International Journal of Database Theory and Application
Vol. 11, No.2 (2018), pp. 1-10
hhtp://dx.doi.org/10.14257/ijdta.2018.11.2.01

Towards Outlier Detection: A Survey

Muhammad Arif [1,2]

Department of Computer Science University of Gujrat, Pakistan
Faculty of computer science and technology university Malaya
arifmuhammad36@hotmail.com

Abstract

The objective of this paper is to present a survey on Outlier Detection Algorithms for High Dimensional Data in different applications to solve the real-world problems. The aim to present this paper is to facilitate the new researches of this field as it includes a lot of application in which Outlier Detection Algorithms for High Dimensional has been used. This paper also contains the experimental results of each problem which is solved by Outlier Detection Algorithm for High Dimensional Data.

Keywords: *Outlier Detection Algorithms, High Dimensional Data*

1. Introduction

Outlier is classify as an observation that is very different from the rest of observations that it stimulates uncertain patterns as it was generated by a different mechanism from large amount data and from data warehouse [16, 21, 22,23,24]. The recognition of outliers can guide to the invention of valuable and important knowledge and has a number of useful applications in different areas. For example, transportation, safety and health of public, location based services and many more. The rest of paper contains survey of different papers presented on Outlier Detection for High Dimensional Data.

A. Examples Based Robust Outlier Detection in High Dimensional Datasets:

In this paper the author discuss the novel algorithm for outlier detection in high dimensional datasets named Robust Algorithm [22]. With the help of this algorithm, he has overcome the problem of detecting outlier for dataset having d dimensions and to find out the hidden k dimensional subspace.

The inputs of this algorithm are as follows:

E: Set of outlier examples

Φ: Number of equi depth ranges for each attributes

The outputs of this algorithm are as follows:

O: Set of outliers

FE: Set of false examples

The body of this algorithm does the following processes:

P: Initial population of solution

Received (December 27, 2017), Review Result (March 3, 2018), Accepted (March 8, 2018)

ISSN: 2005-4270 IJDTA

Under the while loop it performs three processes iteratively:

P <- Selection (p); For SELECTION it used rank selection mechanism for better selection

II. P <- Crossover (p); this function performs two types of CROSSOVER:
- Scattered Crossover
- Optimized Crossover, both are specially designed for outlier detection

III. P <- Mutation (p); and this is the end of this loop.

After this we have following sets:

i. S: Solutions with best fitness value
ii. O: Set of objects which are sparser
iii. FE: Set of examples which sparsity coefficient is positive

And at the end of this algorithm, it returns two sets which are: O and FE.

Simulation Results:

The following graph shoes the accuracy of this algorithm by comparing Time Cost and Accuracy verses Dimensionality.

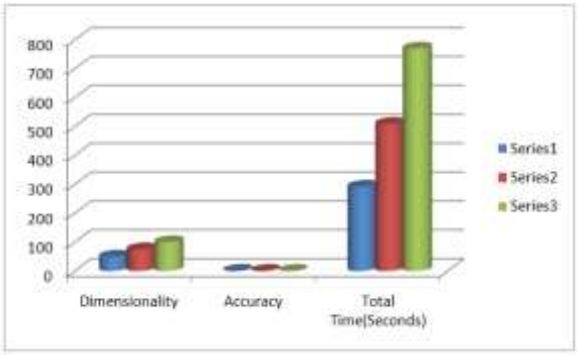

Figure 1. Time, Cost & Accuracy vs. Number of Dimensions

B. Research on Maximal Frequent Pattern Outlier Factor for Online High-Dimensional Time-Series Outlier Detection:

This paper is presented to propose a novel approach named MFPOF (maximal frequent pattern outlier factor) and an algorithm for outlier detection named OODFP for online high-dimensional time-series outlier detection [2,25]. The objective to purpose this approach is to overcome the problem of finding outlier in high dimensional data. Traditional approaches can find outliers only from static data no for dynamic data. First of all, the data streams of time series are processed with sliding window to find out the item sets having maximal frequency. Then the recurrent patterns are arranged and simplified to calculate the maximal frequent pattern outliers of time-series data streams. The main idea behind MFPOF is very simple. If a data objects hold extra frequent patterns, it indicates that it is not likely to be an outlier. A non-empty subset containing maximal frequent pattern define a frequent pattern, and it indicates that the item sets with maximum frequency are already completely holds the set of frequent items. In a set of item with maximum frequency, the number of patterns is much less that the items present in frequent item set. The mentioned MFPOF detect outliers and the purposed algorithm OODEFP is used to discover outlaying objects time series data streams of high

