Implementation and Analysis Optimal Flexible Frequency Discretization (OFFD) Method to Minimize Classification Error at Naïve Bayes Classification

Dita Martha Pratiwi[1], Warih Maharani[2], Intan Nurma Yunita[3]
Fakultas Informatika Institut Teknologi Telkom
Jl. Telekomunikasi Terusan Buah Batu
Bandung 40257 Indonesia
Telp: 62-22-756 4108 Fax: 62-22 756 5200
[1]deeytha3@gmail.com, [2]wmaharani@gmail.com, [3]intanurma@gmail.com

Abstrak

Naïve Bayes merupakan salah satu teknik pengklasifikasian dalam data mining yaitu dengan menerapkan teorema Bayes dalam pengolahannya. Teknik ini akan memberikan hasil optimal bila setiap atribut dalam dataset saling bebas. Namun pada umumnya, suatu dataset memiliki atribut numerik dan atribut nominal yang tidak saling bebas sehingga apabila dianggap saling bebas maka dapat menimbulkan permasalahan classification error. Oleh karena itu dibutuhkan suatu metode untuk meminimalkan nilai error rate pada kesalahan pengklasifikasian tersebut, salah satu metodenya adalah strategi diskritisasi. Diskritisasi adalah sebuah metode yang memetakan beberapa nilai numerik (X) ke sebuah interval bernilai nominal (X) berdasarkan pengaturan frekuensi di dalam satu interval sehingga bisa di dapatkan jumlah interval terbentuk dalam satu atribut numerik.*

Salah satu metode diskritisasi yang diterapkan dalam penelitian ini adalah Optimal Flexible Frequency Discretization (OFFD) yang berbasis sequential search dan wrapper based supervised untuk incremental learning. Pada metode ini dilakukan feature selection dengan teknik wrapper untuk mendapatkan atribut yang optimal berdasarkan parameter fmeasure. Selanjutnya data dengan atribut optimal tersebut akan didiskritisasi secara sequential search untuk nilai frekuensi minimum tiap interval. Berdasarkan hasil pengujian yang dilakukan menunjukkan bahwa metode OFFD dipengaruhi oleh proses pemilihan atribut, dimana dalam penelitian ini digunakan pencarian Best First Search pada proses Wrapper Feature Selection, sehingga berpengaruh terhadap penurunan nilai error.

***Kata kunci** : Wrapper based, Feature Selection, Diskritisasi, Sequential Search, Naïve Bayes, Optimal Flexible Frequency Discretization, frekuensi interval*

Abstract

Naive Bayes is one of the classification techniques in data mining that apply Bayes Theorem in its processing and provide optimal result when each attributes in dataset is independent. But generally, a dataset has numeric attributes and nominal attributes are dependence so that if considered independent, it can cause classification error problems. Therefore, it needs a method to minimize the error rate, the method is discretize strategy. Discretization is a method that maps some numerical values (X) into an interval of nominal value (X) based on the frequency setting in one interval so it can get number of interval formed in one numeric attribute.*

One of discretization method adopted in this research is Optimal Flexible Frequency Discretization (OFFD) based on sequential search and wrapper based supervised for incremental learning. This method will be carried out wrapper feature selection to get optimal attributes based on its fMeasure parameter. Then, optimal dataset will de discrete in sequential search for the minimum frequency on each interval. Based on the results of testing, showed that the OFFD influenced by the process of selecting attributes of Best First Search on the Wrapper Feature Selection, so that influence the decline in the value of the error.

***Keywords** : wrapper based, Feature Selection, discretization, sequential search, Naïve Bayes, Optimal Flexible Frequency Discretization, interval frequency*

1. Pendahuluan

Dewasa ini, perkembangan teknologi yang semakin pesat membuat akumulasi data yang ada semakin besar dan berlebih sehingga membuat para peneliti melahirkan sebuah konsep dan teknologi yang mampu mengolah data dalam jumlah besar dan menampilkan representasi data yang mudah dimengerti, yaitu data mining [6]. Secara sederhana, data mining merupakan proses Knowledge Discovery in Database (KDD) atau ekstraksi pengetahuan dari jumlah data yang besar [4]. Dan ketika ingin membangun sebuah sistem data mining, peneliti harus mengetahui kegunaan pembangunan sistem tersebut, misal pengidentifikasian resiko diabetes (9,6% orang muda dan 20,9% orang tua) pada biomedical. Hal ini agar tidak terjadi misunderstanding pada teknologi data mining [1,4,10].

Pada data mining sendiri ada beberapa fungsionalitas yaitu Characterization and Discriminant, Association analysis, Classification and Prediction, Cluster analysis, Outlier analysis, dan Evolution analysis [5]. Salah satu metode dari fungsionalitas Classification and Prediction adalah Naïve Bayes Classification yang merupakan metode sederhana, mudah diimplementasikan dan mendukung incremental training [3,4,7,8]. Berdasarkan [12], Naïve Bayes Classification akan bernilai optimal ketika atribut-atribut pada suatu dataset merupakan atribut yang diasumsikan sebagai variable acak yang saling bebas. Namun pada umumnya, suatu dataset memiliki atribut numerik dan atribut nominal yang tidak saling bebas sehingga apabila dianggap saling bebas maka dapat menimbulkan permasalahan classification error [7].

