An Implementation of ECODB Algorithm to Identify Outliers on the Data of National Exam Scores, Integrity Index, and Accreditation Level of Senior High Schools in Yogyakarta

1st Angela Mediatrix Melly[1], 2nd Paulina H. Prima Rosa[2]
{angelamediatrixmelly@gmail.com[1], rosa@usd.ac.id[2]}

Department of Informatics Sanata Dharma University, Indonesia[12]

Abstrak. Outlier detection is one field of research in data mining. One of the outlier detection algorithms is the Enhanced Class Outlier Distance Based (ECODB) algorithm. In this paper, the ECODB algorithm is used to identify outliers on national exam results, integrity index and accreditation level of senior high schools in the Yogyakarta province year 2015. An experiment by varying number of outlier and varying number of nearest neighbor were done to identify the value of N and K that result on smallest Class Outlier Factor (COF). The identified N and K were then used to find out schools that are classified as outliers. Further decriptive analysis of the schools were then performed, afterward.

Keywords: Data Mining, Outlier Identification, Enhanced Class Outlier Distance Based Algorithm, National Exam, Integrity Index, Accreditation Level

1 Introduction

Rapid data growth leads to huge data stacks. The rapid growth of data are often regarded as useless because they fill up the storage space and contain uninformative data. Therefore, data mining is needed to change the large and uninformative data into more information and knowledge. In the data mining there are many methods or techniques. Outlier detection is one field of research in data mining. Outliers are data that deviate too far from other data in a dataset. Outliers are often regarded as noise and most algorithms in data mining try to minimize and eliminate outliers [1]. However the outlier may be is representation or event of a unique or rare data that needs to be further analyzed [1].

There are many techniques or methods to detect outliers. Most of these methods identify the outliers regardless the class label of the data set. These methods only identify the outlier with respect to the whole of data set. Class Outlier Mining identifies outliers which behave differently from other data with the same class label. Hewahi & Saad in [2] describe the general definition of class outlier mining problem as “given a set of observations with class labels, find those that arouse suspicions, taking into account the class labels”.

ICSTI 2018, October 19-20, Yogyakarta, Indonesia

DOI 10.4108/eai.19-10-2018.2282539

One of the algorithms to perform class outlier mining is Enhanced Class Outlier Distance Based (ECODB) algorithm [2]. Ajani et.al. [3] has analyzed and implemented ECODB to study the performance of ECODB.

In this paper, ECODB algorithm is used to identify outliers on national exam results, integrity index and accreditation level of senior high schools at Daerah Istimewa Yogyakarta province of year 2015.

This research is expected to provide information of rare occurrence of National Exam data. From this study the government can get information about schools with National Exam data that are rare or unique compared to other schools. The results of this study can be further analyzed by the school as well as the government for school development.

2 Data mining and outlier detection

2.1 Data Mining

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to explore large databases in order to find novel and useful patterns that might otherwise remain unknown. Data mining also provide capabilities to predict the outcome of a future observation. Data mining is an integral part of knowlegde discovery in database, which is the overall process of converting raw data into useful information. Not all information discovery tasks are considered to be data mining. For example, looking up individual record using a database management system or finding particular Web pages via a query to an Internet search engine are tasks related to the area of information retrieval. Nonetheless, data mining technique have been use to enhance information retrieval systems [4].

2.2 Outlier Detection

An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism. Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies. Many data mining algorithms try to minimize the influence of outliers or eliminate them all together. However, it could result in the loss of important hidden information because one person’s noise could be another person’s signal [1]. There are many outlier detection methods. There are two orthogonal ways to categorize outlier detection methods. First, the categorization of outlier detection methods according to whether the sample of data for analysis is given with domain expert–provided labels that can be used to build an outlier detection model. Second, the categorization of methods into groups according to their assumptions regarding normal objects versus outliers.

In addition, most of these methods identify the outliers regardless the class label of the data set. Class Outlier Mining identifies outliers which behave differently from other data with the same class label. One of the algorithms to perform class outlier mining is Enhanced Class Outlier Distance Based (ECODB) Algorithm which is the development of regular outliers detection, Class Outlier Distance Based (CODB) [2].

