ÇOK AMAÇLI GENETİK BULANIK SINIFLAMA İLE SALDIRI TESPİT SİSTEMİ

Sibel Tarıyan Özyer[(1)] **Tansel Özyer**[(2)]
Çankaya Üniversitesi, Bilgisayar Mühendisliği Bölümü[(1)]
TOBB ETÜ Bilgisayar Mühendisliği Bölümü[(2)]
tariyan@cankaya.edu.tr, ozyer@etu.edu.tr

ABSTRACT

The paper includes an intrusion detection system (IIDS) that utilizes data mining techniqes: Classification and association rules mining for predicting different behaviours in network traffic. To achieve this, A boosting genetic fuzzy classifier is proposed. It uses class besed fuzzy association rules to classify instances. Each fuzzy rule is weighted. Fuzzy rules are extracted each time genetic algorithm runs. Near optimal solution at the end of convergence gives the best rule found so far to be involved in fuzzy rle base. KDD Cup 99 data is split as training and test data. Classifier is trained with training data and then tested against test data nad results are given at the experiments.

Key words: Data mining, classification, fuzzy association rule, genetic algorithm, boosting.

1. GİRİŞ

Bu çalışma sınıflandırma ve birliktelik kurallarını içeren veri madenciliği tekniğini kullanarak saldırı tespit sistemi kurmayı hedeflemektedir. Hedeflenen sistem, ağlarla birleştirilmiş olan bilgisayarların bulunduğu ortamda farklı davranışları tahmin etmektedir. Veri madenciliği otomatik olarak ilk bakışta dışarıdan anlaşılamayan yapısal bilginin veri tabanlarından elde edilmesidir. Birliktelik kurallarının amacı veri özellikleri arasında birlikteliklerin üretilmesine dönüktür. Her bir birliktelik kuralının destek ve güvenirlik değeri bulunmaktadır. Bu değerler ilgili eşik değerlerinden fazla olmak zorundadır. Sınıflandırma bilinmeyen objeleri doğru sınıflandırabilmek için mantıksal tanımlamaları vermektedir. Sınıflandırma için üç adet gereksinim bulunmaktadır: Basitlik, tamlık ve verimlilik.

İzinsiz saldırı sistemdeki mevcut bütünlüğe, kaynakların bulunabilirliğine, gizliliğine zarar verici eylemler kümesidir[14]. Denning saldırı mekanizmasını sisteme girilen girdi ve gözlemlenebilen mevcut saldırılarla özetlemektedir [6]. IIDS kullanıcının bilgisayar sistemine giriş davranışlarını belirli kuralları uygulayarak sınırlandırır. Bu kurallar olası saldırı senaryolarını kurgulayan sistemi sömüren etkenleri bulan uzman bilgisine sahip uzmanlar tarafından belirlenmektedir. Sistem kullanıcılar tarafından yapılan tüm saldırıları tanımlar; ardından eylemleri durdurur ya da saldırıyı durdurmak amaçlı tavsiyelerde bulunur.

Saldiri tespit sisteminde amacı dışında kullanma ve anomali tespiti gibi iki farklı yaklaşım yer almaktadır[27]:

Amacı dışında kullanım bilinen örüntülere bağlı kalarak kötü niyetli aktiviteleri belirler. IDS IDS yapanlar kötü niyetli aktivitelerin en güncel hallerini yükleyerek oluşabilecek zararlara karşı mümkün olduğunca korumaya çalışırlar. Anomali tespitinde ise, normal ya da ağın temel çizgisindeki davranışlar(Yük, arıza, protokol, ve tipik paket büyüklükleri) ile anormal durumlar gözlenir.

Saldırı tespit problemlerinin çözümü için birçok yaklaşım bulunmaktadır.

