International Journal of Innovative Technologies
WWW.IJITECH.ORG
ISSN 2321-8665
Vol.05,Issue.01,
January-2017,
Pages:0079-0084

Anti Hub Distance Based Unsupervised Outlier Detection

K. NAGA LAKSHMAN[1], SD. SHAREEF[2]
[1]Research Scholar, Eswar College of Engineering, Narasaraopet, Guntur, AP, India.
[2]Assistant Professor & Research Supervisor, Eswar College of Engineering, Narasaraopet, Guntur, AP, India.

Abstract: As the dimensionality of the data increases due to this all point becomes good outlier the distance based outlier detection methods fails. Reason of these issues is irrelevant and redundant features; nearest neighbor of the point P is K points whose distance to point P is less than all other points. Reverse nearest neighbors (RNN) of Point P is the points for which P is in their k nearest neighbor list. Some points are frequently comes in k-nearest neighbor list of another points referred as hubs and some points are infrequently comes in k nearest neighbor list of different points are called as Anti-hubs. Recent research proposes anti-hub based unsupervised outlier detection methods but these propose are suffered from computation cost of finding anti-hubs. In the case of data which has outstanding dimensionality, computation cost and time requirement to find anti-hubs is high. There is need to remove the redundant features if high dimensional data contains redundant attributes. Reduce the computation cost and time requirement by removal of redundant features to find anti-hubs for outlier detection. For extending anti-hub based outlier detection method for high dimensional data apply feature selection.

Keywords: High-Dimensional, Data Outlier Detection, Reverse Nearest Neighbors.

I. INTRODUCTION

Outlier detection is studied widely in the survey because need of searching intrusion detection and anomaly detection in many applications. There are three main types of outlier detection methods namely, unsupervised, semi-supervised and supervised. These types are divided by labels of instances on which outlier detection is to be applied. Need availability of correct labels of the instances for supervised and semi-supervised outlier detection. For outlier detection availability of labels is not practically possible therefore unsupervised technique is used widely which does not need label to the instances. Most popular and effective method for unsupervised outlier detection is distance based outlier detection [1]. Distance based outlier detection consider that normal instances have small distance among them and outliers have large distance from normal instances. V. Hautamaki et al [2] stated that as the dimensions of the data raises, distances turn useless to find outliers because each point seems as outlier. Unsupervised outlier detection confronts some challenges in high-dimensionality. Regardless of the common notion that all points in a high-dimensional data-set seem to turn outliers, Milos Radovanovic et al [20] showed that unsupervised methods can detect outliers under the assumption that all (or most) data attributes are purposeful, i.e. not noisy. The relation between the high dimensionality and outlier nature of the instances investigates by Milos Radovanovic et al [20]. K-nearest neighbor of the point P is K points whose distance to point P is less than all other points. Reverse nearest neighbors (RNN) of Point P is the points for which P is in their k nearest neighbor list. Some points are frequently comes in k -nearest neighbor list of other points and some points are infrequently comes in k nearest neighbor list of some other points are called as Anti-hubs.

Density Based Local Identifiers (LOF) [9] its variants are proposed in literature. Also Angle-Based Outlier Detection is available in the literature [10]. For outlier detection RNN concept is used in literature [2] [4], but there is no theoretical proof which explores the relation between the outlier natures of the points and reverse nearest neighbors. Gustavo H. Orair et al[6] stated that reverse nearest count is get affected as the dimensionality of the data increases, so there is need to investigate how outlier detection methods bases on RNN get affected by the dimensionality of the data. Milos Radovanovic et al [20] discusses

- In high dimensionality the problems in outlier detection and shows that how unsupervised methods can be used for outlier detection.
- How Anti-hubs are related to outlier nature of the point is investigates.
- For outlier detection Based on the relation anti-hubs and outlier two methods are proposed for high and low dimensional data for showing the outlierness of points, beginning with the method ODIN (Outlier Detection using in-degree Number).

In existing system it takes large computation cost, time to calculate the reverse nearest neighbors of the all points. Use of antihubs for outlier detection is of high computational task. Computation complexity increases with the data dimensionality. For this there is scope to removal of irrelevant features before application of Reverse Nearest Neighbor. So to overcome this problem, feature selection is applied on the data. In this step, all features are rank according to their importance and required features are selected for finding reverse nearest neighbors. To find reverse nearest neighbor using Euclidean distance and outlier score is calculated by

using technique from existing system. According to studies, if system does not know about the distribution of the data then euclidean distance is the best choice. Proposed scheme deals with curse of dimensionality efficiently. We discussed existing system, problem statement and proposed scheme with detailed structure and algorithms.

