International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 3 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

Intelligent data analysis for medicinal diagnosis

Renuka Devi J[1], Sandhiya A[1], Sandhiya D[1], Maheswari M[2]

[1,2]*Computer Science and Engineering, Panimalar Engineering College, Chennai-600123, India*

Abstract—*Clinical choice emotionally supportive network, which utilizes progressed information mining procedures to help clinician make legitimate choices, has gotten extensive consideration as of late. The focal points of clinical choice emotionally supportive network incorporate not just enhancing analysis exactness additionally lessening conclusion time. In particular, with a lot of clinical information produced regular, na¨ıve Bayesian grouping can be used to exhume profitable data to enhance clinical choice emotionally supportive network. In this paper, we propose another security saving patient-driven clinical choice emotionally supportive network, which helps clinician integral to analyze the danger of patients' ailment in a protection safeguarding way. In the proposed framework, the past patients' verifiable information are put away in cloud and can be utilized to prepare the na¨ıve Bayesian classifier without releasing any individual patient therapeutic information, and afterward the prepared classifier can be connected to register the malady hazard for new coming patients furthermore permit these patients to recover the top-k sickness names as per their own inclinations Also, to influence the spillage of na¨ıve Bayesian classifier, we present a protection safeguarding top-k sickness names recovery convention in our framework. In addition to the present, the user will chat with the offered doctor for valuable suggestion relating to the treatment*

Keywords—Privacy; Medicinal diagnosis; Na¨ıve Bayesian classifier; Clinical Decision Support System.

1. INTRODUCTION

HEALTHCARE business, extensively distributed within the international scope to produce health services for patients, has never faced such a colossal amounts of electronic knowledge or old such a pointy rate of information these days. As declared by the Institute for Health Technology Transformation (iHT2), U.S. health care knowledge alone reached one hundred fifty exabytes (1018 bytes) in 2011 and would before long reach zettabyte (1021 bytes) scale and even yottabytes (1024 bytes) within the future [1]. However, if no applicable technique is developed to search out nice potential economic values from huge attention knowledge, these knowledge won't solely become vacuous however additionally need an out sized quantity of area to store and manage. Over the past 20 years, the miraculous evolution of information mining technique has obligatory a significant impact on the revolution of human's style by predicting behaviors and future trends on everything, which might convert keep knowledge into meaningful information. These techniques are well appropriate for providing decision support within the healthcare industry. To hurry up the diagnosis time and improve the diagnosis accuracy, a replacement system in health care business ought to be practicable to produce means a far cheaper and quicker way for diagnosis. Clinical decision support system (CDSS), with varied data mining techniques being applied to help doctors in identification of patient diseases with similar symptoms, has received an excellent attention recently [2]–[4]. Na¨ıve Bayesian classifier, one among the popular machine learning tools, has been widely used recently to predict varied diseases in CDSS.

We propose privacy-preserving patient-centric clinical decision system, known as PPCD, that is based on na¨ıve Bayesian classification to assist doctor to predict disease risks of patients in an exceedingly privacy-preserving means.

We propose a secure PPCD ,that permits processing to diagnose patient's disease without leaking any patient's medical information In PPCD, the past patient's historical medical information is often utilized by processing unit to coach the na¨ıve theorem classifier. Then, processing unit will use the trained classifier to diagnose patient's diseases in line with his symptoms in a very privacy-preserving manner. Finally, patients will retrieve the diagnosed results in line with his own preference in camera while not compromising the service provider's privacy.

Second, since individual historical medical knowledge can disclose patient's privacy revelation, we have a tendency to additionally introduce a brand new aggregation technique referred to as additive homomorphic proxy aggregation(AHPA) ,that permits service supplier to create na¨ıve Bayesian classifier while not unseaworthy somebody historical medical knowledge. Even the service supplier and cloud platform (CP) interact, no party will get any information concerning the individual historical medical knowledge apart from the owner himself , and only the aggregative knowledge is accessed by the service supplier.

1.1.RELATEDWORK

In 2010, Hamed Monkaresi, et.al [1] proposed a machine learning approach to enhance the accuracy of Heart rate detection in realistic measurements. This technology measures pulse rate and respiration using video solely. In this paper, we tend to evaluate the technique for measuring Heart rate in a very controlled manner, in a naturalistic computer interaction session and in an exercise situation. For comparison, heart rate was measured simultaneously using an diagnostic technique (ECG) device throughout all sessions. The proposed method didn't show positive ends up in naturalistic HCI and indoor exercise situations.

