PENERAPAN ALGORITMA C4.5 BERBASIS *ADABOOST* UNTUK PREDIKSI PENYAKIT JANTUNG

Abdul Rohman[1], Vincent Suhartono[2], Catur Supriyanto[3]
[123]Pasca Sarjana Teknik Informatika Universitas Dian Nuswantoro

ABSTRACT

Heart disease is the occurrence of partial or total blockage of a blood vessel over, as a result of the self peyumbatan deep chemical energy supply to the heart muscle is reduced, resulting in impaired balance between supply and needs .Research in predicting heart disease have been carried out by several previous investigators. In this study will be done for heart disease prediction algorithm using C4.5 and improved the performance of C4.5 algorithm using Adaboost method is implemented on the data of heart disease patients. From the test results by measuring method using a C4.5-based Adaboost, confusion matrix, and the ROC curve, it is known that C4.5 algorithms yield accuracy values 86,59%, AUC values obtained after 0.957 and optimized by using the method to be 92,24% Adaboost, the AUC to 0.982. by looking at the accuracy and AUC values after the optimizations, the algorithmbased C4.5 classification Adaboost into the category of groups is very good, because AUC values between 0.90 – 1.00

Keyword: prediction, C4.5, Adaboost, heart disease

1. LATAR BELAKANG

Industri kesehatan memiliki sejumlah besar data kesehatan, namun sebagian besar data tersebut tidak diolah untuk mengetahui informasi tersembunyi untuk dijadikan pengambilan keputusan yang efektif oleh para praktisi kesehatan. Pengambilan keputusan atas dasar data dan informasi yang akurat akan menghasilkan keputusan dan prediksi penyakit menjadi tepat sasaran.

Penyakit jantung di Indonesia merupakan penyakit nomor satu yang mendorong angka kematian yang cukup tinggi, sehingga sampai sekarang penyakit tersebut ditakuti oleh manusia. Oleh karena itu penyakit jantung perlu diprediksi yaitu dengan menggunakan klasifikasi *data mining* sehingga praktisi kesehatan dalam pengambilan keputusan bisa lebih tepat dan akurat.

Banyak penelitian prediksi penyakit jantung dengan teknik klasifikasi *Data mining,* diantaranya dilakukan oleh Palaniappan dan Awang [1] dengan melakukan komporasi 3 metode yaitu *Naives Bayes, Decision Tree*, dan *Artificial Neural Network (ANN)* dengan total kasus 909 dan 15 atribut. Hasil dari penelitian tersebut metode *Decision tree* menghasilkan nilai terbaik.

Penelitian yang dilakukan oleh Anbarasi dkk. [2] dalam memprediksi kelangsungan hidup penyakit jantung dengan berdasarkan 909 kasus dan 6 Atribut dengan menggunakan metode *Naïve Bayes, Decision Tree* dan *Clasification Via Clustering*. Hasil penelitian tersebut metode *Decision Tree* menghasilkan nilai terbaik.

Selain itu juga S.B. Kotsiantis dalam review papernya menjelaskan, bahwa metode *Decision Tree* mempunyai kelebihan-kelebihan dalam mengolah *Dataset* penyakit jantung yaitu dari segi; kecepatan dalam klasifikasi, tiap atribut bersifat diskrit, binari dan kontinue, serta transparansi pengetahuan atau klasifikasi [3]

Berdasarkan atas penelitian di atas, peneliti memilih metode *Decision Tree* dalam memprediksi penyakit jantung. Dalam penelitian ini dilakukan penerapan algoritma *Decision Tree (C4.5)* menggunakan metode *Adaboost* dengan mengoptimalkan atribut-atribut yang berasal dari *Dataset* yang terpercaya untuk memprediksi penyakit jantung dengan tujuan agar akurasi menjadi meningkat.

Berdasarkan hasil penelitian sebelumnya algoritma C4.5 dalam prediksi penyakit jantung akurasinya masih belum mencapai *level excellence* untuk itu akurasi model *C4.5* perlu ditingkatkan dengan metode *Adaboost* dalam memecahkan masalah prediksi jantung.

