**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI**

**INSTITUT TEKNOLOGI SEPULUH NOPEMBER**

**USULAN TUGAS AKHIR**

# **IDENTITAS PENGUSUL**

Nama : **Azwar Rizal Alfarisi**

NRP : **5108 100 117**

Dosen Wali : **Imam Kuswardayan, S.Kom, M.Kom**

1. **JUDUL TUGAS AKHIR**

***“Implementasi Algoritma Mutual K-Nearest Neighbor Classifier Untuk Klasifikasi”***

1. **LATAR BELAKANG**

Klasifikasi data adalah sebuah teknik *supervised learning* yang bertujuan untuk menetapkan kategori pada suatu data baru berdasarkan hasil pelatihan sekumpulan data yang kategorinya sudah diketahui. Klasifikasi data telah dipelajari secara luas dalam berbagai bidang, seperti bioinformatik dan klasifikasi dokumen.

Meskipun telah banyak metode klasifikasi telah diperkenalkan, seperti *decision tree* dan *support vector machine* (SVM), pengklasifikasi berjenis *instance-based* tetap banyak digunakan karena sederhana dan efektif. Salah satunya adalah algoritma *k-nearest neighbor* (KNN). KNN termasuk algoritma *lazy learning* karena menggunakan semua data latihnya sebagai model untuk menghitung kemiripan dengan data baru. Untuk menghitung kemiripan antar data digunakan *euclidean distance* untuk data numerik dan *jaccard distance* untuk data diskrit.

Walaupun mudah untuk diimplementasikan dan rata-rata mempunyai akurasi yang cukup tinggi, KNN mempunyai kelemahan yang bisa menyebabkan performa yang buruk. Secara umum, performa KNN bergantung pada tiga aspek, yaitu nilai k, jumlah data training dan pemilihan metode untuk menghitung kemiripan, dimana nilai k paling krusial. Biasanya dengan memperbesar nilai k akan membuat model lebih bebas dari gangguan data *noise*, walaupun akan memperbesar *cost* komputasi[1].

Hampir semua dataset mempunyai data *noise* karena data diperoleh dari sumber yang heterogen. Keheterogenan sumber data itulah yang menyebabkan adanya data yang tidak konsisten. Data *noise* cukup berpengaruh pada akurasi hasil klasifikasi dan akan menyebabkan bertambahnya kompleksitas algoritma *learning* dalam membangun model klasifikasi. Oleh karena itu, diperlukan suatu teknik yang efektif untuk menghilangkan data *noise* dari dataset[2].

Algoritma *mutual nearest neighbor* (MNN) adalah algoritma hasil pengembangan dari algoritma KNN. Berbeda dengan KNN yang menggunakan k tetangga terdekatnya untuk memprediksi label kelas data baru *x0*, MNN menggunakan tetangga mutual terdekat dari *x0* untuk mendapatkan label kelas *x0*. Keuntungan dari metode ini adalah mampu mendeteksi suatu data *noise* yang berupa *outlier* untuk dieliminasi. Selain itu, algoritma MNN bisa digunakan untuk melakukan imputasi terhadap *missing values*.

Berdasarkan kelebihan-kelebihan algoritma MNN tersebut, maka akan dibuat model klasifikasi baru yang memanfaatkan algoritma MNN untuk klasifikasi sekaligus untuk *anomaly detection* dan menangani *missing values* yaitu algoritma *mutual k-nearest neighbor classifier* (MkNNC).

Tugas akhir ini akan mengimplementasikan algoritma MkNNC tersebut untuk melakukan klasifikasi pada berbagai macam data. Yaitu data dengan tipe data kontinu, diskrit dan gabungan kedua tipe data tersebut. Selain itu, ujicoba juga akan dilakukan terhadap dataset dengan jumlah label kelas biner dan label kelas berjumlah banyak.

