International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 18 (2018) pp. 13961-13968

Performance Measurement of Various Threshold Values for Discrimination Removal and Data Quality Percentage by Different Discrimination Measures

Mr. Manoj Ashok Wakchaure[1]*, Prof. Dr. S.S.Sane[2]

[1] Research Scholar, Department of Computer Engineering, KKWIEER, Nashik, SPPU, Pune, India.

[2]Professor& Head, Department of Computer Engineering, KKWIEER, Nashik, SPPU, Pune, India.

**Corresponding author*

Abstract

Now a day Discrimination Prevention area is important sub area in data mining. Researchers are working in this area for getting some improved results. Several techniques are already developed by researchers. Rules are consider if various measures and Threshold values are to evaluated .*elift, slift, clift, olift* and *elb* are various discrimination measures are ruminate for evaluation of alpha rules ,Red lining rules ,Protect rules and Non redlining Rules. Both Parameters are calculated by the various Results from alpha.

This Paper Deals with the Experiments of Certain Measures by values of alpha between 1 to 1.5. The Performance shows that while increasing the value of alpha, count of rules will Decreases.

Keywords: Discrimination Prevention, Discrimination Measures, Threshold

INTRODUCTION

Discrimination Prevention in Data mining is playing vital role in the Data science. Discrimination is defined as "the movement of unlawfully treating people on the basis of their belonging to a definite group, namely race, ideology, etc"[1],[2],[10],[11]. This comprises refusing chance to associates of one class that are accessible to associates of other class. It is essential to make the use of some anti-discriminatory actions, which are the rules and regulations proposed to stop discrimination. Numerous decision-making areas offer themselves to discrimination, such as Educational Institutions, Financial Institutions, Health Insurance and Human Resource Management, Home on Rent etc.. The usage of these data systems in information technology mining for automated decision-making has enticed the consideration of several individuals.[4] The aim of this field is to develop the discrimination free decisions so as to attain impartiality in most of the decision making areas.

This paper deals with study of use of discrimination measures such as elift, slift, clift and olift. This paper is based on experiments by various values of alpha. Rest of the paper is organized as follows. Section 1 provides introduction, Literature survey is presented in section 2. Section 3 deals with performance evaluation work carried out it also deals with experimental results and discussions. Conclusions are provided in section 4.

LITERATURE REVIEW

In the field of information mining which aims to identify hypothetically valuable and eventually reasonable patterns, many authors have given their contribution. Data processing concept strives to simplify and enhance the information quality, consequently establishing it as more dependable [3],[4]. "Data preprocessing" does this by eliminating the unnecessary data and drawing out the important features of the data. This makes pattern discovery simple without ignoring any crucial information. A survey on discrimination prevention methods in data mining exhibits new discrimination prevention method. Various transformations are used for the discovery of discrimination.[1] The procedure measures the discriminative bias and identifies the groupings by decision-making processes. Data models that are discrimination-free can be created from the changed dataset without genuinely harming the quality of Data. More data can be controlled and the system result is accurate. Another study views the recent best in class methodologies for "antidiscrimination" systems and likewise concentrates on "discrimination" discovery and deterrence in" data mining"[5],[6]. Furthermore, convey a theoretic suggestion for improving the outcomes of the information quality. And also discusses about the most ideal approach to clean data collection and outsourced informational indexes such that immediate discrimination choice guidelines are changed over to authentic characterization rules[8]. There was a research to build a different pre-processing discriminatory bias deterrence technique together with various "data transformation" techniques which can prevent "direct" discriminatory bias, "indirect" discriminatory bias or together at the similar time[7],[9]. To accomplish this aim, the primary step is to calculate the amount bias then recognizes classes and collections of individuals which may be straightforwardly and indirect victimized in decision making process[10]. The subsequent step is data transformation now is a appropriate approach to expel each one of those discrimination predispositions at long last, data models that are discrimination free can be delivered from the changed data without genuinely harming information quality.

