Fırat Üniversitesi Müh. Bil. Dergisi
33(1), 295-301, 2021

Benzetimli Tavlama Algoritması İle Eksik Veri Tamamlama

Serkan METİN*
Yönetim Bilişim Sistemleri Bölümü, Sosyal ve Beşeri Bilimler Fakültesi, Malatya Turgut Özal Üniversitesi, Malatya, Türkiye
serkan.metin@ozal.edu.tr

(Geliş/Received: 31/10/2020; Kabul/Accepted: 27/11/2020)

Öz: İstatiksel birçok yöntem eksik değerlere sahip veri setleri üzerinde çalışma kapasitesine sahip değildir. Bu nedenle, girdi olarak yalnızca tam veriyi kabul eden modellerin tahmin performansı önemli ölçüde düşmektedir. Eksik verilerin tamamlanması bunun için veri analizlerinde önemli bir yere sahiptir. Bu çalışmada kullanılan veri seti üzerinde eksik olan verilerin tamamlanma probleminin çözümünde sezgisel optimizasyon yöntemi olan Benzetimli Tavlama Algoritması(BTA) kullanılmıştır. Modern sezgisel teknikler, bir problem çözümünde, kendi yerel arama sistemleri ile en iyi sonuca ulaşmayı amaçlamaktadırlar. BTA performansını etkileyen en önemli değer başlangıç sıcaklık değeri (T0) olduğundan üç farklı sıcaklık değeri ile sonuçlar alınmıştır. To=100.000 değeri için %68, To=10.000 için %51 ve To=1.000 için %46'lik bir başarı elde edilmiştir.

Anahtar kelimeler: Eksik veri, genetik algoritma, benzetimli tavlama algoritması, sezgisel yöntemler

Completion of Missing Data with the Simulated Annealing Algorithm

Abstract: Many statistical methods are not capable of working on datasets with missing values. Therefore, the forecasting performance of models that accept only full data as inputs drops significantly. For this reason, completing missing data has an important place in data analysis. Simulated Annealing Algorithm (SAA), a heuristic optimization method, was used to solve the problem of completing the missing data on the data set used in this study. Modern heuristic techniques aim to achieve the best results with their local search systems when solving a problem. Since the most important value affecting SAA performance is the initial temperature value (T_0), results have been obtained with three different temperature values. The following success rates were obtained: 68% for T_0=100.000, 51% for T_0=10.000 and 46% for T_0=1.000.

Key words: Missing data, genetic algorithm, simulated annealing algorithm, heuristic methods

1. Giriş

Gözlemsel verileri analiz ederken karşılaşılan en yaygın sorun eksik verilerdir. Gerçek veri kümeleri, çeşitli nedenlerden dolayı eksik verilerden oluşabilir [1]. Endüstri, tıp, ticaret ve bilimsel araştırmalar gibi [2] çok farklı kaynaklardan alınarak oluşturulan veri setlerinde eksik bilgiler ile karşılaşılabilir [3]. Toplanan verilerde tutarsızlıklar, hatalar, aykırı değerler ve eksik değerler gibi çeşitli kusurlar olabilir. Özellikle eksik veri oranı çok yüksek olan veri setlerinde, veri madenciliği veya makine öğrenmesi yöntemleri uygulanırken performans düşerken [4], istatiksel yöntemlerde ise eksik veriler ile tahmin yapmak oldukça sorunludur [5]. Girdi olarak yalnızca tam veriyi kabul eden algoritmalar eksik bir veri ile test edildiklerinde tahmin sonuçlarında önemli bir hata oluşmaktadır [6]. Bu nedenle, veri kalitesini artırmak için eksik veriler tamamlanmalıdır [7]. Eksik veri terminolojisi ilk kez Little ve Rubin tarafından kullanılmıştır [8]. Eksik değerleri ele almak için iyi bilinen ve hesaplama açısından basit birkaç yaklaşım vardır [1]:

- Eksik kayıtları göz ardı etmek.
- Boş değerleri manuel olarak doldurmak.
- Eksik veriyi ortalama veya medyan değeri ile doldurmak

Eksik değerlere yaklaşımlarda yapılacak ilk adım verinin örüntü varlığını incelenmektir [9]. Veri setinde yer alan eksik verilerin belli bir örüntü oluşturmadığı durumlarda farklı çözüm yöntemleri önerilmektedir [10]. Bu yöntemler silme, yaklaşık değer atama [11] ve model tabanlı atama yöntemleri olarak sınıflandırılabilir.

