Perbandingan Kinerja Algoritma Decision Tree dan Naive Bayes dalam Prediksi Kebangkrutan

Dian Oktafia
Sistem Informasi Universitas Gunadarma
Jakarta
Universitas Gunadarma Depok, Indonesia
doktafia@yahoo.com

D. L. Crispina Pardede
Sistem Komputer Universitas Gunadarma
Jakarta
Universitas Gunadarma Depok, Indonesia
pardede16@staff.gunadarma.ac.id

Abstract

Kinerja algoritma data mining menjadi satu pertimbangan dalam pemilihan algoritma untuk memprediksi kebangkrutan. Penelitian ini mengukur kinerja an membandingkan hasil pengukuran tingkat akurasi algoritma decision tree dan naïve bayes. Hasil pengukuran menunjukkan persentase akurasi dan eror dari data training dan data tes yang digunakan. Kinerja setiap algoritma diukur berdasarkan persentase akurasi dan error. Variabel-variabel yang digunakan dalam klasifikasi adalah rasio keuangan yang dihitung berdasarkan laporan keuangan tahunan dari periode 2005 s/d 2007 untuk perusahaan yang masih aktif dan laporan keuangan satu tahun sebelum terjadi kebangkrutan untuk perusahaan yang sudah bangkrut. Jumlah data sampel yang digunakan adalah 33 perusahaan yang terdiri dari 22 perusahaan aktif dan 11 perusahaan yang mengalami kebangkrutan. Jenis perusahaan yang digunakan adalah perusahaan manufaktur, grosir (wholesale), pengecer (retail) dan jasa. Alat bantu yang digunakan adalah WEKA 3-4. Penelitian ini menunjukkan bahwa algoritma yang memiliki kinerja yang lebih unggul adalah algoritma naive bayes dengan tingkat akurasi mencapai 100%.

Kata Kunci: Decision Tree, Naive Bayes, Kebangkrutan.

1 Pendahuluan

Berdasarkan penelitian yang dilakukan oleh Mous (2005), kinerja algoritma *decision tree* lebih baik jika dibandingkan dengan *Multiple Discriminant Analysis* (MDA) (Altman,1968) dalam memprediksi kebangkrutan perusahaan. Algoritma *decision tree* juga merupakan algoritma paling popular dalam teknik klasifikasi.

Sedangkan menurut penelitian yang dilakukan oleh Lili Sun dan Shenoy (2004) algoritma yang paling tepat dalam memprediksi kebangkrutan perusahaan adalah algoritma *bayesian network* yang difokuskan pada model *naïve bayes*. Model ini mudah untuk diimplementasikan dan telah terbukti memiliki kinerja yang baik dalam memprediksi kebangkrutan. Pertanyaan yang kemudian mengemuka adalah "Bagaimana tingkat akurasi algoritma *decision tree* jika dibandingkan dengan algoritma *naïve bayes* dalam menghasilkan model prediksi kebangkrutan?"

Variabel yang digunakan adalah rasio keuangan (*financial ratio*) dari laporan keuangan tahunan pada 33 perusahaan. Variabel tersebut digunakan untuk memprediksi apakah perusahaan akan mengalami kebangkrutan atau masih dalam keadaan "sehat". Kinerja algoritma akan terukur berdasarkan keaku-

ratan dan eror yang dihasilkan. Semakin besar keakuratan, kinerja algoritma semakin baik. Semakin kecil eror, menunjukkan bahwa kinerja algoritma tersebut semakin baik.

Tujuan dari penelitian ini adalah membandingkan tingkat akurasi yang dimiliki oleh teknik/model data mining, yaitu teknik *decision tree* dan *naïve bayes*, dalam memprediksi kebangkrutan perusahaan.

2 Tinjauan Pustaka

Teknik data mining sering digunakan dalam dunia bisnis khususnya yang terkait dengan masalah memperediksi kebangkrutan. Teknik *data mining* yang digunakan dalam beberapa penelitian sebelumnya dalam memprediksi kebangkrutan antara lain *neural networks* (Zhang, 1999), *instance based learners* (Park dan Han, 2002), *Bayesian model* (Sarkar dan Sriram, 2001), *rule learners* (Thomaidis, 1999), *decision tree algorithms* (Mckee dan Greenstein, 2000), *support vector machines* (Shin, 2005). Algoritma yang digunakan dalam penelitian ini dijelaskan sebagai berikut :

1. *Decision Tree*

   Decision tree adalah algoritma yang paling banyak digunakan untuk masalah klasifikasi. Sebuah *decision tree* terdiri dari beberapa simpul yaitu *tree's roo*, *internal nod* dan *leafs*. Konsep entropi digunakan untuk penentuan pada atribut mana sebuah pohon akan terbagi (*split*). Semakin tinggi *entropy* sebuah sampel, semakin tidak murni sampel tersebut. Rumus yang digunakan untuk menghitung *entropy* sampel S adalah

   $$Entropy(S) = -p_1 log_2 p_1 – p_2 log_2 p_2 \quad (1)$$

   Pada persamaan 1, p_1 adalah proporsi sampel atau grup yang akan jatuh bangkrut dan p_2 adalah proporsi untuk perusahaan yang tidak akan jatuh bangkrut.

