Incorporating Negative Association Rules to discover meaningful Outlier from Non_Reduct Computation : A Medical Predicitve Analysis

Faizah Shaari[1],Azmi Ahmad[2]
[1]Res. & Inno. Unit. , Polytehnic S.Salahuddin Abd.A.Shah, Shah Alam, Selangor, Malaysia.
[2]MIIT, University Kuala Lumpur, Malaysia
[1]sfaizah@psa.edu.my, [2]azmiahmad@miit.unikl.edu

Azuraliza Abu Bakar[3], Abd Razak Hamdan[3]
[3]Center for Artificial Intelligence Technology, Fac. of Tech. & Inf. Sci., Univ.Kebangsaan Mal. Bangi, Selangor DE, Malaysia
{aab,arh}@ftsm.ukm.edu.my

***Abstract*— Outlier Mining has always attract much attention among the data mining community. This paper discusses on the discovery of meaningful outlier detection based on Non_Reduct computation by incorporating the Negative Association Rules. Non_Reduct computation is proposed to detect outliers from rare classes. These outliers may have meaningful knowledge having incorporating the concept of Negatives rules to the outlier rules. Thus, a meaningful and comprehensive knowledge is expected to obtain for medical reasoning and predictive analysis by the experts in the field.**

Keywords-Negative Association Rules; Outlier; Non-Reduct; Infrequent ; Frequent itemsets;

INTRODUCTION

Much effort have been done by researchers to improve techniques in predictive mining especially for medical diagnosis. Among common and popular intelligence techniques for medical predictive analysis are Neural Network, Bayesian classifier, Genetic algorithm, Decision Tree, Fuzzy Logic and Rough Sets Theory(RST). RST has been proposed by [1] to discover the knowledge of breast cancer data using classification approach. More interesting, [2] proposed a predictive medical analysis from Outlier Mining concept named Rough_Outlier Algorithm(RSetAlg). RSetAlg uses the concept of Non-Reduct computation from Rough Set Theory approach. Non_Reduct computation followed the computation of reduct as in Rough Set Theory. The Non_Reduct is computed to detect outliers from rare cases. The computation involves the formulation of Indiscernibility Matrix Modulo Decision (iDMM *D)* and Indiscerniblity Function Modulo D(iDFM *D*). The results showed that the computation of Non_Reduct is able to diagnose symptom of sickness in patients through abnormal data and rare case at fast speed.

In this study, the results from Rough_Outlier algorithm is explored to improve the predictive medical analysis by incorporating Negative Association Algorithm(NAR) to the rules obtained. Study over NAR has been done by researchers and it is believed that NAR rules are hidden pattern discovered among infrequent and high-relational itemsets. NAR is important in many applications where it enables to identify items which are rarely occur together with common itemsets. Hence, a total overview and better coverage of search space would be able to generate and discover significant knowledge for predictive medical analysis.

The following section II discusses on Outlier mining methods including the computation of Non_Reduct based on RST to detect outliers. NAR algorithm is discussed in section III. While, section IV describes the experiment and results obtained. The conclusion will finally discuss in section V.

OUTLIER MINING

Outlier mining focuses on the rare data whose behaviour is very exceptional when compared with the rest of the large amount of data. This exception identification can lead to a discovery of unexpected knowledge.

Outlier mining has been realized from several approaches or technologies in the field of statistics, machine learning, artificial intelligence, visualization and database management. Finding these outliers in large dataset has drawn increasing attention among researchers [4-9]. In detecting outliers based on Rough_Outlier Algorithm(RSetAlg), the concept of Non_Reduct computation as in [11] is explained. Generally, Non_Reduct computation followed the computation of reduct as in Rough Set Theory. The Non_Reduct is computed to detect outliers from rare cases. The computation involves the formulation of Indiscernibility Matrix Modulo Decision (iDMM *D)* and Indiscerniblity Function Modulo D(iDFM *D*).

A measurement, *RSetOF* is used to detect outliers. *RSetOF* with large value indicates that the object is unlikely to be an outlier. In contrast, objects with small *RSetOF* values is detected as outlier.

MINING NEGATIVE ASSOCIATION RULES

Positive Association Rules

Association Rules or Positive Association Rules(PAR) have been extensively studied in the literature for their usefulness in many application domains such as recommender system, medical diagnosis, intrusion

detection and telecommunication. Association Rules task is to find the relationship between items in a dataset. For a huge and massive databases, this task is able to show interesting relationships among the itemsets which then uncover hidden knowledge between the known values. Typical example of Association rules mining is market basket analysis.