dimensions. The proposed approach has the advantage of being efficient and suitable for online high-dimensional time-series outlier detection. We have found this to be very important in real-world applications of outlier detection. In the near future, we would like to apply this approach to real world applications. The proposed technique has the advantage of efficiency and is appropriate for outlier detection of online high-dimensional time-series.

Simulation Results:

Simulations results demonstrate that this purposed approach not only grant higher effectiveness, but also comparable accurateness.

C. Outlier Detection in High Dimension Based On Projection:

A new approach is discussed here called Outlier Detection in High Dimension based on Projection (ODHDP) to finds the projection based outliers from the data set. In this presented paper, a projection based approach called ODHDP, which is particularly suitable for problem of high dimensional. This approach consists of following three steps shown in flow chart:

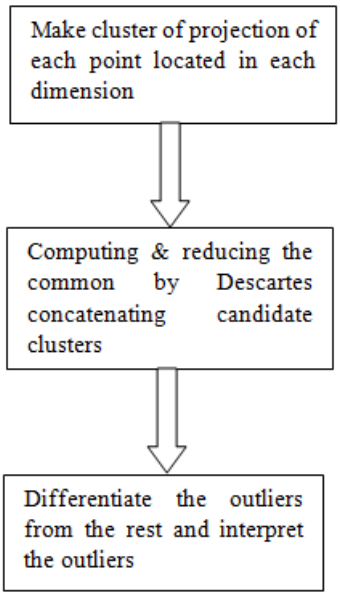

Figure 2. Flowchart of Purposed Approach

The complicatedness in the purposed technique is that prune strategy should be chosen cautiously in Descartes concatenation to avoid amalgamation explosion [3]. This approach used the following parameters:

i. k : initial number of clusters
ii. σ : pruning threshold
iii. τ0: outlier threshold

Simulation Results:

Experiments show that this approach is feasible. And it can also be observe that it is important to choose the number of first clusters and the reduce threshold which can enhance the performance of the Outlier Detection in High Dimension based on Projection algorithm appreciably.

D. OutRank: Ranking Outliers in High Dimensional Data:

In this paper, the author proposes a novel approach named Out Rank for ranking of outliers in heterogeneous data of high dimensional. They present a reliable model for dissimilar attribute types which are continuous and categorical attributes, of high dimensional data [4]. The scoring functions of this technique convert the analyze structure of the data into a significant ranking. Scoring function must replicate the deviation of objects i.e. to perform a ranking of outliers that can be calculated by cataloging objects in rising order according to their scores. Therefore, they expand a latest subspace clustering model [4,11,28,29] for heterogeneous data in a constant manner for attributes of both types.

Simulation Results:

Results in promising testing shows that this approach is successful for outlier ranking in high dimensional data [4]. Current work contract with in detail study of two different scoring functions and their performance. Approaching into their weaknesses and their strengths, could be merging into a framework that works for a wide range of applications of various domains.

E. A Novel Method for Detecting Outlying Subspaces in High dimensional Databases Using Genetic Algorithm:

Here a novel method has been purposed that use patterns of genetic algorithm (GA) for exploration of outlying subspaces competently. The developed method efficiently computing the upper/lower limits of the distance among a certain point and kth neighbor that is nearest to it in every possible subspace. These limits are employed to accelerate the fitness evaluation of the genetic algorithm that designed for detection of outlying subspace. In this paper, an arbitrary sampling technique is presented to further lessen the computation of the GA. To guarantee the accurateness of the result, the best possible number of sampling data is specified [5]. The major task behind this approach is to discover the subspaces in which every data point shows considerable deviation from the remaining population. An outlying subspace of any data point indicates this data point can be a candidate to become an outlier. The formulation of outlying subspace detecting problem is as follows:

i. A data point is given
ii. Subspaces are found in which this point is greatly dissimilar which are incomparable or incompatible according to the rest of population in the database [5, 12].

