.

Solution 1 (R1 is more interesting):
If we consider R1 is very interesting, we should send the letters to the students who play basketball, i.e., student ID 5001-8000. Then, the respond rate should be 66.7%, i.e. 2000 returns will be received!

Solution 2 (R2 is more interesting):
How about using R2? Based on R2, the students who play basketball do not eat cereal. Those students should have lower priority for sending them promotion pamphlets. In other words, we should choose students with ID 8001-10000, altogether 2000 students. According to the past statistics, the response rate should be 1750/2000, i.e. 1750 returns will be received. As we can send out 3000 pamphlets, we can randomly select 1000 students from the remaining students, i.e., student ID 5001-8000, to send out the invitation letters. The number of returns will be 1000x66.7%=667. Hence, the total number of returns will be 1750+667=2417.

Obviously, we should use solution 2. The interestingness of R1 is not as interesting as expected.