Oleh karena itu dibutuhkan suatu cara untuk mengoptimalkan pengklasifikasian dan meminimalkan error rate pada Naïve Bayes Classification, salah satu metodenya adalah discretization strategy. Cara tersebut adalah mengubah setiap atribut numerik (X) menjadi atribut nominal (X*) dengan memetakan nilai-nilai atribut numerik ke satu nominal yaitu dengan mengatur nilai ukuran atau frekuensi atribut numerik di dalam satu interval (interval frequency) sehingga diketahui jumlah interval yang terbentuk (interval number) [11]. Metode pendiskritisasian harus mampu melakukan pengaturan interval yang optimal agar tidak muncul efek simpangan dan selisih diskritisasi berlebih yang mempengaruhi nilai error rate [12]. Selain itu, disarankan untuk melakukan proses pemilihan atribut (feature selection) yang relevan terlebih dahulu pada dataset sehingga dapat diketahui atribut mana saja yang akan didiskritkan.

Pada penelitian ini, digunakan metode Optimal Flexible Frequency Discretization (OFFD) yang berbasis sequential search dan wrapper based supervised untuk incremental learning [8,9]. Pendekatan feature selection yang ada di metode OFFD ini merupakan wrapper model dimana algoritmanya melakukan pencarian menggunakan algoritma induksinya sendiri untuk mendapatkan atribut yang optimal terhadap label pengklasifikasian [5]. Dan menurut [8], OFFD akan mengatur nilai ukuran atau frekuensi atribut numerik di dalam satu interval (interval frequency) dengan melakukan sequential search terhadap nilai frekuensi minimum dari 1 sampai 40 (minBinsize=1-40) dan batas nilai frekuensi maksimumnya sebesar 2 kali nilai minimum frekuensinya (maxBinsize=2 x minBinsize) agar diketahui interval frequency dan interval number optimalnya. Secara garis besar langkah – langkahnya adalah wrapper-based feature selection untuk mendapatkan atribut-atribut yang optimal terhadap label kelas. Selanjutnya adalah diskritisasi terhadap atribut numeriknya dengan menelusuri nilai minBinsize untuk mendapatkan nilai error rate yang kecil.

Tujuan dari penelitian ini adalah Menganalisis kerja metode OFFD untuk meminimalkan classification error pada Naïve-Bayes Classification serta hasil error rate nya dengan atau tanpa metode OFFD, Menganalisis pengaruh keoptimalan parameter minBinsize dan maxBinsize pada OFFD terhadap pengaruh tingkat kesalahan Naïve-Bayes Classification, Menganalisis pengaruh Wrapper Feature Selection terhadap hasil error rate pada metode Optimal Flexible Frequency Discretization (OFFD).

2. Naïve Bayes Classifier

Naïve Bayes merupakan teknik pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal

sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan ”naive” dimana diasumsikan kondisi antar atribut saling bebas. Atribut dianggap saling bebas apabila nilai-nilai yang dimiliki oleh atribut satu dengan atribut lainnya pada dataset yang sama tidak memiliki keterkaitan sehingga saat pemrosesan tidak melibatkan nilai yang ada di atribut lain.

Pada sebuah dataset, setiap baris/dokumen I diasumsikan sebagai vector dari nilai-nilai atribut $<x_1,x_2,...,x_3>$ dimana tiap nilai-nilai menjadi peninjauan atribut X_i ($i \in [1,n]$)). Setiap baris mempunyai label kelas $c_i \in \{c_1,c_2,\ldots,c_k\}$ sebagai nilai variabel kelas C, sehingga untuk melakukan klasifikasi dapat dihitung nilai probabilitas $p(C=c_i|X=x_j)$, dikarenakan pada Naïve Bayes diasumsikan setiap atribut saling bebas, maka persamaan yang didapat adalah sebagai berikut :

$$P(C = c_i \mid I) \tag{2.1}$$
$$= \frac{P(C=c_i)P(I \mid C=c_i)}{P(I)} \tag{2.2}$$
$$\propto P(C = c_i)P(I|C = c_i) \tag{2.3}$$
$$= P(C = c_i)P(< x_1, x_2, .., x_n > |C = c_i) \tag{2.4}$$
$$= P(C = c_i)\prod_{j=1}^{n} P(X_j = x_j|C = c_i) \tag{2.5}$$

Peluang $p(C=c_i|X=x_j)$ menunjukkan peluang bersyarat atribut X_i dengan nilai x_i diberikan kelas c, dimana dalam Naïve Bayes, kelas C bertipe kualitatif sedangkan atribut X_i dapat bertipe kualitatif ataupun kuantitatif. Ketika atribut X_i bertipe kuantitatif maka peluang $p(X=x_i|C=c_j)$ akan sangat kecil sehingga membuat persamaan peluang tersebut tidak dapat diandalkan untuk permasalahan atribut bertipe kuantitatif. Maka untuk menangani atribut kuantitatif, ada beberapa pendekatan yang dapat digunakan seperti Distribusi Normal (Gaussian).
Ataupun Kernel Density Estimation (KDE) :

$$\hat{f} = \frac{1}{n_c}\sum_{j} N(X_i; \mu_{ij}, \sigma_c), \tag{2.7}$$

Selain dua pendekatan distribusi tersebut, ada mekanisme lain untuk menangani atribut kuantitatif (numerik) yaitu Strategi Diskritisasi. Proses diskritisasi sendiri terjadi saat proses persiapan data atau saat data preprocessing, dimana atribut numerik X diubah menjadi atribut nominal X*. Performansi klasifikasi Naive Bayes akan lebih baik ketika atribut numerik didiskritisasi daripada diasumsikan dengan pendekatan distribusi seperti di atas [Dougherty]. Nilai-nilai numerik akan dipetakan ke nilai nominal dalam bentuk interval yang tetap memperhatikan kelas dari tiap-tiap nilai numerik yang dipetakan, penggambaran perhitungan Naive Bayesnya terlihat pada tabel kontingensi seperti berikut:

