

Vol. 7 No.1

ISSN 2085 - 3092

SPIRIT

Sarana Pengembangan Informasi & Intelektual Terkini



STMIK YADIKA JOURNAL OF COMPUTING AND CYBERNETIC SYSTEM

S T M I K Y A D I K A B A N G I L

JURNAL
INFORMATIKA

Vol. 7

No.1

Halaman 1-70

Pasuruan
MEI-2015

ISSN
2085 - 3092

JURNAL SPIRIT

Sarana Pengembangan Informasi & Intelektual Terkini

Vol. 7, No. 1, Mei 2015

ISSN 2085 – 3092

JURNAL SPIRIT
JURNAL MANAJEMEN INFORMATIKA
Volume 7, Nomor 1

Ketua Redaksi
Dr. Syaiful Bukhori, M.Kom

Redaksi Pelaksana
Dr. Moh. Aris Sufagi, S.Pd, MT (STMIK Yadika)
Yusron Rijal, S.Si, MT (STMIK Yadika)
Muhammad Noval Riswanda, M.Kom (STMIK Yadika)
Teguh Pradana, M.Kom (STMIK Yadika)

Pelaksana Tata Usaha
Mohammad Imron

Alamat Sekretariat / Redaksi :
Pusat Penelitian STMIK Yadika
Jl. Bader 09 Kalirejo, Bangil 67153, Indonesia
Telp. 0343-742070, Fax. 0343-748914 E-mail : jurnal_informatika@stmik-yadika.com
<http://www.stmik-yadika.com>

Jurnal "SPIRIT" merupakan Jurnal Ilmiah untuk mengembangkan ilmu di bidang Teknik dan Manajemen Informatika

Jurnal Informatika diterbitkan oleh Jurusan Teknik Informatika dan Lembaga Pusat penelitian STMIK Yadika.

Redaksi mengundang para profesional dari dunia usaha, pendidikan dan peneliti untuk menulis mengenai perkembangan ilmu di bidang Teknik dan Manajemen Informatika.

Jurnal Informatika diterbitkan 2 (dua) kali dalam 1 tahun pada bulan Mei dan Nopember.

JURNAL SPIRIT

Sarana Pengembangan Informasi & Intelektual Terkini

3092

Vol. 7, No. 1, Mei 2015

ISSN 2085-3092

DAFTAR ISI

PEROTASI OTOMATIS PADA SISTEM PEMERIKSA LEMBAR JAWABAN

Yusron Rijal , Nur Choliz.....1-7

PENGGUNAAN DSS AHP, SAW DAN SMART UNTUK MENENTUKAN PRIORITAS PEMBANGUNAN DI KOTA MALANG

Deddy Kusbianto Purwoko Aji.....8-13

KLASIFIKASI BERBASIS GRAVITASI DATA DAN PROBABILITAS POSTERIOR

Muhamad Arief Hidayat , Arif Djunaidy.....14-22

PENERAPAN ALGORITMA C4.5 PADA SISTEM INFORMASI KLINIK MATA FITUR REGISTRASI PASIEN

Anang Andrianto,ST., MT , Zakiyah Qurrotul Aini , M. Arief Hidayat,S.Kom., M.Kom.....23-35

ANALISIS PENGGUNAAN SISTEM INFORMASI PADA PENGGUNA WEBSITE eCOMMERCE DENGAN PENDEKATAN MODEL UTAUT

(Studi Kasus Pengguna Toko *Fashion Online* www.zalora.co.id)

Hari Moertii , Endang Siti Astuti , Imam Suyadi36-47

KLASIFIKASI GAMBAR FOTO BERDASARKAN TEMPAT PENGAMBILAN DENGAN MENGGUNAKAN TRANSFORMASI HOUGH

Moch. Ali Rokhib, ²⁾Endang Setyati48-53

PENERAPAN ASPEK PEDAGOGIK UNTUK MEMBANGUN KOMPUTER GAME MERUPAKAN INTI DARI GAME PENDIDIKAN

Moh. Aries Syufagi.....54-64

PEMANFAATAN METODE MOMENT INVARIANT DAN MORFOLOGI DALAM MENENTUKAN KLASIFIKASI KANKER PAYUDARA

Nurul Fuad , Abdul Rokhim.....65-70

KLASIFIKASI BERBASIS GRAVITASI DATA DAN PROBABILITAS POSTERIOR

Muhamad Arief Hidayat¹⁾, Arif Djunaidy²⁾

¹⁾ Program Studi Sistem Informasi, Universitas Jember

²⁾ Jurusan Sistem Informasi, Institut Teknologi Sepuluh Nopember

email: arief.hidayat@unej.ac.id

Abstract:

The classification method based on data gravitation (DGC) is one of the new classification techniques that uses data gravitation as the criteria of the classification. In the case of DGC, an object is classified on the basis of the class that creates the largest gravitation in that object. However, the DGC method may cause inaccurate result when the training data being used suffer from the class imbalanced problem. This may be caused by the existence of the training data containing a class having excessively big mass that will in turn tend to classify an unknown object as a member of that class due to the high degree of the data gravitation produced, and vice versa.

