ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DETECTION OF OUTLIERS BY MAKING DISTANCE-BASED METHOD

Dr.K.V.Ranga Rao[*1] & Dr.B.Vijayakumar[2]
[*1]Professor,CSE Dept,Vidya jyothi institute of technology,Aziz nagar,c.b.post.
[2]Professor and HOD,CSE Dept,Vidya jyothi institute of technology,Aziz nagar,c.b.post a

DOI: 10.5281/zenodo.1199285

ABSTRACT

Outlier detection in high-dimensional data presents various challenges resulting from the curse of dimensionality. A prevailing view is that distance concentration, the tendency of distances in high-dimensional data to become indiscernible, hinders the detection of outliers by making distance-based methods label all points as almost equally good outliers. In this paper, we provide evidence supporting the opinion that such a view is too simple, by demonstrating that distance-based methods can produce more contrasting outlier scores in high-dimensional settings. Furthermore, we show that high dimensionality can have a different impact, by reexamining the notion of reverse nearest neighbors in the unsupervised outlier-detection context. Namely, it was recently observed that the distribution of points' reverse-neighbor counts becomes skewed in high dimensions, resulting in the phenomenon known as hubness. We provide insight into how some points like antihubs appear very infrequently in k-NN lists of other points, and explain the connection between antihubs, outliers, and existing unsupervised outlier-detection methods. By evaluating the classic k-NN method, the angle-based technique designed for high-dimensional data, the density-based local outlier factor and influenced outlierness methods, and antihub-based methods on various synthetic and real-world data sets, we offer novel insight into the usefulness of reverse neighbor counts in unsupervised outlier detection.

Keywords: Aantihub,Distance-Based method, K-nn lists.

I. INTRODUCTION

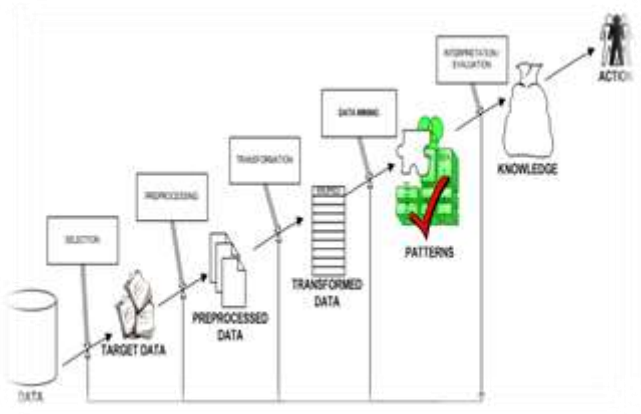


Figure1:Structure of Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Data mining major elements:

1) Extract, transform, and load transaction data onto the data warehouse system.
2) Store and manage the data in a multidimensional database system.
3) Provide data access to business analysts and information technology professionals.
4) Analyze the data by application software.
5) Present the data in a useful format, such as a graph or table.

II. EXISTING SYSTEM

The task of detecting outliers can be categorized as supervised, semi-supervised, and unsupervised, depending on the existence of labels for outliers and/or regular instances. Among these categories, unsupervised methods are more widely applied because the other categories require accurate and representative labels that are often prohibitively expensive to obtain. Unsupervised methods include distance-based methods that mainly rely on a measure of distance or similarity in order to detect outliers. A commonly accepted opinion is that, due to the "curse of dimensionality," distance becomes meaningless, since distance measures concentrate, i.e., pair wise distances become indiscernible as dimensionality increases. The effect of distance concentration on unsupervised outlier detection was implied to be that every point in high-dimensional space becomes an almost equally good**.**

III. PROPOSED SYSTEM

It is crucial to understand how the increase of dimensionality impacts outlier detection. As explained in the actual challenges posed by the "curse of dimensionality" differs from the commonly accepted view that every point becomes an almost equally good outlier in high-dimensional space. We will present further evidence which challenges this view, motivating the (re)examination of methods.Reverse nearest-neighbor counts have been proposed in the past as a method for expressing outlierness of data points but no insight apart from basic intuition was offered as to why these counts should represent meaningful outlier scores. Recent observations that reverse-neighbor counts are affected by increased dimensionality of data warrant their re examination for the outlier-detection task. In this light, we will revisit the ODIN method.Demonstration of one plausible scenario where the methods based on antihubs are expected to perform well, which is in a set-ting involving clusters of different densities. For this reason, we use synthetic data in order to control data density and dimensionality.

IV. SYSTEM REQUIREMENT:

It specifies the hardware and software requirements that are required in order to run the application properly. The Software Requirement Specification (SRS) is explained in detail, which includes overview of this dissertation as well as the functional and non-functional requirement of this dissertation

SRS for Supporting Privacy Protection in Personalized Web Search

Functional	Admin login by using valid user name & password, admin can view the user details, add contents, providing general key & personalized key to users, admin can recover the attacked contents, Data sharing between Admin to user, user login by using authorized user name & password, User can search contents, user can request general key & personalized key, End user authentication by admin, finding the attackers.
Non- Functional	Admin never monitors the user activities
External interface	LAN , WAN
Performance	Admin login, User login, add contents, view all content details, search contents, providing general key & personalized key, view search history, recover contents, view users, search based on query, request, response, view user details, view attacker details.
Attributes	Privacy protection, personalized web search, utility, risk, profile, profile based personalization, Admin, users.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

V. SYSTEM SPECIFICATIONS

Hardware Requirements:

- System : Pentium IV 2.4 GHz
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb
- Monitor : 14' Colour Monitor.
- Mouse : Optical Mouse.
- RAM : 512 MB

Software Requirements:

- Operating system : Windows XP or Windows 7, Windows 8.
- Coding Language : Java and J2EE
- Data Base : My Sql .
- Documentation : MS Office
- IDE : Netbeans 7.4
- Development Kit : JDK 1.6

VI. IMPLEMENTATION

1. Modules

Admin module:
In this module, the Admin has to login by using valid user name and password. After login successful he can do some operations such as add contents, view all contents, list all searching history, list ranking of images, list of all personalized search, attacker details, recover contents, list of all user and logout.

Add contents:
In this module, the admin can add n-number of contents. If the admin want to add a new content, then admin will enter a URL, domain, title, description, uses, related images of the particular content ,then submit and that data will stored in data base. If admin want view to the newly added content, then click on view contents button, it will display the all contents & with their tags, the initially rank will be zero.

List of users:
In this module, the Admin can view list of all users. Here all register users are stored with the details such as user ID, user name, E mail ID, mobile no, Location, date of birth, address, pin code, general key and personalized key.

View list all searching history:
This is controlled by admin; the admin can view the all searching history. If admin clicks on search history button, then the server will display the all searching history with their tags such as user name, key word used, field searched, time & date.

Attacker details:
In this module, the admin can view the attacker details. If admin clicks on attacker details button, the admin will get attacker information with their tags such as attacker name, attacked content URL and attacked content ID. After attacking content, the admin will recover the content.

2. User module

In this module, there are n numbers of users are present. User should register before doing some operations. After registration successful he has to login by using authorized user name and password. Login successful he will do some operations such as view my details, query search, personalized search, personalized search comparisons, attack content details, request for general key, request for personalized key and logout. If user clicks on my details button, then the server will give response to the user with their tags such as user ID, name, mobile no, address, pin code and email ID.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

Query Search:
In this module, the user can search query. Before searching any query, the user should request general key, then admin will provide a general key. Then enter general key, select field to search, enter key word and search, it will display all related contents with their tags. After searching a content rank will be increased.

Personalized Search:
In this module, the user can search contents. Before searching contents, the user should request personalized key, then admin will provide personalized key, then enter key and enter keyword, then user will get a related contents with their tags. After searching content the rank will be increased.

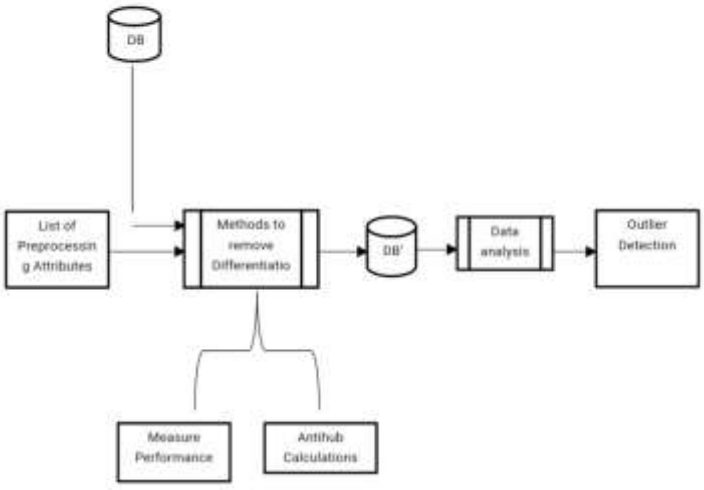
Personalized Search Comparison:
In this module, the user can view the comparison between greedy DP & greedy IL. After personalized searching, the greedy IL will be generated. If the user clicks on personalized search button, it will display all personalized search details with their tags such as user name, keyword used, date, time, using greedy DP and using greedy IL.

3. Time delay Generation chart module:
In this module, we can view the time delay Generation chart results. This chart shows the time delay by using greedy DP and time delay using greedy IP. After viewing or search the content, rank will be increased and also the time delay will be display, the time variation can be shown in this chart.

4. Attack content module:
In this module, user can attack contents, and then user should enter content URL to attack, then user will get all information about content, then user can add malicious data and click on attack button. After attacking successful, the attacker details will send to admin.