2. Naive Bayes

   Klasifikasi Bayesian adalah klasifikasi statistik yang bisa memprediksi probabilitas sebuah class. Klasifikasi Bayesian ini dihitung berdasarkan Teorema Bayes

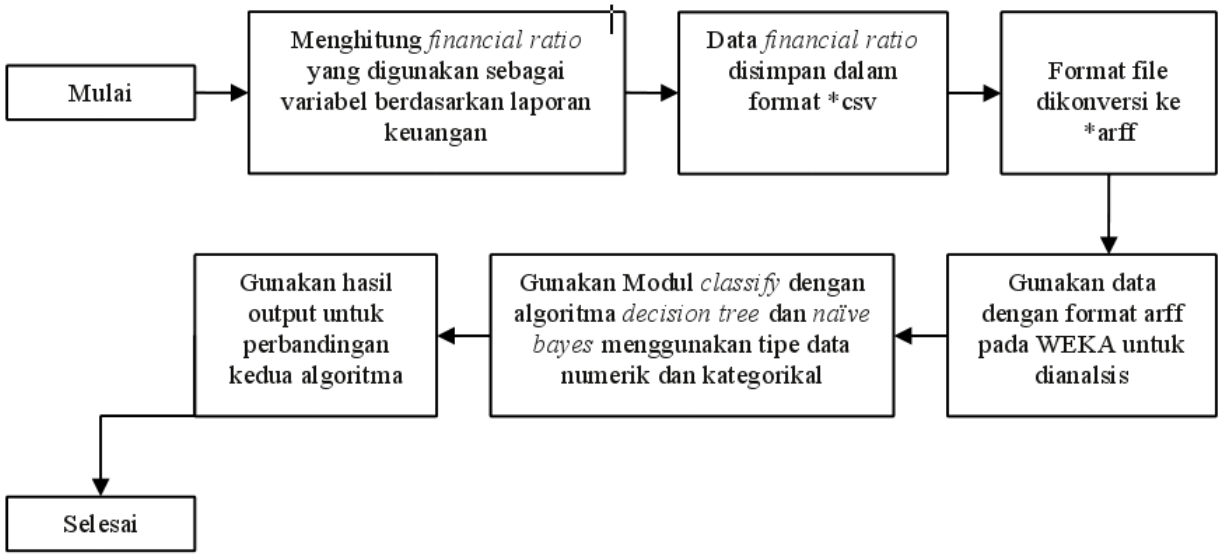
   $$P(H|X) = \frac{P(H|X)P(H)}{P(X)} \quad (2)$$

   Dalam konteks prediksi kebangkrutan, berdasarkan rumus di atas, kejadian H merepresentasikan kebangkrutan perusahaan. $P(H)$ adalah prior probability di mana dalam kasus ini merupakan probabilitas perusahaan yang mendeklarasikan bangkrut.$P(H|X)$ merefleksikan probabilitas perusahaan dengan data X akan mengalami kebangkrutan. $P(X|H)$ adalah posterior probability yang menunjukkan kemungkinan terjadinya kebangkrutan berdasarkan prediktor X. $P(X)$ merupakan prior probability dalam hal ini adalah probabilitas sebuah perusahaan dengan kriteria X.

3 Metode Penelitian

Penelitian yang dilakukan untuk membandingkan algoritma *decision tree* dan *algoritma naïve bayes* (Gambar 1) terdiri atas tiga langkah:

1. Menyiapkan data berupa rasio keuangan yang dihitung berdasarkan laporan keuangan dari jenis perusahaan manufaktur, grosir (*wholesale*), pengecer (*retail*) dan jasa. Laporan keuangan bagi perusahaan yang masih aktif diperoleh dari perusahaan-perusahaan yang terdaftar di Bursa Efek Indonesia (BEI). Sedangkan data perusahaan yang mengalami kebangkrutan diambil dari data yang digunakan pada penelitian sebelumnya yang dilakukan oleh Cindy Yoshiko Shirata (1998). Perusahaan yang dijadikan sampel adalah sebanyak 33 perusahaan, terdiri dari 22 perusahaan yang masih aktif dan 11 perusahaan yang sudah bangkrut. Jumlah sampel yang digunakan berdasarkan rasio 1 : 2 yang dikemukakan oleh Wo-Chiang Lee (2004). Mengingat waktu/periode perusahaan yang bangkrut



Gambar 1: Skema Penelitian

tidak sama, maka untuk perusahaan tersebut digunakan data keuangan satu tahun sebelum perusahaan mengalami kebangkrutan. Sedangkan data perusahaan yang aktif digunakan data keuangan perusahaan tahun 2005 s/d 2007.

