PREDIKSI NILAI DENGAN METODE SPECTRAL CLUSTERING DAN CLUSTERWISE REGRESSION

Ahmad Yusuf[1*)], Handayani Tjandrasa[1)]
[1)]Teknik Informatika, Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
[*)]ahmad.yusuf.its@gmail.com

Abstrak. *Prediksi nilai adalah hal yang terus dikembangkan dalam penggalian data. Regresi linier merupakan metode dasar dalam memprediksi nilai berdasar variabel-variabel pada data. Salah satu hal yang mempengaruhi kualitas dari hasil regresi adalah persebaran data latih. Data latih terkadang membuat persamaan regresi kurang optimal. Hal ini dapat diantisipasi dengan mengelompokkan data terlebih dahulu kemudian membangun model regresi dari masing-masing kelompok. Pengelompokan data dilakukan dengan menggunakan algoritma Spectral Clustering, sedangkan model regresi dibangun dengan algoritma Clusterwise Regression. Hasil prediksi merupakan hasil perkalian keanggotaan fuzzy data uji dengan persamaan regresi pada masing-masing kelompok. Metode ini diujicobakan terhadap beberapa dataset yang bervariasi yang dibandingkan dengan metode regresi linear biasa. Ukuran pengujian yang digunakan adalah Root Mean Square Error yang menghitung kesalahan dari hasil prediksi. Semakin kecil nilai RMSE suatu metode maka metode tersebut semakin baik. Berdasar pada uji coba yang dilakukan, penggunaan metode yang diusulkan mampu memprediksi nilai dengan kesalahan sekitar 3 sampai 6 persen. Parameter jumlah cluster juga berpengaruh terhadap hasil prediksi yaitu berbanding terbalik dengan nilai RMSE.*

Kata Kunci: Clusterwise Regression, Pengelompokan, Penggalian Data, Prediksi, Regresi, Spectral Clustering.

Salah satu aplikasi data mining adalah prediksi nilai dari sejumlah variabel data tertentu. Model yang umumnya digunakan dalam prediksi nilai adalah Regresi Linier. Dengan regresi akan didapatkan nilai prediksi sesuai dengan variabel-varibel yang mempengaruhi pada data. Hanya saja permasalahannya kualitas regresi biasanya tergantung pada model training. Persebaran data yang tidak merata pada data training dapat menyebabkan model regresi yang dibentuk kurang akurat, sehingga prediksi yang dihasilkan kurang baik. Persamaan regresi dapat diperbaiki dengan mengelompokkan data dalam membangun model regresi. Dengan data yang lebih seragam diharapkan dapat membentuk model regresi yang lebih baik. Pemodelan regresi dengan pengelompokan data selanjutnya disebut dengan *Clusterwise Regression* [1,2].

Pada metode *Clusterwise Regression*, inisialisasi *cluster* dilakukan secara acak [1]. Hal ini dapat menyebabkan komputasi dalam pembentukan model regresi cukup besar untuk mencapai konvergen. Oleh karena itu, sebaiknya dilakukan inisialisasi *cluster* dalam pemodelan *Clusterwise Regression* selanjutnya. Salah satu cara dalam inisialisasi *cluster* adalah dengan metode *Clustering*.

Algoritma *Clustering* yang sangat umum digunakan adalah K-Means *Clustering*. Metode K-means mudah dalam implementasi serta memiliki waktu komputasi yang cukup cepat. Tetapi metode ini mempunyai kelemahan dalam menganalisis persebaran data serta bergantung pada inisialisasi centroid. K-means hanya melihat jarak data ke masing-masing centroid pada setiap *cluster* [3]. Salah satu metode *Clustering* lain yang pernah dikembangkan dalam memperbaiki akurasi regresi adalah *Spectral Clustering*. *Spectral Clustering* mengelompokkan data berdasarkan kesamaan dari tiap data [3].

Penelitian ini bertujuan untuk membangun model yang mampu melakukan prediksi nilai dengan metode *Clusterwise Regression* dan *Spectral Clustering*. Pada penelitian ini dilakukan pengoptimalan model prediksi dengan melakukan inisialisasi *cluster* dalam

pemodelan regresi dan melakukan pembobotan pada tahap regresi menggunakan metode *Clusterwise Regression*. Hasil algoritma ini nantinya dapat digunakan dalam memprediksi nilai dalam berbagai bidang misalnya pendidikan (prediksi nilai siswa, dsb), ekonomi (prediksi nilai saham, dsb), dan lain-lain.

