PEMBELAJARAN ALGORITMA K-NN UNTUK BIG DATASET MENGGUNAKAN METODE SAMPLE BOOTSTRAP DAN WEIGHTED GINI INDEX

Bijanto[1], Zainal Abidin[2], Teguh Tamrin[3]
[1, 2] Sekolah Tinggi Teknik Pati
[3] Universitas Islam Nahdtatul Ulama Jepara
biyantokakoi@gmail.com, zainal.frsd@yahoo.com, teguh@unisnu.ac.id

Abstract

A dataset that has a large number of records or attributes can also be called a big dataset. Large dataset sizes have quantities ranging from terabytes to petabytes. Processing large datasets requires computers that have high specifications. To classify new objects based on the training sample data attributes, you can use the k-NN algorithm. One of the advantages of the kNN algorithm is that it is effective and is often used to manage problems regarding classification. The long computation time is one of the weaknesses of the KNN algorithm. This is due to the calculation of the kNN algorithm for a large dataset. From the problems that arise, the researchers propose the KNN learning system using boostraping and the Weighted Gini Index as a solution for handling large dataset processing problems. KNN learning using the Bootstrap-Weighted Gini Index is used to trim attributes and records based on the results of filtering attributes and records that have a little error quality. This study proves that, the results of the additional accuracy obtained from processing on the Landsat dataset (original accuracy of 91.40% to 94.95%), Thyroid (original accuracy 89.31% to 96.61%), HTRU (original accuracy 96 , 01% to 98.18%) and EEG Eye (original accuracy 97.40% to 97.80%).

Keywords: *Big dataset, Bootstrap, Weighted Gini Index, k-NN*

Abstrak

Dataset yang mempunyai jumlah record atau atribut dalam jumlah besar bisa disebut juga dengan dataset besar. Ukuran dataset besar memiliki jumlah dalam besaran dari terabyte sampai petabyte. Pengolahan dataset besar tersebut membutuhkan komputer yang memiliki spesifikasi tinggi. Untuk mengklasifikasikan objek baru berdasarkan atribut data training sample tersebut bisa menggunakan algoritma k-NN. Salah satu kelebihan algorotma kNN adalah efektif dan sering digunakan untuk mengatur permasalahan mengenai klasifikasi. Cukup lamanya waktu komputasi menjadi salah satu kelemahan algoritma kNN. Hal ini diakibatkan oleh proses kalkulasi algoritma kNN terhadap dataset yang besar. Dari masalah-masalah yang muncul tersebut, maka peneliti mengusulkan sistem pembelajaran kNN menggunakan boostraping dan Weighted Gini Index sebagai solusi untuk penanganan masalah pengolahan dataset besar. Pembelajaran kNN menggunakan Bootstrap-Weighted Gini Index dipakai untuk memangkas atribut maupun record berlandaskan hasil penyaringan atribut dan record yang mempunyai kuwalitas error sedikit. Penelitian ini membuktikan bahwa, hasil penambahan akurasi yang didapat dari pengolahan pada dataset Landsat (akurasi semula sebesar 91,40% menjadi 94,95%), Thyroid (akurasi semula 89,31% menjadi 96,61%), HTRU (akurasi semula 96,01% menjadi 98,18%) dan EEG Eye (akurasi semula 97,40% menjadi 97,80%).

