Enhancing Attribute Oriented Induction Of Data Mining

Safaa O. Al-Mamory
Department of Software ,College of Computer Technology, university of Babylon

Saad Talib Hasson **Mazin K.Hammid**
Department of Computer Science, College of Science , university of Babylon

Abstract

Data summarization is a data mining technique to summarize huge data in few understandable knowledge. Attribute-Oriented Induction(AOI) is a data summarization algorithm, it suffer from overgeneralization problem. In this paper, we use an entropy measure to enhance generalization process, feature selection, and stop condition. Experimental results show that the proposed technique will reduce the effect of overgeneralization problem.

الخلاصة

تلخيص البيانات هي احدى تقنيات تعدين البيانات والتي تعنى بتلخيص كمية هائلة من البيانات واستخلاص معلومات قليلة ومفيدة. طريقة الاستقراء من الصفات تعاني من مشكلة التعميم المفرط. في هذا البحث، استخدمنا مقياس كمية المعلومات لتحسين عملية التعميم واختيار الصفات وشرط التوقف. النتائج العملية اثبتت ان الطريقة المقترحة قد قللت من مشكلة التعميم المفرط.

1.Introduction

Data mining is a kind of data analysis technique, which aims to discover hidden ,previously unknown, and potentially useful patterns from a huge amount of data for decision support[Chung-Chian Hsu,2006]. In other words, " Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [Daniel T. Larose,2005]. It is likely to put data mining activities into one of two categories: Predictive data mining which produces the model of the system described by the given data set such as classification and predicator, and Descriptive data mining, which produces new, nontrivial information based on the available data set[Mehmed Kantardzic ,2003]. Summarization, as a descriptive data mining technique[Mehmed Kantardzic ,2003], is the process by which data is compact in an intelligent and meaningful fashion to its important and relevant features[Sarat M. Kocherlakota,2005] .Summarization techniques characterized as follow Association rule mining[Sarat M. Kocherlakota,2005], Outlier Detection[Sarat M. Kocherlakota,2005], Clustering[Sarat M. Kocherlakota,2005], Aggregation[Sarat M. Kocherlakota,2005], Reducing Dimensionality[Sarat M. Kocherlakota,2005] and Attribute-Oriented Induction(AOI)[R. Angryk ,2006]. The main focusing of this paper is on enhancing AOI. AOI suffers from Overgeneralization [Jiawei Han,2006], Data types[Chung-Chian Hsu,2006], Building Concept Hierarchy[Spits Warnars H.L.H,2010], and Threshold Setting[Yu-Ying Wu,2009].However, a potential weakness of AOI is the Overgeneralization because the overgeneralization problem associated with information validity. In this paper we proposed new method to enhance generalization process , our method depend on Entropy measurement.

The organization of this paper is as follows. Section 2 surveys the related works. Section3 presents AOI algorithm, primitives, and weaknesses. The Entropy measure is stated in Section 4.Section 5 proposes an Enhancement to AOI algorithm, and the accuracy measure. The experiments and results are stated in Section 6. Finally, Section 7 concludes this paper.

2.Related Work

In 1994 Jiawei Han and Yongjian Fu are develop some algorithms for automatic generation of concept hierarchies for numerical attributes based on data distributions and for dynamic refinement of a given or generated concept hierarchy based on a learning request, the relevant set of data and database statistics[Jiawei Han,1994]. In 1997 Micheline Kamber and et al are propose two algorithm MedGen and MedGenAdjust [M. Kamber,1997]. In 2000 David W. Cheung , H.Y. Hwang , ADA W. FU and Jiawei Han are extend the concept generalization to rule-based concept hierarchy to enhances greatly its induction power[David W. Cheung,2000].
In 2002 Klaus Julisch and Marc Dacier are notice that a major source of overgeneralization is noise. They proposed three modifications on classic AOI algorithm [Klaus Julisch ,2002]. In 2006 Chung-Chian and Sheng-Hsuan Wang were noticed that the AOI may fail to disclose major values due to over generalization, they introduced a parameter, majority threshold β [Chung-Chian Hsu ,2006] . In 2009 Yen-Lian Chen and Ray-I Chary were proposed two novel generalized knowledge induction approaches Generalized Positive Knowledge Induction (GPKI) and Generalized Negative Knowledge Induction (GNKI)[Yu-Ying Wu,2009]. In 2010 Spits Warnars H.L.H proposes some improvements on classic AOI [Spits Warnars H.L.H,2010].In 2010 Devi Prasad Bhukya and S. Ramachandram are propose data classification method using AVL trees[Devi Prasad Bhukya1,2010].

