

Veri Akışı Diyagramları Tabanlı Veri Madenciliği Araçları ve Yazılım Geliştirme Ortamları

Yrd.Doç.Dr. T. Tugay Bilgin

Maltepe Üniversitesi, Bilgisayar Mühendisliği Bölümü
ttbilgin@maltepe.edu.tr

Özet: Bu çalışmada, veri akış diyagramları ve veri akışı tabanlı veri madenciliği süreçleri görselleştirilmesi açıklanmıştır. Üç farklı tür veri akışı tabanlı yazılım incelenmiş ve detaylı özellikleri karşılaştırılmıştır.

Abstract: In this study, data flow diagrams and data flow based mining process visualization are explained. Three different examples of data flow based software are reviewed and detailed specifications are compared.

Anahtar Kelimeler: Veri Madenciliği, Görselleştirme, Veri Akışı, Sınıflandırma, Kümeleme.

1. Giriş

Bilgi teknolojileri alanında son yıllarda yaşanan gelişmeler, veri ambarlarında çok büyük miktarda veri depolamaya, düzenlemeye ve gerektiğinde kullanmaya olanak sağlamaktadır. Finansal hizmetlerden telekomünikasyon alanına kadar yüzlerce farklı sektör daha hızlı ve etkin veri analizi konusunda yarışmaktadır [1].

Veri yığınlarının analizi için istatistiksel teknikler ve veritabanı yönetim araçları uzun yıllar boyunca yeterli başarı sağlamışlardır [2]. Son yıllarda çok büyük veri yığınları üzerinde çalışacak yeni nesil araçlar ve teknikler ortaya çıkmaktadır. Bunlar ile birlikte veri madenciliği olarak adlandırılan disiplinler arası bir uzmanlık alanı ortaya çıkmıştır.

Veri madenciliği, veri ambarlarında veya diğer bilgi depolarında tutulmakta olan büyük miktardaki verinin işlenerek içindeki değerli olabilecek bilginin ortaya çıkarılması sürecidir.

Görsel veri madenciliği ise görselleştirmeyi insan ile bilgisayar arasında bir iletişim kanalı olarak kullanarak yeni ve yorumlanabilir ürünler ortaya çıkarma sürecidir.

2. Veri Akışı Diyagramları ve Veri Akış Kavramı

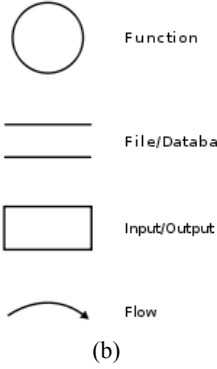
Veri akış diyagramı (data flow diagram – DFD), bir bilgi sistemi içinde verinin akış şeması şeklinde görsel olarak ifade edilmesidir. Veri akış diyagramları algoritma tasarımında kullanılan akış diyagramlarına benzemekle birlikte kullanım amaçları tamamen farklıdır. Algoritma akış diyagramlarında görselleştirilen program akışıdır, veri akış diyagramlarında ise görselleştirilen veri akımıdır.

Veri akış diyagramları aynı zamanda veri işleme sistemlerinin görselleştirilmesi için de kullanılmaktadır (structured design).

DFD'ler bir sistemin nasıl bölümlere ayrıldığı ve bu bölümler arasındaki veri akışının hangi yönde olduğunu göstermek üzere tasarlanmıştır. DFD kavramı ilk olarak "Structured Design" [3] adlı makalesinde Larry Constantine tarafından kullanılmıştır.



(a)



Şekil 1. (a) DFD örneği, (b) Yourdon Notasyonuna göre DFD nesnelere [4].

Edward Yourdon “Just Enough Structured Analysis” adlı kitabında [4] DFD oluşturma yaklaşımlarını iki temel gruba ayırmıştır:

- Yukarıdan aşağı yaklaşımı (top-down)
- Olay bölümlenme yaklaşımı (event partitioning)

Şekil-1.’de Yourdon notasyonuna göre basit bir akış diyagramı görülmektedir.