2. Data yang ada disimpan dalam format CSV (*Comma Separated Value*, *.csv) yang kemudian akan di convert menjadi format ARFF (*Attribute Relation File Format*, *arff) agar dapat dikenali oleh alat bantu yang digunakan, yaitu WEKA 3-4.

3. Data dalam format ARFF dapat dianalisis menggunakan modul WEKA 3-4. Modul yang digunakan adalah modul *Classify* dengan algoritma *decision tree* dan *naïve bayes*. Kedua algoritma diuji menggunakan dua tipe data, yaitu data numerik dan data kategorikal. Output yang dihasilkan mengandung persentase akurasi dan eror yang menjadi pembanding antara kedua algoritma tersebut.

4 Hasil dan Pembahasan

Rasio keuangan yang digunakan sebagai variabel diambil berdasarkan penelitian yang dilakukan oleh Shirata (1998) yang dijelaskan sebagai berikut :

X2	=	Retained Earnings/ Total Assets
X10	=	(Current period liabilities and shareholders equity / Previous period liability and shareholders equity) - 1
X24	=	Interest and discount expense / (Short term loans + long term borrowings + corporate bond + convertible bond + note receivable discounted)
X36	=	(Note payable + accounts payable) x 12 / Sales

Berdasarkan rumusan di atas, maka didapat nilai variabel untuk masing-masing perusahaan. Data yang diolah menggunakan Microsoft Excel dikonversi menjadi format csv yang kemudian diubah menjadi format file yang dikenali oleh WEKA yaitu arff. Data dengan format file arff, sudah bisa diproses menggunakan modul-modul yang ada di WEKA. Hasil pemrosesan data pada tipe numerik dan kategori diringkas

Tabel 1: Perbandingan Decision Tree (DT) dan Naïve Bayes (NB) Pada Data Numerik

	Use Training Set		Cross Validation		Percentage Split	
	DT	NB	DT	NB	DT	NB
Correctly Classified	100%	100%	96.97%	100%	100%	100%
Incorrectly Classified	0%	0%	3.03%	0%	0%	0%
Kappa Statistic	1	1	0.9333	1	1	1
Mean Absolute Error	0	0	0.0303	0.001	0	0
Root Mean Squared Error	0	0.0001	0.1741	0.0047	0	0.0001
Relative Absolute Error	0	0.00%	6.75%	0.22%	0	0.011
Root Relative Squared Error	0	0.01%	36.80%	0.99%	0	0.03%

Tabel 2: Perbandingan Decision Tree (DT) dan Naïve Bayes (NB) Pada Data Kategori

	Use Training Set		Cross Validation		Percentage Split	
	DT	NB	DT	NB	DT	NB
Correctly Classified	87.88%	90.91%	84.85%	87.88%	100%	100%
Incorrectly Classified	12.12%	9.09%	15.15%	12.12%	0%	0%
Kappa Statistic	0.7	0.7805	0.6154	0.7	1	1
Mean Absolute Error	0.2051	0.127	0.2265	0.1998	0.1778	0.1392
Root Mean Squared Error	0.3203	0.2719	0.3465	0.3263	0.1988	0.1552
Relative Absolute Error	45.83%	28.38%	50.41%	44.48%	40.40%	31.64%
Root Relative Squared Error	67.92%	57.66%	73.25%	68.99%	44.44%	34.71%

dalam Tabel 1.

Secara keseluruhan algoritma naïve bayes kinerjanya lebih baik dibandingkan dengan kinerja algoritma decision tree. Tapi untuk jenis pengujian percentage split dan use training set, kinerja decision tree lebih baik. Hal ini ditunjukkan dengan nilai eror dari decision tree adalah 0. Sedangkan algoritma naïve bayes menghasilkan root mean squared error, relative absolute error, relative squared error untuk pengujian use training set dan percentage split masing-masingnya adalah 0.0001, 0.011, 0.03% dan 0.0001, 0.0039, 0.01%. Perbandingan algoritma decision tree dan naïve bayes pada data kategori dapat dilihat pada Tabel 2 .

Penggunaan data kategori dalam kasus memprediksi kebangkrutan ternyata membuat kinerja masing-masing algoritma menurun jika dibandingkan dengan tipe data numerik. Hasil pada tabel menunjukkan perubahan nilai statistiknya menjadi lebih tidak baik.