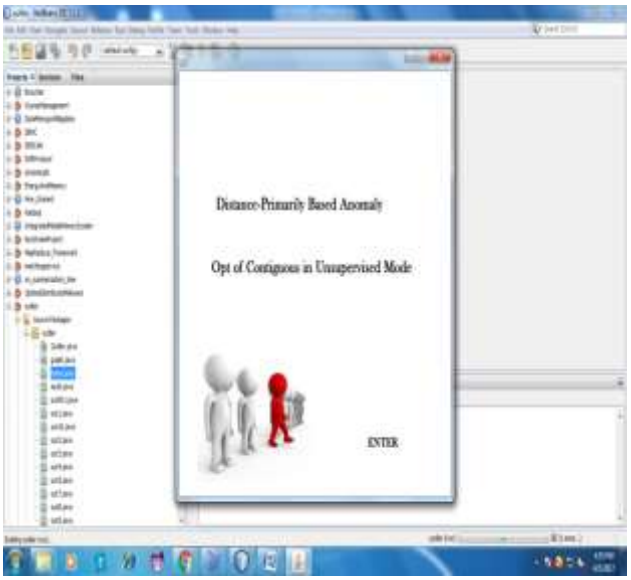
VII. BLOCK DIAGRAM

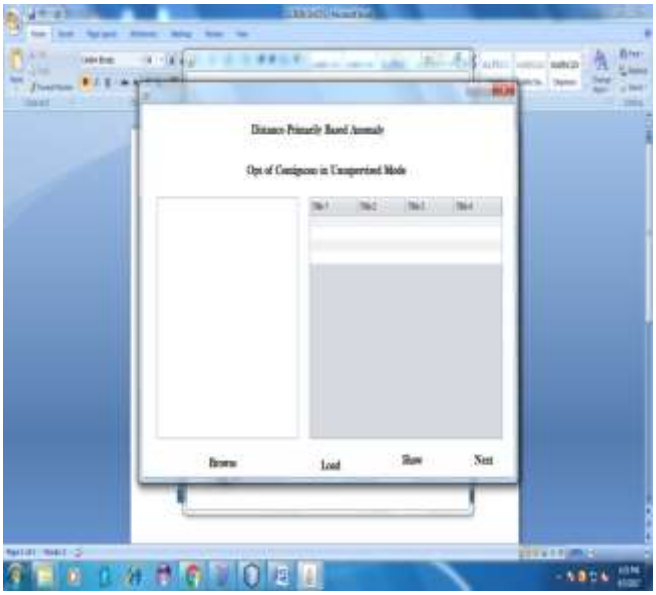
BLOCK DIAGRAM:



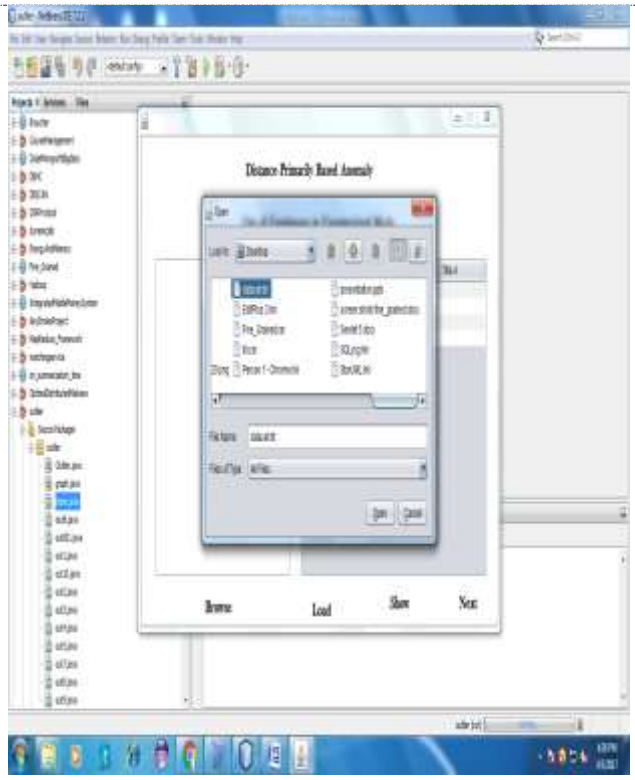
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

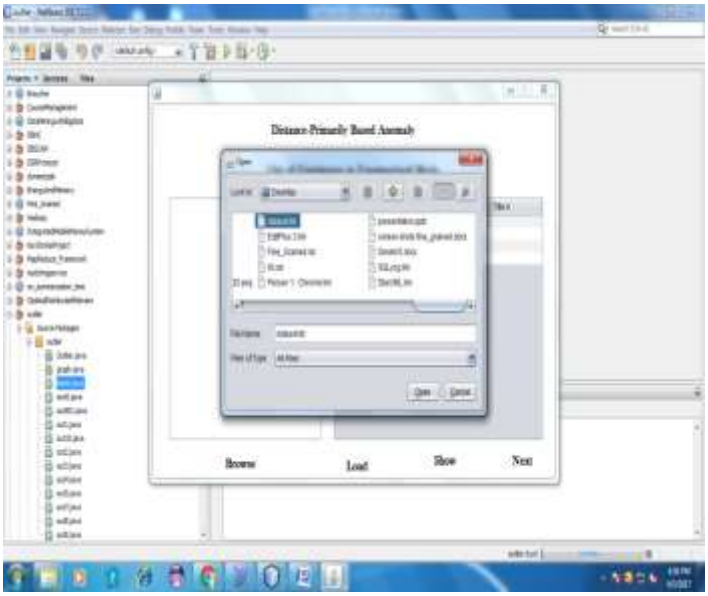
VIII. SCREEN SHOTS



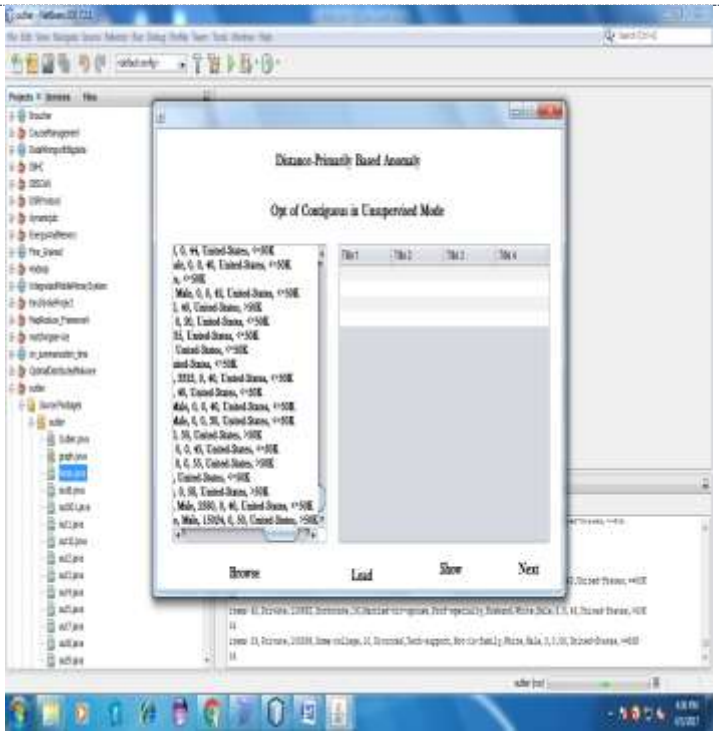


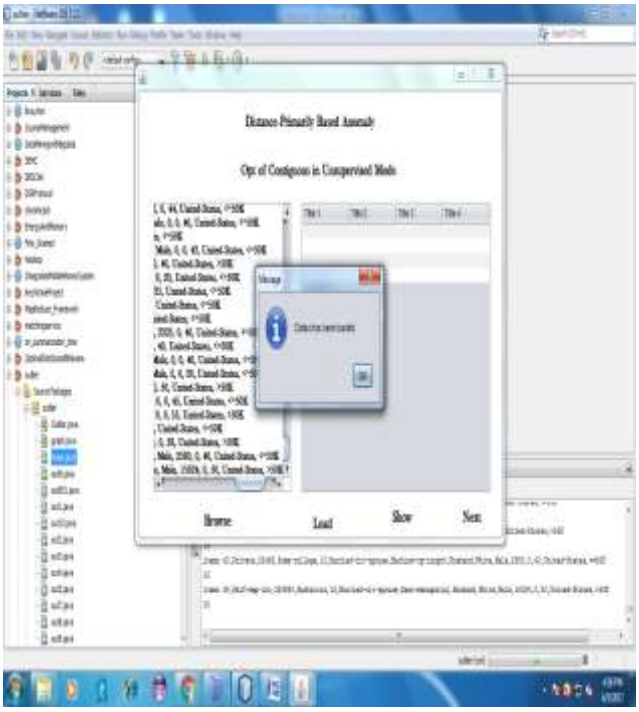
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



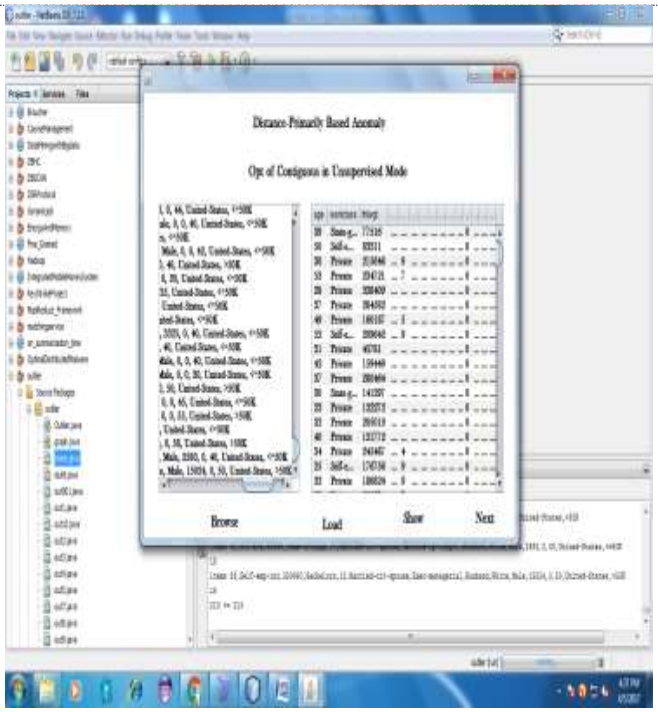


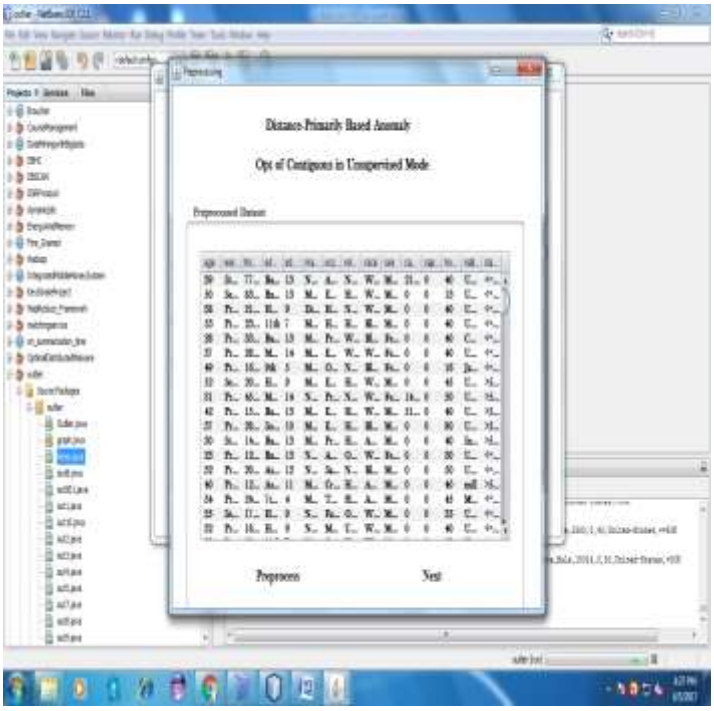
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



