ÖZET

KİŞİSEL VERİLERİN ANONİMLEŞTİRİLMESİNİN İYİLEŞTİRİLMESİNE YÖNELİK BİR MODEL GELİŞTİRİLMESİ VE E-DEVLET ALANINDA UYGULANMASI

Mustafa AFYONLUOĞLU

Doktora, Elektrik ve Elektronik Mühendisliği Bölümü

Tez Danışmanı: Prof. Dr. Ali Ziya ALKAR

Haziran 2019, 341 sayfa

Günümüzde, e-Devlet hizmetleri sayesinde ortaya çıkan verinin istatistik, araştırma-geliştirme, yapay zekâ öğrenimi, hizmet iyileştirme, projeksiyon geliştirme başta olmak üzere pek çok alanda değerli bir kaynak olduğunun fark edilmesi, bu veriyi işlemek üzere elde etme ihtiyacını arttırmıştır. Ancak e-Devlet hizmetlerinden doğan bu büyük hacimli verinin önemli bir bölümü kişisel verilerden oluşmaktadır ve mahremiyetin korunmasına yönelik olarak gerek Avrupa Birliği gibi uluslararası kuruluşlarda Genel Veri Koruma Tüzüğü (GDPR) ile gerekse ülkemizde 6698 sayılı Kişisel Verilerin Korunması Kanunu gibi mevzuat altyapısı ile korunmakta ve verinin paylaşımı için anonimleştirme şart koşulmaktadır. Verinin belli kısımlarının kapatılması veya genelleştirilmesi gibi yöntemlerle gerçekleştirilen anonimleştirme ile söz konusu veri, bir gerçek kişi ile ilişkilendirilemeyecek hale getirilmektedir. Bu konudaki mahremiyet standartlarının en bilineni olan k-anonimlik (kişileri aynı yarı tanımlayıcı değere sahip en az k tane kayıt

grupları şeklinde tutarak mahremiyeti sağlar), ℓ-çeşitlilik (her bir denklik sınıfının her hassas özellik için en az ℓ adet değere sahip olmasını gerektirir) ve t-yakınlık (herhangi bir denklik sınıfında hassas özniteliğin dağılımının, özniteliğin veri setindeki dağılımına yakın olmasını, yani bu iki dağılım arasındaki mesafenin t eşik değerinden fazla olmamasını gerektirir) uygulamasında, veri üzerinde yapılan genelleştirme işlemleri sebebiyle, belirli seviyede veri kaybı oluşmakta ve bu durum, sonuç veri kümesinden elde edilmesi beklenen faydayı azaltmaktadır.

Bu doktora çalışmasında, veri niteliğini dikkate alarak ve kayda özelleşmiş grupları adaptif şekilde oluşturarak veri fayda kaybını en aza indirgeyen ve tam baskılanan kayıtların sayısında yüksek oranda iyileştirmeler sağlayan, hedef odaklı, e-Devlet verisi dahil olmak üzere her türlü veri setine uygulanabilen, yenilikçi bir adaptif ve dinamik anonimleştirme modeli (ADAM) ortaya konulmuştur. e-Devlet veri kümeleri büyük hacimlerden oluştuğu için, ortaya konulan modelin, yüksek veri hacimlerinde de beklenen iyileştirmeyi sağlaması gerekmektedir. Bu sebeple, bu buluşsal yöntemin sağlayacağı iyileştirme seviyesini ölçebilmek için, e-Devletin önemli uygulama alanlarından birisi olan sağlık alanında sentetik veri üreten bir uygulama geliştirilmiş ve 1.000 kişilik veri setlerinden başlayarak kademeli olarak 100.000 kişilik sentetik veri setleri oluşturularak ADAM algoritması uygulanmış, mevcut anonimleştirme yöntemlerine kıyasla önerilen modelin önemli ölçüde iyileştirmeler sağlayabildiği gösterilmiştir.