Tabel 3 Dataset yang digunakan pada penelitian

No	Dataset	Jumlah *Record*	Jumlah Atribut	*Missing Value*
1	Nilai Praktikum Struktur Data Teknik Informatika ITS 2008/2009	180	6	Tidak Ada
2	Nilai Praktikum Struktur Data Teknik Informatika ITS 2010/2011	153	4	Tidak Ada
3	Data Auto MPG - UCI	398	8	Ada
4	Data Housing – UCI	506	14	Tidak Ada
5	Data Computer Hardware – UCI	209	8	Tidak Ada

Data

Data masukan pada penelitian ini adalah kumpulan data yang berasal dari berbagai sumber dimana variabel yang diprediksi bersifat kontinu. Pada penelitian ini digunakan 5 dataset, 2 dataset merupakan data nilai praktikum mahasiswa dan 3 sisanya merupakan dataset yang bersumber dari uci dataset repository [4]. Dataset yang digunakan dapat dilihat pada Tabel 1.

Dataset 1 dan 2 merupakan nilai praktikum struktur data yang berisi nilai modul-modul praktikum sebagai nilai prediktor dan satu nilai proyek akhir sebagai variabel respon. Dataset 3, 4, dan 5 merupakan dataset yang bersumber dari uci dataset repository yang bersifat data regresi.

Dataset Auto MPG (dataset 3) adalah data penggunaan bahan bakar yang dihitung dengan satuan mpg (miles per gallon) [4]. Atribut mpg merupakan nilai penggunaan bahan bakar yang merupakan variabel respon dari dataset tersebut. Tujuh atribut lainnya merupakan atribut prediktor yang mempengaruhi penggunaan bahan bakar yang meliputi silinder yang digunakan kendaraan, tenaga kuda dari kendaraan, dan sebagainya. Atribut nama mobil tidak digunakan dalam penelitian ini karena bernilai string yang berbeda-beda dari setiap *record* pada Dataset Auto MPG.

Dataset ke-4 adalah dataset Housing yang merupakan data nilai kepemilikan rumah di sebagian wilayah Boston, Amerika Serikat [4]. Dataset Housing terdiri dari 13 atribut kontinu dan 1 atribut yang bersifat diskrit. Variabel respon dari dataset ini adalah nilai median kepemilikan rumah pada suatu kota dalam satuan 1000 US Dollar yang ditunjukkan pada atribut MEDV. Variabel prediktornya merupakan atribut lainnya antara lain rata-rata kejahatan pada kota bersangkutan (CRIM), proporsi tanah residential (ZN), proporsi bisnis non-retail (INDUS), dan sebagainya.

Dataset Computer Hardware merupakan dataset ke-5 yang digunakan dalam uji coba pada penelitian ini. Dataset Computer Hardware merupakan data performa relatif dari cpu [4]. Dataset ini terdiri dari 9 atribut, dimana Atribut PRP merupakan performa relatif CPU yang digunakan sebagai variabel respon. Atribut lainnya selain Vendor Name dan Model Name digunakan sebagai variabel prediktor. Atribut Vendor Name dan Model Name tidak digunakan karena memiliki nilai string. Variabel prediktor yang digunakan meliputi cycle time mesin (MYCT), memory minimal (MMIN) dan maksimal (MMAX), dan sebagainya.

METODOLOGI

Metode prediksi menggunakan Spectral Clustering dan Clusterwise Regression terdiri dari dua tahap yaitu tahap latih dan tahap uji. Pada tahap latih terdiri dari tahap pra proses, *clustering*, serta tahap regresi. Tahap uji merupakan tahap prediksi saat model telah dibangun pada tahap latih.

Tahap Pra Proses

Pra proses yang dilakukan adalah penanganan missing value. Nilai-nilai yang kurang pada data akan diganti dengan rata-rata nilai pada atribut yang bersesuaian jika atribut tersebut merupakan numerik, dan akan diganti

dengan modus nilai atribut yang bersesuaian jika nominal.