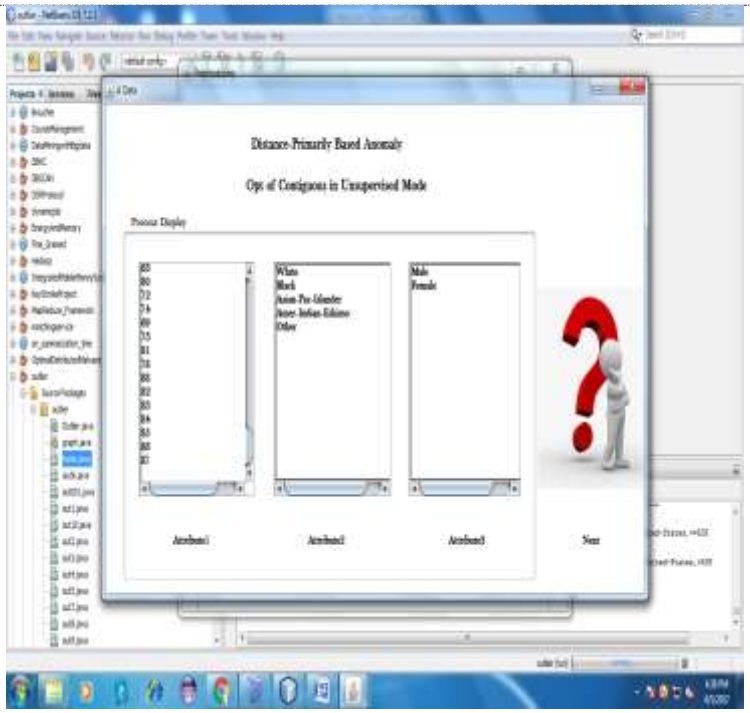


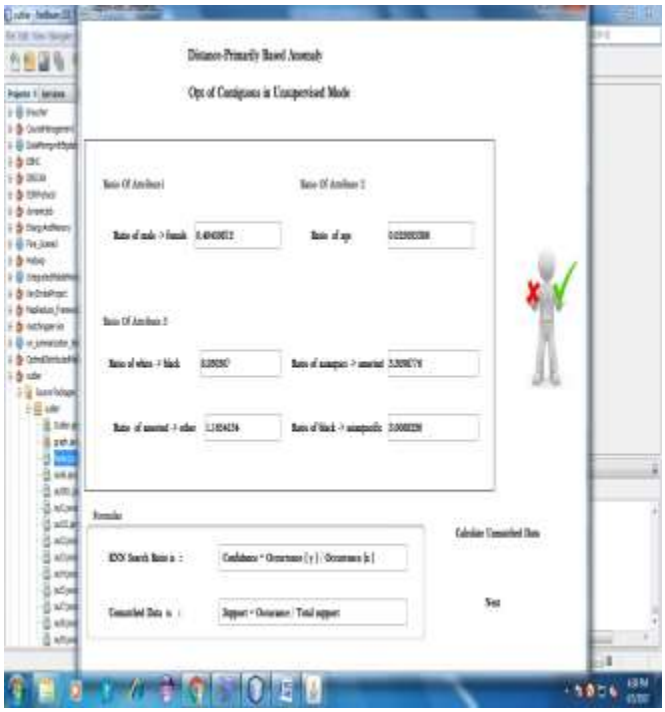
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



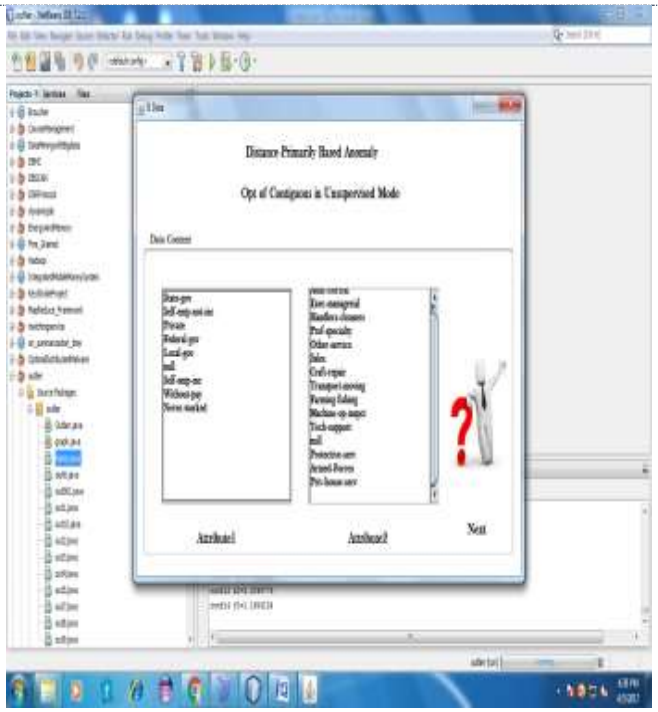


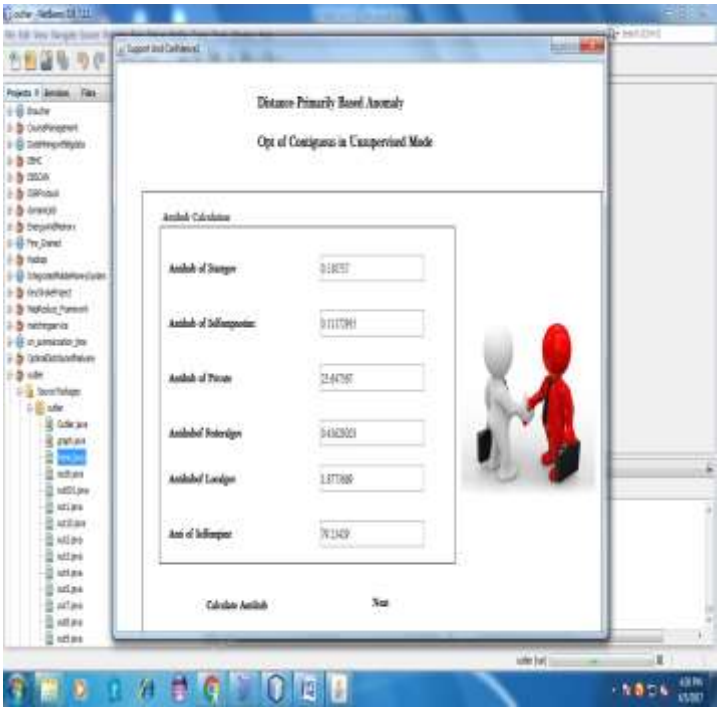
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



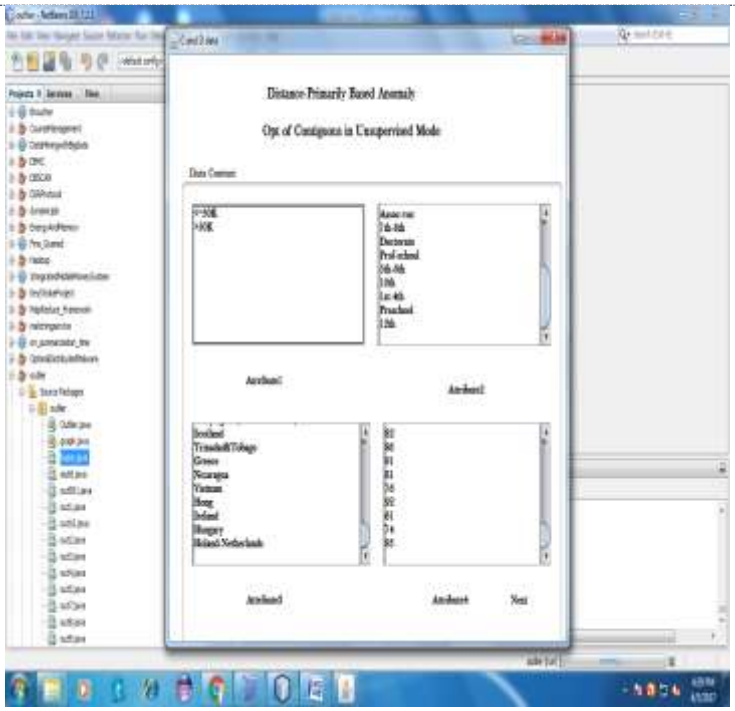


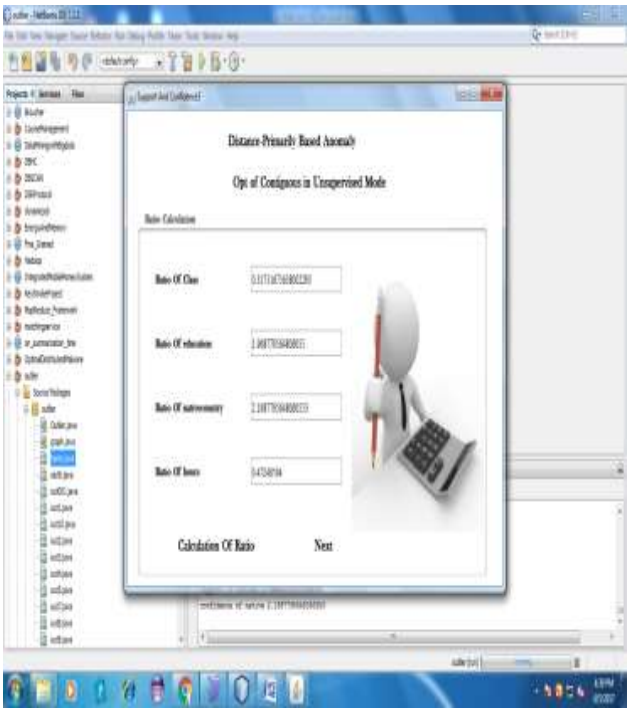
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



