PREDIKSI CHURN PELANGGAN TELEKOMUNIKASI SELULAR MENGGUNAKAN METODE K NEAREST NEIGHBOR

Eka Kartika Kusumaningdewi[1], Moch. Arif Bijaksana[2], Zk. Abdurahman Baizal[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Churn prediction merupakan salah satu jenis task data mining, yaitu klasifikasi, yang bertujuan untuk memprediksi pelanggan potensial churn pada industri telekomunikasi selular. Permasalahan prediksi churn adalah imbalance class, dimana terjadi ketidakseimbangan tajam antara jumlah suatu kelas dengan jumlah kelas lainnya pada data training. Classifier cenderung mengasumsikan data dalam kondisi balance sehingga mengakibatkan pembiasan prediksi kelas minor ke kelas mayor dan kemungkinan menganggap kelas minor hanya sebagai outlier (untuk data minor dengan jumlah sangat kecil).

Tugas Akhir ini akan mengimplementasikan metode K-Nearest Neighbor untuk imbalance class yaitu modifikasi metode K-Nearest Neighbor dengan berusaha lebih 'memihak' pada kelas minor. K-Nearest Neighbor akan membentuk center dinamis sesuai distribusi kelas pada proses learning, dan memilih K buah tetangga terdekat diutamakan dari satu cluster terdekat. Dengan modifikasi ini, diharapkan akan meningkatkan akurasi prediksi untuk kelas minor tanpa mengorbankan prediksi untuk kelas mayor. Akurasi hasil klasifikasi dari KNearest Neighbor untuk imbalance class akan dibandingkan dengan hasil klasifikasi oleh classifier lazy learner IBk pada tool Weka 3.5.6 serta dengan classifier populer lain dari tools Clementine 10.1 dalam bentuk top decile lift, lift curve dan gini coefficient.

Kata Kunci : churn prediction, imbalance class, K-Nearest Neighbor

Abstract

Churn prediction represents one type of data mining's task, called classification, that aim to predict potential churn customer at cellular telecommunication industry. Problem in churn prediction is imbalance class, where the distribution of training data isn't balance, sum of one class is much greater than another class. Classifier tends to assume that data is in balance condition, so it will bias the prediction for minority class belong to prediction for majority class, and possibly will judge minority class only as outlier (for minority class in too litte amount).

Churn prediction will be solved by implementing K-Nearest Neighbor for imbalance class method which modify K-Nearest Neighbor in order to more consider to minority class. At learning process, K-Nearest Neighbor will build dynamic cluster according to its class distribution and choose its K-Nearest Neighbor priority from the same nearest cluster. Wish this modification able to increase minority class prediction accuracy without decreasing majority class prediction accuracy. Classification result accuracy from K-Nearest Neighbor for imbalance class will be compared with the classfication result from lazy learner classifier Ibk on Weka 3.5.6 tool and another populer classifier on Clementine 10.1 tool in top decile lift, lift curve also gini coefficient forms.

Keywords : churn prediction, imbalance class, K-Nearest Neighbor

1. Pendahuluan

1.1 Latar Belakang Masalah

Industri penyedia jasa telekomunikasi merupakan industri yang terus berkembang dan selalu dibutuhkan masyarakat. Dengan semakin banyaknya jumlah perusahaan telekomunikasi baik penyedia layanan GSM (*Global System Mobile*) maupun CDMA (*Code Division Multiple Access*), masing-masing akan saling menerapkan strategi untuk memperebutkan perhatian pelanggan dan menguasai pasar. Berbagai cara dilakukan dalam mendukung strategi tersebut, seperti: penerapan tarif murah, penyediaan layanan/ fitur khusus kepada pelanggan, undian berhadiah, bonus pulsa, jaminan minimalisasi *call drop*, ataupun lainnya. Semua hal tersebut bertujuan untuk mempertahankan atau menambah *revenue* yang didapat oleh perusahaan, serta landasan bahwa biaya untuk mempertahankan pelanggan akan lebih murah dibandingkan biaya untuk menarik pelanggan baru [18].

Churn lahir dari fenomena yang terjadi di atas. Pengertian *churn* adalah pemutusan jasa suatu perusahaan oleh pelanggan karena pelanggan tersebut lebih memilih menggunakan layanan jasa perusahaan kompetitor [5,6]. *Churn* sangat berpotensi terjadi pada operator telekomunikasi selular karena kemudahan untuk berganti layanan dari satu operator ke operator yang lainnya [5]. *Churn* harus diwaspadai oleh perusahaan karena dengan bertambahnya jumlah *churn* akan semakin mengakibatkan penurunan *revenue* perusahaan.

