Pencarian Pola Asosiasi Keluhan Pasien Menggunakan Teknik Association Rule Mining

Ulya Anisatur Rosyidah*, Hardian Oktavianto**
* Universitas Muhammadiyah Jember
** Universitas Muhammadiyah Jember
*ulyaanisatur@gmail.com, **hardian.oktavianto@gmail.com

ABSTRAK

Perkembangan dan pertumbuhan data di bidang kesehatan semakin meningkat dan bertambah, baik dari kualitas maupun kuantitas, dilihat dari sisi kualitas, perkembangan data ini mengalami perubahan dari bentuk dokumen tulis menjadi dokumen digital atau yang biasanya kita sebut dengan file. Isu yang muncul adalah apakah informasi yang bisa diambil atau didapatkan dari sekian banyak data medis yang tersedia hanya berupa informasi – informasi pada umumnya, sedangkan dari suatu basis data yang tersedia seringkali memuat beberapa variabel sekaligus, bahkan apabila diteliti lebih jauh lagi, basis data yang berbeda bisa jadi memuat beberapa variabel yang sama, dari isu tersebut maka diperlukan suatu metode untuk bisa menggali lebih dalam informasi – informasi yang belum diketahui. Berkaitan dengan data medis serta data mining, maka penelitian kali ini akan membahas tentang implementasi atau kegunaan dari data mining pada data kunjungan pasien dengan cara menerapkan association rule mining untuk mendapatkan pola – pola asosiasi dari basis data kunjungan pasien yang tersedia menggunakan algoritma apriori dan algoritma FP-*Growth*. Baik algoritma apriori dan algoritma FP-*Growth* menghasilkan output yang sama. Perbedaan hasil uji coba terletak pada jumlah rule asosiasi yang ditemukan, dengan menggunakan algoritma apriori ditemukan 3 buah rule asosiasi, sedangkan ketika digunakan algoritma FP-*Growth* ditemukan 2 buah rule asosiasi, hal ini terjadi pada saat uji coba yang dilakukan menggunakan confidence sebesar 80%.

Keyword: *data mining, association rule, apriori,* FP-*Growth*

1. Latar Belakang

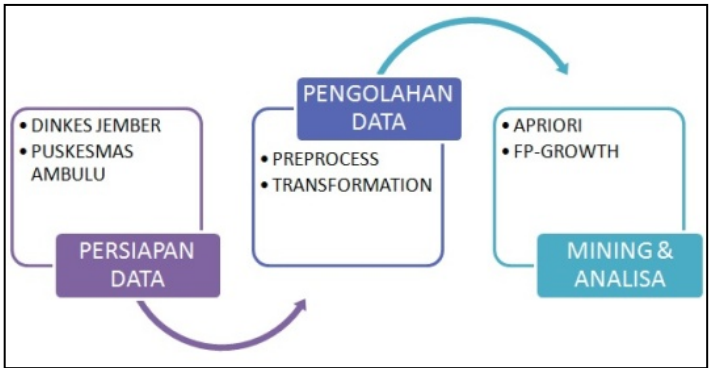
Perkembangan dan pertumbuhan data di bidang kesehatan semakin meningkat dan bertambah, baik dari kualitas maupun kuantitas, dilihat dari sisi kualitas, perkembangan data ini mengalami perubahan dari bentuk dokumen tulis menjadi dokumen digital atau yang biasanya kita sebut dengan *file*, baik file dokumen, gambar, maupun video, sedangkan apabila dilihat dari sisi kuantitas pertumbuhan data di bidang kesehatan mengalami pertambahan jumlah dari waktu ke waktu [2][4]. Sebagai contoh, data yang paling sering kita temui adalah data kunjungan pasien, data rekam medis pasien, data transaksi obat, data penyakit, sampai dengan data citra medis (hasil *USG*, hasil *CT-Scan*, hasil *Rontgen*), pada umumnya tumpukan data ini disimpan dan digunakan oleh pihak terkait untuk berbagai keperluan sesuai kebutuhan, data rekam medis biasanya digunakan sebagai sarana kontrol terhadap pasien, data transaksi obat untuk sumber informasi persediaan obat serta distribusinya, data penyakit untuk mengetahui jenis – jenis penyakit serta cara penanggulangannya. Isu yang muncul adalah apakah informasi yang bisa diambil atau didapatkan dari sekian banyak data medis yang tersedia hanya berupa informasi – informasi pada umumnya, sedangkan dari suatu basis data yang tersedia seringkali memuat beberapa variabel sekaligus, bahkan apabila lebih teliti lagi, basis data yang berbeda bisa jadi memuat beberapa variabel yang sama, dari isu tersebut maka diperlukan suatu metode untuk bisa menggali lebih dalam informasi – informasi yang belum diketahui [1][5][7].

