al_diagnosis_prediction_using_rough-regression_approximation.pdf

by

Medipre: Medical Diagnosis Prediction Using Rough-Regression Approximation

Riswan Efendi
Universiti Tun Hussein Onn Malaysia
Mathematics Department, UIN Sultan Syarif Kasim Riau, Indonesia
+62761589026, 28293
riswan@uthm.edu.my

Noor Azah Samsudin
Universiti Tun Hussein Onn Malaysia
+6074533603, 86400
azah@uthm.edu.my

Mustafa Mat Deris
Universiti Tun Hussein Onn Malaysia
+6074533603, 86400
mmustafa@uthm.edu.my

ABSTRACT

Rough set and regression approximations are useful in establishing decision support system for medical diagnostic applications. However, the data elimination strategy for unclassified elements or patients in the medical diagnostic applications remains as a serious issue to be explored, especially with the aim of achieving higher prediction accuracy. This paper presents step-by-step procedure in building rough-regression approximation based on data elimination strategy. A number of data sets is used to examine our proposed approximation. The result has shown that the proposed rough-regression is capable to improve the prediction accuracy if compared with the existing approximations significantly. The proposed approximation can improve the performance of medical diagnosis prediction system. Therefore, it may help inexperienced doctors and patients for preliminary diagnosis.

CCS Concepts

• **Mathematics of computing~Regression analysis** • **Applied computing~Health care information systems** • **Applied computing~Forecasting** • ***Information systems~Data mining***

Keywords

"Rough set", "regression model", "data reduction", "unclassified element", "rough-regression approximation", "medical diagnosis prediction"

1. INTRODUCTION

Monitoring of health condition and level is very essential activity for patients to continue their life. Besides that, this monitoring is also help the counterparts in providing information for decision making and health budget planning.

For illustration, regression models have been presented to investigate the factors (variables), such as, blood pressure [1], sleeping hour [2],[3], weight or obesity [4] and calorie level [5] which affect the patient cholesterol level. However, the statistical assumptions are strictly required for row data (attributes) and unclassified data are not eliminated in the regression models. Additionally, these models are also not appropriate for categorical data analysis.

Rough set is commonly used in conjunction with other techniques connected to discretization on the data set. The main advantage of rough set data analysis is both non-invasive and notable ability to handle categorical data. This fits into most real life applications effectively [6].

Rough set theory has been widely applied to solve complex problems by researchers in recent years. For example, pattern recognition, emergency room diagnostic medical, acoustical analysis, power system security analysis, spatial-meteorological pattern classification, intelligent control systems and measure the quality of a single subset [6].

In medical diagnostic applications, the rough set approaches can be used to assist such inexperienced doctors in diagnosing based on clinical decision support model of disease symtoms [6, 7, 8, 9, 10]. The decision support model can be used to determine whether a patient can be discharged, requires further investigation, or consultation.

Some decision support model have the potential to provide accuracy recommendation as good as medical experts [6]. There are existing rough set applications in medical diagnostic procedure to detect diseases [7, 8] such as, dengue [6], diabetes mellitus [7, 8], chikungunya [7], and other. However, the step-by-step procedure in determining suitable rules for the medical diagnostic applications remains an interesting issue since the ultimate goal is to achieve accurate prediction results.

Motivated by application of rough set theory in various medical diagnostic applications [1-10, 16], we are interested to propose rough-regression approximations (RRA) using data reduction and elimination strategy. Furthermore, the proposed RRA can be used to improve the performance of Medipre system in providing preliminary diagnosis (information) of the patients.

This paper is organized as follows. In section 2, the principles of rough set and regression theories is presented. In sections 3 and 4, the proposed approximation and implementation are discussed. The summary is concluded in section 5.

2. FUNDAMENTAL THEORIES OF ROUGH SET AND REGRESSION MODEL

2.1 Rough Set Theory

The rough set theory has been introduced by Pawlak [11] and well divided by researchers into information systems, indiscernibility

HP3C 2018, March 15–17, 2018, Hong Kong, Hong Kong
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6337-2/18/03…$15.00
https://doi.org/10.1145/3195612.3195619

relation, set approximations, rough clustering, and others. An information system $S = (U, \Omega, V_q, f_q)$ consists of [12, 13]:

U : a nonempty, finite set called the universe;

Ω : a nonempty, finite set of attributes;

$\Omega = C \cup D$, in which C is a finite set of condition attributes and D is a finite set of decision attributes;

for each $q \in \Omega, V_q$ is called the domain of q;

f_q : an information function $f_q : U \rightarrow V_q$.

