PREDIKSI PELANGGAN 2G YANG BERPOTENSI BERALIH KE LAYANAN 3G PADA JARINGAN TELEKOMUNIKASI BERGERAK DENGAN METODE SUPPORT VECTOR MACHINE (STUDI KASUS PAKDD 2006 KOMPETISI DATA MINING)
PREDICTION OF 2G CUSTOMER WHICH POTENCY TO SWITCH

Eko Setiawan Rufiantino[1], Moch. Arif Bijaksana[2], Dade Nurjanah[3]

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

Abstrak

Perkembangan teknologi telekomunikasi saat ini berkembang semakin pesat. Perusahaan-perusahaan yang bergerak di bidang jasa telekomunikasi juga mulai mengembangkan jaringan telekomunikasi generasi ke-tiga (3G) untuk memanjakan para pelanggannya dengan fitur-fitur yang lebih canggih. Seiring dengan kemajuan teknologi, kebutuhan akan penyimpanan data juga semakin besar. Namun, data yang besar tidak akan berguna jika informasi yang terkandung di dalamnya tidak diketahui. Oleh karena itu digunakan teknik data mining untuk mengakses informasi tersebut. Dalam tugas akhir ini, data mining digunakan untuk memprediksi para pelanggan 2G yang berpotensi beralih ke jaringan 3G. Metode yang digunakan dalam memecahkan permasalahan tugas akhir ini adalah dengan menggunakan SVM. Hal ini dikarenakan SVM merupakan salah satu metode unggulan dalam bidang pattern recognition. Di samping itu, SVM merupakan metode machine learning yang selalu berusaha menemukan hyperplane terbaik untuk memisahkan dua buah kelas pada input space.

Hasil pemodelan classifier yang diperoleh pada tahap training akan digunakan untuk memprediksi kelas pelanggan. Dalam hal ini, penggunaan kernel RBF mampu memberikan hasil yang optimal dibandingkan dengan kernel linear dan polynomial. Semakin besar peluang pelanggan 2G yang beralih ke jaringan 3G, maka pihak perusahaan pun akan semakin intensif dalam menembak target pasar yang potensial. Diharapkan semakin bertambahnya pelanggan 3G akan semakin meningkatkan profit bagi perusahaan.

Kata Kunci : data mining, feature selection, diskretisasi, SVM, PAKDD 2006 Kompetisi Data Mining

Abstract

The growth of telecommunication technology in this time are expands fast progressively. Companies, which are active in telecommunication service, also begin expand telecommunication network of third-generation (3G) to pampering its customers with more sophisticated features. Along with progress of technology, the requirement data repository is progressively big. But big data will not useful if the information which consists in data repository is unknown.

In this final project, data mining is used to predict 2G's customers which potency to switch 3G network. The method which used for solving this final project problem use SVM. This matter caused SVM is one of pre-eminent method in pattern recognition field. Besides, SVM is machine learning method which always trying to find the best of hyperplane to separate two classes at input space.

The result of classifier model which is obtained at training phase will be used for predicting customer class. In this case, usage of RBF kernel can give optimal result rather than linear or polynomial kernel. More greater of opportunity of 2G customers who switch to 3G network, hence the company even will be intensive progressively in shooting potential market goals. It is expected progressively increasing of 3G customers will improve the profit of company.

Keywords : data mining, feature selection, discretization, SVM, PAKDD 2006 Data Mining Competition

1. Pendahuluan

1.1 Latar Belakang

Perkembangan sistem komunikasi pada jaringan bergerak telah berkembang semakin pesat seiring dengan kemajuan tekonologi komunikasi itu sendiri. Terdorong oleh pertumbuhan kebutuhan para pelanggannya, membuat para *provider* layanan telekomunikasi melakukan berbagai macam terobosan dengan fitur-fitur dan fasilitas lainnya yang memanjakan para pelanggannya. Kesuksesan peluncuran sistem komunikasi 2G telah mendorong penyebaran dan penyempurnaan kapabilitas untuk memenuhi harapan para pelanggan untuk mendapatkan layananan informasi yang lebih canggih. Kecanggihan akan fitur-fitur yang telah dikembangkan pada jaringan 2G terus dikembangkan hingga saat ini telah memasuki era 3G (*Third Generation*). Jaringan 3G akan membawa perubahan secara revolusioner pada para pengguna *mobile phone* yang sebelumnya berbasiskan layanan suara menjadi layanan yang berbasis multimedia dan akan digunakan secara luas di dunia telekomunikasi.