Data Mining merupakan salah satu teknik untuk menemukan, mencari, atau menggali informasi atau pengetahuan baru dari sekumpulan data yang sangat besar, dengan integrasi atau penggabungan dengan disiplin ilmu lain seperti statistika, kecerdasan buatan, serta *machine learning*, menjadikan *data mining* sebagai salah satu alat bantu untuk menganalisa data yang kemudian menghasilkan informasi yang berguna [4]. Fungsi *data mining* yang sering digunakan adalah untuk klasifikasi, klasterisasi, estimasi, prediksi, serta penemuan pola asosiasi. *Association rule mining* merupakan salah satu teknik dalam *data mining* yang berguna untuk menemukan pola asosiasi tersembunyi dalam suatu basis data, pola yang dimaksud disini adalah keterkaitan atau korelasi antara tiap tiap *item* yang berbeda pada setiap *record* di dalam basis data. Pola asosiasi yang ditemukan nantinya berupa *rule – rule* dengan masing – masing nilai bobot asosiasinya, *rule* yang terbentuk biasa dinotasikan dengan X → Y dimana X dan Y disini adalah *itemset*, bobot asosiasi disini berupa nilai *support* yang menjelaskan berapa kali sebuah *itemset* tercatat atau muncul dari sejumlah dataset dan nilai *confidence* yang menjelaskan seberapa kuat hubungan diantara *itemset* X dan Y[6][8].

Berkaitan dengan data medis serta *data mining*, maka penelitian kali ini akan membahas tentang implementasi atau kegunaan dari *data mining* pada data kunjungan pasien dengan cara menerapkan *association rule mining* untuk mendapatkan pola – pola asosiasi dari basis data kunjungan pasien yang tersedia menggunakan algoritma apriori dan algoritma *FP-Growth* serta *software* bantu WEKA [5], kedua algoritma ini sengaja dipakai untuk menunjukkan sejauh mana kemampuan masing – masing algoritma di dalam penemuan pola – pola asosiasi, sampai saat ini apriori merupakan algoritma yang paling sering dipakai karena kemudahan implementasinya, sedangkan algoritma *FP-Growth* merupakan revisi dari apriori yang mana dilakukan perbaikan ketika proses *scanning dataset* dilakukan[3].

2. Metode Penelitian

Untuk dapat menemukan pola asosiasi keluhan pasien maka tentunya harus didapatkan dulu data yang akan diproses, setelah pengumpulan data telah selesai maka akan dilakukan *preprocess* data kemudian dilanjutkan dengan *data transformation*, langkah selanjutnya yaitu melakukan *association rule mining* sehingga pola – pola asosiasi keluhan pasien bisa didapatkan. Deskripsi alur penelitian bisa dilihat pada gambar 1, yang diikuti dengan penjelasan mengenai tahapan – tahapan penelitian. Pada penelitian ini dibagi menjadi 3 tahap utama yaitu : Tahap persiapan data, tahap pengolahan data, serta tahap mining dan analisa hasil keluaran.



Gambar 1. Alur Penelitian

Tahap Persiapan Data

Data yang digunakan pada penelitian ini adalah data kunjungan pasien puskesmas Ambulu pada rentang waktu Agustus 2011 sampai januari 2012 dan diambil dari Dinas Kesehatan Kabupaten Jember. Data yang diambil termasuk data sekunder karena data ini merupakan hasil olahan Dinas Kesehatan yang berasal dari laporan bulanan tiap – tiap puskesmas di Kabupaten Jember. Data yang diperoleh berbentuk file .xls (format Microsoft Excel), jumlah record sebanyak 16.169 buah, yang mempunyai 3 atribut yaitu nomor pasien, keluhan, serta kode penyakit. Data yang diperoleh ini tidak bisa langsung digunakan dalam proses association *rule* mining, sebelumnya data harus diperiksa apakah tidak ada data yang berulang, apakah terdapat duplikasi data, atau bahkan apakah ada data yang tidak lengkap, selain itu diperiksa juga apakah dari ketiga atribut tersebut dipakai semuanya ataukah tidak. Proses tersebut akan dilakukan pada tahap selanjutnya yaitu tahap pengolahan data, dimana nantinya akan ada 2 sub proses, *preprocess*, dan *data transformation*.