Tujuan dari penelitian ini adalah melakukan optimasi algoritma C4.5 berbasis *Adaboost* dengan melakukan perulangan *(iteration)* dan *attribute wighting* untuk meningkat akurasi dalam prediksi penyakit jantung.

Penelitian ini diharapkan agar dapat digunakan oleh para praktisi kesehatan seperti dokter untuk sebagai masukan untuk prediksi penyakit jantung dan memberikan sumbangsih bagi pengembangan ilmu pengetahuan yang berkaitan dengan prediksi penyakit jantung dengan menggunakan algoritma *C4.5* berbasis metode *Adaboost*.

2. LANDASAN TEORI

2.1. Algoritma C4.5

Algoritma *decision tree* digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar [4] Pohon keputusan adalah salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi oleh manusia. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan.

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma *C4.5* [4] yaitu:

a. Mempersiapkan data *training*, dapat diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokan dalam kelas-kelas tertentu.

b. Menentukan akar dari pohon dengan menghitung nilai *gain* yang tertinggi dari masing-masing atribut atau berdasarkan nilai *index entropy* terendah. Sebelumnya dihitung terlebih dahulu nilai *index entropy*, dengan rumus:

$$Entropy(i) = \sum_{j=1}^{m} \mathrm{f}(i,j).\,2\ \mathrm{f}[(i,j)] \qquad (2.1)$$

c. Hitung nilai *gain* dengan rumus:

$$\mathrm{gain} = -\sum_{i=1}^{p} \frac{ni}{n}.IE(i) \qquad (2.2)$$

d. Untuk menghitung gain ratio perlu diketahui suatu term baru yang disebut Split Information dengan rumus:

$$SplitInformation = -\sum_{t=1}^{c} \frac{S_1}{S} log_2 \frac{S_1}{S} \qquad (2.3)$$

e. Selanjutnya menghitung gain ratio

$$Gainratio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \qquad (2.4)$$

f. Ulangi langkah ke-2 hingga semua *record* terpartisi

Proses partisi pohon keputusan akan berhenti disaat:

1) Semua tupel dalam *record* dalam simpul m mendapat kelas yang sama
2) Tidak ada atribut dalam *record* yang dipartisi lagi
3) Tidak ada *record* didalam cabang yang kosong.

2.2. Metode C 4.5 berbasis *Adaboost*

Setelah melakukan tahapan dalam membuat sebuah pohon keputusan dengan algoritma *C4.5*, dilakukan pemberian bobot pada pohon tunggal sehingga menghasilkan hipotesa baru dan sebuah pohon keputusan baru dengan langkah-langkah sebagai berikut [5]:

Metode *Adaboost* adalah sebagai berikut:

a. Inisialisasi bobot data $\{W_n\}$ dengan $W_n^{(m)}$ untuk n = 1, 2, ……,N.

b. For m = 1 ……..,M.

1) *Training* $y_m(x)$ dengan meminimalkan fungsi kesalahan (*error function*) sebagai berikut:

$$J_m = \sum_{n=1}^{N} W^{(m)}_n I\,(y_m(x_n) \neq t_n \qquad (2.5)$$

2) Evaluasi kesalahan

$$\varepsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} (y_m(x_n) \neq t_n}{\sum_{n=1}^{N} w_n^{(m)}} \qquad (2.6)$$

3) Dan kemudian digunakan evaluasi

$$a_m = In\left\{\frac{1-\varepsilon_m}{\varepsilon_m}\right\} \qquad (2.7)$$

c. Memperbaiki (update) bobot data

$$w_n^{(m+1)} = w_n^{(m)} \exp(a_m I(y_m(x_n) \neq t_n)\} \qquad (2.8)$$

d. Membuat prediksi menggunakan model terakhir sebagai berikut

$$Y_m(x) = sign\,(\sum_{m=1}^{M} a_m y_m\,(x)) \qquad (2.9)$$

2.3. Algoritma C4.5 Berbasis *Adaboost* untuk Prediksi Penyakit Jantung

Kerangka pemikiran penelitian ini berawal dari masalah masih kurang optimalnya akurasi prediksi penyakit jantung. Jantung adalah organ berupa otot, berbentuk kerucut, berongga dengan basisnya di atas dan puncaknya di bawah. Yang fungsinya untuk memompa bersih ke seluruh tubuh dan darah kotor ke paru-paru. Jika terjadi gangguan pada jantung maka fungsi pemompaan darah akan terganggu bahkan bisa mengakibatkan kematian.