In 2005, C A M Schurink, et.al [2] proposed models and computer-based decision-support systems that are developed to help ICU physicians within the management of infectious diseases. we have a tendency to discuss the historical development, potentialities, and limitations of assorted computer-based decision-support models for infectious diseases, with special emphasis on theorem approaches. Though Bayesian decision-support systems are potentially useful for medical higher cognitive process in communicable disease management, clinical experience with them is restricted and prospective analysis is required to see whether or not their use will improve the quality of patient care

In 2013, Yousef Elmehdwi, et.al [3] proposed Secure k-Nearest Neighbor query over Encrypted data in Outsourced Environments. As a result of the increase of varied privacy issues, sensitive knowledge (e.g., medical records) have to be compelled to be encrypted before outsourcing to the cloud. Additionally, query processing tasks ought to be handled by the cloud; otherwise, there would be no purpose to outsourcing the data to the cloud. In this paper, we tend to focus on solving the k-nearest neighbor (kNN) query drawback over encrypted database outsourced to a cloud: a user issues an encrypted query record to the cloud, and also the cloud returns the k nearest records to the user. The primary protocol leaks some data to the cloud. The second protocol protects the confidentiality of the data however it's costlier.

In 2000, Pascal Paillier [4] proposed a completely unique computational problem, namely the Composite Residuosity class problem, and its applications to public-key cryptography. we have a tendency to propose a replacement trapdoor mechanism and derive from this technique3encryption schemes : a trapdoor permutation and 2 homomorphic probabilistic encryption schemes computationally corresponding to RSA. Our cryptosystems, based on usual modular arithmetic, are provably secure beneath acceptable assumptions within the standard model. It doesn't offer any proof of security against chosen cipher text

In 2015, C.Vanathy et.al [5] proposed k-NN Classification over Semantically Secure Encrypted relational information ,which defines a secure k-NN classifier over encrypted information within the cloud. The planned protocol protects the confidentiality of information, privacy of user's input query, and hides the information access patterns. To the most effective of our knowledge, our work is the first to develop a secure k-NN classifier over encrypted information under the semi-honest model. we have a tendency to empirically analyze the potency of our proposed protocol employing a real-world dataset under totally different parameter settings

Jingnian Chen, et.al [6] proposed Feature selection for text classification with Naïve bayes in 2008. as the Naïve theorem classifier is incredibly straight forward and economical and sensitive to feature selection, so the analysis of feature selection specially for it is significant. This paper presents 2 feature evaluation metrics for the Naïve theorem classifier applied on multiclass text datasets: Multi-class Odds ratio (MOR), and class Discriminating measure (CDM). Experiments of text classification with Naïve theorem classifiers were carried out on 2 multi-class texts collections. It doesn't work well for big multi class information sets

In 2009, Igor Kononenko [7] proposed Learning for Medical Diagnosis: History, State of the Art and Perspective provides an summary of the development of intelligent data analysis in medication from a machine learning perspective. The paper isn't

supposed to produce a comprehensive summary however rather describes some subareas and directions that from my personal point of view appear to be necessary for applying machine learning in diagnosing. the primary describes a recently developed methodology for handling reliability of decisions of classifiers, that looks to be promising for intelligent information analysis in medication. The second describes an approach to using machine learning in order to verify some unexplained phenomena from complementary medicine, that isn't (yet) approved by the orthodox health profession but could in the future play a crucial role in overall diagnosing and treatment.

In 2008, Riccardo Bellazzi, et.al [8] proposed data mining in clinical medicine: Current issues and guidelines. The goal of this review is to debate the extent and role of the analysis area of predictive data mining and to propose a framework to address the issues of constructing, assessing and exploiting It reviews the recent relevant work revealed within the space of predictive data processing
in clinical medicine, highlighting vital problems and summarizing the approaches in a set
of learned lessons. The paper provides a comprehensive review of the state of the art of predictive data processing in clinical drugs and provides tips to hold out data processing studies during this field.

1.3 PRELIMINARIES

1.3.1 Na¨ıve Bayesian Classifier

Na¨ıve Bayesian classifier is a very attractive classifier, which has been proved to be effective in several practical applications, together with text classification medical diagnosis and systems performance management

The na¨ıve Bayesian classifier is briefly reviewed as follows.

There are n classes which are represented as C1,C2,...,Cn.

Each sample is represented by n-dimensional vector $\vec{X}$= {X1,...,Xn}, depicting n measured values of the n-attributes T1,...,Tn, respectively. The classifier needs to predict $\vec{X}$ be-longs to the class with the highest a posteriori probability, i.e.,

$\vec{X}$ is predicted to lie in the class Ci if and only if there exists i, such that

P(Ci|$\vec{X}$)>P(Cj|$\vec{X}$), for all 1≤j≤n, j=i.

