APLIKASI KORELASI *PEARSON* DALAM MEMBANGUN MODEL *TREE-AUGMENTED NETWORK* (TAN) (Studi Kasus Pengenalan Karakter Tulisan Tangan)

Irwan Budi Santoso
Jurusan Teknik Informatika, Sains dan Teknologi
Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang
irwan.budi331177@gmail.com

Abstrak- *Langkah pertama dalam membangun model pengenalan Tree-Augmented Network (TAN) dengan mengukur besarnya hubungan diantara pasangan fitur objek. Salah satu metode yang dapat digunakan mengukur besarnya keeratan hubungan secara linier diantara pasangan fitur adalah Korelasi Pearson. Aplikasi Korelasi Pearson dalam membangun model Tree-Augmented Network (TAN) dalam penelitian ini, akan diujicobakan pada kasus membangun model pengenalan karakter tulisan tangan. Data fitur karakter tulisan tangan untuk kasus ini, diasumsikan mengikuti distribusi gaussian karena estimasi parameter model pengenalannya menggunakan estimator Maximum Likelihood (ML). Hasil eksperimen dengan menggunakan data training yang terdiri dari 5 jenis karakter tulisan tangan, menunjukkan untuk dimensi fitur karakter tulisan tangan 10x30 (30 fitur), akurasi sistem Korelasi Pearson dalam membangun model TAN untuk mengenali karakter tulisan tangan sebesar 88 %.*

Kata Kunci : *Korelasi Pearson, Tree-Augmented Network, Karakter Tulisan Tangan, Maximum Likelihood*

1. Pendahuluan

Salah satu tujuan utama dalam pengembangan metode pengenalan adalah untuk mendapatkan atau meningkatkan akurasi dalam pengenalan. Akurasi suatu metode dalam melakukan pengenalan sangat dipengaruhi oleh seberapa baik model yang dihasilkan metode tersebut dan seberapa baik estimator yang digunakan. *Tree-Augmented Network* (TAN) adalah salah satu model pengenalan yang secara konsep dibangun dengan memperhatikan hubungan atau dependensi diantara pasangan fitur-fitur objek (Irwan, 2012) dan *Maximum Likelihood* (ML) adalah salah satu *estimator* model yang sering digunakan.

Secara konsep keberhasilan dalam meningkatkan akurasi sistem atau pengenalan suatu objek tergantung dari dua hal yaitu metode yang digunakan dalam membangun model pengenalan serta *estimator* model yang digunakan. Dalam penelitian ini, akan dicoba membahas bagaimana aplikasi motode Korelasi Pearson dalam membangun model TAN.

Model *Tree-Augmented Network* (TAN) adalah model pengenalan atau klasifikasi yang secara teori dibangun dengan memperhatikan hubungan atau dependensi diantara pasangan fitur objek, sehingga salah satu tahapan penting dalam membangun model pengenalan ini adalah dengan mengukur besarnya hubungan diantara pasangan fitur objek. Pada penelitian ini, akan diaplikasikan metode Korelasi Pearson untuk mengukur keeratan hubungan secara linier diantara pasangan fitur objek sebagai tahapan dalam membangun model TAN serta seberapa besar tingkat akuasi sistem yang dihasilkannya.

2. Korelasi *Pearson*

Korelasi *Pearson* digunakan untuk mengetahui tingkat atau keeratan hubungan secara linier antara dua variabel

atau dua fitur objek. Selanjutnya besarnya hubungan ini ditunjukkan dengan koefisien korelasi yang disimbolkan dengan ρ untuk populasi dan r untuk sampel (DeCoursey, 2003). Bila diketahui fitur X_i ={ x_{i1} , x_{i2} ,..., x_{iN} } dan fitur X_j ={ x_{j1} , x_{j2} ,..., x_{jN} } maka koefisien korelasi dirumuskan sebagai berikut:

$$\rho(X_i, X_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}} \quad (1)$$

dengan nilai σ_i^2 , σ_j^2 dan σ_{ij} merupakan elemen-elemen dari matrik kovarian dari X_i dan X_j yang ditulis sebagai berikut:

$$\sum_{ij} = \begin{bmatrix} \sigma_i^2 & \sigma_{ij} \\ \sigma_{ji} & \sigma_j^2 \end{bmatrix} \quad (2)$$

Koefisien korelasi yang digunakan untuk mengukur keeratan diantara pasangan fitur objek ini, untuk selanjutnya sebagai bobot pada *egde* dari sebuah *graph*.

