International Journal of
Applied Mathematics,
Electronics and Computers
ISSN:2147-8228 www.atscience.org/IJAMEC
Advanced Technology and Science
Original Research Paper

Big Data: Controlling Fraud by Using Machine Learning Libraries on Spark

Ferhat Karataş*[1], Sevcan Aytaç Korkmaz[2]

DOI: 10.18100/ijamec.2018138629

Abstract: Continuous changes and the high calculation volume in network data distribution have made it more difficult to detect abnormal behaviors within and analyze data. For this cause, large data solutions have gained important. With the advancement of internet technologies and the digital age, cyber-attacks have increased steadily. The k-Means clustering algorithm is one of the most widely used algorithms in the world of data mining. Clustering algorithms are algorithms that automatically divide data into smaller clusters or sub-clusters. The algorithm places statistically similar records in the same group. In this article, we have used k-Means method from the Machine Learning libraries on Spark to determine whether the incoming network values are normal behavior. 400 thousand network data were used in this article. This data was obtained from KDD Cup 1999 Data. We have detected 10 abnormal behaviors from 400 thousand network data with k-means method.

Keywords: *k-Means, Spark, Machine Learning, Anomaly Detection, Big Data*

1. Introduction

Continuous changes and the high calculation volume in network data distribution have made it more difficult to detect abnormal behaviors within and analyze data. For this cause, large data solutions have gained importance [1]. In 2016, the US Department of State Security Network Security Dept. budget request was $ 479.8 million [1,2]. Norton has reported that victims of the cybercriminals have spent $ 126 billion globally since 2015 [1,3]. The increasingly intelligent, complex and destructive nature of cybercrime has led to an increase in these large cybercsecurity investments. Therefore, it is necessary to identify the abnormalities in the network data with the help of a computer in order to reduce these investments and to provide better security of the country. Spark is an open-source platform developed at UC Berkeley AMPLab in 2010. The goal is to perform iterative and efficient computation on large datasets. [4]. Abnormal behaviors can occur in many industries (banking, insurance, network security, etc.). These; forgery on credit cards, forgery on insurance policies, abnormal packet exchanges on the network or potential attacks. Such cases are called fraud or anomaly detection. And it can cause some problems in every sector (such as material losses, reputation losses). Cyber-attacks with the advancement of Internet technologies have increased steadily. These attacks use short networks of networks to access unauthorized to sensitive information. It is very important to intervene before any attack takes place. Big data is the data community that includes a diverse type of datasets [1,5]. Network-based intrusion detection is required to detect unusual behaviors of network users. And it is necessary to perform large data analyzes when making this determination [1].

A threat or attack attempt is to change the information that unauthorized people access and information of the system. They create an anomaly behavior [1,6].

Fraud detection, attack detection, and prevention of data leakage are anomaly detection approaches [1,7]. When the literature is examined in terms of detection of large network anomalies, it can be seen that extensive data network analysis has been performed, such as CTU-13 data [1]. There are many anomaly detections works in the literature. These are network-wide anomaly detection with PCA [8], Kalman filters [9], single-link traffic measurements [10]-[11], Hough Transform [12], sketches [13], and equilibrium properties [14] methods [15]. The study [16] provides a comprehensive summary of techniques for detecting general anomalies [15]. Another study [17] has a specific questionnaire on detection and detection of an anomaly in Internet traffic [15]. Lakhina et al. [18] have used the PCA-based method [19] but they used several entropy metrics based on source. They are advised to reuse these entropy measurements to classify the anomalies [15]. Xu et al. [20] have applied clustering to entropy metrics similar to [21] to classify abnormal events and construct a traffic model [15]. Fernandes et al. [22] have suggested NADA, a signature-based tool that separates abnormalities into different categories [15]. Silveira et al. [23] have suggested URCA, a method for determining the underlying causes of abnormal events [15]. Spark also has very few studies done with k-means. R. Kumari et al. arc. have discussed how these interventions can be detected by using k-means clustering-based machine learning approach using large data analytical techniques, and that the attacks advocate experimental results for multi-core prevention [24]. In one study, abnormal behaviors have been detected using the Principal Component Analysis (PCA) method. The accuracy result is 96%. This approach has been implemented on public Net Flow data [1].

