

ANALISIS DAN IMPLEMENTASI TEKNIK FUZZY K-NEAREST NEIGHBOR PADA KASUS IMBALANCE CLASS

Yuke Cipta Mandiri Putri¹, Imelda Atastina², Intan Nurma Yulita³

¹Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Salah satu metode dalam klasifikasi yaitu K-Nearest Neighbor (K-NN) yang berusaha mengklasifikasikan data dengan memilih data sejumlah K yang letaknya terdekat dari data baru tersebut. Metode Fuzzy K-Nearest Neighbor (Fuzzy K-NN) merupakan metode klasifikasi yang menggabungkan teknik Fuzzy dan K-Nearest Neighbor. Pada tugas akhir ini akan diimplementasikan metode Fuzzy K-NN pada data yang imbalance yaitu pada kasus churn prediction. Hasil klasifikasi dari Fuzzy K-NN akan dihitung nilai performansinya berupa precision, recall, f-measure, lift curve dan dibandingkan dengan hasil klasifikasi menggunakan metode K-NN. Hasil dari penelitian ini menunjukkan bahwa metode Fuzzy K-NN lebih baik dibandingkan dengan K-NN walaupun masih terjadi misklasifikasi.

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

Abstract

One method of classification is K-Nearest Neighbor (K-NN) which attempted to classify the data by selecting K data that is closest to the new data. One variant of K-Nearest Neighbor is Fuzzy K-Nearest Neighbor (Fuzzy K-NN) which is a classification method that combines Fuzzy and K-Nearest Neighbor method. In this final project, Fuzzy K-NN will be implemented on imbalance data which is churn prediction. The performance value like precision, recall, f-measure, lift curve from Fuzzy K-NN will be calculated and compared with classification result using K-NN method. The result of this study indicates that the Fuzzy K-NN method is better than K-NN, although there are still misclassification occurred.

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

Telkom
University

1. Pendahuluan

1.1 Latar belakang

Imbalance class merupakan kondisi adanya ketidakseimbangan dalam jumlah data training antara dua kelas yang berbeda, salah satu kelasnya merepresentasikan jumlah data yang sangat besar (*majority class*) sedangkan kelas yang lainnya merepresentasikan jumlah data yang sangat kecil (*minority class*). *Imbalance class* ini biasanya ditemukan pada kasus-kasus anomali, misalnya: *fraud detection*, *churn prediction*, *network intrusion detection*, dan *thyroid disease* [7].

Imbalance class dalam data mining perlu dipelajari karena *minority class* lebih sulit untuk diprediksi daripada *majority class*. Padahal, terkadang kelas minoritas inilah yang lebih memiliki informasi yang berharga. Oleh karena itu, kita memerlukan algoritma yang dapat memprediksi label kelas yang tepat agar memperoleh hasil dengan nilai akurasi yang tinggi.

Salah satu task yang dapat dilakukan dalam data mining yaitu klasifikasi. Metode klasifikasi dapat digunakan untuk memprediksi kasus-kasus *imbalance class*. Dalam proses klasifikasi yang dilakukan adalah mencari nilai dari salah satu atribut dari data yang tersedia yang disebut atribut *class*. Nilai dari atribut *class* yang akan kita cari dalam proses klasifikasi ini dapat diperoleh dengan mengacu pada nilai-nilai atribut *class* dari data yang telah diketahui nilai-nilai atribut *class*-nya, data ini disebut *training set*.

Ada beberapa metode yang digunakan dalam proses klasifikasi, salah satu metode dalam klasifikasi yaitu K-Nearest Neighbor atau yang disingkat K-NN yang merupakan metode pengklasifikasian data yang bekerja relatif dengan cara yang lebih sederhana dibandingkan dengan metode pengklasifikasian data lainnya. Algoritma ini berusaha mengklasifikasikan data baru yang belum diketahui class-nya dengan memilih data sejumlah K yang letaknya terdekat dari data baru tersebut. Class terbanyak dari data terdekat sejumlah K tersebut dipilih sebagai class yang diprediksikan untuk data yang baru [6].

