Integrasi Metode *Sample Bootstrapping* dan *Weighted* Principal Component Analysis untuk Meningkatkan Performa k Nearest Neighbor pada Dataset Besar

Tri Agus Setiawan, Romi Satria Wahono dan Abdul Syukur
Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
tri.triagus.setiawan45@gmail.com, romi@romisatriawahono.net, abah.syukur01@gmail.com

Abstract: Algoritma k Nearest Neighbor (kNN) merupakan metode untuk melakukan klasifikasi terhadap objek baru berdasarkan k tetangga terdekatnya. Algoritma kNN memiliki kelebihan karena sederhana, efektif dan telah banyak digunakan pada banyak masalah klasifikasi. Namun algoritma kNN memiliki kelemahan jika digunakan pada dataset yang besar karena membutuhkan waktu komputasi cukup tinggi. Pada penelitian ini integrasi metode *Sample Bootstrapping* dan *Weighted* Principal Component Analysis (PCA) diusulkan untuk meningkatkan akurasi dan waktu komputasi yang optimal pada algoritma kNN. Metode *Sample Bootstrapping* digunakan untuk mengurangi jumlah data training yang akan diproses. Metode *Weighted* PCA digunakan dalam mengurangi atribut. Dalam penelitian ini menggunakan dataset yang memiliki dataset *training* yang besar yaitu Landsat Satellite sebesar 4435 data dan Tyroid sebesar 3772 data. Dari hasil penelitian, integrasi algoritma kNN dengan *Sample Bootstrapping* dan *Weighted* PCA pada dataset Landsat Satellite akurasinya meningkat 0.77% (91.40%-90.63%) dengan selisih waktu 9 (1-10) detik dibandingkan algoritma kNN standar. Integrasi algoritma kNN dengan *Sample Bootstrapping* dan *Weighted* PCA pada dataset Thyroid akurasinya meningkat 3.10% (89.31%-86.21%) dengan selisih waktu 11 (1-12) detik dibandingkan algoritma kNN standar. Dari hasil penelitian yang dilakukan, dapat disimpulkan bahwa integrasi algoritma kNN dengan *Sample Bootstrapping* dan *Weighted* PCA menghasilkan akurasi dan waktu komputasi yang lebih baik daripada algoritma kNN standar.

Keywords: algoritma kNN, *Sample Bootstrapping*, *Weighted* PCA

1 PENDAHULUAN

Data mining merupakan suatu proses untuk mengidentifikasi pola yang memiliki potensi dan berguna untuk mengelola dataset yang besar (Witten, I. H., Frank, E., & Hall, 2011). Dalam data mining ada 10 algoritma teratas yang paling berpengaruh yang dipilih oleh peneliti dalam komunitas data mining, dimana 6 (enam) diantaranya adalah algoritma klasifikasi yaitu C4.5, Support Vector Machines (SVM), AdaBoost, k Nearest Neighbor (kNN), Naïve Bayes dan CART (Fayed & Atiya, 2009).

Salah satu algoritma yang banyak diteliti adalah algoritma klasifikasi kNN (Wan, Lee, Rajkumar, & Isa, 2012). Algortima kNN merupakan sebuah metode untuk melakukan klasifikasi terhadap objek baru berdasarkan (k) tetangga terdekatnya (Witten, I. H., Frank, E., & Hall, 2011)(Amores, 2006)(Morimune & Hoshino, 2008). Tujuan dari algoritma kNN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training sample (Morimune & Hoshino, 2008)(Han, J., & Kamber, 2012), dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada kNN.

Algoritma kNN memiliki kelebihan karena sederhana, efektif dan telah banyak digunakan pada banyak masalah klasifikasi (Wu, Xindong & Kumar, 2009). Namun algoritma kNN memiliki kelemahan jika digunakan pada database yang besar karena membutuhkan waktu komputasi cukup tinggi (Fayed & Atiya, 2009)(Wan et al., 2012)(Neo & Ventura, 2012). Adapun dataset yang besar berupa volume yang banyak, label data yang banyak, kecepatan tinggi, dana / atau aset informasi yang membutuhkan bentuk-bentuk baru dari pengolahan untuk pengambilan keputusan, penemuan wawasan dan optimasi proses (O'Reilly, 2012)(Zikopoulos, Eaton, & DeRoos, 2012).