Munculnya teknologi komunikasi generasi ke-tiga ini diharapkan dapat meningkatkan skala bisnis bagi para *provider* layanan komunikasi dengan tujuan untuk meningkatkan profit perusahaan. Beberapa langkah yang bisa diterapkan dalam mencapai tujuan tersebut adalah : (1) Perumusan target, yaitu memilih pelanggan-pelanggan potensial untuk beralih ke jaringan 3G yang menjadi target pemasaran. Peningkatan profit perusahaan diharapkan semakin besar apabila jumlah pelanggan jaringan 3G semakin meningkat. (2) Estimasi dan prediksi. Estimasi adalah menerka sebuah nilai yang belum diketahui berdasarkan informasi yang telah tersedia dan prediksi adalah memperkirakan nilai estimasi tersebut untuk masa yang akan datang.

Untuk mendukung solusi-solusi tersebut maka digunakan teknologi *data mining*. Dengan menggunakan teknologi ini perusahaan akan lebih intensif untuk menembak target pasar yang potensial. Dengan demikian diharapkan semakin besar peluang para pelanggan 2G yang akan beralih ke layanan 3G semakin besar pula profit yang akan diperoleh perusahaan. Oleh karena itu, dalam tugas akhir ini akan dibuat suatu implementasi perangkat lunak yang dapat memperkirakan para pelanggan 2G yang potensial beralih ke layanan jaringan 3G pada studi kasus PAKDD *(Pasific-Asia Conference on Knowledge Discovery and Data Mining)* 2006 Kompetisi Data Mining.

1.2 Perumusan Masalah

Tujuan utama dalam pengerjaan tugas akhir ini adalah memprediksi pelanggan 2G yang potensial untuk beralih ke layanan 3G. Rumusan masalah untuk tugas akhir ini akan dilakukan, yaitu (1) menentukan profil pelanggan berdasarkan kelas 2G atau 3G dan (2) memprediksi pelanggan 2G yang potensial berdasarkan pemodelan *classifier* yang telah diperoleh pada saat pelatihan sistem.

Namun, masalah yang dihadapi untuk fase ini adalah dimensi *dataset* pelanggan cukup tinggi dan tidak efisien jika seluruh atribut digunakan untuk memprediksi. Untuk mengatasi hal ini, yang pertama dilakukan adalah melakukan pencarian atribut yang relevan dalam menentukan kelas pelanggan dengan teknik *feature selection*. Kemudian untuk melakukan klasifikasi dan prediksi pelanggan akan dilakukan dengan metode *Support Vector Machine* (SVM). Hasil klasifikasi berupa *false positive* akan dianggap sebagai pelanggan 2G yang potensial. Dengan mengacu pada latar belakang masalah di atas, maka permasalahan yang akan dibahas dan diteliti adalah :

1. Bagaimana melakukan *feature selection* yang relevan terhadap label kelas dengan *dataset* yang berdimensi tinggi
2. Bagaimana menerapkan metode *Support Vector Machine* sebagai *classifier* untuk melakukan klasifikasi terhadap *dataset* pelanggan sehingga menghasilkan kelas pelanggan 2G dan 3G dengan akurat.
3. Bagaimana menentukan parameter-parameter masukan SVM agar hasil klasifikasi yang diperoleh hasil yang optimal.

1.3 Tujuan

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

1. Menganalisis data pelanggan yang terdiri dari data yang telah memiliki kelas (data *training*) dan data yang belum memiliki kelas (data *testing*).
2. Menerapkan metode *feature selection* untuk tahap *pre-process* untuk menangani dimensi data yang tinggi.
3. Menerapkan metode SVM sebagai *classifier* untuk klasifikasi serta prediksi.
4. Melakukan perbandingan akurasi *classifier* terhadap penggunaan *kernel linear, polynomial*, dan RBF.
5. Menghasilkan *list* prediksi pelanggan mana saja yang termasuk ke dalam kelas positif (3G) dan kelas negatif(2G).