Prediksi *churn* (*Churn Prediction*) merupakan salah satu jenis *task* pada *data mining*, yaitu klasifikasi. *Data mining* sendiri merupakan ilmu yang berkembang akibat semakin menumpuknya kuantitas data dan diperlukan suatu teknik untuk mendapatkan informasi dari tumpukan data tersebut. *Churn* harus bisa diprediksi oleh perusahaan dalam rangka penerapan strategi untuk mempertahankan pelanggan yang potensial *churn* dan agar pemberian retensi atau wujud layanan tambahan lainnya dapat tepat pada sasaran. Kebutuhan aplikasi prediksi *churn* bagi perusahaan telekomunikasi yaitu mampu menghasilkan daftar nama-nama pelanggan potensial *churn* diurutkan secara *descending* berdasarkan nilai *confidence*-nya (bobot *churn*). Adapun parameter *predictors* yang digunakan yaitu data demografis, data *traffic* (data *Call Detail Record*), dan data *payment* (pembayaran) [7].

Permasalahan prediksi *churn* yang akan dipaparkan adalah *imbalance/ unbalance class*. *Imbalance class* merupakan salah satu masalah yang vital dalam klasifikasi dimana pada *training set* terdapat ketidakseimbangan yang tajam tentang jumlah *record* suatu kelas dengan *record* pada kelas lainnya. Kelas dengan jumlah *record* lebih besar disebut dengan kelas mayor, sedangkan kelas dengan jumlah *record* lebih kecil disebut dengan kelas minor. *Imbalance* menjadi sebuah masalah karena prediksi untuk kelas minor lebih menarik daripada prediksi untuk kelas mayor. Dari banyak penelitian didapat kesimpulan bahwa pada kondisi data *imbalance*, *classifier* cenderung membias *record* yang seharusnya

merupakan kelas minor menjadi kelas mayor dan kemungkinan menganggap kelas minor hanya sebagai *outlier* (untuk data minor dengan jumlah sangat kecil).

Metode yang akan digunakan adalah *K-Nearest Neighbor* untuk *imbalance class* yaitu modifikasi metode *K-Nearest Neighbor* yang berusaha lebih 'memihak' pada kelas minor. *K-Nearest Neighbor* dianggap sebagai salah satu algoritma yang ampuh dalam mengatasi permasalahan klasifikasi. Hal ini dikarenakan objek diklasifikasikan sesuai label kelas terbanyak yang terdapat pada area k buah tetangga terdekat (k buah *data learning*), sehingga bersifat *local approximation*. Penentuan tetangga terdekat dihitung dengan rumus perhitungan jarak antar objek tertentu seperti *Euclidean Distance*. Dengan modifikasi KNN yang dilakukan, diharapkan akan memperbaiki akurasi dari prediksi pada data minor terutama jika dibandingkan dengan penggunaan *classifier lazy learner* IBk (nama lain KNN) pada *tool* Weka 3.5.6 serta *classifier* populer lain dari *tool* Clementine 10.1.

1.2 Perumusan Masalah

Dengan mengacu pada latar belakang masalah diatas, maka permasalahan yang akan dibahas dan diteliti adalah :

1. Bagaimana menganalisa dan mengimplementasikan metode *K-Nearest Neighbor* untuk *imbalance class* dalam kasus *churn prediction*.
2. Bagaimana akurasi yang didapat dari *K-Nearest Neighbor* untuk *imbalance class* jika dibandingkan dengan *classifier lazy learner* IBk pada *tool* Weka 3.5.6 serta dengan *classifier* populer lain dari *tool* Clementine 10.1 dalam bentuk *top decile lift*, *lift curve* dan *gini coefficient*.

Batasan masalah yang akan dibahas dalam penelitian tugas akhir ini adalah :

1. *Preprocessing* berupa *filter nominal(*kategoris*) to binary* dan normalisasi ditangani oleh sistem, sedangkan *preprocessing* lainnya menggunakan Weka 3.5.6, Clementine 10.1 dan Microsoft Excel.
2. *Postprocessing* menggunakan Weka 3.5.6 dan Microsoft Excel.
3. Tidak membahas penggunaan *tool* Clementine 10.1 dan *Weka 3.5.6* dalam pengujian sebagai pembanding terhadap akurasi dari perangkat lunak yang dihasilkan.