By Bayes' theorem

$$P(Ci|\vec{X}) = \frac{P(\vec{X}|Ci)P(Ci)}{P(\vec{X})}$$

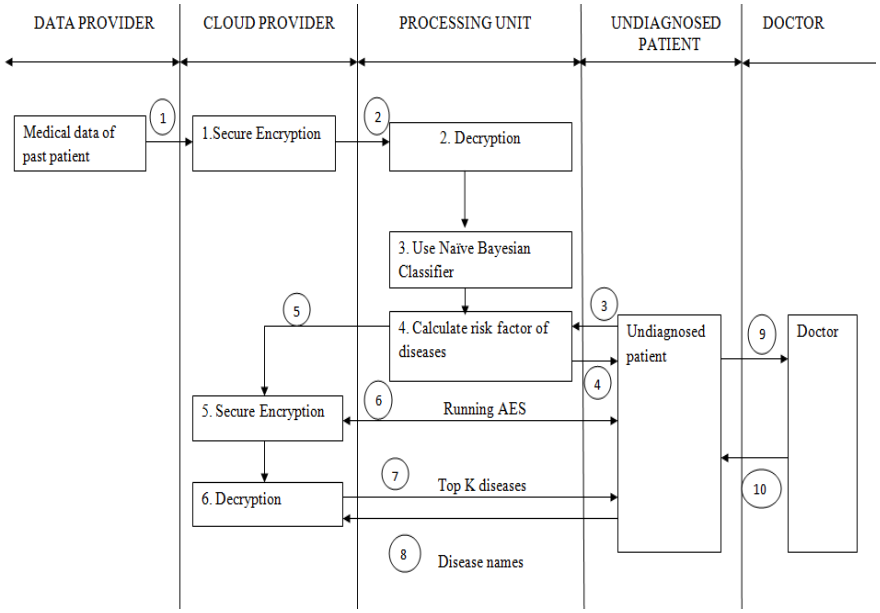
we can see that $P(\vec{X})$ is the same for all classes, only

P($\vec{X}$|Ci)P(Ci) needs to be maximized. In order to evaluate P($\vec{X}$|Ci)P(Ci), the na¨ıve assumption on class conditional independence is made.

P($\vec{X}$|Ci) ≈ $\prod_{k=1}^{n}$ P(Xk|Ci)

The probabilities P(X_1|Ci),P(X_2|Ci),...,P(X_n|Ci) can easily be estimated from the training set. In all, the classifier predicts that the class label of X is Ci if and only Ci has maximized the value P($\vec{X}$|Ci)P(Ci) among P($\vec{X}$|Cj)P(Cj), (1≤j≤f).

1.4.PROPOSEDSYSTEM



PRIVACY REQUIREMENTS

In our privacy model we tend to think that Data Provider is trustable which provides correct historical medical knowledge. The inner party Processing Unit is considered to be is considered as curious but honest, that is fascinated by Data Provider's individual medical knowledge and Undiagnosed patient's medical knowledge but strictly follows the protocol within the system. Both Cloud Provider and Undiagnosed patient are curious but honest parties during the system. Undiagnosed patient is curious about PU's Classifier, whereas Cloud provider is interested by all the other parties knowledge within the system.

The proposed system has the following modules.

1. User interface design
2. Hospital management
3. Trust Authorizes
4. Symptoms solutions
5. Chatting techniques
6. Reviews

User interface design

To connect with server user should offer their username and password then solely they'll be able to connect the server. If the user already exist user will directly login into the server else user should register their details like username, password and Email id, into the server. Server can produce the account for the whole user to take care of transfer and transfer rate. Name are going to be set as user id. Logging in is sometimes accustomed enter a selected page to attach with server user should offer their username and password then solely they'll be able to connect the server. If the user already exist

user will directly login into the server else user should register their details like username, password and Email id, into the server. Server can produce the account for the whole user to take care of transfer and transfer rate. Name are going to be set as user id. Logging in is sometimes accustomed enter a selected page.

Hospital management

To connect with server admin should offer the username and password then solely they will ready to connect the server. If the admin have solely the login method don't register the admin. when work it'll move to the admin page that point admin can also use the method. The method is to register the trust Authorizes and doctor.

Trust Authorizes

Trust Authorizes provides their username and password then solely they'll be ready connect with the server. The trust authorizes will collect past data .The Authorizes will get knowledge and transferinformation for information.

User Symptoms solutions

The user enters the user page and search the symptoms by patient that is accustomed. find the user resolution. The user also notice worth symptoms that's accustomed worth the user symptoms.

Chatting techniques

The user will chat with doctor that is employed for their Verification. The symptoms based result's correct or not may be verified with the actual specialist .Doctor reply question that are accustomed .

Reviews

We will exploit privacy conserving patient-centric clinical call support systems with data processing techniques. We do reviews with the method.

1.5.ALGORITHM

Length of the input and output block and therefore the State is 128 for AES algorithm. It is represented by Lb. For the AES algorithm 128,192 or 256 bits is the length of the Cipher Key K .Here the length of secret key is 128 bits. The key length of value reflects the amount of 32 bit words within the cipher Key.

For the AES algorithm rule, throughout the execution of algorithm the numbers of rounds to be performed are dependent on the key size. Lr is used to represent the number of rounds.AES algorithmic rule uses a round function for both its cipher and inverse cipher. This function consists of 4 different byte oriented transformations:

a. Using Sub Bytes (S box) substitution

b. Shifting Rows of the state array

c. Mix the data with each Column

d. Add Round Key Routine.