In this research, a modification to the DGC method is performed by constructing a classification method that is based on both the data gravitation and posterior probability (DGCPP). In DGCPP, the mass concept defined in the DGC method as the prior probability is replaced by the posterior probability. By using this modification, data gravitation calculation process is expected to produce more accurate results in compared to those produced by the DGC method. In addition, by improving the data gravitation calculation, it is expected that the DGCPP method will produce more accurate classification results in compared to those produced by the DGC method for both normal dataset as well as dataset having class imbalanced problems. A thorough tests for evaluating the classification accuracy are performed using a ten-fold cross-validation method on several datasets containing both normal and imbalanced-class datasets. The results showed that DGCPP method produced positive average of accuracy differences in compared to those produced by the DGC method. For the tests using the entire normal datasets showed that the average of accuracy differences are statistically significant with a 95% confidence level. In addition, results of the tests using the four imbalanced-class datasets also showed that the average accuracy differences are statistically significant with a 95% confidence level. Finally, results of the tests for evaluating the computing times required by the classification program showed that the additional computing time needed by DGCPP method to perform the classification process is insignificant and less than the human response time, in compared to that needed by DGC method for running all datasets being used.

Keywords—data gravitation-based classification, class imbalanced problem, posterior probability

1. Pendahuluan

Klasifikasi merupakan kegiatan untuk menggolongkan sebuah obyek sebagai kelas tertentu. Proses klasifikasi dilakukan dengan menggunakan model klasifikasi. Sebuah obyek yang belum diketahui kelasnya diprediksi kelasnya oleh model klasifikasi berdasar nilai fitur - fiturnya.

Saat ini terdapat banyak algoritma pembelajaran untuk membangun model klasifikasi seperti *Hierarchical SVM* [1], *Two-Stage Fuzzy Classification Model* [2], *Alert Classification Model* [3] dan lain - lain. Beberapa algoritma pembelajaran seperti *Nearest Class Mean* menggunakan kriteria *distance* terdekat antara obyek dengan pusat massa kelas sebagai kriteria klasifikasi sebuah obyek.

Klasifikasi Berbasis Gravitasi Data atau *Data GravitationBased Classification* (DGC) [4][5] merupakan algoritma pembelajaran atau teknik klasifikasi yang dapat dianggap sebagai pengembangan teknik klasifikasi berbasis *distance*. Pada metode DGC, selain *distance* ditambahkan konsep massa yaitu banyaknya data pelatihan yang menjadi anggota sebuah kelas. Terinspirasi dari teori gravitasi newton, metode DGC mengusulkan lebih jauh bahwa terdapat gravitasi data antara obyek yang akan diklasifikasi dengan kelas - kelas yang ada. Proses klasifikasi pada metode DGC dilakukan dengan menggunakan kriteria gravitasi data terbesar untuk mengklasifikasikan sebuah obyek.

Metode DGC memberikan hasil klasifikasi yang baik untuk data pelatihan normal. Namun

metode DGC memiliki kekurangan yaitu memberikan akurasi yang rendah jika data pelatihan yang digunakan tidakimbang [5]. Jika pada data pelatihan terdapat kelas yang massanya sangat kecil atau sangat besar dibandingkan kelas – kelas lain, maka akurasi klasifikasi metode DGC menjadi rendah.

Pada penelitian ini diajukan modifikasi metode DGC, yaitu metode Klasifikasi Berbasis Gravitasi Data dan Probabilitas Posterior (DGCPP), untuk meningkatkan akurasi DGC pada *dataset*imbang maupun yang mengalami *class imbalance problem*. Pada modifikasi yang diajukan, konsep massa kelas diinterpretasikan sebagai probabilitas prior kelas tersebut. Dengan demikian massa sebuah kelas merepresentasikan probabilitas sebuah obyek (yang akan diklasifikasi) adalah anggota kelas tersebut. Dengan menginterpretasikan massa sebagai probabilitas prior, muncul gagasan untuk mengganti penggunaan massa atau probabilitas prior pada DGC dengan probabilitas posterior yang lebih baik untuk klasifikasi. Penggantian massa atau probabilitas prior dengan probabilitas posterior ini diharapkan dapat meningkatkan kualitas perhitungan gravitasi data. Dengan meningkatnya kualitas perhitungan gravitasi data, diharapkan proses klasifikasi yang menggunakan kriteria gravitasi data juga memberikan hasil yang lebih baik. Uji coba menggunakan metode *Ten Fold Cross Validation* pada 4 *dataset* normal dan 4 *dataset* yang mengalami *class imbalance problem* menunjukkan metode DGCPP memiliki *mean* selisih akurasi positif dari metode DGC. Dari 4 *dataset* normal, 3 *dataset* nilai *mean* selisih akurasi signifikan secara statistik pada confidence level 95%. Dari 4 *dataset* yang mengalami *class imbalance problem*, 2 *dataset* nilai *mean* selisihnya akurasi signifikan secara statistik pada confidence level 95% dan 98%.

