Indonesian Journal of Electrical Engineering and Computer Science
Vol. 15, No. 1, July 2019, pp. 527~534
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v15.i1.pp527-534

Protecting sensitive information utilizing an efficient association representative rule concealing algorithm for imbalance dataset

Mylam Chinnappan Babu, Sankaralingam Pushpa
Dept. of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, India

Article Info

Article history:

Received Jan 2, 2019
Revised Mar 2, 2019
Accepted Mar 10, 2019

Keywords:

Accuracy
Direct Discrimination
Discrimination
EARRC (Efficient Association Representative Rule Concealing algorithm)
Indirect Discrimination
NMI (Normalized Mutual Information)
RR (Representative Rule)

ABSTRACT

In data mining, discrimination is the detrimental behavior of the people which is extensively studied in human society and economical science. However, there are negative perceptions about the data mining. Discrimination has two categories; one is direct, and another is indirect. The decisions depend on sensitive information attributes are named as direct discrimination, and the decisions which depend on non-sensitive information attributes are called as indirect discrimination which is strongly related with biased sensitive ones. Privacy protection has become another one of the most important problems in data mining investigation. To overcome the above issues, an Efficient Association Representative Rule Concealing (EARRC) algorithm is proposed to protect sensitive information or knowledge and offer privacy protection with the classification of the sensitive data. Representative rule concealing is one kind of the privacy-preserving mechanisms to hide sensitive association rules. The objective of this paper is to reduce the alternation of the original database and perceive that there is no sensitive association rule is obtained. The proposed method hides the sensitive information by altering the database without modifying the support of the sensitive item. The EARRC is a type of association classification approach which integrates the benefits of both associative classification and rule-based PART (Projective Adaptive Resonance Theory) classification. Based on Experimental computations, proposed EARRC+PART classifier improves 1.06 NMI and 5.66 Accuracy compared than existing methodologies.

Corresponding Author:

Mylam Chinnappan Babu,
Department of Computer Science and Engineering,
St. Peter's Institute of Higher Education and Research,
Chennai, Tamil Nadu, India.
Email: mcbabu.phd@gmail.com

1. INTRODUCTION

The word discrimination invents from the Latin discriminate, which means to differentiate among discrimination functionalities. The social and financial discrimination is the unfair treatment of people on the basis of their type. At investigation part, the discrimination has become an issue in credit, finance, insurance, labor marketplace, education and other human being actions which has attracted much investigator preference in financials and social science. There are numerous decision-making processes available and it offers themselves to discrimination, e.g., education, loan granting, health insurances and employee's selection. An automated framework decides whether the customer is to be suggested for credit or some kinds of life insurance in a specific set of data items for the available customer.

Problem: Privacy protection has become one of the most important problems in data mining. Several privacy-preserving data mining mechanisms have been proposed in which the existing literature is based on

either a cryptographic or a statistical method. Privacy-preserving association rule protects sensitive data item from unnecessary or illegal discovery. The secure multi-party method utilized in the cryptographic mechanism which ensures strong confidentiality and accurateness. However, the technique usually suffers from privacy and time complexity. Most existing methods are utilized for resolving discrimination issues such as preprocessing, in-processing and post-processing approach. Generally, Rule-based frameworks are deployed IDPSes (Internally Displaced Persons) and achieve better results; when the signature data is precise. The rules derived from them which are accurately built by a rule generator. However, a physical attack is determined in derived rules and it alters using few previous rules.

Background: In [1] explained direct and indirect system-level discrimination in the training information. The method proposed in this work for expands the non-discrimination outcome from the training information for data prediction. The group-level direct discrimination and individual-level direct discrimination were studied. In [2] addressed the two-phase co-occurrence association rule mining method to recognize implicit aspects. It contained two stage rule generation. The first stage of rule generation was happened in an explicit ruling in the corpus for every opinion words. The second stage of rule application was clustered the rule consequent (explicit attributes) to create more robust rules for each opinion word. In [3] discussed a discrimination discovery approach that depends on modeling of possibility sharing of a context utilizing Bayesian networks. It computed the consequence of a protected feature in a subset of the dataset. A classification technique corrected the determined discrimination without utilizing protected features in the decision process. In [4] explained a Data Envelopment Analysis (DEA) that evaluates the rank of association rules with various kinds of criteria's for example as support and confidence. In [5] discussed data transformation approaches such as rule protection and rule generalization which depends on direct and indirect discrimination with numerous discriminatory products.