[1] *Fırat University / Biotechnology Department, Elazığ, TURKEY*
[2] *Fırat University, Electronic and Automation Dept, Elazığ, TURKEY*
[1] **Corresponding Email: ferhat@bilgipark.com*

Theory and Method of the study are given in Section 2. Experiment results are described in Section 3. And conclusions are explained in Section 4.

The purpose of this article is to determine what is different from the millions of network movements. The novelty of this article, anomaly detection in spark is performed on KDD Cup 1999 network data using k-means. K-means is used on many platforms. Spark is a current issue in the big data field. And so far it has not been applied for anomalous detection on KDD Cup 1999 network data

2. Theory and Method

Network movements (datasets) are than subjected to a normalization process after they are collected. In this process, abnormal data is removed from the system in order not to be used in modeling. The model is being created after normalization. In the next stage, anomaly detection is made by asking every data to the data model. The algorithm we'll use here is k-Means. K-Means is in Spark's Machine Learning library. If attention is paid to the clusters, similar ones are gathered together. Some points will stay away from the center of the clusters. These points will be defined as abnormal movements.

2.1. What is k-Means?

The k-means method, a multivariate statistical technique, is used to classify homogeneous subgroups according to their similarities. One of the most well-known clustering methods is the k-means method. [25]. In this method, we start with the determination of the centers of the predetermined A units and each variable is assigned to the nearest cluster center according to the similarity [26]. After assigning each variable in the input data set to a cluster, the cluster center is recalculated for each cluster so that the variables can be assigned to different new clusters depending on the location of these new cluster centers. This process is repeated until there is no change in cluster membership. In an examined problem, a T data set with K feature vectors and n variables can be defined as $T = \{t_k | k = 1,2, \ldots . A\}$. In this data set, k. the feature vector can be written as $X_k = [x_{k1}, x_{k2}, \ldots . x_{kn}]$, $x_k \in R^d$ [25,27]. In Equation (1), the data set is divided into the smallest cluster. For the calculation of the distance measure, the Euclidean distance criterion is given in Equation (2) is used [27,28].

$$J(S:T) = \sum_{i=1}^{A} \sum_{k=1}^{B} n_{ik}^2 (t, S_i) \quad (1)$$

The equation given in Equation (1) n_{ik}^2 is defined as follows.

$$n_{ik}^2 = \left\| t_k^{(i)} - S_i \right\|^2 \quad (2)$$

$$S_i = \frac{1}{B} \sum_{k=1}^{B} t_k^{(i)} \quad (3)$$

A key advantage of Apache Spark for k-Means is that its machine learning library (MLlib) and its library for Spark Streaming are built on the same core architecture for distributed analytics. This facilitates adding extensions that combine components with novel ways [29]

K-means in Apache Spark is a cluster computing platform that is used for general purposes and designed to be fast. On the speed side, Spark for k-means expands MapReduce model to support more types of computing like interactive queries and stream processing. Speed in spark for k-means is very important inprocessing large datasets, as it means the difference between exploring data interactively and waiting minutes or hours. One of the main features that Spark for k-means proposes for speed is the ability to compute in memory [30].

2.1. Data processing

Each point has x and y values. We define random centers on these points. By doing iterations at a later stage, a new center point is determined according to the distance of the points. In the last stage, the structure we build will be our model. And the incoming data will now be interpreted according to this model. We need a datasheet for our article. We used KDD Cup 1999 Data for this [31]. When we look at the data, we see that each line shows the details of data exchange on the network. The KDD Cup names in the URL we give will give you a list of which columns these lines correspond to. For example:

Table 1. Examples of the KDD Cup 1999 Data

protocol_type: symbolic.
service: symbolic.
flag: symbolic.
src_bytes: continuous.
dst_bytes: continuous.
land: symbolic.
wrong_fragment: continuous.
urgent: continuous.
hot: continuous.
num_failed_logins: continuous.
logged_in: symbolic.
num_compromised: continuous.
root_shell: continuous.
su_attempted: continuous.
num_root: continuous.
num_file_creations: continuous.
num_shells: continuous.
num_access_files: continuous.
num_outbound_cmds: continuous.
is_host_login: symbolic.
is_guest_login: symbolic.
count: continuous.
srv_count: continuous.
serror_rate: continuous.
srv_serror_rate: continuous.
rerror_rate: continuous.