Tabel 1 : Kontingensi counter

	Interval 1 (i_1)	Interval 2 (i_2)
Kelas 1 (c_1)	Counter: penambahan/incremental jumlah anggota di tiap interval berdasarkan kelas nya.	
Kelas 2 (c_2)		

Maka, rumus Naive Bayes yang dipakai menjadi :

$$p(I=i_j|C=c_i) = \frac{p(I=ij)\,p(C=ci|I=ij)}{p(C=ci|)} \tag{2.8}$$

Ket :
$p(I=i_j|C=c_i)$: peluang interval i ke-j untuk kelas c_i
$p(C=c_i|I=i_j)$: peluang kelas c_i pada interval i ke-j
$p(I=i_j)$: peluang sebuah interval ke-j pada semua interval yang terbentuk

$p(C=c_i)$: peluang sebuah kelas ke-i untuk semua kelas yang ada di dataset

3. Optimal Flexible Frequency Discretization

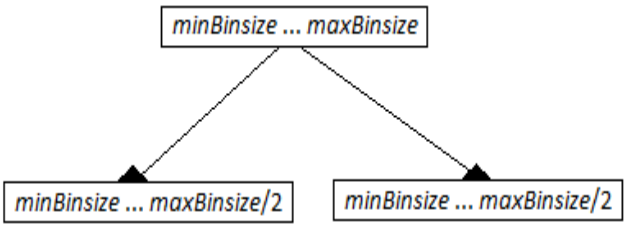
OFFD merupakan metode pengembangan dari IFFD (Incremental Flexible Frequency Discretization) dimana data akan direduksi dengan cara diskritisasi secara sekuensial berdasarkan nilai minimum frekuensi interval yang diinputkan. Nilai baru akan diinputkan ke dalam sebuah interval tertentu selama frekuensi di dalam interval tersebut masih diantara nilai minimum frekuensi sampai 2 kali minimum frekuensi (maksimum frekuensi).

Apabila frekuensi melebihi atau sama dengan nilai maksimum frekuensi maka interval tersebut akan di split menjadi 2 interval. Pendiskritisasian akan dilakukan pada setiap atribut dan berulang sebanyak atribut yang dimiliki oleh dataset. Jika semua atribut sudah terdiskrit, maka lakukan kembali secara berulang untuk nilai minimum frekuensi 1 sampai 40. Sebelum proses pendiskritisasian, akan dilakukan proses feature selection dengan metode wrapper hal ini untuk memperkecil ruang pencarian sistem ini sehingga waktu pengeksekusian lebih efisien serta didapatkan atribut-atribut yang akan mengoptimalkan hasil pengolahan knowledge data.

3.1 Diskritisasi Sekuensial

Diskritisasi merupakan sebuah pendekatan yang terkenal untuk memetakan atribut numerik ke sebuah atribut nominal untuk Naive Bayes Classification. Pada diskritisasi untuk Naive Bayes ini ada dua terminologi yaitu frekuensi interval (banyaknya nilai dalam satu interval) dan jumlah interval (banyaknya interval yang terbentuk oleh algoritma diskritisasi tertentu) [2]. Dua terminologi tersebut yang bisa menimbulkan permasalahan classification error karena tidak tepatnya penyetingan nilainya, disebut dengan permasalahan diskritisasi bias dan varians [4]. Bias adalah komponen error yang hasilnya dari kesalahan sistematik algoritma yang digunakan. Varians adalah komponen error yang hasilnya dari variasi acak pada data training dan sifat acak pada algoritmanya demikian mengurus seberapa sensitif suatu algoritma diubah pada data training.

Oleh karena itu, dibutuhkan metode diskritisasi yang tepat agar nilai error optimal terkecil yang didapat. Pada penelitian ini metode pendiskritisasian yang digunakan yaitu metode OFFD (Optimal Flexible Frequency Discretization). Metode ini akan menguji secara sequential search setiap nilai minBinsize (1-40) untuk dilakukan pendiskritisasian yaitu memetakan nilai atribut pada instance secara satu per satu hingga batas nilai maxBinsize (nilainya 2x minBinsize) [8]. Apabila frekuensi di satu interval sudah mencapai maxBinsize maka interval tersebut akan dibagi menjadi 2 interval (splitting).