We refer to the round keys as K_0, K_1,K_2, ...K_{Lr}
4.1 The Algorithm:
The input (block size Lb, also known as plaintext) of the AES algorithm is converted into a 4 x 4 matrix, called a state.
Four transformations,
AddRoundKey, SubBytes, ShiftRows and

MixColumns, perform various operations on the state to calculate the output (cipher text).
InvMethod(Method(a)) = a
If AddRoundKey operates on a variable twice, the variable itself is returned

4.2 Transformation in AES
Some mathematical operations are needed to understand to perform these four transformations which are given as below

a. SubBytes :

Subbyte is the SBOX for AES. It operates on every byte in the state and performs a non-linear substitution in the GF(28) field, that is what makes AES a non-linear cryptographic system. In order to be invertible every value of b' should be generated from a singular worth of b. A look up table will additionally be implemented for SubBytes. SubByte operation performs an affine transformation on the inverse of byte b, and adds it to 0xC6.

b. ShiftRows:

ShiftRows performs operations on individual rows. It provides diffusion throughout the AES algorithm. the primary row isn't modified. The second row is shifted one byte to the left, with the left most byte wrapping around. The third row shifts 2 bytes to the left, and also the fourth row shifts 3 bytes to the left with acceptable wrapping to the right. This description is for AES-128,
the number of shifts for every row changes based on the key size.

Lb	**Row 0**	**Row 1**	**Row 2**	**Row 3**
4	0	1	2	3
6	0	1	2	3
8	0	1	3	4

Table 1: No. of Shifts

c. Mix Columns

MixColumns performs operations on individual columns of the state. The columns are considered polynomials over GF(28) and multiplied
modulo y4+1 with a(y)
where a(y) = {03}y3+ {01}y2+ {01}y + {02}
NOTE: y4+1 is relatively prime to a(y). This can be represented as a matrix equation:

$$\begin{bmatrix} a'_0 \\ a'_1 \\ a'_2 \\ a'_3 \end{bmatrix} = \begin{bmatrix} 02 & 03 & 01 & 01 \\ 01 & 02 & 03 & 01 \\ 01 & 01 & 02 & 03 \\ 03 & 01 & 01 & 02 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

d. AddRoundKey Routine

The AddRoundKey is used for adding(XOR addition) the round key to the output of the previous step during the forward process

$K_{0,0}$	$K_{0,1}$	$K_{0,2}$	$K_{0,3}$
$K_{1,0}$	$K_{1,1}$	$K_{1,2}$	$K_{1,3}$
$K_{2,0}$	$K_{2,1}$	$K_{2,2}$	$K_{2,3}$
$K_{3,0}$	$K_{3,1}$	$K_{3,2}$	$K_{3,3}$

$\otimes$

$b_{0,0}$	$b_{0,1}$	$b_{0.2}$	$b_{0,3}$
$b_{1,0}$	$b_{1,1}$	$b_{1,2}$	$b_{1,3}$
$b_{2,0}$	$b_{2,1}$	$b_{2,2}$	$b_{2,3}$
$b_{3,0}$	$b_{3,1}$	$b_{3,2}$	$b_{3,3}$

$\Downarrow$

$A_{0,0}$	$A_{0,1}$	$A_{0,2}$	$A_{0,3}$
$A_{1,0}$	$A_{1,1}$	$A_{1,2}$	$A_{1,3}$
$A_{2,0}$	$A_{2,1}$	$A_{2,2}$	$A_{2,3}$
$A_{3,0}$	$A_{3,1}$	$A_{3,2}$	$A_{3,3}$

The AddRoundKey is performed at first and at the end in order to provide initial and final randomness to the algorithm. Without this, a third party can easily deduct the first or final portion of the cipher, and therefore it would be irrelevant to the privacy of the cipher.

Example

INPUT TEXT: swelling

ENCRYPTED TEXT:1kXNDBvlkTiq46RMCbjaA=

DECRYPTED TEXT: same as input message.