2. Tinjauan Pustaka

A. Konsep Gravitasi Data

Gravitasi data merupakan konsep yang diinspirasi dari teori gravitasi Newton [4][5]. Konsep ini menyatakan bahwa antara sebuah obyek yang akan diklasifikasi dan sebuah kelas

pada data pelatihan terdapat gravitasi data yang besarnya ditentukan oleh *distance* obyek dengan pusat massa partisi kelas dan massa partisi kelas tersebut. Berikut ini akan didefinisikan beberapa terminologi yang digunakan pada konsep gravitasi data

Definisi 1 (Partikel data) partikel data adalah partisi data pelatihan yang memiliki kelas sama dan *distance* antara sembarang dua anggotanya kurang dari ambang batas tertentu. Sebuah partikel data dibuat dengan menggunakan prinsip *Minimum Distance Principle* (MDP). Sebuah anggota data pelatihan dipilih secara acak sebagai anggota awal partikel data tersebut. Kemudian dicari anggota data pelatihan lain yang kelasnya sama dan *distancenya* kurang dari radius tertentu dari data pelatihan yang telah terpilih. Bila terdapat data pelatihan lain yang memenuhi syarat tersebut, maka dimasukkan sebagai anggota partikel data kemudian pusat massa partikel data diupdate. Hal yang sama dilakukan ulang sampai tidak ditemukan data pelatihan yang memenuhi kriteria.

Definisi 2 (massa) massa sebuah partikel data adalah banyaknya data pelatihan yang menjadi anggota partikel data tersebut.

Definisi 3 (pusat massa) pusat massa sebuah partikel data adalah pusat geometris dari partikel data tersebut. Misalnya terdapat sebuah partikel data X pada *data space* berdimensi n . Partikel X terdiri atas m anggota (data pelatihan) yaitu X_1, X_2, \dots dan X_m . Pusat massa dari X , $X_0 = (X_{01}, X_{02}, \dots, X_{0n})$, dihitung dengan persamaan

$$x_{0j} = \frac{\sum_{i=1}^m x_{ij}}{m}, i = 1, 2, \dots, m \quad j = 1, 2, \dots, n \quad (1)$$

Dengan X_{0j} merupakan nilai pusat massa untuk atribut ke j dan X_{ij} adalah nilai atribut ke j pada anggota partikel ke i .

Definisi 4 (partikel data tunggal) partikel data tunggal adalah partikel data yang massanya 1. Sebuah obyek yang akan diklasifikasi dapat dipandang sebagai partikel data tunggal.

Definisi 5 (gravitasi data) gravitasi data merupakan ukuran similarity antara partikel data dan merupakan besaran skalar. Inilah perbedaan antara gravitasi data dengan gravitasi newton yang merupakan besaran vektor. Hukum gravitasi data menyatakan bahwa gravitasi antara dua partikel data pada *data space* merupakan

rasio dari perkalian massa dua partikel tersebut dengan kuadrat *distance* antara pusat massa dua partikel tersebut. Secara matematis,

$$F = \frac{m_1 m_2}{d^2} \quad (2)$$

F adalah gravitasi data antara partikel 1 dan 2, m_1 merupakan massa partikel 1, m_2 merupakan massa partikel 2 dan d merupakan *euclidean distance* antara pusat massa dua partikel.

B. Klasifikasi Berbasis Gravitasi Data

Metode DGC menggunakan gravitasi data sebagai kriteria klasifikasi [4][5]. Pada metode DGC, obyek diklasifikasikan sebagai kelas yang menghasilkan gravitasi data terbesar pada obyek tersebut.

Misalnya pada data pelatihan terdapat k kelas, yaitu C_1, C_2, \dots dan C_k . Masing masing kelas memiliki anggota sebanyak L_1, L_2, \dots dan L_k . Masing – masing kelas dipartisi menjadi T_1, T_2, \dots dan T_k partikel data. Sebuah data atau obyek X yang akan diklasifikasi dapat dianggap sebagai partikel data tunggal dengan nilai pusat massa sama vektor fiturnya. Gravitasi data kelas C_i pada obyek yang akan diklasifikasi dapat dihitung menggunakan persamaan

$$F_i = \sum_{j=1}^{T_i} \frac{m_{ij}}{|x_{ij} - x|^2} \quad (3)$$

Dengan F_i adalah superposisi atau total gravitasi data kelas i pada X , m_{ij} adalah massa partikel j pada kelas C_i dan X_{ij} merupakan pusat massa partikel tersebut. Dengan menggunakan persamaan 3 dapat dicari kelas yang menghasilkan gravitasi data terkuat pada obyek yang akan diklasifikasi.

Pada *distance* $|x_j - x|^2$ di persamaan 3 dimasukkan faktor bobot tiap fitur untuk meningkatkan akurasi klasifikasi. Pada DGC, pembobotan dilakukan dengan metode TRFS (*Tentative Random Selection Features*) [5] yang mensimulasikan proses mutasi algoritma genetic untuk mencari kombinasi bobot terbaik.

C. Metode TRFS Untuk Pembobotan Atribut

Pada tabel 1 ditunjukkan algoritma TRFS (*Tentative Random Feature Selection*) yang digunakan untuk mencari bobot atribut data yang menghasilkan akurasi terbaik pada DGC.