Kata kunci: Dataset Besar, Bootstrap, Weighted Gini Index, k-NN

PENDAHULUAN

Algoritma klasifikasi yang sering dipakai oleh sebgaian besar peneliti dalam menjalankan penelitian adalah algorotma kNN (Fayed & Atiya, 2009). Suatu metode klasifikasi yang dipakai untuk mengelompokkan obyek tertentu bedasarkan nilai tetangga terdekat (k) yaitu algoritma kNN. (Heriyanto & Wisnu, 2008; Wan, *et al.,* 2012; Amores, 2006). Salah satu kelebihan algoritma kNN yaitu lumayan mudah, efektif, kemudian paling sering dipakai untuk melakukan kasus-kasus yang berhubungan dengan pengelompokan atau klasifikasi. (Han & Kamber, 2012). Algoritma kNN memang mempunyai kelebihan, tetapi disisi lain juga mempunyai kekurangan yaitu dalam pemprosesan database yang jumlahnya besar, hal tersebut diakibatkan tempo komputasi kNN lumayan tinggi. (Witten *et al*., 2011; Fayed & Atiya, 2009; Wu *et al*., 2009). Sebuah dataset yang mempunyai ukuran besar, mempunyai jumlah label data yang besar, membutuhkan komputasi berkecepatan tinggi, biaya atau aset informasi yang memerlukan sesuatu yang baru pada waktu pemrosesan dengan maksud untuk mengambil keputusan, penemuan ilmu pengetahuan yang baru, dan optimasi proses disebut dataset besar. (Neo *et al*., 2012; O'Reilly, 2012). Seperti penelitian yang dilakukan oleh Dedic, N dan Stanier. C menyangkut "parameter" mempunyai target yang berkesinambungan berjalan (dinamis), yang mulai pada tahun 2012 dari ukuran beberapa lusin terabyte sampai ukuran banyak petabyte data. (Han *et al*., 2012). Saat bertambahnya jumlah nilai yang paling banyak mendominasi, letaknya berjauhan dan tidak relevan akan mengakibatkan waktu proses klasifikasi *nearest neighbor* semakin lama dan proses komputasi tidak bisa maksimal. (Han *et al*., 2012).

Oleh sebab itu, untuk meminimalisir jumlah data training bisa menggunakan metode Sample Bootstrapping. (Zikopoulos *et al*., 2012; Wan *et al*., 2012; Weidong *et al*., 2014). Metode boostrap bisa untuk mengurangi data maupun atribut yang tidak sesuai dan tidak berpengaruh, sehingga waktu lamanya proses komputasi dan tingakat error bisa diminimalisir (Morimune *et al*., 2008). Shankar dan G. Karpys menerapkan Gini Index untuk pembobotan pengkategorian fitur. (Larose, 2005). Untuk teknik pembobotan yang efektif dan memberikan bobot yang berbeda,Gini Index diintergrasikan dengan beberapa algoritma untuk klasifikasi. (Tang *et al*., 2005). Penelitian ini bertujuan untuk menganalisa peningkatan akurasi dan semakin cepatnya waktu komputasi pada algoritma k-NN pada saat preprosessing data dengan dilakukan penerapan metode metode Sample Bootstrapping yang digunakan sebagai pengurang jumlah data training yang akan diproses dan kemudian untuk memilih jumlah attribut yang memiliki bobot terbaik, menggunakan metode Weighted Gini Index.

TINJAUAN PUSTAKA

Sesuai dengan keempat penelitian yang terkait, dalam pengolahan jumlah data yang besar pada algoritma k-NN, penggunaaan metode untuk meningkatkan akurasi algoritma k-NN untuk menentukan hasil atau kesimpulan, dilakukan dengan berbagai macam metode. Pendekatan Template Reduction yang digunakan untuk menghilangkan nilai yang jaraknya jauh dari batasan threshold kurang signifikan pengaruhnya terhadap klasifikasi k-NN. (Witten *et al*., 2011). Penerapan algoritma Direct Boosting dalam penelitian yang dilakukan oleh Wu *et al*. (2009) dengan memodifikasi pembobotan jarak pada data latih dengan local warping of distance matric yang berguna untuk meningkatkan akurasi k-NN. Klasifikasi hybrid dengan menggabungkan antara algoritma SVM dan k-NN (SVM-NN) dalam mengatasi ketergantungan parameter yang rendah untuk menghasilkan akurasi yang terbaik pada pemprosesan dataset besar. (Fayed & Atiya, 2009). Penelitian yang menerapkan metode bootstrapping dan Weighted Principal

Component Analysis (WPCA) pada algoritma k-NN digunakan untuk mengurangi jumlah data.

METODE PENELITIAN

Data yang dipersiapkan untuk eksperimen menggunakan data yang sudah ada yaitu dari UCI antara lain Landsat, Thyroid, HTRU dan EEG Eye. Hal tersebut seperti yang dilakukan penlitian sebelumnya yaitu : (Witten *et al*., 2011; Fayed & Atiya, 2009; Wu *et al*., (2009) tentang kNN dan Weidong *et al*., (2014) tentang bootstrapping yang menggunakan dataset pada Tabel 1.