Gambar 1 : Proses pemotongan sebuah interval (splitting)

Apabila di dalam 1 interval yang akan melalui proses splitting terdapat nilai yang sama maka nilai tersebut tidak boleh terpisah atau berada di interval terpecah yang berbeda. Nilai-nilai yang sama tersebut harus berada di dalam 1 interval yang sama. Lalu tentukan titik potong baru pada masing-masing interval yang terpisah. Titik potong disini adalah batas atas dan batas bawah di dalam 1 interval, misal:

4.2	4.5	4.5

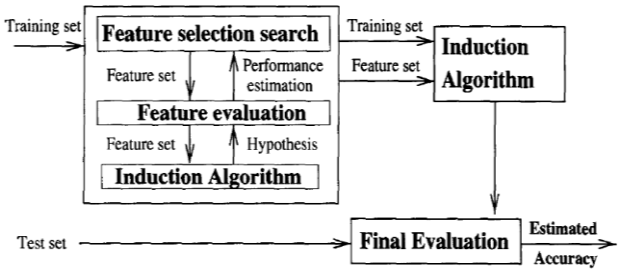
Gambar 2 : Interval terbentuk

Maka titik potongnya adalah [4.2 , 4.5] yang artinya nilai baru (X) yang akan masuk ke interval tersebut bernilai diantara 4.2 ≤ X ≤ 4.5.

3.2 Wrapper Feature Selection

Merupakan suatu pendekatan untuk mengatasi permasalahan dalam algoritma klasifikasi yang melibatkan atribut-atribut untuk fokus di dalam algoritma tersebut. Pada suatu algoritma pembelajaran, untuk mendapatkan akurasi yang tinggi saat membangkitkan classifier dibutuhkan kumpulan subset atribut terbaik. Dimana pemilihan atribut menggunakan algoritma induksi sebagai blackbox serta sebagai bagian dari fungsi evaluasi dan melakukan pencarian atributnya menggunakan algoritma pencari seperti Best First Search.

Pendekatan wrapper melakukan pencarian atribut di ruang parameter-parameter yang mungkin, seperti ruang kondisi, kondisi awal, kondisi penghentian, dan mesin pencari. Pencarian atribut relevan juga dilakukan secara forward selection dimana pencarian dimulai dari node (kondisi awal) kosong atau tidak ada kelompok kombinasi atribut yang terpilih.



Gambar 3 : Flowchart untuk Wrapper Feature Selection [5]

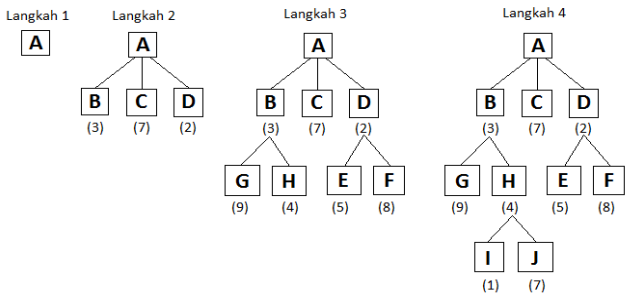
Hal yang ingin dicapai dalam pencarian subset atribut dengan pendekatan wrapper adalah mendapatkan node dengan nilai evaluasi tertinggi menggunakan fungsi heuristik dan fungsi pengevaluasian salah satu fungsinya adalah k-fold cross-validation. Pengevaluasian ini melakukan pengecekan berulang sebanyak nilai k yang diinputkan.

3.3 Best First Search

Merupakan cara pencarian yang menggabungkan kelebihan Breadth-First Search dan Depth-First Search. Pada setiap langkah proses pencarian terbaik pertama, kita memilih node-node dengan menerapkan fungsi heuristik yang memadai pada setiap node/simpul yang kita pilih dengan menggunakan aturan-aturan tertentu untuk menghasilkan penggantinya. Fungsi Heuristik yang digunakan merupakan prakiraan (estimasi) cost dari initial state ke goal state, yang dinyatakan dengan : **F' = G + H'** dimana

F' = prakiraan cost dari initial ke goal
G = cost dari initial state ke current state
H' = prakiraan cost dari current state ke goal state

Adapun contoh prosesnya adalah sebagai berikut **:**



Gambar 4 : Proses pencarian dengan metode BFS

3.4 Pengukuran Evaluasi

$$Precision = \frac{categories\ found\ and\ correct}{total\ categories\ found}\ Recall = \frac{categories\ found\ and\ correct}{total\ categories\ correct}$$

Untuk menguji keefektifan atau kualitas hasil suatu klasifikasi dibutuhkan suatu pengukuran yang disebut evaluation measure

- Precision : persentase instance yang diklasifikasikan dengan benar diantara semua instance yang ditentukan kategorinya oleh classifier.
- Recall : persentase instance yang dikategorikan dengan benar diantara instance yang benar untuk kategori tersebut.
- Macro fMeasure : nilai rata-rata dari nilai fMeasure setiap kelas yang ada di dalam dataset. Maka perumusan nya adalah sebagai berikut

$$\text{Macro fMeasure} = \frac{\sum fmeasure(i)}{\sum class}$$

dengan persamaan fMeasure nya adalah sebagai berikut

$$\text{fMeasure} = \frac{2(precision \times recall)}{(precision + recall)}$$

Selain pengukuran nilai Macro fMeasure untuk mendapatkan subset atribut terbaik, akan diukur nilai perubahan error berdasarkan sekuensial minimum frekuensinya.

$$\text{Error rate} = 1 - \frac{TP+TN}{TP+TN+FP+FN}$$

Tabel 2 : Confusion Matrix

		Predicted Class	
		Yes	no
Actual Class	Yes	True Positif (TP)	False Negative (FN)
	No	False Positif (FP)	True Negative (TN)

