Outlier Detection Using Association Rule Mining for Information Quality Improvement

Dwi Welly Sukma Nirad and Kridanto Surendro

***Abstract*—** The abundance of data production in the latest decade encourages researchers to create a variety of opportunities. This shall be interspersed with the strong willing of the society to change their perspective regarding the availability of data. Once data viewed as a dull stack of letters and numbers, they must now be considered as a fresh resource capable of contributing values and benefits for the owners. Good data will produce high-quality information, which then will direct to a good decision-making. One of the efforts to improve data quality is by performing an outlier detection. There are few methods to do that, but none is considered the best knowing each retains its own strength when applied to different cases. Therefore, this research aims at proving the capability of Association Rule Mining method in detecting outlier when applied to a case requiring interpersonal relationship definition.

***Keywords*—** outlier detection, data mining, data science, information quality

I. INTRODUCTION

THE abundance of data these days is living proof of technology capitalism, as supposed to the fact that they are produced every single year whether one realizes it or not. The Guardian states that global data supply reached 2.8 zettabytes (ZB) in 2012, or equivalent to 2.8 trillion gigabytes [1]. International Data Corporation (IDC) research even estimates total amount of data can reaches up to 40 trillion gigabytes in the 2020, doubling up from 2012 [2].

The increase of data amount goes hand in hand with increase of information. Recognizing this fact shall, indeed, arise some kind of awareness for business actors to start using data and information overflow. Thus, it is time to change the way to see data not only as a complement in the company but also as resources that conceive beneficial as well as profits. The activity of guiding data extraction process, which becomes a particular kind of knowledge, now is more ubiquitously known as data science.

The main idea of data science foregrounds a set of problem-solving methods, techniques, and arrangements—all aiming at producing worthwhile knowledge for decision-making [3]. Knowledge-producing data extraction is done using data mining technique.

Outlier detection is one of the data mining task. It is a set of data deviant from others and thus is referred to as some kind of abnormality, oddity, deviance, or anomaly [4]. Moreover, outlier could be caused by failure of measurement, including wrong data entry or different population from the existing data [5]. Although outliers are few, the existence of outliers can influence the decision-making process. Having that in mind, this research will focus on solving problem regarding the association rule mining technique in detecting outlier to produce high-quality information.

Dwi Welly Sukma Nirad is Lecturer, STMIK Indonesia Padang, Indonesia (corresponding author's e-mail: tri.sundara@stmikindonesia.ac.id).

Kridanto Surendro is with the Bandung Institute of Technology, Indonesia.

II. LITERATURE REVIEW

Data mining is an analysis based on dataset (typically big ones) observation, to identify unexpected relation and infer data in new ways understandable by and useful for the data owner [6]. There are some processes considered as the general standard to achieve best practice in data mining, one of which is the Cross-Industry Standard Process Data Mining (CRISP-DM) process, which has been proposed since early 1990s. It is described as follows.

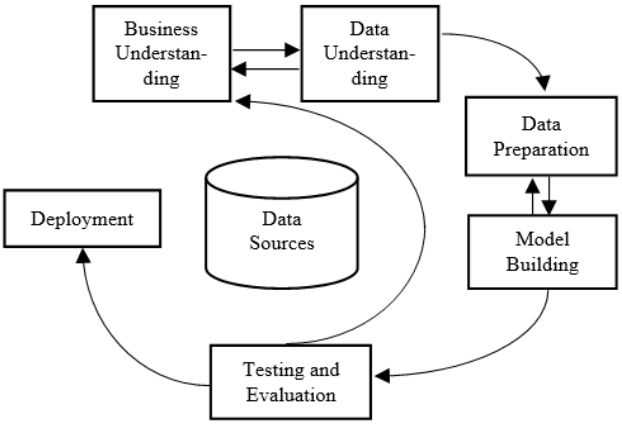

Fig. 1: CRISP-DM Process

Association rules are one of the simple but useful data mining techniques in identifying complicated patterns in large database. Association rule is another way of saying X→Y, where X is a set of items, and Y an another single item [7]. So, X and Y is a disjoint itemset.

When it comes to determining association rule, there is an interestingness measure which is obtained from data processing with particular kind of measurement technique. There divided two kinds:

- Support: part of transactions that containing both X and Y.

$$supp\,(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|} \quad (1)$$

- Confidence: frequency of Y item appears in the X-containing transaction.

$$conf\,(X \Longrightarrow Y) = \frac{supp\,(X \cup Y)}{supp\,(X)} \quad (2)$$

The association rule mining process, which takes account into both measurements above, consists of two mains stages. The stages are as follows [8]:

- Identifying all frequent item sets: each will occur at least as frequent as the predetermined minimum support value.
- Crafting a rigid association rules from the frequent itemsets: rules must always fulfill the minimum support value and the minimum confidence value, thus also known as strong rules.