DFD’ler Yapısal Sistem Analizi ve Dizayn metodunun (Structured Systems Analysis and Design Method SSADM) üç temel yaklaşımından biridir.

3. Veri Akışı Tabanlı Veri Madenciliği Yazılımları

Bu çalışmada veri akışı (Data Flow – DF) kullanılan 3 farklı veri madenciliği yazılımı incelenmiştir. Bu yazılımların tümü açık kaynak kodlu geliştirme ortamları kullanılarak geliştirilmiş olup hepsi Linux, Mac ve Win32 sistemlerde çalışabilmektedirler.

İncelenen yazılımlardan Knime ve Keel java tabanlı, Orange ise C++ ve Python tabanlı geliştirilmiştir.

4. KEEL

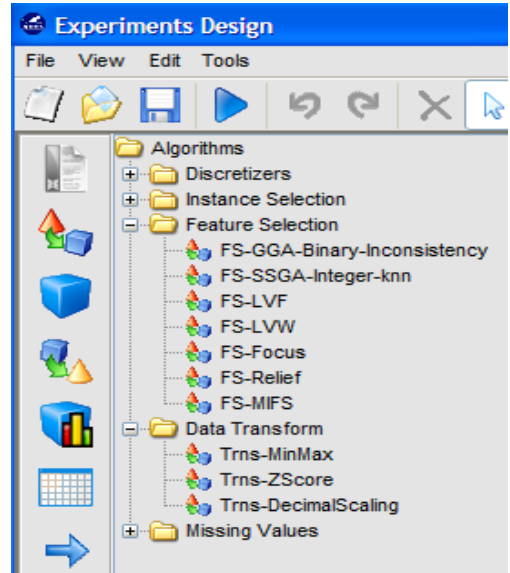
KEEL aracı [5] İspanya Ulusal Bilim Projeleri Kurumunun desteği ile Granada Üniversitesi tarafından geliştirilmektedir. Tamamı Java dili ile kodlanmıştır. Yalnızca bir veri madenciliği yazılımı değil aynı zamanda veri madenciliği eğitimi alanında eğitsel demo araçlarına da sahiptir.

Regresyon, sınıflandırma, kümeleme gibi klasik veri madenciliği algoritmalarının yanı sıra yapay zeka tabanlı algoritmalar ile genetik ve yapay sinir ağlarından oluşan hibrid algoritmaları da kullanmaya olanak sağlamaktadır.

4.1 Veri Kaynakları

KEEL, çok farklı biçimlerde veriyi import etme özelliğine sahiptir. Bu türler arasında CSV, TXT, PRN, XLS, DIF, XML ve HTML bulunmaktadır. Ayrıca SQL veritabanlarından ve WEKA adlı veri madenciliği yazılımından import işlemini desteklemektedir.

4.2. Veri Önleme

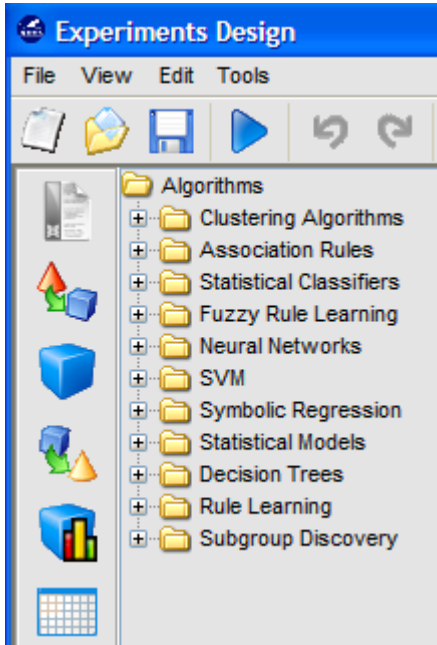


Şekil 2. KEEL programında Veri önleme algoritmaları seçim paneli.

Veri ayrıklaştırıcı (discretizer) algortimalarının ve özellik seçme (feature selection) algortimalarının birçoğunu desteklemekte, ayrıca MinMax, Z-Score ve Decimal Scaling gibi Veri Dönüşüm (transformasyon) araçlarını da içermektedir. Ayrıca eksik değerler (missing values) içeren veri setleri için de çeşitli araçlar barındırmaktadır.