Beberapa peneliti telah melakukan penelitian tentang pengurangan jumlah data dan waktu komputasi. Penelitian Fayed (Fayed & Atiya, 2009) menggunakan pendekatan *Novel Template Reduction* yang digunakan untuk membuang nilai yang jauh dari batasan *threshold* dan memiliki sedikit pengaruh pada klasifikasi kNN. Penelitian Wan (Wan et al., 2012) menggunakan Support Vector Machines-Nearest Neighbor (SVM-NN) dengan pendekatan klasifikasi *hybrid* dengan tujuan bahwa untuk meminimalkan dampak dari akurasi klasifikasi. Penelitian Koon (Neo & Ventura, 2012) menggunakan algoritma *Direct Boosting* untuk meningkatkan akurasi klasifikasi kNN dengan modifikasi pembobotan jarak terhadap data latih.

Oleh karena itu perlu adanya metode untuk mengurangi jumlah *data training* untuk diproses dan mengurangi atribut sehingga mampu meningkatkan akurasi dan meminimalkan waktu komputasi.

Metode *Sample Bootstrapping* digunakan untuk mengurangi jumlah data *training* yang akan diproses (Dudani, 1976)(Amores, 2006). Untuk dapat mengatasi dataset yang besar maka perlu perlu adanya sampel data (*sampling*) secara acak agar data yang akan diproses menjadi lebih kecil (Liaw, Wu, & Leou, 2010)(Morimune & Hoshino, 2008), sedangkan untuk mengukur jarak tetangga terdekat digunakan *euclidian distance* (Han, J., & Kamber, 2012) dalam proses klasifikasi.

Dalam menentukan waktu komputasi dalam proses klasifikasi kNN yang akan dicari adalah nilai mayoritas sehingga dapat dihitung nilai *query instance*, pada tahapan ini semakin banyak nilai mayoritas data yang tidak dekat dan tidak relevan maka akan mengakibatkan proses klasifikasi kategori *nearest neighbor* semakin lama dan proses komputasi tidak dapat optimal (Larose, 2005). Untuk mengatasi masalah tersebut maka data yang tidak penting ataupun relevan harus dieleminasi sehingga waktu komputasi dan *error* dapat dikurangi (Han, J., & Kamber, 2012). Adapun untuk mengurangi atribut dalam mengolah data yang besar maka dapat menggunakan metode Principal Component Analysis (PCA) (Neo & Ventura, 2012)(Han, J., & Kamber, 2012).

Namun PCA memiliki kekurangan dalam kemampuan memilih fitur yang tidak relevan dari dataset (Kim & Rattakorn, 2011), karena bisa saja fitur yang dibuang ternyata adalah fitur yang berpengaruh. Untuk mengatasi masalah tersebut maka dapat dilakukan seleksi fitur dengan melakukan pembobotan atribut yaitu *Weighted* PCA (Kim & Rattakorn, 2011)(Liu & Wang, 2012) berdasarkan nilai *threshold*, dimana fitur yang nilainya kurang dari batas *treshold* akan dieliminasi. Dengan menggunakan metode Weighted PCA dapat mengurangi waktu komputasi (Kim & Rattakorn, 2011) sehingga efisien untuk menangani dataset yang memiliki dimensi yang tinggi.

Dari penelitian yang sudah dilakukan belum ditemukan model yang menggunakan kombinasi pengurangan jumlah data *training* dan pengurangan atribut dalam proses klasifikasi kategori *nearest neighbor*. Oleh karena itu, akan dilakukan integrasi metode *Sample Bootstrapping* dengan *Weighted* PCA sehingga mampu meningkatkan akurasi dan waktu komputasi yang optimal pada algoritma kNN.

Dalam penulisan ini dibagi menjadi beberapa bagian. Pada bagian 2, menjelaskan tentang peneltian terkait.Pada bagian 3, menjelaskan metode yang diusulkan. Hasil penelitian dan dan pembahasan mengenai komparasi metode yang diusulkan dijelaskan dalam bagian 4. Penutup, pada bagian ini akan menjelaskan tentang kesimpulan dan saran dari penelitian.

2 PENELITIAN TERKAIT

Dalam penelitian yang dilakukan oleh Fayed et al. (Fayed & Atiya, 2009) menggunakan pendekatan *Novel Template Reduction* yang digunakan untuk membuang nilai yang jauh dari batasan *threshold* dan memiliki sedikit pengaruh pada klasifikasi kNN. Adapun untuk pengujian waktu proses klasifikasi menggunakan metode *condensed set* dengan melakukan pengurangan terhadap data yang tidak terpakai sehingga dapat meningkatkan akurasi

Adapun penelitian yang dilakukan oleh Wan et al. menyajikan pendekatan klasifikasi *hybrid* dengan menggabungkan algoritma Support Vector Machine (SVM) dan algoritma kNN pada ketergantungan parameter yang rendah (Wan et al., 2012), untuk mendapatkan akurasi terbaik dengan menggunakan training dataset yang besar. Dalam model *hybrid* SVM-kNN, SVM digunakan untuk mengurangi data training ke Support Vectors (SVs) dari masing-masing kategori, dan algoritma *nearest neighbor*, kemudian digunakan untuk menghitung jarak rata-rata antara pengujian titik data ke set SVs dari kategori yang berbeda. Langkah selanjutnya menentukan kategori data baru yang tidak berlabel berdasarkan jarak rata-rata terpendek antara SVs kategori dan titik data baru, kemudian menghitung jarak rata-rata untuk masing-masing kategori dengan menggunakan rumus *euclidean distance*.