Berdasarkan *Dataset* penyakit jantung di UCI (Univercity of California Irvine) terdapat 14 atribut yaitu umur, jenis kelamin, jenis sakit dada, tekanan darah, kolestrol, kadar gula, elekrokardiografi, tekanan darah, angina induksi, oldpeak, segmen_st, flaurosopy, denyut jantung dan hasil sebagai label yang terdiri atas *healthy* (sehat) dan *sick* (sakit). Semua atribut tersebut selain hasil merupakan hal-hal yang mempengaruhi terjadinya penyakit jantung.

Decision Tree digunakan untuk membangun sebuah pohon keputusan namun karena kurang akuratnya penerapan algoritma C4.5 dalam memprediksi penyakit jantung, maka digunakan *Adaboost* untuk pemberian bobot pada pohon agar diperoleh akurasi yang meningkat.

3. METODE PENELITIAN

Penelitian ini menggunakan data pasien yang melakukan pemeriksaaan penyakit jantung yang didapat dari UCI (*Universitas California, Invene*) *Machine Learning Repository* [6]. Hasil yang didapat sebanyak 867 orang yang diperiksa dan sebanyak 364 pasien terdeteksi sakit, sehingga 503 pasien terdeteksi sehat. *Dataset* tersebut adalah penggabungan antara *Dataset* dari Cleveland yang terdiri dari 303 pasien, data dari statlog yang terdiri dari 270 pasien, dan data dari hungaria terdiri dari 294 pasien.

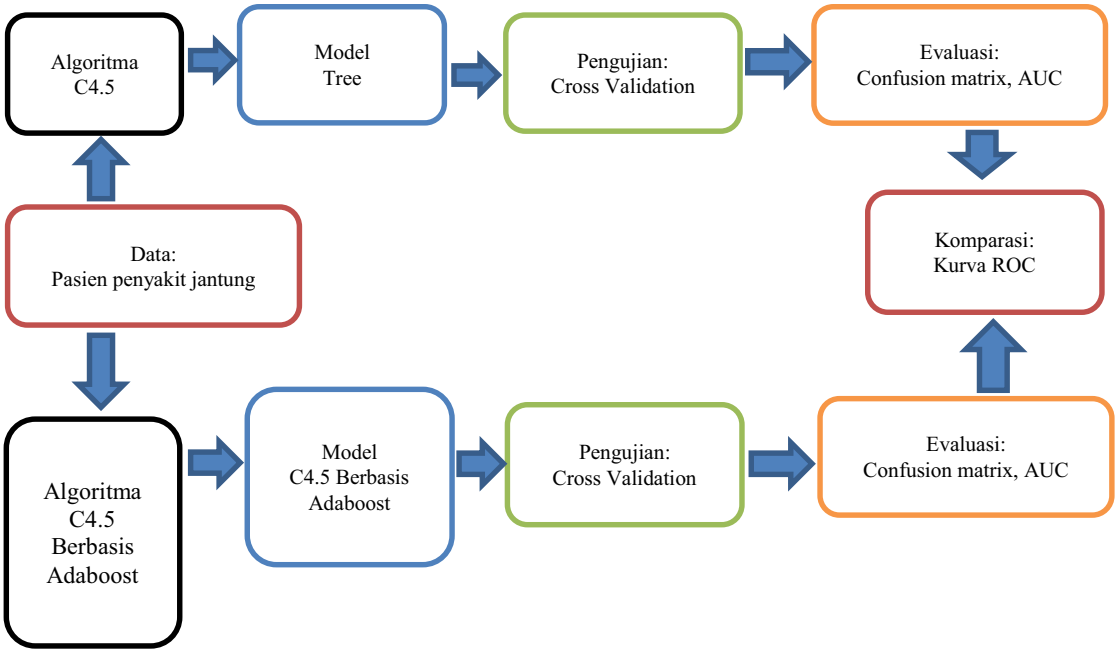
Penelitian ini adalah penelitian *experiment* yang melibatkan penyelidikan tentang perlakuan pada parameter dan variabel yang semuanya tergantung pada peneliti itu sendiri. *software* dan *hardware* sebagai alat bantu dalam penelitian ini adalah sebagai berikut.