In identifying the frequent item sets, some common algorithms have been used. In this research, one that is chosen to use is apriori algorithm, which is the basic algorithm for association rule in data mining proposed by R. Agrawal and R. Srikant in 1994 [9].

III. Research Methodology

There are many techniques in data mining commonly used to extract data, including detecting outlier, one of which is association rule mining. However, it is not functioning well in identifying the maximum outlier, and thus a special formula has been added to detect outlier more accurately.

The research model proposed to achieve improved information quality is as follows.

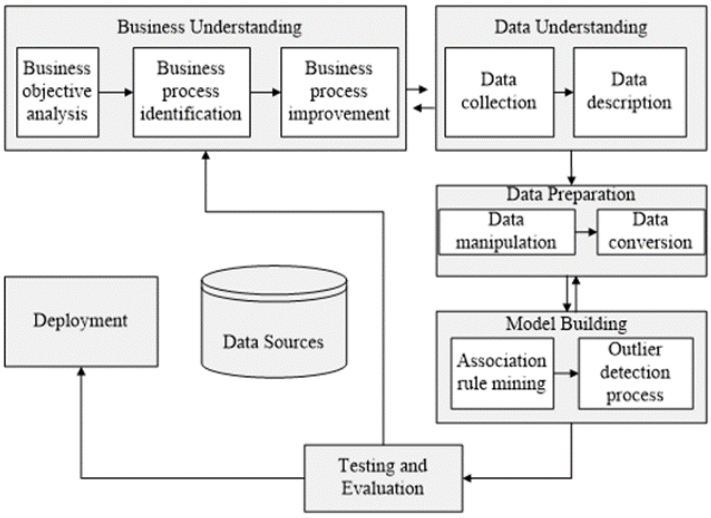

Fig. 2: The Proposed Research Model based on CRISP-DM Process

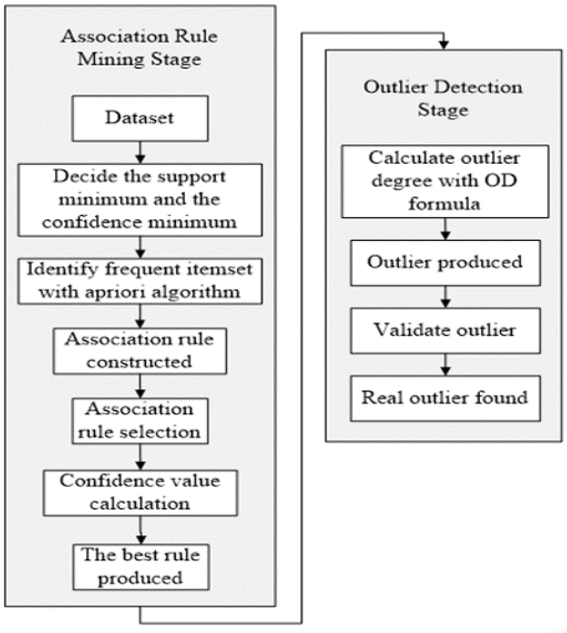

Fig. 3: Association Rule Mining and Outlier Detection Workflow

In the data modeling process of this research, a combination has been done between association rule mining technique and outlier degree formula. This step is adopted from [10], where they offer a method to detect outlier in transactional database. That step is as follows.

The calculation formula of outlier degree by is as follows [10].

$$od(t) = \frac{|t^{+} - t|}{|t^{+}|} \quad (3)$$

Where

od(*t*) : *outlier degree*
t : *associative closure*
t : *transaction*

IV. DISCUSSION

The research was done by raising academic cases in a particular university that drops out students when they either have reached the maximum studying time or have not reached the minimum standard grades. In fact, if seen more closely, many factors should be considered before dropping out students from college, such as knowing some of their inner

skills may have remained unnoticed.

Research data used 1,942 samples, and after carefully processed, 110 of which found as outliers with 80% accuracy. The table elaborates it.

From the analysis result of outlier finding, some facts are obtained as below and are offered recommendations for later decision-making. They are shown in this table.