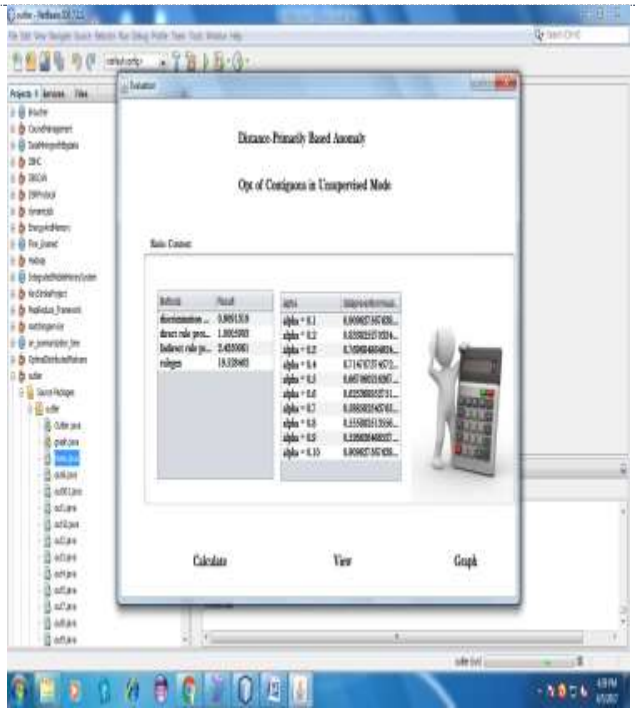


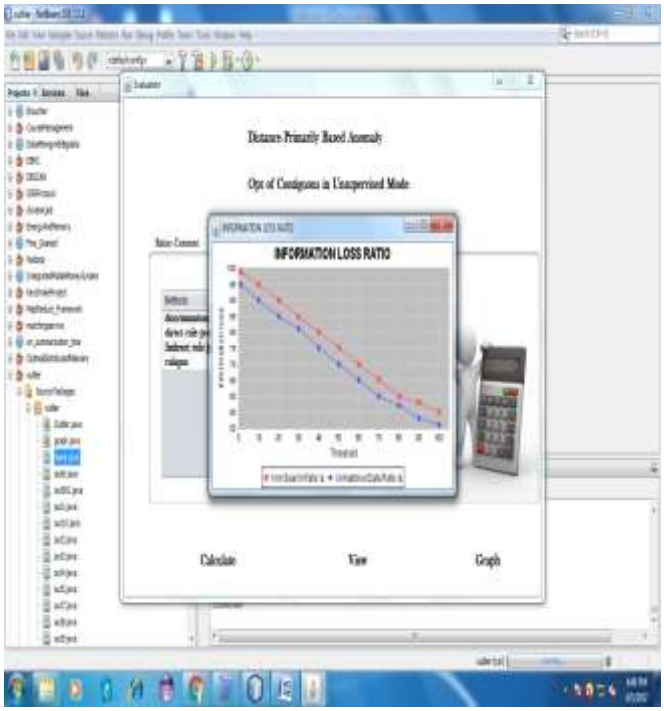
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



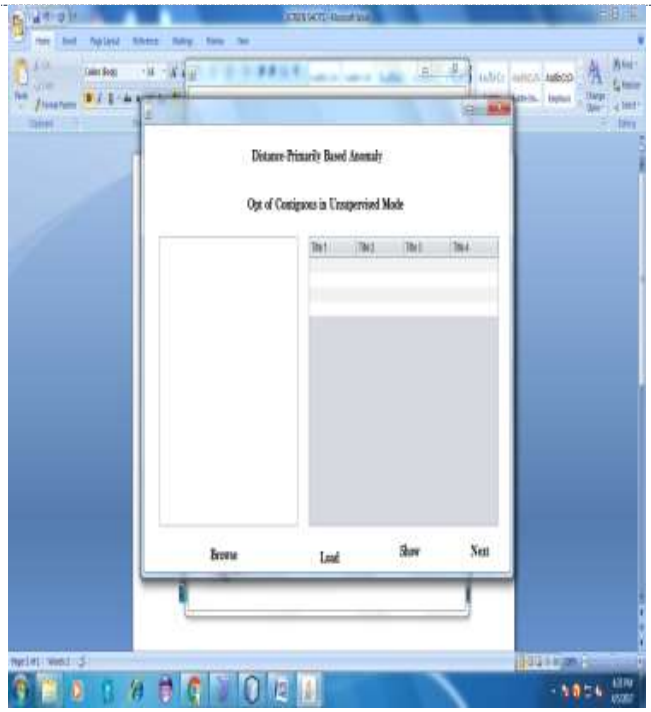


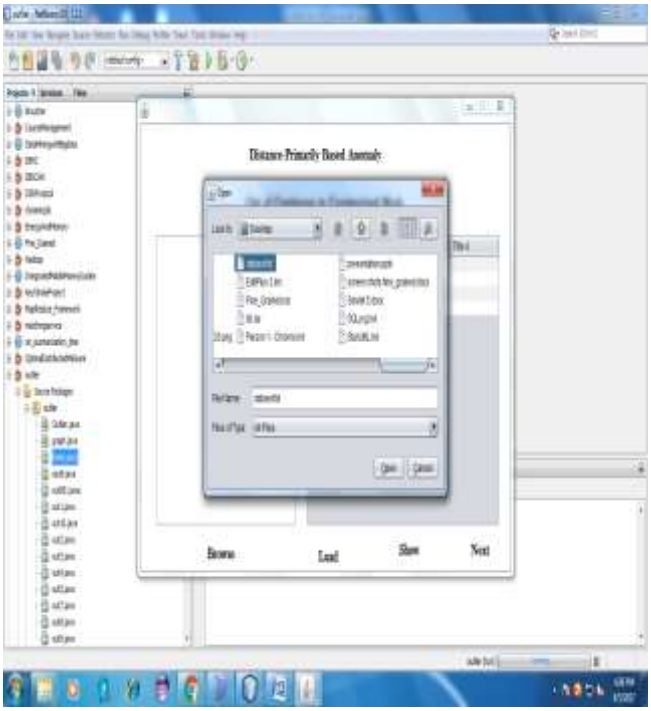
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



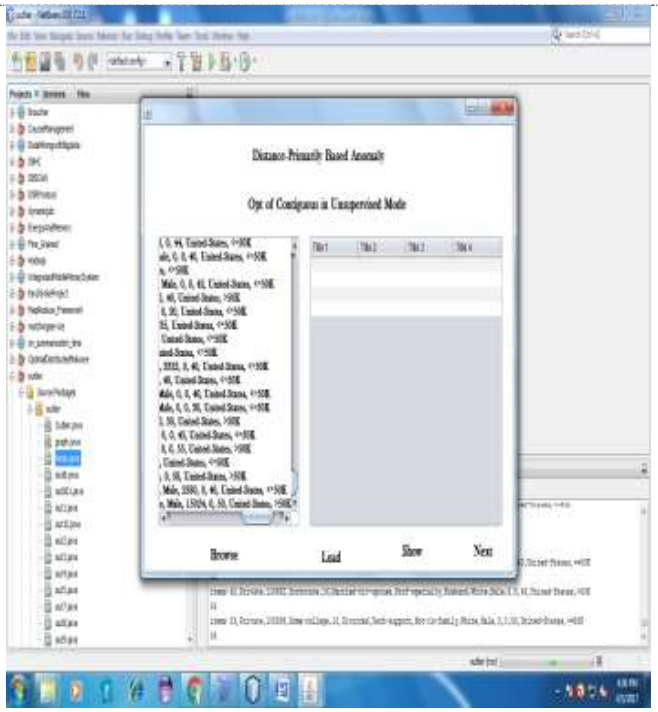


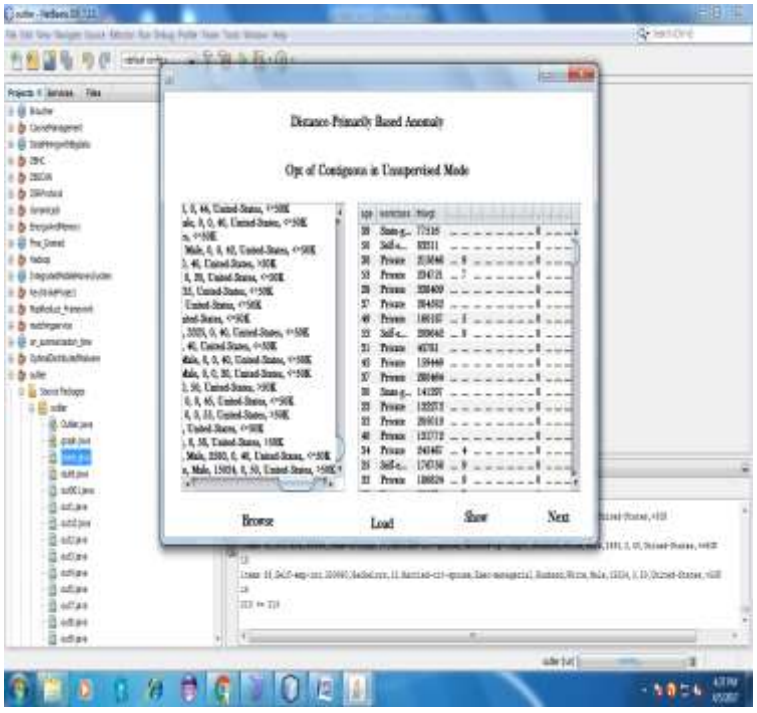
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



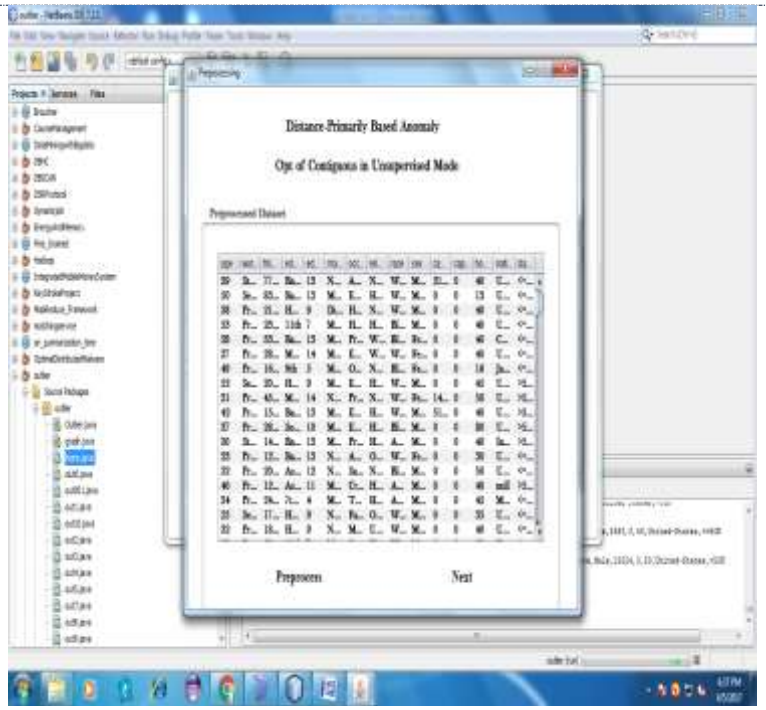


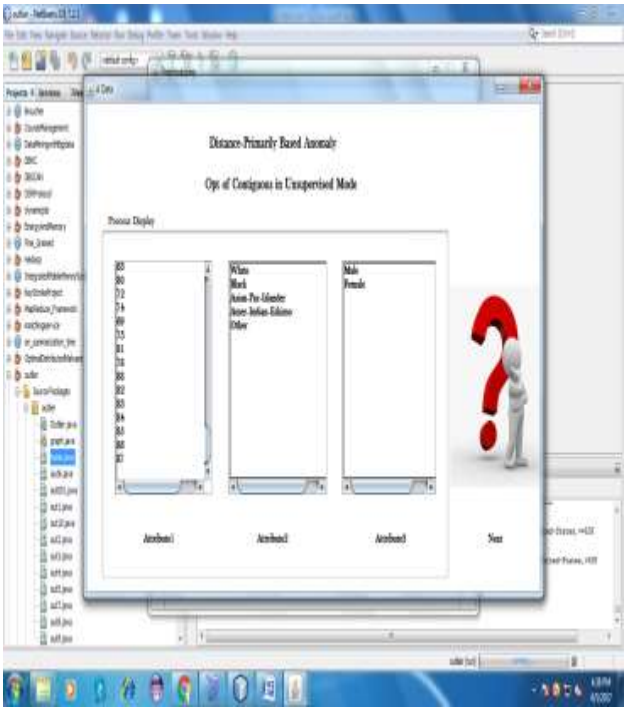
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