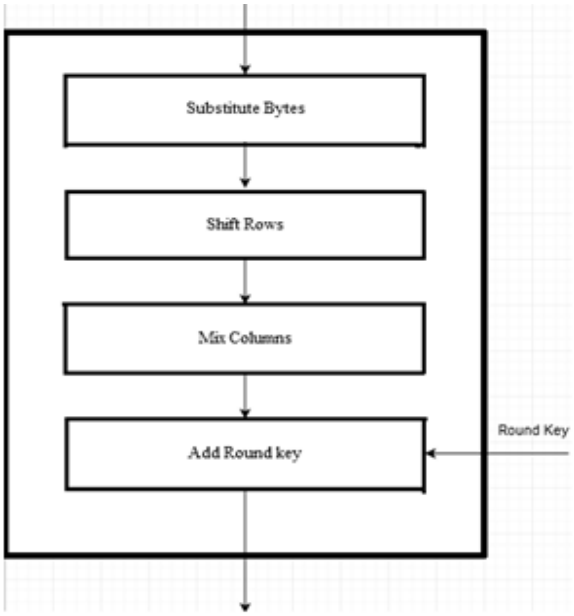


Fig 1 AES Encryption

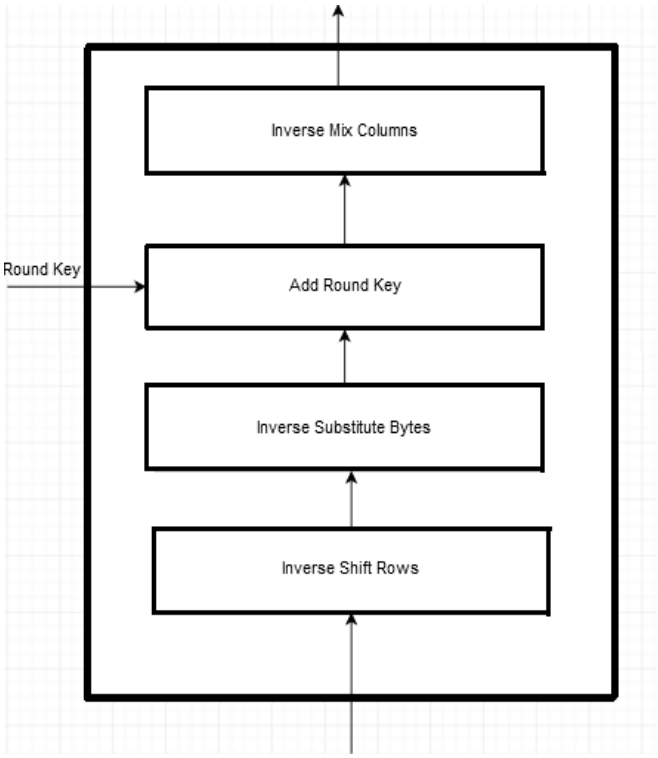


Fig 2 AES Decryption

2. ABBREVIATIONS

CDSS-Clinical Decision Support System

AES-Advanced Encryption Standard

3.CONCLUSION

In this paper we've got projected a PPCD victimization na¨ıve theorem classifier. By taking the advantage of rising cloud computing technique system will use massive medical dataset hold on in CP to coach na¨ıve theorem classifier, then apply the classifier for illness identification while not compromising the privacy of stateless person, additionally the patient will firmly retrieve the top-k identification results consistent with their own preference in our system. Since all the information square measure processed within the encrypted kind, our system are able to do patient- centric diagnose result retrieval in privacy preserving means. For the longer term work, we are going to exploit PPCD with different advanced data processing technique.

4.REFERENCES

[1] H. Monkaresi, R. A. Calvo, and H. Yan, "A machine learning approach to improve contactless heart rate monitoring using a webcam," *IEEE*
J. Biomed. Health Informat., vol. 18, no. 4, pp. 1153–1160, Jul. 2014.

[2] C. Schurink, P. Lucas, I. Hoepelman, and M. Bonten, "Computer-assisted decision support for the diagnosis and treatment of infectious iseases in intensive care units," *Lancet Infectious Dis.*, vol. 5, no. 5, pp. 305–312, 2005.

[3] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," *Artif. Intell.Med.*, vol. 23, no. 1, pp. 89–109, 2001.

[4] N. Lavraˇc, I. Kononenko, E. Keravnou, M. Kukar, and B. Zupan, "Intelligent data analysis for medical diagnosis: Using machine learning and temporal abstraction," *Artif. Intell.Commun.*, vol. 11, no. 3, pp. 191–218, 1998.

[5] Y. Elmehdwi, B. K. Samanthula, andW. Jiang, "Secure k-nearest neighbor query over encrypted data in outsourced environments," in *Proc. IEEE 30th Int. Conf. Data Eng.*, pp. 664–675, 2014s.

[6] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proc. Adv. Cryptol. Int. Conf. Theory Appl. Cryptograp.Techn*., Prague, Czech Republic, May 2–6, 1999, pp. 223–238.

[7] Y. Elmehdwi, B. K. Samanthula, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted relational data,"*IEEE Trans. Knowledge Data Eng.*, (2015).[Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6930802.

[8] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," *Int. J. Med. Informat.*, vol. 77, no. 2,pp. 81–97, 200

Abstract — Outlier is defined as an event that deviates too much from other events. The identification of outlier can lead to the discovery of useful and meaningful knowledge. Outlier means it's happen at some time it's not regular activity. Research about Detection of Outlier has been extensively studies in the past decade. However, most existing research focused on the algorithm based on specific knowledge, compared with outlier detection approach is still rare. In this paper mainly focused on different kind of outlier detection approaches and compares it's prone and cones. In this paper we mainly distribute of outlier detection approach in two parts classic outlier approach and spatial outlier approach. The classical outlier approach identifies outlier in real transaction dataset, which can be grouped into statistical approach, distance approach, deviation approach, and density approach. The spatial outlier approach detect outlier based on spatial dataset are different from transaction data, which can be categorized into spaced approach and graph approach. Finally, the comparison of outlier detection approaches.