Mining of Outlier can be profitable from detection of outlying subspace with respect to many aspects which are:

i. Outlying subspace detection can be put in to a better categorization of detected outliers. This categorization of outliers mostly involves presenting that subspaces which contains these outlier. In high-dimensionality, it is significant to not just extract outliers but to find the situation in which these outliers present.

ii. The purposed technique makes it possible to mine outliers more accurately. It also makes it possible to detect outliers from one or more subspaces which allow the applicability of this technique in various real time problems such as: credit card fraud detection.

Simulation Results:

The experiment set that has been accomplished on both artificial and real world data sets; show that the method presented here is effective and efficient in handling problems of detecting outlying subspace [5]. The results exhibits that this method performs accurately and effectively in handling the problems of outlying subspace detection.

F. Outlier Mining in Large High-Dimensional Data Sets:

Here purposed new description of distance-based outlier and an algorithm, named HilOut [6], designed to discover the top n outliers of data set of large and high-dimensional efficiently. According to this approach, outliers are those points having the highest values of weight and to calculate weight an integer value of k should be given and the weight of any point is defined by computed distances of that point from its k nearest-neighbors [6]. The purposed method based on the concept of space-filling curve to make the data set linear, and it consists of two steps whish are:

i. Within an uneven factor, give an estimated solution, after completion of at most d þ 1 sorts and then scans of the whole data set, with temporal cost quadratic in d (number of dimensions) and linear in k and also in N (number of points in that data set) [6]. At this phase, the purposed algorithm separates the point that is a candidate to become an outlier and decreases the size of this set at each iteration. If, after reduction, the size of the set becomes n, then this algorithm discontinues reporting the accurate solution.

ii. In the second phase, it computes the accurate solution with a concluding scan analyzing more point that is candidate outliers that remained after the first step.

Simulation Results:

Testing results demonstrate that during the first phase, the HilOut must discontinue reporting the accurate solution [6]. It is tested on both an in-memory and disk-based implementation [6] of the HilOut algorithm and a detailed evaluation for real time and artificial data set prove that the algorithm performs well in both cases.

G. Monitoring High-Dimensional Data for Failure Detection and Localization in Large-Scale Computing Systems:

A novel technique is proposed here to design the geometry of fundamental of data generation and detect abnormality based on that design [7]. In this paper, both data generation models are considered here either linear or nonlinear. Following two statistics are used to reflect data difference inner and outer the model:

i. Hotelling T2
ii. Squared prediction error (SPE)

They follow the probabilistic density of mined statistics to observe the fitness of the system. After the detection of a failure, they purposed a localization process to discover to discover the most doubtful attributes interrelated to the failure. They begin their work with the state where the examining data is produced from linear structure low-dimensions. SVD (Singular value decomposition) is utilized to find out the linear subspace that holds the greater part of data. The Hotelling T2 and Squared Prediction Error are then developed from the geometry aspects of every measurement according to that subspace

[7]. Then they expand their work to the situation where the essential data structure is nonlinear [7], which is frequently encountered in information systems because of nonlinear method i.e. g. queuing etc in the system. In this paper, a statistical test algorithm is also presented to make a decision that whether the nonlinear or linear model is appropriate, known the measurement data. After this, they calculate the values of two statistics of every measurement and approximate their probabilistic density [7]. The detection of failure based on the variation of recently computed statistics of each upcoming measurement with according to learned density. When a failure is detected, they purposed a localization procedure to expose the main guarded attributes based on the values of desecrated statistics [7]. To get the advantages of the EIV model and fusion in restructurings in manifold [7] they used synthetic data.

Simulation Results:

In this paper, they evaluate the results of the LLE algorithm on the original measurements [7] and the data that has been preprocessed by the EIV model [7] and fusion and evaluation results prove that both presented methods are essential to attain a correct restructuring of the nonlinear manifold. Results of experiments, examined on real time applications of e-commerce and synthetic data, show the efficiency of this approach purposed to detecting and localize failures in computing systems [7].