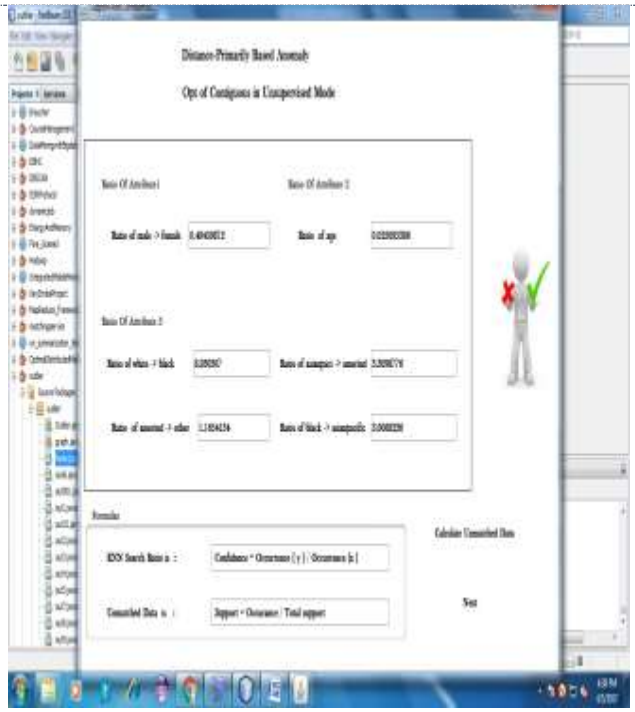


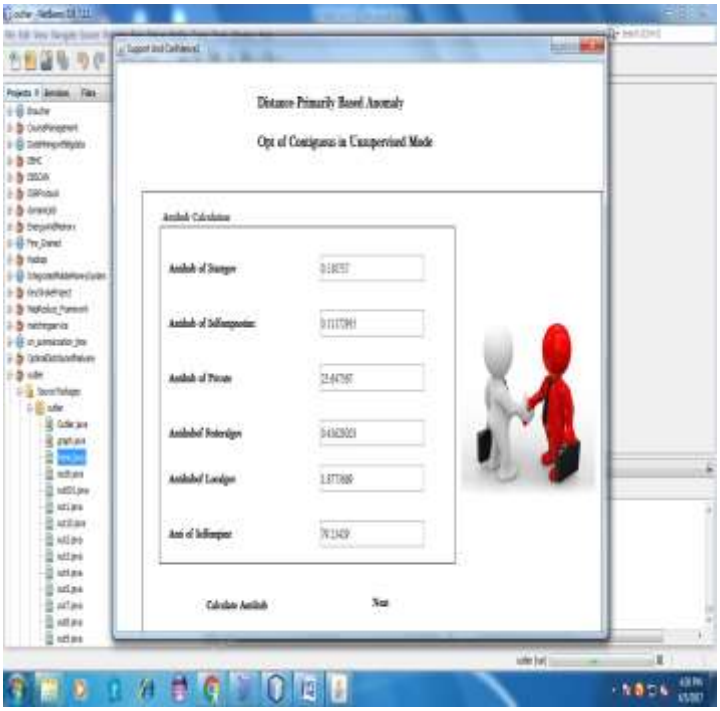
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



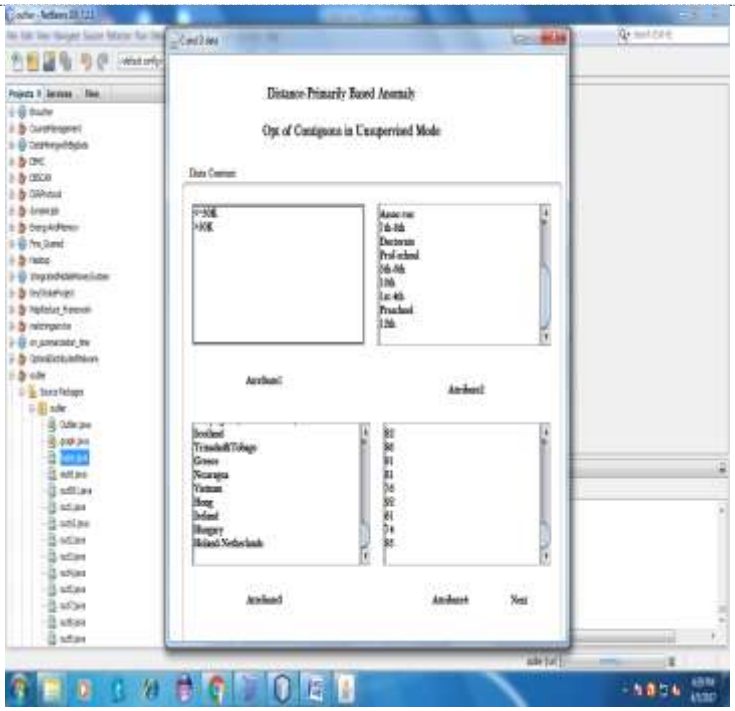


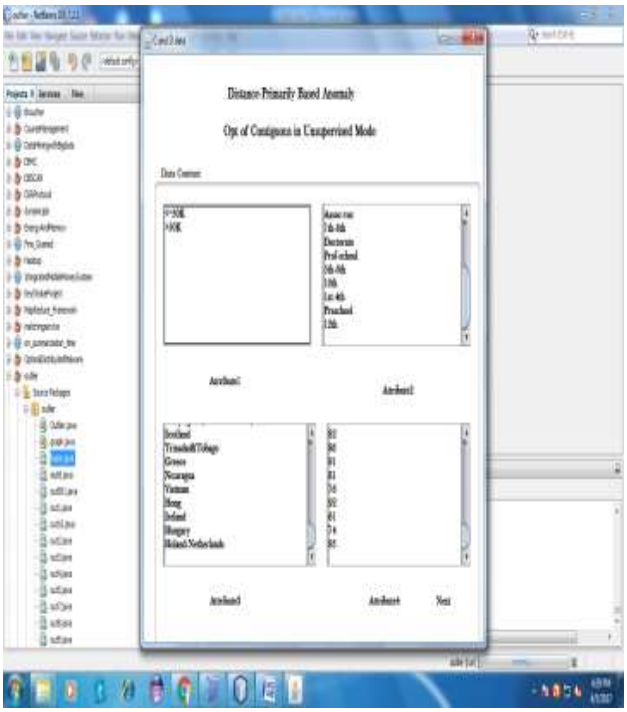
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



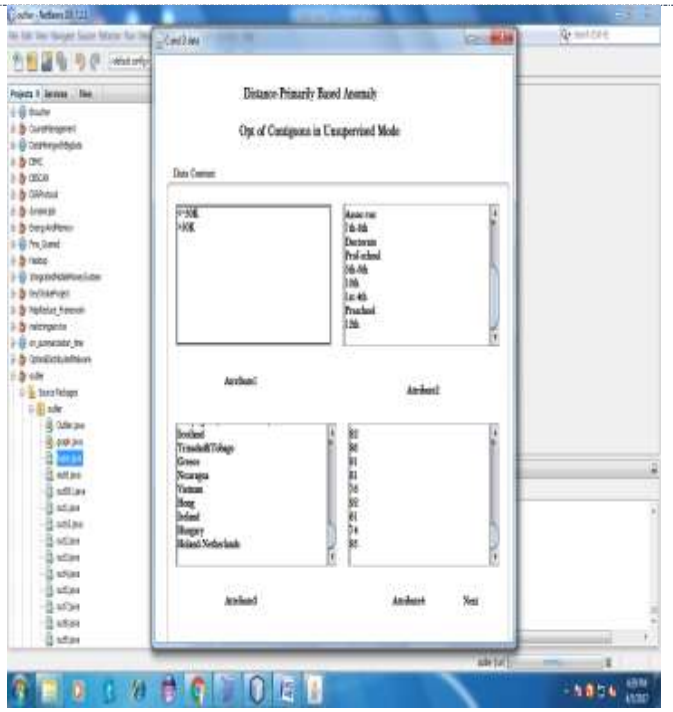


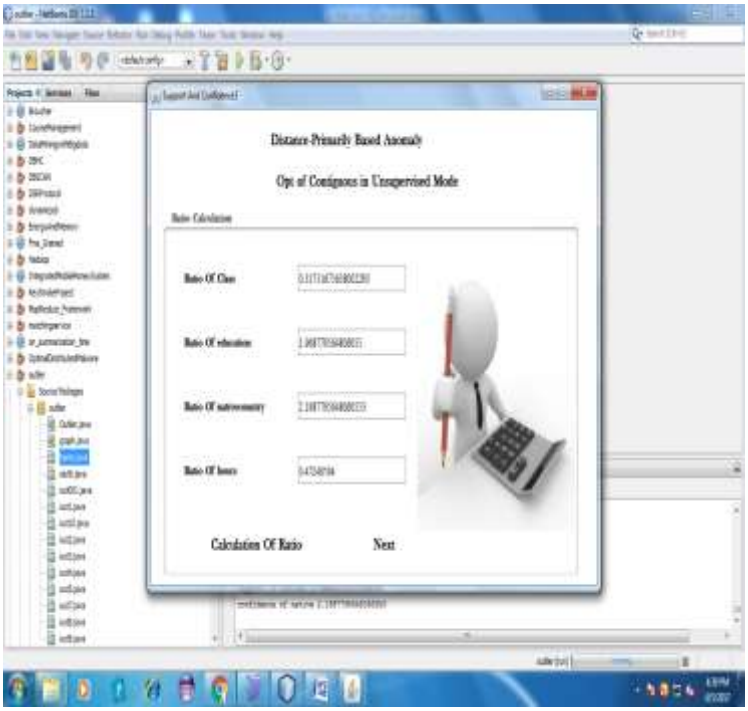
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