Keywords— *outlier detection; spatial data, transaction data.*

I. INTRODUCTION

Data mining is a process of extracting valid, previously unknown, and ultimately comprehensible information from large datasets and using it for organizational decision making [10]. However, there a lot of problems exist in mining data in large datasets such as data redundancy, the value of attributes is not specific, data is not complete and outlier [13].Outlier is defined as an observation that deviates too much from other observations that it arouses suspicions that it was generated by a different mechanism from other observations [21]. The identification of outliers can provide useful, sufficient and meaningful knowledge and number of applications in areas such as climatology, ecology public health, transportation, and location based services. Recently, a few studies have been
conducted on outlier detection for large dataset [4]. However, most existing study concentrate on the algorithm based on special background, compared with outlier identification approach is comparatively less. This paper mainly discusses about outlier detection approaches from data mining perspective. The inherent idea is to research and compare achieving mechanism of those approaches to determine which approach is better based on special dataset and different background.

The rest of this paper is organized as follows. Section 2 reviews related work in outlier detection. We would like to discuss different method of outlier detection which can be differentiating based on: classic outlier technique based on real time dataset and spatial outlier technique based on spatial dataset which is discuss in section 3. The classic outlier approach can be grouped into statistical-based approach, distance-based approach, deviation -based approach, density based approach. The spatial outlier approach can be grouped into space-based approach and graph-based approach. Comparison of outlier detection is provided in Section 4. Finally, Section 5 concludes with a summary of those outlier detection algorithms.

II. PREVIOUS WORK

The classic definition of an outlier is due to Hawkins [21] who defines "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism".

Most approaches on outlier mining in the early work are based on statistics which use a standard distribution to fit the dataset. Outliers are describing based on the probability distribution. For example, Yamanishi et a1. Used a Gaussian mixture model to describe the normal behaviors and each
object is given a score on the basis of changes in the model
[22].

Knorr et al. proposed a new definition based on the concept of distance, which regard a point p in data set as an outlier with respect to the parameters K and λ, if no more than k points in the data set are at a distance λ or less than p [6].

Arning et a1. Proposed a deviation-based method, which identify outliers by inspecting the main characteristics of objects in a dataset and objects that "deviate" from these features arc considered outliers [1].

Breunig et al. introduced the concept of local outlier, a kind of density-based outlier, which assigns each data a local outlier factor LOF of being an outlier depending on their neighborhood [13]. The outlier factors can be computed very efficiently only if some multi-dimensional index structures such as R-tree and X-tree [17] are employed. A top-n based local outlier mining algorithm which uses distance bound micro cluster to estimate the density was presented in [9].

Lazarevic and Kumar proposed a local outlier detection algorithm with a technique called "feature bagging" [4]. Shekhar et al. [6] proposed the definition of spatial outlier: "A spatial outlier is spatially associating objects whose non spatial rank values are much distinct to those of other spatially assign objects in its spatial neighborhood".

Kou et al. developed spatial weighted outlier detection algorithms which use properties such as center distance and common border length as weight when comparing non-spatial attributes [2]. Adamet al. proposed an algorithm which considers both spatial relationship and semantic relationship among neighbors [5]. Liu and Jezek proposed a method for detecting outliers in an irregularly distributed spatial data set

[11].

III. OUTLIER DETECTION APPROACHES

Outlier detection has been extensively studied in the past decennium and numerous methods have been created. Outlier detection approach is differentiating in two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density based approach. The spatial outlier approaches analyze outlier based on spatial dataset, which can be grouped into space based approach and graph-based approach, as illustrated in Figure 2.

3.1. CLASSIC OUTLIER

Classic outlier approach analyzes outlier based on transaction dataset, which consists of collections of items. A typical example is market basket data, where each transactions is the group of products purchased by a customer in a one transaction. Such data can also be augmented by additional "items" describing the customer or the context of the transaction. Commonly, transaction data is relative to other data to be simple for the outlier detection. Thus, most outlier approaches are researched on transaction data.

(1) Statistical Approach

Statistical approaches were the oldest algorithms used for outlier identification, which cause a distribution model for the given data set and using a discordance test they
detect outliers. In fact, many of the techniques described in both Barnett and Lewis [20] and Rousseeuw and Leroy [18] are single dimensional. However, with the dimensions increasing, it becomes more difficult and inaccurate to make a model for dataset.

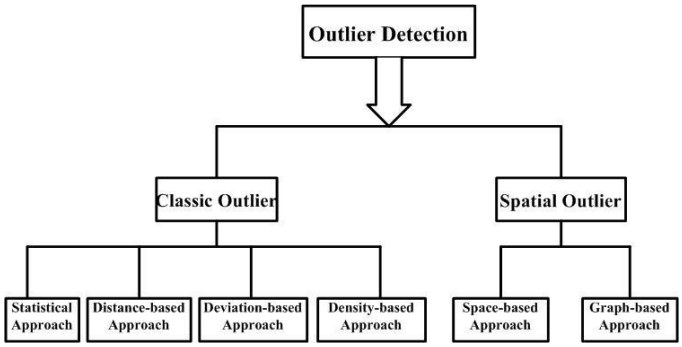


Fig. 2 Classification of Outlier Approach

(2)Distance-based Approach

The concept of distance-based outlier relies on the notion of the neighborhood of a point, typically, the knearest neighbors, and has been first introduced by Knorr and Ng [16,14]. Distance-based outliers are those points for which there are less than k points within the distance in the input data set. Distance-based approach is not providing required knowledge about a ranking of outlier detection but it's used to define a preferable rank of the parameter.