3.5 Tahapan OFFD

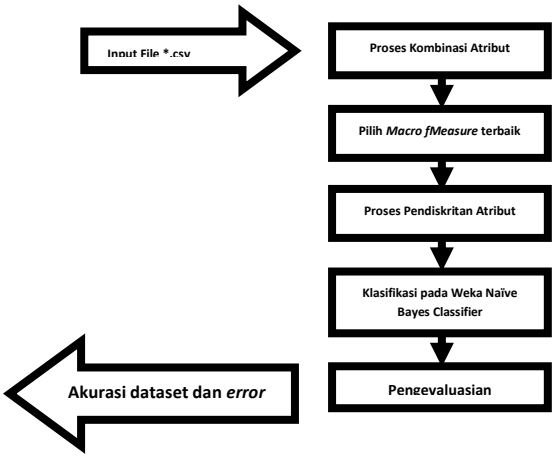
Suatu dataset akan mengalami pemilihan atribut terlebih dahulu secara wrapper based selain untuk mengurangi ruang pencarian, pemilihan atribut dapat meningkatkan akurasi dikarenakan atribut-atribut tidak relevan dihilangkan sehingga atribut tersisa akan didiskritkan.

Adapun tahap pendiskritan:

a. Sebuah baris/dokumen data uji dengan nilai v untuk atribut V sebagai inputan lalu setting nilai frekuensi minimum di dalam 1 interval.
b. Pengecekan, jika v > titik potong terakhir pada sekelompok interval maka nilai v akan masuk ke interval terakhir, lalu update frekuensi di interval tersebut dan rekam kondisi ini.
c. Sedangkan untuk sebaliknya, akan melakukan penelusuran dari interval pertama lalu melakukan pengecekan. Jika nilai v ≤ titik potong di interval tersebut maka nilai v masuk ke interval tersebut lalu frekuensi di interval ditambah. Lakukan penambahan counter [j] [label kelas] dan rekam interval keberapa yang mengalami perubahan.
d. Setelah keluar dari perulangan pencarian interval, maka cek apakah frekuensi di interval tersebut ≥ 2 x minimum frekuensi. Jika ia maka interval tersebut dibagi dua dengan frekuensi masing-masing tidak boleh kurang dari frekuensi minimum.
e. Lalu tentukan titik potong baru di masing-masing interval split, hitung counter [interval split1] dan counter [interval split2].
f. Ulangi dari semua proses di pseudo code untuk atribut-atribut numerik selanjutnya.

g. Setelah semua atribut numerik terdiskrit berdasarkan nilai frekuensi minimum yang saat ini di uji, hitung error rate dengan Naïve Bayes Classifier dan simpan menjadi classification error saat ini.
h. Proses tersebut diulangi untuk nilai minBinsize 1 sampai 40.

4. Perancangan Sistem



Gambar 5 : Flowchart Metode OFFD

5. Pengujian

5.1 Dataset

Pengujian dalam penelitian ini menggunakan data set dari UCI KDD Archieve Repository yang tersedia di http://archieve.ics.uci.edu/ml/data_sets.html. Tedapat empat buah data set yang akan digunakan, dengan detail dari masing-masing data set adalah sebagai berikut :

Tabel 3 : Dataset Pengujian

Nama Dataset	Jml Record	Jml Atribut	Distinct	Jml Kelas	Selisih (%)
Wine	178	13	98	3	18
Sonar	208	60	188	2	11,1
Vowel	990	10	809	11	12,2
Wdbc	569	30	511	2	11,1

5.2 Skenario Pengujian

Pengujian dilakukan dengan memroses keempat dataset secara bergantian menggunakan sistem OFFD dimana atribut numerik akan didiskritisasi terhadap nilai minBinsize 1-40 untuk melihat pengaruhnya terhadap perubahan error rate yang dihasilkan. Selain itu, karakteristik dataset dianalisis untuk melihat pengaruh perubahan nilai minBinsize nya terhadap perubahan error rate yang dihasilkan. Karakteristik yang diperhatikan adalah perbandingan antara nilai rata-rata distinct yang dimiliki masing-masing atribut dengan jumlah record yang dimiliki dataset.

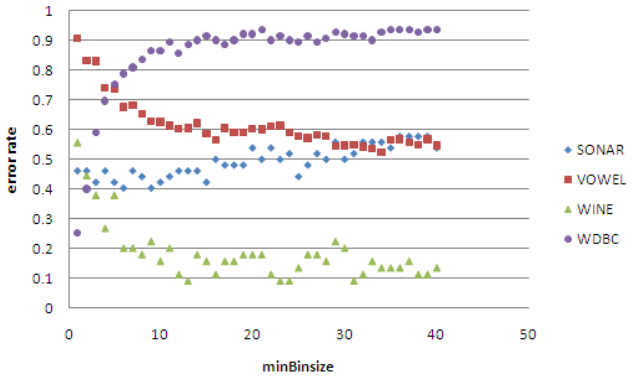
Proses pendiskritisasian keempat dataset dilakukan dua kali yaitu dengan pemilihan atribut terlebih dahulu serta tanpa pemilihan atribut sebelum didiskritisasi. Hal ini untuk melihat proses pada metode OFFD yang mempengaruhi penurunan error rate yang dihasilkan.