3.Attribute-Oriented Induction(AOI)

Attribute-oriented induction(AOI) is a descriptive data mining technique which compresses the original set of data into a generalized relation, providing summarative and concise information about the massive set of the original data[R. Angryk ,2006]. The induction method mainly includes two steps, attribute removal and attribute generalization[Jiawei Han ,2006].

- Attribute removal is based on the following rule: If the attribute of the initial working relation contains a large set of distinct values but either (1) there is no generalization operator on the attribute (e.g., there is no concept hierarchy defined for the attribute, or (2)its higher-level concepts in terms of other attributes are expressed , then the attribute from the working relation should be removed[Jiawei Han ,2006].
- Attribute generalization is based on the following rule: If the attribute of the initial working relation contains a large set of distinct values and there exists a set of generalization operators on the attribute, then a generalization operator to the attribute should be selected and applied[Jiawei Han ,2006].

3.1Primitives in AOI

- Data relevant to the discovery process: A database typically stores a large amount of data of which only a portion may be relevant to a definite learning task. For example, to characterize the features of graduate students in science only the data relevant to graduates in science are suitable in the learning process. To collect task-relevant data from the database, a query can be used [Jiawei Han,1992].
- Background knowledge: Concept hierarchies characterize necessary background knowledge which controls the generalization process. Different levels of concepts are often organized into a categorization of concepts[Jiawei Han,1992].

The concept categorization can be partially ordered according to a general-to-specific ordering. The too high general concept is the null description (described by a reserved word "ANY") and the most specific concepts correspond to the specific values of

attributes in the database. By knowledge engineers or domain experts the concept hierarchies can be provided [Jiawei Han,1992].

- Representation of learning results: From a logical point of view, each tuple in a relation is a logic formula in conjunctive normal form and by a large set of disjunctions of such conjunctive forms a data relation is characterized[Jiawei Han,1992].

3.2 AOI Problems

AOI algorithm suffers from the following problems:

- Overgeneralization: Overgeneralization caused by the following motivation:(If the attribute is generalized excessively it may lead to overgeneralization and the resulting rules may not be very informative. On the other hand, if the attribute is not generalized to a suitably level then under generalization may result where the rules obtained may not be informative either).[Jiawei Han ,2006].
- Data Types: AOI is suffer from the problems of constructing subjectively numeric concept hierarchies and improperly generalizing boundary values near discretization points[Chung-Chian Hsu,2006].
- Concept Hierarchy: Concept hierarchy stores as a relation in the database provides essential background knowledge for data generalization and multiple level data mining [Spits Warnars H.L.H,2010].
- Threshold Setting: If the users of standard AOI are set different thresholds, they will obtain different sets of generalized tuples that also describe the major characteristics of the input relation [Yu-Ying Wu,2009].

4.ENTROPY

The entropy is a measure of uncertainty of a random variable. Let R be a discrete random variable with alphabet α and probability mass function p(r) = Pr{R = r}, r $\in$ α [Thomas M. Cover and Joy A. Thomas ,1991].

The entropy H(R) of a discrete random variable X is defined by[Thomas M. Cover and Joy A. Thomas ,1991]

$$H(R) = -\sum_{r \in \alpha} p(r) \log_2 p(r) \qquad (1)$$

The Entropy is also known as **information gain** of database which is used as attribute selection measure by classification data mining algorithms[Jiawei Han,2006].

Example 1: Let

$$\mathbf{P}=\begin{cases} a \text{ with probability } \frac{1}{2}, \\ b \text{ with probability } \frac{1}{4}, \\ d \text{ with probability } \frac{1}{8}, \\ e \text{ with probability } \frac{1}{8}. \end{cases}$$

The entropy of P is

$$H(P)= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} = \frac{7}{4}\text{bits.}$$

The graph of the function H(P) is shown in Figure 1. The figure illustrates some of the necessary properties of entropy, when p = 0 or 1 it is a concave function of the distribution and equals 0.This makes sense because when p = 0 or 1 there is no uncertainty and the variable is not random. In the same way the uncertainty is maximum when p = 1 / 2 which also corresponds to the maximum value of the entropy[Thomas M. Cover and Joy A. Thomas ,1991].