In [6] described sensitive attributes like gender, religion, race, etc. that influence the discriminatory decisions. The decisions were made on the basis of biased sensitive attributes and non-sensitive attributes. In [7] addressed a causal Bayesian networks technique; where, the method captured discrimination based on a legally grounded situation testing methodology. The method utilized the causal Bayesian Networks and associated with causal inference guidelines. In [8] focused on the cleaning and outsourcing of training datasets using legitimated classification rules to extract the discriminating rules. Legitimated classification rules utilized to predict intrusion, fraud or crimes; where to be highly focused on sensitive attributes. In [9] reviewed discrimination and estimated the performance of discrimination aware predictive models. It reviewed and discussed for measuring the procedures and expressed the recommendations for practitioners in the domain of data mining, machine learning, pattern recognition, statistical modeling; that are developing non-discriminatory predicative models. In [10] developed a discrimination-aware data mining (DADM) method for deriving the patterns. The technique does not discriminate "unjust grounds" like gender, ethnicity or nationality.

In [11] illustrated a concurrent chronic disease in the course of treatment, it took two types of comorbid datasets as resultant input. Several popular machine learning techniques such as Logistic Regression (LR), Random Forest (RF), etc. applied to build predictive models. In [12] explained discrimination aware association rule classifier (DAAR) is used to filter out the discrimination issues. Discrimination aware measurements are incorporated and associated with rule mining algorithm. In [13] surveyed various discrimination discovery and discrimination prevention methods to identify the feature and limitation of technique. The paper has explained the antidiscrimination technique for compromising the discrimination discovery and prevention. In [14] discussed the evaluation results on over four types of discrimination, i.e., direct discrimination, indirect discrimination, individual-level discrimination, and group-level discrimination. The technique preferred casual networks to capture the existence of discrimination patterns that provided quantitative evidence of discrimination in decision making. In [15] elaborated the WEKA workbench and organized data preprocessing tools for state-of-art machine learning algorithms. The system offers a convenient graphical user interface for data exploration, larger implementations setup on distributed computing environments with configured streaming for data processing. In [17] illustrated integration of Adaptive Weight Ranking Policy (AWRP) with intelligent classifiers (NB-AWRP-DA and J48-AWRP-DA) through dynamic aging feature to enhance classifiers power of prediction. The schemes are utilized to select the best subset of aspects. In [18] studied to detect the best classifiers for class imbalanced health datasets through a price depended comparison of classifier performance. The uneven misclassification prices were characterized in a cost matrix, and cost-benefit. In [19] discussed the WEKA tool for higher education institutes utilize a data mining tools and techniques for academic development of the student performance and to prevent drop out.

Proposed Solution: The research aims to design an Efficient Association Representative Rule Concealing (EARRC) algorithm is proposed for protecting sensitive information or knowledge and offers privacy protection with the classification of the sensitive data. The method Representative rule concealing is one kind of the privacy-preserving mechanisms to hide sensitive association rules. The objective of this method

is to reduce the alternation of the original database and perceive that there no sensitive association rule is obtained. The proposed method hides the sensitive information by altering the database without modifying the support of the sensitive item. The technique is used to enhance the domain of the lost rule and ghost rule side effects. The lost rule is hiding sensitive rules completely. It is not affected the non-sensitive rules. In hiding process, no extra fake rules are incorrectly extracted; it is called Ghost rule. It is an evolutionary mechanism to resolve the compound issues and require optimal sanitization. Degradation of information is computed in two dimension aspects. The first dimension computes the confidential information protection and second calculates the loss of functionality. The proposed work discusses effective mechanism for privacy preservation and discrimination prevention to be deployed. The EARRC is a type of association classification approach which integrates the benefits of both associative classification and rule-based PART classification. The PART is a rule-based classifier to predict the performance. The method prevents discrimination prevention and improves the accuracy:

a. To develop Efficient Association Representative Rule Concealing (EARRC) algorithm that is utilized for protecting sensitive information or knowledge and to hide sensitive association rules.
b. To offer privacy preservation with the prediction of the sensitive data
c. To alter the original database and perceive that there is no sensitive association rule obtained.
d. To compute the confidential information protection and the missing functionality.
e. To improve the Normalized Mutual Information (NMI) and Accuracy compared than their existing methods