Setelah setiap atribut terdiskritisasi, atribut-atribut pada masing-masing dataset dimodelkan menggunakan alat bantu Weka dengan klasifikasi Naïve Bayes untuk didapatkan akurasinya. Akurasi yang akan dianalisis adalah akurasi dataset sebelum diproses dengan OFFD, setelah diproses dengan OFFD tanpa pemilihan atribut serta dengan pemilihan atribut terlebih dahulu. Perubahan akurasi yang terjadi memperlihatkan

karakteristik dataset yang bagaimana yang didiskritisasi terlebih dahulu sebelum diklasifikasi atau yang tanpa proses diskritisasi.

5.3 Hasil Pengujian

Tujuan pengujian yang dilakukan adalah dengan membandingkan perubahan nilai error rate untuk setiap data set terhadap perubahan nilai minBinsize nya sehingga dapat dianalisis pengaruh nilai minBinsize terhadap karakteristik data set nya serta nilai error rate yang didapatkan

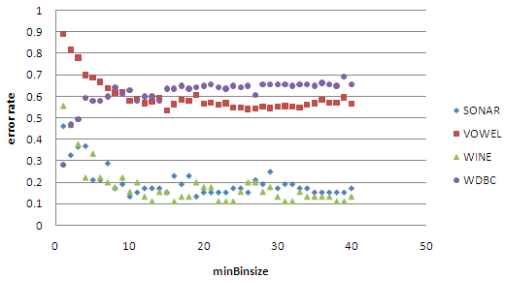


Gambar 6. Perubahan nilai error rate tanpa Wrapper Feature Selection berdasarkan perubahan nilai minbinsize

Pada dataset wdbc besar selisih antara jumlah record yang dimiliki dengan rata-rata distinct tiap atribut nya kecil yang artinya nilai yang ada di dalam dataset tersebut banyak yang bernilai sama, begitupun pada dataset sonar. Namun, titik awal atau pada minbinsize 1 untuk kedua dataset tersebut berbeda hal ini dikarenakan pada dataset wdbc nilai untuk masing-masing atribut perbedaan nilainya sangat jauh sedangkan pada dataset sonar perbedaannya dekat sehingga pola yang terbentuk tidak terlalu jauh berbeda dari satu titik dengan titik lainnya.

Sedangkan pada dataset vowel dan wine, karakteristik data yang dimiliki berdasarkan nilai selisih antara jumlah record dan rata-rata distinct adalah kecil. Nilai selisih yang kecil tersebut mengartikan karakter data yang dimiliki oleh kedua dataset tersebut banyak nilai yang sama sehingga nilai minbinsize yang baik digunakan adalah minbinsize besar. Hal ini agar tiap data yang bernilai sama terletak pada 1 interval dan dianggap sebagai sebuah nominal yang sama.

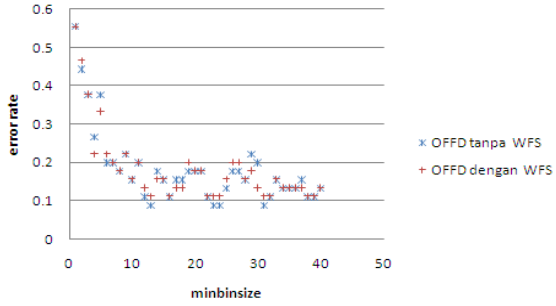
Sehingga dengan karakteristik masing-masing data set yang berbeda pada tabel 4.2 serta gambar grafik 4.1 dapat dikatakan bahwa semakin besar perbedaan antara rata-rata nilai distinct dengan jumlah record yang dimiliki data set tersebut maka nilai minbinsize yang baik untuk digunakan dalam proses pendiskritisasian adalah nilai minbinsize besar atau mendekati 40 hal ini dikarenakan semakin kecil kemungkinan suatu nilai yang sama berada di dalam interval yang berbeda karena frekuensi minimum di dalam satu interval besar sehingga data bernilai sama berpeluang besar untuk berada di interval yang sama.



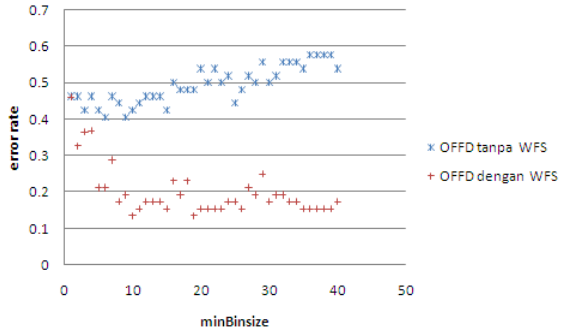
Gambar 7. Perubahan nilai error rate dengan Wrapper Feature Selection berdasarkan perubahan nilai minbinsize

Pada gambar di atas perubahan nilai untuk data set wdbc dan sonar khususnya mengalami perubahan dari gambar grafik 4.1. dikarenakan pada percobaan ini setiap dataset mengalami pereduksian atribut terlebih dahulu. Pada dataset sonar, dari jumlah atribut yang dimiliki 60 direduksi hingga mencapai 1 atribut terpilih hal ini tentu berpengaruh terhadap perubahan minbinsize yang cocok untuk karakter data yang dimiliki. Berdasarkan hasil pereduksian ternyata untuk kedua dataset tersebu, atribut yang terpilih memiliki nilai selisih antara jumlah record dengan jumlah distincnya adalah kecil sehingga nilai error akan semakin menurun jika nilai minbinsize yang digunakan semakin besar. Sedangkan untuk dataset vowel dan wine, perubahan yang terjadi tidak terlalu signifikan dikarenakan jumlah atribut yang tereduksi hanya 2-3 atribut saja.