* Sorumlu yazar: serkan.metin@ozal.edu.tr Yazarların ORCID Numarası: 0000-0003-1765-7474

Eksik verilerin değerlendirilmesindeki ilk yöntem kayıp veri olan kaydı yok saymaktır. Ancak eksik verinin çok olduğu ya da az kayıta sahip testlerde bu çözüm sonucu yanlış değerlere saptırmaktadır. Bu tür veri setlerinde eksik değer yerine yeni bir değer ataması yapmak çok daha iyi bir yaklaşımdır [12]. Bu gibi veri setlerindeki eksik sütun değerlerini tahmin etmek için hem sezgisel hem de model tabanlı gösterim yöntemleri kullanılır [13]. Eksik veri tamamlama ile ilgili kullanılabilecek yöntemler Şekil 1'de verilmiştir.

Şekil 1. Kayıp/Eksik Veri Çözümleme Stratejileri [14]

2. Benzetimli Tavlama Algoritması

Benzetimli Tavlama Algoritması (BTA), metallerin tavlama sürecinden esinlenerek [15] belirli bir maliyet fonksiyonunun küresel optimumuna yaklaşmak için tasarlanmış [16] sezgisel bir optimizasyon tekniğidir [17]. BTA, katıların ısıtılması ve ardından yavaşça soğutulması esasına dayanır [18]. Isıtılan katıların sıcaklığı düştüğünde, katının iç parçacıkları her sıcaklıkta bir denge durumuna ulaşır [19]. BTA, ısı arttıkça, en iyi yerel optimayı bulmak için komşu bölgeye gidecektir. Yavaş yavaş soğumaya başladığında ise en iyi yerel optimada durmaya çalışacaktır [20]. Bu yaklaşım, optimizasyon problemine en çözümü bulmak için kullanılır [21].

BTA, çözümde rastgele değişiklik yapabildiğinden yerel olarak optimal bir çözüme düşme olasılığı az olduğundan [22] geleneksel optimizasyon algoritmalarından daha güvenilir [23] ve problem için daha iyi sonuçlar verir [24]. BTA'nın yerel çözümlerde takılı kalmaması için bir P kabul olasılığı tanımlanır [25]:

$$P = e^{-\Delta E/T} \quad (1)$$

ΔE farklı zaman aralıklarında malzemenin enerji değişimini, T ise sıcaklık değerini temsil eder. Başlangıç değeri olan T'nin değeri her yinelemede yavaş yavaş azalacaktır. BTA'nın kurallarına göre birincil çözüm tamamen rastgele oluşmaktadır [26]. Bu neden ile BTA'nın performansı büyük ölçüde başlangıç değerine bağlıdır. Başlangıç değerinin kalitesi zayıfsa, sonuç yetersiz olur [27]. Bu sebeple, aramaya yeteri kadar yüksek bir sıcaklık değeri ile başlamak gereklidir [28].

3. Yöntem

3.1. Benzetimli Tavlama Algoritması Yöntemi

Kirkpatrick 1983'de, kombinatoryal optimizasyon problemini çözmek için ilk kez BTA algoritmasını kullanmıştır [29]. BTA, pek çok farklı alandaki optimizasyon problemlerine kapsamlı bir şekilde uygulanan [30] etkili bir optimizasyon algoritmasıdır [31]. BTA yöntemine ait çözüm adımları [32] :

Adım 1: s = S_0 kriterlere uyan herhangi bir çözüm

Adım 2: t = t_0 başlangıç sıcaklık değeri.

Adım 3: α sıcaklık düşürme kuralını belirle.