Lee et al[21] birliktelik kurallarını ve bölüm sıklığı tekniklerini sistemi denetleme ve saldırı tespiti modelini kurmak için kullanmışlardır. Öznitelikler ile kısıtların oluşturulması ilgili örünütlerin bulunması ve seviye bazında yinelemeli olarak seviye bazlı madencilik işleminden geçirilerek düşük sıklıkta örüntülerin yarı otomatik bir şekilde saptanmıştır. NIDES[20] sistemi istatistiksel metodları kullanarak konu ve profil aktivitelerini tutarak anomali tespiti yapmaya çalışmaktadır. Aktivite yoğunluğu, denetlenen kayıt
dağılımı, kategoriksel ve sıralı veriler üzerinde istatistiksel ölçümler yapılmaktadır. Yapay sinir ağları kullanılarak soyut komutlar sıralı birim bilgi
cinsinden belirtilir. Verilen *k* adet komut ile sonraki komut tahmin edilir. Bunun için sinir ağları eğitilir. Kullanıcı profilleri belirlenir. Test adımında beklenen davranışlardan sapma durumlarında anomaly oluşmuştur diye ifade edilir[7,23]. Kısa dizilerden oluşam system çağrıları tahmin işlevini yürütür. Hamming uzaklık ile eşikten sapma miktarı belirlenir [12, 3]. Doğal bağışılık sistemi dağıtık yapıda önerilen başka bir metottur. Kendinden ve kendinden olmayan davranışlar dağıtık
pozitif ve negative belirleyiciler kullanılarak birbirinden ayırt edilmektedir[10]. Çok etmenli mimari kullanılarak da bağımsız varlıklar, otonom etmenler kollektif çalışarak saldırı tespit edebilirler[2]. Bir başka uygulamada etmenler genetik programlama üzerine inşa edilmişlerdir[4]. Genetik programlamanın öğrenme gücüne dayandırarak etmen arası sıralama sonucu en yüksek

skoru alan etmenler saldırı tespit için seçilirler. Bunun dışında, kümeleme teknikleri kullanılarak etiketlenmemiş veriler üzerinde anomali tespiti yapılır[22, 26].
Evrimsel bulanık sınıflandırıcılar da saldırı tespit sistemlerinin çözümü için kullanılmışlardır[9]. Sistem eğitilmesi için log verileri kullanılır. Bu sayede kurallar elde edilir. Elde edilen kurallar normal ve normal olmayan davranışlara ilişkin kurallardır.
İlgili makalenin amacı çok amaçlı evrimsel ve bulanık mantık yöntemlerini kullanarak normal ve normal olmayan kurallar için kurallar oluşturmak; ilgili kuralları kullanarak mevcut verileri sınıflandırmaktır.

Bu çalışma ilgili ağ trafik davranışlarını farklı sınıflara göre etiketlendirir. Birliktelik kurallarının karmaşık olmaması ve 2-6 arasında olması düşünülmüştür. Beş farklı sınıf bilgisi bulunmaktadır: Normal, PRB-probe, DOS-denial of service, U2R-user to root and R2L- remote to local [28]. Güçlendirilmiş sınıflandırma algoritması ile 2 adet alt uygunluk fonksiyonu birleştirilerek en iyi skoru alan kurallar kural veri tabanına eklenmektedir. Genetik algoritmayı her çalıştırışta objelerin dağılımı farklı olmaktadır. Güçlendirilmiş sınıflandırıcı bir önceki aşamada sınıflandıramadığı objelere daha fazla önem vermektedir. Bu da her obje ile ilişkilendirilen ağırlık bilgisiyle kontrol edilmektedir. Stem sabit diske bağlı olarak ölçelenebilir bir yapıya sahiptir. Tüm veriler veri tabanında saklanmaktadır.

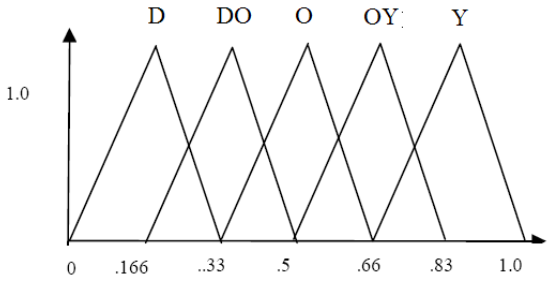
2. TEMEL ÇALIŞMALAR

2.1 Bulanık Mantık

Bulanık mantıkta bulanık kümeler bilinen klasik küme teorisinden farklı olarak küme elemanlarının sadece ve sadece bir kümenin elemanı olmasını değil; bütün kümelere farklı derecelerde eleman olması durumu da belirtilebilmektedir[17]. Buna göre her hangi bir X tanım kümesindeki bir x elemanı için belirtilen A bulanık kümesinde elemanların değer, üyelik değer ikilisi şöyle gösterilir:

$$A = \{(x, \mu_A(x)) \mid x \in X\}$$

Bulanık mantık kümeleri farklı şekillerde parametrik olarak gösterilebilirler.