Pada penelitian yang dilakukan Konn at el (Neo & Ventura, 2012) menyajikan pendekatan menggunakan algoritma *Direct Boosting* untuk meningkatkan akurasi klasifikasi kNN dengan modifikasi pembobotan jarak terhadap data latih dengan *local warping of distance matric*. Metode *local warping of the distance matric* digunakan untuk merubah bobot jarak setiap data latih, kemudian memodifikasi klasifikasi kNN dengan memberi bobot jarak 1/d untuk mengklasifikasikan setiap data latih menggunakan sisanya setiap iterasi. Dalam melakukan validasi sehingga menghasilkan akurasi terbaik menggunakan metode 10-*fold cross validation* untuk setiap melakukan iterasi melakukan validasi sehingga menghasilkan akurasi terbaik menggunakan metode 10-*fold cross validation* untuk setiap melakukan iterasi.

Dalam penelitian ini kita akan melakukan perbaikan metode dengan melakukan integrasi metode *Sample Bootstrapping* dan *Weighted* Principal Component Analysis (PCA) diusulkan untuk meningkatkan akurasi dan waktu komputasi yang optimal pada algoritma kNN. Metode *Sample Bootstrapping* digunakan untuk mengurangi jumlah data *training* yang akan diproses. Metode *Weighted* Principal Component Analysis (PCA) digunakan untuk mengurangi atribut. Untuk pengujian akurasi hasil klasikasi dilakukan menggunakan metode *confusion matrix* (Witten, I. H., Frank, E., & Hall, 2011)(Maimon Oded, 2010) dan uji efisiensi (lamanya waktu proses klasikasi) dinyatakan dalam waktu (detik).

3 METODE YANG DIUSULKAN

Untuk melakukan penelitian ini menggunakan spesifikasi komputer Intel Core i5-2557M 1.7GHz, RAM 2 GB, operating system Microsoft Windows 7 Home Premium.Untuk pengembangan sistem menggunakan Rapid Miner 5.3.015.

Proses eksperimen dan pengujian model menggunakan bagian dari dataset yang ada. Data yang digunakan dalam penelitian ini menggunakan dataset Landsat Satellite dan Thyroid, hal ini berdasarkan penelitian-penelitian sebelumnya (Fayed & Atiya, 2009)(Wan et al., 2012)(Neo & Ventura, 2012) tentang kNN menggunakan datataset tersebut seperti pada Tabel.1.

Tabel 1 Dataset yang Digunakan di Penelitian

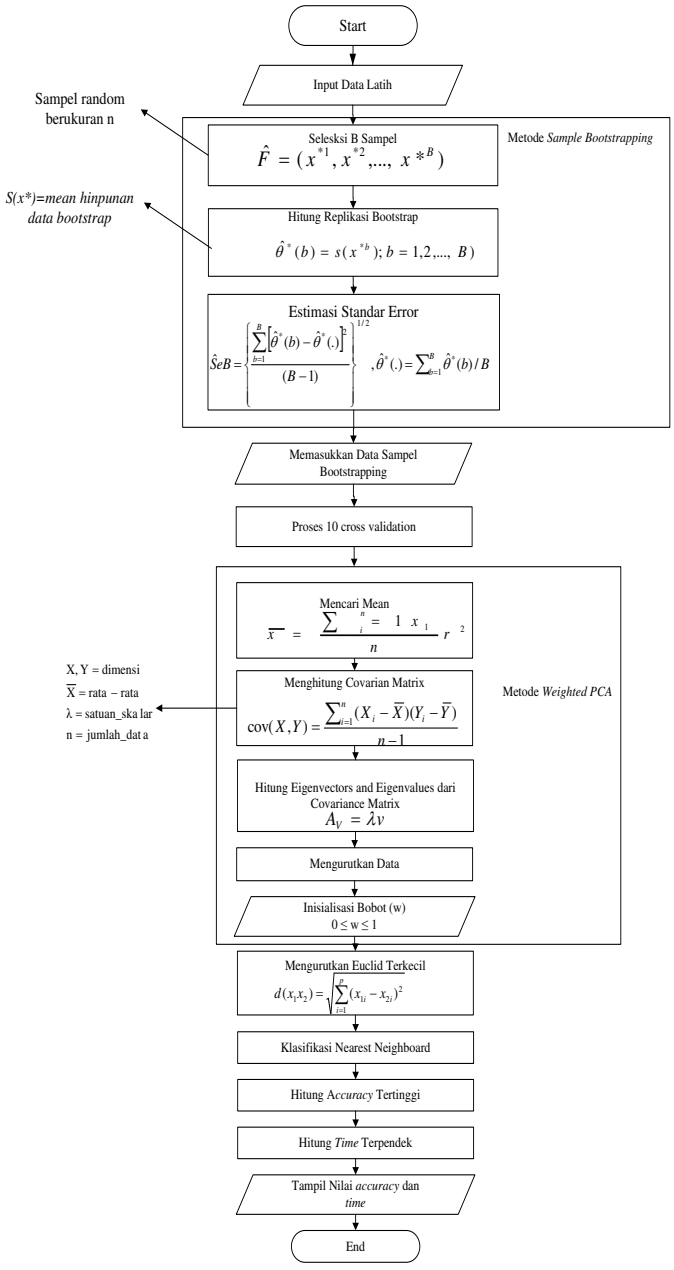
No	Name	Type	Record	Dimension	Class
1	Landsat Satellite	Classification	6435	36	6
2	Thyroid	Classification	7200	21	3