We will create these fields as an object when doing java coding. Below are datasets that sample these areas.

Table 2. Examples of the Normal Network Data in Java Coding

0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8, 0.00,0.00,0.00,0.00,1.00,0.00,0.00,9,9,1.00,0.00,0.11,0.00, 0.00,0.00,0.00,0.00,normal.
0,tcp,http,SF,239,486,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00, 0.00,0.00,0.00,1.00,0.00,0.00,19,19,1.00,0.00,0.05,0.00,0.00, 0.00,0.00,0.00,normal.
0,tcp,http,SF,235,1337,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,8,8,0.00, 0.00,0.00,0.00,1.00,0.00,0.00,29,29,1.00,0.00,0.03,0.00,0.00, 0.00,0.00,0.00,normal.

2.2. Coding

The program code used in this article consists of 5 steps.

Step 1: Download the kddcup.data_10_percent.gz file from KDD Cup 1999 Data [31].

Step 2: Create a Maven article on Eclipse.

Step 3: After creating the article, our first task will be to configure our pom.xml file. The dependency libraries we will use in the article will be loaded into the article. [32]

With spark-core_2.10, spark-sql_2.10, spark-mllib_2.10 we have added spark's machine learning libraries. better-monads are used to translate json objects into java classes.

Step 4: Creating the class Kdd.java. Our goal is to convert every line in the dataset to a java object. [33]

Step 5: We are making the following improvements on the main class called App.java. Each section is described in the comments line what it did. [34]

3. Experimental Results

When the program code is run, the following 10 results have been seen on the screen. The 10 anomaly values founded on the KDD Cup 1999 data are below.

Anomaly
data[6.9337564E8,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,57.0,3.0,0.79,0.67,0.21,0.33,0.05,0.39,0.0,255.0,3.0,0.01,0.09,0.22,0.0,0.18,0.67,0.05,0.33]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,3.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,255.0,1.0,0.0,0.08,0.32,0.0,0.01,0.0,0.0,0.0]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,1.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,2.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,7.0,1.0,0.14,0.29,1.0,0.0,0.0,0.0,0.0,0.0]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,2.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,1.0,27.0,1.0,0.0,1.0,0.22,0.0,0.04,0.0,0.0]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,28.0,1.0,0.0,1.0,0.11,0.0,0.0,0.0,0.0]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,3.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,1.0,34.0,1.0,0.0,1.0,0.15,0.0,0.0,0.0,0.0]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,4.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,255.0,66.0,0.26,0.13,0.36,0.0,0.02,0.02,0.0,0.0]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,3.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,2.0,42.0,1.0,0.0,1.0,0.1,0.0,0.0,0.0,0.0]

Anomaly
data[5135678.0,0.0,0.0,0.0,0.0,4.0,0.0,1.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,1.0,1.0,0.0,0.0,0.0,0.0,1.0,0.0,0.0,5.0,30.0,1.0,0.0,1.0,0.17,0.0,0.0,0.0,0.0]

When we examine the dataset we use, we see that the first value after SF is usually below 1000. We see that records that are detected as anomalies are in 50.000. This indicates that something else is happening within the network. With the k-Means algorithm we use, it is possible to evaluate the given data according to a model and to take necessary actions.

.

4. Discussion and Conclusion

Until today, many field machine learning techniques have been used [35-46]. Cyber-attacks with the advancement of Internet technologies have increased steadily. These attacks use short networks of networks to access unauthorized to sensitive information. It is very important to intervene before any attack takes place. Big data is the data community that includes a diverse type of datasets [1,5]. Network-based intrusion detection is required to detect unusual behaviors of network users. And it is necessary to perform large data analyzes when making this determination [1]. A threat or attack attempt is to change the information that unauthorized people access and information of the system. They create an anomaly behavior [1,6]. Fraud detection, attack detection, and prevention of data leakage are anomaly detection approaches. Continuous changes and the high calculation volume in network data distribution have made it more difficult to

detect abnormal behaviors within and analyze data. For this cause, large data solutions have gained importance. With the advancement of internet technologies and the digital age, cyber-attacks have increased steadily. The k-Means clustering algorithm is one of the most widely used algorithms in the world of data mining. Clustering algorithms are algorithms that automatically divide data into smaller clusters or sub-clusters. The algorithm places statistically similar records in the same group.