Şekil 1 Beşe bölünmüş bulanık uzay (D: Düşük, DO: Düşük Orta, O:Orta, OY: Orta Yüksek, Y: Yüksek)

IF x_1 is Aq_1 and x_2 is Aq_2 and … and x_n is Aq_n THEN class is c_q kuralında

Bulanık mantık ile herhangi bir x= (x_1, x_2…,x_n) objesi her boyutta öznitelik sayısal değer içermektedir. Her bir Aq_i bulanık küme ve üyelik derecesi μ_i ise, koşul kısmı ilgili kuralın gücünü gösterir. $\mu_{Aq} = \min(\mu_1, \mu_2, ..., \mu_n)$.
Bulanık mantık saldırı tespit için iki nedenle uygundur: 1) Saldırı tespit verisinde sayısal değerler yer almaktadır. 2) Bulanık mantık ile kesin çizgilerle sayısal verileri sınıflandırmak yerine bir çok nitelikten oluşmuş verinin pürüzsüz bir şekilde sınıflandırmasına olanak sağlar[3].

Birliktelik kuralları öznitelikler içindeki değerler arasında ilişkilerin tanımanması için kullanılır. En bilinen örneği market satış verilerinde elde edilen bira, bebek bezi arasında ilişkidir. Buna göre bira alan müşteriler aynı zamanda bebek bezi de almaktadırlar. Bu durum pazarlama stratejisi olarak belirlenip market içinde ilgili ürünlerin yakın yerleştirilmesi düşünülmüştür. Bulunan kural, bira→bebek bezi şeklinde gösterilir.
Birliktelik kuralında destek ve güvenirlik parametreleri kullanılır. Uygulama tarafından verilmiş olan minimum destek ve güvenirlik parametreleri eşik değer olarak kullanılarak verilern veri kümesi içindeki anlamlı birliktelikler ve sonrasında kurallar yinelemeli bir işlem ile elde edilir.
Aq→Cq şeklinde kuralın solunda öznitelik değerleri ve sağında ilgili sınıf bilgisi yer almaktadır. Sol bir kural için bulanık mantık kullanarak birliktelik kuralları elde edilebilir[15, 13].

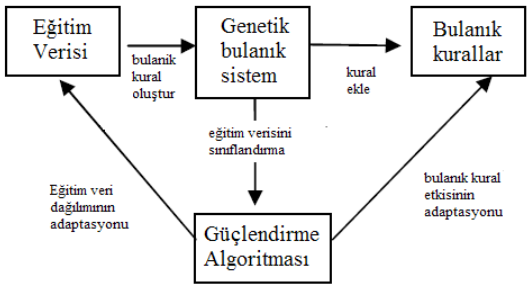
$$destek(Aq \rightarrow C) = \sum_{p \in classCq} \mu_{Aq}(x_p) / m$$

$$guvenirlik(Aq \rightarrow C) = \sum_{p \in classCq} \mu_{Aq}(x_p) \Big/ \sum_{p=1..m} \mu_{Aq}(x_p)$$

Elde edilecek olan kurallar ile verileri sınıflandırma amaçlanmaktadır. Buna göre, veri kümesindeki verilerden kurallar elde edilip sınıflandırma amacıyla kullanılacaktır. Elde edilen kurallardan ilk *n* tanesi kullanılarak karar mekanizmasında arama uzayı mümkün olduğunca küçültülecektir.

2.3 Güçlendirilmiş Evrimsel Bulanık Sınıflandırıcı

Bulanık bilgi tabanını kullanan evrimsel algoritmalar konusunda çalışmalar vardır[5]. Evrimsel algoritmaların amacı ya bulanık sistem parametrelerinin iyileşmesine yöneliktir ya da bulanık bilgi tabanı tasarımının otomatikleşmesine yardımcı olmaktır. Güçlendirilmiş sınıflandırma yinelemeli kural öğrenme ile her adımda Önceki adımlarda verilerden öğrenilmeyen hipotezleri öğrenerek kendini geliştirir[11]. Artarak öğrenme yaklaşımı benimsemiştir. Fazla sayıda hipotezin birleşiminden oluşmuştur. Evrimsel algoritmalarda sınıflandırma için bulanık mantık ile oluşturulan kuralların eniyilemesine ilişkin uygulamalar da mevcuttur[24, 26].