Pada algoritma TRFS, mula – mula data pelatihan dipartisi menjadi 2 partisi secara proporsional. Setelah dipartisi, dilakukan iterasi untuk pembobotan. Pada setiap iterasi, dipilih secara acak bobot atribut tertentu untuk diubah nilainya. Nilai bobot yang baru dievaluasi dengan mekanisme *cross validation* menggunakan 2 partisi data yang telah dibuat. Jika rata – rata akurasi yang dihasilkan mekanisme *cross validation* lebih baik dari akurasi bobot sebelumnya, maka bobot yang baru digunakan. Hal yang sama dilakukan pada setiap iterasi sampai mencapai iterasi maksimal atau bobot yang didapatkan mencapai akurasi yang diharapkan.

Tabel 1 Algoritma TRFS

1	Split data pelatihan menjadi dua subset Ta dan Tb
2	W0, Pi / banyaknya atribut, fo
3	For i to i = i_max or f < fo
4	Pilih Wx dari W secara acak dengan mempertimbangkan P
5	W' = W + ε
6	Evaluasi w' menggunakan <i>cross validation</i> pada Ta dan Tb, hasilnya adalah f'
7	If f < f'
8	W = W'
9	f = f'
10	Px = Px + δ
11	Else
12	If Px >
13	Px = Px - δ
14	Else
15	Px = 0
16	End if
17	End if
18	End for

D. Perhitungan Probabilitas Posterior

Anggaplah bahwa X adalah himpunan atribut sebuah obyek dan Y merupakan kelas obyek tersebut. Jika antara X dan Y tidak terdapat hubungan deterministik, maka X dan Y dapat diperlakukan sebagai variabel acak. Sebagai variabel acak, hubungan keduanya dapat dinyatakan dengan $P(Y|X)$. $P(Y|X)$ melambangkan peluang obyek tersebut merupakan kelas Y jika diketahui nilai atribut – atributnya adalah X . $P(Y|X)$ disebut probabilitas

posterior. Terdapat juga probabilitas prior $P(Y)$, yaitu peluang obyek tersebut merupakan kelas Y tanpa mempertimbangkan nilai X

Probabilitas posterior digunakan untuk memprediksi kelas Y alih alih probabilitas prior karena memasukkan faktor nilai attribute sehingga lebih presisi [6]. Metode *Naive bayes* juga menggunakan probabilitas posterior untuk melakukan proses klasifikasi. Sebuah obyek diklasifikasikan sebagai kelas yang probabilitas posteriornya paling besar. Pada metode *naive bayes* probabilitas posterior dihitung dengan persamaan

$$P(Y|X) = P(Y) \prod_{i=1}^d P(X_i|Y) \quad (4)$$

$P(Y)$ atau probabilitas prior merupakan proporsi data pelatihan yang berkelas Y dan $P(X_i|Y)$ merupakan proporsi nilai attribute X_i pada data pelatihan yang memiliki jenis kelas Y .

Untuk menghitung $P(X_i|Y)$ pada atribut kontinyu digunakan persamaan sebagai berikut

$$P(X = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \quad (5)$$

Parameter μ_{ij} dapat diestimasi berdasarkan sampel mean $X_i(x)$ untuk seluruh data pelatihan yang berkelas y_j . Dengan cara sama, σ_{ij}^2 dapat diestimasi dari sampel varian (s^2) data pelatihan yang berkelas y_j .

Persamaan 4 mengasumsikan bahwa atribut – atribut data tidak berkorelasi. Secara teoritis, bila metode *Naive bayes* diterapkan pada data yang atribut – atributnya berkorelasi akan menurunkan akurasi klasifikasi. Meskipun demikian, hasil uji coba secara empiris menunjukkan bahwa metode *Naive bayes* secara mengejutkan memberikan hasil baik jika diuji coba pada data yang atribut – atributnya berkorelasi [7][8][9]. Untuk mendapatkan hasil yang lebih baik jika atribut – atribut datanya berkorelasi, dapat digunakan persamaan *multivariate gaussian distribution*

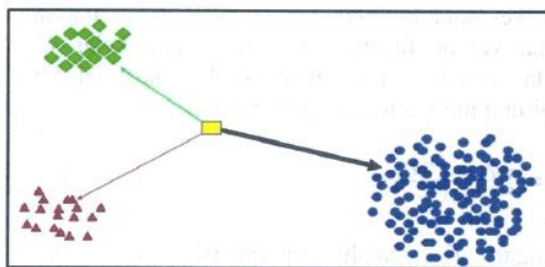
E. Penelitian Terkait Sebelumnya

Metode DGC merupakan metode klasifikasi baru yang diajukan [4][5]. Metode ini dapat

dianggap sebagai pengembangan teknik klasifikasi berbasis *distance*. Pada DGC, selain *distance*, ditambahkan konsep massa dan gravitasi data. Klasifikasi dilakukan menggunakan gravitasi data.

Metode DGC memberikan hasil klasifikasi yang baik untuk data pelatihan normal. Kelebihan lain dari metode DGC adalah efisien dan prinsip yang mendasari metode tersebut mudah dipahami serta mudah diimplementasikan.