1.3 Tujuan

Berdasarkan rumusan masalah di atas, maka tujuan dari tugas akhir ini adalah:

1. Mengimplementasikan metode *K-Nearest Neighbor* untuk *imbalance class* dalam kasus *churn prediction*.
2. Menganalisis akurasi hasil klasifikasi yang dihasilkan oleh perangkat lunak *Churn Prediction* menggunakan algoritma *K-Nearest Neighbor* untuk *imbalance class*.
3. Membandingkan akurasi hasil klasifikasi dari algoritma *K-Nearest Neighbor* untuk *imbalance class* dengan hasil klasifikasi oleh *classifier*

lazy learner IBk pada *tool* Weka 3.5.6 serta dengan *classifier* populer lain dari *tool* Clementine 10.1 dalam bentuk *top decile lift, lift curve* dan *gini coefficient*.

1.4 Metodologi Penyelesaian Masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah menggunakan metode studi pustaka atau studi literatur dan analisis dengan langkah kerja sebagai berikut :

1. Studi Literatur :
   a. Pencarian referensi yang layak dan berhubungan dengan *imbalance class problem, churn prediction,* dan *K-Nearest Neighbor*.
   b. Pendalaman materi, mempelajari dan memahami materi yang berhubungan dengan tugas akhir.
   c. Pencarian manual dan referensi tentang penggunaan *tool* Clementine 10.1 dan Weka 3.5.6.
2. Analisis Permasalahan :
   a. Mencari data pelanggan dan memahaminya.
   b. Mempelajari konsep dari *K-Nearest Neighbor* untuk *imbalance class* yang akan digunakan dalam implementasi perangkat lunak.
   c. Menganalisis algoritma *K-Nearest Neighbor* untuk *imbalance class* dalam perancangan perangkat lunak
   d. Simulasi dan analisa data dengan *classifier lazy learner* IBk pada *tool* Weka 3.5.6 serta dengan *classifier* populer lain dari *tool* Clementine 10.1.
3. Mengumpulkan *requirement* terhadap perangkat lunak yang akan dibangun.
4. Melakukan disain/ perancangan perangkat lunak dengan teknik *object oriented*.
5. Melakukan implementasi perancangan perangkat lunak.
6. Melakukan pengujian perangkat lunak dengan memasukkan data yang sudah di-*preprocessing* serta menganalisis hasil keluaran program.
7. Menganalisis hasil prediksi dari perangkat lunak dengan hasil prediksi oleh *classifier lazy learner* IBk pada *tool* Weka 3.5.6 serta dengan *classifier* populer lain dari *tool* Clementine 10.1.
8. Pengambilan kesimpulan dan penyusunan laporan tugas akhir.

5. Penutup

5.1 Kesimpulan

1. KNN untuk *imbalance class* yang diimplementasikan terbukti baik dalam memprediksi banyaknya kelas minor pada permasalahan *imbalance class* data *churn* perusahaan telekomunikasi dan data *churn tournament*. Akan tetapi, KNN untuk *imbalance class* mempunyai kelemahan, yaitu banyak melakukan *misclassification* kelas mayor ke kelas minor sehingga menurunkan akurasi prediksi kelas mayor.
2. Nilai *top decile, lift curve*, dan *gini coefficient* untuk evaluasi *churn prediction* tidak mempunyai hubungan *dependency* dengan nilai K. Oleh karena itu, untuk mendapatkan nilai *top decile, lift curve,* dan *gini coefficient* terbesar, satu-satunya cara adalah mencoba satu persatu parameter nilai K.
3. Penggunaan KNN untuk *imbalance class* tidak berhasil memberikan peningkatan akurasi untuk permasalahan *churn prediction.*
4. Pada data perusahaan telekomunikasi, IBk memberikan hasil terbaik dibanding KNN untuk *imbalance class, Neural Network, Chaid, Oversampling + C5.0* dan *SVM Imbalance Gaussian*. Sedangkan pada data *churn tournament, Neural Network* memberikan hasil terbaik dibanding IBk, KNN untuk *imbalance class, Chaid* dan *Oversampling + C5.0.*

4.2 Saran

1. Mengembangkan metode *K-Nearest Neighbor* untuk *imbalance class* yang mampu menempatkan *actual churn* pada prosentase *customer* lebih awal terutama 10% *customer* dan meminimalkan pembiasan prediksi kelas mayor ke kelas minor.
2. Karena *K-Nearest Neighbor* hanya menerima atribut dengan tipe data numerik, maka *preprocessing* untuk atribut dengan tipe data nominal harus dilakukan secara sangat hati-hati dan penuh pemahaman agar hasil prediksinya baik.