$t=t-\alpha$ (2)

$t=t*\alpha$ (3)

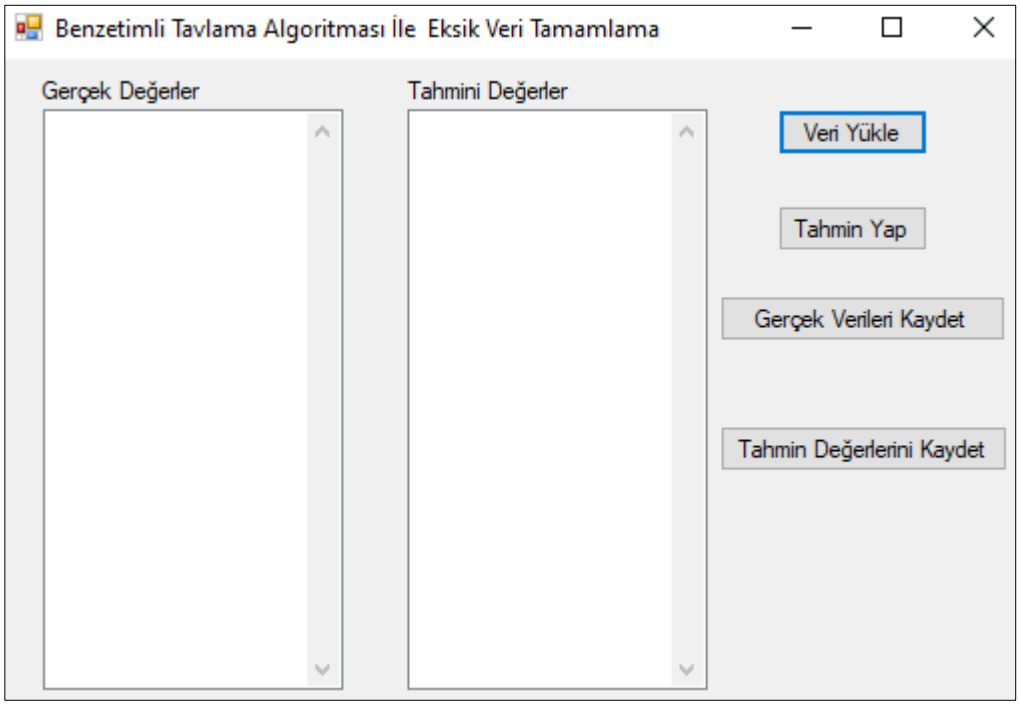
$t=t/(1+\beta t)$ β rastgele bir sayı (4)

Adım 4: İlk sıcaklıktan başlayarak, 5. Adımın n yinelemesini tekrarlayın ve ardından sıcaklığı α'ya göre düşürlür.

Adım 5: N (s) çözümlerinin komşuluğunu göz önünde bulundurarak, çözümlerden birini seçin ve eski çözüm ile yeni komşu çözüm arasındaki maliyet farkı hesaplanır.

Adım 6: Eski ve yeni çözüm arasındaki maliyet farkı 0'dan büyükse yeni çözümü farkı 0'dan düşükse eski çözüm kabul edilir.

Uygulama için Türkiye'ye ait Sağlık Bakanlığı tarafından yayınlanan Covid-19 verileri kullanılmıştır. Veri setinden rastgele olarak değerler silinerek yeni veri seti oluşturulmuştur. Veri seti içerisindeki "Tanı Sayısı" sütununa ait veriler Şekil 2'de verilmiş olan programa ait arayüzde bulunan veri yükle seçeneği ile programa dahil edilmiştir.