Namun untuk data pelatihan yang tidak seimbang, metode DGC memberikan hasil yang buruk [6]. Pada gambar 1 ditunjukkan terdapat sebuah kelas pada data pelatihan yang massanya atau banyaknya data pelatihan yang menjadi anggota kelas tersebut sangat besar (berwarna biru). Akibatnya, gravitasi data kelas tersebut menjadi sangat kuat. Semua obyek pada data uji cenderung diklasifikasi sebagai kelas yang memiliki massa sangat besar tersebut. Hal yang sebaliknya juga berlaku jika terdapat kelas yang massanya sangat kecil



Gambar 1 Pada metode DGC, bila terdapat sebuah kelas yang massanya sangat besar, semua data uji cenderung diklasifikasikan sebagai kelas tersebut

3. Pengembangan Klasifikasi Berbasis Gravitasi Data

Untuk meningkatkan akurasi metode DGC jika data pelatihan yang digunakan tidak seimbang, dilakukan beberapa modifikasi sebagai berikut

- 1) Menginterpretasikan konsep massa kelas sebagai probabilitas prior

Jika massa sebuah kelas pada persamaan gravitasi data diganti dengan proporsi kelas tersebut pada data pelatihan, maka nilai gravitasi datanya memang berubah namun hasil klasifikasi akhir tetap. Hal ini disebabkan karena massa setiap kelas proporsional dengan proporsinya pada data pelatihan. Dengan demikian persamaan

gravitasi data dapat ditulis ulang dengan mengganti massa dengan proporsi. Pada kasus data pelatihan di mana kelas C_i hanya memiliki satu partikel data, gravitasi data C_i pada obyek yang akan diklasifikasi dinyatakan dengan persamaan

$$F_i = \frac{\text{proporsi}(C_i)}{d_i^2} \tag{6}$$

Dari sudut pandang metode Naive bayes, proporsi sebuah kelas pada data pelatihan dianggap probabilitas prior obyek X merupakan anggota kelas tersebut [6]. Sehingga pada perhitungan gravitasi data, massa kelas atau proporsi dapat diganti dengan probabilitas prior kelas tersebut. Persamaan 6 dapat ditulis ulang menjadi

$$F_i = \frac{P(C_i)}{d_i^2} \tag{7}$$

Hasil klasifikasi bila gravitasi data dihitung menggunakan persamaan 7 tidak berubah meskipun nilai gravitasi data tidak sama. Hal ini menunjukkan bahwa konsep massa pada DGC dapat diinterpretasikan sebagai probabilitas prior. Interpretasi dan penggantian massa dengan probabilitas prior tersebut merupakan hal penting karena membuka peluang untuk mengeksplorasi konsep gravitasi data melalui sudut pandang *bayesian learning*.

2) Mengganti massa (probabilitas prior) dengan probabilitas posterior

Pada metode *bayesian learning*, probabilitas posterior dianggap lebih akurat untuk klasifikasi dibandingkan dengan probabilitas prior. Karena itu, modifikasi kedua yang diajukan untuk meningkatkan akurasi metode DGC adalah mengganti massa atau probabilitas prior dengan probabilitas posterior.

Karena probabilitas posterior lebih akurat dibanding probabilitas prior, diharapkan penggunaan probabilitas posterior untuk menggantikan massa (yang ekuivalen dengan probabilitas prior) akan meningkatkan kualitas perhitungan gravitasi data sekaligus meningkatkan akurasi klasifikasi metode DGC pada kasus data pelatihan yang tidakimbang.

Dengan demikian, pada kasus data pelatihan di mana kelas C_i hanya memiliki satu partikel data, gravitasi data C_i pada obyek yang akan diklasifikasi dapat ditulis ulang menjadi

$$F_i = \frac{P(C_i)}{d_i^2} \tag{8}$$

Persamaan 8 tidak memberikan hasil klasifikasi sama dengan persamaan 6 dan 7

3) Modifikasi persamaan gravitasi data

Penggantian massa pada DGC dengan probabilitas posterior membutuhkan beberapa perubahan pada persamaan untuk menghitung gravitasi data jika kelas terdiri atas banyak partikel data. Pada DGC massa sebuah kelas dipartisi menjadi partikel - partikel data. Karena pada modifikasi metode DGC massa diganti dengan probabilitas posterior, maka probabilitas posterior tersebut juga harus dipartisi menjadi partikel partikel. Setiap partikel mendapat potongan probabilitas posterior sesuai dengan proporsi massanya pada kelas tersebut. Potongan probabilitas posterior tersebut menggantikan massa partikel pada persamaan untuk menghitung gravitasi data

Misalnya pada data pelatihan terdapat kelas, yaitu C_1, C_2, \dots dan C_k . Masing masing kelas memiliki anggota sebanyak L_1, L_2, \dots dan L_k . Masing - masing kelas dipartisi menjadi T_1, T_2, \dots dan T_{L_i} partikel data. Sebuah data atau obyek X yang akan diklasifikasi dapat dianggap sebagai partikel data tunggal dengan nilai pusat massa sama vektor fiturnya. Menggunakan prinsip superposisi pada persamaan 3, gravitasi data kelas C_i pada obyek X dapat dihitung menggunakan persamaan

$$F_i = \sum_{j=1}^{L_i} \frac{P(C_i|X) \frac{m_{ij}}{L_i}}{d_{ij}^2} \tag{9}$$

F_i adalah superposisi atau total gravitasi data kelas i pada X , $P(C_i|X)$ merupakan probabilitas posterior X merupakan anggota kelas i , m_{ij} adalah massa partikel T_j pada kelas C_i , L_i merupakan banyaknya data pelatihan pada kelas i dan d_{ij} adalah *distance* antara X dengan pusat massa partikel T_j pada kelas C_i . $P(C_i|X)$ dihitung dengan

persamaan 6. Sedangkan d_p dihitung menggunakan persamaan 3.