Tabel 1 Dataset yang digunakan untuk eksperimen

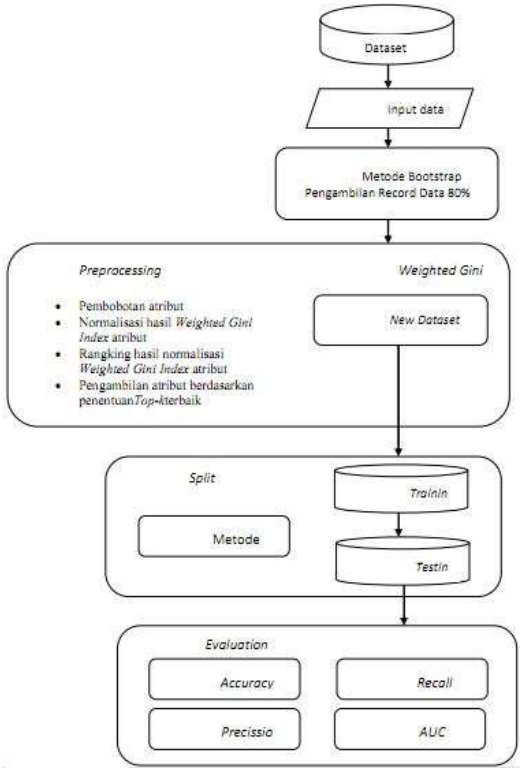
Dataset	Jumlah Record	Jumlah Atribut	Jumlah Atribut Nominal	Jimlah Atribut Numerik	Missing Value	Jumlah Class
Landsat Satellite	6435	36	1	35	0	7
Theroid	7200	21	16	6	0	3
HTRU	17898	8	1	8	0	2
EEG Eye	14980	14	1	14	0	2

Untuk peningkatan akurasi dan semakin cepatnya waktu komputasi pada algoritma k-NN pada saat preprosessing data diusulkan menggunakan penerapan metode metode Sample Bootstrapping yang digunakan sebagai pengurang jumlah data training yang akan diproses dan kemudian untuk memilih jumlah atribut yang memiliki bobot terbaik, menggunakan metode Weighted Gini Index. Metode Sample Bootsrapping digunakan untuk mengurangi jumlah data training yang akan diproses. Dengan berkurangnya jumlah data yang diproses, maka akan semakin singkat pula waktu yang diperlukan. Disamping itu, spesifikasi komputer yang diperlukan juga tidak setinggi pada saat data masih utuh atau belum dikurangi.Untuk mengurangi jumlah data training yang akan diproses, Metode Bootstrapping bisa digunakan. (Zikopoulos *et al*., 2012; Wan *et al*., 2012; Weidong *et al*., 2014).

Penelitian yang telah dilakukan oleh Shankar telah membahas mengenai penerapan prinsip Gini Indeks dalam pilihan teks dan masalah penyesuaian bobot, namun cakupannya terbatas pada klasifikasi berbasis centroid (T. Pang-Ning *et al*., 2006). Kemudian Analisis prinsip dan fitur teks Gini-index, akan membangun fungsi evaluasi secara langsung pada fitur asli untuk pemilihan fitur, lalu memilih subset fitur yang paling signifikan berpengaruh. (Breiman *et al*., 1984). Pendekatan tersebut juga baik untuk pengklasifikasi teks lain yang ada, seperti kNN, SVM, LLSF, Bayes dan sebagainya. (Breiman *et al*., 1984).

Berikut ini adalah tahapan mengenai ekperimen penelitian :

1. Dataset yang akan digunakan untuk penelitian disiapkan kemudian melakukan pemilihan data mengunakan sampel bootstrap yang berfungsi untuk mendapatkan bootstrap data.
2. Seteleh selesai pada tahap bootstrapping, dilakukan filtering dengan cara pendekatan pembobotan atribut dengan metode Weighted Gini Index.
3. Data-data besar yang sudah terfilter dengan metode Weighted Gini Index sesuai atribut masing-masing, akan dilajutkan ke training dan testing terhadap algoritma k-NN dan kemudian melakukan pencatatan hasil kinerja pengklasifikasian yang diantaranya akurasi, presisi, recall dan AUC.
4. Membandingkan hasil akurasi dengan metode penelitian sebelumnya.
5. Mengamati dan menganalisa metode yang lain untuk melakukan pendeteksian *big data* yang mengandung outlier pada tahapan preprocessing yang terbaik untuk k-NN.