Şekil 2. Uygulama arayüzü

BTA ile tahmin işlemi başlatıldığında rastgele silinmiş olan satırdaki bilgiler var olan bilgilerden faydalanılarak tespit edilmiş ve eksik veriler tamamlanmaya çalışılmıştır. Uygulamada kullanılan BTA kod yapısı [33]:

1 Başlangıç sıcaklığık değeri $\mathbf{T_0}$

2 İlk çözümü oluşturun ***s***

3 **While** sonlandırma şartlar oluşmazsa

4 **v**= Komşuseçim(***s***)
5 ***f***= Değerlendirme(**v**)
6 If ***f*** tatmin olasılıklı kabul kriteri
7 ***s*** **= v**
8 Tavlama programına göre **To** güncelle
9 Son
10 Çıktı ***s***

Eksik veri tamamlamada kullanılan BTA algoritmasının doğru sonuca ulaşmasındaki en büyük etken başlangıç sıcaklık değeridir. Çalışmamızda 3 farklı başlangıç sıcaklık değeri kullanılarak veri seti içerisindeki eksik veriler tamamlanmıştır. İlk örnek setinde başlangıç sıcaklı değeri sırası ile 100.000, 10.000 ve 1.000 olarak belirlenmiştir. Verilen başlangıç değerine göre elde edilen başarı oranları sırası ile %68, %51 ve %46 olarak tespit edilmiştir. Tablo 1'de elde edilen sonuçlara ait değerler Tablo 1, Tablo 2 ve Tablo 3'te verilmiştir.

Tablo 1. Başlangıç değeri=100.000 için bulunan değerler

Gerçek Değerler	Tamamlanan Değerler	Gerçek Değerler	Tamamlanan Değerler	Gerçek Değerler	Tamamlanan Değerler
343955	40974	370832	20163	38226	38226
308069	45459	47	47	6	6
349519	36538	170132	170132	186493	186493
15679	15679	52167	52167	190165	190165
155686	155686	158762	158762	20921	20921
169218	169218	185245	185245	314433	48467
222402	32177	23934	23934	366208	23892
355528	32091	301348	42853	120204	120204
1529	1529	138657	138657	209962	32707
7402	7402	135569	135569	117589	117589
244392	33198	127659	127659	82329	82329
112261	112261	10827	10827	306302	43506
42282	42282	284943	43786	98674	98674
182727	182727	95591	95591	90980	90980
126045	126045	281509	43376	86306	86306
361801	27296	164769	164769	191657	191657
114653	114653	174023	174023	261194	40525
176677	176677	324443	46929	195883	195883
197239	197239	237265	33296	156827	156827
5	5	191	191	308069	47167

Tablo 2. Başlangıç değeri=10.000 için bulunan değerler

Gerçek Değerler	Tamamlanan Değerler	Gerçek Değerler	Tamamlanan Değerler	Gerçek Değerler	Tamamlanan Değerler
201098	201098	56956	56956	223315	30702
47029	47029	244392	35731	302867	46243
312966	46360	373154	18105	219641	29610
173036	173036	167410	167410	162120	162120

98	98	13531	13531	149435	149435
368513	22123	255723	39016	299810	45579
288126	42132	274943	41402	212993	31282
179831	179831	353426	33648	278228	41696
154500	154500	181298	181298	345678	39386
159797	159797	233851	32973	139771	139771
238450	34133	150593	150593	107773	107773
1236	1236	148067	148067	283270	43570
291162	44515	184031	184031	279806	40508
289635	43708	163103	163103	152587	152587
110130	110130	226100	31904	5698	5698
131744	131744	104912	104912	244392	34501
947	947	334031	51052	327557	51500
193115	193115	224252	30606	267064	41780
178239	178239	69392	69392	2433	2433
218717	32032	320070	43990	30217	30217

Tablo 3. Başlangıç değeri=1.000 için bulunan değerler

Gerçek Değerler	Tamamlanan Değerler	Gerçek Değerler	Tamamlanan Değerler	Gerçek Değerler	Tamamlanan Değerler
332382	52642	188897	188897	292878	44735
9217	9217	133721	133721	229891	32208
224252	31958	65111	65111	221500	32105
18	18	168340	168340	163942	163942
175218	175218	225173	30574	166422	166422
122392	122392	144749	144749	194511	194511
227982	30644	302867	44622	18135	18135
3629	3629	251805	36560	311455	46089
236112	31669	253108	36994	253108	35652
74193	74193	240804	32113	265515	41537
199906	199906	227019	31977	239622	34420
146457	146457	214993	32280	215940	30832
230873	32390	1	1	171121	171121
211981	31472	124375	124375	129491	129491
187685	187685	137115	137115	228924	29455
240804	33406	670	670	241997	32406
157814	157814	153548	153548	323014	50459
250542	37489	274943	42700	359	359
1872	1872	248117	36381	271705	42406
265515	38960	223315	32033	288126	44198