Salah satu varian dari algoritma k-nearest neighbor yaitu Fuzzy K-Nearest Neighbor atau yang disebut Fuzzy K-NN yang merupakan metode klasifikasi yang menggabungkan teknik fuzzy dan K-Nearest Neighbor classifier. Fuzzy K-Nearest Neighbor memiliki dua keunggulan utama daripada algoritma K-NN [1]. Pertama, algoritma ini mampu mempertimbangkan sifat ambigu dari tetangga jika ada. Algoritma ini telah dirancang sedemikian rupa agar tetangga yang ambigu tidak memainkan peranan penting dalam klasifikasi saat ini. Keunggulan kedua yaitu sebuah *instance* akan memiliki derajat nilai keanggotaan pada setiap kelas sehingga akan lebih memberikan kekuatan atau kepercayaan suatu *instance* berada pada suatu kelas.

Dari penelitian yang telah dilakukan sebelumnya, telah terbukti bahwa metode K-NN untuk *imbalance class* baik dalam memprediksi banyaknya kelas minor pada permasalahan *imbalance class*, tetapi melakukan misclassification (pembiasan) kelas

mayor ke kelas minor yang menyebabkan nilai akurasi kelas mayor menurun [7]. Sehingga diharapkan dengan penggunaan metode Fuzzy K-NN dengan beberapa keunggulannya ini mampu mengatasi kekurangan yang dihadapi metode K-NN pada kasus *imbalance class*.

1.2 Perumusan masalah

1. Bagaimana menerapkan Fuzzy K-Nearest Neighbor pada *imbalance class*?
2. Seberapa besar peningkatan akurasi dalam klasifikasi menggunakan teknik Fuzzy pada K-Nearest Neighbor classifier?
3. Bagaimana melakukan pengujian dan analisis dari implementasi teknik Fuzzy K-Nearest Neighbor dalam klasifikasi?

Batasan Masalah :

1. Data training sudah mengalami data preprocessing
2. Data yang digunakan untuk analisis merupakan data yang supervised (memiliki class label)
3. Dataset yang digunakan merupakan data *imbalance class*
4. Evaluasi diperoleh berdasarkan nilai *recall*, *precision*, dan *f-measure*
5. Data yang digunakan yaitu data churn perusahaan telekomunikasi

1.3 Tujuan

1. Mengimplementasikan metode Fuzzy K-NN pada kasus *imbalance class*
2. Mengetahui besarnya peningkatan akurasi dalam klasifikasi menggunakan metode Fuzzy K-NN
3. Melakukan pengujian dan analisis dari implementasi teknik Fuzzy K-NN

1.4 Metodologi penyelesaian masalah

1. Studi literatur
Pencarian referensi, mencari referensi dan sumber-sumber lain yang layak yang berhubungan dengan data mining, *imbalance class*, dan Fuzzy K-NN
2. Pencarian dan pemahaman data
Mencari data. Setelah data didapatkan maka langkah selanjutnya adalah data understanding yaitu memahami data yang diperoleh
3. Analisis dan perancangan perangkat lunak
Menganalisis permasalahan yang akan diselesaikan dan menganalisis tahapan-tahapan yang diinginkan untuk menyelesaikan permasalahan dengan metode prosedural
4. Implementasi sistem
Melakukan implementasi sistem dengan membangun perangkat lunak sesuai dengan perancangan yang telah dilakukan
5. Pengujian sistem dan analisa hasil

Melakukan proses pengujian terhadap keakuratan hasil dan performansi algoritma Fuzzy K-NN, serta melakukan analisis terhadap kelebihan dan keterbatasan algoritma Fuzzy K-NN pada kasus *imbalance class*

6. Pengambilan keputusan dan penyusunan tugas akhir
Pengambilan kesimpulan berdasarkan pengujian dan analisis disertai dengan penyelesaian laporan tugas akhir.