Pada tahap pra proses juga dilakukan normalisasi data. Normalisasi data dilakukan untuk menyeragamkan interval data. Keseragaman data berpengaruh dalam perhitungan matriks-matriks pada data. Normalisasi yang dilakukan merupakan normalisasi kolom/atribut. Normalisasi dilakukan berdasar persamaan 1

$$\bar{x}_{i,j} = \frac{x_{i,j} - \min(x_j)}{\max(x_i) - \min(x_j)} \quad (1)$$

Dimana merupakan nilai normalisasi data pada baris i dan atribut j, xi,j merupakan data aslinya, dan min(xj) merupakan nilai minimal dari atribut j, sedangkan max(xj) merupakan nilai maksimalnya. Keluaran dari proses normalisasi ini berupa data yang memiliki interval 0 sampai 1.

Tahap Clustering

Clustering merupakan salah satu metode eksplorasi data yang digunakan dalam mencari pola yang ada pada suatu dataset. Pada umumnya pola tersebut dapat dilihat dari kesamaan sifat, karakteristik, atau ciri dari *record-record* pada dataset [5,6].

Salah satu metode *Clustering* adalah *Spectral Clustering*. Pengelompokkan *Spectral Clustering* dilakukan berdasarkan atas kesamaan antara setiap data. Kesamaan tersebut dilihat dari keterkaitan antara setiap data. Pada *Spectral Clustering* akan dibentuk sebuah graf dari data yang ada. Dimana verteks dari graf tersebut merupakan setiap *record* pada data. *Edge*-nya berupa hubungan antar data yang biasanya bernilai jarak dari dua *record* yang berhubungan [7].

Diagram alir dalam melakukan *Spectral Clustering* ditunjukkan pada Gambar 1. Langkah-langkah pengelompokan berdasar pada diagram alirnya adalah sebagai berikut [3,8]:

1. Kontruksi graf similaritas dari dataset training. verteks pada graf tersebut merupakan representasi dari setiap *record* pada data training. Bobot dari tiap *edge* merupakan jarak antara satu verteks dengan verteks lainnya. Perhitungan jarak antar verteks menggunakan persamaan jarak exponential yang tertulis pada persamaan 2 [3].

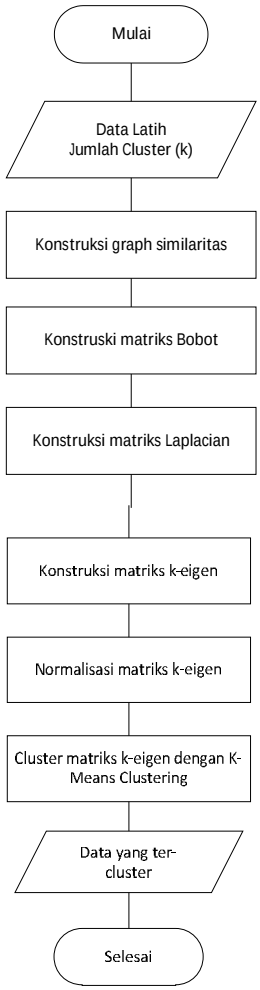
$$w_{ij} = \exp\left(\frac{-\|s_i - s_j\|^2}{2\sigma^2}\right) \quad (2)$$

Setelah itu, bobot dari setiap *edge* yang ada dibentuk menjadi matriks *weight*. Dengan begitu matriks *weight* merupakan representasi graf similaritas dari dataset.

2. Dari matriks *weight* dihitung derajat dari setiap verteks dengan menjumlahkan bobot dari *edge* yang terhubung pada verteks yang bersangkutan. Dari derajat verteks tersebut dapat dibentuk matriks degree yang merupakan matriks diagonal yang berisi bobot setiap verteks.
3. Dibentuk normalisasi matriks *Laplacian* dengan menggunakan matriks *weight* (W) dan matriks degree (D) yang telah dihitung sebelumnya. Perhitungan matriks *Laplacian* (L) dengan rumus pada persamaan (3) [9].

$$L = D - W \quad (3)$$

Gambar 1 Diagram Alir Algoritma *Spectral Clustering*



4. Dihitung k *eigenvector* pertama dari matriks *Laplacian*, dimana k merupakan parameter jumlah *cluster*. Maka terbentuklah matriks k-

eigen yang merupakan k *eigenvector* permata dari matriks *Laplacian*. Matriks k-eigen berukuran n x k, variabel n merupakan jumlah *record* pada data masukan.

5. Normalisasi data dengan matriks k-eigen sehingga akan terbentuk k kolom yang merepresentasikan setiap nilai normalisasi eigen pada setiap kolomnya.
6. Hasil dari data normalisasi kemudian di-*cluster* dengan K-Means *Clustering*. Data normalisasi mewakili masukan data latih. Data latih ke-i akan dimasukkan pada suatu *cluster* jika dan hanya jika data hasil normalisasi ke-i masuk pada *cluster* yang sama.