The rest of paper is organized as: Section 2 describes the literature study with the closest conventional method. Section 3 describes the proposed methodology with implementation details. Section 4 discusses implemented result and comparative study with the conventional technique.

2. RESEARCH METHOD

This research work proposes an Efficient Association Representative Rule Concealing (EARRC) algorithm to protect sensitive information or knowledge for hiding sensitive association rules and offering privacy protection with sensitive data predictions. EARRC is divided into following modules like loading data, preprocessing of data, Frequent Itemset Generation, rule generation, Classification, EARRC Algorithm. The workflow diagram of the proposed system is illustrated in Figure 1 stepwise.

2.1. Implementation Pre-processing Steps

2.1.1 Loading Data

Loading data is a process to browse the biased data set in the proposed framework. The data contains the file name, file size, time, the total number of attributes, and the total number of records. The method predicts the attributes of sensitive information that contains the column; attribute name, description.

2.1.2 Preprocessing and Data Cleaning

The method processes the data with discriminatory biases that is comprised of the original sensitive information. It eliminates zero unfair decision rules which can be extracted from the transformed sensitive information. The method acquires discrimination free information and applies some standard data mining algorithm. The sensitive information transformation and frequent item set generalization can be adapted from the privacy preservation utilizing EARRC methodology.

2.1.3 Frequent Item set Generation

The EARRC algorithm extracts the recurrently occurring item sets in a specific biased data set. The input is a set of transactions with sensitive items, and the output is the sensitive items with a constraint confident of item sets. It generates a set of candidate item sets and counts.

2.1.4 Representative Rule generation

Representative Rule generation is generating the improved privacy of association rules for each frequent item set; where, each rule is a binary partition of a frequent item set. The method considers reliable, sensitive information and creating a universal statement of each item. The EARRC technique evaluates common ideas by abstracting the general properties (name, country, profession, DOB, income, addresses, etc.) form of the training dataset. The method applies nominal attribute of the biased data set and transforms numeric feature into a range of information.