From the original *apriori* algorithm there have been a remarkable number of variants and improvements of association rule mining algorithms[14,15,16,17]. The algorithm finds large quantity of interesting relationship or correlation among item sets and is represented in form of rules. These rules indicate that the presence of some itemsets will imply the presence of other itemsets within the same transactions. For an association rule A=> B, it is used to predict that 'if A occurs, then B generally also occurs'.

Negative Association Rules

Mining traditional association rules is extended by many researchers into mining Negative Association rules. Currently, Negative Association Rules(NAR) is found to play important roles in decision making with cutting of search space of a different angle and perspective than the Positive Association Rule which is already known as a useful decision making algorithm. The Negative Association rules describe negative relationships in the itemsets and imply that the occurrence of some itemsets by the absence of others. In other word, these itemsets are not positively associated with any of the itemsets within the same transaction. One simple example that can be used here is when a patient who complains of headache does not have a throbbing pain, migrane should not be suspected with high probability[18,19,20].

Example of NAR rule can be in the form of A=> ¬ B or ¬A=>B. In addition to positives associations, the negative can provide valuable information in devising many important decision making system like marketing strategies, investment analysis and many more.

In this study, the Rough_Outlier detection method will generate list of outliers rules that belong to rare class. These rules are small and differ from the rest of the dataset. Although they are small in size and rare, these rules are assumed as frequent rules in the sense that it was generated and processed from outlier mining concept which is significant and have important knowledge to uncover. In enhancing the Rough_Outlier algorithm, NAR is incorporated to the method by assuming that the hidden NAR rules can be discovered from the infrequent itemsets generated.

Searching for outlier using Non-Reduct computation is an advantage as it search space is small as attributes of non-interesting is reduced and they are of rare class in comparison when searching using positive association rules which has exponential rules generated. As the result, the computation and the processing time is reduced, which allows better time for detection rate of outliers. By incorporating NAR, the results of outliers detected is expected to give a comprehensive and significant knowledge from the medical reasoning and interpretation from the experts and professional of that field.

According to [18], medical reasoning can be obtained from positives and negatives rules which not only to reflect experts' decision but also for domain experts to interpret both which are important to enhance the discovery process through both cooperation. [21] describes that positive and negatives rules gives better classification accuracy thus helps in reasoning with less classification time. In devising marketing strategies, [22] highlights the importance of NAR besides PAR where better decision can be made.

EXPERIMENTAL AND RESULTS

By following the work in [22], a discovery from PAR and NAR rules are observed and followed. First, Rough_Outlier Method is experimented upon Heart Disease dataset. The following sub-section *A* describes the process to discover the outliers rules.

Experimental Results from Rough_Outlier Detection Method

Rough_Outlier Method is tested on Heart Disease dataset to detect outliers. The results from the Non-Reduct computation is analysed. There are 344 rules generated from the computation based on Non_Reduct[11]. These rules obtained are outliers rules or frequent rules generated from Non_Reduct computation. The rules are ranked using RSetOF value to look for outliers. As dicussed in section II, any object or equivalence class which has more support would unlikely to be outlier. Otherwise the objects are outliers. Results shows that based on Rough_Outlier Algorithm detection method , five outliers which belong to rare class are identified with least RSetOF value and are detected at top ratio 6.25%. Table I illustrates seven outliers detected from Rough_Outlier Algorithm Method which are e133, e134, e135, e136, e137, e138 and e139. The second column in the Table I denotes outliers rules of each equivalence class obtained. There is only one rule for each equivalence class generated. Only partial of frequent rules obtained for each class obtained are showed.

TABLE I. OUTLIERS RULES GENERATED FROM ROUGH_OUTLIER ALGORITHM METHOD FOR HEART DISEASE DATASET

Tid	Outliers Rules
e133	[Age<50]^ [male]^[chest pain: Type 3 Or 2] => HeartDisease
e134	**[Age>50] ^ [male]^ [chest pain: Type 4] => HeartDisease**
e135	**[60<Age>62] ^ [male]^[chest pain: Type 4]=> HeartDisease**
e136	[63<Age>71]^ [female^[chest pain: Type 4]=> HeartDisease
e137	**[60<Age>62]^[male]^[chest pain: Type 1] => HeartDisease**
e138	**[57<Age>59)]^[male]^[chest pain: Type 4]=> HeartDisease**
e139	**[Age <57)]^ [male]^[chest pain:Type 4] =>=> HeartDisease**

Observation I and Discussion

Regardless of age, there are more male patients whom are diagnosed with heart disease than female patients. The male patients whom are diagnosed with the presence of heart disease are found not to have any chest pain symptom. These are shown as the patients in TID e134, e135, e138, e139 whom have chest pain of type 4 that is asymptomatic.