2.3 Enhanced Class Outlier Distance Based Algorithmn

ECODB deals with class label. A class label is an attribute chosen among a set of attributes in a given database based on the request of the user and type of the application. A class label could be medical diagnoses, credit or loan approval decision, customer classification… etc.) [2]. The conventional Methods (*Outlier Mining*) is to find exception or rare cases in a dataset irrespective of the class label of these cases, they are considered rare events with respect to the whole dataset. *Class Outlier Mining* is to find suspicious instances taking into account the class label. *Outlier Mining* cannot detect cases which behave differently from its class, whereas the *Class Outlier Mining* can do [5].

Hewahi and Saad proposed a novel definition for class outlier and new method for mining class outlier based on distance-based approach and nearest neighbors. This method is called *CODB* algorithm and is based on the concept of Class Outlier Factor (*COF*) which represents the degree of being a class outlier for a data object. ECODB algorithm is enhancement from Class Outlier Distance Based (CODB) algorithm. The ECODB algorithmn steps are as follows [5] :

2.3.1 For a given dataset, compute *PCL(T, K)* for all instances. *PCL(T, K)* is the probability of the class label of the instance *T* with respect to the class labels of its *K* Nearest Neighbors. The *PCL(T, K)* formula is defined as follows :

$$PCL(T,K) = \frac{N_T}{K} \qquad \text{... (1)}$$

Where:

- N = the number of instances that have the same class label as the instance T
- K = the number of nearest neighbors of instance T

For example, suppose we are working with 7 nearest neighbors of an inatance *T* (including itself) on a dataset with two class labels x and y, where 5 of these neighbors have the class label x, and 2 have the class label y. The instance *T* has the class label y, which means the *PCL* of instance *T* is 2/7.

2.3.2 Maintain a list of top *N* instance with least *PCL(T, K)* value.

2.3.3 For each instance in the top *N* list, compute *Deviation(T)* and *KDist(T)*. *Deviation(T)* is how much the instance *T* deviates from the instances that have the similar class label of that istance *T*. The *Deviation(T)* is calculated by summing the distance between the T instances and each instance that has the same class as the T instance. *Deviation(T)* formula is defined as follows :

$$Deviation(T) = \sum_{i=0}^{n} d(T, t_i) \qquad \text{... (2)}$$

Where :

- n = the number of instances that have the same class label as the instance *T*
- $d(T, t_i)$ = the distance between instances that have the same class as instance *T*
- *KDist(T)* is the distance between the instance *T* and its *K* nearest neighbors. *KDist(T)* formula is defined as follows

$$KDist(T) = \sum_{i=0}^{K} d(T, t_i) \quad \text{... (3)}$$

Where :

- *K* = the number of nearest neighbors
- *d(T, t$_i$)* = the distance between the nearest neighbor to instance T

Then do normalization on *Deviation* and *KDist* for Deviation and KDist in range 0-1. Normalization *Deviation* and *KDist* formula is defined as follows :

$$norm(Deviation(T) = \frac{(Deviation(T) - MinDev)}{(MaxDev - MinDev)} \quad (4)$$

$$norm(KDist(T)) = \frac{(KDist(T) - MinKDist)}{(MaxKDist - MinKDist)} \quad \text{...(5)}$$

Where :

- *norm(Deviation(T))* = normalized Deviation (T) values
- *norm(KDist(T))* = normalized *KDist(T)* value
- *MaxDev* = the highest deviation value for top *N* class outliers
- *MinDev* = the lowest deviation value for top *N* class outliers
- *MaxKDist* = the highest KDist value for top *N* class outliers
- *MinKDist* = the lowest KDist value for top *N* class outliers

2.3.4 Compute *COF* value for all instances in the top *N* list. *COF(T)* formula id defined as follows :

$$COF(T) = K \times PCL(T, K) - norm(Deviation(T)) + norm(KDist(T)) \quad (6)$$

Where :