4.3. Veri Madenciliği Algoritmaları

KEEL, kümeleme ve sınıflandırma gibi klasik veri madenciliği algoritmaları açısından zengin değildir. Bunların yerine Fuzzy sınıflandırıcılar, Yapay zeka tabanlı sınıflandırma ve Kural tabanlı kümeleme algoritmalarının birçok çeşidini içermektedir.



Şekil 3. KEEL programında veri madenciliği algoritmaları seçim paneli.

4.4. Veri Akışı tasarımı

Veri akışı tasarımı, sürükle-bırak yöntemi yerine tıkla-ekle yöntemini kullanmaktadır. Şekil-3'deki gibi bir liste kutusunda yöntem seçimi yapılmakta ve seçili yöntem tasarım kanvasının tıklanan noktasına eklenmektedir. Ayrıca

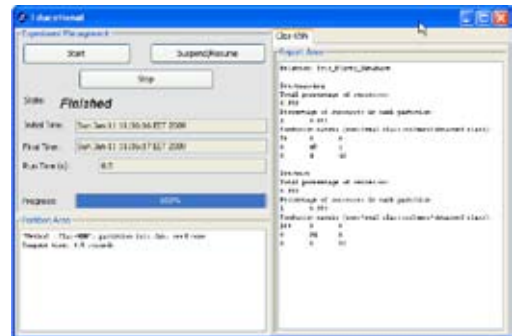
veri akış yolunun tanımlanması da şekil-3'de sol panelde bulunan ok simgesi ile iki nesne birleştirilerek gerçekleştirilmektedir. Şekil-4'de KEEL programı ile K-en yakın komşu (K-Nearest Neighbor) algoritmasının gerçekleştirilmesi görülmektedir.



Şekil 4. KEEL programında veri akış yolu tasarımı ve kanvas yapısı.

4.5. Görselleştirme yöntemleri

Karar verme sürecinin en önemli aşaması olan veri görselleştirme KEEL programının en zayıf yönlerinden biri olarak gözlemlenmiştir. Elde edilen sonuçlar yalnızca tabloid yapıda görselleştirilmektedir (Şekil-5).



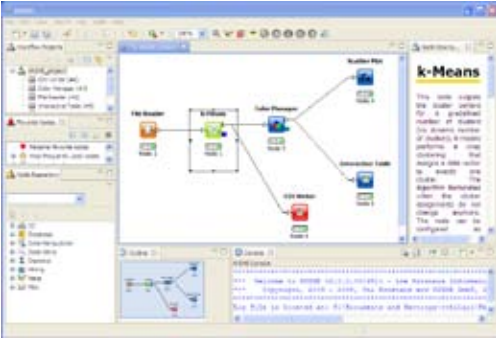
Şekil 5. KEEL programında sonuçların raporlaştırılması.

5. KNIME

Konstanz Information Miner (KNIME) [6] adlı araç Konstanz Üniversitesi görsel veri madenciliği araştırma grubu tarafından Eclipse Rich Client Platform üzerinde geliştirilmiştir.

KNIME, Eclipse tabanlı olmasının sağladığı avantaj sayesinde modüler ve görsel veri akışı sistemi geliştirme ortamı sunmasının yanı sıra eğitim ve araştırma amaçlı ortak çalışma ortamı da sunmaktadır.

KNIME aracı genişletilebilir özellikleri ile ön plana çıkmaktadır. Bu çalışmada incelenen yazılımlar içerisinde kullanıcılara bir yazılım geliştirme kiti (software development kit) sunarak kullanıcıların kendi modüllerini yazabilmelerini sağlayan tek uygulamadır.



Şekil 6. KNIME programı ekran görüntüsü.