Şekil 2 Güçlendirilmiş genetik bulanık sınıflandırıcı

Şekil 2' de güçlendirilmiş genetik bulanık sınıflandırıcı mimarisi gösterilmektedir. Her bir yineleme adımında, mevcut veri dağılımı önceki adımlarda elde edilen hipotezlerle yapılan sınıflandırma sonucu yapılan yanlış tahminler sonucu değişecektir. Eldeki mevcut veri dağılımı baz alınarak genetik algoritma ile bulanık kurallar elde edilecektir. Elde edilenlerin en iyi sınıflayanı Daha önce elde edilen bulanık kuralların olduğu veri tabanına aktarılacaktır.
Genetik algoritmada çok boyutlu eniyilemede amacı eldeki kromozomu arama uzayı koordinatlarına geri döndürerek en iyi ya da en iyiye yakın çözümleri elde etmektir.
Genetik algoritma başlangıç populasyonu ile başlar. Populasyonda kromozomlar bulunmaktadır. Her kromozom bir tane çözüm sunmaktadır. Önerilen sistemde her bir kromozom içerisinde bir adet bulanık kural bilgisi yer almaktadır. Populasyonda her kromozomda başlangıç çözümleri yer almaktadır. Her adımda populasyon yeniden birleşme(Seçim, çapraz geçiş ve mutasyon) geçirirler. Belirli sayıda yineleme sonucunda evrimleşen kromozomlar eniyiye yakın bir değere yakınsarlar.

IF x1 is Aq1 and x2 is Aq2 and … and xn is Aqn THEN class is cq şeklinde bir bir kural için kromosome bilgisi şöyledir:

Gene1		Gene2		..		Gene k		..	Gene n	
x_1	Aq_1	X_2	Aq_2			X_k	*		X_n	Aq_n

Şekil 3 Kromozom Gösterimi

Şekil 3'teki gösterimde '*' sembolü ilgili özniteliğin dikkate alınmadığını belirtmektedir.
Fitness fonksiyonu iki adet alt fonksiyon içermektedir. Her ikisi de birer amaçtır.

Veri kümesi içinde her bir *k* objesinin öznitelik sayısı ilgili veri kümesindeki boyut sayısını verir. Şekil 3teki gibi kromozom içinde gerekli görülen öznitelik değerleri kural oluşturmak amacıyla doldurulur.

$$f_1 = \frac{\sum_{k:c_k=c_i} w_k \mu_{Ri}(x_k)}{\sum_{k:c_k=c_i} w_k}$$ fonksiyonu verilen kurala göre pozitif kaplı alan değerini vermektedir.

$$f_2 = \frac{\sum_{k:c_k\neq c_i} w_k \mu_{Ri}(x_k)}{\sum_{k:c_k=c_i} w_k \mu_{Ri}(x_k)}$$ fonksiyonu verilen kuralın negatif kaplı bölümünü sonuç olarak döndürmektedir.

$$f = \begin{Bmatrix} 0, f_2 > k_{\max} \\ f_1 * (1 - f_2 / k_{\max}), f_2 \le k_{\max} \end{Bmatrix}$$

Genetik algoritmada kullanılacak olan uygunluk(fitness) fonksiyonu önceki iki fonksiyon kullanılarak tanımlanmıştır.

$$E(R_t) = \frac{\sum_{k:c_k\neq c_i} w_k \mu_{Ri}(x_k)}{\sum_{k} w_k \mu_{Ri}(x_k)}$$ Bir kuralın hata oranı değerini döndürmektedir.

$$w_k(t+1) = \begin{cases} w_k(t), c_i \neq c_k \\ w_k(t) * \beta_k, c_i = c_k \end{cases}$$ k nesnesinin veri kümesi içerisinde ağırlığı hesaplanır. Birçok hipotezle sınıflandırması sonucu obje belli bir ağırlığa sahip olacaktır. Buna göre sınıflandırılamayan nesnelerin ağırlığı sonraki yinelemelerde güçlendirilmiş sınıflandırma oluşturmak için arttırılacak; doğru sınıflandırma yapılan objelerin ağırlığı da azalacaktır.