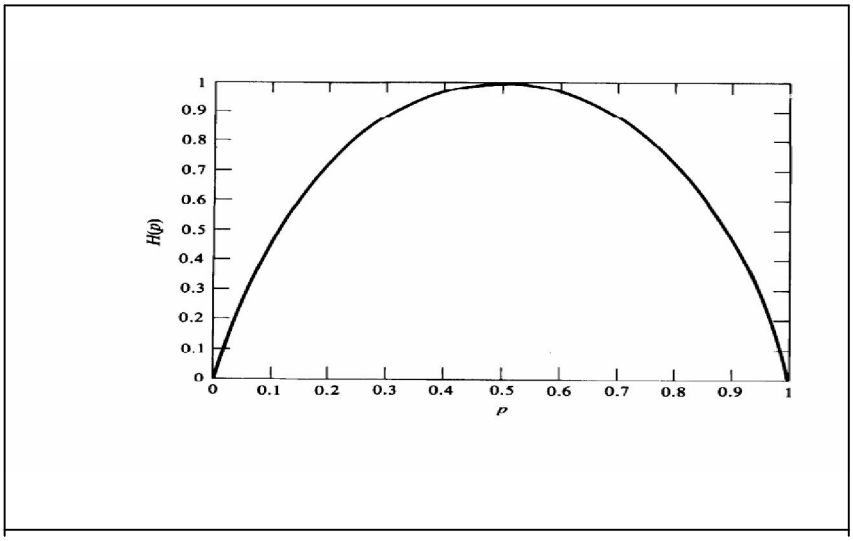

Figure 1 *H(P)* versus *p*.

5.The Proposed AOI

The Block diagram of the Proposed AOI steps is shown in Figure 2.

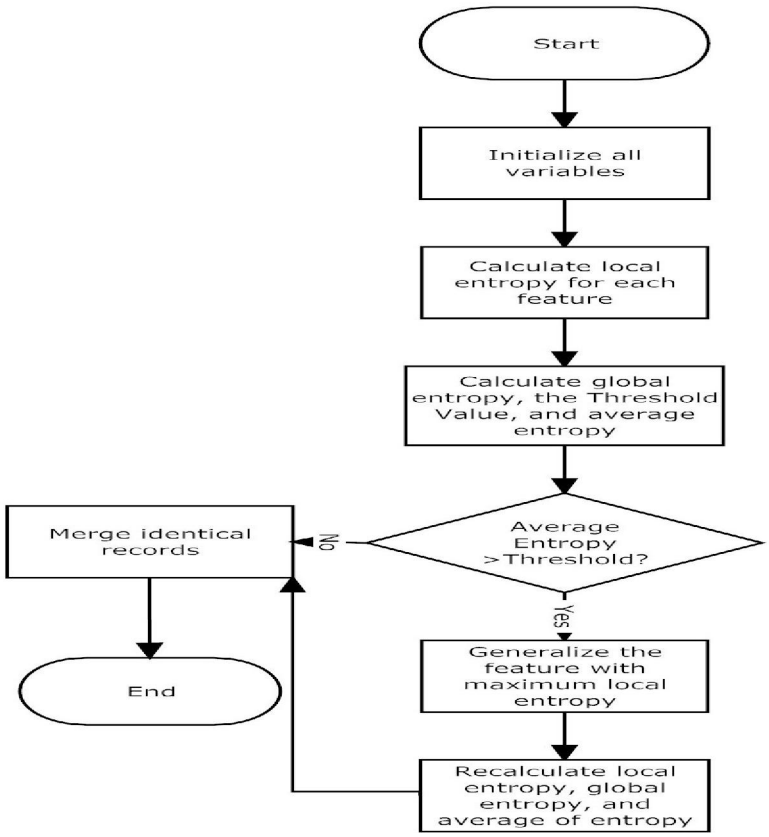

Figure 2 The Block diagram of the Proposed AOI steps.

The basic idea of *Proposed AOI* is summarized in the following algorithm as it is shown in the Figure 3.

Definition1: Local Entropy is the entropy of the individual attribute.
Definition2:

$$\text{Global Entropy} = \sum_{i=1}^{n} Local\ Entropy \qquad (2)$$

Where *n* is number of attributes in relational database.

Algorithm: Entropy Based AOI.

Input: Relational database, Generalization Hierarchies, Entropy_Threshold .

Output: Generalized relation rules.

Method:

Begin

0: *n* = # features in the relational database;
1: ***Foreach*** feature F_i ($1 \le i \le n$) in the relational database ***do***
2: Calculate Local Entropy for feature F_i;
3: Calculate Global entropy;
4: AvgEntropy= Global entropy / *n*;
5:Threshold= AvgEntropy * **Entropy_Threshold**;
6:*While* (**AvgEntropy** > **Threshold**) ***do*** **{**
7: Select *Fi* with maximum entropy, where *Fi* is not considered previously;
8: ***Foreach*** distinct value *fi* in feature *Fi* **do**
9: Generalize feature's value *fi*;
10: Recalculate the Local Entropy of *Fi* after generalization;
11: Recalculate the Global Entropy;
12: AvgEntropy =Global Entropy/*n*; **}**
13: Merge identical tuples;
End

Figure 3 The Proposed AOI

In Step 1 and Step 2, we have to compute the entropy for each attribute in relational database. The Global entropy of relational database was computed in Step 3. Step 4 computes the average entropy of relational database divided by n.