Our study have been compared with similar studies in the literature. The results in similar studies in the literature are below.

In a article study [47] has been suggested the parallel version of K-means implemented on Hadoop MapReduce1. According to this, when spark has been compared with Hadoop, Spark is more suitable for parallelizing the iterative algorithms such like K-means. The distributed memory abstraction called as resilient distributed datasets (RDDs) can be cache both intermediate data and input data in memory [48,49]. it has been discussed how to parallelize K-means-based algorithms on Spark in a paper [49]. According to this, K-means-based clustering algorithms include two iterative procedures: centroid updating and distance computation. Also, technical details of two phases have been discussed. Especially, it has been given an implementation detail for parallelizing K-Means-based clustering on Spark. Further, it has been illustrated the their alternative strategies and technical barrier for each step. Experiments on text datasets and large-scale UCI datasets have been suggested that the effectiveness of the algorithms demonstrated [49]. In another paper [50], it have been presented a new K-Means based algorithm implemented on Spark. The this algorithm has been indicated to automate the input of number of clusters in advance, which is the major drawback of the classical K-Means algorithm. The proposed algorithm has also been indicated to tackle the resolution problem. it have been shown with experimental results that proposed algorithm works efficiently on large scale data sets and outperforms the K-Means algorithm implemented in Spark Machine Learning Library. Moreover the algorithm has been scaled gracefully on adding more machines to cluster and increasing the data size. In another article [51] have been successfully designed intelligent k-means based on spark. It has been runned in Hadoop environment. the algorithm has been designed using batch of data. Also, it has been compared with the version of algorithm without using batch of data. It has been suggested with experiment results that design can be speed up computational time in big data problem. In addition, authors have suggested that design have higher silhouette value than original k-means using synthetic data. In this article, we have used k-Means method from the Machine Learning libraries on Spark to determine whether the incoming network values are normal behavior. 400 thousand network data were used in this article. This data was obtained from KDD Cup 1999 Data. We have detected 10 abnormal behaviors from 400 thousand network data with k-means method.

In future work, anomaly detection in spark will perform on KDD Cup 1999 network data using different methods.

References

[1] Terzi, Duygu Sinanc, Ramazan Terzi, and Seref Sagiroglu. "Big data analytics for network anomaly detection from netflow data." Computer Science and Engineering (UBMK), 2017 International Conference on. IEEE, 2017.

[2] Budget-in-Brief Fiscal Year 2016, US Department of Homeland Security, Editor. 2016.

[3] 2016 Norton Cyber Security Insights Report. 2016.

[4] Meng, Xiangrui, et al. "Mllib: Machine learning in apache spark." The Journal of Machine Learning Research 17.1, 1235-1241, 2016.

[5] Terzi, Duygu Sinanc, Ramazan Terzi, and Seref Sagiroglu. "Big data analytics for network anomaly detection from netflow data." Computer Science and Engineering (UBMK), 2017 International Conference on. IEEE, 2017.

[6] Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. "Network anomaly detection: methods, systems and tools." IEEE communications surveys & tutorials 16.1, 303-336, 2014.

[7] Goldstein, Markus, and Seiichi Uchida. "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data." PloS one 11.4, 2016.

[8] Lakhina, Anukool, Mark Crovella, and Christophe Diot. "Diagnosing network-wide traffic anomalies." ACM SIGCOMM Computer Communication Review. Vol. 34. No. 4. ACM, 2004.

[9] Soule, Augustin, Kavé Salamatian, and Nina Taft. "Combining filtering and statistical methods for anomaly detection." Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement. USENIX Association, 2005.

[10] Barford, Paul, et al. "A signal analysis of network traffic anomalies." Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment. ACM, 2002.