Daftar Pustaka

[1] Abe, Naoki. *Sampling Approach to Learning from Imbalanced Data Sets*. IBM T.J Watson Research Center USA. 2003

[2] Barandela R, J.S. Sanchez, V. Garcia, and E. Rangel. *Strategies for Learning in Class Imbalance Problems*. Instituto Technologico de Toluca, Mexico and Department de Lienguatges i sistemes informatics, Universitat Jaume, Castellon, Spain.

[3] Batista, Gustavo E.A. *A Study of The Behavior of Several Methods for Balancing Machine Learning Training Data*. University of Ottawa

[4] Berson, Alex. Stephen Smith and Kurt Thearling. *An Overview of Data Mining Techniques*. 2005. http://www.thearling.com

[5] Cardell, N Scott. Mikhail Golovnya and Dan Steinberg. *Churn Modelling for Mobile Telecommunications: Winning the Duke/ NCR Teradata Center for CRM Competition.* Salford Sysem. 2003. http://www.salford-systems.com

[6] Euler, Timm. *Churn Prediction in Telecommunications Using Mining Mart.* Computer Science VIII, University of Dortmund, D-44221 Dortmund, Germany. 2005. http://www-ai.cs.uni-dortmund.de

[7] Fahrudin, Tora. *Analisis dan Implementasi Metode Databoost-IM (Studi Kasus Churn Prediction Mobile Telecommunication).* Buku TA. STT Telkom Bandung. 2007

[8] Guo, Hongyu and Herna L Viktor. *Learning from Imbalance Data Sets with Boosting and Data Generation : The DataBoost-IM Approach*. School of Information Technology and Engineering, University of Ottawa, Canada

[9] Hand, David and Veronica Vinciotti. *Choosing K for Two-Class Nearest Neighbor Classifiers with Unbalanced Classes.* Department of Mathematics, Imperial College, UK. 2002. http://www.ComputerScienceWeb.com

[10] Japkowich, Nathalie. *Learning from Imbalance Data Sets : A Comparison of Various Straregies*. Faculty of Computer Science, Daltech/ Dalhousie University, 6050 University, Canada

[11] Jiawei Han and Micheline Kamber. *Data Mining : Concepts and Techniques.* Intelligent Database Systems Research Lab, School of Computing Science, Simon Fraser University

[12] Lazarevic Aleksandar, Arindam Banerjee, Varun Chandola, Vipin Kumar and Jaideep Srivastava. *Data Mining for Anomaly Detection.* Tutorial at the European Conference on Principles and Practice of KDD. Atwerp, Belgium.

[13] Lemmend, Aurelié and Christopher Croux. *Bagging and Boosting Classification Tress to Predict Churn.* Department of Applied Economics, K.U. Leuvan, Belgium

[14] Mennicke, Jörg. *Classifier Learning for Imbalance Data with Varying Misclasification Costs, A Comparison of KNN, SVM, and Decision Tree Learning*. University of Bamberg in cooperation with Fraunhofer Institute for Integrated Circuits (IIS), Erlangen. 2006

[15] Mitchell, T. *Machine Learning.* Mac Graw Hill. 1997

[16] Pang-Ning Tan. Michael Steinbach and Vipin Kumar. *Introduction to Data Mining.* University of Minnesota and Army High Performance Computing Research Center

[17] Prihandini, Quvin Nola. *Analisis dan Implementasi Churn Prediction Menggunakan Algoritma Genetika (Studi Kasus Pelanggan Flexi PT. Telkom).* Buku TA. STT Telkom. 2007

[18] Richeldi, M., Perrucci, A.: "*Mining Mart Evaluation Report*". Deliverable D17.3, IST Project MiningMart, IST-11993 (2002).

[19] Sastrawan, Angelina Sagita. *Analisis Pengaruh Sampling dalam Churn Prediction.* Buku TA. IT Telkom. 2008

[20] Teknomo, Kardi, PhD. *K-Nearest Neighbor Tutorial.* http://people.revoledu.com/kardi/tutorial/KNN/index.html

[21] Teknomo, Kardi, PhD. *Similarity Measurement.* http://people.revoledu.com/kardi/tutorial/Similarity/NominalVariables.html

[22] Ye, Nong and Xiangyang Li. *A Scalable Clustering Technique for Intrusion Signature Recognition.* Proceeding of the 2001 IEEE, Workshop on Informastion Assurance and Security, United States Military Academi, West Point, NY. 2000

[23] Yunanto, Ikhwan. *Metode Support Vector Machine dengan Pruning pada Support Vector dalam Churn Prediction pada Telekomunikasi.* Buku TA. IT Telkom. 2009