Tahap Regresi

Selanjutnya dari masing-masing *cluster* yang dihasilkan akan diproses dengan *Clusterwise Regression*. Pada tahap regresi masukan berupa data yang telah terkelompokkan dalam k *cluster*. Pada tahap tersebut akan dihasilkan k model regresi yang dapat digunakan pada tahap selanjutnya yaitu tahap uji.

Clusterwise Regression merupakan metode regresi yang menggunakan lebih dari satu persamaan regresi pada tahap latih. Data latih akan dikelompokkan terlebih dahulu sehingga masing-masing kelompok akan dibentuk persamaan regresi.

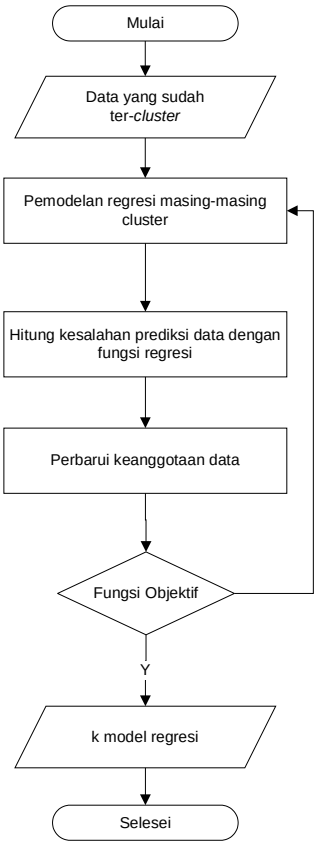
Langkah-langkah dari algoritma *Clusterwise Regression* berdasar diagram alir pada Gambar 2 adalah sebagai berikut [1,2]:

1. Data dibagi menjadi kelompok awal Q1, Q2, … , Qk
2. Masing-asing kelompok dibangun persamaan regresi linier.
3. Masing-masing data latih akan dihitung nilai kesalahan yang paling kecil dari persamaan regresi yang ada sesuai dengan persamaan 4.

$$Min(Z) = \sum_{k=1}^{K} \left\| X^k b_k - y^k \right\|^2 \quad (4)$$

4. Ubah keanggotaan kelompok pada masing-masing data berdasarkan dengan kesalahan prediksi yang minimal.
5. Ulangi langkah 2 sampai fungsi objektif yang ditentukan.

Clusterwise Regression akan memodelkan n persamaan regresi dimana n adalah jumlah *cluster* dengan meminimalisasi SSE. Pada setiap iterasi akan dimodelkan n persamaan regresi kemudian dihitung nilai kesalahan dan keanggotaan masing-masing data. Persamaan regresi, nilai kesalahan, dan keanggotaan akan diperbarui pada tiap iterasi hingga mencapai model yang optimal atau jumlah iterasi tertentu.



Gambar 2 Diagram Alir Algoritma Clusterwise Regression

Tahap Prediksi

Pada tahap uji masing-masing data akan dihitung nilai fuzzy keanggotaannya dengan masing-masing *cluster* yang ada. Nilai fuzzy keanggotaan (Zi) dari data (x) dihitung berdasarkan persamaan fuzzy k-nearest neighbor yang ditunjukkan pada persamaan 5 [10].

$$Z_i(x) = \frac{\sum_{j=1}^{k} Z_{i,j} \left(\frac{1}{\left| x - x_j \right|^{2/m-1}} \right)}{\sum_{j=1}^{k} \left(\frac{1}{\left| x - x_j \right|^{2/m-1}} \right)} \quad (5)$$

dimana k yaitu jumlah tetangga terdekat, Zi,j merupakan nilai keanggotaan data ke-j (xj) pada *cluster* i, dan m merupakan parameter bobot. Hasil prediksi merupakan pernjumlahan dari

perkalian antara bobot masing-masing *cluster* dan hasil prediksinya [11].

HASIL DAN PEMBAHASAN

Uji coba dilakukan dengan memprediksi dataset dengan metode yang diusulkan. Metode Cross Validation digunakan dalam pengujian model yang dihasilkan.

