KLASIFIKASI BERBASIS GRAVITASI DATA DAN PROBABILITAS POSTERIOR

Muhamad Arief Hidayat[1)], **Arif Djunaidy**[2)]

[1)] Program Studi Sistem Informasi, Universitas Jember
[2)]JurusanSistem 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 uknown 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 classificaion 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 addtion, 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 olehmodel 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 pembelajaranseperti *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 algoritmapembelajaran 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 tidak imbang [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 kelasdiinterpretasikan 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*yangmengalami *class imbalance problem* menunjukkan metode DGCPP memiliki *mean* selisih akurasi positif dari metode DGC. Dari 4 *dataset* normal, 3 *dataset* nilai *mean* selisih akurasinya signifikan secara statistik pada confidence level 95%. Dari 4 *dataset* yang mengalami *class imbalance problem*, 2*dataset*nilai *mean* selisihnya akurasinya signifikansecara 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 datapelatihan 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 datapelatihan dipilih secara acak sebagai anggota awal partikel data tersebut. Kemudian dicari anggota data pelatihan lain yang kelasnya sama dan *distance*nya 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 adalahbanyaknya data pelatihan yang menjadi anggota partikel data tersebut.

Definisi 3 (pusat massa) pusat massa sebuah partikel dataadalah 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,... dan X_m. Pusat massa dari X, X_0= (X_{01}, X_{02}, ,..., X_{0n}),dihitung dengan persamaan

$$x_{oj} = \frac{\sum_{i=1}^{m} x_{ij}}{m}, i = 1, 2, \dots, m \;\; j = 1, 2, .., n \qquad (1)$$

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

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

Definisi 5 (gravitasi data) gravitasi data merupakan ukuransimilarity 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 *dataspace* merupakan

rasio dari perkalian massa dua partikeltersebut 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, $m_1$merupakanmassa 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,... dan C_k. Masing masing kelas memiliki anggota sebanyak L_1, L_2, ... dan L_k. Masing – masing kelas dipartisi menjadi T_1, T_2, ... dan $T_k$partikel data. Sebuah data atau obyek *X* yangakan 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

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

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* $| x_{ij}-x |^2$ di persamaan 3 dimasukkan faktorbobot tiap fitur untuk meningkatkan akurasi klasifikasi. Pada DGC, pembobotan dilakukan dengan metode TRFS (*TentativeRandom Selection Features*) [5] yang mensimulasikan prosesmutasi algoritma genetic untuk mencari kombinasi bobot terbaik.

C. Metode TRFS Untuk Pembobotan Attribut

Pada tabel 1 ditunjukkan algoritma TRFS (*TentativeRandom Feature Selection*) yang digunakan untuk mencaribobot attribut 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 attribut 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 *crossvalidation* lebih baik dari akurasi bobot sebelumnya, makabobot 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
2	Tb
3	W0, Pi / banyaknya attribut, f0
4	**For** i_0**to** i = i_{max}**or** f < fo
5	Pilih W_x dari W secara acak dengan
6	mempertimbangkan P
7	W' = W + Ɛ
8	Evaluasi w' menggunakan *cross*
9	*validation* pada Ta dan Tb,
10	hasilnya adalah f'
11	**If** f<f'
12	W = W'
13	f = f'
14	$P_x = P_x + \delta$
15	**Else**
16	**If** P_x>
17	$P_x = P_x - \delta$
18	**Else**
19	$P_x = 0$
20	**End if**
	End if
	End for

D. Perhitungan Probabilitas Posterior

Anggaplah bahwa *X* adalah himpunan attribut 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 attribut – attributnya 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 *naïvebayes* 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 pelatihanyang berkelas Ydan $P(X_i|Y)$merupakan proporsi nilai attibutke iobyek Xpada data pelatihan yang memiliki jenis kelas Y.

