IJDCST @Feb-March Issue- V-2, I-3, SW-01
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

# Countering Data Miss Usability Using Record Distinguishing Factors

**Ms. N. Madhu Bindu**
**Dept. Computer Science & Engineering,**
**Devineni Venkata Ramana & Dr. Hima Sekhar MIC College of Technology, Vijayawada, India.**

**Abstract:** Assigning a Measurability weight to a given dataset is strongly related to the way the data is presented (e.g., tabular data, structured or free text) and is domain specific. We define misusability assessment data events in real time data records present in domain expert with publishing data with other users. Records Ranking [LR] handles unknown values well with an assumption is made that the relationships between the attributes and the dependent variable are linear. In this paper we propose to develop Angular Metric for Shape Similarity (AMSS) for calculating the individual misusability score in reliable data record with different users. AMSS is designed to recognize similarity between shapes with limited influence on spatial variation and not to be much affected by outliers.

**Index Terms: Misuseability-based access control, SVM Classifier, K-Anonymity, Privacy Preserving.**

## I. INTRODUCTION

Data represent today an important asset for companies and organizations. To assure the security and privacy of data assets is a crucial and very difficult problem in our modern networked world. [7] Privacy preservation has become a major issue in many data mining applications. Overall, data security has a central role in the larger context of information systems security. The development of Database Management Systems (DBMS) with high- assurance security is a central research issue; it requires a revision of architectures and techniques adopted by traditional DBMS. [2] Relational database management systems (RDBMS) are the fundamental means of data organization, storage and access in most organizations, services, and applications. Most statistical solutions concern more about maintaining statistical invariant of data.

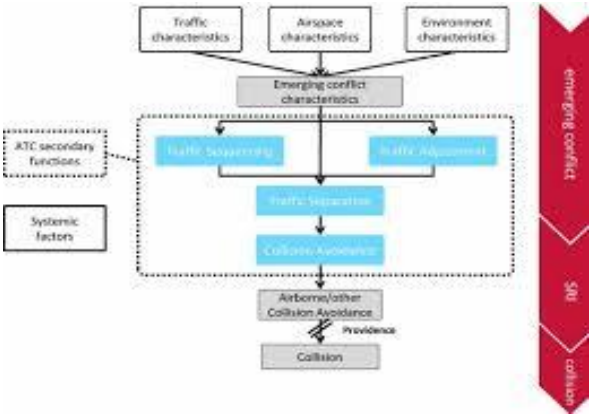


**Figure 1: Data assessment process in real time database applications.**

The ubiquity of RDBMSs led to the prevalence of security threats against the systems. Consider, an intruder from the outside may be able to gain unauthorized access to data by sending carefully crafted queries to a back-end database of a Web application. The data mining community has been studied at building strong privacy-preserving models and designing efficient optimal and scalable heuristic solutions.

Traditionally more number of techniques were introduced for doing data assessment in real time database applications. The most commonly used technique was used for representing user behavioral profiles for analyzing the user behavioral with semantic data representation in various applications. And analyzing actual data representations of each user perspectives with profile registration. These methods consider the different sensitive levels of attributes within exposed.

This factor has a great impact in estimating the damage that can be caused to an organization when data is leaked or misused. Security-related data measures including *k*-Anonymity, *l*-Diversity and (*α*,*k*)-Anonymity are mainly used for privacy-preserving and are not relevant when the user has free access to the data. Therefore, we present a new concept, Misuse ability Weight, which assigns a sensitivity score to datasets, thereby estimating the level of harm that might be inflicted upon the organization when the data is leaked.

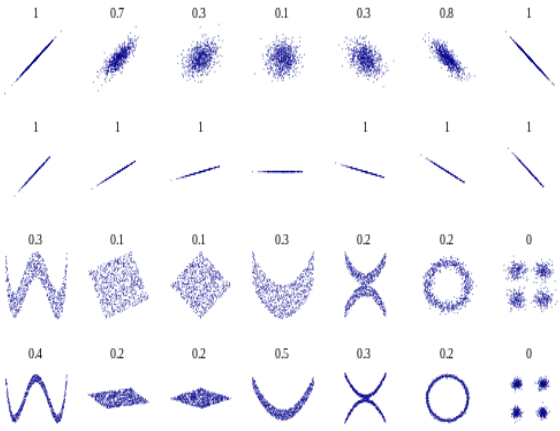


**Figure 2: Correlation time series similarity measures in data assessment process.**

For doing these aspects efficiently, in this paper we introduce to develop a similarity measure called the Angular Metric for Shape Similarity (AMSS). This technique can be help to retrieve relevant results regarding the data sets present in the database applications. A time series is a sequence of time intervals present in the data base applications. Such time series data can arise in any disciplines, such as agriculture, chemistry, demography, and finance. Privacy is the necessary from different data record with sequential format generation to various users present in preferred data sets.