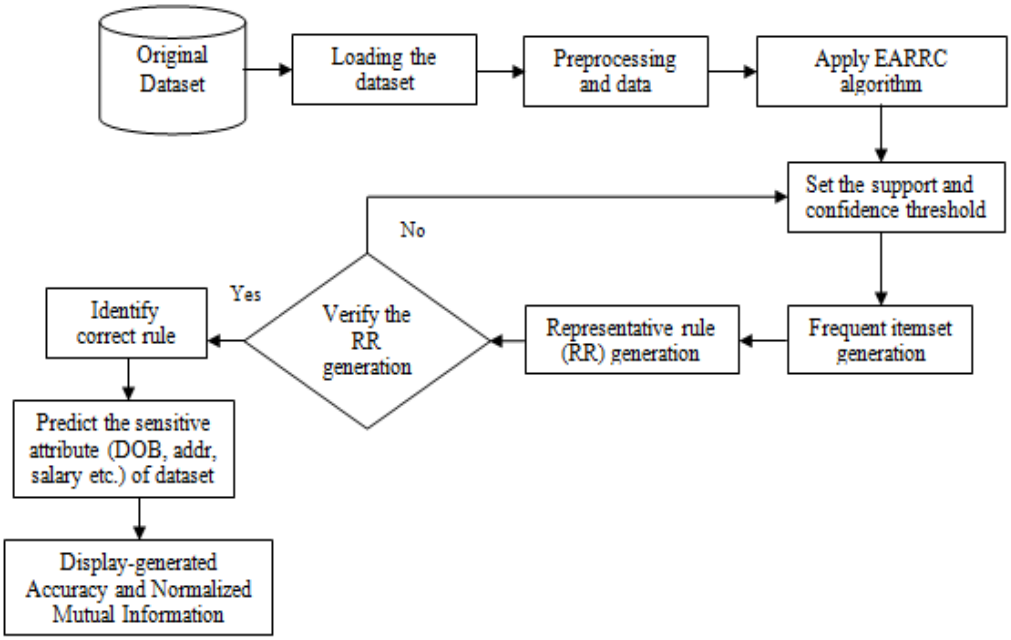

Figure 1. Workflow diagram of the proposed system

2.1.5 Data Prediction

Data prediction is processed to compute group assignments or membership for sensitive information occurrences of the training dataset. The prediction is evaluated with the reference of the original data set. The aim of the classification mechanism to analyze the input data set and develop a correct model for every grouping attributes which are available in the present in the sensitive information.

2.2. Efficient Association Representative Rule Concealing (EARRC) Algorithm

The Efficient Association Representative Rule Concealing (EARRC) algorithm is implemented to protect sensitive information or knowledge for hiding sensitive association rules and offering privacy protection with sensitive data predictions. The rules are described in representative rules (RR) sensitive data on the left or right-hand side of the rules. The technique selects a rule from the set of RR's which comprises sensitive data. The method selects database operations which includes all the sensitive data in the RR. The proposed EARRC method hides the sensitive data by altering the database without modifying the support of the sensitive data.

The association rules are determined in a given dataset. RR is a set of rules which allows for assuming all association rules without accessing a data set. The cover operator C initiated for a dynamic set of association rules from a provided association rule. Representative rules creating process is decomposed into two sub-procedures such as frequent item-sets generation and RR prediction from frequent item-sets. The frequent item set is $\varphi \neq A \subset B$. The association rule $A \Rightarrow Z/B$ is the representative rule; if there is no association rule $(A \Rightarrow Z\ '/ A)\ '$. Where $Z \subset Z'$, and there is no association rule $(A\ ' \Rightarrow Z / A')$ such that $A \supset A\ '$. A set of representative rules (RR) for a provided association rules (AR) can be described as (1).

$$RR = \{r \in AR | \neg \exists r' \in AR, r' \neq r \ and \ r \in C(r') \quad (1)$$

The C is the Candidate item set. Every rule in RR is called representative association rule. There is no representative rule may suitable in the coverage of another association rule. An imbalanced biased dataset, minimum support, and confidence are provided as an input of the algorithm.

The pseudo code of proposed algorithm is given below in details:

The Input: S is an imbalanced biased data set, mi_support, mi_confidence, and F is a set of sensitive data items.

Output: A transformed database S' where representative rules (RR) including F and visualize Normalized Mutual Information (NMI) and Accuracy

Procedure:

```
Start;
        Compute item sets from Dataset S;
        Every sensitive data item f∈ F;
        {
            If f is a small item set then
                F=F-{f};
            If F is null then EXIT;
                Select a representative rule RR from the dataset;
                Arrange RR in descending order by supported items;
                Choose r (association rule) from RR
                Estimate confidence of rule r;
            If conf>mi_conf then
                { //modify the place of sensitive information item f.
                Find Ti = {t (subset) in S |t completely supports RR;
            If t comprises attribute and f then
                Eliminate f from t;
            Else
                Find Ti = {t in S|t does not support and partially supports attributes;
                Add f to t
                Select the first rule from RR;
                Compute confidence of r;
                Until (RR is empty);
            End If
            If conf>mi_conf
                Update S with new item transaction t;
                Calculate and visualize Normalized Mutual Information (NMI)
                and Accuracy
            Else
                It failed to compute and visualize Normalized Mutual Information (NMI)
                and Accuracy
            End If
            End
```

3. RESULTS AND ANALYSIS

3.1. Programming Environment

The implementation work is deployed on Intel i6th processor, 8 GB RAM and 500 GB memory with the windows7 ultimate operating system. The proposed framework is developed in JAVA programming language, JDK 1.8, NETBEANS 8.0.2, with MYSQL database. The proposed technique is used WEKA library with Dataset.