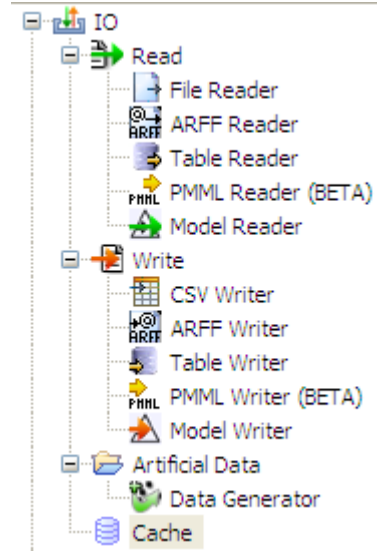
Bu araç Eclipse platformunun desteklediği tüm işletim sistemlerinde kurulum gerektirmeden çalışabilmektedir. Kendi içinde bir JRE barındırmaktadır. Bu sayede işletim sisteminde Java kurulu olmasına gerek yoktur.

5.1 Veri Kaynakları

KNIME programı metin dosyadan (TXT) veya ARFF, TABLE formatında veri alabilmektedir (Şekil-6).

Veri ayrıklaştırıcı (discretizer) algoritmalarının ve özellik seçme (feature selection) algo-

ritmalarının birçoğunu desteklemekte, ayrıca MinMax, Z-Score ve Decimal Scaling gibi Veri Dönüşüm (transformasyon) araçlarını da içermektedir. Ayrıca eksik değerler (missing values) içeren veri setleri için de çeşitli araçlar barındırmaktadır.



Şekil 7. KNIME programında veri kaynağı ekleme nesneleri.

Import edilen verilerin ne kadarını bellekte tutabileceğini, ne kadarını diskte tutabileceğini ayarlamaya da olanak sağlamaktadır. Bu özellik büyük veritabanlarında çalışırken bellek bitmesi sorunu ile karşılaşma ihtimalini düşürmektedir.

Ayrıca veritabanı sunuculardan SQL sorgulama ile veri alma işlemini ve veri madenciliği ve istatistik uygulamaları arasında veri transferine olanak sağlayan PMML (Predictive Model Markup Language) adlı XML tabanlı dilde veri import işlemini de desteklemektedir.

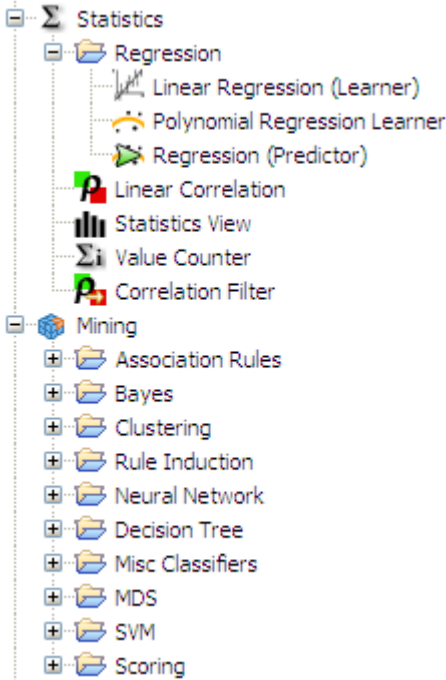
KNIME, veri import işleminin yanı sıra aynı biçimde veri export etmeyi sağlayan Data Write bileşenlerine de sahiptir (Şekil-7).

5.2. Veri Önışleme

KNIME, önışleme adı altında kendi başına bir bölüm içermez fakat Mining başlığı altında kullanılan algoritmaların bazıları veri önışleme amaçlı da kullanılabilir.

5.3. Veri Madenciliği Algoritmaları

Bu yazılımda sık kullanılan veri madenciliği yöntemlerinin tamamına yakını mevcuttur. Bunlar arasında Destek vektör makinaları, Bayes ve Multi dimensional Scaling (MDS) gibi yöntemler de bulunmaktadır.



