Exploring Different Aspects of Data Mining for Business

Mazin Kadhum Hameed
Department of Software, Collage of Information Technology, university of Babylon
Mazin_kadum2000@yahoo.com

Abstract

Data Mining, also commonly known as Knowledge Discovery in Databases (KDD), mentions to the indirectly mining of implied, earlier unknown and possiblyvaluable information from data in databases.In this paper we focused on two important problems, the first students of information technology collages have greater capability in their specialist but its un clear, the second the trend in this time is using the electronic achieve but is generate huge data resulted form daily work which is difficult to control on it.in this paper we using data mining techniques rule based classification to classify the student of information technology to classes according to scientific capabilities and using Attribute Oriented Induction(AOI) data mining techniques to solve problem of the electronic achieve. Experimental results show that the proposed techniques enhancement the method of student classification and reduce the problem that resulted from using electronic achieve., that would help improve customer experiences and decision-making.

Keywords: Data Mining, Knowledge Discovery in Databases (KDD), Attribute Oriented Induction (AOI)

الخلاصة

يشير استخراج البيانات، المعروف أيضا باسم اكتشاف المعرفة في قواعد البيانات (كدد)، إلى التعدين غير المباشر من ضمنية، معلومات غير معروفة في وقت سابق ويمكن تقديرها من البيانات في قواعد البيانات. في هذا البحث ركزنا على مشكلتين مهمتين، المشكلة الاولى هي ان طلبة كلية تكنلوجيا المعلومات لديهم قدرة كبيرة في اختصاصهم ولكنها مبهمة وغير واضحة، اما المشكلة الثانية هو انه في الوقت الحاضر يتم استخدام الارشفة الإلكترونية ولكنها تولد بيانات ضخمة نتيجة عن العمل اليومي الذي من الصعب السيطرة عليها.في هذه البحث نستخدم تقنيات تعدين البيانات القائمة على التصنيف لتصنيف طلبة تكنولوجيا المعلومات إلى طبقات وفقا لقدراتهم العلمية واستخدام السمات الموجه التعريفي (AOI لحل مشكلة الارشفة الإلكترونية. وتبين النتائج التجريبية أن التقنيات المقترحة تحسين عملية اختيار افكار ومشاريع تخرج تتناسب مع المفردة العلمية للطلبة وتقلل من المشكلة الناتجة عن استخدام الارشيف الإلكتروني والتي من شأنها أن تساعد على تحسين عمل المستخدمين وصنع القرار.

1. Introduction

Data mining technology [Divya and Vijendra, 2016] appear as a talented field and is used in extensive application areas like bank transactions, scientific experiments ecommerce microarray gene expression data, etc [Divya and Vijendra, 2016].The data mining is mentioned as detection of relations in huge databases routinely and in various circumstances, data mining is used for concluding relations depend on the consequences revealed. It acts a significant task in several applications as health care industry, e-commerce, scientific and engineering, business organizations [Vijayarani, & Sudha, 2013].

The data mining approaches, used for mining veiled patterns from the data, are categorized into the following two groups: prediction methods and description methods. Prediction-oriented methods wish to automatically construct an interactive

model, which gets new and hidden models and is capable to guess values of one or more variables associated to the model. Description methods are concerned with the data analysis, which emphases on considerate (by visualization for example) the way the original data relates to its parts [Oded and Lior 2010].

Summarization is data mining technique that includes methods for discovery a compressed notation of a dataset.[Sayal and Scheuermann, 2001]

Summarization is the procedure to summary data in a important and well informed way, to its significant and related structures. [Kocherlakota & Healey, 2005]

The classification is data mining technique that used to supervised learning or unsupervised learning [Neelamegam, & Ramaraj,2013].

The main focusing of this paper is classify the students of the IT collage in an intelligent way and manage the electronic archive of the Babylon university in the applicable manner that would help improve customer experiences and decision-making.

This structured of this paper is as follow. Part 2 surveys Rule Based Classification. Part3 presents AOI algorithm, primitives, and weaknesses. The classification of students and electronic archives are stated in Part4 and 5. The experiments and results are stated in Part 6. Finally, Part 7 concludes this paper.