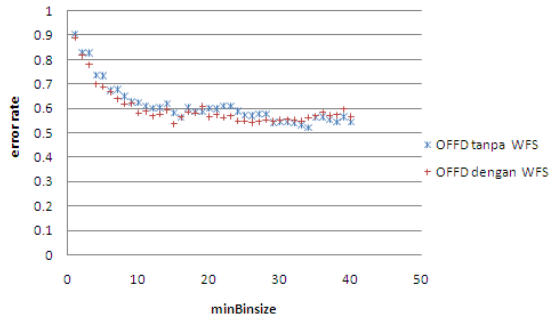
Jadi, apabila data set melalui proses pemilihan atribut relevan terlebih dahulu lalu dilakukan proses diskritisasi maka ada beberapa data set yang mengalami perubahan nilai error berbalik dengan kondisi sebelumnya terlihat pada grafik di gambar 4.2. Hal ini dikarenakan pada saat proses pemilihan atribut dengam metode Wrapper Feature Selection ada atribut yang terbuang dan meninggalkan atribut yang karakternya memiliki jumlah distinct kecil. Sehingga apabila data set tersebut melalui proses diskritisasi, nilai error akan semakin menurun apabila proses diskritisasi dilakukan dengan nilai minimum frekuensi (minbinsize) nya semakin besar



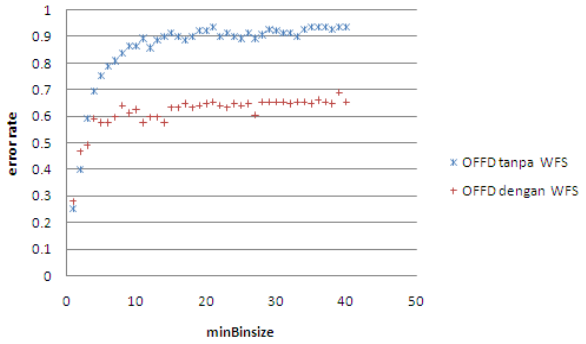
Gambar 8. Perubahan nilai error rate pada data set Wine



Gambar 9. Perubahan nilai error rate pada data set Sonar

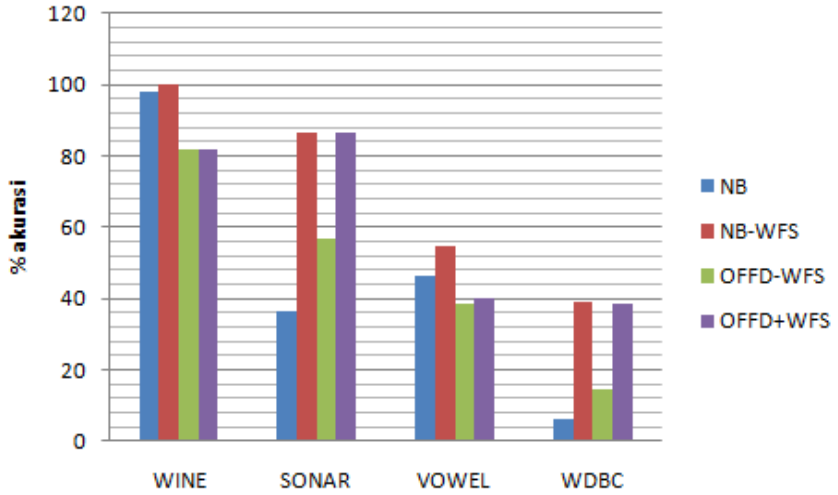


Gambar 10. Perubahan nilai error rate pada data set Vowel



Gambar 11. Perubahan nilai error rate pada data set Wdbc

Berdasarkan ke empat grafik di atas, terlihat bahwa pada metode OFFD (Optimal Flexible Frequency Discretization) yang lebih berpengaruh terhadap penurunan nilai error adalah tahap pemilihan atribut dengan metode wrapper feature selection hal ini dikarenakan atribut-atribut yang menyimpan informasi salah telah terbuang, untuk selanjutnya sistem hanya akan melakukan diskritisasi pada data set dengan atribut relevan. Jadi, selain untuk mengurangi waktu dalam ruang pemrosesan diskritisasi, proses wrapper feature selection juga mengurangi nilai error pada sebuah dataset.



Gambar 12. Grafik perubahan nilai error berdasarkan metode yang digunakan

Berdasarkan grafik di atas, untuk nilai akurasi hasil pemrosesan dengan sistem OFFD akan rendah apabila pada dataset tersebut memiliki nilai parameter selisih tinggi begitupun sebaliknya. Hal ini dikarenakan pada dataset dengan nilai selisih tinggi karakteristik nilai yang dimiliki sedikit yang bernilai sama sehingga ada kemungkinan salah pemetaan nilai ke dalam interval. Salah pemetaan nilai ke dalam sebuah interval dapat menghasilkan kesalahan informasi yang digunakan untuk pengklasifikasian sehingga bisa menimbulkan akurasi rendah. Selain itu, jumlah atribut yang dimiliki dataset akan mempengaruhi akurasi dataset tersebut sehingga untuk dataset dengan atribut numerik banyak preprocessing dengan metode OFFD akan meningkatkan akurasi dataset tersebut.