The cases, states, processes, patients, companies, and observations can be interpreted as objects or elements of rough sets. The attributes of each element can be assumed as symptoms, factors, and characteristic information. A relationship between conditional attributes and decision attribute can be explained using information table. In this table, the row and column correspond to objects and attributes, respectively. The starting point of rough set theory is the indiscernibility relation, generated by information about objects of interest.

Let $S = (U, \Omega, V_q, f_q)$ be an information system, then any subset B of A determines a binary (equivalence) relation IND(B) on U, which will be called B-indiscernibility relation, and is defined as follows:

$$\mathrm{IND}(B) = \{(x, y) \in U^2 : \forall a \in B, a(x) = a(y)\}, \qquad (1)$$

where $a(x)$ denotes the value of attribute a for element x in U. The collection of all equivalence classes determined by IND(B), denoted by U/B. An equivalence class of U/B, containing x, is denoted by $[X]_B$. In rough set theory, an equivalence class is the basic concepts of our knowledge. The indiscernibility relation will be used next to define approximations, basic concepts of rough set theory.

Let $S = (U, \Omega, V_q, f_q)$ be an information system and let $B \subseteq A$ and $X \subseteq U$. We can approximate X using only the information contained in B by constructing the B-lower and B-upper approximations of X. Both approximations are denoted as:

$$\underline{B}(X) = \{\{x \in U | [x]_B \subseteq X\}\}, \qquad (2)$$

and

$$\bar{B}(X) = \{x \in U | [x]_B \cap X \neq \emptyset\}, \qquad (3)$$

where $[x]_B$ is an equivalence class containing x. While, the difference between both approximations and its accuracy can be written:

$$\mathrm{BND}(X) = \underline{B}(X) - \bar{B}(X), \qquad (4)$$

2.2 Linear Regression Model

A simple linear regression can be written as [14]:

$$Y = \beta_0 + \beta_1 X + e, \qquad (5)$$

where Y is a dependent (endogenous) variable, X is an independent (exogenous) variable, β_0 and $\beta_1$are intercept and slope, while e is error of model. The algorithm in building Eq. (5) can be explained by the following steps:

Step 1: Check linear correlation between variables Y and X.

Step 2: Estimate intercept and slope, respectively using ordinary least square (OLS) method.

Step 3: Verify the significance X to Y using ANOVA and F-test.

Step 4: Verify the significance intercept and slope using t-test.

Step 5: Test the normality of residuals model Y.

3. BUILDING MEDIPRE USING ROUGH-REGRESSION APPROXIMATION

In building Medipre system, we apply Cross Industry Standard Process for Data Mining (CRISP-DM) which describes the life cycle of a data mining project in form of six different phases, such as, business (any area) understanding, data understanding, data preparation, modeling, evaluation and deployment [15]. The CRISP-DM will be implemented to build rough-regression approximation in improving Medipre system. While, rough-regression was implemented in determining the decisive criteria of cholesterol levels of patients recently [16]. However, the removing unclassified patients do not yet considered in building rough-regression model.

3.1 Data Preparation

This section provides rough-regression data and information that can be used to explain the correspondence between conditional attributes (exogenous variables) and decision attribute (endogenous attribute) as shown in Table 1.

Table 1. Rough-Regression data preparation

Rough Data Form	

Conditional attributes	Symptom A: categorical data; {Yes, No}
	Symptom B: categorical data; {Yes, No}.
	…
	Symptom P: categorical data; {Normal, High, Very High}

Decision attribute	Decision of disease: categorical data; {Yes, No}
Regression Data Form	
Exogenous variables	Symptom $A = x_1$; numerical data; {1, 0}. Symptom $B = x_2$: numerical data; {1, 0}. … Symptom $P = x_n$: numerical data; {1, 2, 3}
Endogenous variable	Decision of disease ; numerical data; {1, 0}.