Untuk menghitung $P(X_i|Y)$pada attribut kontinyu digunakanpersamaan 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. Dengancara sama, μ^2_{ij}dapat diestimasi dari sampel varian (s^2) datapelatihan yang berkelas y_j.

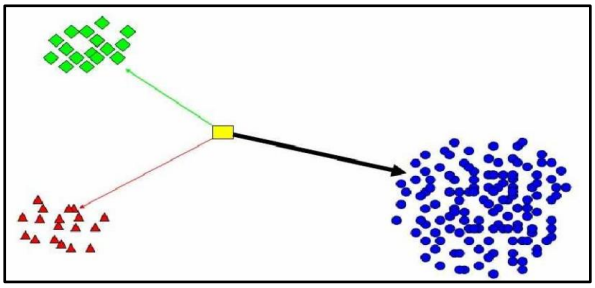
Persamaan 4 mengasumsikan bahwa attribut – attribut datatidak berkorelasi. Secara teoritis, bila metode Naive bayesditerapkan pada data yang attribut – attributnya berkorelasiakan menurunkan akurasi klasifikasi. Meskipun demikian,hasil uji coba secara empiris menunjukkan bahwa metodeNaive bayessecara mengejutkan memberikan hasil baik jikadiuji coba pada data yang attribut – attributnya berkorelasi[7][8][9].Untuk mendapatkan hasil yang lebih baik jika attribut -attribut datanya berkorelasi, dapat digunakan persamaan*multivariate gaussian distribution*

E. *Penelitian Terkait Sebelumnya*

Metode DGC merupakan metode klasifikasi baru yangdiajukan [4][5]. Metode ini dapat dianggap sebagaipengembangan teknik klasifikasi berbasis *distance*. PadaDGC, selain *distance*, ditambahkan konsep massa dangravitasi data. Klasifikasi dilakukan menggunakan gravitasidata.

Metode DGC memberikan hasil klasifikasi yang baik untukdata pelatihan normal. Kelebihan lain dari metode DGCadalah efisien dan prinsip yang mendasari metode tersebutmudah dipahami serta mudah diimplementasikan.

Namun untuk data pelatihan yang tidak imbang, metodeDGC memberikan hasil yang buruk [6]. Pada gambar 1ditunjukkan terdapat sebuah kelas pada data pelatihan yangmassanya atau banyaknya data pelatihan yang menjadianggota kelas tersebut sangat besar (berwarna biru).Akibatnya, gravitasi data kelas tersebut menjadi sangat kuat.Semua obyek pada data uji cenderung diklasifikasi sebagaikelas yang memiliki massa sangat besar tersebut. Hal yangsebaliknya juga berlaku jika terdapat kelas yang massanyasangat 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 datapelatihan yang digunakan tidak imbang, dilakukan beberapamodifikasi 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

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

Dari sudut pandang metode Naive bayes, proporsi sebuahkelas pada data pelatihan dianggap probabilitas prior obyek Xmerupakan anggota kelas tersebut [6]. Sehingga padaperhitungan gravitasi data, massa kelas atau proporsi dapatdiganti dengan probabilitas prior kelas tersebut. Persamaan 6dapat ditulis ulang menjadi

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

Hasil klasifikasi bila gravitasi data dihitug menggunakanpersamaan 7 tidak berubah meskipun nilai gravitasi datanyatidak sama. Hal ini menunjukkan bahwa konsep massa padaDGC dapat diinterpretasikan sebagai probabilitas prior.Interpretasi dan penggantian massa dengan probabilitas priortersebut merupakan hal penting karena membuka peluanguntuk mengekslorasi konsep gravitasi data melalui sudutpandang *bayesian learning*.

2) Mengganti massa (probabilitas prior) dengan probabilitasposterior

Pada metode *bayesian learning*, probabilitas posteriordianggap lebih akurat untuk klasifikasi dibandingkan denganprobabilitas prior. Karena itu, modifikasi kedua yang diajukanuntuk meningkatkan akurasi metode DGC adalah menggantimassa atau probabilitas prior dengan probabilitas posterior.