4. Hasil Penelitian

A. Data Dan Skenario Uji Coba

Uji coba yang dilakukan pada penelitian ini menggunakan 8 dataset dari dataset yang digunakan pada [5]. Empat dataset yang digunakan merupakan dataset normal, yaitu segment, sonar, vehicle dan wine. Empat dataset yang lain mengalami class imbalance problem, antara lain glass, ionosphere, pimdian WBCD.

Uji coba dilakukan dalam 2 skenario,

1) Skenario 1

Uji coba skenario 1 dilakukan dengan membandingkan akurasi dan waktu klasifikasi metode DGC dan DGCPP menggunakan metode Full Train Full Test dengan parameter radius yang bervariasi antara radius minimum 0 hingga radius K di mana sebuah kelas pada data pelatihan menjadi anggota dari sebuah partikel data

2) Skenario 2

Uji coba skenario 2 dilakukan dengan membandingkan akurasi metode DGC dan DGCPP menggunakan metode Ten Fold Cross Validation dengan parameter terbaik, yaitu radius 0

B. Hasil Dan Pembahasan Uji Coba Skenario 1

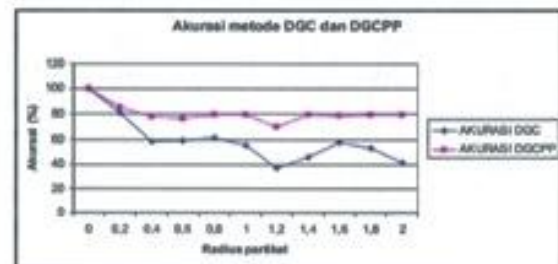
Tabel 2 menunjukkan akurasi uji coba skenario 1 metode DGC dan DGCPP untuk dataset normal. Dari tabel 2 dapat diamati bahwa untuk dataset normal, metode DGCPP hampir selalu memiliki akurasi lebih baik dari atau sama dengan metode DGC pada hampir semua nilai radius.

Tabel 3 menunjukkan akurasi uji coba skenario 1 metode DGC dan DGCPP untuk dataset yang mengalami class imbalance problem. Dari tabel 3 dapat diamati bahwa untuk dataset yang mengalami class imbalance problem, metode DGCPP hampir selalu memiliki akurasi lebih baik dari atau sama dengan metode DGC pada hampir semua nilai radius.

Gambar 2 menunjukkan grafik akurasi uji coba skenario 1 untuk dataset vehicle. Gambar 2 mewakili karakteristik hampir semua hasil uji

coba skenario 1. Dari gambar 2 dapat diamati beberapa karakteristik akurasi klasifikasi metode DGC dan DGCPP

- 1) Nilai akurasi terbaik metode DGC dan DGCPP pada uji coba skenario 1 yang menggunakan metode Full Train Full Test relatif sama. Namun hal ini tidak berarti bahwa pada metode uji coba lain memberi hasil yang sama seperti yang ditunjukkan pada uji coba skenario 2.
- 2) Akurasi metode DGC dan DGCPP semakin rendah bila radius partikel yang digunakan semakin besar
- 3) Grafik akurasi metode DGC dan DGCPP memiliki bentuk hampir sama, namun grafik akurasi metode DGCPP berada di atas grafik akurasi metode DGC
- 4) Makin besar ukuran radius partikel data, selisih akurasi metode DGC dan DGCPP semakin besar. Akurasi metode DGC dan DGCPP



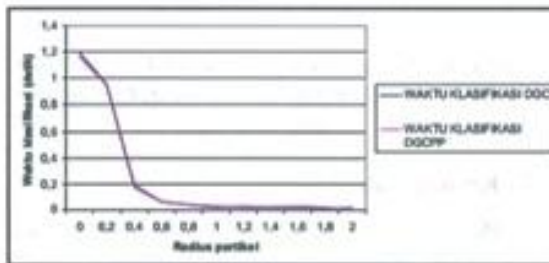
Gambar 2 grafik akurasi metode DGC dan DGCPP untuk dataset vehicle pada skenario 1

Tabel 4 menunjukkan waktu klasifikasi uji coba skenario 1 metode DGC dan DGCPP untuk dataset normal. Dari tabel 4 dapat diamati bahwa untuk dataset normal, kedua metode memiliki waktu klasifikasi hampir sama. Selisih waktu klasifikasi untuk semua data pelatihan bernilai kurang dari human response time.