3.2 Prediction Procedure

Step 1: In the first step, we have to determine boundary region (BR) between conditional attributes and decision attribute. Based on BR, we have to eliminate the elements of attributes which cannot be classified, since they have different decision attribute as presented in Table 2.

Table 2. Elimination of unclassified elements attributes

Before elimination					**After elimination**				
x_1	x_2	⋯	x_n	y	x_1	x_2	⋯	x_n	y
No	No	⋯	No	No	No	No	⋯	No	No
No	Yes	⋯	No	No	No	Yes	⋯	No	No
Yes	No	⋯	Yes	Yes	Eliminated information				
Yes	Yes	⋯	Yes	Yes	Yes	Yes	⋯	Yes	Yes

No	Yes	⋯	Yes	Yes	Eliminated information				
Yes	No	⋯	No	No	Yes	No	⋯	No	No
Yes	No	⋯	Yes	No	Eliminated information				
Yes	Yes	⋯	No	Yes	Yes	Yes	⋯	No	Yes
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
No	Yes	⋯	Yes	No	Eliminated information				
Yes	Yes	⋯	Yes	Yes	Yes	Yes	⋯	Yes	Yes

Step 2: Based on Table 2, the regression equations (before and after eliminations process) can be determined using the steps given in Section 2.2. In this case, we eliminate information of attributes using data reduction of rough set in building a new regression equation. Thus, the new equation can be used to predict the decision attribute more accurate if compared with regression equation before information elimination. In this case, we assumed that unclassified elements can be interpreted as noise information in data sets. So that, we have to eliminate them first. Moreover, the new regression equation after elimination can be called as rough-regression model (RRM).

Step 3: Based on Table 1, provide data reduction in information system until equivalent (intersection) data can be obtained.

Step 4: Based on Step 3, generate decision support rules.

Step 5: Based on Steps 2 and 4, predict the actual decision.

Step 6: Based on Step 5, verify the prediction error from the proposed model with the existing models based on average of percentage error as follows:

$$Average\ accuracy = \frac{number\ of\ corrected\ value}{total\ observation} \times 100\%. \qquad (6)$$

4. IMPLEMENTATION OF PROPOSED MEDIPRE SYSTEM

This section explains application of Medipre system for patient with dengue based on steps given in Section 3.2. The patient data set with possible dengue symptoms is presented in Table 3 [6].

Table 3. Patient data set with dengue symptoms

Patient code	Conditional attributes			Decision attribute
	BRS	MPA	T	Dengue
P1	No	No	Normal	No
P2	No	No	High	No
P3	No	No	Very High	Yes
⋯	⋯	⋯	⋯	⋯
P19	Yes	No	Normal	No
P20	No	Yes	Normal	No

Step 1: Determine boundary region (BR) as presented in Table 4.

Table 4. Lower-upper approximations and boundary region

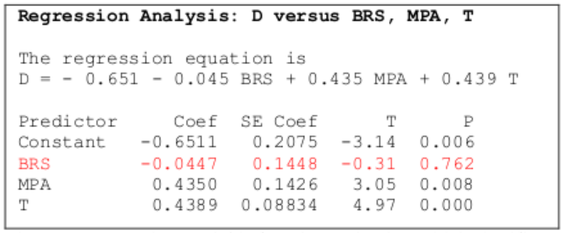
Rough set approximation	
Lower approximation	Upper approximation
The patients that are definitely have dengue = {P3, P4, P5, P6, P7, P13, P18}	The patients that possibly have dengue = {P3, P4, P5, P6, P7, P9, P13, P18}.
The patients does not have dengue = {P1, P2, P8, P10, P12, P14, P15, P16, P17, P19, P20}.	The patient that possible does not have cancer = {P1, P2, P8, P10, P11, P12, P14, P15, P16, P17, P19, P20}.
Determining BR	
BR for definitely have dengue	BR for possibly have dengue

BR = {P3, P4, P5, P6, P7, P13, P18}- {P3, P4, P5, P6, P7, P9, P13, P18}.
= {P9}
	BR = {P1, P2, P8, P10, P12, P14, P15, P16, P17, P19, P20}- {P1, P2, P8, P10, P11, P12, P14, P15, P16, P17, P19, P20}
	= {P11}

Table 4 shows that elements (patients) P9 and P11 cannot be classified, because they possess the same conditional symptoms (attributes), but with different conclusion in the decision attribute. Therefore, the symptoms of patients P9 and P11 are both inconclusive for next data reduction.