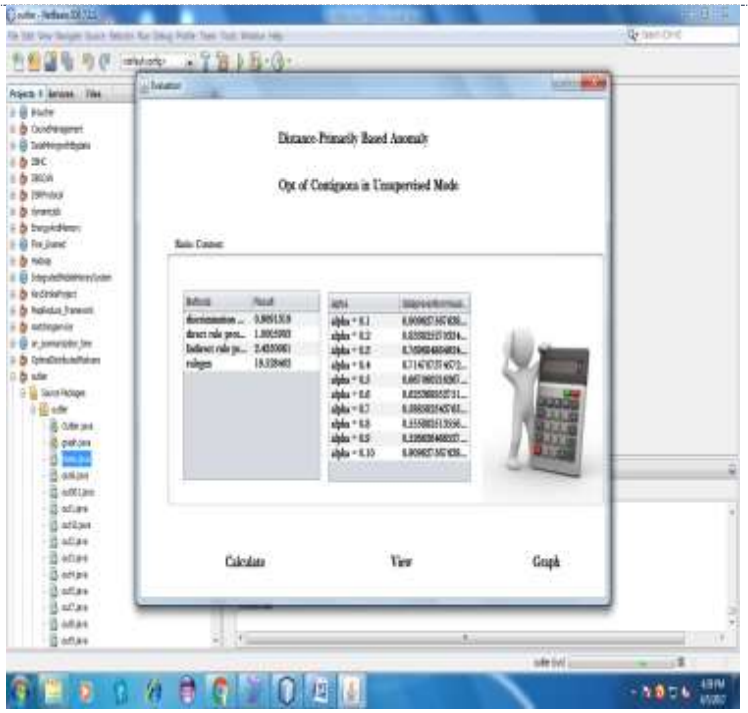


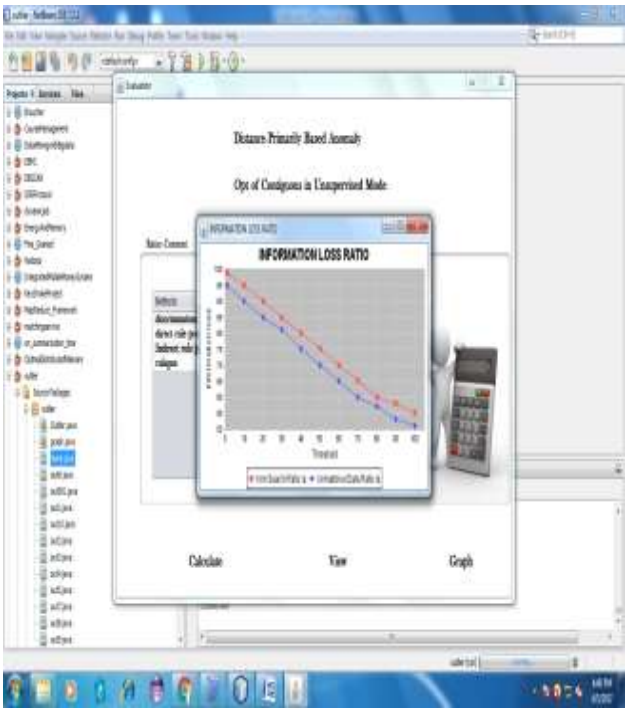
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



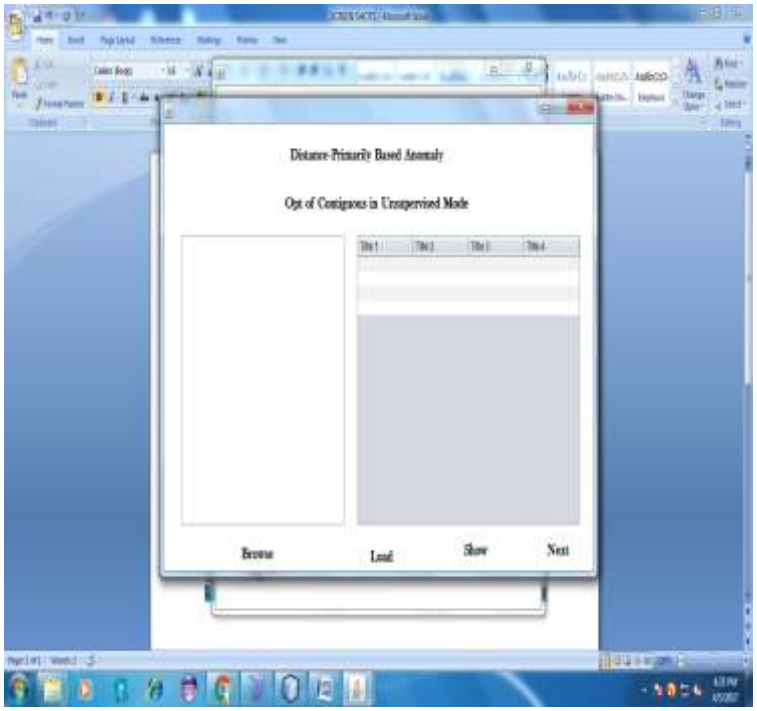


ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

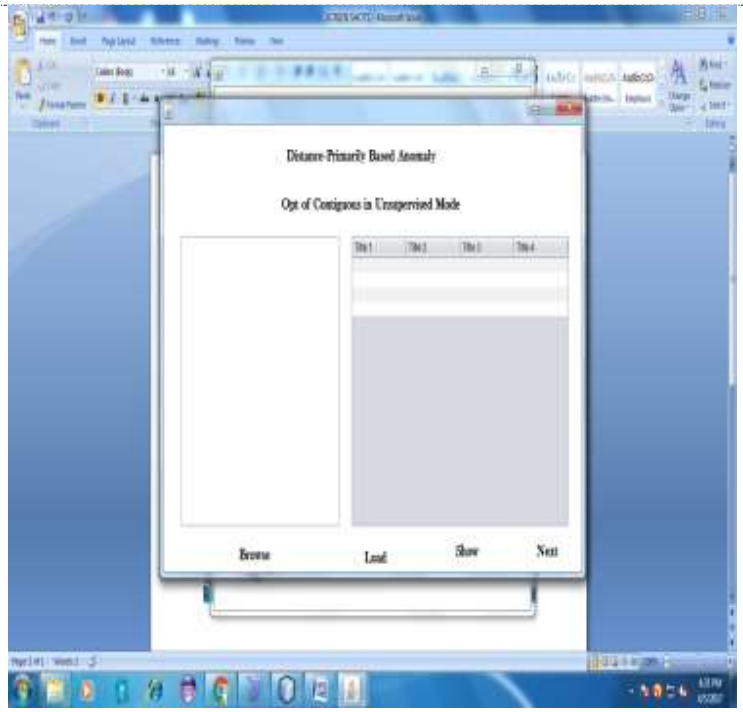


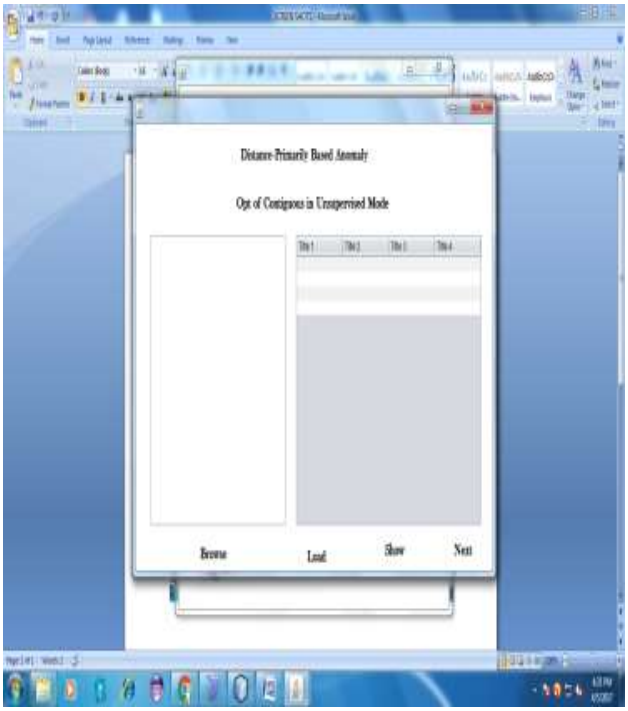


ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

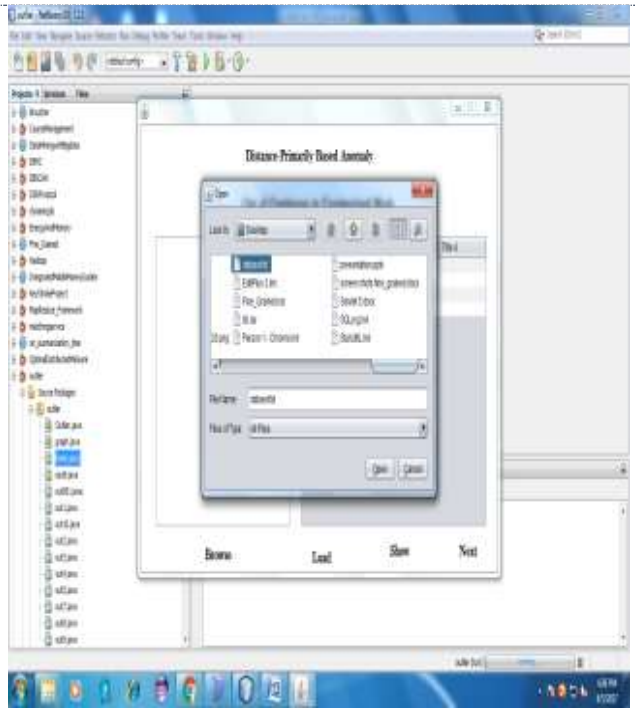


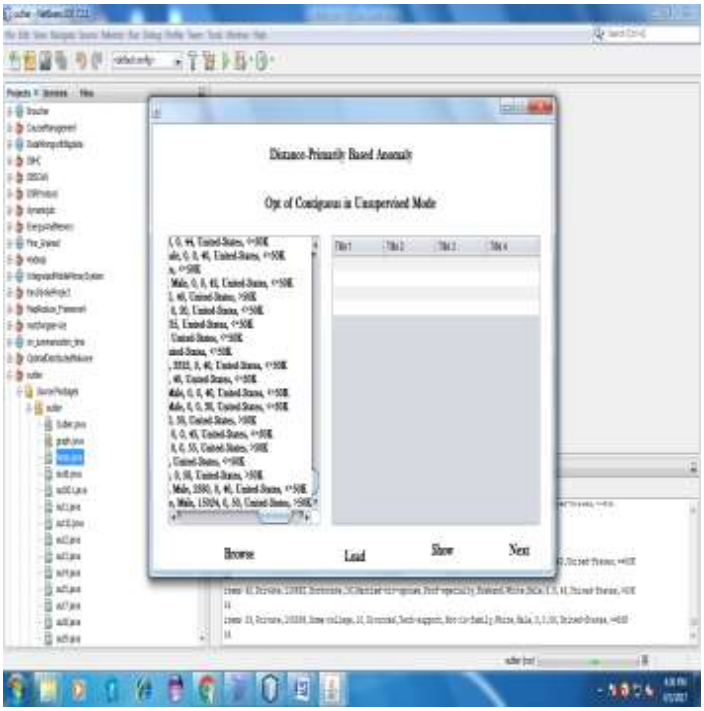
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