Önerilen sistemde ilk adımda bulanık birliktelik kuralları elde edilmektedir. Hesaplamanın karmaşıklığı nedeniyle kuraldaki eleman sayısının en fazla üç olarak belirlenmiştir. Her bir sınıf için birliktelik kuralları elde edilmiştir.

Kurallar destek ile güvenirlik çarpımı sonucuna göre sıralanmıştır. Elde edilen kurallar her bir sınıf etiketi için genetik algoritmanın ilk populasyonuna rastgele birleşik olarak karıştırılarak yerleştirilir. Her adımda yeniden birleştirme ile çapraz değiştirme ve mutasyon işlemlerinden geçirilir. Genetik algoritmada en iyileme sonucunda en son populasyonda en iyi sonucu döndüren kural bulanık kural veri tabanına aktarılır. Aktarılan kural genetik algoritma en iyilemesi sonucunda elde edilen en iyi kural sıfırdan yüksek uygunluk fonksiyonu içerdiği müddetçe bulanık kural kümesine kabul edilir.

Kural veri tabanına aktarılacak olan her kural için genetik algoritma en iyileme yapar. En iyileme esnasında populasyon içerisinde kromozomlar jenerasyonlar boyunca yeniden birleşme adımlarından geçerler. Populasyon içerisinde belirlenen kromozomlardan k_{elite} tanesi sonraki adımdaki kuşak populasyon içine atılır. Eniyileme ile uygunluk fonksiyonu en iyi olan kural seçilir. Onun hata oranı hesaplandıktan sonra verilerin ağırlığı yen

3. DENEYLER

Deneyler Intel Xeon 1.40 GHz CPU, 512 MB RAM Windows XP Dell PC bilgisayarda gerçekleşmiştir. Oracle database 8i Personal Server Edition kullanılmıştır. KDD Cup 99 verisi kullanılmıştır. Birliktelik kuralı için A-priori algoritması kullanılmıştır [1] Genetik algoritma için GAlib, C++ kütüphanesi kullanılmıştır [8].

Sistem 5 defa çalıştırılmıştır. Her çalıştırmada %2 örnekleme ile birliktelik kuralları çıkartılmıştır. KDD Cup verisi ağ trafik davranışlarını içerir. Her objede 42 öznitelik vardır[18]. Eğitim verisinde 494014 obje vardır. Test verisinde 311029 obje bulunmaktadır. Sınıf etiketleme bilgileri awk kodu ile elde edilmiştir [19].

Tablo 1 Eğitim Verisi

Sınıf	Sınıf Adı	Obje Sayısı	%
0	Normal	97271	19,6
1	Probe	4107	0,83
2	DOS	391458	79,1
3	U2R	59	0,01
4	R2L	1119	0,2

Tablo 2 Test Verisi

Sınıf	Sınıf Adı	Obje Sayısı	%
0	Normal	60593	19,4
1	Probe	4166	1,33
2	DOS	231455	74,4
3	U2R	88	0,028
4	R2L	14727	4,73

Tablo 3 Sınıflandırma Doğruluk Sonucu

No	Sınıf Adı	Doğruluk
0	Normal	95.80
1	Probe	54.10
2	DOS	97.40
3	U2R	10.9
4	R2L	6.9

Tablo 3'te sınıflandırma sonucunu her bir etiket bazında doğruluk sonucu verilmiştir.

4. SONUÇLAR

Güçlendirilmiş sınıflandırıcı uygulaması gerçekleştirilmiştir. Bu sistem genetik algoritma ve bulanık mantık ile birliktelik kurallarını kullanarak kural tabanı oluşturmayı hedeflemiştir. Oluşturulan kural tabanı ile ağ saldırılarını sınıflandırmak amacı güdülmüştür.

KAYNAKLAR

[1] Agrawal R. and Srikant R., "Fast Algorithms for Mining Association Rules," Proc. of the International Conference on Very Large Dataases, 1994.

[2] Balasubramaniyan J., et al, "An Architecture for Intrusion Detection using Autonomous Agents," Proc. of the Annual Computer Security Applications Conference, pp. 13-24, 1998.
10

[3] Bridges S. and Vaughn R. B., "Fuzzy Data Mining and Genetic Algorithms Applied to Intrusion Detection," National Information Systems Security Conference, 2000

[4] Crosbie M., "Applying Genetic Programming to Intrusion Detection," Proc. of AAAI Fall Symposium Series, 1995.