The average entropy of relational database multiplied by Entropy_Threshold was computed in Step 5.The stop condition is represented in Step 6.Step7, was designed to select the attribute with maximum entropy and not considered previously for the next generalization process.

Step 8 and Step 9 represents generalization process(*Attribute generalization)*. In Step 10 a new entropy value for the attribute in process was computed. Step 11 a new Global entropy of relational database was computed. Step 12 estimates new relational database average entropy value. Finally, Step 13 merge identical records into a single one whose count value equals the sum of the constituent counts.

5.1Coverage And Accuracy Measures

A rule *R* was assessed by its coverage and accuracy, given a tuple , ***X***, from a class labeled data set *D* let ⎪D⎪ be the number of tuples in *D n covers* be the number of tuples covered by *R n correct* be the number of tuples correctly classified by *R*. The coverage and accuracy of *R* was defined as[Jiawei Han ,2006]:

$$Coverage(R) = \frac{n\,\mathrm{cov}\,ers}{|D|} \quad \textbf{(3)}$$

$$Accuracy(R) = \frac{ncorrect}{n\,\mathrm{cov}\,ers} \quad \textbf{(4)}$$

6.Experiments and Results

We address the efficiency of generalization process regarding standard AOI by proposing a technique composed of the following improvments:1) Feature selection for generalization process depends on feature entropy measurement, 2) Run time for the proposed algorithm is less than the run time of standard AOI 3)The proposed algorithm needs only one threshold number but standard AOI needs two threshold numbers. The proposed algorithm is applied on employees data set of Babylon university which is sql server type with 7 features(attributes) which is place of birth, employment degree, current status, certificate, title, faculty membership, martial state and using the class attribute as class label this features were selected because to characterize the general properties of Babylon university employees, and 4330 records.

It is worth mentioning that the number of generalization hierarchy trees used is 7 and all are building manually which is generalization hierarchy tree for place of birth with maximum height is 4 and generalization hierarchy trees for employment degree, current status, certificate, title, faculty membership and martial state with maximum height is 2.

A generalization hierarchy trees for *current status* and *faculty membership* are shown in Figure 4.

Part a of Figure 4 we shows a concept tree for *current status*, while in part b we summarize the concept tree for *faculty membership*.