TABLE I

Validation	Outlier Finding Calculation Result
Number of outlier found	259
Number of appropriate outlier finding based on reality	110
Number of real outlier	138
Accuracy percentage	80%

TABLE I
FACT, CONCLUSION, AND RECOMMENDATION

Fact	Conclusion	Recommendation
Most students are detected as outliers having GPA lower than 2.75 and IQ more than 120. However, only half has relatively good work ethics.	GPA does not always reflect one's level of intelligence. However, those with high IQ do not always mean they have high Emotional Quotient (EQ), which is reflected from work ethics.	• Consider many other factors or parameters before deciding to drop students out. • Direct student's organizations to become a platform in training their work ethics and emotional intelligence. • Double up seminars on student's Emotional Quotient (EQ) improvement. • Give insight about work ethics in either general lectures or seminars. • Do a routine EQ examination for students. • Evaluate and control the result of EQ examination routinely.
77 of 110 outlier data includes less active students in organization during their studies.	The more active a student is in an organization, the less chance of them to be an outlier.	• Make organizations top priority for students. • Do a routine in-depth control of organizational activities in campus. • Raise student's awareness on the importance of joining organization during college period.
The most data-containing work status in outliers is unemployed. However, the percentage of outlier students choosing entrepreneurship as their financial source is higher than those who do not	The "have not passed" status and low GPA increases the chance of unemployment. The potential of outlier students to become an entrepreneur.	• Do preventive action to lower the chance of students dropping out. • Give students lectures on entrepreneurship, albeit in general lectures or in mandatory courses. • Give students in every major insights about the chance of entrepreneurship. • Raise awareness for entrepreneurship in the future rather than for employment. • Encourage students to go entrepreneurship for career alternative supposing one day they (whether outliers or not) are not employed in any company.
The difference between students who work suitable with their major in normal data is balanced. However, data outliers are dominated by not appropriate category. About 67% and 79% of normal and outlier students are not satisfied with their current job.	Most of the graduates are not satisfied with their current job. This is in line with the suitability between work and college, where most of them work in a field completely out of their major.	• Give insight about some available college major as well as career opportunity and consequence to occur during both college and work. • Explore student's passion with a more representative material and method, therefore could lead to more accurate reference when choosing a major. • Advise students when choosing what their majors will be in college. • If the student has zero interest in coming to the classes, recommend students to resign at the first year. • Coordinate with parents or guardians in evaluating the student's passion. • Advise students regarding emotional intelligence to more grateful on the life path they've chosen.

Data spread of normal students within the parameter of position and work status is not equal. There lies a relatively high hap between work status (employed) and other work statuses like working students (69%). The same gap is also apparent within the position parameter, where normal students become staff than other positions (69%). Unlike outlier students, the both parameter data spread of students are relatively balanced.	Most normal graduates have chosen to work, and because of their "fresh graduate" status they are placed merely as staff. However, outlier students are considered more creative in pursuing their career. Such limited access, due to their incapability of getting their degree, encourages them to find new ways to pursue their career.	Train students's creativity in entrepreneurship by holding creative and innovative classes.
The average salary of these graduates exists in a mediocre category, both normal and outlier. However, for bonus, 73% of outlier graduates get mediocre bonus. This goes against 50% of normal students who don't get any bonus in their workplace.	There are enough outlier students who can get mediocre range bonus in their work. One of the achievement factors is IQ, knowing that outlier students getting mediocre bonus are those who have an IQ above 120.	Give insights about range of salary expectation and bonus in the company commonly filled with the campus alumni and what they do to go through.

V. CONCLUSION

From what has been done in the research, it is recognized that the utilization of outlier degree formula in association rule mining technique is able to detect outlier with 80% in accuracy. A properly data identification will significantly help selecting a proper technique for research data, therefore a high-quality information could be achieved. However, an in-depth analysis is necessary as well before determining the outlier definition. This is because there has been a difference of outlier definition, which will affect the result. The proposed model in this research can applied in other non-academic cases to identify the capability of the model in various fields.

ACKNOWLEDGMENT

We would like to thank STMIK Indonesia Padang for financial support. We also would like to thank all staff of the case study organisation.

REFERENCES

[1] The guardian.com, "The Guardian," 19 Desember 2012. [Online]. Available: https://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume. [Accessed Januari 2015].

[2] J. Gantz and D. Reinsel, "EMC Corporation," Desember 2012. [Online]. Available: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf. [Accessed Juli 2015].

[3] F. Provost and T. Fawcett, Data Science for Business, Sebastopol: O'Reilly Media, Inc, 2013.

[4] C. C. Aggarwal, Outlier Analysis, New York: Springer, 2013.

[5] S. Seo, A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets, Master's Program Thesis ed., Seoul: Kyung Hee University, 2002.

[6] D. Hand, H. Mannila and P. Smyth, Principles of Data Mining, Cambridge: MIT Press, 2001.

[7] D. L. Olson and D. Delen, Advanced Data Mining Techniques, Berlin: Springer, 2008

[8] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, USA: Elsevier, 2012.

[9] P. Jardosh and D. Ganatra, "A Comprehensive Survey: Association Rule Mining From XML," *International Journal of Innovative and Emerging Research in Engineering*, vol. 2, no. 4, pp. 103-106, 2015. J.

[10] K. Narita and H. Kitagawa, "Outlier Detection for Transaction Databases using Association Rules," *in The Ninth International Conference on Web-Age Information Management*, 2008.