Tabel 4 Perbandingan RMSE Clusterwise Regression – Spectral Clustering *dengan parameter jumlah* cluster *yang bervariasi*

Dataset	Jumlah *Cluster*	RMSE *Clusterwise Regression – Spectral Clustering*
Nilai Praktikum Struktur Data Teknik Informatika ITS 2008/2009	1	4.428
	2	3.566
	3	2.526
	4	2.872
	5	2.879
Nilai Praktikum Struktur Data Teknik Informatika ITS 2010/2011	1	5.509
	2	3.828
	3	3.606
	4	3.280
	5	3.300
Data Auto MPG - UCI	1	3.873
	2	4.148
	3	3.623
	4	2.953
	5	2.595
Data Housing – UCI	1	5.587
	2	5.749
	3	4.175
	4	3.211
	5	5.054
Data Computer Hardware – UCI	1	59.228
	2	55.113
	3	55.745
	4	46.553
	5	50.625

Untuk mengukur performa yang dihasilkan, digunakan Root Mean Squared Error (RMSE) yang ditunjukkan dengan persamaan 6 [12].

$$RMSE = \frac{\sum_{i=1}^{n} \sqrt{(x_i - f_i)^2}}{n} \quad (6)$$

Dimana variabel n merupakan jumlah data, dan variabel x merupakan nilai prediksi yang dihasilkan model sedangkan variabel f merupakan nilai sebenarnya yang diinginkan.

Untuk setiap pasangan data masukan dan data keluaran pada setiap parameter jumlah *cluster* akan dihitung kesalahan prediksinya menggunakan RMSE. Semakin keci nilai RMSE pada kesalahan hasil prediksi semakin kecil, begitu juga sebaliknya.

Hasil uji coba merupakan hasil prediksi dari algoritma *Clusterwise Regression* dengan *Spectral Clustering* dengan beberapa nilai parameter jumlah *cluster*.

Metode *Clusterwise Regression- Spectral Clustering* menggunakan parameter jumlah *Clustering*. Pada ujicoba dilakukan perbandingan kesalahan RMSE dengan nilai *cluster* bervariasi yaitu satu, dua, tiga, empat, dan lima *cluster*. Hasil uji coba ditunjukkan pada Tabel 2.

Pada hasil uji coba terlihat bahwa RMSE terkecil diperoleh pada parameter jumlah *cluster* yang berbeda pada tiap dataset. Pada dataset nilai praktikum struktur data tahun 2008/2009, nilai RMSE terkecil dicapai pada parameter jumlah *cluster* sama dengan 5 yaitu sebesar 2.526. Rentang nilai mahasiswa dari satu sampai 100 sehingga nilai kesalahan relatifnya sebesar 2.53 persen. RMSE terkecil pada dataset ke 2 yaitu dataset nilai praktikum struktur data tahun 2010/2011 dicapai pada parameter jumlah *cluster* sama dengan 4 sebesar 3,28 atau kesalahan relatifnya 3,28 persen.

Pada data Auto MPG nilai RMSE terkecil diperoleh pada parameter 5 *cluster* yaitu 2,595. Rentang nilai mpg pada dataset Auto MPG 0 sampai 46.6 sehingga nilai kesalahan relatifnya 5,57 persen. Pada dataset Housing nilai RMSE terkecil diperoleh pada parameter jumlah *cluster* 4 yaitu 3,211 atau kesalahan relatifnya 6,42 persen (rentang 0-50). Dataset kelima yaitu Computer Hardware RMSE terkecil dicapai pada parameter jumlah *cluster* sama dengan 4 yaitu 46,553. Kesalahan relatif pada dataset Computer Hardware sebesar 4,05 persen (0 - 1150). Dari hasil analisis diatas nilai kesalahan relatif yang dicapai dari dataset yang digunakan sekitar 3 sampai 6 persen.

Dari perubahan jumlah *cluster* secara umum terlihat bahwa semakin besar jumlah *cluster* maka nilai kesalahan yang dihasilkan semakin kecil sampai pada titik optimal. Dari kelima dataset yang diujicobakan menunjukkan tren yang relatif sama yaitu nilai RMSE berbanding terbalik dengan jumlah *cluster*. Pada beberapa kasus terlihat nilai RMSE yang lebih besar parameter jumlah *cluster* tertentu dari jumlah *cluster* -1 dan jumlah *cluster* +1. Hal ini dapat terjadi karena hasil *Clustering* yang digunakan selanjutnya dalam proses prediksi kurang optimal untuk dimodelkan persamaan regresi.

