RINGKASAN

PENENTUAN PELUANG KEPINDAHAN PELANGGAN DENGAN MENGGUNAKAN MODEL REGRESI LOGISTIK UNTUK DATA BERUKURAN BESAR, Eto Wuryanto, Dyah Herawatie dan Rimuljo Hendradi, 2006, 32 halaman.

Pemodelan regresi logistik sulit dilakukan jika data yang digunakan termasuk *massive dataset* yang pada dasarnya berkaitan dengan jumlah data yang sangat besar. Oleh karena itu, perlu dilakukan penyamplingan terhadap data induknya dan menurut penelitian sebelumnya data *Squashing* merupakan metode yang baik untuk mereduksi data yang telah dilakukan oleh Eto et al (2005) dan Rimuljo et al (2005).

Dalam mendapatkan data *Squashing* ada dua tahap yang sulit yaitu : Pertama, pengelompokan atau pembuatan partisi terhadap data induk. Walaupun penelitian sebelumnya telah membahas data *Squashing* tetapi hanya melibatkan data dengan satu variabel. Sementara penggunaan model regresi logistik datanya minimal harus memuat 2 variabel sehingga akan terjadi kesulitan untuk mendapatkan partisi yang cocok. Akibatnya timbul persoalan, bagaimana cara pembuatan partisi untuk 2 variabel atau lebih. Kedua, metode apa yang bisa dipakai untuk menentukan nilai *pseudo point* dan pembobotnya untuk setiap partisi. Setelah diperoleh data *Squashing*, bagaimana cara menerapkan data *Squashing* tersebut pada model regresi logistik dan mengestimasi parameternya.

Penelitian ini bertujuan mendapatkan data *Squashing* dengan dua variabel atau lebih dan menentukan estimator parameter model regresi logistik untuk data berukuran besar dengan program dalam bahasa C++ sebagai langkah awal dalam pembuatan *software datamining*.

Untuk menentukan data *Squashing* dapat dilakukan dengan cara : Pertama, urutkan data berdasarkan variabel respon, sehingga didapat kelompok data dengan variabel respon nol dan kelompok data dengan variabel respon satu. Kedua, pada masing-masing variabel respon urutkan variabel prediktor berdasarkan varibel yang bertype non rasio. Ketiga, urutkan data yang ada di variabel rasio dari kecil ke besar. Keempat, pilih kelompok yang mempunyai anggota terbesar kemudian bagi ke dalam

dua kelompok dan lakukan proses pembagian ini sampai diperoleh sebanyak kelompok yang diinginkan. Kelima, hitung nilai *mean* dan jumlah anggota dari tiap-tiap kelompok yang masing-masing merupakan nilai pengamatan baru (*pseudo data point*) dan pembobot. Sedangkan untuk mengestimasi parameter β pada Model Regresi Logistik dipakai metode MLE dan jika diketahui hasilnya berupa fungsi implisit maka digunakan metode Newton-Raphson.

Implementasi algoritma metode *squashing* dan penentuan estimator parameter β pada model regresi logistik dilakukan dengan menggunakan program C++. Sedangkan data yang digunakan pada penelitian ini adalah data bangkitan berjumlah 100.000 data pengamatan dengan 3 variabel prediktor yang terdiri dari satu variabel berskala rasio (X_1) , satu variabel berskala nominal 0,1 dan 2 (X_2) , dan yang terakhir berskala nominal 0 dan 1 (X_3). Penerapan program terhadap data bangkitan dengan 4 jenis kelompok data *squashing* menghasilkan nilai MSE semakin kecil seiring dengan membesarnya jumlah kelompok di data *squashing*. Sehingga dapat dikatakan bahwa secara umum estimator $\hat{\beta}$ dari data *squashing* dengan jumlah kelompok lebih banyak semakin mendekati nilai estimator $\hat{\beta}$ dari data induk.

Adapun saran yang ingin disampaikan adalah : pembahasan ini dapat dikembangkan : pertama, untuk metode *Squashing* dengan jumlah variabel prediktor yang berskala rasio lebih dari satu. Kedua, perlu dibuatkan program untuk meng-*update* data *squashing* baik jika ingin menambah jumlah kelompok maupun kalau ada data baru yang masuk dapat di *update* secara otomatis tanpa harus menghitung dari awal.

Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Airlangga, Nomor Kontrak : 615 / JO3.2 / PG / 2006, Tanggal : 7 Juni 2006.

SUMMARY

DETERMINATION OF THE PROBABILITY OF REMOVING CUSTOMER USING LOGISTIC REGRESSION MODEL FOR LARGE DATA, Eto Wuryanto, Dyah Herawatie and Rimuljo Hendradi, 2006, 32 pages.

Logistic regression modeling is difficult to be done if it is related to a massive dataset which basically is a large data. So, it is necessary to make sample from the main data. The previous researchs of squashing data by Eto et al (2005) and Rimuljo et al (2005) represent that Squashing is a good methods to reduce data.

In getting squashing data, there is two difficult phases that is : First, grouping or making partition to main data. Although the previous researchs have studied squashing data but only include one variable. While using logistic regression model need two variables or more in the data. Therefore the problem will be happened to get the appropriate partition, how to make partition for two variables or more? Second, what methods can be utilized to calculate the value of pseudo point and weighted for each partition? After squashing data is obtained, how to apply the the squashing data on logistics regression model and estimate its parameter?

This research aim to get the squashing data with two variable or more and determine the estimator parameter of the logistics regression model for the large data by using the C++ program as the early step in making of data mining software.

To obtain the squashing data can be done by : First, sort the data based on dependent variable so that is got a data group with the dependent variable zero and the other with the dependent variable one. Second, for each dependent variable sort the independent variable pursuant to variable that has non ratio scale. Third, do ascending sort to the values of ratio variable. Fourth, choose the group having biggest member then divide it into two group and do this division process so that is obtained the certain number of group. Fifth, compute the mean value and the number of member for every group which is each representing new observation value (pseudo data point) and weighted. On the other hand, to estimate the parameter of

logistics regression model is used the MLE method and if the result of the methods is an implicit function then Newton-Raphson method can be utilized.

Implementation of the squashing's algorithm and the estimation of parameter of logistics regression model's algorithm is done by writing it to C++ program. The data that is taken in this research is a generated data with 100.000 rows. The data contain three independent variable consist of one variable with ratio scale (X_1), one variable with nominal scale 0,1,2 (X_2) and the last with nominal scale 0,1 (X_3). Application of the program to the data for four kinds of group yield that increasing of number of group cause decreasing of the MSE value of squashing data. On the other word, the value of estimator $\hat{\beta}$ of squashing data with the largest number of element will approach the one of main data.

Any advice for the next research can do any work as the below : First, you are able to use squashing methods for two or more variable with ratio scale. Second, add a program for squashing data upgrading : add a new number of groups and entering a new observation that is automatically updated without doing computation from zero.

Mathematics Department, Faculty of Mathematics and Sciences, Airlangga University, Contract's Number : 615/JO3.2/PG/2006, Date : 7 Juni 2006.