Anahtar Kelimeler: Kişisel Veri, Anonimleştirme, e-Devlet, k-Anonimlik, Adaptif Dinamik Anonimleştirme

5. SONUÇLAR

5.1. Sonuçlar

Anonimleştirme alanında birçok akademik çalışma olmakla birlikte, bu çalışmalar, kişisel verinin ve buna bağlı olarak hassas verilerin ayrıştırılabilmesi ve ayırt edilebilmesini engelleyen modellere odaklanmış olup, anonimleştirme verimini artıran, yani bu işlemi yaparken ortaya çıkan veri kaybını en aza indirgeyerek faydayı maksimum seviyeye taşımayı hedefleyen, anonimleştirilecek verinin karakteristiğini dikkate alarak veri kümesine özelleştirilmiş yöntem ileri süren bir çalışmanın olmadığı görülmektedir. Ayrıca böyle bir çalışmanın asıl kullanım yeri olan ve verinin en büyük kaynağı olan e-Devlet ile ilişkilendirilmesi halinde, özellikle kavram ispatı çalışmalarında daha gerçekçi sonuçlar çıkacağından, çalışma sahası olarak e-Devlet alanının ele alınması elzem olup, bu yöndeki bir çalışmanın eksikliği dikkati çekmiştir.

Bu doktora çalışmasında, kişisel verilerin imha sürecindeki bütünlüğünün korunmasını bünyesinde barındıracak bir e-Devlet yazılım çerçevesi (Şekil 2.6) ile buna entegre olacak şekilde kişisel verilerin imhasında bütünlüğünün korunmasına yönelik bir model (Şekil 4.42) ve e-Devlet hizmetleri tarafından üretilen kişisel verilerin anonimleştirilmesinde veri fayda kaybını en aza indirecek hedef odaklı bir anonimleştirme modeli (Bölüm 4) ortaya konulması hedeflenmiş, ayrıca bir kavram ispatı çalışması (Bölüm 4.4) ile bu anonimleştirme modelinin ortaya koyacağı verim ölçümlenmeye çalışılmıştır.

Ortaya konulan Kişisel Veri İmha Merkezi modeli ile, veri imhasında kurum içi ve kurum dışı bütünlüğün sağlanması hedeflenmiş ve modelin sağlayacağı faydayı ispat için yapılan kavram ispatı çalışmasında, üretilen sentetik veriler ile verinin üretilmesinden, paylaşımına, imhasına ve kurumlar arası senkronizasyonuna kadarki tüm aşamaları içeren toplam 14 senaryo (Bölüm 4.4.7) kurgulanarak uygulanmış, beklenen tüm bütünlük gereksinimlerinin karşılandığı görülmüştür.

Öngörülen iteratif model ile, kişisel verilerin korunması bakımından ülkelerin mevzuatlarında esas alınan k-anonimlik ölçütünü sağlayacak şekilde, genelleştirme ve

baskılama yöntemlerinin başta alfa sayısal, sayısal ve zaman alanlarında dinamik olarak uygulanmak üzere, "her bir yarı tanımlayıcı veri kümesinin içerik karakteristiğine özel" olarak yapılacak bir anonimleştirme modeli (Dinamik Anonimleştirme Modeli, DAM) ortaya konulmuştur. DAM ile, öncelikle anonimleştirme şartlarını sağlayan kayıtlar ayıklanır, geriye kalan kayıtların veri karakteristiğine uygun şekilde maske boyutu ayarlanır. İkinci nesil DAM çözümünde, maske birleşik olma kısıtı da kaldırılarak her bir kayda özel en kısa maske uygulanır. Böylece kayıt bazlı özel genelleştirme uygulandığı için tüm veri kümesinde, verinin içerik karakteristiğine uygun olarak dinamik şekilde belirlenmiş olur.

Adaptif ve dinamik anonimleştirme modelinin (ADAM) uygulamasında ise, DAM'daki uygulama, bir uzunluğundaki maskeden (ve hiyerarşik veriler için birinci seviye hiyerarşi uygulamasından) itibaren maske uzunluğu (ve seviye) bir arttırılarak kümeye tekrar tekrar uygulanır ve her durum için, kümeye uygun uzunluk ve desendeki maske ile k-anonimlik şartını sağlayan alt veri kümeleri tespit edilir. Böylece, tüm küme için en uygun tek dinamik maske yerine, farklı alt kümelere daha düşük boyutlu dinamik maskeler uygulanarak daha da verimli bir netice elde edilir.

Geliştirilen ADAM modeli ve kişisel veri bütünlük (KİM) modeli ile yazılım çerçevesini ilgili bileşenlerini de içeren üst model, e-Devlet verileri dahil her türlü veri setine uygulanabilecek genel bir modeldir.