3. Tree-Augmented Network (TAN)

Tree-Augmented Network (TAN) sangat berkaitan dengan *naïve Bayes* klasifier, karena *Tree-Augmented Network* adalah hasil pengembangan dari *naïve Bayes.* Pada *naïve Bayes* klasifier, diasumsikan bahwa diantara fitur objek bersifat independen. Adanya asumsi independen inilah yang menyebabkan *naïve Bayes* klasifier tidak realistis untuk diterapkan karena pada faktanya hampir setiap fitur memiliki hubungan atau bersifat dependen terhadap fitur yang lainnya. Sedangkan pada *Tree-Augmented Network* (TAN) secara konsep dibangun dengan memperhatikan hubungan atau dependensi diantara pasangan fitur-fitur objek. Dalam prakteknya untuk membagun *Tree-Augmented Network* (TAN) dilakukan dengan menemukan *Bayesian network* yang baik dengan varibel kelas sebagai *root*-nya serta diperbolehkannya saling mempengaruhi diatara fitur atau variabel (Friedman,1997).

Jika diketahui $U = \{X_1,..,Xn\}$ adalah sekumpulan data training, selanjutnya *bayesian network* dianggap sebagai *Directed Acyclic Graph* (G) dan merupakan join distribusi probabilitas U. Bila diketahui parameter *network* adalah Θ maka bayesian network dapat ditulis $B =< G,\Theta >$, dan dalam join distribusi probabilitas U (Friedman,1997) ditulis sebagai berikut:

$$P_B(X_1,..., X_n) = \prod_{i=1}^{n} P_B(X_i \mid \prod_{X_i}) = \prod_{i=1}^{n} \theta_{(X_i|\prod_{X_i})} \quad (3)$$

dengan parameter *network* (Θ) berisi parameter $\theta_{x_i|\prod_{x_i}} = P_B(x_i \mid \prod_{xi})$ untuk setiap nilai $x_i \in X_i$ dan $\prod_{xi} \in \prod_{X_i}$, dan $\prod_{X_i}$ merupakan *parent* dari X_i didalam G.

4. Membangun Model *Tree-Augmented Network*

Membangun model TAN (Amy, 2005) (Friedman, 1997) (Murphy,2001) dilakukan berdasarkan prosedur Chow dan Liu yang mempunyai lima tahap, sebagai berikut:

1. Menghitung besarnya dependensi diantara setiap pasangan atribut yaitu X_i dan X_j , $i \neq j$.
2. Membangun *graph* komplit tak berarah dengan *node*-nya merupakan atribut $X_1,..,X_n$. Sedangkan bobot pada *edge* X_i X_j adalah besarnya dependensi antara X_i dan X_j.
3. Membangun *maximum weighted spanning tree* (MWST) dengan algoritma Prim's (Levitin,2003).
4. Membangun sebuah *tree* berarah sebagai hasil transformasi dari *tree (graph)* tak berarah menjadi sebuah *tree* berarah dengan memilih sebuah *root* variabel dan mensetting arah *edge* dari *root* variabel tersebut.
5. Membangun model TAN, dengan menambah simpul (*vertex*) yang diberi label C dan menambah *edge* atau *arc* dari C ke setiap X_i.

5. Estimasi Parameter Model TAN

Dengan memperhatikan pada bab-bab sebelumnya maka nilai parameter Θ dari model TAN dapat ditentukan dengan menggunakan persamaan 4 (Friedman, 1997) (Jesus,1999)

$$\theta_{x_i|\prod_{x_i}|C} = \hat{P}_D(x_i|\prod_{x_i}|C) = \frac{\hat{P}_D(x_i,\prod_{x_i}|C)}{\hat{P}_D(\prod_{x_i}|C)} \qquad (4)$$

dengan $\hat{P}_D(x_i,\prod_{x_i}|C) \approx N(\mu_{X_j\Pi_{X_I}|C},\Sigma_{X_j\Pi_{X_I}|C})$ dan $\hat{P}_D(\prod_{x_i}|C) \approx N(\mu_{\Pi_{X_I}|C},\Sigma_{\Pi_{X_I}|C})$ (irwan, 2012) . sedangkan untuk mendapat parameter dari setiap distribusi normal (*gaussian*) dengan menggunakan *estimator Maximum Likelihood*.