5 Penutup

Perbandingan kedua algoritma ini menggunakan tiga model pengujian yaitu use training set, cross validation dan percentage split. Dari ketiga model pengujian ini, model yang paling direkomendasikan adalah cross validation karena pada model ini setiap data yang ada di data sampel mempunyai peluang yang sama untuk menjadi data training dan data tes. Hasil pengujian menggunakan cross validation yang dilihat adalah dari nilai correctly classified. Pada algoritma decision tree, nilai akurasi untuk data numerik adalah sebesar 96.97% dan data kategori sebesar 84.85%. Sedangkan algoritma naive bayes menghasilkan nilai yang lebih besar dibandingkan algoritma decision tree yaitu 100% untuk data numerik dan 87.88% untuk data kategori.

Kesimpulan untuk perbandingan dua algoritma ini adalah secara keseluruhan, kinerja algoritma naïve

bayes lebih baik dibandingkan dengan algoritma decision tree. Kinerja naïve bayes masih tetap unggul ketika pengujian dilakukan pada tipe data kategori.

Dari pengujian yang dilakukan juga dapat disimpulkan bahwa penggunaan data kategori untuk kasus prediksi kebangkrutan perusahaan kurang tepat. Hal ini dapat dilihat dari nilai correctly classified yang menjadi lebih kecil dibandingkan dengan pengujian pada data numerik. Misalnya pada pengujian algoritma decision tree model training set, nilai correctly classified pada data numerik mencapai 100%. Setelah dilakukan pengujian pada data kategori nilainya turun menjadi 87.88%. Nilai mean absolute error pada data numerik menunjukkan angka 0, pada data kategori nilai eror menjadi lebih besar yaitu 0.127. Hal ini menunjukkan penurunan kinerja algoritma.

Pengukuran kinerja sebuah algoritma data mining dapat dilakukan berdasarkan beberapa kriteria antar lain akurasi, kecepatan komputasi, robustness, skalabilitas dan interpretabilitas. Penelitian ini baru menggunakan satu kriteria yaitu berdasarkan akurasi. Akan lebih baik jika semua kriteria diuji coba agar algoritma yang diteliti lebih teruji kinerjanya. Akurasi sebuah algoritma bisa ditingkatkan dengan menggunakan beberapa teknik antara lain teknik bagging dan boosting. Penelitian ini juga belum menggunakan kedua teknik tersebut untuk meningkatkan akurasi karena penelitian ini hanya terbatas pada perbandingan algoritma decision tree dan naïve bayes . Penelitian ini juga menggunakan data sampel yang cukup terbatas yaitu 33 perusahaan yang terdiri dari perusahaan bangkrut dan tidak bangkrut. Untuk mengestimasi akurasi sebuah algoritma akan lebih baik jika jumlah data sampel yang digunakan mendekati populasi yang ada. Diharapkan pada penelitian selanjutnya, data perusahaan yang digunakan lebih banyak dibandingkan penelitian ini agar pengklasifikasian data jauh lebih akurat.

Pustaka

[1] dan Benjamin Tai Andrada Anghelescu. Bankruptcy prediction in the high-tech industry, 2005.

[2] dan Ita Rulina D. Hadad Muliaman, Wimboh Santoso. Indikator kepailitan di indonesia: An additional early warning tools, pada stabilitas sistem keuangan. 2003.

[3] Prakash P. Shenoy dan Lili Sun. Using bayesian networks for bankruptcy prediction : Some methodological issues. In *European Journal of Operational Research*, volume 18, pages 738–753, 2007.

[4] Mandar Haridas. Step by step tutorial for weka, 2007.

[5] dan Eibe Frank Ian H. Witten. *Data Mining, Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann Publisher, San Francisco, 2005.

[6] Daniel T. Larose. *Discovering Knowledge In Data, an Introduction to Data Mining*. Wiley Interscience, New Jersey, 2005.

[7] Wo Chiang Lee. Genetic programming decision tree for bankruptcy prediction.

[8] dan Jiawei Han Michelin Kamber. *Data Mining Concepts and Techniques, Second Edition*. Morgan Kaufmann Publisher, San Francisco, 2006.

[9] Lonneke Mous. Predicting bankruptcy with discriminant analysis and decision tree using financial ratios, 2005. Faculty of Economics at Erasmus University Rotterdam.

[10] Iko Pramudiono. Pengantar data mining: Menambang permata pengetahuan di gunung data, 2003.

[11] Budi Santosa. *Data Mining, Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Graha Ilmu, Yogyakarta, 2007.

[12] Cindy Yoshiko Shirata. Financial ratio as predictors of bankruptcy in japan: An empirical research. In *Apira98*, pages 1–17, 1998.