Tabel 1. Spesifikasi *Hardware* dan *Software*

Software	Hardware
Sistem operasi: Windows XP SP III 32 bit	CPU: Dual Core 1,7 Ghz Ram 2 GB, Hdd 160Gb
***Data mining*: RapidMiner Versi 5**	

Data yang diperoleh dari UCI akan di *preprocessing* terlebih dahulu supaya data berkualitas dengan cara manual. Jadi data yang diolah dan diteliti sebanyak 567 dengan keadaan sakit sejumlah 257 orang dan keadaan sehat sejumlah 310 orang.

Model yang diusulkan pada penelitian ini adalah menggunakan algoritma *C4.5* dan algoritma *C4.5* berbasis *Adaboost* yaitu:



Gambar 1. Metode Penelitian

4. HASIL PEMBAHASAN

Dalam pengujian K-*Fold Cross Validation* Algoritma C4.5 dam Algoritma C4.5 berbasis *Adaboost*, peneliti juga menggunakan 10 kali percobaan dengan sampling type Stratified (bertingkat-tingkat) dengan menggunakan use local random seed karena hasil akurasinya juga lebih tinggi.

Metode klasifikasi bisa dievaluasi berdasarkan beberapa kriteria seperti tingkat akurasi, kecepatan, kehandalan, skalabilitas, dan interpretabilitas [7]. Hasil pengujian model yang dilakukan dalam bab tiga adalah untuk mengukur tingkat akurasi dan AUC (*Area Under Curve*) dari prediksi penyakit jantung dengan metode *cross validation.*

4.1. Hasil Pengujian Model Algoritma C4.5

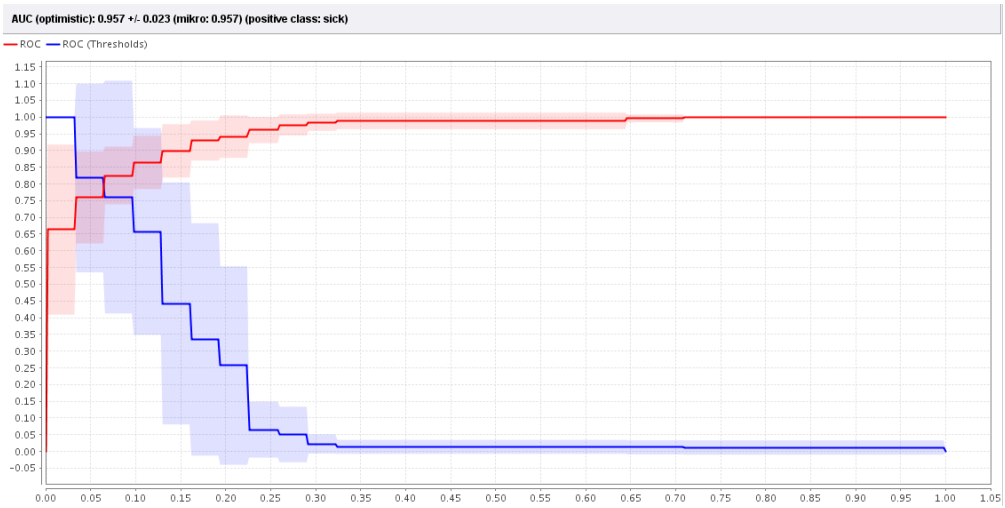
Hasil dari pengujian model yang telah dilakukan adalah untuk mengukur tingkat akurasi dan AUC (*Area Under Curve*).

Tabel 2. Model *Confusion Matrix* untuk Algoritma C4.5

accuracy: 86.59% +/- 4.12% (mikro: 86.60%)

	true healthy	true sick	class precision
pred. healthy	270	36	88.24%
pred. sick	40	221	84.67%
class recall	87.10%	85.99%	

Grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.957 terlihat sebagai berikut.