Tahap Pengolahan Data

Data yang sudah diperoleh kemudian akan dipersiapkan untuk proses *association rule mining* melalui *preprocess* dan *transformation* [4][6][8], kedua proses ini menggunakan alat bantu yaitu *software* Microsoft Excel 2007 (*evaluation copy*). *Preprocess*, bertujuan untuk membersihkan dan memilih data, yang dimaksud dengan membersihkan data adalah memeriksa apakah ada duplikasi data, kesalahan pengetikan, serta data yang tidak lengkap, sedangkan yang dimaksud dengan pemilihan data adalah memilih variabel data yang diperlukan saja. Setelah melalui *preprocess* maka hanya tersedia 670 data saja yang siap untuk proses selanjutnya, hal ini disebabkan karena pada variabel "keluhan" banyak berisi keluhan tunggal, keluhan berupa rujukan, dan keterangan yang kosong. *Transformation*, bertujuan untuk merubah format data yang sudah ada menjadi format yang bisa diproses, format yang dipakai adalah format biner, dimana pada satu transaksi berisi angka 0 (nol) atau 1 (satu), dimana 0 (nol) merepresentasikan "tidak ada", sedangkan 1 (satu) merepresentasikan "ada".

Tahap Mining dan Analisis

Pencarian pola asosiasi diantara keluhan pasien akan dikerjakan dengan dua algoritma, apriori dan FP-*Growth*, dengan batasan nilai *support* dan *confidence* yang diinputkan oleh *user*, setelah pola asosiasi didapatkan maka tahap yang terakhir adalah melakukan analisa terhadap *rule – rule* yang terbentuk, sehingga diharapkan akan ditemukan informasi baru yang menarik dan berguna. Berdasarkan dataset yang diberikan *rule* yang terbentuk biasa dinotasikan dengan X → Y dimana X dan Y disini adalah *itemset*, bobot asosiasi

disini berupa nilai *support* yang menjelaskan berapa kali sebuah *itemset* tercatat atau muncul dari sejumlah dataset dan nilai *confidence* yang menjelaskan seberapa kuat hubungan diantara *itemset* X dan Y.

3. Hasil dan Analisis

Pada bab ini akan disampaikan tentang hasil penelitian serta analisis dari hasil penelitian tersebut. Bab ini dibagi menjadi 2 sub bab, yang menjelaskan tentang dua tahapan penelitian yaitu tahap pengolahan data dan tahap mining dan analisis.

3.1. Tahap Pengolahan Data

Tahap pengolahan data ini terbagi menjadi 2 sub proses yaitu *preprocess* dan *transformation*. Tahap *preprocess*, bertujuan untuk membersihkan dan memilih data, yang dimaksud dengan membersihkan data adalah memeriksa apakah ada duplikasi data, kesalahan pengetikan, serta data yang tidak lengkap, sedangkan yang dimaksud dengan pemilihan data adalah memilih variabel data yang diperlukan saja. Tahap *transformation* merubah format data yang sudah ada menjadi format yang bisa diproses, format yang dipakai adalah format biner, dimana pada satu transaksi berisi angka 0 (nol) atau 1 (satu), dimana 0 (nol) merepresentasikan "tidak ada", sedangkan 1 (satu) merepresentasikan "ada".

PUSING	PANAS	FLU	BATUK	SESAK	PILEK	MUAL
1	0	0	1	0	0	0
0	1	0	0	0	1	0
1	0	1	0	0	0	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
1	0	1	0	0	0	0
0	1	0	1	0	0	0
0	1	1	0	0	0	0
0	1	0	1	0	0	0
0	1	1	0	0	0	0
0	0	1	1	0	0	0
0	1	1	0	0	0	0
1	0	1	0	0	0	0
1	1	0	0	0	0	0
0	0	1	0	1	0	0

Gambar 2. Alur Penelitian

Pada gambar 2 merupakan potongan hasil dari proses pengolahan data, dimana 0 (nol) merepresentasikan "tidak ada", sedangkan 1 (satu) merepresentasikan "ada", jadi dari gambar 2 dapat diartikan bahwa pada tiap baris mewakili pasien, baris pertama, si pasien mengalami pusing dan batuk, sedangkan pada baris kedua pasien mengalami panas dan pilek, begitu juga seterusnya.