H. Cluster PCA for Outliers Detection in High-Dimensional Data:

A new method is introduced here to detect various outliers in data set of high-dimensional based on the concepts of analysis of principal component and hierarchical clustering [8]. The presented method is quick computationally and proposed algorithm is computationally fast and strong to detect outliers. Comparisons with presented approaches are executed on high and also on low dimensional datasets. Here they present a novel approach based on distance, referred to as principal component analysis of cluster (cluster PCA). The objective to proposed this method is to recognize outliers in data sets of both low/ high dimensions with the mixture of concepts of hierarchical clustering [8,13] and PCA [8,14] in order to get better performance while keeping the computation time and complexity time comparatively low. The presented method is based on hierarchical clustering in contrast with PCA to attain a strong subset of observations used to discover outliers. Cluster Principal Component Analysis is a model for transforming the observations in a dataset into new observations which are unconnected with each other and report for lessening proportions of the total variance of the original variables [8]. Every new observation indicates a linear combination of the original observations. If the variables are in dissimilar units or while the variance of the unlike columns of the data is significant then standardize the data is preferable.

Simulation Results:

Results of tests show that the purposed algorithm achieves improved performances as compare to already available approaches for outlier detection.

I. SPOT: A System for Detecting Projected Outliers from High-dimensional Data Streams:

In this paper, a novel approach is presented named SPOT (Stream Projected Outlier deTector), to contract with problems of outlier in high-dimensional data streams. The reason to present this approach is following:

i. SPOT utilize a novel a time model based window and decomposing cell summaries to confine statistics from the data stream [9]

ii. SST (Sparse Subspace Template) a set of top sparse subspaces find by one or both learning processes whish are unsupervised/ supervised is developed in SPOT for projected outlier detection effectively. Multi-Objective [9]
ii. SST is capable to accomplish online self-progression to cope with data streams dynamics [9].

SPOT flows among following processes [9]:

i. Time Model
ii. Data Synapses
iii. Stages of SPOT which are:
    - Fixed SST Subspace
    - Clustering Based SST Subspace
    - Outlier Driven SST Subspace
iv. Unsupervised Learning
v. Supervised Learning

Simulation Results:

They do not present any result of experiment or testing phase in their paper.

J. Efficient and Effective Clustering Methods for Spatial Data Mining:

In this paper, they discover whether clustering techniques have an important role to play in mining of spatial data. At the end they developed a novel method for clustering called CLAHANS [18] based on arbitrary search. They also developed the algorithms for two spatial data mining [18] that utilize CLAHANS. Analysis and results of experiments demonstrate that in combination of CLAHANS both algorithms perform very effectively and can help to explore that is difficult to find with already purposed algorithm of spatial data mining [18].

Simulation Results:

They have present results of experiment results show that CLABANS is more competent than already presented methods of clustering. So that, CLARANS has established its own self as a very capable tool for effective and efficient mining of spatial data [10, 23,26, 28,]. Furthermore, experiments perform to compare the performance of CLAHANS with existing methods of clustering, confirm that CLAHANS is the much efficient [10].

K. Simultaneously Removing Noise and Selecting Relevant Features for High Dimensional Noisy Data:

This paper is presented to introduce a novel approach to get better quality of training data sets with noisy dependent variable and high dimensionality [15] by concurrently eliminating noisy occurrences and choosing related features for classification. This technique relies on two GA [32] (Genetic Algorithms) from which one is used for detection of noise and the second for selection of feature, and also allows these algorithms to swap their results sporadically at definite generation intervals. Selection of prototype is used to enhance the performance beside with GA in method of noise detection. GA-ND [15] presents the results in binary encoding representation. The size of the individuals represents the number of occurrence of any point in training data set and every position (gene) in an individual corresponds to the number of an instance in the training data set. If the value of each position is 0 it means the instance is noise free and if the value is 1 its

means the instance is noisy [15]. GA-FS also uses binary encoding representation. Same as GS-ND if the value is 0 it represents selection of the features and 1 defines removal of features from experimental data set.

Simulation Results:

Experimental results shows that this purposed technique achieve better quality of high dimensional noisy training data sets and substantially enhance accuracy of classification. Following graph shows percentage of error reduction.