Metode yang diusulkan dalam penelitian ini yaitu dengan melakukan integrasi *Sample Bootstrapping* dan *Weighted* PCA dalam meningkatkan akurasi dan menentukan waktu komputasi pada algoritma kNN. Metode *Sample Bootstrapping* digunakan untuk mengurangi jumlah data *training* yang akan diproses (Dudani, 1976)(Amores, 2006), sedangkan untuk mengurangi jumlah atribut dalam mengolah data yang besar maka dapat menggunakan metode *Weighted* Principal Component Analysis (PCA) (Neo & Ventura, 2012)(Han, J., & Kamber, 2012) karena mampu mereduksi untuk data yang memiliki dimensi tinggi. Adapun tahapan eksperimen pada penelitian ini adalah:

1. Menyiapkan dua dataset untuk eksperimen yang diambil dari University of California, Irvine (UCI)
2. Melakukan pengujian menggunakan algoritma kNN menggunakan dataset Landset Satellite dan Thyroid kemudian dari hasil pengujian tersebut dicatat hasil yang diperoleh
3. Melakukan pengujian menggunakan algoritma kNN dengan *Sample Bootstrapping* menggunakan dataset Landsat Satellite dan Thyroid kemudian dari hasil pengujian tersebut dicatat hasil yang diperoleh
4. Melakukan pengujian menggunakan algoritma kNN dengan Weighted PCA menggunakan dataset Landsat Satellite dan Thyroid kemudian dari hasil pengujian tersebut dicatat hasil yang diperoleh
5. Melakukan pengujian menggunakan algoritma kNN dengan *Sample Bootstrapping* dan *Weighted* PCA menggunakan dataset Landsat Satellite dan Thyroid

kemudian dari hasil pengujian tersebut dicatat hasil yang diperoleh
6. Membandingkan hasil akurasi terbaik dan waktu komputasi minimal dan mengambil hasil terbaik
7. Mengintegrasikan hasil algoritma klasifikasi terbaik.

Adapun algoritma yang diusulkan dalam penelitian ini seperti pada Gambar 1, diawali dengan memasukkan dataset baik *data training* maupun *data testing*, kemudian melakukan transformasi dimana metode *Sample Bootstrapping* digunakan untuk mengurangi jumlah data *training* yang akan diproses kemudian menghitung validitas data *training*, setelah itu menghitung kuadrat jarak *euclidian* (*euclidean distance*) masing-masing objek terhadap data sampel yang diberikan. Kemudian menghitung nilai *distance weighted* yang didapat dari memasukkan nilai validitas dan nilai *eucledian*, setelah melakukan pembobotan atribut dan diperoleh klasifkasi *nearest neighbor*. Metode *Weighted* Principal Component Analysis (PCA) digunakan untuk mengurangi atribut.



Gambar 1 Algoritma *Sample Bootstrapping* Weighted PCA

4 HASIL PENELITIAN

Dalam penelitian ini akan dilakukan komparasi antara algoritma kNN dengan algoritma kNN dan *Sample Bootstrapping* dan *Weighted* PCA. Metode *Sample Bootstrapping* digunakan untuk mengurangi jumlah data *training* yang akan diproses dan metode *Weighted* Principal Component Analysis (PCA) digunakan untuk mengurangi atribut serta mengitegrasikan metode *Sample Bootstrapping* dan *Weighted* PCA diusulkan untuk meningkatkan akurasi dan waktu komputasi yang optimal pada algoritma *kNN* pada dataset Thyroid dan Landsat Satellite.