4. Sonuçlar

Bu çalışmada veri madenciliği, makine öğrenmesi ve istatiksel yöntemlerde kullanılan veri setleri içerisindeki eksik verilerin sezgisel optimizasyon tekniği olan BTA ile tamamlanması amaçlanmıştır. Çalışma kapsamında literatür incelenerek eksik veri tamamlama yaklaşımları belirlenmiştir. Eksik veri tamamlama yöntemleri kullanılırken var olan verilerden faydalanmak genel bir yaklaşımdır. BTA'nın tercih edilmesinin sebebi algoritmanın yerel optimum noktalarında takılı kalmamasıdır.

BTA incelendiğinde başlangıç değeri başarı oranını doğrudan etkilediğinden uygun bir başlangıç değerinin seçilmesi gerekmektedir. Algoritma bir optimizasyon yöntemi olduğu için uygulanacak olan veri setindeki eksik verilerin neden kaynaklandığı belirlenmelidir. Eksik veriye ait örüntü çıkarıldıktan sonra yöntem uygulanmalıdır.

Elde edilen sonuçlara bakıldığında en yüksek olarak %68'lik bir başarı yakalanmıştır. Başarı oranını artırabilmek için optimum başlangıç değerini ayarlayabilecek yöntemler geliştirilebilirse eksik veri tamamlama konusunda önemli katkılar sağlayacağı düşünülmektedir. BTA algoritmasının dezavantajı ise soğuma işleminin uzun sürmesidir.

Kaynaklar

[1]. Sefidian A.M, Daneshpour N. Estimating missing data using novel correlation maximization based methods, Applied Soft Computing Journal, 2020; 91: 106249.

[2]. Rahman M.G, Islam M.Z. Missing value imputation using a fuzzy clustering-based EM approach, Knowl. Inf. Syst. 2016; 46 (2): 389–422.

[3]. Gopalakrishnan R, Guevara C.A, Akiva M. Combining multiple imputation and control function methods to deal with missing data and endogeneity in discrete-choice models, Transportation Research Part B, 2020; 142: 45–57.

[4]. Ye C, Wang H, Li J, Gao H, Cheng S. Crowdsourcing-Enhanced missing values imputation based on Bayesian network, in: International Conference on Database Systems for Advanced Applications, Springer; 2016: 67–81.

[5]. Mercaldo S.F, Blume J.D. Missing data and prediction: the pattern submodel, Biostatistics, 2020; 21(2): 236–252.

[6]. Zhiyong C, Longfei L, Ziyuan P. Yinhai Wang Graph Markov network for traffic forecasting with missing data, Transportation Research Part C; 2020.

[7]. Qin Y, Zhang S, Zhu X, Zhang J, Zhang C. POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases, Expert Syst. Appl., 2009; 36 (2): 2794–2804.

[8]. Molenberghs G, Thijs, H, Jansen I, Beunckens, C, Kenward, M.G, Mallinckrodt, C, Carroll, R.J. Analyzing Incomplete Longitudinal Clinical Trial Data; 2004.

[9]. Sayın A, Yandı A, Oyar E. Kayıp Veri ile Baş Etme Yöntemlerinin Madde Parametrelerine Etkisinin İncelenmesi, Journal of Measurement and Evaluation in Education and Psychology, 2017; 8(4): 490-510.

[10]. Carpita M, Manisera M. On the imputation of missing data in surveys with Likert-type scales. Journal of Classification, 2011; 28(1): 93-112.

[11]. Demir E, Parlak B. Türkiye'de eğitim araştırmalarında kayıp veri sorunu. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 2012; 3(1): 230-241.

[12]. Sezgin E, Çelik Y.Veri Madenciliğinde Kayıp Veriler İçin Kullanılan Yöntemlerin Karşılaştırılması, Akademik Bilişim Konferansı, Akdeniz Üniversitesi, 2013.