## II. EXISTING SYSTEM

Mitigating leakage or misuse incidents of data stored in databases (i.e., tabular data) by an insider having legitimate privileges to access the data is a challenging task. Security-related data measures including k-Anonymity, l-Diversity and [8] (α, k)-Anonymity are mainly used for privacy-preserving and are not relevant when the user has free access to the data. The most common approach for representing user behavioral profiles is by analyzing the SQL statement submitted by an application server to the database (as a result of user requests), and extracting various features from these SQL statements. [11] The main goal of this experiment was to find whether the M-score fulfills its goal of measuring misusability weight. An implementation of the above approach validates the current systems efficiency in identifying the potential data misuse.

**Records Ranking**

In this approach, the domain expert is requested to assign a sensitivity score to individual records.

Thus, the domain expert expresses the sensitivity level of different combinations of sensitive values. Records Ranking [LR] handles unknown values well with an assumption is made that the relationships between the attributes and the dependent variable are linear. Pair wise Comparison [AHP] uses analytic hierarchy process AHP tree structures and hence produces optimized results faster. Records Ranking [CART] makes no assumption that the relationships between the attributes and the dependent variable are linear and hence takes much time. Records Ranking [LR] and Pair wise Comparison [AHP] significantly outperformed the Records Ranking [CART] in expert scoring and hence the chosen mode of knowledge model. Knowledge acquired from one expert is sufficient to calculate the M-score for the entire domain.

## III. PROPOSED SYSTEM

Traditional techniques are used one domain expert with semantic data natures in sensitive data records. Also we plan to extend the computations of sensitivity level of sensitive attributes to be objectively obtained by using machine learning techniques such as SVM classifier along with expert scoring models.

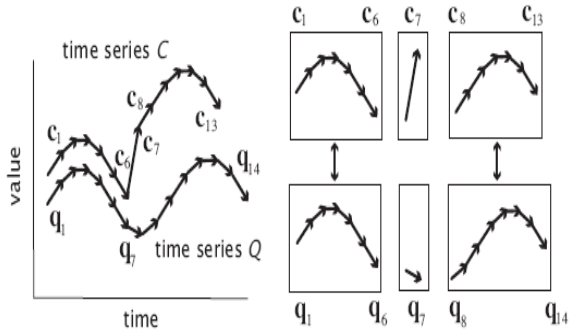


**Figure 3: Two time series data compared by AMSS.**

AMSS treats a time series as vector sequence. AMSS calculates similarity between two vector sequences based on vectors' directions, not on the actual locations of the data points in the d-dimensional space. The above figure explains individual arrangement of time efficiency which includes 1-dimensional feature extends and figure 3 defines C contains successive vectors c1 to c6 which appear to be similar to vectors q1 to q6 in Q. Although subsequent vectors, c7 to c13, in C are not quite similar to q7 to q14 in Q partly due to a sudden increase by c7, AMSS will be less affected by the large discrepancy. Although c7 and q7 are quite different in direction and would (locally) lead to low similarity, AMSS can properly com pute global similarity between C and Q by pairing similar subsequences and by focusing on the shapes of the subsequences represented by vector directions.

## IV. PERFORMANCE ANALYSIS

In this section we describe the efficiency of the user intension with the description of various data records present real time data applications. In this paper we construct data analysis of sensitive records. These records are sensitive and quasi identifier attributes representation. Data client data can be achieved with systematic formation related to mobile communication representation.

| Cid | Name | Last Name | Average Bill | Account |
|---|---|---|---|---|
| 1 | Ernest | Velasquez | 991.0 | Gold |

| | | | | |
|---|---|---|---|---|
| 2 | Wayne | Guerrero | 973.0 | Gold |
| 3 | Mayo | Share | 258.0 | White |
| 4 | Clint | Hernandez | 965.0 | Silver |

**Table 1: Client data for accessing services based on m-score representation.**

As shown in the above diagram, it shows relevant data into publication purpose to use data event generation. This data presents the relative event generation of the each client present in systematic procedures.