[11] Fontugne, Romain, et al. "Random projection and multiscale wavelet leader based anomaly detection and address identification in internet traffic." Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015.

[12] Fontugne, Romain, and Kensuke Fukuda. "A Hough-transform-based anomaly detector with an adaptive time interval." ACM SIGAPP Applied Computing Review 11.3, 41-51, 2011.

[13] Kanda, Yoshiki, et al. "ADMIRE: Anomaly detection method using entropy-based PCA with three-step sketches." Computer Communications 36.5, 575-588, 2013.

[14] Silveira, Fernando, et al. "ASTUTE: Detecting a different class of traffic anomalies." ACM SIGCOMM Computer Communication Review 40.4, 267-278, 2010.

[15] Mazel, Johan, et al. "Hunting attacks in the dark: clustering and correlation analysis for unsupervised anomaly detection." International Journal of Network Management 25.5, 283-305, 2015.

[16] Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3, 15, 2009.

[17] Marnerides, Angelos K., Alberto Schaeffer-Filho, and Andreas Mauthe. "Traffic anomaly diagnosis in internet backbone networks: a survey." Computer Networks 73, 224-243, 2014.

[18] Lakhina, Anukool, Mark Crovella, and Christophe Diot. "Mining anomalies using traffic feature distributions." ACM SIGCOMM Computer Communication Review. Vol. 35. No. 4. ACM, 2005.

[19] Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies. Proceedings of the 4th conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM), 2004.

[20] Xu, Kuai, Zhi-Li Zhang, and Supratik Bhattacharyya. "Internet traffic behavior profiling for network security monitoring." IEEE/ACM Transactions On Networking 16.6, 1241-1252, 2008.

[21] Lakhina, Anukool, Mark Crovella, and Christophe Diot. "Mining anomalies using traffic feature distributions." ACM SIGCOMM Computer Communication Review. Vol. 35. No. 4. ACM, 2005.

[22] Fernandes, Guilherme, and Philippe Owezarski. "Automated classification of network traffic anomalies." International Conference on Security and Privacy in Communication Systems. Springer, Berlin,