6. METODE PENELITIAN

6.1. Membangun Model TAN dengan Korelasi *Pearson*

Langkah-langkah membangun Model TAN dengan Korelasi *Pearson* untuk pengenalan karakter tulisan tangan dalam bentuk image) dapat dilihat secara detail dapat Algoritma Membangun Model TAN dengan Korelasi *Pearson* (Irwan,2012).

Algoritma Membangun Model TAN dengan Korelasi *Pearson*

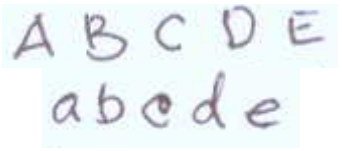
Model – TAN(D)
for setiap X_i, X_j *do*
Tentukan w_{ij} (bobot)
$\{ w_{ij} = \rho(X_i, X_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}} \}$
$GTB \leftarrow GraphTakBerarah(w)$
$BN \leftarrow MWST(GTB)$
{maximum weighted spanning tree}

$TreeB \leftarrow GraphBerarah(BN, root)$
$TAN \leftarrow TambahC(TreeB)$
return TAN

Dari algoritma tersebut, Korelasi *Pearson* digunakan untuk menentukan bobot untuk membentuk *GTB* yaitu graph tak berarah dimana bobot tersebut adalah bobot dari *edge* sutau *graph*. Dengan menggunakan *maximum weighted spanning tree* selanjutnya dibentuk *BN* yaitu *bayesian network*. Dan dengan *BN* dan *root* yang dipilih, dapat dibangun *TreeB* yang merupakan tree berarah dan dari *TreeB* selanjutnya dibangun model *TAN*.

6.2. Data Eksperimen

Untuk membangun model TAN dengan mengaplikasi Korelasi *Pearson* digunakan data nyata dalam bentuk *image* karakter tulisan tangan yang terdiri dari 5 kelas (karakter) yaitu a/A , b/B, c/C, d/D dan e/E (Irwan, 2012) seperti pada **Gambar** 1. yang masing-masing kelas memiliki sampel berukuran 10 pengamatan tanpa membedakan huruf besar atau kecil.



Gambar 1. Data nyata dengan lima jenis Objek karakter tulisan tangan (Irwan,2012)

Selanjutnya data nyata tersebut sebagai data training untuk membentuk model TAN. Untuk dimensi fitur objek setiap data training dibuat sama misalnya lima jenis objek karakter tulisan tangan (a/A , b/B, c/C, d/D dan e/E) dalam bentuk *image* dengan setiap objek ukuran dimensi atau fiturnya sama. Sebagai misal percobaan dengan dimensi objek 10x3 (30 fitur), maka dimensi *image* untuk lima jenis objek yang lain juga dibuat sama dengan ukuran 10x3 (30 fitur).

6.3 Pelaksanaan Eksperimen

Pelaksanaan eksperimen dalam penelitian ini, dengan menggunakan data nyata yaitu lima jenis karakter tulisan tangan dalam bentuk *image* atau citra dilaksanakan dalam beberapa tahap. Adapun langkah-langkah pelaksanaan eksperimen adalah sebagai berikut:

1. Menentukan dimensi fitur objek percobaan yaitu banyaknya fitur

objek yang diekstraksi dari objek (*image*) data training.

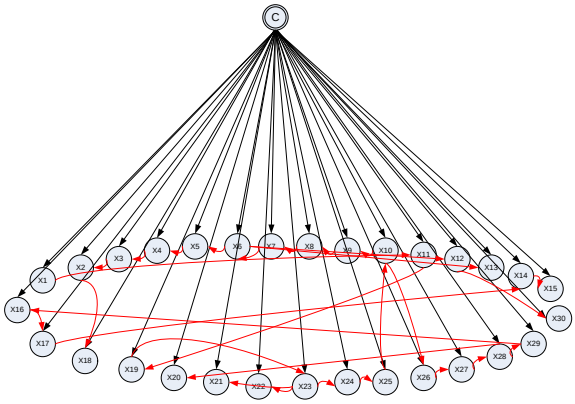
2. Melakukan ekstraksi fitur objek berdasarkan dimensi fitur yang dipilih
3. Membangun Model TAN dengan mengaplikasikan Korelasi Pearson berdasarkan algoritma Membangun Model TAN dengan Korelasi Pearson pada sub-bab 6.1
4. Estimasi parameter model dengan mengguanakan persamaan 4.
5. Pengenalan objek data *training* berdasarkan model yang telah dibangun.
6. Menghitung akurasi sistem dalam mengenali karakter tulisan tangan.