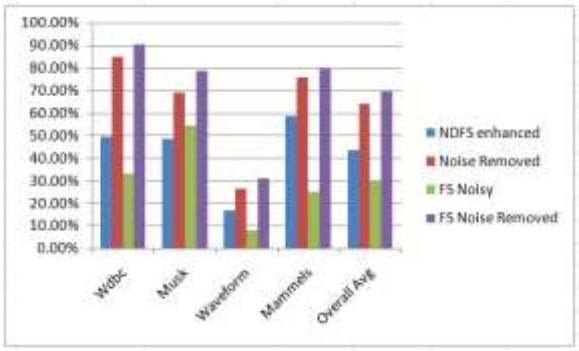

Figure 3. Error Detection Percentage

L. Feature Extraction for Outlier Detection in High-Dimensional Spaces:

In this paper the author introduce an efficient feature extraction that fetch nontrivial enhancements in accuracy of detection when it's applied on two common detection techniques.

This paper has explore the domains of feature extraction on outlier detection research and presents a novel approach called DROUT [20] to achieve the objective. In short, DROUT performs function in two phases which are:

i. Regularization of eigen- space
ii. Discriminant feature extraction [20,30]

In phase one, Decompose the data eigen-sapace in to three parts where various polices of regularization are applied and no subspace is removed. This facilitates DROUT to protect the information that is discriminant in the data earlier than entering the concrete feature extraction process [20].

In the next phase, obtained discriminant features by solving the conventional eigen-value problem on the regularized entire scatter matrix, from the regularized eigen-space.

Further advantage of this approach is that its phases makes DROUT superior tuned to detection techniques of outlier as compare to other presented techniques because they carried-out on the weight-adjusted scatter matrices [20].

Simulation Results:

Examining real data sets show the achievability of feature extraction in outlier detection.

They also examine other promises of reduction of dimensionality for outlier detection separately from their proposed technique [20]. This will help to superior choose appropriate ways for dealing with the curse of dimensionality.

2. Conclusion

In this presented paper, a lot of work has been done in field of Outlier Detection from High Dimensional Data with different changes in. It is used in many applications and provides accurate results.

3. Future Work

Using Outlier Detection Approaches, we can achieve our goal with accuracy and in effective way, when we have to face data of high dimensional. Above research work is very helpful to use these algorithms in future work without take care of the domain that we choose for future work as these algorithms gives better performance in solving problem of any domain.

References

[1] C. Zhu, H. Kitagawa and C. Faloutsos, "Example-based robust outlier detection in high dimensional datasets", In Data Mining, Fifth IEEE International Conference on IEEE, **(2005)** November, pp. 4.
[2] F. Lin, W. Le and J. Bo, "Research on maximal frequent pattern outlier factor for online high-dimensional time-series outlier detection", Journal of convergence information technology, vol. 5, no. 10, **(2010)**, pp. 66-71.
[3] P. Guo, J. Y. Dai and Y. X. Wang, "Outlier Detection in High Dimension Based on Projection", In Machine Learning and Cybernetics, 2006 International Conference on IEEE, **(2006)** August, pp. 1165-1169).
[4] E. Muller, I. Assent, U. Steinhausen and T. Seidl, "OutRank: ranking outliers in high dimensional data", In Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on IEEE, **(2008)** April, pp. 600-603.
[5] J. Zhang, Q. Gao and H. Wang, "A novel method for detecting outlying subspaces in high-dimensional databases using genetic algorithm", In Data Mining, 2006, ICDM'06, Sixth International Conference on IEEE, **(2006)** December, pp. 731-740.
[6] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets", Knowledge and Data Engineering, IEEE Transactions, vol. 17, no. 2, **(2005)**, pp. 203-215.
[7] H. Chen, G. Jiang and K. Yoshihira, "Monitoring high-dimensional data for failure detection and localization in large-scale computing systems", Knowledge and Data Engineering, IEEE Transactions, vol. 20, no. 1, **(2008)**, pp. 13-25.
[8] G. Stefatos and A. B. Hamza, "Cluster pca for outliers detection in high-dimensional data", In Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on IEEE, **(2007)** October, pp. 3961-3966.
[9] J. Zhang, Q. Gao and H. Wang, "Spot: A system for detecting projected outliers from high-dimensional data streams", In Data Engineering, 2008, ICDE 2008. IEEE 24th International Conference on IEEE, **(2008)** April, pp. 1628-1631.
[10] Y. Zhang, N. Meratnia and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey", Communications Surveys & Tutorials, IEEE, vol. 12, no. 2, **(2010)**, pp. 159-170.
[11] I. Assent, R. Krieger, E. Muller and T. Seidl, "DUSC: Dimensionality unbiased subspace clustering", In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on IEEE, **(2007)** October, pp. 409-414.
[12] J. Zhang and H. Wang, "Detecting outlying subspaces for high-dimensional data: the new task, algorithms, and performance", Knowledge and information systems, vol. 10, no. 3, **(2006)**, pp. 333-355.
[13] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern classification", John Wiley & Sons, **(2012)**.
[14] I. Jolliffe, "Principal component analysis", John Wiley & Sons, Ltd., **(2002)**.
[15] B. Byeon and K. Rasheed, "Simultaneously removing noise and selecting relevant features for high dimensional noisy data", In Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on IEEE, **(2008)** December, pp. 147-152.
[16] D. M. Hawkins, "Identification of outliers", London: Chapman and Hall, vol. 11, **(1980)**.
[17] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki and D. Gunopulos, Online outlier detection in sensor data using non-parametric models, In Proceedings of the 32nd international conference on Very large data bases, VLDB Endowment, **(2006)** September, pp. 187-198).
[18] M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", In Kdd, vol. 96, no. 34, **(1996)** August, pp. 226-231.
[19] C. Aggarwal and S. Yu, "An effective and efficient algorithm for high-dimensional outlier detection", The VLDB Journal-The International Journal on Very Large Data Bases, vol. 14, no. 2, **(2005)**, pp. 211-221.
[20] H. V. Nguyen and V. Gopalkrishnan, "Feature extraction for outlier detection in highdimensional spaces", Journal of Machine Learning Research, vol. 10, **(2010)**, pp. 66-75.