Karena probabilitas posterior lebih akurat dibandingprobabilitas prior, diharapkan penggunaan probabilitasposterior untuk menggantikan massa (yang ekivalen denganprobabilitas prior) akan meningkatkan kualitas perhitungangravitasi data sekaligus meningkatkan akurasi klasifikasimetode DGC pada kasus data pelatihan yang tidak imbang.

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

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

Persamaan 8 tidak memberikan hasil klasifikasi sama denganpersamaan 6 dan 7

3) Modifikasi persamaan gravitasi data

Penggantian massa pada DGC dengan probabilitas posteriormembutuhkan beberapa perubahan pada persamaan untukmenghitung gravitasi data jika kelas terdiri atas banyakpartikel data. Pada DGC massa sebuah kelas dipartisi menjadipartikel – partikel data. Karena pada modifikasi metode DGCmassa diganti dengan probabilitas posterior, maka probabilitasposterior tersebut juga harus dipartisi menjadi partikelpartikel. Setiap partikel mendapat potongan probabilitasposterior sesuai dengan proporsi massanya pada kelastersebut. Potongan probabilitas posterior tersebutmenggantikan massa partikel pada persamaan untukmenghitung gravitasi data

Misalnya pada data pelatihan terdapat kkelas, yaitu C_1, C_2,... dan C_k. Masing masing kelas memiliki anggota sebanyakL_1, L_2, ... dan L_k. Masing – masing kelas dipartisi menjadi T_1,T_2, ... dan $T_k$partikel data. Sebuah data atau obyek Xyangakan diklasifikasi dapat dianggap sebagai partikel data tunggaldengan nilai pusat massa sama vektor fiturnya. Menggunakanprinsip superposisi pada persamaan 3, gravitasi data kelas $C_i$pada obyek Xdapat dihitung menggunakan persamaan

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

$F_i$adalah superposisi atau total gravitasi data kelas i pada X,P(Ci|X)merupakan probabilitas posterior Xmerupakan 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 Xdengan pusat massa partikel $T_j$padakelas C_i. $P(C_i|X)$dihitung dengan

persamaan 6. Sedangkan d_{ij}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, pimadan WBCD.

Uji coba dilakukan dalam 2 skenario,

1) Skenario 1

Uji coba skenario 1 dilakukan dengan membandingkanakurasi dan waktu klasifikasi metode DGC dan DGCPPmenggunakan metode *Full Train Full Test*dengan parameterradius yang bervariasi antara radius minimum 0 hingga radiusK di mana sebuah kelas pada data pelatihan menjadi anggotadari sebuah partikel data

2) Skenario 2

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

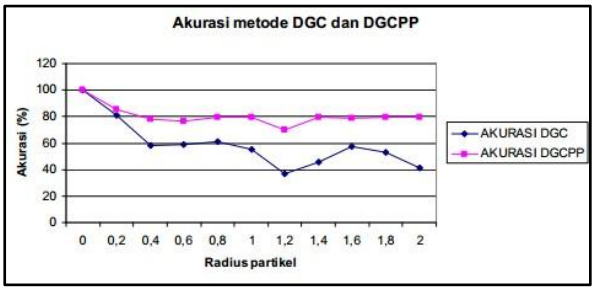
B. Hasil Dan Pembahasan Uji Coba Skenario 1

Tabel 2 menunjukkan akurasi uji coba skenario 1 metodeDGC dan DGCPP untuk *dataset*normal. Dari tabel 2 dapatdiamati bahwa untuk *dataset*normal, metode DGCPP hampirselalu memiliki akurasi lebih baik dari atau sama denganmetode DGC pada hampir semua nilai radius.