1. **TUJUAN TUGAS AKHIR**

Tugas akhir ini bertujuan mengimplementasikan sistem perangkat lunak yang menerapkan algoritma *mutual k-nearest neighbor classifier* untuk klasifikasi berbagai macam data.

1. **MANFAAT TUGAS AKHIR**

Tugas akhir ini dikerjakan dengan harapan dapat memberikan manfaat pada perkembangan ilmu klasifikasi data karena algoritma yang diimplementasikan bisa dijadikan pertimbangan untuk penelitian di bidang klasifikasi data kedepannya.

1. **RUMUSAN MASALAH**

Untuk mengimplementasikan algoritma *mutual k-nearest neighbor* ini, maka dirumuskan permasalahan sebagai berikut.

1. Bagaimana mengimplementasikan metode *mutual nearest neighbor* untuk imputasi data yang mempunyai *missing values*.
2. Bagaimana melakukan *anomaly detection* menggunakan algoritma *mutual nearest neighbor*.
3. Bagaimana mengimplementasikan algoritma *mutual k-nearest neighbor classifier* untuk klasifikasi berbagai macam data.
4. Bagaimana mengevaluasi model yang telah dibuat. Evaluasi akan dilakukan sebagai berikut.
   1. Membandingkan hasil imputasi dengan metode *mutual nearest neighbor* terhadap metode imputasi mencari *mean* dari jumlah nilai atribut data lain.
   2. Membandingkan akurasi algoritma KNN biasa terhadap algoritma KNN dengan dataset yang telah dihilangkan *outlier*nya.
   3. Membandingkan akurasi algoritma MkNNC dengan algoritma KNN biasa dan algoritma KNN yang datasetnya tanpa *outlier*.
   4. Membandingkan akurasi algoritma MkNNC dengan model klasifikasi lain.
5. **BATASAN MASALAH**

Permasalahan yang dibahas dalam Tugas Akhir ini memiliki beberapa batasan, yaitu sebagai berikut.

1. Sistem perangkat lunak dibuat dengan menggunakan bahasa pemrograman Java SE dan perangkat lunak NetBeans IDE.
2. Menggunakan library Weka untuk melakukan perbandingan terhadap metode lain pada saat evaluasi model.
3. Dataset yang digunakan diperoleh dari UCI *Repository of Machine Learning Databases*[3]. Setiap dataset mempunyai keunikan sendiri. Seperti tipe data atribut, adanya *missing value* dan jumlah label kelas.
4. **RINGKASAN TUGAS AKHIR**

Dalam implementasi tugas akhir ini dataset yang digunakan adalah sebagai berikut.

**Tabel 1 Dataset yang Digunakan**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Nama Dataset | Tipe Data | *Missing Values?* | Jumlah Label Kelas |
| 1 | *Breast Cancer Wisconsin (Diagnostic)* | *Real* | Tidak Ada | Biner |
| 2 | *Iris* | *Real* | Tidak Ada | Banyak |
| 3 | *Breast Cancer Wisconsin (Prognostic)* | *Real* | Ada | Biner |
| 4 | *Chess (King-Rook vs. King Pawn)* | *Categorical* | Tidak Ada | Biner |
| 5. | *Soybean (Large)* | *Categorical* | Ada | Banyak |
| 6. | *Lymphography* | *Categorical* | Tidak Ada | Banyak |
| 7. | *Credit Approval* | *Categorical, Real* | Ada | Biner |
| 8. | *Statlog (Heart)* | *Categorical, Real* | Tidak Ada | Biner |
| 9. | *Cleveland* | *Categorical, Real* | Ada | Banyak |

Dalam pengimplementasian Tugas Akhir ini akan mengikuti diagram alir sebagai berikut.



Gambar 8.1 Diagram alir secara umum

Berdasarkan diagram alir diatas, ada 4 tahap yang akan dilakukan pada implementasi tugas akhir ini, yaitu.