EXPERIMENTAL RESULTS

We take support 2 and confidence 10 Percent for Adult Data set and support 5 and confidence 10 percent for German Credit dataset respectively. Various measures are used to find MR, PR, RR& NR rules to calculate DR and DQ. Our experiments on Direct and Indirect Rule Protection Algorithm

Table 1: Adult Data Set: Utility Measures for Minimum Support 2 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Elift

α	No. Red lining Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
				Direct		Indirect		MC	GC
				DDPD	DDPP	IDPD	IDPP		
1	43	71	804	89.45	100	96.20	100	0	0.01
1.1	21	30	280	100	100	100	100	0	0
1.2	9	14	140	100	100	100	100	0	0
1.3	0	0	67	100	100	NA	100	0	0.01
1.4	0	0	32	100	100	NA	100	0	0
1.5	0	0	07	NA	100	NA	100	0	0
No. of Frequent Classification Rules : 5,092				No. of Background Know. Rules : 2,089					

Comments: if elift measure is used to calculate values then for α=1 ,RR rules are 43 ,MR rules are 804 and DDPD &IDPD are 89.45 & 96.20 Respectively .for 1&1.2 GC is 0.01 all other values Data quality is 100% .if α=1.1,RR=21,MR=280 . Discrimination Removal is 100% from α=1.1 onwards .FR rules & BK rules are 5092&2089 respectively. For support 5 , α=1,MR=135 .

Table 2: Adult Data Set: Utility Measures for Minimum Support 2 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Slift

α	No. Red lining Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
				Direct		Indirect		MC	GC
				DDPD	DDPP	IDPD	IDPP		
1	43	71	807	91.60	97.45	100	100	0	0
1.1	21	30	282	100	100	100	100	0	0
1.2	9	14	142	100	100	100	100	0	0
1.3	0	0	69	100	100	NA	100	0	0.02
1.4	0	0	31	100	100	NA	100	0	0
1.5	0	0	07	100	100	NA	100	0	0
No. of Frequent Classification Rules : 5,092				No. of Background Know. Rules : 2,089					

Comments: if slift measure is used to calculate values then for α=1 ,RR rules are 43 ,MR rules are 807 and DDPD &IDPP are 91.60 & 97.45 Respectively .for α =1.3, GC is 0.02 all other values Data quality is 100% .if α=1.1,RR=21,MR=282 . Discrimination Removal is 100% from α=1.1 onwards .FR rules & BK rules are 5092&2089 respectively. For support 5 , α=1,MR=137 .

Table 3: Adult Data Set: Utility Measures for Minimum Support 2 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Clift

α	No. Red lining Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
				Direct		Indirect		MC	GC
				DDPD	DDPP	IDPD	IDPP		
1	43	71	807	90.25	95.55	100	100	0	0
1.1	21	30	283	100	100	100	100	0	0
1.2	9	14	142	100	100	100	100	0	0
1.3	0	0	72	100	100	NA	100	0	0.01
1.4	0	0	34	100	100	NA	100	0	0.01
1.5	0	0	07	100	100	NA	100	0	0
No. of Frequent Classification Rules : 5,092				No. of Background Know. Rules : 2,089					

Comments: if Clift measure is used to calculate values then for α=1, RR rules are 43, MR rules are 807 and DDPD &DDPP are 90.25 & 95.55 Respectively. for 1.3&1.4 GC is 0.01 all other values Data quality is 100% .if α=1.1, RR=21, MR=283. Discrimination Removal is 100% from α=1.1 onwards .FR rules & BK rules are 5092&2089 respectively. For support 5, α=1, MR=136.

Table 4: Adult Data Set: Utility Measures for Minimum Support 2 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Olift

α	No. Red lining Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
				Direct		Indirect		MC	GC
				DDPD	DDPP	IDPD	IDPP		
1	43	71	802	91.30	97.90	100	100	0	0
1.1	21	30	282	100	100	100	100	0	0
1.2	7	14	142	100	100	100	100	0	0
1.3	0	0	68	100	100	NA	100	0	0.03
1.4	0	0	36	100	100	NA	100	0	0
1.5	0	0	08	100	100	NA	100	0	0
No. of Frequent Classification Rules : 5,092				No. of Background Know. Rules : 2,089					

Comments: if olift measure is used to calculate values then for α=1 ,RR rules are 43 ,MR rules are 802 and DDPD &DDPP are 91.30 & 97.90 Respectively .for 1.3 GC is 0.03 all other values Data quality is 100% .if α=1.1,RR=21,MR=282 . Discrimination Removal is 100% from α=1.1 onwards .FR rules & BK rules are 5092&2089 respectively. For support 5 , α=1,MR=133 .