II. LITERATURE SURVEY

The problem arises due to increase in dimensionality of the data the problems arises due to increase in dimensionality of the data investigated by M. E. Houle et al [1]. Poor discrimination was caused by presence of the redundancy of attributes, presence of the irrelevant features and concentration. These issues reduce the usability of the similarity and distance measures. They evaluated that secondary measures like shared-neighbor were still useful in such condition. V. Hautamaki et al [2] used reverse nearest neighbor count is to score outlier nature of the point. User defined threshold was used to take decision about outlier nature of the point. Method proposed in this paper is named as Outlier Detection using Indegree Number (ODIN). If score is less than threshold then the point is said to be an outlier otherwise it is normal point. The link between the reverse nearest neighbor count and outlier nature of the point investigated by V. Hautamaki et al. J. Lin et al [3] proposed special case of ODIN [2] where point was considered outlier if reverse nearest neighbor count of the point is zero. They does not provide any mathematical explanation or proof why point which has reverses nearest neighbor count is outlier. They mainly focused on the speed and scalability.

The method to find reverses nearest neighbor of the point in metric spaces described by Y. Tao et al [4]. Proposed algorithms do not necessitate representation of the instances i.e. objects. Proposed technique uses metric index therefore it affirms by recurring to the insertion/deletion operations of the index. C. Lijun et al [5] explored the relation between outlier and RNN but there was no research study how high dimensionality was connected with reverse nearest neighbors. They focused on data stream application and reducing execution time for finding reverse nearest neighbor of point. Outlier detection was the process of discovering observations which noticeably deviates from other observations and also it was a fundamental approach in data analysis task described by Gustavo H. Orair et al [6]. Applications range from financial fraud detection to clinical diagnosis of diseases and network intrusion detection. They described and evaluated several distance based outlier detection approaches. They presented the study to understand the impact of optimizations strategies and tried to consolidate them. K. S. Beyer et al [7] tried to finding the effective answers for the problem of nearest neighbor. This problem is specified as, finding the data point that was closest to the query point by giving an aggregation of points of data and a query point in a multidimensional metric space. They analyzed the effects of dimensionality on Nearest Neighbor queries. They observed that as there is increase in the dimensionality, the distance to the neighbor advances to the distance to the farthest neighbor. Conducted the experiments to find out the proportion at which the NN breaks down and also explored the situations where even on dimensionality NN queries do not break down low dimensions and LB-ABOD suitable for high dimensional data.

Numeral data analysis tools and nearest neighbor search mostly based on the use of euclidean distance describe by D. Franc et al [11]. In case of broad dimensionality, though all distances amongst different couple of date elements appears similar; the euclidean distance appear to concentrate. Therefore the distance's relevancy has been doubted in the past, and fractional norms were brought to overcome this problem. They suggested the use of alternative distances to agitate the concentration. For the purpose of large spatial databases A. Nanopoulos et al[12] introduced a clustering algorithm C2P, which uses techniques of spatial access for the purpose of determining closest pairs and also introduced the extensions for scalable clustering in huge databases containing clusters of outliers and different shapes. The proposed algorithm has advantages of the hierarchical clustering and graph theoretic algorithms which give the efficiency. A method for judging outlier-ness and this method is named as Local Correlation Integral (LOCI) proposed by S. Papadimitriou et al [13]. LOCI are highly effective with best previous methods for detection of outliers and group of outliers. It also extends an automatic data- dictated cut off to find out whether a point was an outlier or not. Yunjun Gao et al [14] studied a new form of nearest neighbor queries which is called as Mutual Nearest Neighbor (MNN) search in a spatial database. But existing spatial query processing approaches cannot handle MNN queries with effectiveness. They introduced a work for dealing with MNN queries efficiently.