1. Imputasi terhadap *missing value*.
2. Eliminasi *outlier*.
3. Membangun model algoritma *mutual k-nearest neighbor classifier*.
4. Mengevaluasi model dengan membandingkan akurasinya terhadap model-model lain menggunakan metode *k-fold cross validation*.

Tahap pertama dalam pengerjaan setelah didapatkan dataset adalah melakukan imputasi terhadap data yang mempunyai *missing value*. Metode yang digunakan adalah gabungan antara *mutual nearest neighbor* dengan *mean* data untuk data numerik dan *modus* data untuk data diskrit*.* Prosedur langkah per langkah pendekatan metode ini dijelaskan sebagai berikut.

* Langkah 1 : Inputkan *k*, cari *k* *mutual nearest neighbor* dari data *X* dengan cara.
  + Hitung jarak antara data *X* dengan data lain. Urutkan dari yang mempunyai kemiripan paling besar.
  + Cari *k* tetangga yang salah satu *k* tetangga dekatnya adalah data *X.*
* Langkah 2 : Hitung *mean* nilai data atribut tersebut untuk data numerik dan *modus* nilai data untuk data diskrit.
* Langkah 3 : Masukkan hasil yang didapat dari langkah 2 ke nilai yang kosong.
* Langkah 4 : Ulangi langkah 1 sampai langkah 3 sampai tidak ada *missing value*.

Tahap selanjutnya adalah melakukan eliminasi data *outlier* dengan *mutual nearest neighbor*. Algoritma KNN akan mencari *k* tetangga terdekat dari data *X,* namun belum tentu data-data tetangga tersebut adalah tetangga sebenarnya dari *X.* Dalam metode algoritma MNN disebut tetangga bayangan. Oleh karena itu, dari *k* tetangga terdekat yang didapatkan akan diperiksa kembali apakah *X* juga *k* tetangga terdekat dari data-data yang ada. Jika data-data tersebut tidak mempunyai *k* tetangga terdekat berupa data *X* maka *X* dianggap sebagai data *outlier*. Berikut *pseudocode* metode menghilangkan data *outlier*.

procedure anomalies\_removal

**Input**: dataset D dan nilai *nearest neighbors* k;

**Output**: Reduksi D;

For tiap *instance* x pada D

Do

Dapatkan MNN dari x, Misal M(x);

if M(x) = null;

Hapus x dari D

Otherwise

Biarkan x di D.

End

Kembalikan D sebagai hasil reduksi.

Selanjutnya adalah mengimplementasikan algoritma *mutual k-nearest neighbor classifier* (MkNNC) untuk klasifikasi. Mula-mula, algoritma ini membuat himpunan label kelas *C(x)*, yang nantinya digunakan untuk memprediksi label *x*. Selanjutnya, mendapatkan *k* tetangga terdekat dengan teknik KNN tradisional. Kemudian algoritma mengidentifikasi tetangga mutual terdekat dari *x* dan menyimpan labelnya. Lebih khusus, untuk tiap-tiap tetangga terdekat *y*, cari *k* tetangga terdekatnya dari *D* untuk memastikan apakah salah satu *k* terdekat dari *y* adalah *x*. Jika benar, y adalah salah satu tetangga mutual terdekat dari *x* dan labelnya akan menjadi kandidat kelas dari *x*. Akhirnya, label kelas *C(x)* ditentukan dengan menggunakan strategi mayoritas. Berikut adalah *pseudocode* dari metode ini.

procedure MkNNC

**Input**: dataset D, *instance* x dan nilai *k* nearest neighbors;

**Output**: prediksi label c(x);

Inisialisasi parameter relatif, seperti, C(x)= null;

Dapatkan k tetangga terdekat Nk(x) dari D;

For tiap data tetangga y ϵ Nk(x)

Do

Dapatkan k tetangga terdekat Nk(y) dari D;

If x juga k tetangga terdekat Nk(y)

Tambahkan label informasi y ke C(x);

End

Label kelas C(x) ditentukan sebagai berikut:

c(x)= arg max , dimana I(.) adalah fungsi indikasi.