3.2. Tahap Mining dan Analisis

Skenario uji coba memakai 670 record data, dengan 5 nilai *confidence* yang berbeda – beda, yaitu 100%, 90%, 80%, 70%, dan 60%, dengan nilai support yang sama yaitu 1%. Alasan mengapa uji coba ini hanya dilakukan dengan nilai *confidence* yang berbeda – beda adalah karena biasanya pola asosiasi yang menarik ada pada kecenderungan pasangan itemset yang sering terjadi (definisi dari *confidence*). Uji coba nantinya akan dilakukan bergantian, yaitu uji coba terlebih dahulu menggunakan algoritma apriori kemudian menggunakan algoritma FP-*Growth*.

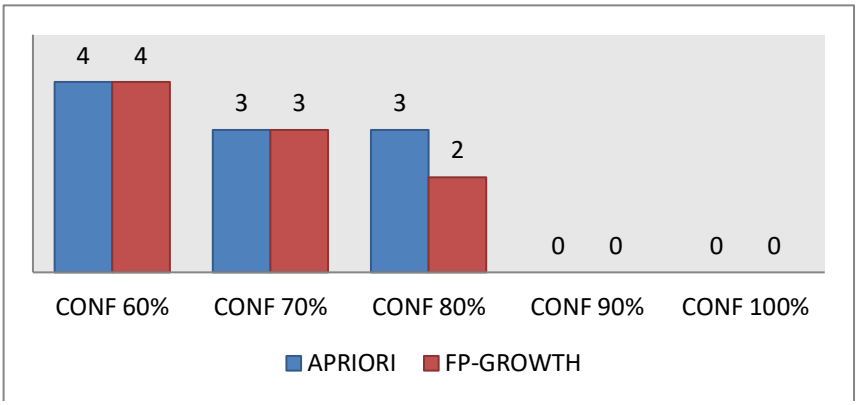
Tabel menunjukkan hasil uji dengan variasi *confidence* yang berbeda – beda, secara konsep, rule – rule yang dihasilkan juga ditampilkan secara berbeda oleh kedua algoritma.

Tabel. 1 Hasil Uji Coba

Uji ke-	Confidence	Hasil Apriori	Hasil FP-Growth

1	60%	Generated sets of large itemsets: Size of set of large itemsets L(1): 6 Size of set of large itemsets L(2): 8 Best rules found: 1. PILEK=t 44 ==> PANAS=t 37 conf:(0.84) 2. FLU=t 344 ==> PANAS=t 288 conf:(0.84) 3. SESAK=t 45 ==> BATUK=t 36 conf:(0.8) 4. BATUK=t 259 ==> PANAS=t 168 conf:(0.65)	FPGrowth found 4 rules (displaying top 4) 1. [PILEK=t]: 44 ==> [PANAS=t]: 37 <conf:(0.84)> lift:(1.08) lev:(0) conv:(1.21) 2. [FLU=t]: 344 ==> [PANAS=t]: 288 <conf:(0.84)> lift:(1.07) lev:(0.03) conv:(1.32) 3. [SESAK=t]: 45 ==> [BATUK=t]: 36 <conf:(0.8)> lift:(2.07) lev:(0.03) conv:(2.76) 4. [BATUK=t]: 259 ==> [PANAS=t]: 168 <conf:(0.65)> lift:(0.83) lev:(-0.05) conv:(0.62)
2	70%	Generated sets of large itemsets: Size of set of large itemsets L(1): 6 Size of set of large itemsets L(2): 8 Best rules found: 1. PILEK=t 44 ==> PANAS=t 37 conf:(0.84) 2. FLU=t 344 ==> PANAS=t 288 conf:(0.84) 3. SESAK=t 45 ==> BATUK=t 36 conf:(0.8)	FPGrowth found 3 rules (displaying top 3) 1. [PILEK=t]: 44 ==> [PANAS=t]: 37 <conf:(0.84)> lift:(1.08) lev:(0) conv:(1.21) 2. [FLU=t]: 344 ==> [PANAS=t]: 288 <conf:(0.84)> lift:(1.07) lev:(0.03) conv:(1.32) 3. [SESAK=t]: 45 ==> [BATUK=t]: 36 <conf:(0.8)> lift:(2.07) lev:(0.03) conv:(2.76)
3	80%	Generated sets of large itemsets: Size of set of large itemsets L(1): 6 Size of set of large itemsets L(2): 8 Best rules found: 1. PILEK=t 44 ==> PANAS=t 37 conf:(0.84) 2. FLU=t 344 ==> PANAS=t 288 conf:(0.84) 3. SESAK=t 45 ==> BATUK=t 36 conf:(0.8)	FPGrowth found 2 rules (displaying top 2) 1. [PILEK=t]: 44 ==> [PANAS=t]: 37 <conf:(0.84)> lift:(1.08) lev:(0) conv:(1.21) 2. [FLU=t]: 344 ==> [PANAS=t]: 288 <conf:(0.84)> lift:(1.07) lev:(0.03) conv:(1.32)
4	90%	Generated sets of large itemsets: Size of set of large itemsets L(1): 6 Size of set of large itemsets L(2): 8 Best rules found:	No rules found!
5	100%	Generated sets of large itemsets: Size of set of large itemsets L(1): 6 Size of set of large itemsets L(2): 8 Best rules found:	No rules found!