Table 5: German Credit Data Set: Utility Measures for Minimum Support 5 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Elift

α	No. Red lining Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
				Direct		Indirect		MC	GC
				DDPD	DDPP	IDPD	IDPP		
1	37	42	3715	99.97	99.98	100	100	0	2.09
1.1	0	0	1860	100	100	NA	100	0	2.00
1.2	0	0	991	100	100	NA	100	0	0.02
1.3	0	0	415	100	100	NA	100	0	0.01
1.4	0	0	207	100	100	NA	100	0	0.44
1.5	0	0	120	100	100	NA	100	0	0.06
No. of Frequent Classification Rules : 32,340				No. of Background Know. Rules : 22,763					

Comments: if elift measure is used to calculate values then for α=1 ,RR rules are 37 ,MR rules are 3715 and DDPD &DDPP are 99.97 & 99.98 Respectively .for 1,GC is 2.09,for1.1,GC is 2,for 1.2, GC is 0.02,for 1.3 GC is 0.01,for 1.4 GC is 0.44,for 1.5,GC is 0.06 .for MC, Data quality is 100% .if α=1.,RR=37 . Discrimination Removal is 100% from α=1.1 onwards .FR rules & BK rules are 32340&22763 respectively. For support 2, α=1,MR=9033 .

Table 6: German Credit Data Set: Utility Measures for Minimum Support 5 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Slift

α	No. Red lining Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
				Direct		Indirect		MC	GC
				DDPD	DDPP	IDPD	IDPP		
1	37	42	3812	99.98	100	100	100	0	2.00
1.1	0	0	1920	100	100	NA	100	0	2.00
1.2	0	0	1003	100	100	NA	100	0	0.01
1.3	0	0	470	100	100	NA	100	0	0.01
1.4	0	0	224	100	100	NA	100	0	0.54
No. of Frequent Classification Rules **:** 32,340				No. of Background Know. Rules **:** 22,763					

Comments: if slift measure is used to calculate values then for α=1 ,RR rules are 37 ,MR rules are 3812 and DDPD is 99.98 .for 1,GC is 2,for1.1,GC is 2,for 1.2, GC is 2,for 1.3 GC is 0.01,for 1.4 GC is 0.54,f..for MC, Data quality is 100% .if α=1.,RR=37 . Discrimination Removal is 100% from α=1.1 onwards .FR rules & BK rules are 32340&22763 respectively. For support 2 , α=1,MR=9102 .

Table 7: German Credit Data Set: Utility Measures for Minimum Support 5 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Olift

α	No. Red lining Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
				Direct		Indirect		MC	GC
				DDPD	DDPP	IDPD	IDPP		
1	37	42	3770	99.98	100	100	100	0	2.01
1.1	0	0	1893	100	100	NA	100	0	2.02
1.2	0	0	1017	100	100	NA	100	0	0.01
1.3	0	0	473	100	100	NA	100	0	0.01
1.4	0	0	233	100	100	NA	100	0	0.60
No. of Frequent Classification Rules : 32,340				No. of Background Know. Rules : 22,763					

Comments: if olift measure is used to calculate values then for α=1 ,RR rules are 37 ,MR rules are 3770 and DDPD is 99.98 .for **α** =1,GC is 2.01,for1.1,GC is 2.02,for 1.2, GC is 0.01,for 1.3 GC is 0.01,for 1.4 GC is 0.60 .for MC, Data quality is 100% .if α=1.1.,RR=Nil . Discrimination Removal is 100% from α=1.1 onwards .FR rules & BK rules are 32340&22763 respectively. For support 2 , α=1,MR=9117 .

Table 8: German Credit Data Set: Utility Measures for Minimum Support 5 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Clift

α	No. Red lining Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
				Direct		Indirect		MC	GC
				DDPD	DDPP	IDPD	IDPP		
1	37	42	3795	100	99.80	NA	100	0	2.01
1.1	0	0	1872	100	100	NA	100	0	2.02
1.2	0	0	1022	100	100	NA	100	0	0.01
1.3	0	0	461	100	100	NA	100	0	0.01
1.4	0	0	227	100	100	NA	100	0	0.29
No. of Frequent Classification Rules : 32,340				No. of Background Know. Rules : 22,763					

Comments: if clift measure is used to calculate values then for α=1 ,RR rules are 37 ,MR rules are 3795 and DDPP is 99.80 .for 1,GC is 2.01,for1.1,GC is 2.02,for 1.2, GC is 0.01,for 1.3 GC is 0.01,for 1.4 GC is 0.29,f..for MC, Data quality is 100% .if α=1.1.,RR=Nil. Discrimination Removal is 100% from α=1.1 onwards .FR rules & BK rules are 32340&22763 respectively. For support 2 , α=1,MR=9087 .