W. Jin et al [15] found local outlier's needs estimation of density distribution which is founded on density distribution of its k-nearest neighbors. But results may be wrong when outliers in the location where there is different density distribution in the neighborhood. To tackle this, they introduced a measure which considers both neighbors and reverse neighbors of an object. Hub ness was caused because of huge dimensionality problem intrinsic nearest neighbor methods presented by N. Tomasev et al [16]. For exploiting the hub ness process they presented new approach in k-nearest neighbor classification. They introduces an algorithm named, Hun ness Information k-nearest Neighbor (HIKNN), this algorithm introduced the k occurrence informative-ness into the hub ness aware k-nearest neighbor voting framework. Formalized view of study which is useful for theoretical comparison of many existing methods described by E. Schubert et al [17]. The provided view improved the ability of interpreting the differences of outlier detection models and shared properties. The presented model alleviates the expression of abstract framework for many special data types which requires specialized algorithms to deal with them. Some algorithms which are used recently describe by C. C. Aggarwal et al[18]. In order to find the outliers based on outlier's relationship, this algorithm used concepts of closeness to the rest of the data. Still, in high dimensional

space, the data is sparse and the impression of closeness i.e. proximity fails to hold back its significance. In fact, the scarcity of multidimensional data entails that from the view of definitions which are based on proximity, every point is an almost equally good outlier.

At the time of comparing clustering results, metric which is used for evaluation metric decomposes the available entropy i.e. information to a single number describe by E. Achtert et al [19]. However, usable metrics for evaluation are not always agreeable and are hard to explain in evaluating the correspondence of a pair of clustering. For the purpose of comparing multiple clustering, authors provided the tool to visually support the judgment of clustering results. Milos Radovanovic et al[20] discoursed issues in outlier detection in the case of eminent data dimensionality and showed the way outlier detection in high dimensional data can be made using unsupervised methods. It also enquires how Anti-hubs are associated to the point's outlier nature.

III. SYSTEM ARCHITECTURE

A. Existing System

- From set of instances i.e. outlier detection existing system consist of the process of finding irregular instances and it aims at make the use of outlier detection in finding intrusion detection and anomaly detection in many applications.
- Existing system discussed the issues in outlier detection in high dimensionality and shows that how unsupervised methods can be used for outlier detection in high dimensional data.
- It also investigated how Anti-hubs are related to outlier nature of the point and Based on the relation anti-hubs and outlier, two ways of using k-occurrence information are proposed for outlier detection for high and low dimensional data for showing the outlier-ness of points, beginning with the method ODIN (Outlier Detection using in-degree Number).

Limitation Of Existing System:

- In existing system it takes high computation cost, time to calculate the reverse nearest neighbors of the all points.
- Use of Antihubs for outlier detection is of high computational task
- Computation complexity increases with the data dimensionality.

B. Proposed System

Proposed system is designed for removing the drawback of exiting system. Proposed system consists of following steps as follows:-

Feature Selection: To deal with the Curse of dimensionality proposed system is designed. It takes high computation cost, time to calculate the reverse nearest neighbors of the all points in existing system. Feature selection is applied on the datato overcome this problem. In this step, all features are rank according to their importance and required features are selected for finding reverse nearest neighbors. Importance of the feature is calculated using the Mutual Information (MI) measure. Mutual Information is one most important feature which calculates the mutual dependence between two features. The mutual information between feature A and feature B calculated by Equation 1 where PRBR (b).PRAR (a) is marginal probability distribution and PRABR (a, b) is joint probability distribution. To calculate the MI of A, sum of MI of A with all other features is taken,

$$MI(A) = \qquad (MI\ (A,i)) \tag{1}$$

After calculation of MI values of all features, features with MI values less than threshold values are discarded from further process.

Find Reverse Nearest Neighbor: In this step, data of selected features will be considered for finding the reverse nearest neighbor. To determine the reverse nearest neighbor, first k-nearest neighbors of each point is evaluated. Existing system used euclidean measure for calculating the distance between two

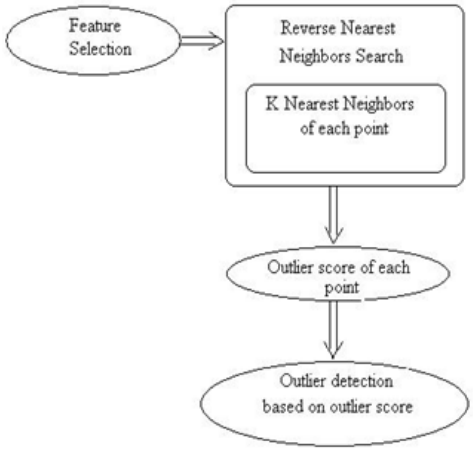

Fig. 1. System Architecture.