Step 2: build regression models before-after data reduction based on steps given in Section 2.2.

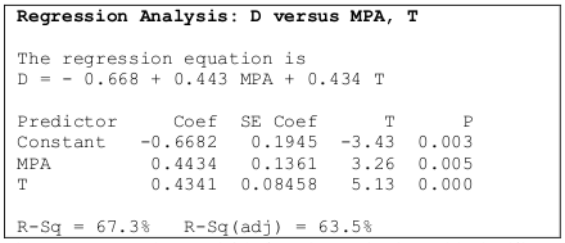
Regression model before data reduction is shown in Figure 1.



```
Regression Analysis: D versus BRS, MPA, T

The regression equation is
D = - 0.651 - 0.045 BRS + 0.435 MPA + 0.439 T

Predictor      Coef  SE Coef      T      P
Constant    -0.6511   0.2075  -3.14  0.006
BRS         -0.0447   0.1448  -0.31  0.762
MPA          0.4350   0.1426   3.05  0.008
T            0.4389  0.08834   4.97  0.000
```

Figure 1. Regression model before data reduction with insignificant variable.

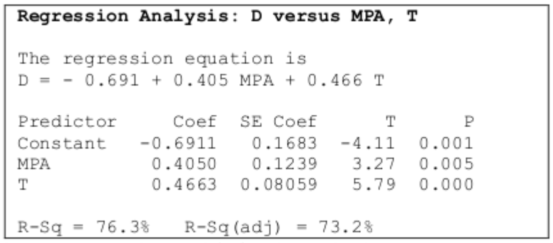
Regression after removing insignificant attribute is shown in Figure 2.



```
Regression Analysis: D versus MPA, T

The regression equation is
D = - 0.668 + 0.443 MPA + 0.434 T

Predictor      Coef  SE Coef      T      P
Constant    -0.6682   0.1945  -3.43  0.003
MPA          0.4434   0.1361   3.26  0.005
T            0.4341  0.08458   5.13  0.000

R-Sq = 67.3%   R-Sq(adj) = 63.5%
```

Figure 2. Regression model before data reduction with significant variables.

Regression model after data reduction (P9 and P11) is shown in Figure 3.



```
Regression Analysis: D versus MPA, T

The regression equation is
D = - 0.691 + 0.405 MPA + 0.466 T

Predictor      Coef  SE Coef      T      P
Constant    -0.6911   0.1683  -4.11  0.001
MPA          0.4050   0.1239   3.27  0.005
T            0.4663  0.08059   5.79  0.000

R-Sq = 76.3%   R-Sq(adj) = 73.2%
```

Figure 3. Regression model after data reduction

Based on Figures 2 and 3, the regression models are similar between before and after data reduction, except the constant (intercept) in term of *p*-value. Moreover, the correlation values from both models are different significantly. In Figure 2, the correlation value between decision (D) and attributes (MPA, T) is 0.673, this values indicates that 67.3% contribution MPA and T into D. While, 76.3% contribution MPA and T into D in Figure 3. In this case, we can interpret that data reduction procedure is very significant strategy in improving regression model performance.

Step 3: data reduction in information system. By following all steps given in data reduction and extraction [6], we only present the final result of intersection data with symptoms and decision attributes on Table 5.

Table 5 Final result of intersection data [6]

Patient code	Conditional attributes			Decision attribute
	BRS	MPA	T	Dengue
P1	No	No	Normal	No
P3	No	No	Very High	Yes
P4	No	Yes	High	Yes

Step 4: generate the decision support rules based on Table 5 as follows:

Rule-1 (R1):
If a patient with
Symptom BRS: No, and
Symptom MPA: No, and
Symptom T: Normal
Then decision of dengue: No.

Rule 2 (R2):
If a patient with
Symptom BRS: No, and
Symptom MPA: No, and
Symptom T: Very High
Then decision of dengue: Yes.

Rule 3 (R3):
If a patient with
Symptom BRS: No, and
Symptom MPA: Yes, and
Symptom T: High
Then decision of dengue: Yes.