2. Rule-Based Classifier

Classification is data mining technique that used to guess collection participation for a data. Classification includes seeking for rules that representative information and splits the data into disjoint clusters. Set of IF-THEN rules are treatment by a rule-based classifier for classification [Thangaraj & Vijayalakshmi,2013].A rule based classifier is a collection of rules of type(if (condition) –then conclusion).

The rules of classification are establish in two ways a) Indirect method b) Direct method [Thangaraj &Vijayalakshmi, 2013].

Indirect methods are citation rules as decision trees, e.g. C4.5rules whereas the direct or consecutive approaches are citation rules straightly from data. Rule-based classifications have the following advantages [Thangaraj, & Vijayalakshmi, 2013]:

- Simple to comprehend
- Simple to create
- Greatly crossing as decision trees
- Can organize new cases rapidly

Pseudo code for sequential covering

Algorithm : Sequential Covering (DD, At-valss)
In: A relation dataset DD with tuples with class-labeled, fields and their values At-valss.
Out : A collection of rules RR(if-then).
Process:
RR = clear set; // original group of rules learn is clear
For every class cc do
Repeat
R = Learn one rule (DD, At_valss, cc)
DD = DD- Rule // eliminate the rows cover by R from DD
Until rows in DD = NULL;
R R= R R+ R; // enter new rule to RR
End For

Return RR;
End

3.AOI (Attribute Oriented Induction)

AOI was advanced to learn altered types of knowledge rules such as classification rules, characteristic rules, discernment rules, cluster description rules, [Han, *et al.,* 1993], association rules and data development regularities rules [Han & Fu, 1995]. All data that stored in database are satisfied by the notions that defined by the characteristic rule in AOI. This rule offers general concepts about a feature, which can aid persons, identify the mutual structures of data in a class, an example of such state is the indication of the exact illness [Cai,1989].

Attribute oriented induction (AOI) uses quality rule to distinguish, knowledge and mining as an exact nature for each of feature as their particular mining description with assist of concept hierarchy as the normal saving surroundings knowledge to discover goal class as an optimistic knowledge [Warnars, 2015]. Attribute oriented induction (AOI) algorithm has seven approach steps in procedure of simplification [Han, *et al.,* 1992] and they are:

1) Simplification on the minimum destructible ingredients where simplification must be done on the minimum destructible Ingredients on data relative.
2) Attribute elimination, the attribute must be removed throughout generalization if there is dissimilar values for an attribute.
3) Concept tree ascension, the replacement of the value by its higher-level concept hierarchy tree is done.
4) Vote generation, by depending on the majority value which is counting tuples value, the majority value was estimated after merger the same tuples in the simplification.
5) Threshold manage each feature, additional generalization on this attribute must be done if the determined threshold value < n.of distinct values in a derived relation.
6) Threshold manage generalized relatives,if the n.of tuples is > the particular threshold value, then extra simplification will be made depend on the assured feature and the duplicate tuples must be combination.
7) Rule conversion, will be done by varying last simplification to quantitative rule and qualitative rule from a tuple (conjunctive) or several tuples (disjunctive).

The quality rule of AOI algorithm [S. Warnars,2010] as shown below has two sub processes which is threshold manage each feature and number of tuples[S.Warnars, 2015]. Firstly, AOI algorithm started with input threshold with range integer number greater than zero and this threshold shown in line number two and nine. If this threshold is small number this guide to easy rule with additional ANY value that is predictable as unexciting value. However, if this threshold is bigger number this will lead to little generalization. The process of selecting the appropriate number that is set to threshold will affect the number of distinct values in every feature.

AOI algorithm

1) For each field Aii (i >= 1 and i <= n, n = n.of fields) in the GR(generalized relation)
2) { While
   n.of distinct values in the field Aii greater than threshold
3) {If there are no higher level concept in concept hierarchy tree for the field Ai
4) Then delete field Aii
5) Else replaces the value of the field Aii by its equivalent smallest generalized concept
6) combine the same tuples
7) }
8) }
9) While n.of tubles in GR greater than threshold
10) {Select the generalize fields
11) combine the same tuples
12) }

The initial line of AOI algorithm shows the looping process depending on a number of features in database. The procedure continue among the line number 2 and 9 to check if a number of distinct values for each feature in dataset is greater than threshold value otherwise the process will return to the 1 line and assigned to the next feature.