**4.1. Measurability Weight Measure:** Data stored in organizations into different with extremely imported data that embedded with organizations power distribution. On the other hand, this data is necessary for daily work processes. Users within the organization's perimeter (e.g., employees, sub-contractors, or partners) perform various actions on this data (e.g., query, report, and search) and may be exposed to sensitive information embodied within the data they access**.**

**4.2. M-Score Measure:** To measure the m-score in different aspects related to misusability present imported data. The *M*-score measure is tailored for tabular datasets (e.g., result sets of relational database queries) and cannot be applied to non-tabular data such as intellectual property, business plans, etc.

| Client | Existing | Proposed |
|---|---|---|

| data request | Approach(M-Score) Formation | Approach(M-Score formation) |
|---|---|---|
| 1 | 1.9654 | 1.2365 |
| 2 | 2.7896 | 2.1234 |
| 3 | 0.9874 | 0.6984 |
| 4 | 5.1234 | 3.1654 |

**Table 2: Comparison results of the exisitng and proposed acceptance of the data representation.**

Calculating m-score for combining table contents as follows:

$$\boldsymbol{MScore} = r^{1/x} \times RS = r^{1/x} \times \max_{0 \le i \le r} \left( \frac{RRS_i}{D_i} \right).$$

Where *r* is the number of records in the table, *x* is a given parameter and *RS* is the final Record Score presented in original data records.

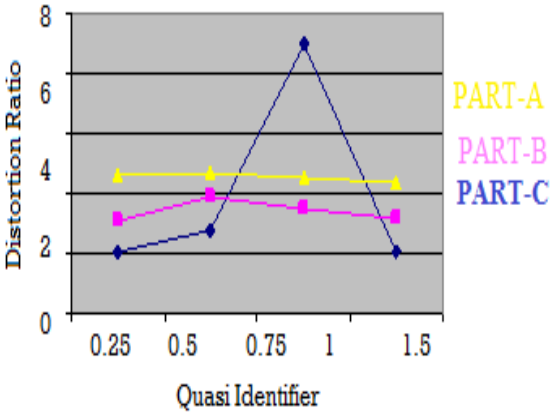


**Figure 1: Distortion Ratio versus Quasi identifier.**

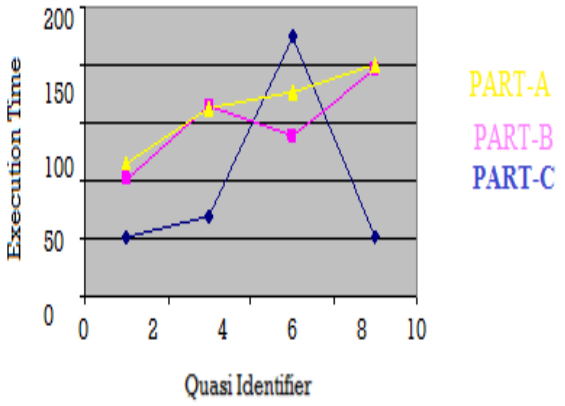


**Figure 2: Execution Time and Distortion Ratio versus Quasi identifier Size and a (k = 2).**

The above diagram shows the execution procedure for identifying quasi identifier in our presented data set by comparison of each record attribute values. [5] In this section we are measuring the each tuple space allocation. Initially it is α is greater gradually it is association with generalized values present in our data set.

**Data Assessment:** This is the procedure in accessing services from data publisher, if we want data records with simple and other data accessing services. In this section we give assessment process to user who have these requirements efficiently, then we calculate the system process of other data representation to find the duplicate and tuples records of original user present in the data assessment process in sufficient record generations. In this section we describe 4 types of questionnaires' are described to find the user intension in realistic event generation they are Part A, Part B, Part C, Part D. In Part A processing we identify the quasi identifier generation and in Part B declares the sufficient sensitive attributes presentation with quasi attributes and sensitive attributes. Part D selects the system process with suitable generation.

These comparisons gives unique results between each tuple, and find the preserving results of each user based on his/her intention for our analysis. This overall observation gives average companion time with distortion ratio and Quasy identifiers. Finally our approach mainly focused on user behavior to the inter related environment in each user. In this paper we also calculate and define individual m-score measurements with suitable process generated in the data representation. Time series is an important aspect in present days. These results are developed with systematic data events. This techniques can also developed individual data assessment process generation with real time data record can be stored in data records.