Pada gambar 3 disajikan grafik perbandingan secara persentase tentang jumlah rule yang dihasilkan berdasarkan skenario uji coba. Perbedaan hasil uji coba terletak pada jumlah *rule* asosiasi yang ditemukan, hal ini terjadi pada saat uji coba yang dilakukan menggunakan *confidence* sebesar 80%, dengan menggunakan algoritma apriori ditemukan 3 buah *rule* asosiasi, sedangkan ketika digunakan algoritma FP-*Growth* ditemukan 2 buah *rule* asosiasi.



Gambar 3 Perbandingan Uji Coba

4. Kesimpulan

Algoritma apriori dan algoritma FP-*Growth* dapat digunakan untuk mencari *rule* asosiasi dalam konteks *association rule mining*, algoritma apriori menggunakan prinsip apriori dalam pencarian *frequent itemset*, yaitu semua subset yang tidak kosong dari sebuah *frequent itemset* pasti juga akan merupakan *frequent*, begitu juga dengan semua superset yang tidak kosong dari sebuah non-*frequent itemset* pasti juga akan merupakan non-*frequent*. Sedangkan algoritma FP-*Growth* menggunakan konsep pembangunan *tree* dalam pencarian *frequent itemset*s yang memungkinkan dapat secara langsung membentuk *frequent itemset* dengan menerapkan prinsip *divide and conquer*.

Baik algoritma apriori dan algoritma FP_*Growth* menghasilkan output yang sama. Perbedaan hasil uji coba terletak pada jumlah *rule* asosiasi yang ditemukan, dengan menggunakan algoritma apriori ditemukan 3 buah *rule* asosiasi, sedangkan ketika digunakan algoritma FP-*Growth* ditemukan 2 buah *rule* asosiasi, hal ini terjadi pada saat uji coba yang dilakukan menggunakan *confidence* sebesar 80%.

Penelitian ini bisa dikembangakan dengan menggunakan kombinasi algoritma yang lain dalam bidang *association rule mining*, selain itu bisa juga dilakukan analisa tentang perbandingan waktu proses ketika melakukan pencarian *rule* asosiasi.

Daftar Pustaka

[1] Danapana, H., Roy, M. S., Effective Data Mining Association Rules for Heart Disease Prediction System, IJCST Vol. 2, October – December, 2011.
[2] Erwin, Analisis Market Basket Dengan Algoritma Apriori dan FP-Growth, Jurnal Generic Vol. 4 No. 2, Juli 2009.
[3] Han, J., et al, Mining Frequent Pattern Without Candidate Generation A Frequent-Pattern Tree Approach, Data Mining and Knowledge Discovery, 8, 53–87, 2004.
[4] Han, J., Kamber, M., Pei, J., Data Mining Concepts and Techniques Third Edition, Morgan Kaufmann Publisher, 2012
[5] Ordonez, C., Santana, C. A., de Braal, L., Discovering Interesting Association Rules in Medical Data, Proccedings of ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 2000, Hal. 78 – 85.
[6] Sumathi, S., Sivanandam, S. N., Introduction to Data Mining and its Applications, Springer, 2006.
[7] Srinivas, K., Rao, G. R., Govardhan, A., Mining Association Rules from Large Datasets Towards Disease Prediction, International Conference on Information and Computer Networks, 2012.
[8] Witten, I. H., Frank, E., Data Mining Practical Machine Learning Tools and Technique, Morgan Kaufmann Publishers, 2005