Pada eksperimen pertama akan melakukan perhitungan menggunakan algoritma kNN dengan dataset Thyroid dan Thyroid.Adapun proses perhitungan kNN sebagai berikut:
1. Menyiapkan dataset Thyroid dan Landsat Satellite, kita lakukan validasi dengan *cross validation* dimana dataset kita bagi menjadi data *training* dan data *testing*
2. Menentukan nilai*k*, pada penentuan *k* dilakukan input antara 1…7200 (dataset Thyroid) dan 1…6435 (dataset Lansat Satellite)
3. Menghitung kuadrat jarak euclid (*query instance*) masing-masing objek terhadap sampel data yang diberikan dengan menggunakan *euclidian distance* denga parameter *numeric* dengan rumus:

$$d_i = \sqrt{\sum_{1=i}^{p}(x_{1i} - x_{2i})^2}$$

4. Mengurutkan objek-objek termasuk ke dalam kelompok yang mempunyai jarak euclid terkecil
5. Menghitung Akurasi
   Untuk menghitung nilai akurasi digunakan *confusion matrix* dengan rumus:

$$\text{akurasi} = \frac{Jumlah\ Data\ Benar}{Jumlah\ Data} \text{X } 100\%$$

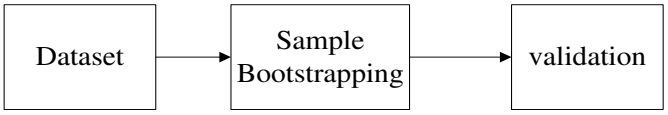
Hasil perhitungan yang dilakukan untuk algoritma kNN mendapatkan nilai akurasi terbaik pada k=1 dengan waktu komputasi minimal adalah 1 detik baik untuk dataset Thyroid maupun Landsat Satellite seperti pada Tabel 2.

Tabel 2 Akurasi Algoritma kNN Dengan Dataset Thyroid dan Landsat Satellite

Dataset	Nilai Akurasi (%)	Waktu (detik)
Thyroid	86.21	10
Landsat Satellite	90.63	12

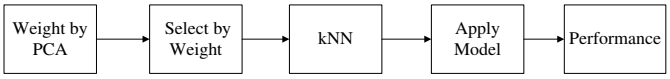
Pada eksperimen kedua akan melakukan perhitungan menggunakan algoritma kNN dengan *Sample Bootstrapping* dan *Weighted* PCA pada dataset Thyroid dan Landsat Satellite. Adapun proses yang dilakukan adalah:

1. Melakukan preprocessing menggunakan metode *sampling*. Algoritma yang dipakai yaitu *Sample Bootstrapping*, kemudian memilih parameter *sample* yaitu *relative* (sampel dibuat sebagai sebagian kecil dari jumlah total contoh dalam sampel data) dan nilai sampel rasio yang diinput antara 0-1. Setelah dilakukan *sampling* maka data *bootstrap* tersebut divalidasi dengan *cross validation* sebagaimana ditunjukkan dalam Gambar 2.



Gambar 2 Pengujian Performa Algoritma kNN dengan *Sample Bootstrapping* untuk Dataset Thyroid

2. Langkah berikutnya yaitu melakukan normalisasi terhadap *attribute class* pada dataset dan melakukan pembobotan terhadap *attribute class* dengan *weighted relation*. *Weighted relation* mencerminkan relevansi bobot atribut dengan nilai attribut class 0 sampai 1.0, pada penelitian ini bobot atribut diisi, kemudian menentukan k. Adapun proses Weighted PCA kNN seperti pada Gambar 3.



Gambar 3 Pengujian Performa Algoritma kNN dengan Weighted PCA untuk dataset Thyroid

Dalam hal ini bobot atribut dan nilai k sangat berperan dalam memdapatkan akurasi dan waktu yang baik. Nilai akurasi dan waktu yang optimal dari *confusion matrix* tersebut seperti dalam Tabel 3 dengan rumus:

$$\text{akurasi} = \frac{Jumlah\ Data\ Benar}{Jumlah\ Data} \text{X } 100\%$$

Tabel 3 Hasil Akurasi dan Waktu KomputasiAlgoritma kNN dengan *Sample Bootstrapping* dan *Weighted* PCA Pada Dataset Thyroid dan Landsat Satellite