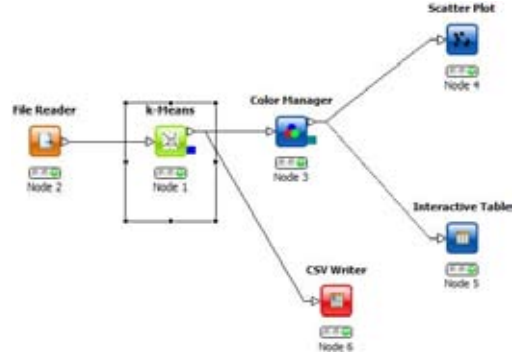
Şekil 8. KNIME programında veri madenciliği algoritmaları seçim paneli.

KNIME, Regresyon, korelasyon ve korealsyon filtresi gibi istatistik tabanlı yöntemlerin de veri akış tasarımında kullanılmasına olanak sağlamaktadır.

5.4. Veri Akış tasarımı

Programda “node repository” bölümünde listelenen bütün nesnelere sürükle-bırak yöntemiyle

kanvas üzerine kolayca yerleştirilebilmektedir. Her bir düğüm (node) arasındaki bağlantı fare ile seçilen düğümden hedef düğüme sürüklenme ile gerçekleşmektedir.



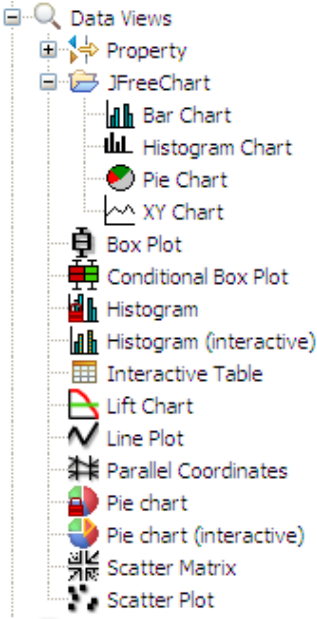
Şekil 9. KNIME programı veri akış diyagramı.

Veri akış diyagramının çalışma yapısı her bir düğümün tek-teke çalıştırılmasına dayanır. Her bir düğüm çalıştırıldığında işlem hatasız tamamlanmış ise düğümün alt bölümünde yeşil ışık yanacaktır. Bu durumda bir sonraki düğümün konfigürasyonu yapılır ve çalıştırılabilir. Bir düğüm, kendinden önceki bir düğüm yeşil ışık durumunda değilse çalıştırılmaz. Veri akış diyagramı Şekil-9’da görülmektedir.

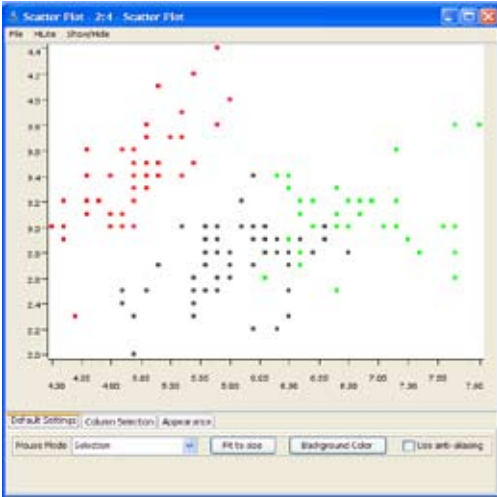
5.5. Görselleştirme yöntemleri

KNIME, incelenen yazılımlar içerisinde en zengin görselleştirme araçları sunan yazılımdır. Scatter Plot, Parallel Coordinates, Box Plot ve Histogram gibi bir çok görselleştirme aracının yanı sıra JFreeChart bileşeni sayesinde çok daha detaylı Java tabanlı görselleştirme düğümleri içermektedir (Şekil-10).

Şekil-9’da KNIME ile Iris veri seti üzerinde 3 küme için K-means algoritmasının gerçekleştirilmesi ile elde edilen saçılım grafiği görülmektedir. Bu veri akış diyagramının sonunda Scatter Plot düğümü eklenerek sonuç görselleştirilmiştir (Şekil-11).



Şekil 10. KNIME programında görselleştirme araçları.



Şekil 11. KNIME aracında Iris veri seti üzerinde K-means algoritmasının çalıştırılması ile elde edilen saçılım grafiği.