Gambar 1. Kerangka kerja metode yang diusulkan

HASIL DAN PEMBAHASAN

Pada penelitian ini melakukan komparasi antara algoritma k-NN dengan pendekatan sampel bootstrapping dan Weighted Gini Index. Sampel bootstrapping digunakan untuk mengurangi jumlah data berdasarkan nilai error yg terkecil. Sedangkan pendekatan Weighted Gini Index digunakan untuk memilih jumlah atribut yang memmiliki tingkat error terkecil pula.

Pada eksperimen pertama melakukan penghitungan mengunakan algoroitma k-NN dengan dataset Landsat. Untuk perhitungan k-NN sebagai berikut:

1. Dataset yang akan digunakan untuk penelitian disiapkan antara lain Landsat, Thyroid, HTRU dan EEG Eye yang kita validasi menggunakan spit validation dimana data tersebut terdiri dari data training dan data testing.
2. Menentukan nilai k, untuk penentuan k kemudian melakukan input antara 1…6435 (dataset Lansat Satellite), 1…7200 (dataset Thyroid), 1...17898 (dataset HTRU), 1...14980 (dataset EEG Eye).
3. Melakukan penghitungan kuadrat jarak euclid (query instance) pada masing-masing objek terhadap sampel data yang diberikan dengan menggunakan euclidian distance menggunakan parameter numeric dengan formula:

$$D(x,y) = \sqrt{\sum_{k-1}^{n}(x_k - y_k)^2}$$

4. Mengurutkan beberapa objek ke dalam kelompok yang memiliki jarak euclidian terkecil.
5. Melakukan penghitungan akurasi dengan confusion matriks mekai rumus:

$$Akurasi = \frac{\text{jumlah data yang diprediksi dengan benar}}{Jumlah\ prediksi\ yang\ dilakukan} \text{ x } 100\%$$

Hasil penghitungan akurasi dengan nilai k=1 dengan menggunakan dataset Landsat, Thyroid, HTRU dan EEG Eye.

Tabel 2 Hasil akurasi menggunakan Algoritma k-NN

No	Dataset	k-NN
1	Landsat	90,63%
2	Thyroid	68,21%
3	HTRU	96,01%
4	EEGEye	97,40%

Pada eksperimen berikutnya melakukan penghitungan k-NN yang dikomparasikan dengan pendekatan bootstrapping dan Weighted Gini Index yang menggunakan dataset Landsat, Thyroid, HTRU dan Eeg Eye. Mempersiapkan dataset Landsat, Thyroid, HTRU dan EEG Eye sebagai input data.

Data yang disudah disiapkan dilakukan preprocessing menggunakan bootstrap yang digunakan untuk mengurangi jumlah data dengan menggunakan pendekatan relative 80%.

e-ISSN. 2548-4168

Berikut hasil perbandingan jumlah data menggunakan bootstrapping:

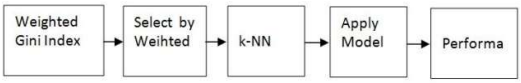
Tabel 3. Perbandingan sebelum dan sesudah bootstrap

Dataset	Jumlah Sample Data Sebelum Bootstrap	Jumlah Sample Data Sesudah Bootstrap
Landsat	6435	5148
Thyroid	7200	5760
HTRU	17898	14318
EEGEye	14980	11984

Adapun alurnya adalah sebagai berikut:

Gambar 2. Alur pengurangan jumlah data dengan menggunakan metode bootstrap

Pada tahap selanjutnya melakukan pembobotan atribut menggunakan Weighted Gini Index yang dilanjutkan dengan normalisasi sampai pemilihan jumlah atribut yang diperlukan untuk pengolahan tahap berikutnya. Pada pembobotan ini penentuan nilai top k yang artinya menentukan jumlah atribut terbaik (memiliki tingkat error rendah) juga sangat berpengaruh terhadap akurasi dan waktu komputasi.