instances. Euclidean distance measure works fine for two and three dimensional data but is gets negatively affected with high dimensionality. According to studies, if system doesn't know about the distribution of the data then euclidean distance is the best choice. Number of occurrences of point P in the k nearest neighbor list of the all other points is called as k-occurrence. Points in the dataset for which point's P is k-nearest neighbor are reverse nearest neighbor for point P. From the k-nearest neighbor list of each point, reverse nearest neighbor list of each point is calculated.

Outlier Score of Each Point: Previous methods than existing system considered k-occurrence of the point as an outlier score. Less k-occurrence indicates more outlier score of the point. Proposed system will follow existing system to calculate the outlier score of the point. Sum of k-occurrence score of k-nearest neighbors of the point P is outlier score of the point P.

Outlier Score (P) =(koccurence(pi))

where pi is the nearest =0 point of point P. If Outlier scores (P) is larger than the threshold then Point P is *h* considered as outlier.

IV. IMPLEMENTATION DETAILS

A. Algorithm 1

Algorithm: AntiHub 2 with feature selection

It works under the following
stages 1: Select features
2: Computation of mutual dependence of two random variable using equation

MI (A, B) = (.)

Equation 1 calculates the mutual information between feature A and feature B. where $P_A(a).P_B(b)$ is marginal probability distribution and $P_{AB}(a,b)$ is joint probability distribution

3: Then MI of one feature with all other features is computed using the relation:

$MI_{fi} = MI_{fj}$ ..... (2)

Where i,j=1,2,.....ft with i!=j and ft is total number of features
4: Then MI of each feature is use to rank the feature
5: a= $AntiHub_{dist.}$ (D,k)
6: For each i ϵ (1,2, ...,n)
7: anni =$\sum_i \epsilon$ $NN_{dist.}$(k, i)a_i .where $NN_{dist.}$ (k, i) is the set of indices of k nearest neighbors of x_i
8: disc=0
9: For each α (0, step, 2*step ....1)
10: For each i ϵ (1,2 ... n)
11: ct_i= (1-α).a_i + α.ann_i
12: cdisc = discScore (ct, p)
13: If cdisc> disc
14: t= ct, disc = cdisc
15.For each i ϵ (1, 2... n)
16. s_i= f(t_i) where f : R ⟶ R is a monotone function

B. Mathematical Model

Let, S be Anti hub based fast unsupervised outlier detection scheme having Input, Processes and Output it can be represented as, S = (I, P, O) Where, I, is a set of inputs given to the System, O is a set of outputs given by the System, P is a set of processes in the System.

I = (I1, I2, I3, I4)
I1- is set of input data D with m number of features with n number of instances.
I2- k for knn
I3- Mutual Information threshold
I4- Outlier score threshold
P= (P1, P2, P3, P4, P5, P6, P7)
P1- Find the Mutual Information between two random variables A and B
PRAR (a) is marginal probability distribution and
PRABR (a, b) is joint probability distribution
Output will be O1
P2 - Find Mutual Information of Feature
Features with high MI than threshold MI is selected for farther process
If MI (Ai) ≥Threshold MI
Then Select Ai Else discard Ai

P3 - To find the distance between two instances Euclidean distance is used

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

Where d (p,q) is Euclidean distance between p and q points , both points has n dimensions

Output will be O3
P4 - Find k-nearest neighbor of each point Knn (P) = (p1, p2, p3, p4, ….., pk)
List of k nearest neighbor points is calculated. P5 - Find RNN list of each point

RNN (P) = Set of points for which P is in their knn list P6 - Outlier score of each point

Outlier Score (P) = =0 (koccurrence (pi)) Where k indicates k nearest neighbors of point p P7 - Outlier detection

If Outlier Score (P) ≥ threshold then P is outlier O1 - List of MI of among all features in D
O2 - List of selected features
O3 - Euclidean distance
O4 - List of list of knn points for each point O5 - List of RNN of each point is calculated O6 - List of outlier score of each point
O7- List of outliers

C. Experimental Setup

The scheme is implemented using Java framework (version jdk 1.8) on Windows platform. The Net bean IDE (version 8.0.2) is applied as a development tool. The scheme doesn't need any particular hardware to run; any standard machine can be able to run the application.