In line number 3 if each feature/column has higher level concept hierarchy then line number 5 will be done, otherwise the feature will eliminate. Later, the line number six will be done so as to combine the same tuples. Line numbers between 9 and 12 are discussed in the above [Warnars,2010]. The little number of level and amount of concepts in concept tree will have easy simplification procedure except great number will have extra simplification process shows in Figure 1 [Warnars, 2010].

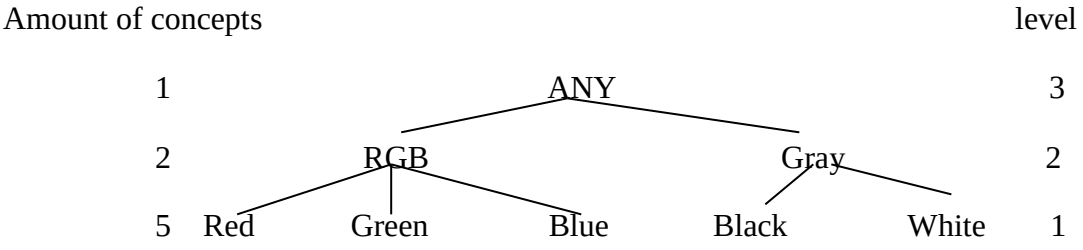

Figure 1 Amount and level of Concept tree

4. Classification of Students

In this research, a sample of students from the College of Information Technology is taken and worked on it. Each student receives the grades obtained in the first, second, third and forall subjects that he studied.
This data is then stored in the database, which will include data for all students and for the three years.
In this research, students are divided into seven types :

1- Programmer
2- Database
3- Languages
4- Mathematics
5- Computer skills
6- Image processing
7- Compilers

4.1Programmer

In this type the student must obtain good or very good or privilege for the following materials.

1. Fundamentals of Programming
2. Object Oriented
3. Artificial intelligence
4. Applications of artificial intelligence

4.2 Database

In this type the student must obtain good or very good or privilege for the following materials.
1-Data structures

2-Concepts of databases

4.3 Mathematics

In this type the student must obtain good or very good or privilege for the following materials.

1-Intermittent structures

2 - Slaked structures 2

3-Differentiation or integration

4-Differentiation or integration

5- Operations research

6- Information theory

7- Numerical analysis

8- Statistics and Probability

4.4 Computer Skills

In this type the student must obtain good or very good or privilege for the following materials. The materials that a student must obtain with good grades are:

1-Architectural

2. Parallel processors

3 - microprocessors

4- Logical design

5- Computer installation

6 - computer skills

4.5 Image Processing

In this type the student must obtain good or very good or privilege for the following materials.

1 - Drawing computer

2-Communications

3-Analysis and design of algorithms

4- Image processing

4.6 Compilers

In this type the student must obtain good or very good or privilege for the following materials.

The materials that a student must obtain with good grades are:

1- Compilers 1-2

2 - Calculation 1-2

3. Data structures

5. Electronic Archives

The electronic archive is one of the most important and modern applications used in various fields. It is used in the exchange of official books issued to different facilities of the society. The benefits of using the electronic archive are the speed and accuracy of sending and receiving official books. However, huge data are resulting from using the electronic archive on a daily basis, this huge data make difficult to control on it about storage and decision making. In this research, AOI technique is

used to obtain a summary of data without losing important information. In this research work was done on the electronic archive of the University of Babylon where a database was created containing the main details of the official books issued and received between the presidency of the university and all colleges. A concept trees for department and Presidency University are shown in Figure2.