Gambar 2. Nilai AUC dalam Grafik ROC Algoritma C4.5

4.2. Hasil Pengujian Model Algoritma C4.5 berbasis *Adaboost*

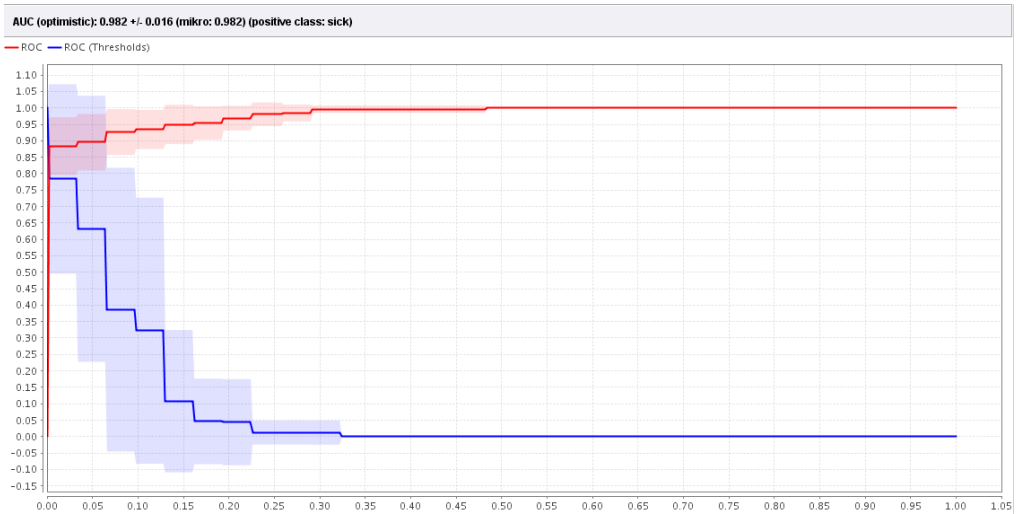
Hasil dari pengujian model yang telah dilakukan adalah untuk mengukur tingkat akurasi dan AUC (*Area Under Curve*).

Tabel 3. Model *Counfusion Matrix* untuk Algoritma C4.5 Berbasis *Adaboost*

accuracy: 92.24% +/- 4.17% (mikro: 92.24%)

	true healthy	true sick	class precision
pred. healthy	291	25	92.09%
pred. sick	19	232	92.43%
class recall	93.87%	90.27%	

Gambar berikut ini menunjukan grafik ROC dengan nilai AUC (*Area Under Curve*) sebesar 0.982



Gambar 3. Nilai AUC dalam Grafik ROC Algoritma C4.5 Berbasis *Adaboost*

4.3. Analisis Evaluasi dan Validasi Model

Dari hasil pengujian di atas, baik evaluasi menggunakan *counfusion matrix* maupun ROC *curve* terbukti bahwa hasil pengujian algoritma C4.5 berbasis *Adaboost* memiliki nilai akurasi yang lebih tinggi dibandingkan dengan algoritma C4.5. Nilai akurasi untuk model algoritma C4.5 sebesar 86,59% dan

nilai akurasi untuk model algoritma C4.5 berbasis *Adaboost* sebesar 92.24% dengan selisih akurasi 5,65%, dapat dilihat pada Tabel di bawah ini.

Tabel 4. Pengujian Algoritma C4.5 dan C4.5 berbasis *Adaboost*

	Accuracy	AUC
C4.5	86,59	0,957
C4.5 Berbasis *Adaboost*	92,24	0,982

Untuk evaluasi menggunakan ROC *curve* sehingga menghasilkan nilai AUC (*Area Under Curve*) untuk model algoritma C4.5 menghasilkan nilai 0.957 dan algoritma C4.5 berbasis *Adaboost* menghasilkan nilai 0.982 dengan nilai diagnosa *Excellent Classification*, dan selisih nilai keduanya sebesar 0.025. Jadi akurasi dari Algoritma C4.5 ke Algoritma C4.5 berbasis *Adaboost* meningkat 6,42% dan nilai AUC dari Algoritma C4.5 ke Algoritma C4.5 berbasis *Adaboost* meningkat 0,26%.