- *COF(T)* = Class Outlier Factor of instance T

2.3.5 Sort the top *N* list in ascending order according to their *COF* value.

3 Methodology

This research utilized the data of national exam scores, integrity index and accreditation level of senior high schools in the Daerah Istimewa Yogyakarta province of year 2015. The data in the form of spreadsheet were obtained from three different sources. The data of national exam scores was taken from the official website of the Badan Penelitian Pendidikan dan Pengembangan Kementrian Pendidikan dan Kebudayaan (http://118.98.234.50/lhun/daftar.aspx). The data of school accreditation level was taken from the official website of the Badan Penelitian Pendidikan dan Pengembangan Kementrian Pendidikan dan Kebudayaan (http://bansm.or.id/sekolah/sudah_akreditasi/4). The integrity index value data was taken from the Badan Penelitian Pendidikan dan Pengembangan Kementrian Pendidikan dan Kebudayaan http://puspendik.kemdikbud.go.id/hasil-un/.

The data was divided into two datasets. Dataset I is the data containing schools having natural science major. Dataset II is the data containing schools having social science major. The data undergoes a process as described in the following sections.

3.1 Preprocessing

Several processes were performed in this stage, namely data cleaning, data integration and data selection. Data cleaning is a process to eliminate noise and inconsistent data. Since not all schools have integrity index, so the data of those schools are eliminated from dataset. Data integration is the process of combining multiple data sources. Since the data used were obtained from three different sources therefor those data were integrated into one file. Data selection is the process of selecting data or attributes that are relevant for this research. In this process 8 relevant attributes were selected from each dataset. Table I and Table II show the list of selected attributes and short description about the attributes.

Table 1. Selected Attributes of Dataset I

Attribute Name	Information	Attribute Type
Indonesia	Indonesia Language National Exam Score	Numeric
English	English National Exam Score	Numeric
Mathematics	Mathematics n National Exam Score ational exam score	Numeric
Physics	Physics National Exam Score	Numeric
Chemistry	Chemistry National Exam Score	Numeric
Biology	Biology National Exam Score	Numeric
Integrity Index	Integrity Index of National Exam	Numeric
Rank	Level of Accreditation	Character (as class label)

Table 2. Selected Attributes of Dataset II

Attribute Name	Information	Attribute Type
Indonesia	Indonesia Language National Exam Score	Numeric

English	English National Exam Score	Numeric
Mathematics	Mathematics National Exam Score	Numeric
Economics	Economics National Exam Score	Numeric
Sociology	Sociology National Exam Score	Numeric
Geography	Geography National Exam Score	Numeric
Integrity Index	Integrity Index of National Exam	Numeric
Rank	Level of Accreditation	Character (as class label)

3.2 Data Mining

At this stage, outlier identification was performed using ECODB algorithm. A software was developed using Java programming language to perform the mining.

The input of the software is a file with .xls extension that can be directly selected by the user. The user entry the value of N and the value of K to be used in the outlier identification process. N is the number of expected outliers, while K is the number of nearest neighbors. The output is a list of schools identified as outliers. Black box testing was also performed towards the software to validate the result of the software.

Furthermore, mining towards the two datasets were undertook using the software. An experiment by varying number of outlier (N=5, N=10, N=15) and varying number of nearest neighbor (K=10, K=15, K=20) were done to identify the value of N and K that result on smallest COF. The identified N and K were then used to find out schools that are classified as outliers. Further decriptive analysis of the schools were then performed, afterward.

4 Results and discussion

4.1 Experiment using Dataset I

Table III describes the result of outlier identification using Dataset I with *N* and *K* varies.

Table 3. Experiment to Find N and K of Dataset I

N	K	Average of COF	Minimum of COF
5	10	0.921947183	0,511034296
	15	0,916100119	0,511034296
	20	0,923700086	0,627533939
10	**10**	**0.899195646**	**0.390370826**
	15	1.31391831	0.710666006

	20	1.385858649	0.833082635
15	10	1.30013443	0.641078874
	15	1.638104533	0.710666006
	20	1.924481519	0.833082635

Table III shows that the smallest COF value and smallest average value of COF was obtained when K = 10 and N = 10.

Table IV elaborates the schools that are identified as outliers when K=10 and N=10. Six out of ten outlier schools are schools having accreditation level "Not Accredited", while 4 out of 10 schools have accreditation level "B". By examining the nearest neighbors of the outlier schools, it can be found that most of neighbors having accreditation level "A". Therefore it is clear that those schools were identified as class outlier.