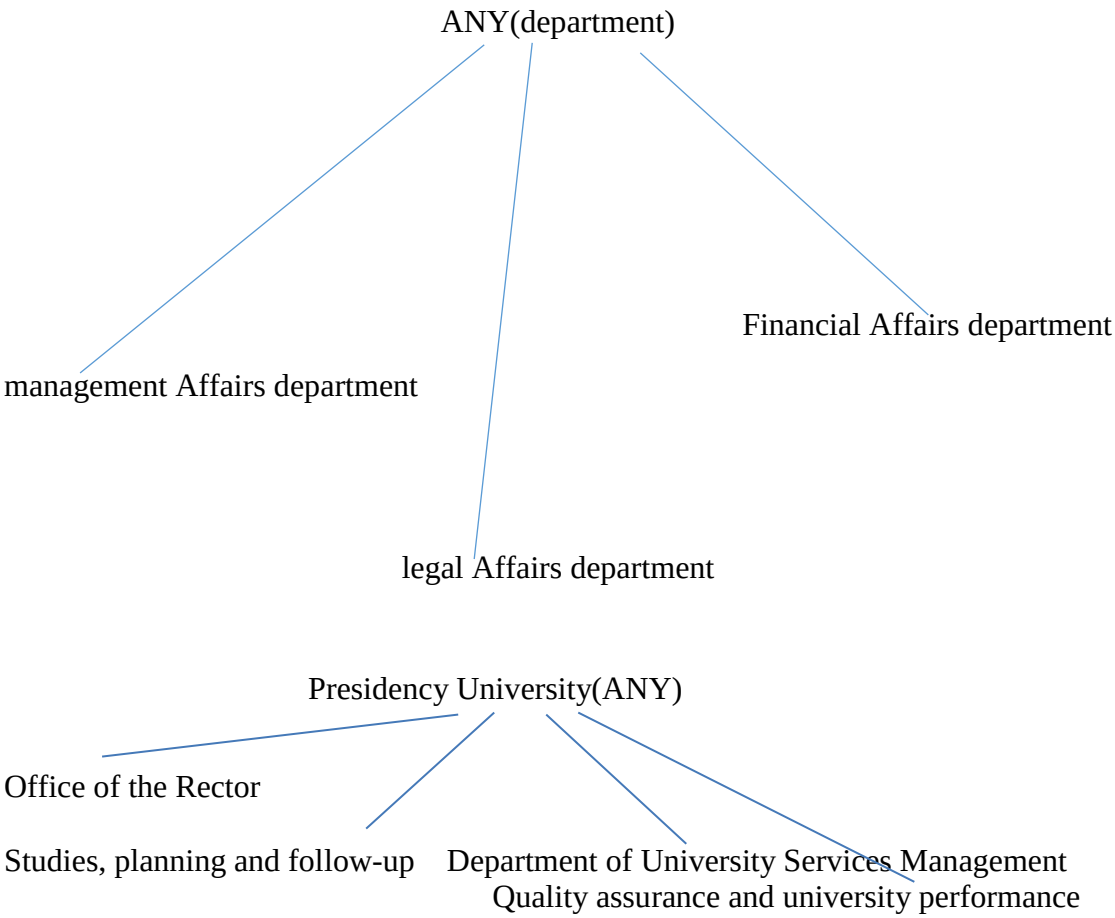

Figure 2 A concept trees for department and Presidency University

6. Experiments and Results

The process of selecting a graduate project or the idea of research and assigning it to students is a complex process due to the intellectual and scientific diversity of the students. In this research, the results proved that the decision making process of assigning research idea to students becomes easy. In the case of the electronic archive where the results proved the process of decision-making process and the search process is enhancement. Figure 3 shows the result of AOI. Experimental results show

that the process of assigning graduate research to students depends mainly on the students' scientific competence.