V. EXPERIMENTAL RESULTS

The reason of the conducting experiments is to check the effect of feature selection before anti-hub based outlier detection on high dimensional data. To see the effectiveness accuracy, memory and time requirement of antihub based outlier detection i.e. Antihub2 [20] and Proposed method is compared. For experiment purpose, we used KDD dataset. Dataset contains 1050 instances, 42 attributes and 1.456% outliers. Minor class category considered as outlier class. Table 1 shows the actual results.

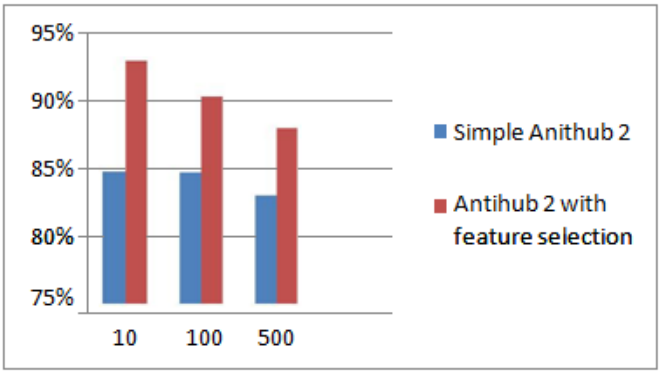

Fig. 2. Accuracy comparison with k variation.

TABLE I: Accuracy Comparison with K Variation

K	Simple Anithub 2	Antihub2 with feature selection
10	84.79 %	93.40%
100	84.70 %	90.34 %
500	83.37 %	88.04 %

TABLE II: Time Comparison with K Variation

K	Simple Anithub 2 time in sec.	Antihub2 with feature selection time in sec.
10	1.31	1.21
100	1.42	1.37
500	1.34	1.29

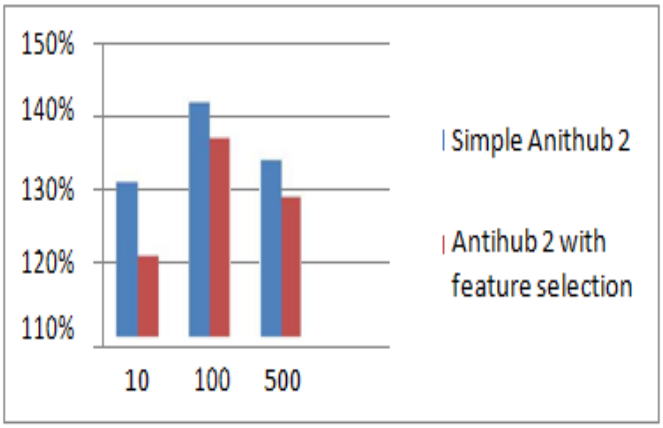

Fig. 3. Time comparison with k variation.

TABLE III: Memory Comparison With K Variation

K	Simple Anithub 2 memory in byte	Antihub2 with feature selection memory in byte
10	8.75	8.23
100	7.73	7.68
500	9.37	9.13

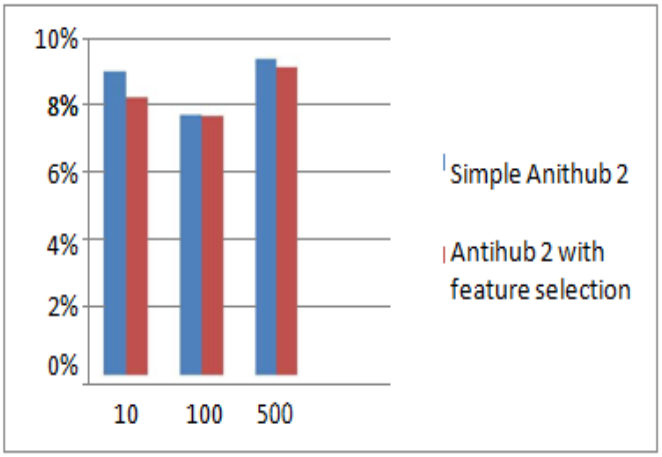

Fig. 4. Memory comparison with k variation.