Table 9: Adult Data Set: Utility Measures for Minimum Support 2 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Elift, slift, Clift, Olift Measures

Measures	Alfa	No. RR Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
					Direct		Indirect		MC	GC
					DDPD	DDPP	IDPD	IDPP		
Results in Literature	1	43	71	804	89.45	100	95.35	100	0	0.03
	1.1	21	30	280	100	100	100	100	0	0
	1.2	9	14	140	100	100	100	100	0	0
	1.3	0	0	67	100	100	NA	100	0	0.01
	1.4	0	0	32	100	100	NA	100	0	0
	1.5	0	0	7	100	100	NA	100	0	0
Elift	1	43	71	804	89.45	100	96.20	100	0	0.01
	1.1	21	30	280	100	100	100	100	0	0
	1.2	9	14	140	100	100	100	100	0	0
	1.3	0	0	67	100	100	NA	100	0	0.01
	1.4	0	0	32	100	100	NA	100	0	0
	1.5	0	0	07	NA	100	NA	100	0	0
slift	1	43	71	807	91.60	97.45	100	100	0	0
	1.1	21	30	282	100	100	100	100	0.11	0
	1.2	9	14	142	100	100	100	100	0	0
	1.3	0	0	69	100	100	NA	100	0	0.02
	1.4	0	0	31	100	100	NA	100	0	0
	1.5	0	0	07	100	100	NA	100	0	0
clift	1	43	71	807	90.25	95.55	100	100	0	0
	1.1	21	30	283	99.78	100	100	100	0.16	0
	1.2	9	14	142	100	100	100	100	0	0
	1.3	0	0	72	100	100	NA	100	0	0.01
	1.4	0	0	34	100	100	NA	100	0	0.01
	1.5	0	0	07	100	100	NA	100	0	0
Olift	1	43	71	802	91.30	97.90	100	100	0	0
	1.1	21	30	282	100	100	97.50	100	**0**	**0**
	1.2	9	14	142	100	100	100	100	0	0
	1.3	0	0	68	100	100	NA	100	0	0.03
	1.4	0	0	36	100	100	NA	100	0	0
	1.5	0	0	08	100	100	NA	100	0	0
No. of Frequent classification Rules : 5092						**No. of Background Know. Rules : 2,089**				

Observations : for Adult Dataset ,RR rules are 43 and decreases when we increase value of α.for α is greater than 1, discrimination removal is 100% and comparatively elift and slift gives good result but No single measure is performing best result for all Parameters and for all α .

Table 10: German Credit Data Set: Utility Measures for Minimum Support 5 Percent and Confidence 10 Percent for Direct and Indirect Rule Protection; Columns Show the Results for Different Values of α using Elift, slift, Clift, Olift Measures

Measures	Alfa	No. RR Rules	No. Indirect Alfa Discrimination Rules	No. direct Alfa Discrimination Rules	Discrimination Removal				Data Quality	
					Direct		Indirect		MC	GC
					DDPD	DDPP	IDPD	IDPP		
Results in Literature	1	37	42	499	99.97	100	100	100	0	2.07
	1.1	0	0	312	100	100	NA	100	0	2.07
	1.2	0	0	26	100	100	NA	100	0	1.01
	1.3	0	0	14	100	100	NA	100	0	1.01
	1.4	0	0	9	100	100	NA	100	0	0.69
Elift	1	37	42	3715	99.97	100	100	100	0	2.26
	1.1	0	0	1860	100	100	NA	100	0	2.00
	1.2	0	0	991	100	100	NA	100	0	0.02
	1.3	0	0	415	100	100	NA	100	0	0.01
	1.4	0	0	207	100	100	NA	100	0	0.44
slift	1	37	42	3812	99.97	100	100	100	0	1.86
	1.1	0	0	1920	100	100	NA	100	0	2.00
	1.2	0	0	1003	100	100	NA	100	0	0.01
	1.3	0	0	470	100	100	NA	100	0	0.01
	1.4	0	0	224	100	100	NA	100	0	0.54
clift	1	37	42	3795	100	99.75	100	100	0	2.00
	1.1	0	0	1872	100	100	NA	100	0	2.02
	1.2	0	0	1022	100	100	NA	100	0	0.01
	1.3	0	0	461	100	100	NA	100	0	0.01
	1.4	0	0	227	100	100	NA	100	0	0.29
Olift	1	37	42	3770	100	100	99.80	100	0	2.23
	1.1	0	0	1893	100	100	NA	100	0	2.02
	1.2	0	0	1017	100	100	NA	100	0	0.01
	1.3	0	0	473	100	100	NA	100	0	0.01
	1.4	0	0	233	100	100	NA	100	0	0.60
No. of Frequent classification Rules : 32,340						**No. of Background Know. Rules : 22,763**				