3.2. Data Set

In The paper utilizes two real datasets, Adult and Dutch Census, from the UCI Repository of Machine Learning Databases. These two datasets are usually utilized in a discrimination investigation. The Adult dataset comprises 48861 tuples (after eliminating those tuples with missing qualities) with 14 attributes. The analytical task is to classify people into high and low salary classes. It is outstanding that various attributes in the Adult dataset are weakly relevant to gender, for example, work class, education, job, race, capital loss, native.

Dutch Dataset: For the Dutch dataset, our fixed hierarchical log-linear model is class variable described whether a people's occupation was high income or low income, and its sensitive attribute illustrated the people's gender. The size of the dataset was 32,584, and the number of non-sensitive attributes was 10. Note that all attributes are categorical and were transformed into multiple binary attributes by a log-linear model (1-of-K) method.

3.3. Normalized Mutual Information (NMI)

The NMI is to measure the results among 0 (no mutual information) and 1 (perfect correlation). NMI is described by normalizing the mutual information into a range [0, 1]. The proposed approach is defined as a mathematical model for Normalized Mutual Information in (2). The Normalized Mutual Information is calculated as:

$$NMI\ (Y,S) = \frac{I(Y;S)}{\sqrt{H(Y)H(S)}} \quad (2)$$

Where I (•; •) and H (•) represent mutual information and entropy, respectively. Where Y is class labels and S is Cluster labels.

3.4. Accuracy

Support Accuracy is defined as computes the ratio of correct or true predictions over the total number of instances estimated. The proposed approach is defined as a mathematical model for accuracy in (3). The accuracy is calculated as:

$$Accuracy = \frac{T_P+T_N}{T_P+F_P+T_N+F_N} \quad (3)$$

TP is true positive values or correctly classified values, and TN is true negative values. FN is false negative values, and FP is false positive values.The proposed EARRC system is computed with following existing methods such as Naïve Bayes (NB) [16], Logistic Regression (LR) [16] and Support Vector Machine (SVM) [16] methods. The proposed EARRC is to protect sensitive information or knowledge. The proposed method also hides sensitive association rules and provides privacy protection with the classification of the sensitive data. Proposed EARRC algorithm is integrated with a rule-based PART classifier to improve the Normalized Mutual Information (NMI) and Accuracy.

According to Figures 2 and 3 observations, the proposed EARRC+PART technique is computed with conventional technique on behalf of Normalized Mutual Information (NMI) and Accuracy. Proposed EARRC+PART algorithm is estimated with Naïve Bayes (NB), Logistic Regression (LR) and Support Vector Machine (SVM) [16] methodologies behalf of on Normalized Mutual Information (NMI) and Accuracy to estimate the efficiency of the proposed technique. The naïve Bayes is a supervised learning classifier, utilizing Bayesian inference and the (often incorrect) assumption that parameters are independent. But, it provides the low Accuracy and NMI for compare than proposed EARRC+PART classifier. Logistic Regression is utilized to explain data and for describing the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It is the nearest competitor on behalf of accuracy. However, it fails to maintain NMI. The SVM is the nearest competitor to a proposed EARRC+PART method for NMI and Accuracy. SVM is supervised learning models with associated learning algorithms that investigate utilized data for classification and regression evaluations. It consumes more time for data processing and does not assure for data accuracy. EARRC+PART algorithm offers the high NMI and Accuracy. Proposed EARRC+ PART improves 1.06 NMI and 5.66 Accuracy. Finally, the paper claims that the proposed EARRC+PART methodology performs best on every evaluation matrix and respective input parameters.