Şekil-12'de ise yine aynı sonuç tablo olarak görselleştirilmiştir. Her satırın başında ilgili satırın hangi kümeye dahil edildiğini renkli olarak göstermektedir.

The image shows a 'Table View' window in KNIME displaying the output of the K-means algorithm. The table has columns for 'Row ID', 'Sepal.Length', 'Petal.Length', 'Sepal.Width', 'Petal.Width', 'Species', and 'Cluster'. The data is grouped into three clusters: Cluster_0 (red), Cluster_1 (green), and Cluster_2 (black).

Row ID	Sepal.Length	Petal.Length	Sepal.Width	Petal.Width	Species	Cluster
Row#1	5.1	3.5	1.6	0.4	iris-setosa	Cluster_0
Row#2	5.1	3.8	1.9	0.4	iris-setosa	Cluster_0
Row#3	4.9	3.7	1.4	0.3	iris-setosa	Cluster_0
Row#4	5.1	3.8	1.6	0.2	iris-setosa	Cluster_0
Row#5	4.8	3.2	1.4	0.2	iris-setosa	Cluster_0
Row#6	5.3	3.7	1.5	0.2	iris-setosa	Cluster_0
Row#7	5	3.3	1.4	0.2	iris-setosa	Cluster_0
Row#8	7	3.2	4.7	1.4	iris-versicol	Cluster_0
Row#9	6.4	3.2	4.5	1.5	iris-versicol	Cluster_1
Row#10	6.9	3.1	5.9	1.5	iris-versicol	Cluster_0
Row#11	5.5	2.3	4	1.3	iris-versicol	Cluster_1
Row#12	6.5	2.8	4.6	1.5	iris-versicol	Cluster_1
Row#13	6.7	2.8	4.5	1.3	iris-versicol	Cluster_1
Row#14	6.3	3.3	4.7	1.8	iris-versicol	Cluster_1
Row#15	4.9	2.4	3.3	1	iris-versicol	Cluster_1
Row#16	6.6	2.9	4.6	1.3	iris-versicol	Cluster_1
Row#17	5.2	2.7	3.9	1.4	iris-versicol	Cluster_1
Row#18	5	2.1	3.5	1	iris-versicol	Cluster_1
Row#19	5.9	3	4.2	1.5	iris-versicol	Cluster_1
Row#20	6	2.2	4	1	iris-versicol	Cluster_1
Row#21	6.1	2.9	4.7	1.4	iris-versicol	Cluster_1
Row#22	6.6	2.9	3.6	1.3	iris-versicol	Cluster_1
Row#23	6.7	3.1	4.4	1.4	iris-versicol	Cluster_1
Row#24	6.6	3	4.5	1.5	iris-versicol	Cluster_1
Row#25	5.8	2.7	4.1	1	iris-versicol	Cluster_1

Şekil 12. KNIME aracında Iris veri seti üzerinde K-means algoritmasının çalıştırılması ile elde edilen sonuçlar tablosu.

6. ORANGE

ORANGE yazılımı [7] Slovenya Ljubljana Üniversitesi Bilgisayar ve Enformatik Bilimleri bölümü yapay zeka araştırmaları ekibi tarafından geliştirilmiştir [8].

Program C++ dili ile, arayüzler ve grafik ortam ise Qt3 kütüphanesi ve Python kullanılarak geliştirilmiştir. Windows kurulumu için Qt3 kütüphanesi ve python ortamını kurulum sırasınca otomatik olarak kurmaktadır.

KNIME yazılımına göre daha zayıf bir görsel veri akış yapısına sahiptir. Arayüz tasarımında da ergonomik açıdan bazı sorunlar bulunmaktadır. Buna rağmen görevini sorunsuz yerine getirmektedir.

6.1 Veri Kaynakları

Program yalnızca metin dosyadan (TXT, TAB) veri import işlemine imkan sağlamaktadır.

6.2. Veri Önileme

Önileme için bir bölüm bulunmamasıyla birlikte kullanılan algoritmaların bazıları veri önileme amaçlı da kullanılabilir.