Ramaswamy et al. [15] modified the definition of outlier introduced by Knorr and Ng and consider as outliers the top n point's p whose distance to their k-th nearest neighbor is greatest. Partition based technique are works as follow: Firstly they divide the input points using clustering technique and then prune that division that cannot contain outlier which is used to detect outliers.

The distanced-based approach is effective in rather low dimensions, because of the sparsity of high dimensional points, the approach is sensitive to the parameter λ and it is hard to figure out a-priori. As the dimensions increase, the method's effect and accuracy quickly decline.

(3) Deviation-based Approach

Arning et a1. proposed a deviation-based method, which identify outliers by inspecting the main characteristics of objects in a dataset and objects that "deviate" from these features arc considered outliers [19].

(4) Density-based Approach

The density-based approach estimates the density
distribution of the data and identifies outliers as those lying in low-density regions. Breunig et al. [13] assign a local outlier

factor (*LOF*) to each point based on the local density of its neighborhood, which is determined by a user-given minimum number of points (*MinPts*). Papadimitriou et al. [7] present *LOCI* (Local Correlation Integral) which uses statistical values based on the data itself to tackle the issue of choosing values for *MinPts*. Density-based techniques have the advantage that they can detect outliers that would be missed by techniques with a single, global criterion. However, data is usually sparse in high-dimensional spaces rendering density-based methods problematic.

3.2. SPATIAL OUTLIER

For spatial data, classic approaches have to be modified because of the qualitative difference between spatial and non-spatial attributes. Spatial dataset could be defined as a collection of spatially referenced objects. Attributes of spatial objects fall into two categories: spatial attributes and non spatial attributes. The spatial attributes include location, shape and other geometric or topological properties. Non spatial attributes include length, height, owner, building age and name. Comparisons between spatially referenced objects are based on non-spatial attributes [8].

Informally, a spatial outlier is a local instability, or an extreme observation with respect to its neighboring values, even though it may not be significantly different from the entire population. Detecting spatial outliers is useful in many applications of geographic information systems and spatial
dataset [6, 8, 12].

The identification of spatial outliers can reveal hidden but valuable information in many applications, For example, it can help locate severe meteorological events, discover highway congestion segments, pinpoint military targets in satellite images, determine potential locations of oil reservoirs, and detect water pollution incidents.

(1) Space-based Approach

Space-based outliers use Euclidean distances to define spatial neighborhoods. Kou et al. developed spatial weighted outlier detection algorithms which use properties such as center distance and common border length as weight when comparing non -spatial attributes [2]. Adamet al. proposed an algorithm which considers both spatial relationship and semantic relationship among neighbors [5]. Liu et al. proposed a method for detecting outliers in an irregularly-distributed spatial data set [11].

(2) Graph-based Approach

Graph-based Approach uses graph connectivity to define spatial neighborhoods. Yufeng Kou et al. proposed a set of graph-based algorithms to identify spatial outliers, which first constructs a graph based on k-nearest neighbor relationship in spatial domain, assigns the non-spatial attribute differences as edge weights, and continuously cuts high-weight edges to identify isolated points or regions that are much dissimilar to their neighboring objects. The algorithms have two major advantages compared with the existing spatial outlier detection methods: accurate in detecting point outliers and capable of identifying region outliers [23].

IV. RECENT ADVANCEMENTS IN OUTLIER DETECTION

4.1 SLOF

Local Outlier Factor was proposed by Markus M. Breunig, Hans-Peter Kriegel, Ray-mond T. Ng and Jörg Sander. This method detects outlier by measuring the local deviation of a given data object with respect to its neighbors. Local outlier factor is based on the concept of local density. The object's neighbor is composed of the object's k-nearest neighbors. SLOF method is a density based out-lier detection method, the outliers detected by SLOF are local outliers. Based on the feature bagging approach, the SLOF method is robust and not quite sensitive to parameter k. The dimensions of the vector describe the features of the object. The objects' local density is calculated by the distances between objects. Finally, SLOF score of each object. If an object's SLOF score is approximate to 1, the object is a normal one, and if an object's SLOF score is significantly larger than 1, the object is an outlier[24].

4.2 Non-Parametric Composite Outlier Detection

Detection of the existence of data streams drawn from outlying distributions among data streams drawn from a typical distribution is investigated. It is assumed that the typical distribution is known and the outlying distribution is unknown. The generalized likelihood ratio test (GLRT) for this problem is constructed. With knowledge of the Kullback-Liebler divergence between the outlier and typical distributions, the GLRT is shown to be exponentially consistent (i.e, the error risk function decays exponentially fast). It is also shown that with knowledge of the Chernoff distance between the outlying and typical distributions, the same risk decay exponent as the parametric model can be achieved by using the GLRT. It is further shown that, without knowledge of the distance between the distributions, there does not exist an

exponentially consistent test, although the GLRT with a diminishing threshold can still be consistent[25].