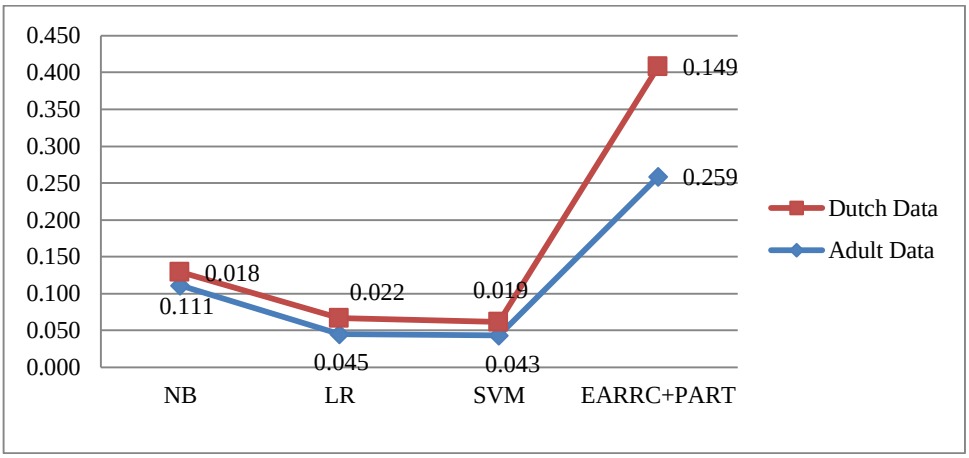

Figure 2. Normalized Mutual Information (NMI) for Adult and Dutch Data set

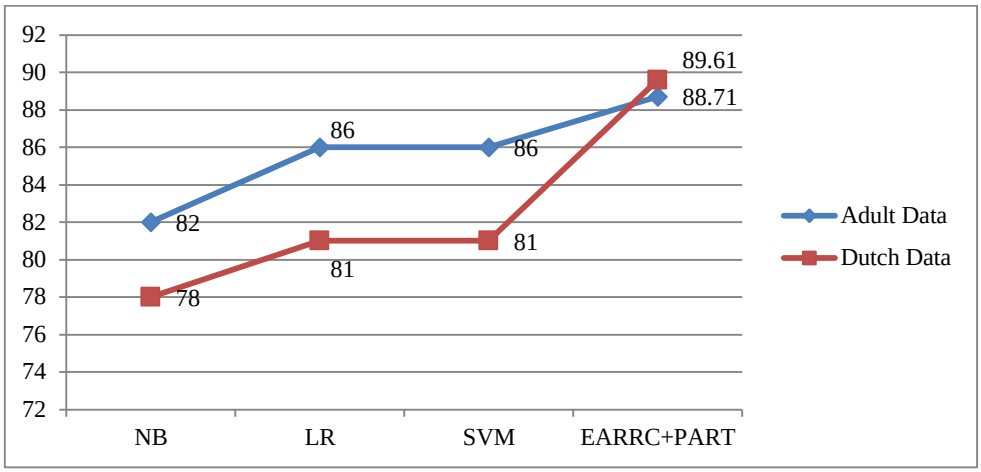

Figure3. Accuracy for Adult and Dutch Dataset

4. CONCLUSION

The paper presents An Efficient Association Representative Rule Concealing (EARRC) algorithm to protect sensitive information or knowledge and provide privacy protection with the classification of the sensitive data. The objective of this paper is to minimize the alteration of the original database and perceive that there is no sensitive association rule is obtained. The proposed method hides the sensitive information by altering the database without modifying the support of the sensitive item. The rules are described in representative rules (RR) sensitive data on the left or right-hand side of the rules. The technique selects a rule from the set of RR's which comprises sensitive data. Representative rules designed two sub-procedures such as frequent item-sets generation and RR prediction from frequent item-sets. Proposed EARRC+PART improve 1.06 NMI and 5.66 Accuracy. Finally, the paper claims that the proposed EARRC+PART methodology performs best on every evaluation matrix and respective input parameters.

In the future, the paper can be improved to apply discrimination technique with content based privacy in an online social network using the Hadoop environment. Due to hues, discrimination occurred in OSN, and it is required to work forward.

REFERENCES

[1] Zhang, L., & Wu, X., "Anti-Discrimination Learning: A Causal Modeling-Based Framework," *International Journal of Data Science and Analytics*, Vol. 4, No. 1, pp. 1-16, 2017.