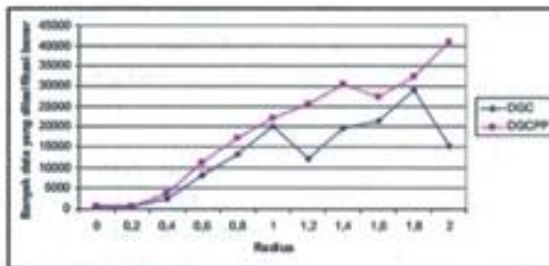
Tabel 5 menunjukkan waktu klasifikasi uji coba skenario 1 metode DGC dan DGCPP untuk dataset yang mengalami class imbalance problem. Dari tabel 5 dapat diamati bahwa untuk dataset yang mengalami class imbalance problem, kedua metode memiliki waktu klasifikasi hampir sama. Selisih waktu klasifikasi untuk semua data pelatihan bernilai kurang dari human response time.

Gambar 3 menunjukkan grafik waktu klasifikasi uji cobaskenario 1 untuk *datasetvehicle*. Gambar 3 mewakilikarakteristik hampir semua hasil uji waktu skenario 1. Padagambar 3 dapat diamati bahwa jika radius partikel bertambah,waktu klasifikasi metode DGC dan DGCPP akan menurun.

Gambar 4 menunjukkan banyaknya data yang diklasifikasibenar per satuan waktu (detik) untuk *datasetvehicle*. Gambar4 menunjukkan salah satu kelebihan metode DGCPPdibanding DGC, yaitu data yang diklasifikasi benar per satuanwaktu DGCPP lebih dari DGC. Hampir semua *dataset*memiliki karakteristik seperti demikian kecuali *datasetpima*.



Gambar 3 grafik waktu klasifikasi metode DGC dan DGCPP untuk *dataset vehicle* pada skenario 1



Gambar 4 Grafik banyaknya data yang diklasifikasi benar per detik untuk *dataset vehicle*

C. Hasil Dan Pembahasan Uji Coba Skenario 2

Tabel 2 Hasil Uji Coba Skenario 2 Untuk *Dataset Normal*

	segment	sonar	vehicle	wine
Meanselisih akurasi DGCPP – DGC	+	+	+	+
FoldDGCPP menang	7	5	9	9
Folddraw	2	1	0	1

FoldDGCPP kalah	1	4	1	0
signifikan	Ya (95%)	tidak	Ya (95%)	Ya (95%)

Tabel 3 Hasil Uji Coba Skenario 2 Untuk *Dataset Normal*

	glass	ionosphere	pima	WBCD
Meanselisih akurasi DGCPP – DGC	+	+	+	+
FoldDGCPP menang	3	3	7	8
Folddraw	5	7	0	2
FoldDGCPP kalah	2	0	3	0
signifikan	tidak	Ya (98%)	tidak	Ya (98%)

Dari rekapitulasi hasil uji coba skenario 2 yang ditunjukkanpada tabel 3 dan 4 dapat diketahui bahwa metode DGCPPmemiliki meanselisih akurasi positif untuk seluruh *dataset*,baik yang normal maupun yang mengalami *class imbalance problem*. Dari 8 *dataset*tersebut, pada 4 *dataset*(segment,vehicle, wine dan WBCD) meanselisih akurasinya sigifikanpada confidence level 95%. Dari 4 sisanya, pada 1 *dataset*(ionosphere) signifikan pada confidence level 98%. Sedangpada *dataset*sonar, glass dan pima meanselisihnya tidaksignifikan secara statistik.

Dari uji coba skenario 2, dapat disimpulkan bahwa secara umum metode DGCPP memiliki akurasi yang lebih baik darimetode DGC pada parameter optimal, yaitu radius partikel 0.perbedaan dengan karakteristik 1 uji coba skenario 1dijelaskan sebagai berikut. Pada metode uji skenario 1 yangmenggunakan metode *Full Train Full Test*, data pelatihayang digunakan sama dengan data pengujian, yaitu seluruh*dataset*. Hal ini mengakibatkan classifier sangat sesuai dengandata uji

D. Analisa Migrasi Klasifikasi

Analisa migrasi klasifikasi dilakukan dengan mengamataidata – data yang diklasifikasi secara salah oleh metode DGCnamun diklasifikasi benar oleh metode DGCPP. Analisa inibertujuan

untuk membuktikan apakah pada *dataset* yang mengalami *class imbalance problem*, klasifikasi menggunakan metode DGC menyebabkan data yang sebenarnya berjeniskelas yang massanya kecil diklasifikasikan sebagai kelas yang massanya besar. Tujuan lain dari analisa migrasi data ini adalah untuk membuktikan apakah pada klasifikasi menggunakan metode DGCPP kesalahan klasifikasi tersebut dapat diperbaiki.

Tabel 4 Recall klasifikasi Metode DGC dan DGCPP *Dataset* Ionosphere pada radius partikel 3

	kelas g (225)		kelas B (126)	
Metode	DGC	DGCPP	DGC	DGCPP
Recall class	0,93	0,96	0,40	0,77

Tabel 5 Recall klasifikasi metode DGC dan DGCPP *dataset* pima pada radius partikel 1,5

	kelas 0 (500)		kelas 1 (268)	
Metode	DGC	DGCPP	DGC	DGCPP
Recall class	0,83	0,792	0,485	0,652