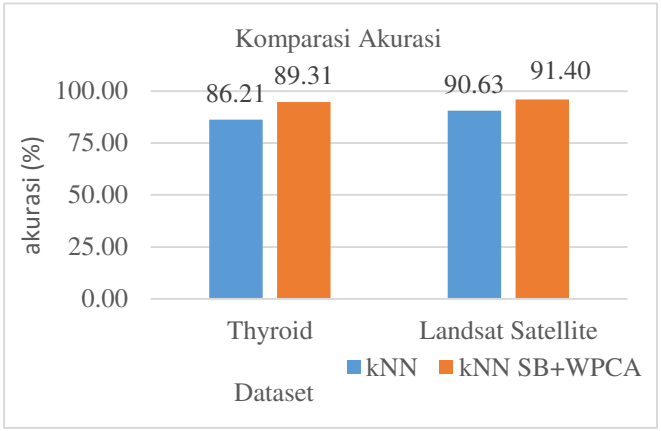
Dataset	Nilai Akurasi (%)	Waktu (detik)
Thyroid	89.31	1
Landsat Satellite	91.40	1

Dari hasil eksperimen tentang nilai akurasi yang dilakukan antara algoritma kNN dengan kNN dengan *Sample Bootstrapping* dan *Weighting* PCA dapat meiningkatkan akurasi (Champagne, Mcnairn, Daneshfar, & Shang, 2014)(Ghaderyan, Abbasi, & Hossein, 2014)(Polat & Kara, 2008) untuk dataset Thyroid dan Lansat Satellite. Pada algoritma kNN data yang digunakan sejumlah data secara keseluruhan tidak ada proses *filtering* maupun sampel data yang digunakan sehingga membutuhkan waktu yang lama sehingga tingkat akurasi menjadi rendah, sedangkan pada algoritma kNN dan *Sample Boostrapping* dan *Weighted* PCA data yang digunakan tidak keseluruhan tetapi dilakukan menggunakan data *sampling* (Witten, I. H., Frank, E., & Hall, 2011) untuk melakukan *filtering* agar mengurangi jumlah data sampel (Champagne et al., 2014)(McRoberts, Magnussen, Tomppo, & Chirici, 2011)(Chen & Samson, 2015).

Dalam metode *Sample Bootstrapping* terdapat rasio parameter *sample* yang berfungsi memberikan nilai jumlah data sample yang digunakan dari seluruh data yang ada dengan nilai 0-1. Dengan metode ini jumlah data yang diproses tidak secara keseluruhan melainkan beberapa data tetapi tidak mengurangi jumlah data yang ada karena setelah data tersebut digunakan maka akan dikembalikan lagi (Tian, Song, Li, & Wilde, 2014). Dari hasil perhitungan dapat dilihat perbandingan berdasarkan akurasi pada Tabel 4 dan Gambar 4. Pada tabel dan gambar ini didapat hasil dimana integrasi algoritma kNN dengan *Sample Bootstrapping* dan *Weighting* PCA mempunyai nilai akurasi yang lebih baik yaitu 3.10% untuk dataset Thyroid dan 0.77% untuk dataset Landsat Satellite

Tabel 4 Komparasi Akurasi Algoritma kNN Dengan *Sample Bootstrapping* dan Weighted PCA (kNN SB+WPCA) Pada Dataset Thyroid dan Landsat Satellite

Dataset	Nilai Akurasi (dalam %)		Kenaikan Akurasi (dalam %)
	kNN	*(kNN SB+WPCA)*	
Thyroid	86.21	89.31	3.10
Landsat Satellite	90.63	91.40	0.77



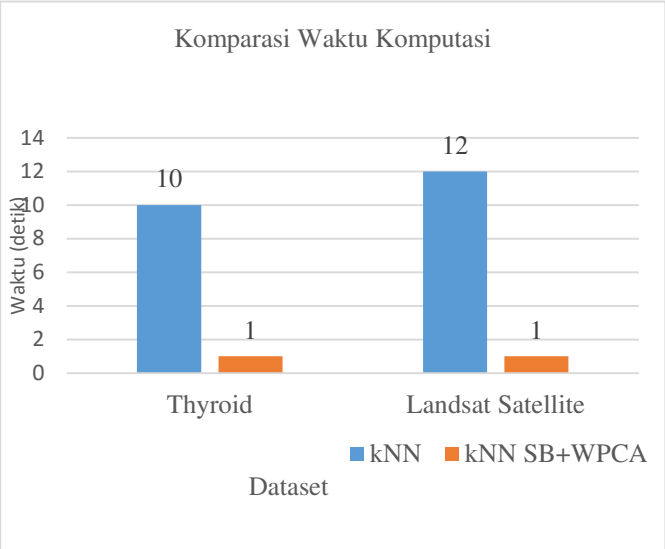
Gambar 4 Komparasi Akurasi Algoritma kNN Dengan *Sample Bootstrapping* dan Weighted PCA Pada Dataset Thyroid dan Landsat Satellite

Dari hasil eksperimen tentang waktu komputasi yang dilakukan antara algoritma kNN dengan kNN dan *Sample Bootstrapping* dan *Weighting* PCA untuk dataset Thyroid keduanya sama-sama menggunakan metode *confusion matrix* untuk pengujian akurasi hasil klasikasi (Witten, I. H., Frank, E., & Hall, 2011)(Maimon Oded, 2010) dan uji efisiensi (lamanya waktu proses klasikasi) dinyatakan dalam *waktu* (detik).