1.4 Batasan Masalah

Dalam Tugas Akhir ini, yang akan dibahas adalah suatu implementasi untuk menemukan pola yang dapat memprediksi para pelanggan 2G yang berpotensi menjadi pelanggan 3G dengan batasan masalahnya sebagai berikut :

1. Sistem yang akan dirancang hanya untuk studi kasus PAKDD 2006 Kompetisi *Data Mining*.
2. Untuk kasus yang lain, sistem bisa digunakan jika memiliki bentuk *dataset* yang sama dengan *dataset* PAKDD 2006 Kompetisi *Data Mining*.
3. *Kernel* yang digunakan pada SVM meliputi *kernel linear*, *polynomial*, dan RBF.
4. Melakukan perbandingan hasil klasifikasi dengan pemenang PAKDD yang menggunakan metode SVM berdasarkan *balanced accuracy* (secara kuantitatif).

5. Jumlah *record* data yang digunakan sebesar 4000 record yang terdiri dari 3000 *record* data *training* dan 1000 *record* data *testing*.

1.5 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
      Mencari referensi dan sumber-sumber lain yang layak yang berhubungan dengan *data mining*, *feature selection*, klasifikasi dengan metode *Support Vector Machine*, dan pemahaman akan tipe pelanggan 2G dan 3G itu sendiri.
   b) Pendalaman materi
      Mempelajari dan memahami materi yang berhubungan dengan tugas akhir ini.
2. Pencarian dan pengumpulan data. Mengumpulkan data yang berupa data profil pelanggan dengan label kelas 2G dan 3G untuk *training* serta data yang belum memiliki label kelas yang akan diklasifikasi dan diprediksi.
3. Mempelajari konsep dari *feature selection* dan *Support Vector Machine* yang akan digunakan dalam implementasi perangkat lunak.
4. Melakukan analisa terhadap *feature selection*, dan *Support Vector Machine* dalam perancangan perangkat lunak.
5. Melakukan implementasi perancangan perangkat lunak dengan menggunakan Matlab 7 dan Microsoft SQL Server 2000 untuk aplikasi *database*.
6. Melakukan proses pengujian perangkat lunak dengan skenario sebagai berikut, yaitu :
   - Melakukan pelatihan agar model sistem yang telah dirancang mampu memberikan hasil yang optimal dan stabil.
   - Tahap pelatihan dilakukan dengan cara membagi data yang telah memiliki label menjadi dua bagian, yaitu sebagai data untuk *training* dan *testing*.
   - Melakukan proses klasifikasi untuk data pelanggan yang belum memiliki kelas.
   - Melakukan prediksi terhadap data pelanggan yang potensial untuk beralih ke jaringan 3G.
7. Mencatat hasil keluaran program.
   Pengambilan kesimpulan dan penyusunan laporan tugas akhir.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Permasalahan klasifikasi *non-linear* dapat dipecahkan dengan menggunakan *kernel trick*.
2. Besar *margin* dipengaruhi oleh nilai C. Semakin besar nilai C maka *margin* yang dihasilkan oleh *classifier* akan semakin kecil. Nilai *margin* yang kecil akan membuat *classifier* semakin fleksibel dan mampu menampung lebih banyak data *training*. Namun, error generalisasi yang ditimbulkan juga akan semakin besar. Sebaliknya *classifier* dengan *margin* yang besar akan mengurangi error generalisasi, tetapi data *training* yang ditampung akan semakin sedikit dan bisa menimbulkan *overfitting*. Oleh karena itu nilai C digunakan untuk mengontrol *tradeoff* antara *margin* dan error generalisasi.
3. Secara umum penggunaan SVM dengan kernel RBF mampu bekerja dan memberikan hasil performansi yang lebih baik dengan mengatur *setting* parameter masukan dengan tepat. Dalam hal ini, akurasi classifier terbaik adalah sebesar 74.78% dengan menggunakan kernel RBF dengan nilai sigma=0.1 dan C=0.1 serta jumlah atribut yang digunakan sama dengan 4. Kesalahan penggunaan parameter masukan pada kernel RBF justru akan menghasilkan performansi yang paling buruk dibandingkan dengan penggunaan *kernel linear* dan *polynomial*.
4. Beberapa pelanggan 3G memiliki karakteristik sebagai pelanggan 2G atau sebaliknya beberapa pelanggan 2G memiliki karakteristik sebagai pelanggan 3G. Karakteristik yang dimiliki masing-masing kelas pelanggan ini menyebabkan *classifier* salah dalam memprediksi pelanggan. Dalam hal ini, yang dianggap sebagai pelanggan potensial adalah pelanggan 2G yang memiliki karakteristik sebagai pelanggan 3G. Pada sistem, pelanggan 2G potensial ini ditandai sebagai data *false positive*.
5. Akurasi terbaik yang diperoleh pada pengujian *classifier* menunjukkan bahwa sampel *dataset* yang digunakan sebesar 4000 *record* cukup representatif terhadap dataset asli. Hal ini bisa dilihat dari rata-rata perolehan akurasi yang dicapai oleh kontestan PAKDD 2006 dengan menggunakan metode SVM adalah sebesar 74.97%. Sedangkan akurasi tertinggi yang diperoleh penulis adalah sebesar 74.78% hampir mendekati nilai rata-rata akurasi yang dicapai oleh kontestan PAKDD 2006.