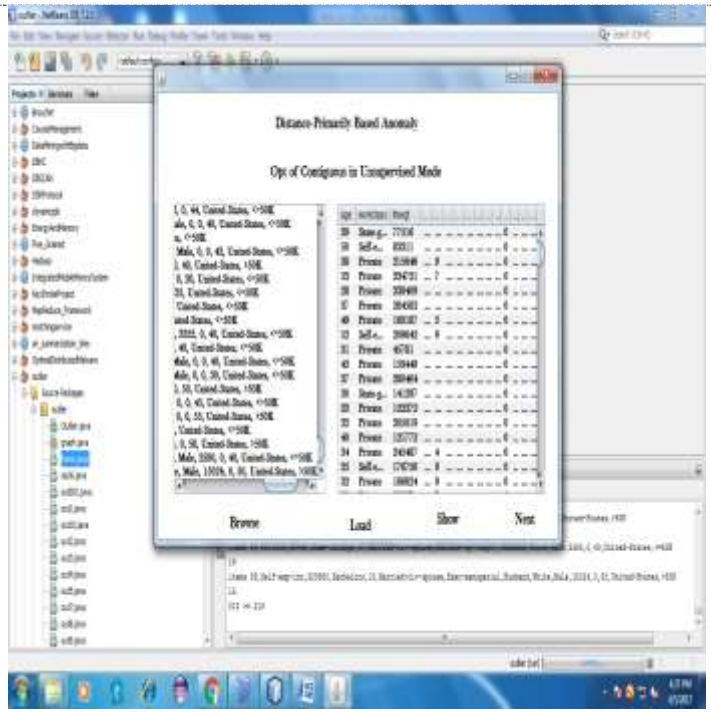


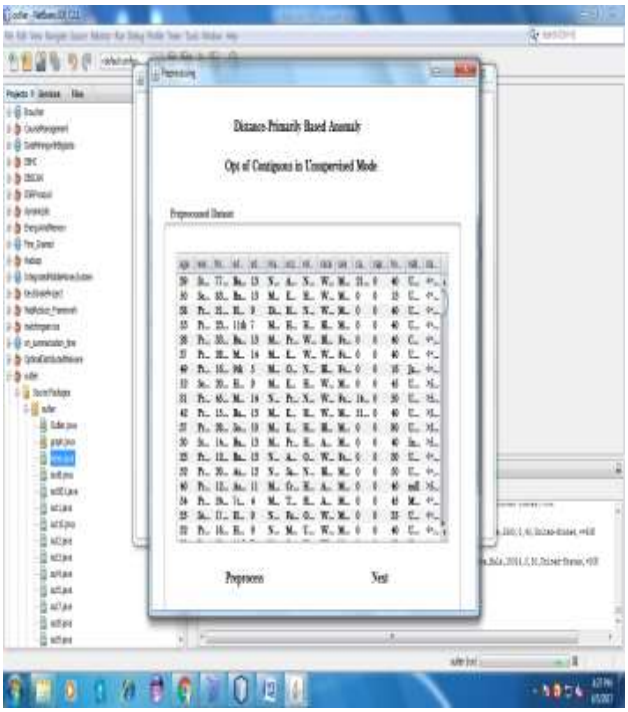
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



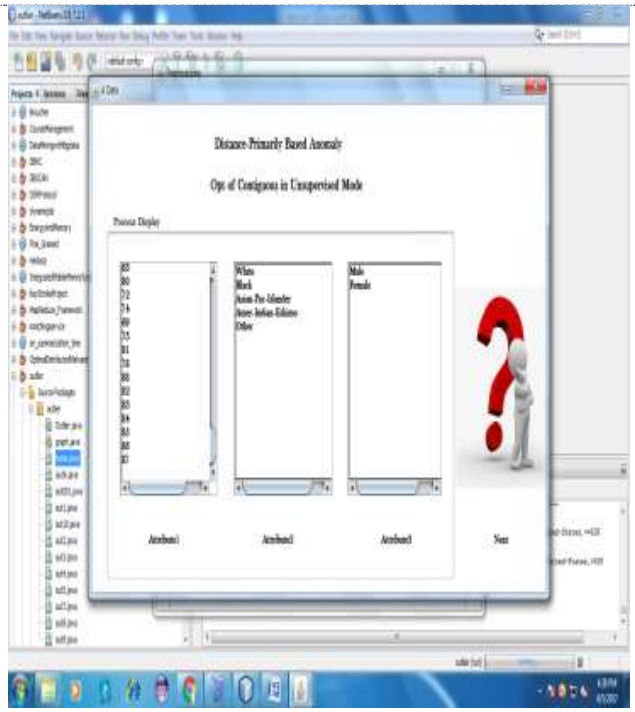


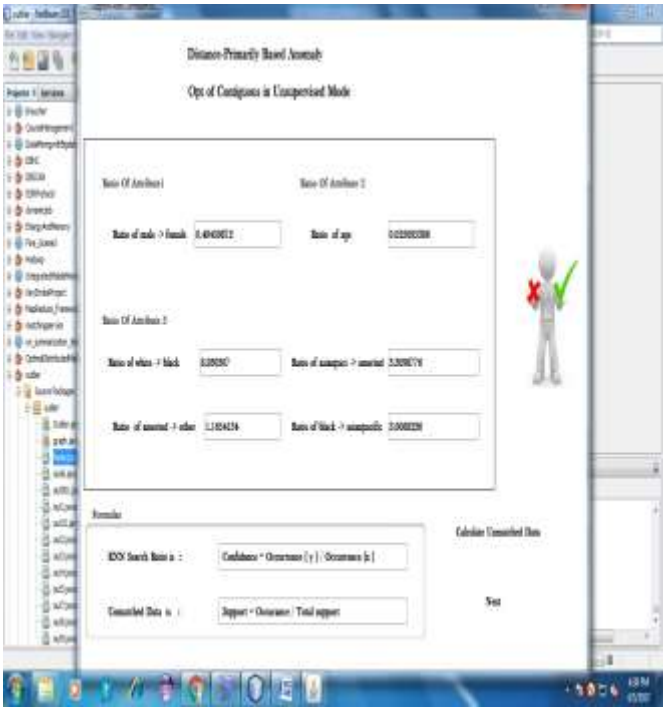
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



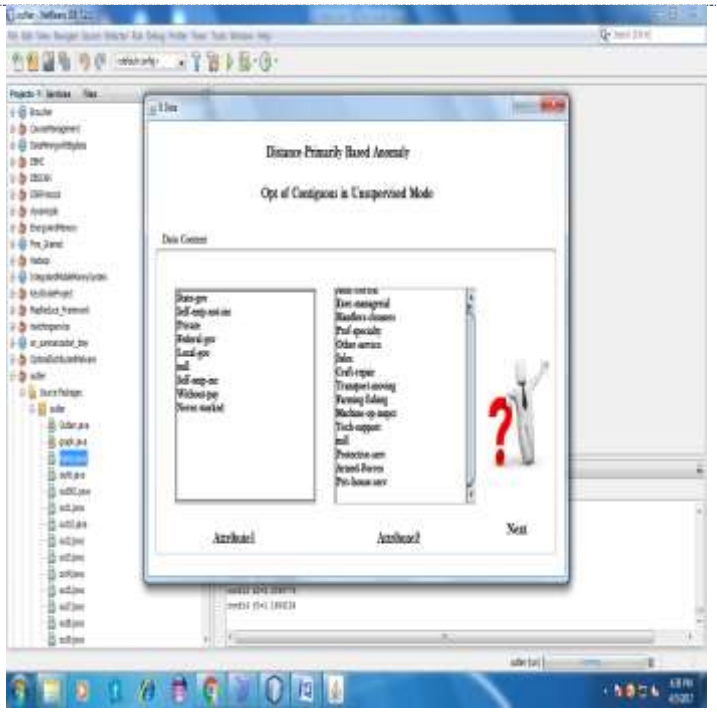


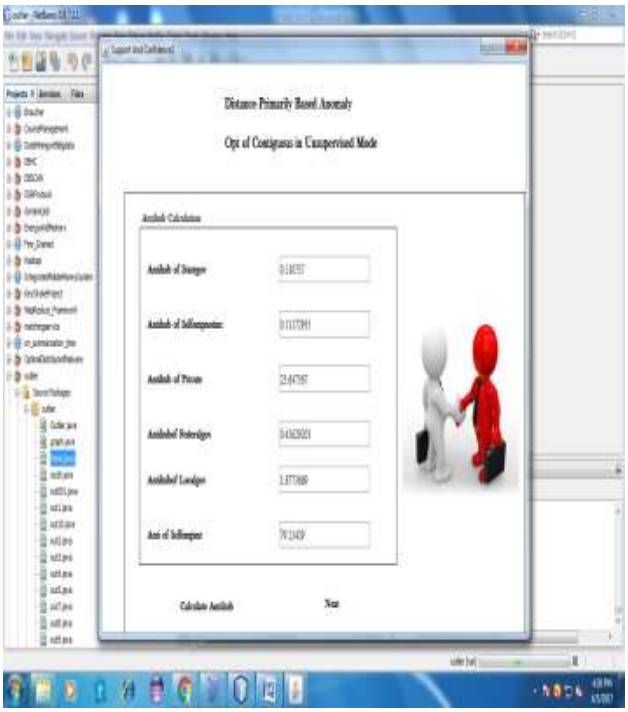
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