Adapun untuk mengurangi jumlah waktu dan memori yang dibutuhkan maka digunakan metode Principal Componen Analysis (PCA) (Amores, 2006)(Morimune & Hoshino, 2008)(Ghaderyan et al., 2014) karena mampu mengurangi atribut pada data yang besar (Han, J., & Kamber, 2012)(Polat & Kara, 2008). Pada algoritma *Sample Bootstrapping* dan *Weighted* PCA diberikan bobot terhadap atribut *class* dengan nilai 0-1, ketika bobot atribut kurang dari nilai *threshold* maka akan dibuang sehingga dapat meningkatkan waktu komputasi. Dalam penelitian ini bobot diinputkan dengan nilai 0-1 dan nilai k. Pada Tabel 5 dan Gambar 5 didapat hasil komparasi antara algoritma kNN dengan algoritma kNN dan *Sample Bootstrapping* dan *Weighted* PCA.

Tabel 5 Komparasi Waktu Komputasi Algoritma kNN Dengan *Sample Bootstrapping* dan Weighted PCA (kNN SB+WPCA) Pada Dataset Thyroid dan Landsat Satellite

Dataset	Waktu Komputasi (detik)		Selisih Waktu Komputasi (detik)
	kNN	*(kNN SB+WPCA)*	
Thyroid	10	1	9
Landsat Satellite	12	1	11



Gambar 5 Komparasi Waktu Komputasi Algoritma kNN Dengan *Sample Bootstrapping* dan *Weighted* PCA Pada Dataset Thyroid dan Landsat Satellite

5 KESIMPULAN

Integrasi algoritma kNN dengan *Sample Bootstrapping* dan *Weighted* PCA untuk dataset Thyroid ada kenaikan akurasi sebesar 3.10% (89.31-86.21%) dengan menggunakan sampel *data training* sebesar 2160 dan selisih waktu komputasi 9 (1-10) detik dengan pengurangan atribut sebanyak 1 atribut, sedangkan untuk dataset Landsat Satellite ada kenaikan akurasi sebesar 0.77% (91.40-90.63%) dengan menggunakan sampel *data training* sebesar 1931 dan selisih waktu komputasi 11 (1-12) detik dengan pengurangan atribut sebanyak 27 atribut. Dari hasil penelitian tersebut dapat disimpulkan bahwa integrasi algoritma kNN dengan *Sample Bootstrapping* dan *Weighted* PCA dapat meningkatkan akurasi dan mengurangi waktu komputasi dibandingkan dengan algoritma kNN standar.

REFERENCES

Amores, J. (2006). Boosting the distance estimation Application to the K -Nearest Neighbor Classifier. *Pattern Recognition Letters*, *27*(d), 201–209. doi:10.1016/j.patrec.2005.08.019

Champagne, C., Mcnairn, H., Daneshfar, B., & Shang, J. (2014). A bootstrap method for assessing classification accuracy and confidence for agricultural land use mapping in Canada. *International Journal of Applied Earth Observations and Geoinformation*, *29*, 44–52. doi:10.1016/j.jag.2013.12.016

Chen, X., & Samson, E. (2015). Environmental assessment of trout farming in France by life cycle assessment : using bootstrapped principal component analysis to better de fi ne system classification. *Journal of Cleaner Production*, *87*, 87–95. doi:10.1016/j.jclepro.2014.09.021

Dudani, S. a. (1976). The Distance-Weighted k-Nearest-Neighbor Rule. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(4), 325–327. doi:10.1109/TSMC.1976.5408784

Fayed, H. A., & Atiya, A. F. (2009). A Novel Template Reduction Approach for the -Nearest Neighbor Method. *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council*, *20*(5), 890–896.

Ghaderyan, P., Abbasi, A., & Hossein, M. (2014). An efficient seizure prediction method using KNN-based undersampling and linear frequency measures. *Journal of Neuroscience Methods*, *232*, 134–142. doi:10.1016/j.jneumeth.2014.05.019

Han, J., & Kamber, M. (2012). *Data Mining Concepts and Techniques*. (M. Han, J., & Kamber, Ed.) (Third Edit.). USA: Morgan Kaufmann Publishers.