5.2 Saran

1. Untuk penelitian selanjutnya *feature selection* yang digunakan pada tahap *pre-process* menggunakan algoritma yang secara otomatis menentukan sendiri jumlah atribut yang terpilih.

2. Menyempurnakan metode SVM agar keseluruhan data yang digunakan untuk studi kasus PAKDD 2006 ini bisa diujikan.
3. Mencoba menggunakan bahasa pemrograman lain selain Matlab untuk mengatasi keterbatasan memori yang digunakan.

Daftar Pustaka

[1] Alexey Tsymbal, Mykola Pechenizkiy, and Pádraig Cunninghamm. *Diversity in Ensemble Feature Selection*. Department of Computer Science, Trinity College Dublin, Ireland.

[2] Burges, Christopher J.C. *A Tutorial on Support Vector Machine for Pattern Recognition*, Data Mining and Knowledge Discovery, 2, Kluwer Academic Publisher, Boston, 1998, pp. 121-167.

[3] Chapman, Sam. *Support Vector Machines*. February 13, 2004.

[4] Gunn, Steve R. *Support Vector Machines for Classification and Regression*. Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton, May 1998.

[5] Hand, David, et all. *Principles of Data Mining*. The MIT Press, Massachusetts, 2001.

[6] I. Kononenko and S.J. Hong. Future Generation Computer Systems, *Attribute Selection for Modelling*, November 1997.

[7] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[8] Kantardzic, Mehmed. *Data Mining : Concepts, Models, Methods, and Algortihm*. John Wiley & Sons, 2003.

[9] M. Dash and H. Liu. *Feature Selection for* Classification, Intelligent Data Analysis, 1, 1997, pp. 131-156.

[10] Mark A. Hall and Geoffrey Holmes. *Benchmarking Attribute Selection Techniques for Discrete Class Data Mining*. IEEE Transaction on Knowledge and Data Engineering, 15, no. 3, June 2003.

[11] Martin, Mario. *On-line Support Vector Machine for Function Approximation*. Software Department, Universitat Politecnica de Catalunya, Spain.

[12] Nugroho, Anto Satrio, 2003, Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika, http://www.IlmuKomputer.com

[13] PAKDD 2006 Data Mining Competition. http://www.ntu.edu.sg/sce/pakdd2006.

[14] Robert J. Vanderbei. 2000. *LOQO User's Manual-Version 4.05*. Princeton University School of Engineering and Applied Science: Department of Operations Research and Financial Engineering.

[15] Tan, Pang-Ning, et all. *Introduction to Data Mining*. Pearson Education, Inc., Boston, 2006.

[16] Webb, Andrew R. *Statistical Pattern Recognition*. 2nd ed., John Willey & Sons, Chichester, England, 2002.

[17] Zalán, Bodò. *Supervised Learning with Support Vector Machines*. Faculty of Mathematics and Computer Science, Babeş-Bolyai University.