[21] M. Arif, "A survey on data warehouse Construction, Processes and Architecture", International Journal of u- and e- Service, Science and Technology, vol. 8, no. 4, **(2015)**, pp. 9-16.
[22] M. Arif and F. Zaffar, "Challenges in efficient Data warehousing", International Journal of Grid and Distributed Computing, vol. 8, no. 2, **(2015)**.
[23] M. Arif and A. Roohani Dar, "Survey on Fraud Detection Techniques Using Data Mining", International Journal of u-and e-Service, Science and Technology, vol. 8, no. 3, **(2015)**, pp. 165-170.
[24] M. Arif and T. Mahmood, "Cloud Computing and its Environmental Effects", International Journal of Grid and Distributed Computing, vol. 8, no. 1, **(2015)**, pp. 279-286.
[25] M. Arif, K. Amjad Alam and M. Hussain, "Crime Mining: A Comprehensive Survey", International Journal of u-and e-Service, Science and Technology, vol. 8, no. 2, **(2015)**, pp. 357-364.
[26] M. Arif and H. Shakeel, "Virtualization Security: Analysis and Open Challenges", International Journal of Hybrid Information Technology, vol. 8, no. 2, **(2015)**, pp. 237-246.
[27] M. Arif, K. Amjad Alam and M. Hussain, "Application of data mining using artificial neural network: Survey", International Journal of Database Theory and Application, vol. 8, no. 1, **(2015)**, pp. 245-270.
[28] Z. Ahmed, "A Comparative Study for Ontology and Software Design Patterns", International Workshop Soft Computing Applications, Springer, Cham, **(2016)**.
[29] A. Ahmed, "MainIndex Sorting Algorithm", International Workshop Soft Computing Applications, Springer, Cham, **(2016)**.
[30] A. Ahmed, "A Smart Way to Improve the Printing Capability of Operating System", International Workshop Soft Computing Applications. Springer, Cham, **(2016)**.
[31] M. Arif, "Maximizing Information of Multimodality Brain Image Fusion Using Curvelet Transform with Genetic Algorithm", Computer Assisted System in Health (CASH), 2014 International Conference on IEEE, **(2014)**.

Author

Muhammad Arif is a PhD student at Faculty of CS and IT, University of Malaya. Currently he is working on Medical image Processing. His research interests include image processing, E learning, Artificial intelligence and data mining. He joined UM as a Bright Spark Scholar in September 2013 for the period of 3 years. Before this he completed masters and bachelor degrees in Pakistan. He received his BS degree in Computer Science from University of Sargodha, Pakistan in 2011. He obtained his MS degree in Computer Science from COMSATS Islamabad 2013 Pakistan.