For doing this work efficiently, in this paper we introduce Angular Metric for Shape Similarity (AMSS). The followings are the constraints AMSS also complies with. The simplest form of ensemble learning trains each classifier separately and integrates the output of all classifiers by, for example, voting. This overall observation gives average companion time with distortion ratio and quasi identifiers.

**Input:** Data sets

**Output:** Publishing record generation

**Step 1:** Import datasets into publisher account.

**Step 2:** Calculate Boundary condition

w1 = (1, 1) and wK = (N,M).

**Step 3:** This constraint requires W to lie across (diagonally) adjacent cells. More formally, given wk = (nk,mk) and wk+1 = (nk+1,mk+1), nk+1 − nk and mk+1 − mk equal either 0 or 1.

**Step 4:** Calculates individual time series generations with systematic data items.

**Step 5:** Publish data with individual data items with questionnaires present in data records.

**Step 6:** Publish data items.

**Algorithm 1: Time individuals with semantic data items.**

AMSS is designed to recognize similarity between shapes with limited influence on spatial variation and not to be much affected by outliers. When AMSS is used for classification, both of the solutions can be implemented in the framework of ensemble learning, which is used to integrate multiple classifiers and generally produces superior performance to single classifiers. The simplest form of ensemble learning trains each classifier separately and integrates the output of all classifiers by, for example, voting. However, this framework is not effective when the accuracy of a classifier varies on different data. More effective learning framework considers the interaction among multiple classifiers.

## V. CONCLUSION

To handle the leakages, to prevent, we have an improved process to identify the incidents by other misuse detection systems by enabling the security officer to focus on incidents involving more sensitive data. The misusability can propose the four optional usages: The data actually exposed by applying the anomalies detection by the learning the normal behavior to an insider in terms of the sensitivity. In this paper we propose to develop Angular Metric for Shape Similarity (AMSS) for calculating the individual misusability score in reliable data record with different users. AMSS is designed to recognize similarity between shapes with limited influence on spatial variation and not to be much affected by outliers. When AMSS is used for classification, both of the solutions can be implemented in the framework of ensemble learning, which is used to integrate multiple classifiers and generally produces superior performance to single classifiers.

## VI. REFERENCES

[1]Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu and Ke Wang, "(α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy-Preserving Data Publishing", KDD'06, August 20–23, 2006.

[2] K. Lefebvre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD Conference*, pages 49–60, 2005.

[3] A. Machanavajjhala, J. Gehrke, and D. Kifer. l-diversity: privacy beyond k-anonymity. In *To appear in ICDE06*, 2006.

[4] M. Bishop and C. Gates. "Defining the insider threat," *Cyber Security and Information Intelligence Research*, 1-3, 2008

[5] Q. Yaseen, and B. Panda, "Knowledge Acquisition and Insider Threat Prediction in Relational Database Systems," *Computational Science and Engineering*, pp. 450-455, 2009.

[6] M.E. Nergiz, et al. "Multirelational k-Anonymity," *IEEE Trans. on Knowledge and Data Engineering*, 21(8):1104-1117, 2009.

[7] C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy preserving data publishing: A survey on recent developments," *ACM Computing Surveys*, 42(4), 2010.

[8] Amir Harel, Asaf Shabtai," M-Score: A Misusability Weight Measure", IEEE Transactions on Dependable and Secure Computing May/June 2012 (vol. 9 no. 3) ISSN: 1545-5971.

[9] Y. Yuan, et al. "Evolution of Privacy-Preserving Data Publishing," *Anti Counterfeiting Security and Identification*, 34-37, 2011.

[10] Prerna Jawdand, Prof. Girish Agarwal, Prof. Pragati Patil," Implementation Of Data Leakage Detection Using Agent Guilt Model", International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 1, January- 2013

[11] K.Sundaramoorthy,, Dr.S.Srinivasa Rao Madhane," Efficient Method of Detecting Data Leakage Using Misusability Weight Measure", www.ijceronline.com ||April||2013||.

[12] Tetsuya Nakamura · Keishi Taki · Hiroki Nomiya · Kazuhiro Seki · Kuniaki Uehara," A Shape-based Similarity Measure for Time Series Data with Ensemble Learning", In Proc. of the IEEE 23rd International Conference on Data Engineering, pages 786–795, 2007.