Based on Rules 1-3, we can revise the same conditional symptom (attribute) as follows:

Refined Rule 1 (RR1)
If a patient with
Symptom MPA: No, and
Symptom T: Normal
Then decision of dengue: No.

Refined Rule 2 (RR2)
If a patient with
Symptom MPA: No, and
Symptom T: Very High
Then decision of dengue: Yes.

Refined Rule 3 (RR3)
If a patient with
Symptom MPA: Yes, and
Symptom T: High
Then decision of dengue: Yes.

Based on RR1-RR3, we revise again the same conditional attributes as follows:

Refined RR1 (RRR1)
If a patient with
Symptom T: Very High
Then decision of dengue: Yes.
Refined Rule 2 (RRR2)
If a patient with
Symptom MPA: Yes, and
Symptom T: High
Then decision of dengue: Yes.
Otherwise, the decision of dengue: No or possible dengue.

Step 5: predict the decision attribute (dengue) using proposed Medipre-RMAR and existing approximations as shown in Table 6.

Table 6. Prediction using proposed approximation

Patient code	Actual dengue	RT rules [6]	Refined Rules (RRT)	Regression model before reduction (RMBR)	Regression model after reduction (RMAR)
P1	No	No	No	No	No
P2	No	*	No	No	No
P3	Yes	Yes	Yes	Yes	Yes
P4	Yes	Yes	Yes	Yes	Yes
P5	Yes	*	Yes	Yes	Yes
P6	Yes	*	Yes	Yes	Yes
P7	Yes	*	Yes	Yes	Yes
P8	No	*	No	No	No
P9	Yes	*	Yes	Yes	Eliminated
P10	No	*	No	No	No
P11	No	*	Yes	Yes	Eliminated
P12	No	*	No	No	No
P13	Yes	Yes	Yes	Yes	Yes
P14	No	*	No	No	No
P15	No	*	No	No	No
P16	No	*	No	No	No
P17	No	*	No	No	No
P18	Yes	*	Yes	Yes	Yes
P19	No	*	No	No	No
P20	No	*	No	No	No
Accuracy		20%	95%	95%	99%
Rank		3	2	2	1

Table 6 shows prediction results using Rissino and Torress (RT) rules [6], refined RT rules, RMBR (Figure 2) and proposed RMAR (Figure 3) in predicting dengue. Moreover, our proposed RMAR able to predict all actual dengue precisely. While, other approximations have accuracy values are less than our proposed approximation. Furthermore, the comparison error is also done for flu, diabetes and cikungunya as presented in Table 7.

Table 7. Comparison accuracy between various approximations

Disease	Average Accuracy			
	RT rules [6]	Refined Rules (RRR)	Regression model before reduction (RMBR)	Regression model after reduction (RMAR)
Flu	57%	88%	89%	93%
Cikungunya	54%	83%	91%	96%
Diabetes	61%	89%	92%	95%

Table 7 indicates that our proposed RMAR in Medipre system has higher prediction accuracy if compared with RT, RRR, and RMBR approximations for flu, cikungunya and diabetes diagnostics.

5. CONCLUSION

In this paper, we proposed rough-regression approximation (RMAR) for improving Medipre system. The data elimination is main strategy for unclassified elements or the patients to build rough-regression approximations. This strategy is very significant in improving the prediction accuracy of RMAR if compared with the existing models and approximations. Additionally, the unclassified elements can be interpreted as "noise element". By using our strategy, this noise can be solved and eliminated.

Based on empirical studies, we can claimed that RMAR is better approximation than RT, RRT and RMBR in boosting Medipre system of dengue, flu, cikungunya and diabetes. Obviously, Medipre system can be developed into higher performance in analysis and accuracy using our proposed approximation. Therefore, the Medipre system can be used by unexperienced doctors and researchers for preliminary diagnostic of the patients.

6. ACKNOWLEDGMENTS

Our thanks to ORICC-UTHM for supporting our paper work under post-doctoral grant Vot. No. D003 at Universiti Tun Hussein Onn Malaysia (UTHM).