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

Gambar 2 menunjukkan grafik akurasi uji coba skenario 1untuk *dataset*vehicle. Gambar 2 mewakili karakteristikhampir semua hasil uji coba skenario 1. Dari gambar 2 dapatdiamati beberapa karakteristik akurasi klasifikasi metode DGCdan DGCPP

1) Nilai akurasi terbaik metode DGC dan DGCPP pada ujicoba skenario 1 yang menggunakan metode *Full Train Full Test*relatif sama. Namun hal ini tidak berarti bahwapada metode uji coba lain memberi hasil yang samaseperti yang ditunjukkan pada uji coba skenario 2.
2) Akurasi metode DGC dan DGCPP semakin rendah bilaradius partikel yang digunakan semakin besar
3) Grafik akurasi metode DGC dan DGCPP memiliki bentukhampir sama, namun grafik akurasi metode DGCPPberada di atas grafik akurasi metode DGC
4) Makin besar ukuran radius partikel data, selisih akurasimetode DGC dan DGCPP semakin besarAkurasi metode DGC dan DGCPP



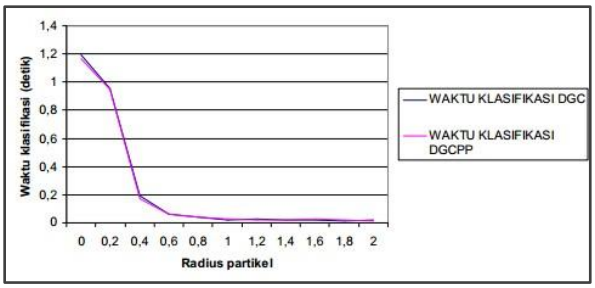
Gambar 2 grafik akurasi metode DGC dan DGCPP untuk *dataset*vehicle pada skenario 1

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

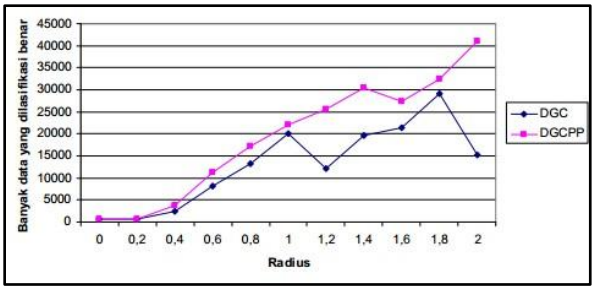
Tabel 5 menunjukkan waktu klasifikasi uji coba skenario 1metode DGC dan DGCPP untuk *dataset*yang mengalami*class imbalance problem*. Dari tabel 5 dapat diamati bahwauntuk *dataset*yang mengalami *class imbalance problem*,kedua metode memiliki waktu klasifikasi hampir sama. Selisihwaktu klasifikasi untuk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

Gambar 3 menunjukkan grafik waktu klasifikasi uji cobaskenario 1 untuk *dataset*vehicle. 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 *dataset*vehicle. 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 *dataset*pima.



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



Gambar 4 Grafik banyaknya data yang diklasifikai 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 secaraumum 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 pelatihanyang 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 mengamatidata – data yang diklasifikasi secara salah oleh metode DGCnamun diklasifikasi benar oleh metode DGCPP. Analisa inibertujuan

untuk membuktikan apakah pada *dataset*yangmengalami *class imbalance problem*, klasifikasi menggunakanmetode DGC menyebabkan data yang sebenarnya berjeniskelas yang massanya kecil diklasifikasikan sebagai kelas yangmassanya besar. Tujuan lain dari analisa migrasi data iniadalah untuk membuktikan apakah pada klasifikasi menggunakan metode DGCPP kesalahan klasifikasi tersebutdapat 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