Pada langkah kedua yaitu ekstraksi fitur dari objek training dilakukan setalah objek atau *image* dirubah dalam bentuk grayscale disesuaikan ukuran *image*-nya berdasarkan dimensi fitur objek yang dipilih. Sedangkan pada langkah 5, proses pengenalan objek data *training* dilakukan setelah proses membangun model TAN dan estimasi parameter modelnya selesai dilakukan. Proses ini dilakukan dalam rangka untuk menghitung tingkat akurasi sistem yang sekaligus menunjukkan seberapa baik performa sistem dalam mengenali karakter tulisan tangan dengan menggunakan model TAN.

7. HASIL DAN PEMBAHASAN

Hasil percobaan dengan menggunakan skenario data *training* seperti pada pemahasan sebelumya, dengan dimensi fitur 10 x 3 (30 fitur) diperoleh struktur model TAN seperti pada **Gambar** 2.

Tingkat akurasi sistem yang dihasilkan dengan menggunakan model tersebut adalah 88% dengan perincian pengenalan dapat dilihat pada **Gambar** 3. Matrik *Confusion* Hasil pengenalan Karakter Tulisan Tangan.



Gambar 2. Struktur Model TAN dengan dimensi objek 10x3 (30 fitur) dengan korelasi

	a/A	b/B	c/C	d/D	e/E
a/A	7	1	0	1	1
b/B	0	9	0	0	1
c/C	0	0	8	2	0
d/D	0	0	0	10	0
e/E	0	0	0	0	10

Gambar 3. Matrik *Confusion* Hasil pengenalan Karakter Tulisan Tangan.

Kesalahan dalam mengenali karakter tulisan tangan ditunjukkan oleh penegenalan terhadap karakter A yaitu satu karakter yang sebenarnya karakter A dikenali sebagai karakter B, satu karakter D dan satu karakter E. Berikutnya kesalahan pada pengenalan karakter B yaitu satu karakter B dikenali sebagai karakter E dan kesalahan pada pengenalan karakter C yaitu dua karakter C dikenali sebagai karakter D. sedangkan untuk karakter D dan E hasil pengenlannya benar semua.

8. KESIMPULAN

Korelasi *Pearson* dapat diaplikasikan untuk membangun model pengenalan *Tree-Augmented Network* (TAN), yaitu untuk mengukur besarnya hubungan

diantara pasangan fitur objek. Dari eksperimen yang dilakukan dengan menggunakan data *training* terdiri dari 5 jenis karakter tulisan tangan dan masing-masing fitur karakter diasumsikan mengikuti distribusi *gaussian*, menunjukkan untuk dimensi fitur 10x3 (30 fitur) model TAN yang dibangun dengan Korelasi *Pearson* mampu menghasilkan akurasi sistem sebesar 88%

9. REFERENSI

Amy Ratnakaran, "Bayesian Network", Applied Statistics Honours, Department of Mathematics and Statistics, University of Melbourne, 2005.

W.J. DeCoursey, "Satistics and Probability for Engineering Application With Microsoft® Excel", Elsevier Science (USA), 2003.

Friedman, N. D. Geiger, and M. Goldszmidt, "Bayesian network classifiers", Machine Learning, vol.29, hal 131–163, 1997.

Irwan B.S, 2012, Model Pengenalan Terbaik dengan Tree-Augmented Network (TAN) dan Estimator Maximum Likelihood (ML) Berdasarkan Fitur Objek, MATICS (jurnal Ilmu Komputer dan Teknologi Informasi UIN Malang), vol.4. no.5, hal 197-203.

Jesus Cerquides, " Applying General Bayesian Techniques to Improve TAN Induction", UBS AG Bahnhofstrasse 45, 1999.

A. Levitin, "Introduction The Design & Analysis of Algorithms", Villanova University, 2003.

K. Murphy (2001), Bayes net matlab toolbox,www.cs.berkeley.edu/»murphyk/Bayes /bnt.html