Table 4. The Outliers of Dataset I

S_id	Indo Lang	Eng	Math	Phy	Chem	Bio	II	Accred Level
S 1009	79,5	54,41	36,71	32,52	42,63	44,2	78,4	Not Accr
S 1058	83,98	66,39	48,12	58,12	59,96	63,39	82,31	Not Accr
S 5004	80,68	51,88	44,95	42,74	42,62	49,03	84,31	Not Accr
S 4054	79,61	62,24	40,42	57,81	41,33	58,94	74,78	Not Accr
S 3013	77,31	53,58	36,47	54,05	53,51	67,42	72,48	Not Accr
S 4048	80,15	48,13	38,54	38,45	53,5	53,13	84,4	Not Accr
S 1030	57,32	63,86	73,81	85,27	22,69	71,37	53,28	B
S 2034	76,55	46,14	30,06	36,16	41,5	43,95	78,62	B
S 4044	72,58	45,54	26,55	59,6	35,31	67,28	71,8	B
S 4057	77,2	48,08	24,21	26,97	29,61	37,69	74	B

4.2 Experiment using Dataset II

The result of outlier identification towards dataset II by varying *N* and *K* can be seen in Table V.

Table 5. Experiment to Find N and K of Dataset II

N	K	Average of COF	Minimum of COF
5	**10**	**0,8717356628**	**0,237431421**
	15	1,2789695082	0,589248799
	20	1,4868797292	0,46560402
10	10	1.080865742	0.491371017
	15	1.661367929	0.704499523
	20	2.028675733	0.465041278
15	10	1.471911781	0.41303264
	15	2.246713192	0.704499523
	20	2.734869769	0.737360023

.

From the result of outlier identification in Table V it can be seen that the smallest COF value and average value of smallest COF was obtained when K = 10 and N = 5. Table VI elaborates the schools that were identified as outliers with K=10 and N=5. As can be seen from Table VI, three outlier schools are schools that are classified as "Not Accredited". These three schools have quite good scores in almost every item, which can be associated as schools with accreditation level "B". Otherwise, two outlier schools have accreditation level "B", but their scores are quite low compare to other schools with the same accreditation level.

Table 6. The Outliers of Dataset II

S_id	Indo Lang	Eng	Math	Eco	Soc	Geo	II	Accred Level
S 4054	80,73	63,36	47,42	50,76	56,82	59,95	72,8	Not Accr
S 1058	80,84	61,08	56,86	46,93	65,14	62,56	83,37	Not Accr
S 4044	62,94	40,72	30,89	35,95	44,09	41,83	71,8	B
S 1027	66,79	35,33	21,66	34,32	42,14	34,84	80,6	B
S 3013	72,66	54,7	45,64	48,07	55,45	46,78	71,14	Not Accr

5 Conclusion

The experiment has proved that ECODB algorithm can be applied to identify outliers on data of National Exam scores, integrity index and accreditation level of senior high schools in Daerah Istimewa Yogyakarta province. By varying K and N, experiment towards the dataset containing schools having natural science major result on the smallest average COF value when K = 10 and N = 10. The same experiment towards the dataset containing schools having social science major result on recommended K = 10 and N = 15.

Descriptive analysis towards schools that are identified as outliers has been performed to understand the reason why those schools are identified as outliers.

Further research to measure the accuracy and running time of the algorithm, especially in large databases to be considered.

References

[1] J. Han and M. Kamber, Data Mining : Concepts and Techniques 3rd Edition. San Fransisco: Morgan Kaufmann Publishers, 2012.

[2] N. M. Hewahi and M. K. Saad, “Class Outlier Mining : Distance-Based Approach,” Int. J. Electr. Comput. Eng., 2007.

[3] M. Ajani, K. R. S. Wiharja, and I. Ataina, “Analisis dan Implementasi ECODB dalam Mendeteksi Class Outlier,” Telkom University, 2011.

[4] P. N.Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Boston: Pearson Addison Weisley, 2006.

[5] N. M. Hewahi and M. K. Saad, “A comparative Study of Outlier Mining and Class Outlier Mining,” Comput. Sci. Lett., vol. 1 (1), 2009.