Kavram ispatı çalışması için, her türlü akademik çalışmada kullanılabilecek, içerik olarak herhangi bir sapma veya eğilim (bias) içermeyen, istenilen büyüklükte veri üretebilen sentetik veri üretim uygulaması yazılmış ve sağlık alanında hasta, muayene ve tahlil veri tablolarının, TÜİK tarafından yayımlanan nüfus istatistikleri ve Sağlık Bakanlığı Ulusal Sağlık Veri Sözlüğü (ve buralarda tanımlanmış uluslararası standartlarla) ile uyumlu olacak şekilde verinin üretilmesi sağlanmıştır. Bu uygulama ile üretilen 1.000, 5.000 ve 10.000 kişilik veri setleri üzerinde, k=2, 3, 4 ve 5 anonimlik seviyeleri için var olan anonimleştirme yöntemleri ve bu çalışmada öngörülen DAM ve ADAM modelleri, tekil veri karakteristiğine sahip alanda (TCKN gibi anonimlik oranı %0 olan) ve uygulamada yaygın olarak gözlenen yapıdaki veri alanında (posta kodu gibi anonimlik oranı %24-%36 arasında) ayrı ayrı uygulanmıştır.

Posta kodu alanında farklı boyutlardaki veri setleri ile gerçekleştirilen kavram ispatı ölçümlemelerinde (EK- 6), 1.000 kişilik örnek sentetik veride k=2-5 aralığı için, bilinen yöntemlerle ölçülen veri fayda kaybı oranı %38,64 - %42,64 iken önerilen model uygulandığında fayda kaybının %17,40 - %14,04 seviyelerine düşerek 2,22 ile 3,03 kat (ortalama %262) iyileştirme olduğu görülmüştür. 10.000 kişilik sentetik veri ile aynı çalışma tekrar edildiğinde, veri fayda kaybının %34,57 - %39,86 aralığından %4 - %8,08 aralığına düştüğü görülmüş ve yaklaşık 8,64 – 4,93 kat (ortalama %679) iyileştirme olduğu ölçülmüştür. Genel olarak yapılan ölçümlemelerde, sağdan genelleştirme yöntemine göre ortalama %298 - %555 iyileştirme olduğu görülmektedir (Şekil 4.62).

ADAM, veri tablolarında en yaygın olarak bulunan alfa nümerik alanlara uygulanabildiği gibi sayısal alanlara da uygulanabilir. Çünkü sayısal alanlar anonimleştirilirken öncelikle alt-üst sınır kodlama gibi yöntemler uygulanır. Mevcut uygulamada, sınırlar ve aralıklar belirlenirken veri karakteristiğine özel sistematik bir yaklaşım bulunmamakta ve yaygın olarak aralıklar sabit alınmaktadır. Zaman (tarih, saat veya her ikisi) bilgisi içeren alanlarda da global kodlama yapılmakta ve aynı sebeplerle verim kayıpları yaşanmaktadır. Bu yöntemden kaynaklı fayda kaybını iyileştirmek için, benzer iteratif yaklaşım ile uygulanacak alt-üst sınırlar ve aralıkların dinamik olarak belirlenmesi için ADAM bu tür alanlarda da kullanılabilir niteliktedir. Dolayısıyla, anonimleştirilecek bir veri setinin barındırdığı her türlü veri tipi için ADAM uygulanabilir bir modeldir.

Şart koşulan k-anonimlik seviyesine ulaşmak için, ADAM'ın getireceği katkıların en başında, orijinal verinin anonimleştirme işlemine tabi tutulurken çok daha az kısmının genelleştirilmesinin yeterli olması, çok daha az sayıda kaydın genel baskılamaya tabi tutulması ve bu sayede ortaya çıkan yüksek fayda kazanımıdır. Ayrıca hedef odaklı maskeleme tercih yaklaşımı, anonim verinin kullanım amacına göre özelleştirilmiş bir anonimleştirme sunmaktadır.

Bu kazanımlar, özellikle yapay zekâ dünyasının öğrenme için girdi olarak ihtiyaç duyduğu verilerde, araştırma-geliştirme çalışmalarında, istatistiki çalışmalarda, hedef belirleme amacıyla gerçekleştirilen projeksiyon çalışmalarında, veriden ekonomik değer yaratan sosyal ve teknolojik projelerde ve kavram ispat çalışmaları ile deneysel çalışmalar için sentetik veri üretemeyen ve/veya gerçek veriye ihtiyaç duyan akademik çalışmalarda büyük önem ifade etmektedir. Bu çalışmada ortaya konulan model, bahsedilen faydalar

için başta e-Devlet verileri olmak üzere her türlü veriye uygulanabilecek genel bir yapıdadır ve çalışmanın toplumsal fayda bileşenini oluşturmaktadır.