V. CONCLUSION

This paper mainly discusses about outlier detection approaches from data mining perspective. Firstly, we reviews related work in outlier detection. Then, we compare and discuss different algorithms of outlier identification which can be classified based on two categories: classic outlier approach and spatial outlier approach. The classic outlier approach analyzes outlier based on transaction dataset, which can be grouped into statistical-based approach, distance-based approach, deviation-based approach, density-based approach.

The spatial outlier approach analyzes outlier based on spatial dataset, which can be grouped into space based approach, graph-based approach. Thirdly, we conclude some advances in outlier detection recently.

References

[1] Agarwal, D., Phillips, J.M., Venkatasubramanian, "The hunting of the bump: on maximizing statistical discrepancy". In: Proc. 17th Ann. ACM-SIAM Symp. On Disc. Alg. pp. 1137–1146 (2006).

[2] Y. Kou, C.-T. Lu, and D. Chen. "Spatial weighted outlier detection". In Proceedings of the Sixth SIAM International Conference on Data Mining,pp. 614–618, Bethesda, Maryland, USA, 2006.

[3] Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, 2005, vol. 14, pp. 211-221.

[4] Lazarevic, A., Kumar" Feature Bagging for Outlier Detection". In: KDD (2005).

[5] N. R. Adam, V. P. Janeja, and V. Atluri., "Neighborhoodbased detection of anomalies in high-dimensional spatiotemporal sensor datasets". In Proceedings of the 2004 ACM symposium on Applied computing, Nicosia, Cyprus, 2004. pp. 576–583

[6] S. C. Shashi Shekhar, "Spatial Databases: A Tour. Prentice Hall", 2003.

[7] Papadimitriou, S., Kitawaga, H., Gibbons, P., Faloutsos, C., "LOCI: Fast outlier detection using the local correlation integral", Proc. of the Int'l Conf. on Data Engineering, 2003.

[8] Chang-Tien Lu, Dechang Chen,Yufeng Kou, "Detecting spatial outliers with multiple attributes", Tools with Artificial Intelligence, 2003. Proceedings. 2003, pp.122–128.

[9] Yu, D., Sheikholeslami, G. and Zang, "A find out: finding outliers in very large datasets". In Knowledge and Information Systems, 2002, pp.387-412.

[10] Jin, W., Tung, A.K.H., Han, J.W. "Mining Top-n Local Outliers in Large Databases". In: KDD (2001).

[11] H. Liu, K. C. Jezek, and M. E. O'Kelly, "Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and gis". International Journal of Geographical Information Science,15(8), 2001. pp.721–741.

[12] Aggarwal, C.C, Yu, P. "Outlier detection for high dimensional data", Proceedings of the ACM SIGMOD International Conference on Management of Data. Santa Barbara, CA, 2001, pp. 37-47.

[13] Breunig, M.M., Kriegel, H.P., and Ng, R.T., "LOF: Identifying density-based local outliers." ACM Conference Proceedings, 2000, pp. 93-104.

[14] E. Knorr, R. Ng, and V. Tucakov, "Distance-Based Outlier: Algorithms and Applications," VLDB J., vol. 8, nos. 3-4 2000, pp. 237-253.

[15] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets,"Proc. Int'l Conf. Management of Data (SIGMOD '00),2000, pp. 427-438.

[16] Knorr, E.M., Ng, R.T., "Finding Intentional Knowledge of Distance-Based Outliers", Proceedings of the 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, pp.211-222, September 1999.

[17] Berchtold, S., Keim, D., Kriegel, H.-P, "The X-tree: An efficient and robust access method for points and rectangles". In: VLDB (1996).

[18] Rousseeuw, P. & Leroy, A. (1996).,"Robust Regression and Outlier Detection", 3rd edn. John Wiley & Sons.

[19] A. Arning, R. Agrawal, and P. Raghavan, "A Linear Method for Deviation Detection in Large Databases," Proc. Int'l Conf. Knowledge Discovery and Data Mining, 1996, pp. 164-169.

[20] Barnett, V. & Lewis, T. (1994).,"Outliers in Statistical Data", 3rd edn. John Wiley & Sons.

[21] D. M. Hawkins, "Identification of Outliers". Chapman and Hall, London, 1980.

[22] Yamanishi. K, Takeuchi. J ,and Williams. G On-line, "unsupervised outlier detection using finite mixtures with discounting learning algorithms". In Proceedings of the Sixth ACM SIGKDDOO, Boston, MA, USA, pp.320-324.

[23] Yufeng Kou, Chang-Tien Lu, Dos Santos, R.F." Spatial Outlier Detection: A Graph-Based Approach", ICTAI 2007, Volume 1, 2007,pp.281 – 288.

[24] Haowen Guan, Qingzhong Li, Zhongmin Yan, Wei Wei." SLOF: Identify Density-based Local Outliers in Big Data ", 2015 12th Web Information System and Application Conference.

[25] Weiguang Wang, Yingbin Liang , H. Vincent Poor "Nonparametric Composite Outlier Detection" 2016 IEEE