Kim, S. B., & Rattakorn, P. (2011). Unsupervised feature selection using weighted principal components. *Expert Systems with Applications*, *38*(5), 5704–5710. doi:10.1016/j.eswa.2010.10.063

Larose, D. T. (2005). *Discovering Knowledge In Data*. USA: John Wiley & Sons, Inc. New York, NY, USA.

Liaw, Y.-C., Wu, C.-M., & Leou, M.-L. (2010). Fast k-nearest neighbors search using modified principal axis search tree. *Digital Signal Processing*, *20*(5), 1494–1501. doi:10.1016/j.dsp.2010.01.009

Liu, N., & Wang, H. (2012). Weighted principal component extraction with genetic algorithms. *Applied Soft Computing Journal*, *12*(2), 961–974. doi:10.1016/j.asoc.2011.08.030

Maimon Oded, R. L. (2010). *Data Mining And Knowledge Discovery Handbook*. (R. L. Maimon Oded, Ed.) (Second Edi.). Israel: Springer.

McRoberts, R. E., Magnussen, S., Tomppo, E. O., & Chirici, G. (2011). Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. *Remote Sensing of Environment*, *115*(12), 3165–3174. doi:10.1016/j.rse.2011.07.002

Morimune, K., & Hoshino, Y. (2008). Testing homogeneity of a large data set by bootstrapping. *Mathematics And Computers In Simulation*, *78*, 292–302. doi:10.1016/j.matcom.2008.01.021

Neo, T. K. C., & Ventura, D. (2012). A direct boosting algorithm for the k-nearest neighbor classifier via local warping of the distance metric. *Pattern Recognition Letters*, *33*(1), 92–102. doi:10.1016/j.patrec.2011.09.028

O'Reilly. (2012). *Big Data Now: 2012 Edition* (First Edit.). O'Reilly Media, Inc.

Polat, K., & Kara, S. (2008). Medical diagnosis of atherosclerosis from Carotid Artery Doppler Signals using principal component analysis (PCA), k -NN based weighting pre-processing and Artificial Immune Recognition System (AIRS). *Elsevier Inc.*, *41*, 15–23. doi:10.1016/j.jbi.2007.04.001

Tian, W., Song, J., Li, Z., & Wilde, P. De. (2014). Bootstrap techniques for sensitivity analysis and model selection in building thermal performance analysis. *Applied Energy*, *135*, 320–328. doi:10.1016/j.apenergy.2014.08.110

Wan, C. H., Lee, L. H., Rajkumar, R., & Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. *Expert Systems with Applications*, *39*(15), 11880–11888. doi:10.1016/j.eswa.2012.02.068

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining*. (M. A. Witten, I. H., Frank, E., & Hall, Ed.) (Third Edit.). USA: Morgan Kaufmann Publishers.

Wu, Xindong & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. (V. Wu, Xindong & Kumar, Ed.). USA: Taylor & Francis Group.

Zikopoulos, P., Eaton, C., & DeRoos, D. (2012). *Understanding big data*. *New York et al: McGraw …*. Mc Graw Hill. doi:1 0 9 8 7 6 5 4 3 2 1

BIOGRAFI PENULIS

Tri Agus Setiawan. Menyelesaikan pendidikan S1 Sistem Informasi di Universitas Dian Nuswantoro Semarang, S2 Magister Teknik Informatika di Universitas Dian Nuswantoro Semarang. Saat ini menjadi dosen Politeknik Pusmanu Pekalongan. Minat penelitian saat ini adalah softcomputing.

Romi Satria Wahono. Memperoleh gelar B.Eng dan M.Eng pada bidang ilmu komputer di Saitama University, Japan, dan Ph.D pada bidang software engineering di Universiti Teknikal Malaysia Melaka. Menjadi pengajar dan peneliti di Fakultas Ilmu Komputer, Universitas Dian Nuswantoro. Merupakan pendiri dan CEO PT Brainmatics, sebuah perusahaan yang bergerak di bidang pengembangan software. Minat penelitian pada bidang software engineering dan machine learning. Profesional member dari asosiai ilmiah ACM, PMI dan IEEE Computer Society.

Abdul Syukur. Menerima gelar sarjana di bidang Matematika dari Universitas Diponegoro Semarang, gelar master di bidang manajemen dari Universitas Atma Jaya Yogyakarta, dan gelar doktor di bidang ekonomi dari Universitas Merdeka Malang. Dia adalah dosen dan dekan di Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, Indonesia. Minat penelitiannya saat ini meliputi decision support systems dan information management systems.