Observations : for German Credit Dataset ,RR rules are 37 and it decreases when we increase value of α. for α is greater than 1, discrimination removal is 100% and comparatively MC and IDPD provides 100% results but No single measure is performing best result for all Parameters and for all α .

CONCLUSION

The conclusions based on experiments carried out are Discrimination Removal parameters DDPD, DDPP, IDPD &IDPP are showing the results more than 99 percentage on an average so it means our accuracy of discrimination removal is greater than 99Percent . As far as Data quality is concern two parameters are considered MC&GC . Misses cost gives better results i.e 100 percent maintaining the data quality except one or two conditions. Also GC provide good results i.e 0.02 when adult dataset is to be considered. If we take German Dataset, GC values will be 1.008 on an average. it means all parameters are providing very good results for Discrimination Prevention. By using this methodology we gives various results by considering various values of alpha and also considered various measures for comparison and analysis .In future will we try to check results of Discriminatory threshold by considering IM_LR and IM_Hybrid i.e fitness method and Aco respectively .

ACKNOWLEDGEMENT

We would like to thanks to all library media who supported our work in this way and help us to get results of better quality. We would also like to show our gratitude to Research Centre K.K.Wagh Institute of Engineering Education & Research, Nashik.We also Thankful to Savitribai Phule Pune university, Pune.

REFERENCES

[1] Sara Hajian and Josep Domingo-Ferrer A Methodology for Direct and Indirect Discrimination Prevention in Data Mining, Data Mining and Knowledge Discovery, vol. 25, no. 7, pp. 1445-1459, 2013

[2] F. Kamiran and T. Calders, Classification with no Discrimination, by Preferential Sampling, Proc. 19th Machine Learning Conf. Belgium and The Netherlands,2010.

[3] F. Kamiran and T. Calders, Classification without Discrimination, Proc. IEEE Second Intl Conf. Computer, Control and Comm.(IC4 09), 2009.

[4] D. Pedreschi, S. Ruggieri, and F. Turini, "DiscriminationAware Data Mining," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 560-568, 2008.

[5] Manoj Ashok Wakchaure,Prof.Dr.Shirish Sane, "Discrimination Prevention by Different Measures in Direct Rule Protection Algorithm "International Journal of Emerging Technology and Advanced Engineering ,Volume 5,Issue 9,2015

[6] T. Calders and S. Verwer, Three Naive Bayes Approaches for Discrimination-Free Classification, Data Mining and Knowledge Discovery, vol. 21, no. 2, pp. 277-292, 2010.

[7] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, Proc. 20th Intl Conf. Very Large Data Bases, pp. 487-499, 1994.

[8] D. Pedreschi, S. Ruggieri, and F. Turini, Measuring Discrimination in Socially-Sensitive Decision Records, Proc. Ninth SIAM Data Mining Conf. (SDM 09), pp. 581-592, 2009.

[9] V. Verykios and A. Gkoulalas-Divanis, A Survey of Association Rule Hiding Methods for Privacy, Privacy Preserving Data Mining, Models and Algorithms, C.C.Aggarwal and P.S. Yu, Springer, 2008.

[10] Manoj Ashok Wakchaure , and Prof.Dr.Shirish S. Sane. A Novel Approach for Discrimination Prevention and Privacy Preservation in Data Mining. *International Journal Of Advance Research And Innovative Ideas In Education* [Internet]. 2016 .

[11] P.Yuvasri,S.Boopathy,"A Method For Preventing Discrimination in Data Mining, "International journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 3,Issue 4,2014.