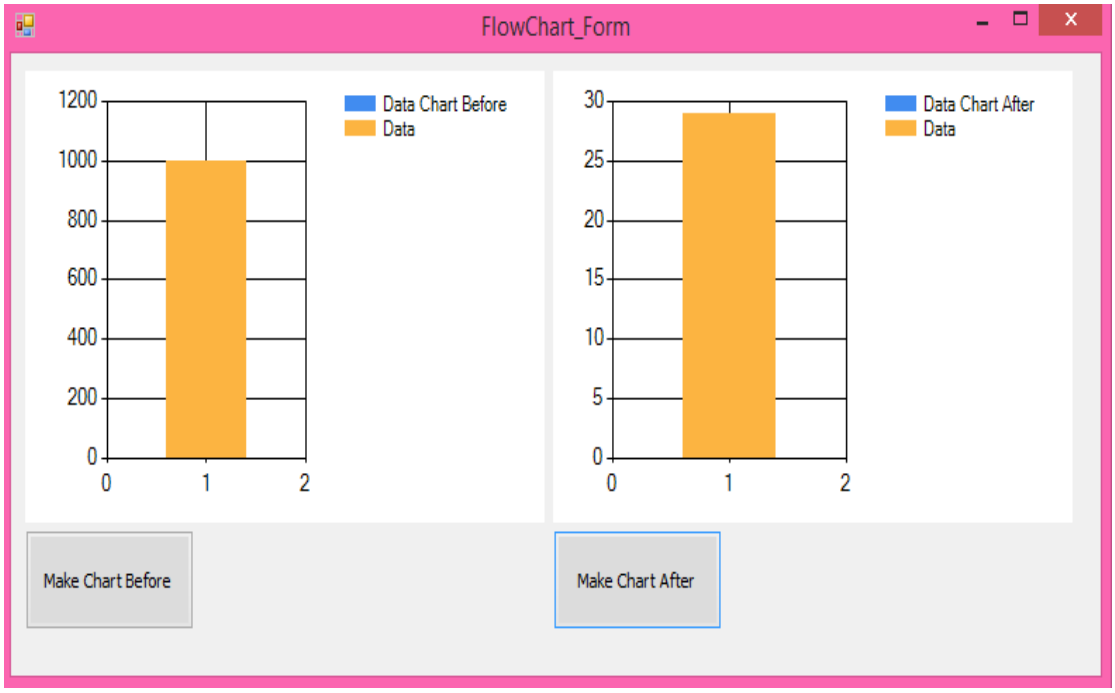

Figure3 result of AOI

The previous figure shows that the size of data entered before the implementation of the algorithm is 1000 rows and after the implementation of the algorithm became 28-row size.

7. Conclusions

The classification process does not affect by the number of students or the number of subjects but is affected if the student obtains equal estimates in all subjects, since it can't be classified into a specific type and will be left to manual classification. In the case of the electronic archive, the algorithm does not influence by the increase in the number of books issued or received as well as there is no overgeneralization problem.

8. References

Divya Jain and Vijendra Singh,2016" Utilization of Data Mining Classification Approach for Disease Prediction: A Survey", I.J. Education and Management Engineering, VOL. 6, pp.45-52.

J. Han, Y. Cai, N. Cercone, 1992,"Knowledge discovery in databases: An attribute-oriented approach", Proceeding of 18th International Conference Very Large Data Bases, Vancouver, British Columbia, 547-559.

J. Han, Y. Cai, N. Cercone, 1993,"Data-driven discovery of quantitative rules in relational databases",IEEE Transction on Knowledge and Data Engineering, 5(1), 29-40.

J. Han, Y. Fu, 1995,"Exploration of the power of attribute-oriented induction in data mining", in U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds. Advances in Knowledge Discovery and Data Mining, 399-421.

M. Sayal and P. Scheuermann,2001,"Distributed web log mining using maximal large item sets"Knowledge and Information Systems, 3(4),pp.389-404.

M. Thangaraj,C.R.Vijayalakshmi,2013," Performance Study on Rule-based Classification Techniques across Multiple Database Relations", International Journal of Applied Information Systems,Vol.5,NO. 4,pp 1-7.

Oded Maimon and Lior Rokach, 2010,"Data Mining and Knowledge Discovery Handbook". Springer Publishing Company, USA, Second Edition, pp.21-29.

S. Warnars,2010," Measuring Interesting rules in characteristic rule". Proceeding of the 2nd International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT), Bali, Indonesia, 152-156.

S.Neelamegam, Dr.E.Ramaraj,2013," Classification algorithm in Data mining: An Overview", International Journal of P2P Network Trends and Technology (IJPTT) ,Vol. 4 ,No. 8,pp. 369-374.

S.Vijayarani, S.Sudha,2013," Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, Vol. 1, NO.3,pp.735-741.

S.Warnars, 2015,"Mining Patterns with Attribute Oriented Induction", Proceeding of International Conference on Database, Data Warehouse, Data Mining and Big Data (DDDMBD), Tangerang, Indonesia, 11-21.

Sarat M. Kocherlakota, Christopher G. Healey,2005," Summarization Techniques for Visualization of Large Multidimensional Datasets", Technical Report,pp.1-18.

Y. Cai,1989, "Attribute-oriented induction in relational databases", Master thesis, Simon Fraser University.