Gambar 3. Alur preprocessing k-NN menggunakan Weighted Gini Index sesudah bootstrap.

Tabel 4. Perbandingan jumlah atribut sebelum dan sesudah Weighted Gini Index

Dataset	Jumlah Atribut Sebelum Weighted Gini Index	Top k	Jumlah Atribut Sesudah Weighted Gini Index
Landsat	36	30	30
Thyroid	21	14	14
HTRU	8	7	7
EEGEye	15	14	14

Penentuan nilai k sangat dominan dalam menentukan akurasi dan waktu yang terbaik. Penghitungan akurasi mengunakan matriks confution dan dapat dikalukan dengan mengggunakan rumus sebagai berikut:

$$Akurasi = \frac{\text{jumlah data yang diprediksi dengan benar}}{Jumlah\ prediksi\ yang\ dilakukan} \text{ x } 100\%$$

Dan berikut hasil akurasi yang didapatkan dari hasil analisa:

Tabel 5 Hasil Penghitungan akurasi masing-masing dataset menggunakan pendekatan Bootstrapping-Weighted Gini Index

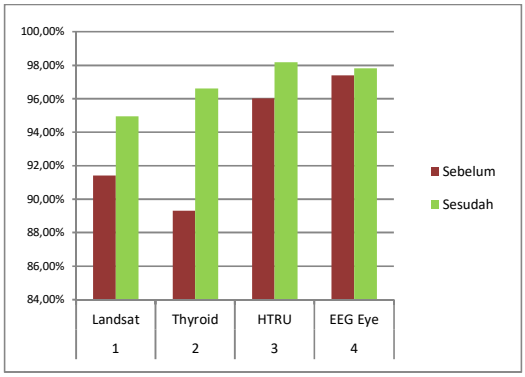
No	Dataset	Akurasi dalam persen	Waktu
1	Landsat	94,95%	1 detik
2	Thyroid	96,61%	1 detik
3	HTRU	98,18%	3 detik
4	EEGEye	97,87%	4 detik

Metode Sample Bootstrapping digunakan untuk mengurangi jumlah data training. (Zikopoulos *et al*., 2012; Wan *et al*., 2012; Weidong *et al*., 2014). Pada penelitian Shankar telah membahas mengenai penerapan prinsip Gini Indeks dalam pilihan teks dan masalah penyesuaian bobot, namun cakupannya terbatas pada klasifikasi berbasis sentroid. (Lin & Dong, 2006). Berdasarkan analisis prinsip dan fitur teks Gini index, akan membangun fungsi evaluasi secara langsung di ruang fitur asli untuk pemilihan fitur, lalu memilih subset fitur yang paling signifikan. (Breiman *et al*., 1984). Sesuai dengan keempat penelitian yang terkait, dalam pengolahan jumlah data yang besar pada algoritma k-NN, penggunaaan metode untuk meningkatkan akurasi algoritma k-NN untuk menentukan hasil atau kesimpulan, dilakukan dengan berbagai macam metode. Pendekatan Template Reduction yang digunakan untuk menghilangkan nilai yang jaraknya jauh dari batasan threshold kurang signifikan pengaruhnya terhadap klasifikasi k-NN. (Witten *et al*., 2011). Penerapan algoritma Direct Boosting dalam penelitian yang dilakukan oleh

Wu *et al*., (2009) dengan cara memodifikasi pembobotan jarak pada data latih dengan local warping of distance matric yang berguna untuk meningkatkan akurasi k-NN. Klasifikasi hybrid dengan menggabungkan antara algoritma SVM dan k-NN (SVM-NN) dalam mengatasi ketergantungan parameter yang rendah untuk menghasilkan akurasi yang terbaik pada pemprosesan *big data* (Fayed & Atiya, 2009). Penelitian yang menerapkan metode bootstrapping dan Weighted Principal Component Analysis (WPCA) pada algoritma k-NN yang digunakan untuk mengurangi jumlah data

Tabel 6. Pendapatan akurasi masing-masing dataset penelitian.