6.3. Veri Madenciliği Algoritmaları

Hiyerarşik algoritmalar, MDS, Partitioning tabanlı algoritmalar, Prediction tabanlı algoritmalar, SVM, C4.5, K-NN gibi algoritmalar mevcuttur.



Şekil 13. ORANGE programında araç kutusu.

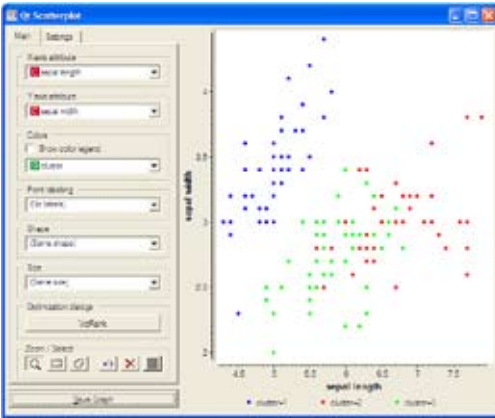
5.4. Veri Akış tasarımı

Programda araç kutusunda listelenen bütün nesnelere tıkladığında kanvas üzerine eklenmektedir. Düğümler arasında veri akışı bir düğüm üzerinden fareyi diğer düğüme sürükleyerek gerçekleştirilmektedir.



Şekil 14. ORANGE programı veri akış diyagramı.

Tüm düğümlerin konfigürasyonu yapıldığında veri akış sistemi otomatik olarak çalışmaktadır. Veri akış diyagramının sonuna yerleştirilen scatterplot çift tıklanarak saçılım grafiği elde edilmiştir.



Şekil 15. ORANGE aracımda Iris veri seti üzerinde K-means algoritmasının çalıştırılması ile elde edilen saçılım grafiği.

Şekil-15'teki grafikte, Iris veri seti üzerinde K-means algoritmasının 3 küme için çalıştırılması ile elde edilen kümeler görülmektedir.

5. Sonuçlar

Veri madenciliği, bilginin her geçen gün daha da hızla çoğaldığı dünyada en güncel araştırma alanlarından biridir. Yakın zamana kadar bu alanda yalnızca bazı firmalar tarafından üretilen pahalı analiz yazılımları bulunmaktaydı. Bu çalışmada, veri madenciliğinde daha kolay ve etkin sonuçlar almayı hedefleyen yeni nesil açık kaynak kodlu yazılımlar incelenmiştir.

Veri akış diyagramlarını temel alan veri madenciliği yazılımları kullanıcıları karmaşık kod kümeleri arasında boğulmaktan kurtararak bilgi keşfi sürecini kısaltmayı hedeflemektedir.

İncelenen yazılımlar içerisindeki KNIME ticari yazılımlarda dahi bulunmayan özellikleriyle veri madenciliği alanında çalışan araştırmacılar için en iyi ortamı sunmaktadır.

Kaynaklar

[1] Bilgin, T.T., "Çok Boyutlu Uzayda Görsel Veri Madenciliği İçin Üç Yeni Çatı Tasarımı Ve Uygulamaları", Doktora Tezi, 2007.

[2] Fayyad, U.M, Piatesky-Shapiro, G., and Smyth P.; "From Data Mining to Knowledge Discovery". An Overview. Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, USA (1996) 1-34.

[3] W. Stevens, G. Myers, L. Constantine, "Structured Design", IBM Systems Journal, 13 (2), 115-139, 1974.

[4] Yourdon, Edward. Just Enough Structured Analysis. <http://www.yourdon.com/jesa/jesa.php>, Chapter 19.

[5] KEEL indirme linki: <http://sci2s.ugr.es/keel/download.php>

[6] KNIME indirme linki:
<http://www.knime.org/downloads/knime>

[7] ORANGE indirme linki:
<http://magix.fri.uni-lj.si/orange/downloads.asp>

[8] Demsar J, Zupan B, Leban G (2004) Orange: From Experimental Machine Learning to Interactive Data Mining, White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.