In differential diagnosis of heart disease where NAR rules is explored, results on transaction e134,e135, e138 and e139 as in Table II are observed.

TABLE II. NAR RULES GENERATED FROM TABLE I

Tid	NAR Rules
~~e133~~	~~[Age<50]^ ¬ [male]^[chestpain: Type 3 Or 2] => HeartDisease~~
e134	**[Age>50] ^ ¬ [male]^ [chestpain: Type 4]=> HeartDisease**
e135	**[60<Age>62] ^ ¬ [male]^[chestpain: Type 4]=> HeartDisease**
~~e136~~	~~[63<Age>71]^ ¬ [female^[chestpain: Type 4]=> HeartDisease~~
~~**e137**~~	~~**[60<Age>62]^ ¬ [male]^[chestpain: Type 1] => HeartDisease**~~
e138	**[57<Age>59)]^ ¬ [male]^[chestpain: Type 4]=> HeartDisease**
e139	**[Age <57)]^ ¬ [male]^[chestpain:Type 4] => HeartDisease**

Table II illustrates Negative rules obtained from the frequent outliers rules from Table I. The occurrence of chest pain itemset with the absence of female itemset gives different reasoning or diagnosis on the rules obtained. Interesting NAR rules are derived for patients e134, e135,e138 and e139. Regardless of age, it is found that the absence of female patients whom are diagnosed with heart disease have no symptom of chest pain.

The results are found true where medical health report in [23] reported that chest pain is the most common symptom of heart attack however not everyone experiences chest pain during a heart attack. Women in fact are less likely than men to feel chest pain during heart attack. Most women experienced so-called "atypical" symptoms such as back pain, nausea or fatigue.

Therefore, in this medical reasoning obtained from both observations of frequent outliers rules and NAR rules, it can be concluded that it is important for women to be diagnosed with various signs and symptoms of a heart attack so that efficient decision can be made.

Observation II and Discussion

The frequent outliers rules as in Table III are found. The rules denoted that male with high cholesterol in their body ,are diagnosed with heart disease. These are shown as in patients e134, e135 and e139. The results are partial of rules in Table I. The combination of both rules from Table I and III for e134, e135 and e139, can be interpreted as : regardless of age, male patients whom has no chest pain symptom but have high cholestrol are diagnosed with heart disease.

TABLE III. NAR RULES GENERATED FROM TABLE I

Tid	Outliers Rules
e133	**[Age<50]^ [male]^[CHOLESTROL:>250mg/dl=> HeartDisease]**
e134	**[Age>50] ^ [male]^[CHOLESTROL:>250mg/dl]=> HeartDisease]**
e135	**[60<Age>62]^[male]^[CHOLESTROL:>250mg/dl]=>HeartDisease**
e136	[63<Age>71]^[female]^[CHOLESTROL:>250mg/dll}=>HeartDisease
e137	[60<Age>62]^[male]^[CHOLESTROL:<250mg/dl]=>HeartDisease
e138	[57<Age>59)]^[male]^[CHOLESTROL:<250mg/dl=>HeartDisease
e139	**[Age<57)]^[male]^[CHOLESTROL:>250mg/dl]=>HeartDisease**

In this domain , it is found that the corelations between the absence of female with high cholesterol is interesting as it indicates the presence of heart disease. With this corelation, it helps to diagnose those patients whom have chest pain of type 4 which not able to tell whether the female patient suffer from heart disease or not.

CONCLUSION

This research intends to prove the hypotheses that the incorporation of Negative Association rule on Non-Reduct outliers rules would able to produce a good associative-outlier detection method that generate interesting knowledge. The above discovery shows the advantage of using NAR in Rough_Outlier Algorithm for medical reasoning by the experts.

It shows that in some important cases, although strong frequent outliers rules generate interesting knowledge with good decision that is predictable but by incorporating, something that is contradict from common belief or knowledge; using NAR; it able to give unexpected comprehensive knowledge for efficient decision making.