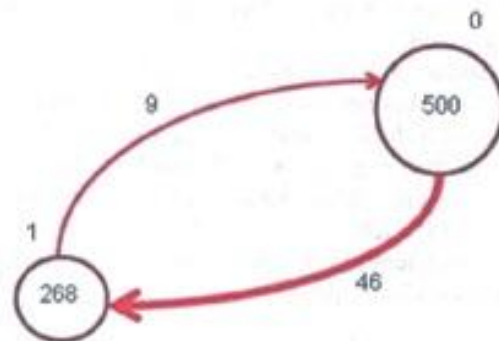
Tabel 6 recall klasifikasi metode DGC dan DGCPP *dataset* Wbed pada radius partikel 1,75

	kelas 2 (444)		kelas 4 (239)	
Metode	DGC	DGCPP	DGC	DGCPP
Recall class	0,986	0,981	0,857	0,924

Dari tabel 4 hingga 6 dapat diamati beberapa hal penting. Hal pertama yaitu *recall* kelas yang massanya kecil cenderung bernilai rendah pada metode DGC. Hal ini dapat diartikan bahwa pada metode DGC, terdapat banyak data uji yang sebenarnya berjenis kelas yang massanya kecil, diklasifikasikan secara keliru sebagai kelas yang massanya besar. Kedua, pada metode DGCPP, *recall* kelas yang massanya kecil naik dibandingkan dengan nilai *recall* pada metode DGC. Hal ini dapat ditafsirkan sebagai berikut. Pada metode DGCPP, sejumlah data dari kelas yang massanya kecil yang diklasifikasikan secara keliru sebagai kelas yang massanya besar oleh metode DGC, diklasifikasikan secara benar sesuai kelasnya oleh metode DGCPP

Banyak data yang awalnya diklasifikasi DGC berkelas yang massanya besar, diklasifikasi sebagai kelas yang massanya kecil. Pada gambar

5 ditunjukkan terdapat 46 data yang sebenarnya berkelas 1 (massa 268) diklasifikasikan sebagai 0 (massa 500) oleh DGC. Oleh DGCPP, data tersebut diklasifikasi secara benar sebagai kelas 1.



Gambar 5 Migrasi data yang diklasifikasi salah oleh DGC dan diklasifikasi benar oleh DGCPP untuk *dataset* pima

5. Kesimpulan

Berdasarkan hasil uji coba yang telah dilakukan dapat diambil kesimpulan sebagai berikut

- 1) Pada parameter radius partikel minimum, metode DGCPP memberikan hasil klasifikasi yang lebih baik dari DGC pada *dataset* normal maupun *dataset* yang mengalami *class imbalance problem* seperti yang ditunjukkan pada uji coba skenario 2.
- 2) Semakin besar ukuran radius partikel yang digunakan, akurasi metode DGC dan DGCPP makin rendah. Namun penurunan akurasi metode DGC lebih cepat dibandingkan dengan metode DGCPP. Akibatnya, semakin besar ukuran radius partikel data yang digunakan, selisih akurasi metode DGCPP semakin melampaui metode DGC
- 3) Penambahan waktu klasifikasi metode DGCPP dibandingkan metode DGC sangat kecil, kurang dari *human response time* untuk klasifikasi seluruh data uji. Dari segi banyaknya data yang diklasifikasi benar persatuan waktu, metode DGCPP mengungguli metode DGC
- 4) Metode DGCPP mengatasi kelemahan misklasifikasi yang dilakukan metode DGC pada *dataset* yang mengalami *class imbalance problem*

Daftar Pustaka

- [1] Hao, Pei-Yi, Chiang, Jung-Hsien dan Tu, Yi-Kun, 2007, "*Hierarchically SVM classification based on support vector clustering method and its application to document categorization*", *Expert Systems with Applications*, 33 (2007), 627-635
- [2] Li, Tzuu-Hseng S., Guo, Nai Ren dan Cheng, Chia Ping, 2008, "*Design of a two-stage fuzzy classification model*", *Expert Systems with Applications*, 35 (2008), 1482-1495
- [3] Jan, Nien-Yi, Lin, Shun-Chieh, Tseng, Shian-Shyong dan P. Lin, Nancy, 2009, "*A decision support system for constructing an alert classification model*", *Expert Systems with Applications*, 36 (2009), 11145-11155
- [4] Peng, Lizhi, Yang, Bo dan Chen, Yuehui 2005, "*A Novel Classification Method Based on Data Gravitation*", *Proc. Of International Conference on Neural Networks and Brain (ICNN&B)*, 667-672, 2005.
- [5] Peng, Lizhi, Yang, Bo, Chen, Yuehui dan Abraham, Ajith, 2009, "*Data Gravitation Based Classification*", *Information Sciences*, 179, 809-819
- [6] Tan, P.N., Steinbach, M. dan Kumar, V., 2006, "*Introduction to Data Mining*", Pearson Education, Inc., Boston.
- [7] Li, Yumei dan Anderson-Sprecher, Richard, 2006, "*Facies identification from well logs: A comparison of discriminant analysis and naive bayes classifier*", *Journal of Petroleum Science and Engineering*, 53 (2006), 149-157
- [8] Rish, Irina, 2001, "*An empirical study of the Naive bayes classifier*", *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*.
- [9] Turhan, Burak dan Bener, Ayse, 2009, "*Analysis of Naive bayes assumptions on software fault data : An empirical study*", *Data & Knowledge Engineering*, 68 (2009), 278-290