Perbandingan nilai RMSE dari *Multiple Regression* (jumlah *cluster* = 1) dan *Clusterwise Regression – Spectral Clustering* dari hasil ujicoba terlihat bahwa nilai RMSE pada hasil prediksi dataset menggunakan metode yang diusulkan lebih kecil dibandingkan dengan RMSE pada hasil prediksi menggunakan *Multiple Regression*. Reduksi kesalahan RMSE yang dihasilkan sekitar 30 hingga 40 persen. Hal ini menunjukkan bahwa metode pada penelitian ini mampu mengoptimalkan performa dari prediksi nilai menggunakan regresi linier.

Pada hasil uji coba nilai kesalahan relatif paling besar (6 persen) didapatkan pada dataset Housing. Dataset Housing memiliki atribut prediktor dibandingkan dengan dataset lainnya yaitu 13 atribut. Hal ini dapat menyebabkan hasil kurang optimal karena memungkinkan adanya atribut-atribut yang tidak relevan.

Berdasarkan hasil uji coba, semakin banyak atribut prediktor yang digunakan maka nilai kesalahan relatifnya semakin besar secara umum. Dataset Auto MPG dan Computer Hardware memiliki 7 atribut prediktor dengan kesalahan relative sekitar 4-5 persen. Dataset nilai praktikum yang memiliki atribut lebih sedikit menunjukkan kesalahan relative yang lebih kecil yaitu 2 sampai 3 persen.

SIMPULAN

Prediksi nilai merupakan salah satu bagian dari penggalian data yang terus dikembangkan. Pada penelitian ini diusulkan algoritma *Clusterwise Regression* dengan *Spectral Clustering*. Metode usulan mampu melakukan prediksi nilai yang telah diujicobakan pada beberapa dataset dengan nilai kesalahan relatif 3 sampai 6 persen. Algoritma *Clusterwise Regression* dengan *Spectral Clustering* mampu mereduksi kesalahan RMSE dari hasi prediksi menggunakan *Multiple Regression* biasa sebesar 30 sampai 40 persen.

Pada metode usulan diperlukan masukan jumlah *cluster* sebagai parameter. Parameter jumlah *cluster* memberikan pengaruh pada hasil prediksi. Pada ujicoba yang dilakukan masing-masing dataset akan menghasilkan model optimal pada parameter jumlah *cluster* tertentu bergantung pada karakteristik data. Jumlah atribut juga mempengaruhi hasil prediksi menggunakan metode yang diusulkan. Pada ujicoba yang dilakukan semakin banyak atribut maka nilai kesalahannya semakin besar. Hal ini dikarenakan adanya kemungkinan atribut-atribut yang tidak relevan dengan variabel respon.

DAFTAR PUSTAKA

[1] Lau, Kin-nam., Leung, Pui-lam., Tse, Ka-kit. (1998). *A mathematical programming approach to Clusterwise Regression model and its extensions*. Elsevier European Journal of Operational Research.

[2] DeSarbo, Wayne S. (1988). *A Maximum Likelihood Methodology for Clusterwise Linear Regression*.

[3] Trivedi, S., A. Pardos, Z., & N. Sar, G. (2008). *Spectral Clustering in Educational Data Mining*.

[4] UCI Machine Learning Repository, <URL: http://archive.ics.uci.edu/ml/ >, diakses pada 1 April 2013.

[5] Tan, P.-N., Steinbach, M., & Kumar, V. (2006). I*ntroduction to Data Mining. Minnesota: Addison-Wesley*.

[6] Zhang, Z., Wu, X., & S. Yu, P. (2006). *Spectral Clustering for Multi-type Relational Data*.

[7] Y. Ng., A., I. Jordan, M., & Weiss, Y. (2002). *On Spectral Clustering : Analysis and Algorithm*.

[8] Von Luxburg, U. (2007). *A Tutorial on Spectral Clustering*. 17(4).

[9] Mohar, B. (1991). *The Laplacian Spectrum of Graphs, In Graph Theory*. 871-898.

[10] Keller, James M., Gray, Michael R., Givens, James A. Jr. (1985). A Fuzzy K-Nearest Neighbor Algorithm. IEEE Transactions on Systems, Mans, and Cybernetics.

[11] Luo, Zairen., Chou, Eddie Y. (2006). *Pavement Condition Prediction Using Clusterwise Regression*.
[12] Karagozog'lu, B., & Turkmen, N. (2007). *A software tool to facilitate design, assessment and evaluation of courses in an educational system.*