Kembalikan c(x) sebagai hasil

.

|  |  |
| --- | --- |
| *d(p,q) =* | (1) |

Untuk menghitung jarak antar data digunakan *euclidean distance* untuk dataset dengan semua tipe data numerik, *jaccard distance* untuk dataset dengan semua tipe data diskrit dan *Heterogeneous Euclidean Overlap Metric* (HEOM) untuk dataset dengan tipe data heterogen[4]. Berikut adalah persamaan-persamaan untuk menghitung jarak ketiga metode diatas.

merupakan persamaan *euclidean distance*  dimana *p* dan *q* adalah data dan *n* adalah jumlah atribut.

|  |  |
| --- | --- |
| *Jd(A,B) =* | (2) |

merupakan persamaan *jaccard distance*  dimana *A* dan *B* adalah data, adalah jumlah atribut dan adalah jumlah atribut yang nilai datanya sama.

|  |  |
| --- | --- |
| *HEOM(x,y) =* | (3) |

dimana *m* adalah jumlah atribut. Dan fungsi mengembalikan nilai jarak data *x* dan *y* pada atribut ke-*a* dengan ketentuan sebgai berikut.

*=*

dimana = 0, jika *x = y* dan = 1 jika sebaliknya.

1. **METODOLOGI**

Metodologi yang akan dilakukan dalam Tugas Akhir ini memiliki beberapa tahapan, diantaranya sebagai berikut:

1. Penyusunan Proposal Tugas Akhir

Tahap awal untuk memulai pengerjaan Tugas Akhir adalah penyusunan Proposal Tugas Akhir. Pada proposal ini, penulis mengajukan gagasan pembuatan sistem untuk memprediksi berbagai macam dataset.

1. Studi Literatur

Pada tahap ini dilakukan pencarian, pengumpulan, penyaringan, pembelajaran dan pemahaman literatur yang berhubungan dengan algoritma *mutual nearest neighbor*, metode imputasi dan metode *anomaly detection* serta pengumpulan dataset. Literatur yang digunakan dalam pengerjaan Tugas Akhir ini sebagian besar berasal dari internet berupa makalah ilmiah, tesis, artikel, materi kuliah, serta beberapa buku referensi.

1. Implementasi

Implementasi merupakan tahap untuk membangun sistem tersebut.

1. Pengujian dan Analisis

Pada tahap ini dilakukan uji coba terhadap sistem yang telah dibuat, mengamati kinerja sistem yang baru dibuat, serta mengidentifikasi kendala yang mungkin timbul.

5. Penyusunan Buku Tugas Akhir

Tahap terakhir merupakan penyusunan laporan yang memuat dokumentasi mengenai pembuatan serta hasil dari implementasi perangkat lunak yang telah dibuat.

1. **JADWAL PEMBUATAN TUGAS AKHIR**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Kegiatan** | **Bulan** | | | | | | | | | | | | |
| **Oktober** | | | | **Nopember** | | | | **Desember** | | | | |
| **1** | Penyusunan Proposal |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **2.** | Studi Literatur |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **3.** | Implementasi |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **4.** | Pengujian dan Analisis |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **5.** | Penyusunan Buku |  |  |  |  |  |  |  |  |  |  |  |  |  |

1. **DAFTAR PUSTAKA**
2. Liu, H., Zhang, S., Zhao, J., Zhao, X., Mo, Y., A New Classification Algorithm using Mutual Nearest Neighbors, 2010 Ninth International Conference on Grid and Cloud Computing 52-57.
3. Liu, H., Zhang, S., Noisy data elimination using mutual k-nearest neighbor for classification mining, The Journal of Systems and Software 85 (2012) 1067-1074
4. UCI Machine Learning Repository, University of California. <http://archive.ics.uci.edu/ml>.
5. Wilson, D.R., Martinez, T.R., 1997, Improved Heterogeneous Distance Functions, Journal of Artificial Intelligence Research 6, 1-34