Heidelberg, 2009.
[23] Silveira, Fernando, and Christophe Diot. "URCA: Pulling out anomalies by their root causes." INFOCOM, 2010 Proceedings IEEE. IEEE, 2010.
[24] Kumari, R., et al. "Anomaly detection in network traffic using K-mean clustering." Recent Advances in Information Technology (RAIT), 2016 3rd International Conference on. IEEE, 2016.
[25] Muda, Z., et al. "Intrusion detection based on K-Means clustering and Naïve Bayes classification." Information Technology in Asia (CITA 11), 2011 7th International Conference on. IEEE, 2011.
[26] Ozcift, Akin, and Arif Gulten. "Assessing effects of pre-processing mass spectrometry data on classification performance." European Journal of Mass Spectrometry 14.5, 267-273, 2008.
[27] FIRAT, Mahmut, et al. "K-ortalamalar yöntemi ile yıllık yağışların sınıflandırılması ve homojen bölgelerin belirlenmesi." Teknik Dergi 23.113 (2012).
[28] Leśniak, Andrzej, and Zbigniew Isakow. "Space–time clustering of seismic events and hazard assessment in the Zabrze-Bielszowice coal mine, Poland." International Journal of Rock Mechanics and Mining Sciences 46.5, 918-928, 2009.
[29] https://databricks.com/blog/2015/01/28/introducing-streaming-k means-in-spark-1-2.html
[30] https://www.researchgate.net/publication/318155071_Comparative_Study_of_Apache_Spark_MLlib_Clustering_Algorithms
[31] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[32] https://gist.github.com/bilgipark/72de4f6b6ef75db4178b12badc7048ce
[33] https://gist.github.com/bilgipark/aca73e9f2259d9005f0912fe0c852914
[34] https://gist.github.com/bilgipark/a21d645eb09e0e3e74ee95dd61180a8c
[35] Korkmaz, Sevcan Aytac, and Mehmet Fatih Korkmaz. "A new method based cancer detection in mammogram textures by finding feature weights and using Kullback–Leibler measure with kernel estimation." Optik-International Journal for Light and Electron Optics 126.20, 2576-2583, 2015.
[36] Korkmaz, Sevcan Aytac, Mehmet Fatih Korkmaz, and Mustafa Poyraz. "Diagnosis of breast cancer in light microscopic and mammographic images textures using relative entropy via kernel estimation." Medical & biological engineering & computing 54.4, 561-573, 2016.
[37] KORKMAZ, Sevcan Aytaç, et al. Recognition of the stomach cancer images with probabilistic HOG feature vector histograms by using HOG features. In: Intelligent Systems and Informatics (SISY), 2017 IEEE 15th International Symposium on. IEEE, p. 000339-000342, 2017.
[38] Korkmaz, Sevcan Aytac, et al. "Diagnosis of breast cancer nano-biomechanics images taken from atomic force microscope." Journal of Nanoelectronics and Optoelectronics 11.4, 551-559, 2016.
[39] Korkmaz, Sevcan Aytaç, and Hamidullah Binol. "Analysis of Molecular Structure Images by using ANN, RF, LBP, HOG, and Size Reduction Methods for early Stomach Cancer Detection."*Journal of Molecular Structure* (2017).
[40] Korkmaz, S. A. (2018). LBP Özelliklerine Dayanan Lokasyon Koruyan Projeksiyon (LPP) Boyut Azaltma Metodunun Farklı Sınıflandırıcılar Üzerindeki Performanslarının Karşılaştırılması. Sakarya University Journal of Science, 22(4), 1-1.
[41] Korkmaz, S. Aytac, and Mustafa Poyraz. "A New Method Based for Diagnosis of Breast Cancer Cells from Microscopic Images: DWEE--JHT." Journal of medical systems 38.9 (2014): 1.
[42] Korkmaz, Sevcan Aytac, and Mustafa Poyraz. "Least square support vector machine and minumum redundacy maximum relavance for diagnosis of breast cancer from breast microscopic images." Procedia-Social and Behavioral Sciences 174 (2015): 4026-4031.
[43] KORKMAZ, Sevcan Aytac; EREN, Haluk. Cancer detection in mammograms estimating feature weights via Kullback-Leibler measure. In: Image and Signal Processing (CISP), 2013 6th International Congress on. IEEE, 2 (2013):1035-1040.
[44] KORKMAZ, Sevcan AYTAÇ. "DETECTING CELLS USING IMAGE SEGMENTATION OF THE CERVICAL CANCER IMAGES TAKEN FROM SCANNING ELECTRON MICROSCOPE." The Online Journal of Science and Technology-October 7.4 (2017).
[45] KORKMAZ, Sevcan Aytaç, et al. Recognition of the stomach cancer images with probabilistic HOG feature vector histograms by using HOG features. In: Intelligent Systems and Informatics (SISY), 2017 IEEE 15th International Symposium on. IEEE, (2017). p. 000339-000342.
[46] Korkmaz, S. A., Poyraz, M., Bal, A., Binol, H., Özercan, I. H., Korkmaz, M. F., & Aydin, A. M. (2015). New methods based on mRMR_LSSVM and mRMR_KNN for diagnosis of breast cancer from microscopic and mammography images of some patients. International Journal of Biomedical Engineering and Technology, 19(2), 105-117.
[47] S. Owen, R. Anil, T. Dunning, and E. Friedman, Mahout in action. Manning Shelter Island, 2011.
[48] M. Zaharia, M. Chowdhury, and T. Das, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in Proceedings of USENIX conference on Networked Systems Design and Implementation, 2012, pp. 2–2.
[49] Wang, Bowen, et al. "Parallelizing k-means-based clustering on spark." Advanced Cloud and Big Data (CBD), 2016 International Conference on. IEEE, 2016.
[50] Sinha, Ankita, and Prasanta K. Jana. "A novel K-means based clustering algorithm for big data." Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on. IEEE, 2016.
[51] Kusuma, Ilham, et al. "Design of intelligent k-means based on spark for big data clustering." Big Data and Information Security (IWBIS), International Workshop on. IEEE, 2016.