REFERENCES

[1] A.E. Hassanien,, J.M.H. Ali : Rough Sets Approach for Generation of classification Rules of Breasr Cancer Data. INFORMATICA. Vol 15,no.1, pp.23-28. Institute of Mathematics and Informatics, Vilnius(2002)

[2] S. Faizah, A.B. Azuraliza, Razak A. H. : A Predicitve Analysis on Medical Data based on Outlier Detection Method using Non-Reduct Computation. Int.Conference Advance Data Mining, ADMA 2009, pp.68–73.

[3] Z. Pawlak, *Rough set theory and its applications to data analysis.* Cybernetics and systems, 1998. **29**(7): p. 661-688.

[4] V. Hodge and J. Austin, *A survey of outlier detection methodologies.* Artificial Intelligence Review, 2004. **22**(2): p. 85-126.

[5] Z. Long, et al., *Multiple Attribute Frequent Mining-Based for Dengue Outbreak*, in *Advanced Data Mining and Applications*, L. Cao, Y. Feng, and J. Zhong, Editors. 2010, Springer Berlin / Heidelberg. P. 489-496.

[6] C.C. Aggrawal, & P. Yu, *Outlier Detection in High Dimensional Data*. In *SIGMOD'01*. 2001. Santa Barbara.

[7] M.M. Breunig, PNg.R.T H.Krigel, & J. Sander, *LOF:Identifying density-based local outlier*. In *Proc. ACM SIGMOD Int Conf on Management of Data*. 2000. Dallas, Texas.

[8] F. Jiang, Y. Sui, and C. Cao, *An information entropy-based approach to outlier detection in rough sets.* Expert Systems with Applications, 2010. **37**(9): p. 6338-6344.

[9] Z. He, X. Xu, and S. Deng, *Discovering cluster-based local outliers.* Pattern Recognition Letters, 2003. **24**(9-10): p. 1641-1650.

[10] A.A. Bakar, *Propotional Satisfiability Method in Rough Classification Modelling for Data Mining*, 2001, University Putra Malaysia: Seri Kembangan.

[11] Shaari, F., A.B. Azuraliza, and A.R. Hamdan, *Outlier detection base on rough set theory.* Intelligent Data Analyisis 2009. **13**: p. 191-206.

[12] Z. He. et al., *A frequent pattern discovery method for outlier detection.* Advances inWeb-Age Information Management, 2004: p. 726-732.

[13] S. Hawkins, H. He, G. William & R.Baxter: Outlier Detection using Replicator Neural Network DaWak 2002 (2002).

[14] J. Liu, J., X. Fan & Z. Qu, A New Interestingness Measure of Association Rules. Second, International Conference on Genetic and Evolutionary Computing, 2008, pp. 393-397.

[15] P. Thongtae, & S, Srisuk. An algorithm for reusable Uninteresting Rules in Association Rule Mining. 2008.

[16] R. Agrawal, T. Imielinski & A. Swami, Mining Association Rules between set of itemsets in large database. Proc. Of the 1993. ACM SIGMOD Conf. 207-216.

[17] R. Agarwal, C. Aggarwal, and V.Prasad, A Tree projection algorithm for generation of frequent itemsets. In J.Parallel and Distributed Computing 2000.

[18] S. Tsumoto, Mining Positive and Negative Knowledge in Clinical Database based on Rough Set Model, PKDD 2001, LNAI 2168.2001.

[19] Y. Gang, & C. Li, A Novel Mining Algorithm for Negative Association Rules. Global Congress on Intelligent Systems, pp 212-219. 2009

[20] Anis Suhaili Abdul Kadir, Azuraliza Abu Bakar, Abdul Razak Hamdan, 2011. Frequent Absence and Presence Itemset for Negative association Rules Mining. *The Int.Conf on Intelligent System Designs and Applications (ISDA11)*. 29 Nov – 1 Dec 2010. Cordoba, Spain.

[21] T. Ma, J. Leng, M., Cui, W. Tian,:Inducing Positive and Negative R ules on Rough Set. , Inf. Tech. Journal Vo8(7): pg 1039-1043. ISSN 1812-5638. 2009.

[22] B. Ramasubbareddy, A. Govardhan, A. Ramohanreddy, Mining Positive and Negative Association Rules. The 5th Int. Conf. On Computer Science & Education Hefei, China. August pg: 24-27, 2010.

[23] B. D'Antono, G. Dupuis, R. Fleet, A. Marchand, D Burelle. Sex differences in chest pain and prediction of exercise-induced ischemia. *Can J Cardiol.* 2003;19:515-522.