[2] Hai, Z., Chang, K., & Kim, J. J., "Implicit Feature Identification Via Co-Occurrence Association Rule Mining," *In International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Heidelberg*, pp. 393-404, 2011.

[3] Mancuhan, K., & Clifton, C., "Combating Discrimination Using Bayesian Networks," *Artificial intelligence and law*, Vol. 22, No. 2, pp. 211-238, 2014.

[4] Chen, M. C., "Ranking Discovered Rules from Data Mining with Multiple Criteria by Data Envelopment Analysis," *Expert Systems with Applications*, Vol. 33, No. 4, pp. 1110-1116, 2007.

[5] Hajian, S., & Domingo-Ferrer, J., "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining," *IEEE transactions on knowledge and data engineering*, Vol. 25, No. 7, pp. 1445-1459, 2013.

[6] Hajian, S., Domingo-Ferrer, J., & Martinez-Balleste, A., "Rule Protection for Indirect Discrimination Prevention in Data Mining," *In International Conference on Modeling Decisions for Artificial Intelligence, Springer, Berlin, Heidelberg*, pp. 211-222, 2011.

[7] Zhang, L., Wu, Y., & Wu, X., "Situation Testing-Based Discrimination Discovery: A Causal Inference Approach," *In IJCAI*, pp. 2718-2724, 2016.

[8] Hajian, S., Domingo-Ferrer, J., & Martinez-Balleste, A., "Discrimination Prevention in Data Mining for Intrusion and Crime Detection," *In Computational Intelligence in Cyber Security (CICS), 2011 IEEE Symposium on IEEE*, pp. 47-54, 2011.

[9] Zliobaite, I., "A Survey on Measuring Indirect Discrimination in Machine Learning," arXiv preprint arXiv:1511.00148, 2015.

[10] Berendt, B., & Preibusch, S., "Exploring Discrimination: A User-Centric Evaluation of Discrimination-Aware Data Mining," *In Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on IEEE*, pp. 344-351, 2012.

[11] Zolbanin, H. M., Delen, D., & Zadeh, A. H., "Predicting Overall Survivability in Comorbidity of Cancers: A Data Mining Approach," *Decision Support Systems*, Vol. 74, pp. 150-161, 2015.

[12] Luo, L., Liu, W., Koprinska, I., & Chen, F., "Discrimination-Aware Association Rule Mining for Unbiased Data Analytics," *In International Conference on Big Data Analytics and Knowledge Discovery, Springer, Cham*, pp. 108-120, 2015.
[13] Gonbare, S., Varma, S., & Deshmukh, M., "Survey on Anti-discrimination in Data Mining," 2015.
[14] Zhang, L., Wu, Y., & Wu, X., "On Discrimination Discovery Using Causal Networks," *In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Springer, Cham*, pp. 83-93, 2016.
[15] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L., "Weka-A Machine Learning Workbench for Data Mining," *In Data mining and knowledge discovery handbook, Springer, Boston, MA*, pp. 1269-1277, 2009.
[16] Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J., "Model-Based and Actual Independence for Fairness-Aware Classification," *Data Mining and Knowledge Discovery*, Vol. 32, No. 1, pp. 258-286, 2018.
[17] Olanrewaju, R. F., & Azman, A. W., "Intelligent Cooperative Adaptive Weight Ranking Policy Via Dynamic Aging Based on NB and J48 Classifiers", *Indonesian Journal of Electrical Engineering and Informatics (IJEEI),*; Vol. 5, No. 4, pp. 357-365, 2017.
[18] Rao, R. R., & Makkithaya, K., "Learning from a Class Imbalanced Public Health Dataset: a Cost-based Comparison of Classifier Performance", *International Journal of Electrical and Computer Engineering*, Vol. 7, No. 4, pp. 2215-2222, 2017.
[19] Hussain, S., Dahan, N. A., Ba-Alwi, F. M., & Ribata, N., "Educational Data Mining and Analysis of Students' Academic Performance using WEKA", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 9, No. 2, pp. 447-459, 2018.