Bunlara ilaveten, sağlık alanındaki her türlü akademik ve sektörel çalışmanın ihtiyacı olan gerçek sağlık verilerinin, demografik bilgileri ve bunun coğrafi dağılımını dikkate alan tüm araştırma ve yatırımların gereksinim duyduğu verilerin, eğitim araştırmalarında ihtiyaç duyulan verilerin, ticaret dünyasında yatırım kararının önemli bir girdisi olan demografik, coğrafi, ekonomik ve diğer sayısal verilerin, yaşam kalitesini arttırmaya yönelik kişiye özel elektronik servis geliştirebilmek için ihtiyaç duyulan konum bilgisi dahil bir çok verinin, bu model ile daha yüksek faydayı sağlayacak şekilde iyileştirilmiş anonimleştirmeye tabi tutulması, çalışmanın sosyal hayata katkısı olarak öne çıkmaktadır.

Bu çalışma kapsamında gerçekleştirilen e-Devlet ölçümleme alanındaki akademik çalışmalara dair SLR çalışması ve bu alandaki uluslararası ölçümleme çalışmalarının kıyaslanmasına ilişkin yapılan yayın, anonimleştirme alanındaki akademik çalışmalara katkı sağlayabilecek sentetik veri üretmek üzere geliştirilmiş kapsamlı bir uygulama, anonimleştirme alanında detaylı analizlerin yapılabildiği, ara adımların izlenebildiği ve istenilen akademik çalışmanın uygulanabileceği şekilde geliştirilmiş olan anonimleştirme ve kavram ispatı aracı, daha önce yapılmamış ve literatüre katkısı olabilecek şekilde geliştirilen özgün ve bütüncül model ile bu model kapsamında geliştirilen yazılım çerçeve modelinin kişisel veri bileşenleri, kişisel veri kayıt imha bütünlük modeli, anonimleştirmede yüksek iyileştirme sağlayan 3 ayrı model, (1. Nesil DAM, 2. Nesil DAM ve ADAM), bu çalışmanın literatüre katkıları olarak değerlendirilebilir.

5.2. Önerilen Çalışmalar

Bu modelin e-Devlet projelerinde uygulanması halinde, çok büyük hacimli veriler üzerinde çalışılacağından, algoritmanın performansının yüksek olması önemli hale gelecektir. Bu durumda, benzer şartlar sağlandığında, ADAM ile gerçekleştirilmiş önceki anonimleştirme deneyimlerinden (ilgili yarı tanımlayıcı veri seti için uygun maskeleme tavsiyelerinden) faydalanılması için merkezi bir yapıda "veri türleri anonimleştirme maske kütüphanesi" modeli üzerinde çalışma yapılmalıdır. ADAM yaklaşımı, veri kümesindeki kayıt sayısından bağımsızdır. Dolayısıyla yarı tanımlayıcının niteliği aynı olan her tabloda aynı maskenin en etkin olacağı açıktır. Örneğin posta kodu daima 5

karakter uzunluğundadır ve içerik karakteristiği standarttır. Dolayısıyla bir defaya mahsus olmak üzere, posta kodu alanı için en verimli maskeler belirlendiğinde, tüm maskeler için iterasyon yapmak yerine sadece bu maskelerin adaptif olarak en uygun alt kümelere uygulanması yeterli olacaktır. Böylece, aynı karakteristiğe sahip yarı tanımlayıcılar için uygun maskelerin listelendiği bir merkezi tablo, daha sonraki anonimleştirmeler için tüm kurumlara önemli iterasyon tasarrufu ve performans artışı sağlayacaktır. Dolayısıyla bu merkez model ile ADAM yapısı kendi kendine öğrenen ve sonra bu öğrendiği verileri, kendi performansını arttırmak için kullanan bir yapıya dönüşecektir.

Performans konusunda yapılabilecek bir diğer çalışma, yarı tanımlayıcı veri setinin tamamında kayıt bazında iterasyona başlamadan evvel, örneklem yöntemi ile veri kümesinden alınacak belirli aralıklardaki kayıtlar için yapay zekâ yaklaşımı ile genel karakteristik eğiliminin belirlenmesi ve böylece yapılacak iterasyon sayısının azaltılmasıdır. Böylece, alan uzunluğuna bağlı olarak uygulanabilecek maske uzunluklarının tüm kombinasyonları yerine, yapay zekâ algoritmasından belirli bir eşik değerini aşarak yüksek potansiyel olduğu belirlenen maske boyutları ve kombinasyonları ile ADAM algoritması uygulanabilecektir. [111]