[13]. Little, R.J.A. , Rubin, D.B. Statistical Analysis with Missing Data: Second Edition. John Wiley and Sons, 2002

[14]. Şener Y. Veri Biliminde Eksik/Kayıp Verilere Yaklaşım Stratejileri ve Python (Pandas) Uygulaması, 2020

[15]. Min L, Yue C, Xiaojing S, Zhishan Z, Xiaoxiao Z, Xiuyu Z, Jun G. Simulated annealing-based optimal design of energy efficient ternary extractive dividing wall distillation process for separating benzeneisopropanol-water mixtures, , Chinese Journal of Chemical Engineering,2020.

[16]. Xiangzhen Z, Sanjiang L, Yuan F. Quantum Circuit Transformation Based on Simulated Annealing and Heuristic Search, , Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020.

[17]. Moriguchi K. Acceleration and enhancement of reliability of simulated annealing for optimizing thinning schedule of a forest stand, Computers and Electronics in Agriculture, 2020.

[18]. Kirkpatrick S, Gelatt C.D, Vecchi, M.P. Optimization by Simulated Annealing. Science new series, 1983; 220(4598): 671–680.

[19]. Songsheng T, Minjun P, Genglei X, Ge W, Cheng Z. Optimization design for supercritical carbon dioxide compressor based on simulated annealing algorithm, Annals of Nuclear Energy, 2020.

[20]. Asrul S.R, Ikram M, Mohd R. Mohd A.A. Energy Management Strategy of HEV based on Simulated Annealing, Int. J. of Integrated Engineering, 2020; 12(2): 30-37.

[21]. Lizhong Z, He M, Wei Q, Haiyan L. Protein structure optimization using improved simulated annealing algorithm on a three-dimensional AB off-lattice model, Computational Biology and Chemistry, 2020.

[22]. Jin C, Bin W. Flocking Control of Mobile Robots via Simulated Annealing Algorithm, Proceedings of the 39th Chinese Control Conference, 2020; 3931- 3935.

[23]. Tatsuya K, Hideharu K, Hiroyuki N, Tatsuhiro T. Using simulated annealing for locating array construction, Information and Software Technology 126, 2020.

[24]. Attiya I, Elaziz M, Xiong S. Job Scheduling in Cloud Computing Using a Modified Harris Hawks Optimization and Simulated Annealing Algorithm, Computational Intelligence and Neuroscience, 2020.
[25]. Hanine M, Benlahmar E H.A Load-Balancing Approach Using an Improved Simulated Annealing Algorithm, J Inf Process Syst, 2020;16:132-144.
[26]. Jafari H, Ehsanifar M, Sheykhan A. Finding Optimum Facility's Layout by Developed Simulated Annealing Algorithm, Int. J. Res. Ind. Eng., 2020; 9(2): 172–182.
[27]. İlhan İ. A population based simulated annealing algorithm for capacitated vehicle routing problem, , Turk J Elec Eng & Comp Sci, 2020; 28: 1217–1235.
[28]. Cayıroglu I. İleri Algoritma Analizi, 2020.
[29]. Xianze M, Yunpeng F, Junsheng Y. Estimating solubilities of ternary water-salt systems using simulated annealing algorithm based generalized regression neural network, Fluid Phase Equilibria, 2020.
[30]. Cunha M, Marques J. A New Multiobjective Simulated Annealing Algorithm—MOSA-GR: Application to the Optimal Design of Water Distribution Networks, Water Resources Research, 2019.
[31]. Minghao G, Chunbo W, Baicheng L, Bin S. Yuanshen Huang Design and implementation of a Placido disk-based corneal topographer optical system based on aberration theory and simulated annealing algorithm,Optics Communications 475, 2020.
[32]. Liang F. Optimization Techniques Simulated Annealing A popular method for optimizing model parameters, 2020.
[33]. Tsai C.W, Hsia CH, Yang SJ, Liu SJ, Fang ZY. Optimizing hyperparameters of deep learning in predicting bus passengers based on simulated annealing, Applied Soft Computing Journal 88, 2020.