NO	DATASET	KNN		BOOTSTRAP-WEIGHTED GINI INDEX-KNN	
		AKURASI	WAKTU	AKURASI	WAKTU
1	LANDSAT	90,52%	2 detik	94,95%	1 detik
2	THYROID	89,31%	2 detik	96,61%	1 detik
3	HTRU	96,01%	4 detik	98,18%	3 detik
4	EEGY Eye	97,40%	6 detik	97,87%	4 detik



Gambar 4. Grafik peningkatan akurasi sebelum dan sesudah menggunakan pendekatan Bootstrap-Weighted Gini Index.

SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan dan pengujian klasifikasi data Landsatyang diolah menggunakan metode integrasi Bootstrap dan Weighted Gini Index pada Algoritma KNN, dapat memberikan kesimpulan bahwa:

1. Didapatkan model klasifikasi data Landsat, Thyroid, HTRU, EEG Eye dengan menggunakan metode integrasi Bootstrap dan Weighted Gini Index pada Algoritma KNN dengan akurasi masing-masing Landsat = 94,95 %, Thyroid = 96,61%, HTRU = 98,18% dan EEG Eye = 97,87%
2. Waktu proses komputasi yang semakin singkat yaitu Landsat = 1 detik dibanding dengan sebelumnya 2 detik, Thyroid = 1 detik dibanding dengan sebelumnya 2 detik, HTRU = 3 detik dibanding dengan sebelumnya 5 detik dan EEG Eye = 4 detik dibanding dengan sebelumnya 6 detik.
3. Dari hasil eksperimen telah diketahui bahwa metode Bootstrap dan Weighted Gini Index dapat berpengaruh terhadap Algoritma KNN.Jika dibandingkan dengan klasifikasi menggunakan metode Preprosesing k-NN menggunakan integrasi Metode Bootstrapping dan WPCA maupun KNN biasa, pada Landsat akurasi dapat meningkat sebesar 3,55% (94.95% - 91,40%), Thyroid 7,30% (96,61%-89,31%), HTRU 2,17% (98,18%-96,01), EEG Eye 0,41% (97,81%-97,40%).

DAFTAR PUSTAKA

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

Breiman L., Friedman J. H., et al, Classification and Regression Trees.Monterey, CA: Wadsworth International Group, 1984.

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

Han, J., & Kamber, Data mining Concept and Techniques (M. Han, J., & Kamber, Ed) (Thirt Edit) USA: Morgan Kaufmann Publishers, 2012.

Heriyanto, M. Ari dan Wisnu AP. 2008. *Pemrograman Bahasa C Untuk Mikrokontroler ATMEGA 8535*. Yogyakarta: ANDI

Lin & Dong at all An Adaptive Fuzzy kNN Text Classifier Based on Gini Index Weight Computers and Communications, 2006.

Morimune, K., & Hoshino, Y.Testing homogenity of a large data set by boostrapping Mathematics Ans Computers In Simaulation, 78,292-302. doi:10.1016/j.matcom.2008.01.021, 2008

Neo, T. K. C., & Ventura, D. A direct boosting algoritm for the k-nearest neightbor classifier via local warping of the distance matric Pattern Recoqnation Latters, 33(1), 92-1-2. doi: 10.1016/j.patrec.2011.09.028, 2012.

O'Reilly, Big Data Now Edition (Fisrs Edit, O'Reilly Media ,Inc, 2012.

S. Shankar, G. Karypis, A Feature Weight Adjustment Algorithm for Document Categor ization. http://www.cs.umm.edu/~karypis..

T. Pang-Ning, M. Steinbach and V. Kumar, Introduction to data mining, Libr. Congr., p. 796, 2006.

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

Weidong, Jingyu & Yongmin, Using Gini-Index for Feature Selection in Text Categorization, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China, College of Economics And Management, Hebei Polytechnic University, 2014.

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

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

Zikopoulos, P., Eaton, C., & DeRoos, D. Understanding big data New York et al:MeGraw Hill. doi:1 0 9 8 7 6.5.4.3.2.1, 2012.