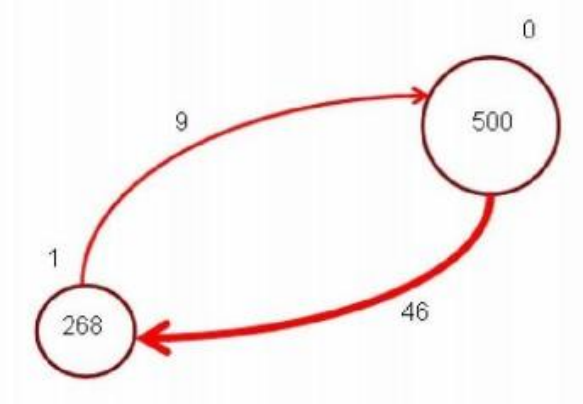
	kelas 0 (500)		kelas 1 (268)	
Metode	DGC	DGCPP	DGC	DGCPP
Recall class	0,83	0,792	0,485	0,652

Tabel 6*recall* klasifikasi metode DGC dan DGCPP *dataset* Wbcd 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 cenderungbernilai rendah pada metode DGC. Hal ini dapat diartikanbahwa pada metode DGC, terdapat banyak data uji yangsebenarnya berjenis kelas yang massanya kecil,diklasifikasikan secara keliru sebagai kelas yang massanyabesar. Kedua, pada metode DGCPP, *recall*kelas yangmassanya kecil naik dibandingkan dengan nilai *recall*padametode DGC. Hal ini dapat ditafsirkan sebagai berikut. Padametode DGCPP, sejumlah data dari kelas yang massanya kecilyang diklasifikasikan secara keliru sebagai kelas yangmassanya besar oleh metode DGC, diklasifikasikan secarabenar sesuai kelasnya oleh metode DGCPP

Banyak data yang awalnya diklasifikasi DGC berkelas yangmassanya besar, diklasifikasi sebagai kelas yang massanyakecil. Pada gambar 5 ditunjukkan terdapat 46 data yangsebenarnya berkelas 1 (massa 268) diklasifikasikan sebagai 0(massa 500) oleh DGC. Oleh DGCPP, data tersebutdiklasifikasi 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 dapatdiambil kesimpulan sebagai berikut

1) Pada parameter radius partikel minimum, metode DGCPPmemberikan hasil klasifikasi yang lebih baik dari DGCpada *dataset*normal maupun *dataset*yang mengalami*class imbalance problem*seperti yang ditunjukkan padauji coba skenario 2.
2) Semakin besar ukuran radius partikel yang digunakan,akurasi metode DGC dan DGCPP makin rendah. Namunpenurunan akurasi metode DGC lebih cepat dibandingkandengan metode DGCPP. Akibatnya, semakin besarukuran radius partikel data yang digunakan, selisihakurasi metode DGCPP semakin melampaui metodeDGC
3) Penambahan waktu klasifikai metode DGCPPdibandingkan 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 misklasifikasiyang dilakukan metode DGC pada *dataset*yangmengalami *class imbalance problem*

Daftar Pustaka

[1] Hao, Pei-Yi, Chiang, Jung-Hsien dan Tu, Yi-Kun, 2007,"*Hierarchically SVM classification based on support vector clusteringmethod and its application to document categorization*", Expert Systemswith 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 withApplications, 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 alertclassification model*", Expert Systems with Applications, 36 (2009),11145–11155

[4] Peng, Lizhi, Yang, Bo dan Chen, Yuehui 2005, "*A NovelClassification Method Based on Data Gravitation*", Proc. OfInternational Conference on Neural Networks and Brain(ICNN&B),667-672, 2005.

[5] Peng, Lizhi, Yang, Bo, Chen, Yuehui dan Abraham, Ajith, 2009, "*DataGravitation Based Classification*", Information Sciences, 179, 809–819

[6] Tan, P.N., Steinbach, M. dan Kumar, V., 2006, "*Introduction to DataMining*", Pearson Education, Inc., Boston.

[7] Li, Yumei dan Anderson-Sprecher, Richard, 2006, "*Faciesidentification from well logs: A comparison of discriminant analysis andnaïve bayesclassifier*", 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 bayesassumptions on software fault data : An empirical study*", Data & Knowledge Engineering, 68 (2009), 278–290