5. Penutup

5.1 Kesimpulan

1. Fuzzy K-NN pada umumnya memberikan hasil klasifikasi yang lebih baik dibandingkan dengan metode K-NN ditunjukkan dengan nilai performansi berupa *precision*, *recall*, dan *f-measure* yang lebih besar daripada *precision*, *recall*, dan *f-measure* K-NN
2. Nilai K paling optimal untuk metode K-NN maupun Fuzzy K-NN adalah $K=1$, hal ini disebabkan karena semakin besar nilai K, akan semakin banyak kelas mayor yang sebenarnya bukan merupakan tetangga terdekat yang mempengaruhi kelas minor, sehingga kelas minor akan memiliki peluang yang besar untuk salah diklasifikasikan.
3. Tingkat *imbalance* data mempengaruhi hasil klasifikasi pada metode K-NN maupun Fuzzy K-NN, semakin tinggi tingkat *imbalance* sebuah data, maka nilai *precision*, *recall*, dan *f-measure* pada kelas mayor semakin meningkat sedangkan untuk nilai *precision*, *recall*, dan *f-measure* pada kelas minor cenderung mengalami penurunan
4. Metode Fuzzy K-NN mampu memprediksi banyaknya *churner* lebih besar pada data yang seimbang / *balance* yaitu sebesar 82% dibandingkan pada data *imbalance* yaitu kurang dari 35%. Namun, tidak lebih besar dibandingkan metode K-NN pada data *imbalance* maupun *balance*.

5.2 Saran

1. Mengembangkan metode Fuzzy K-NN yang mampu meningkatkan hasil klasifikasi kelas minor pada data yang tingkat *imbalance*-nya tinggi dan dapat lebih meminimalkan pembiasan prediksi kelas minor ke kelas mayor
2. Sebaiknya data yang digunakan adalah data numerik sehingga tidak perlu melakukan preprocessing yang mungkin dapat menghilangkan representasi data asli

Daftar Pustaka

- [1] A. Jozwik. *A Learning scheme for A Fuzzy K-NN Rule*. Pattern Recognition Letters, vol 1, pp. 287-289, July 1983
- [2] Bodugula, Rajkumar. *Profiles and Fuzzy k-Nearest Neighbor Algorithm for Protein Secondary Structure Prediction*. Columbia, University of Missouri
- [3] Huang, Wen. 2006. *Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method*
- [4] Keller, James. 1985. *A Fuzzy k-Nearest Neighbor Algorithm*. IEEE Transaction on Systems, Man, and Cibernetic
- [5] Kusumaningdewi, Eka. 2009. *Prediksi churn pelanggan telekomunikasi selular menggunakan metode k nearest neighbor*. Bandung, Departemen Teknik Informatika Institut Teknologi Telkom
- [6] Larose, Daniel T. *Discovering Knowledge in Data : An Introduction to Data Mining*. 2005. A John Wiley & Sons, Inc., Hoboken, New Jersey
- [7] Mattison Rob. *Telco Churn Management : The Golden Opportunity*.
- [8] Pang-Ning Tan. Michael Steinbach an Vipin Kumar. *Introduction to Data Mining*. University of Minnesota and Army High Performance Computing Research Center.
- [9] Singh, Sameer. 1998. *Forecasting Using a Fuzzy Nearest Neighbour Method*. United Kingdom, University of Plymouth
- [10] Sugiharto. 2006. *Pemrograman GUI dengan Matlab*. Andi : Yogyakarta
- [11] T.W. Liao, D.Li. 1996. *Two Manufacturing Applications of The Fuzzy K-NN Algorithm*. Industrial and Mnuufacturing Sysytems Engineering Department, Louisiana State University, Baton Rouge, LA 70803, USA
- [12] V. T. Kissiov, S. T. Hadjitodorov. *A Fuzzy Version of The K-NN Method*. Department of Biotechnics, Technical University,Bulgaria