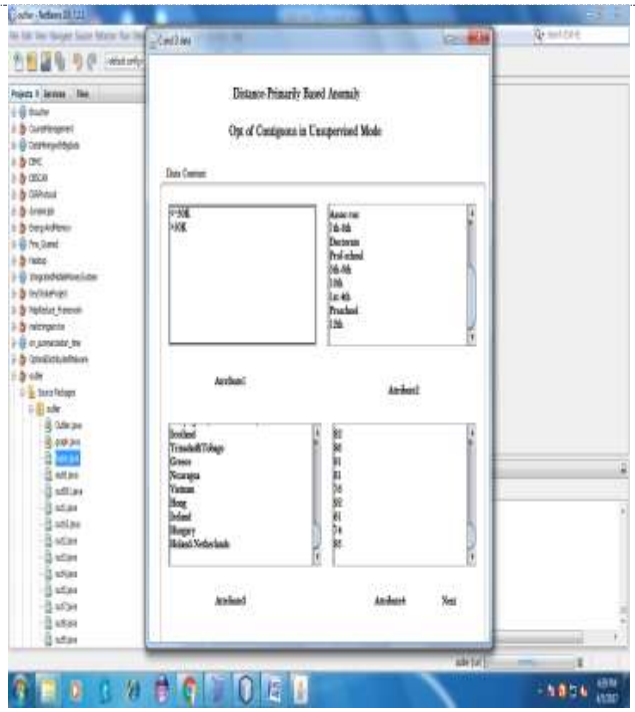


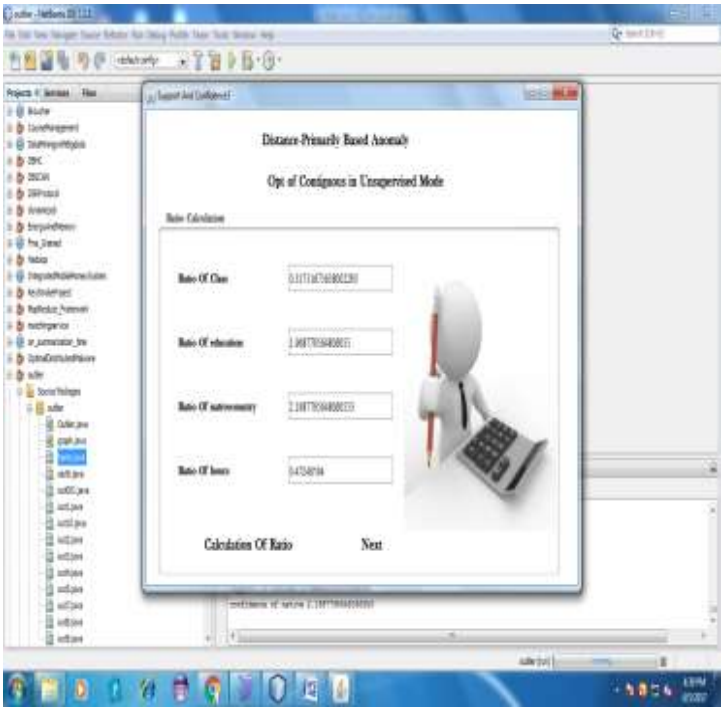
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



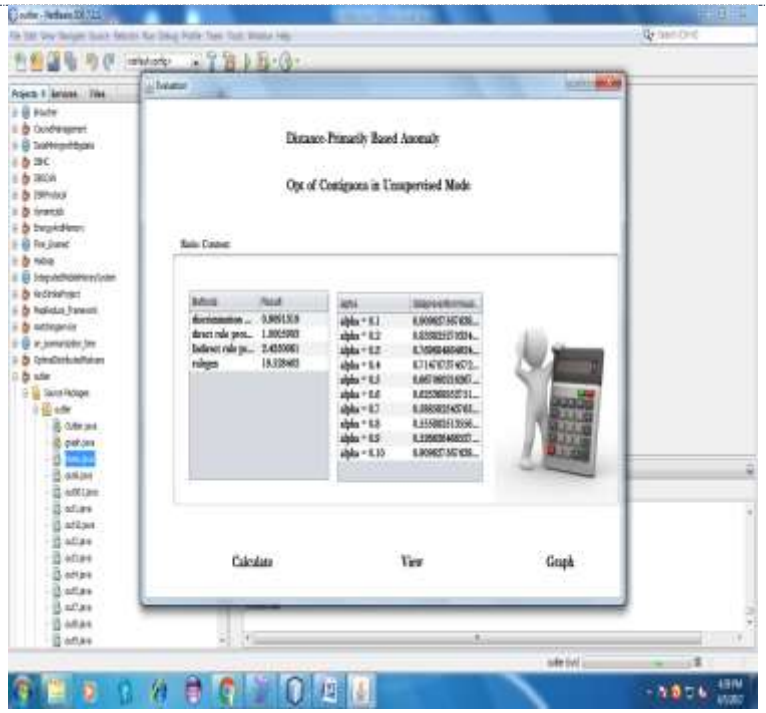


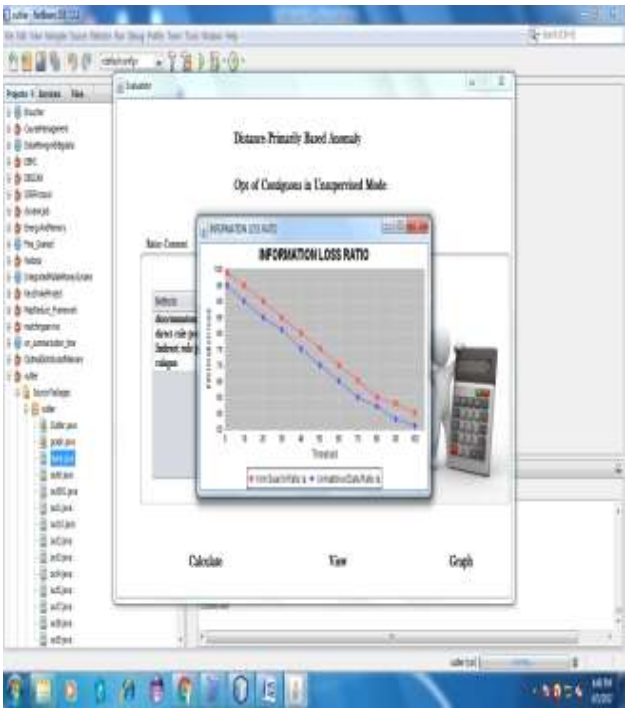
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7



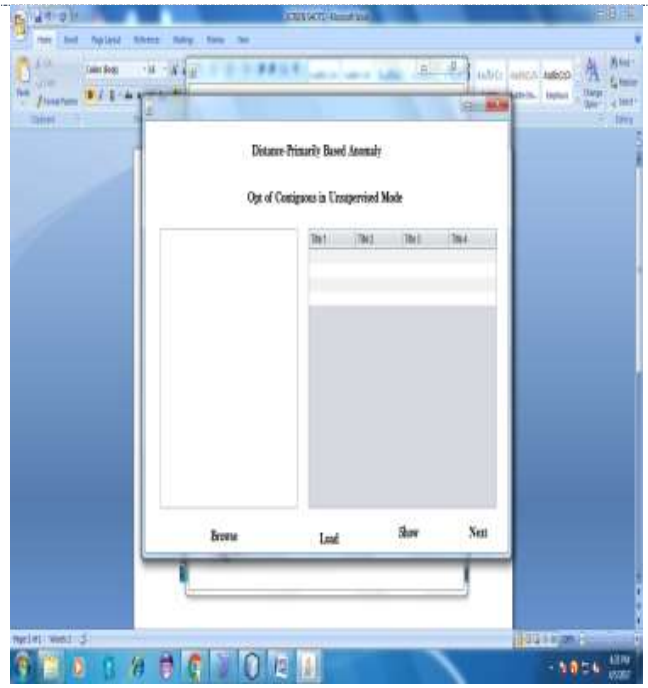


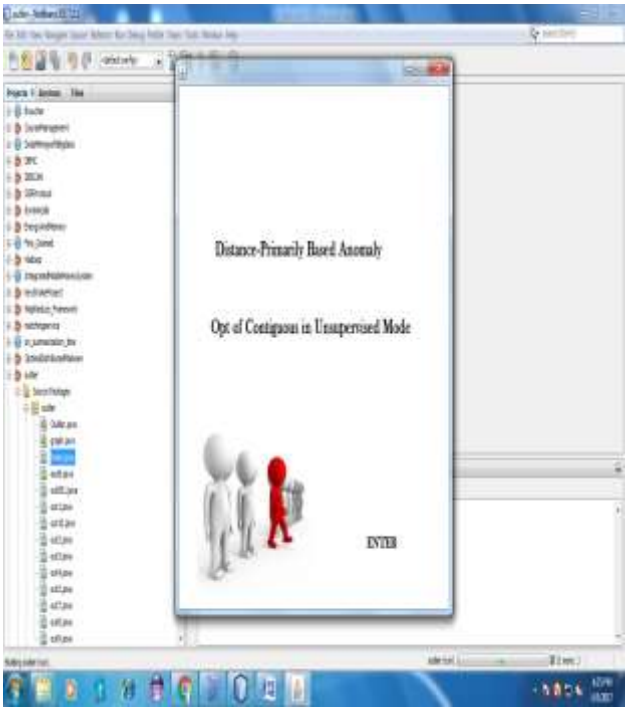
ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7





ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7





IX. CONCLUSION

In this paper, we provided a unifying view of the role of reverse nearest neighbor counts in problems concerning unsupervised outlier detection, focusing on the effects of high dimensionality on unsupervised outlier-detection methods and the hubness phenomenon, extending the previous examinations of (anti)hubness to large values of k, and exploring the relationship between hubness and data sparsity. Based on the analysis, we formulated the AntiHub method for unsupervised outlier detection, discussed its properties, and proposed a derived method which improves discrimination between scores. Our main hope is that this article clarifies the picture of the interplay between the types of outliers and properties of data, filling a gap in understanding which may have so far hindered the widespread use of reverse-neighbor methods in unsupervised outlier detection.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

`The existence of hubs and antihubs in high-dimensional data is relevant to machine-learning techniques from various families: supervised, semi-supervised, as well as unsupervised .In this paper we focused on unsupervised methods, but in future work it would be interesting to examine supervised and semi-supervised methods as well. Another relevant topic is the development of approximate versions of AntiHub methods that may sacrifice accuracy to improve execution speed. An interesting line of research could focus on relationships between different notions of intrinsic dimensionality, distance concentration, (anti)hubness, and their impact on subspace methods for outlie r detection. Finally, secondary measures of distance/similarity, such as shared-neighbor distances [20] warrant further exploration in the outlier-detection context.

X. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Survey, vol. 41, no. 3, p. 15, 2009.
[2] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection. Hoboken, NJ, USA: Wiley, 1987.
[3] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," SIGMOD Rec., vol. 29, no. 2, pp. 427–438, 2000.
[4] 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.
[5] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," VLDB J., vol. 8, nos. 3–4, pp. 237– 253, 2000.
[6] 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.
[7] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in Proc. 8th Int. Conf. Database Theory, 2001, pp. 420–434.
[8] D. Franc¸ois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," IEEE Trans. Knowl. Data. Eng., vol. 19, no. 7, pp. 873–886, Jul. 2007.
[9] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in Proc. 27th ACM SIGMOD Int. Conf. Manage. Data, 2001, pp. 37–46.
[10] A. Zimek, E. Schubert, and H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Statist. Anal. Data Mining, vol. 5, no. 5, pp. 363–387, 2012.
[11] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in Proc 17th Int. Conf. Pattern Recognit., vol. 3, 2004, pp. 430–433.
[12] 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..

CITE AN ARTICLE

Ranga Rao, K., & Vijayakumar, B. (n.d.). DETECTION OF OUTLIERS BY MAKING DISTANCE-BASED METHOD. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *7*(3), 407-435.