Consider Feature selection selects 25 dimensions from 38 dimensions. If existing system needs 1 unit time to process all 28 features then proposed system will required 0.65 unit time. Same as time, memory requirement will be less than existing system.

VI. CONCLUSION

Outlier detection is studied widely because need of finding intrusion detection and anomaly detection in many applications. Exiting method proposed reverse nearest neighbor outlier detection using anti-hubs. But using anti hub for outlier detection is of high computational task. Computational complexity increases with the data dimensionality to avoid this removal of irrelevant features before application of reverse nearest neighbor is introduced. This reduces computational task and improves the efficiency of finding anti-hub and also enhances the anti-hub based unsupervised outlier detection. From actual results it is clear that proposed system inproves the accuracy and also reduces the time and memory requirement for outlier detection.

Future Scope: In future, we enhance the proposed system to handle high dimensional data and high computation complexity for better experimental results, to make an efficient intrusion and anomaly detection system.

VII. REFERENCES

[1]M. E. Houle, H.P.Kriegel, P. Kroger, E. Schubert, and A. Zimek, “Can shared-neighbor distances defeat the curse of dimensionality?,” inProc 22nd Int. Conf. Sci. Statist.DatabaseManage., 2010, pp. 482-500.

[2]V. Hautamaki, I. Karkkainen, and P. Franti,“Outlier detection using k-nearest neigh-bour graph,” in Proc 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 430-433.

[3]J. Lin, D. Etter, and D. DeBarr,“Exact and approximate reverse nearest neighbor search for multimedia data,” in Proc 8th SIAM Int. Conf. Data Mining, 2008, pp. 656-667.

[4]Y. Tao, M. L. Yiu, and N. Mamoulis, “Reverse nearest neighbor search in metric spaces,”IEEE Trans. Knowl. Data Eng.,vol. 18, no. 9, pp. 1239-1252, Sep. 2006.

[5]C. Lijun, L. Xiyin, Z. Tiejun, Z. Zhongping, and L. Aiyong, “A data stream outlier delection algorithm based on reverse k nearest neighbors,” in Proc. 3rd Int. Symp. Comput.Intell.Des., 2010, pp. 236-239.

[6]Gustavo H. Orair, Carlos H. C. Teixeira, Wagner Meira Jr., Ye Wang and Srinivasan-Parthasarathy , “Distance-Based Outlier Detection: Consolidation and Renewed Bear-ing,”2010.

[7]K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, ”When is nearest neighbor meaningful?,”in Proc. 7th Int. Conf. Database Theory, 1999, pp. 217-235.

[8]E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, ”A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data,”in Proc. Conf. Appl. Data Mining Comput. Security, 2002, pp. 78-100.

[9]M. Breunig, H.P.Kriegel, R. T. Ng, and J. Sander, ”LOF: Identifying density-based local outliers,” SIGMOD Rec., vol. 29, no. 2, pp. 93-104, 2000.

[10]Hans-Peter Kriegel,Matthias Schubert and Arthur Zimek, ”Angle-Based Outlier De-tection in High-dimensional Data,” 2008.
[11]D. Francois, V. Wertz, and M. Verleysen, ”The concentration offractional distances,”IEEE Trans. Knowl. Data. Eng., vol. 19, no. 7,pp. 873-886, Jul. 2007.
[12]A. Nanopoulos, Y. Theodoridis, and Y. Manolopoulos, ”C2P:Clustering based on closest pairs,” in Proc 27th Int. Conf. VeryLarge Data Bases, 2001, pp. 331-340.
[13]S. Papadimitriou, H. Kitagawa, P. Gibbons, and C. Faloutsos, ”LOCI: Fast outlier detection using the local correlation integral,”in Proc 19th IEEE Int. Conf. Data Eng., 2003, pp. 315-326.
[14]YunjunGao, BihuaZheng, ”On reverse nearest neighbor queries,”2002.

Author’s Profile:

K.Naga Lakshman is a student pursuing M.Tech (CSE) in Eswar College of Engineering, Narasaraopet, Guntur, India.

Sd.Shareef M.Tech, , is having 05+ years of experience in the field of teaching in various Engineering Colleges . At present he is working as Asst. Prof. in Eswar College of Engineering, Narasaraopet, Guntur, India. He published 1 international journals