[5] Cordon O., et al, Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases, Advances in Fuzzy Systems, World Scientific, Singapore, 2001.
[6] Denning D. E., "An Intrusion Detection Model," IEEE Transactions on Software Engineering, Vol.13, pp.222-232, 1987.
[7] Fox K. L., et al, "A Neural Network Approach towards Intrusion Detection," Proc. of the National Security Conference, pp.125-134, Washington DC, 1990.
[8] Wall M., GAlib A C++ Library of Genetic Algorithm Components http://lancet.mit.edu/ga/
[9] Gomez J. and Dasgupta D., "Evolving Fuzzy Classifiers for Intrusion Detection," Proc. of IEEE Workshop on Information Assurance, United States Military Academy, NY, June 2001.
[10] Hofmeyr S. A., An Immunological Model of Distributed Detection and its Application to Network Security, PhD. Thesis, University of New Mexico, 1999.
[11] Hoffmann, F., "Boosting a Genetic Fuzzy Classifier," Proc. of IFSA/NAFIPS joint conference, Vancouver, 2001.
[12] Hofmeyr S. A., Forrest S. and Somayaji A., "Intrusion Detection Using Sequences of System Calls," Journal of Computer Security, 6:151-180, 1998.
[13] Hong T. P., Kuo C. S and Chi S. C, "Trade-off between Computation Time and Number of Rules For Fuzzy Mining From Quantitative Data," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9. pp.587-604, 2001.
[14] Heady R., et al, "The Architecture of Network Level Intrusion Detection System," Technical Report Department of Computer Science, University of New Mexico, 1990.
[15] Ishibuchi H., Yamamato T. and Natashima T., "Fuzzy Data Mining: Effect of Fuzzy Discretization," Proc. of IEEE International Conference on Data Mining, pp.241-248, 2001.
[16] Ishibuchi H. and Yamamoto T., "Fuzzy rule selection by data mining criteria and genetic algorithms," Proc. of Genetic and Evolutionary Computation Conference, pp.399-406, NY, 2002.
[17] Jang J. S. R. and Sun C. T., "Neuro-Fuzzy Modeling and Control," IEE Proceedings, vol. 83, pp.378-406, Mar. 1995.
[18] KDD Cup 199 Data (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html)
[19] Elkan C.: Results of the KDD'99 Learning Classifier Contest. (http://www-cse.ucsd.edu/users/elkan/clresults.html)
[20] Lunt T., "Detecting intruders in computer systems," Proc. of Auditing and Computer Technology Conference, 1999. (http://www2.csl.sri.com/nides/index5.html)
[21] Lee W., et al, "Mining Audit Data to Build Intrusion Detection Models," Proc. of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.66-72, 1998.
[22] Portnoy L., Eskin E. and Stolfo S., "Intrusion Detection with Unlabeled Data Using Clustering," Proc. of ACM Workshop on Data Mining Applied to Security, 2001.
[23] Ryan J., Lin M. and Miikkulainen R., "Intrusion Detection with Neural Networks," Advances in Neural Information Processing Systems 10, Cambridge MA, MIT Press (1998).
[24] Reyes P. and Sipper M., "A fuzzy Genetic Approach to Breast Cancer Diagnosis," Artificial Intelligence in Medicine, Vol.12, No.2, 1999.
[25] Syswerda G., "Uniform Crossover in Genetic Algorithms," Proc. of the International Conference on Genetic Algorithms, Schaffer, J. (Ed.), Los Altos, CA, pp.2-9, 1989.
[26] Sequeira K. and Zaki M. J., "ADMIT: Anomaly-base Data Mining for Intrusions," Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.
[27] Allen, J., et al, State of the Practice of Intrusion Detection Technologies, CMU/SEI-99-TR-028 (http://www.sei.cmu.edu/pub/documents/99.reports/pdf/99tr028.pdf)., 1999.
[28] Gómez J., et al, "Complete Expression Trees for Evolving Fuzzy Classifiers Systems with Genetic Algorithms and Application to Network intrusion Detection," Proc. of NAFIPS-FLINT joint conference, pp.469-474, New Orleans, LA, 2002.
[29] Goldberg D., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, 1989.