7. REFERENCES

[1] Sakurai, M., Stamler, J., Miura, K., Brown, I.J., Nakagawa, H., Elliot, P., Ueshima, H., Chan, Q., Tzoulaki, I., Dyer, A.R., Okayama, A., Zhao, L.: Relationship of dietary cholesterol to blood pressure: The intermap study. J. Hypertens. **29**, 222-228 (2011)

[2] Wolk, R., Somers, V.K.: Sleep and the metabolic syndrome. Exp. Physiology. **92**, 67-78 (2007)

[3] Kaneita, Y., Uchiyama, M., Yoshiike, N., Ohida, T.: Associations of usual sleep duration with serum lipid and lipoprotein levels. Sleep. **31**, 645-652 (2008)

[4] Miettinen, T. A.: Cholesterol production in obesity. Circulation. **64**, 842-850 (1971)

[5] Ueshima, H., Iida, M., Shimamoto, T., Konishi, M., Tanigaki, M., Doi, M., Nakashini, N., Takayama, Y., Ozawa, H., Komachi, Y.: Dietary intake and serum total cholesterol level: their relationship to different lifestyles in several Japanese populations. Circulation. **66**, 519-526 (1982)

[6] Rissino, S., Torres, G. L. Rough set theory-fundamental concepts, principals, data extraction, and applications, Julio Ponce and Adam Karahoca (Ed), Data Mining and Knowledge Discovery in Real Life App. Inform. (2009) 35-58.

[7] M. L. Thivigar, C. Richard, N. R. Paul. Mathematical innovations of a modern topology in medical events. International Journal of Information Science, 2(4)(2012) 33-36.

[8] R. Ali, J. Hussain, M. H. Siddiqi, M. Hussain, S. Lee. H2RM: A hybrid rough set reasoning model for prediction and management of diabetes mellitus. Sensors, 15(2015) 15921-15951.

[9] Tsumoto, S. Medical diagnosis: Rough set view. Studies in Computational Intelligence, 708(2017) 139 – 156.

[10] Tripathy, B. Application of rough set based models in medical diagnosis. Handbook of research on Computational Intelligence Applications and Bioinformatics, (2016) 108-118.

[11] Pawlak, Z.: Rough sets. Int. J. Compt. Inf. Science. **11**, 341-356 (1982)

[12] Hampton, J.: Rough set theory: the basics (part 1). J. Compt. Intel. Finance. **6**, 35-37 (1998)

[13] Tay, F.E.H., Shen, L.: Economic and financial using rough sets model. European J. Operation Research. **141**, 641-659 (2002)

[14] Wooldridge, M.: Introductory econometrics a modern approach. Third Ed. Thomson, South Western, USA (2006)

[15] Chapman, P., Clinton, J., Khabaza, T., Reinartz, T and Wirth, R. The CRISP-DM Process Model. August 2000

[16] Efendi, R and Deris, M. M: Decision Support Model in Determining Factors and Its Dominant Criteria Affecting Cholesterol Level Based on Rough-Regression. Recent Advance of Soft Computing and Data Mining (SCDM 2018), 243-251 (2018).

al_diagnosis_prediction_using_rough-regression_approximation.pdf

ORIGINALITY REPORT

24%	%	24%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Rasyidah, Nazri Mohd Nawi, Riswan Efendi. "Rough-Regression Model for Investigating Product Attributes and Purchase Decision", 2018 7th International Conference on Computer and Communication Engineering (ICCCE), 2018 Publication	8%
2	Ray, . "Rough Set", Soft Computing and Its Applications Volume One, 2014. Publication	7%
3	Mirza Zaeem Baig, Muhammad Usman Ul Haq, Hafiz Muhammad Umer Surkhail, Rabika Iqbal, Muhammad Mohsin Sheikh. "Bridging the industry-academia collaboration gap a focus towards final year projects", Proceedings of the 2nd International Conference on High Performance Compilation, Computing and Communications - HP3C, 2018 Publication	3%

4	"Advances in Robotics, Automation and Data Analytics", Springer Science and Business Media LLC, 2021 Publication	2%
5	Silvia Rissino, Germano Lambert-Torres. "Chapter 3 Rough Set Theory — Fundamental Concepts, Principals, Data Extraction, and Applications", IntechOpen, 2009 Publication	2%
6	Francis E.H. Tay, Lixiang Shen. "Economic and financial prediction using rough sets model", European Journal of Operational Research, 2002 Publication	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		