5. KESIMPULAN DAN SARAN

Dalam penelitian ini dilakukan pengujian model dengan menggunakan algoritma *C4.5* dan algoritma *C4.5* berbasis *Adaboost* dengan menggunakan data pasien yang menderita penyakit jantung atau tidak. Model yang dihasilkan diuji untuk mendapatkan nilai *accuracy*, dan AUC dari setiap algoritma. Didapat pengujian dengan menggunakan *C4.5* nilai *accuracy*-nya adalah 86,59 % dengan nilai AUC adalah 0.957. sedangkan pengujian dengan mengunakan *C4.5* berbasis *Adaboost* didapatkan nilai *accuracy* 92.24 % dengan nilai AUC adalah 0.982. Selain itu juga peneliti mengkomparasi dengan algoritma C4.5 berbasis Bagging didapat *accuracy* 91,89% dan nilai AUC 0,963. Maka dapat disimpulkan pengujian model penyakit jantung dengan menggunakan algoritma *C4.5* berbasis *Adaboost* lebih baik dari pada C4.5 sendiri, dengan peningkatan akurasi sebesar 6,42% dan peningkatan nilai AUC sebesar 0,26%.

Dengan demikian dari hasil pengujian model di atas dapat disimpulkan bahwa *C4.5* berbasis *Adaboost* memberikan pemecahan untuk permasalahan penyakit jantung lebih akurat.

Dari hasil pengujian yang telah dilakukan dan hasil kesimpulan yang diberikan maka ada saran atau usul yang diberikan antara lain:

a. Dalam Penelitian ini dilakukan dengan menggunakan metode algoritma *C4.5* dan algoritma C4.5 berbasis metode *Adaboost*. Perlu dilakukan penelitian lanjut untuk mengurangi beberapa atribut dan mengujicobakan kembali dengan algoritma lain dengan mengoptimasi selain *Adaboost* yang menghasilkan tingkat akurasi tinggi.
b. Hasil penelitian ini diharapkan bisa digunakan untuk membangun sistem di rumah sakit untuk meningkatkan akurasi dalam prediksi penyakit jantung.

UCAPAN TERIMAKASIH

Penelitian ini dapat terselesaikan karena banyak pihak-pihak yang mendukung, oleh karena peneliti berterimakasih kepada pihak-pihak yang mendukung terlaksananya penelitian yaitu para pembimbing penelitian, UCI sebagai *Dataset* objek penelitian, serta pihak-pihak lain yang mendukung terlaksananya penelitian ini.

PERNYATAAN ORIGINALITAS

“Saya menyatakan dan bertanggung jawab dengan sebenarnya bahwa Artikel ini adalah hasil karya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya”.[ABDUL ROHMAN P31.2011.00870]

DAFTAR PUSTAKA

[1] Sellappan Palaniappan & Rafiah Awang, "Intelligent Heart Disease Prediction System Using *Data mining* Techniques," *IJCSNS International Journal of Computer Science and Network Security*, vol. 8, Agustus 2008.

[2] E. Anupriya & Inyegar M. Anbarasi, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," *International Journal of Engineering Science and Technology*, vol. 02, 2010..

[3] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Department of Computer Science and Technology University of Peloponnese, Greece*, pp. 249-268, juli 2007.

[4] Florin Gorunescu, *Data mining concepts models and technique*. berlin: Springer, 2011.

[5] Wu, Xingdong, *The Top Ten Algorithm in Data mining*. Minnesota: Taylor & Francis Group, 2009

[6] A Jasoni and W Steinbrunn, *UCI MAchine Learning Repository*.: Retrieved from UCI MAchine Learning Repository:http://archive.ics.uci.edu/ml/*Datasets*/Heart+Disease, 2011.

[7] C. Vercellis, *Business Intelligence: Data mining and Optimization for Decision Making Decision Making*. Southern Gate, Chichester, West Sussex, United Kingdom: John Wiley & Sons Ltd, 2009.