Sedangkan untuk nilai selisih rendah, karakteristik yang dimiliki oleh dataset terdapat banyak nilai yang sama sehingga pada proses pendiskritisasian akan dipilih nilai frekuensi besar untuk pembentukan interval sehingga akan terbentuk interval yang sedikit. Dikarenakan sedikitnya pembentukan interval dan pada intervalnya sendiri memiliki frekuensi yang tinggi maka pada saat perhitungan nilai akurasi dengan Naïve Bayes akan menghasilkan nilai peluang yang tinggi.

5.4 Kesimpulan

Berdasarkan pengujian yang dilakukan dalam penelitian ini, dapat disimpulkan bahwa :

a. Kerja metode OFFD (Optimal Flexible Frequency Discretization) ini terhadap keoptimalan penurunan nilai error nya lebih dipengaruhi oleh proses pemilihan atribut dengan pencarian Best First Search pada proses Wrapper Feature Selection.
b. Parameter minbinsize pada proses diskritisasi berbanding terbalik dengan nilai selisih antara jumlah instances dengan rata-rata distinct pada atribut-atribut yang dimiliki masing-masing dataset karena rule terbentuk yang digunakan untuk proses diskritisasi dalam pembentukan sebuah interval.
c. Perubahan nilai akurasi untuk dataset dengan nilai selisih rendah lebih bagus saat tanpa diproses dengan metode OFFD daripada nilai akurasi untuk dataset dengan nilai selisih tinggi saat diproses dengan metode OFFD dikarenakan ada kemungkinan informasi yang salah dalam pembentukan interval masing-masing atribut sehingga bisa membentuk informasi yang tidak tepat saat pengklasifikasian.

5.5 Saran

Saran yang diperlukan untuk percobaan selanjutnya sebagai berikut :

a. Untuk selanjutnya coba gunakan metode pemilihan atribut lainnya seperti Filter Feature Selection atau gunakan metode pemilihan atribut yang sama tetapi dengan algoritma pencarian yang berbeda seperti Hill Climbing.
b. Mencoba metode pendiskritan lainnya seperti Fuzzy Discretization.
c. Penanganan missing value pada dataset yang memilikinya

Daftar Pustaka

[1] Berzal,Fernando , Cubero,Juan-Carlos , Marin,Nicolás , José-Maria , Serrano, 2003 , Usability Issues in Data Mining Systems , Dept. Computer Science and Artificial Intelligence, E.T.S. Ingenieria Informatica, University of Almeria Ctra Sacramento.
[2] Clifton,Chris , Jiang,Wei , Muruguesan,Mummoorthy , Nergiz, M.Ercan , Is Privacy Still an Issue for Data Mining? (Extended Abstract), Dept. of Computer Science, Purdue University.
[3] Dunham, Margaret H. , 2003, New Jersey, Data Mining Introductory and Advanced Topics, Pearson Education Inc.
[4] Han,Jiawei and Micheline Kamber, 2001, Data Mining: Concepts and Techniques First Edition, Morgan Kaufmann Publishers, San Fransisco.
[5] Kohavi,Ron[1], George H John[2], Wrapper for Feature Selection [1]Data Mining and Visualization, Silicon Graphic Inc., 2011, N. Shoreline Boulevard, [2]Epiphany Marketing Software, 2141 landings drive mountain view CA 94043 USA.
[6] Moertini, /v.S., 2002, Data Mining sebagai Solusi Bisnis, Universitas Khatolik Parahyangan, Bandung. http://home.unpar.ac.id/ diunduh 21 November 2010.
[7] Ren,Jiangtao , Chen,Xianlu , Dept. of Computer Science, Sun Yat-sen University, China, Lee Den,Sau , Kao,Ben , Cheng,Reynold , Cheung,David , Dept. of Science The University of Hong Kong, Hongkong, Naïve-Bayes Classification of Uncertain Data. http://www.docjax.com/ diunduh pada 10 Oktober 2010.
[8] Wang,Zhihai , Wang,Song , Min,Fan , dan Cao,Tianyu , 2009 , OFFD: Optimal Flexible Frequency Discretization for Naïve-Bayes Classification, Springer-Verlag, Berlin Heidelberg. http://www.cs.uvm.edu/ diunduh pada 11 Oktober 2010.
[9] Weis, M.Sholom, Indurkhya, Nitin, 1998, Predictive Data Mining, Morgan Kaufmann Publisher, Inc., USA.
[10] W.Seifert,Jeffrey , 16 Desember 2004, “CRS Report for Congres. Data Mining : Overview” , Information Science and Technology Policy, Resources, Science, and Industry Division.
[11] Yang,Ying , I.Webb,Geoffrey , 2002, A Comparatives Study of Discretization Methods for Naïve-Bayes Classifier , School of Computer Science and Software Engineering, Monash University. http://citeseerx.ist.psu.edu/ diunduh pada 8 Oktober 2010.
Yang,Ying , I.Webb,Geoffrey , On Why Discretization Works for Naïve-Bayes Classifier , School of Computer Science and Software Engineering, Monash University. http://citeseerx.ist.psu.edu/ diunduh pada 8 Oktober 2010.