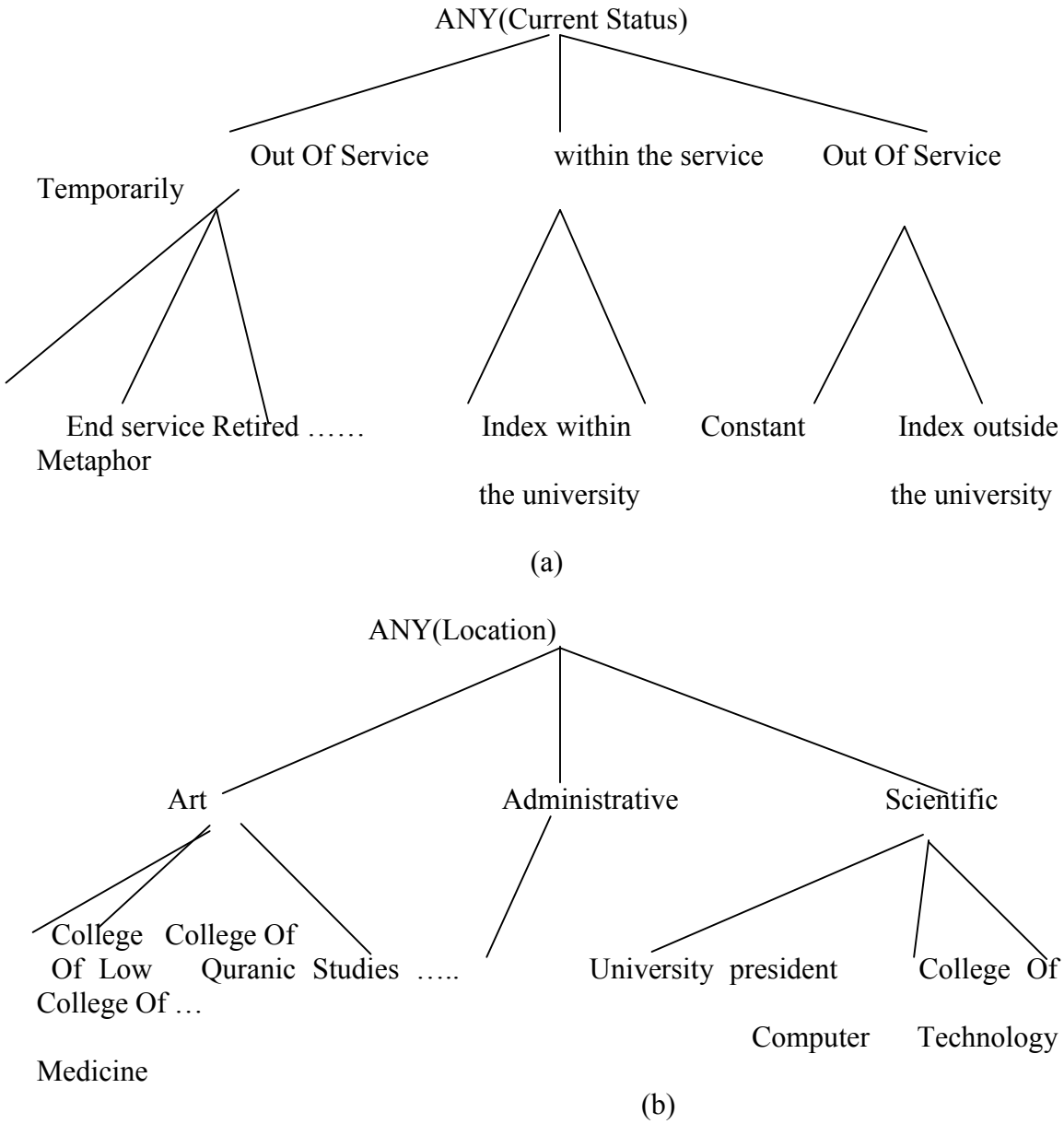

Figure 4 A concept tree for *Current Status* and *Location*.

Measures that are used in testing are, as follows:

- *Threshold* represents a threshold numbers that control on a generalization process.
- *#Records* is the number of resulting records at the end of generalization process. Whenever this value is low means that summarize a large amount of data, this value is affected by a threshold setting process.
- *Average of Entropy* is the average of entropy for the relational data base at the end of generalization process. The lower the entropy, the more concentrated the instantiations and more structured in the database are. The value of Average of Entropy is affected by a threshold setting process.

- *Average of Precision* represents the accuracy of the resulting rule at the end of generalization process. Whenever such a value is 1 means no overlapping of data for the instructor from non. The value of Average of Precision is affected by a threshold setting process.
- *Average of Coverage* represents the overall rate for a number tubles that coverage by each resulting rule at end of generalization process. Whenever this value is high , it means that each rule will cover a large number of the tuble at the end of generalization process. The value of *Average of Coverage* is affected by a threshold setting process.
- *Time* is a run time, where a run time is affected by a threshold setting process, the value of average of entropy for initial relational database , this means larger average of entropy value more run time required and number of records for initial relational database , where it increases when the number of records increased .

These measures have been applied to evaluate the results that obtained by the standard AOI and the Proposed AOI.

Table1 and Table2 are represent the results of Standard AOI and Proposed AOI

Table 1 Results Of Standard AOI With Different Threshold.

No.	Threshold	#Records	Avg. Of Entropy	Avg. Of Precision	Mean Of Coverage	Time
1	٢٠٠٠&&٢٧	١٢٠٦	١.٨٤٤	٠.٩٩٥	٠.٠٠٠٨٣	٢:٢١:٣١٢
2	٢٢٥٠&&٢٧	٢١٨٦	٢.٠٨٥	٠.٩٩٥	٠.٠٠٠٤٦	١:٥٩:٥٠٠
3	٢٧&& ٢٥٠٠	٢١٨٨	٢.٠٨٩	٠.٩٩٥	٠.٠٠٠٤٦	١:٥٩:٧٣٤
4	٢٧٥٠&&٢٧	٢١٨٦	٢.٠٨٥	٠.٩٩٥	٠.٠٠٠٤٦	١:٥٩:٤٠٧
5	٣٠٠٠&&٢٧	٢٨٥٩	٢.٥٤٨	١	٠.٠٠٠٣٥	١:١٣:٣١
٦	٢٠٠٠&&٢٦	١٢٠٨	١.٨٥٠	٠.٩٩٦	٠.٠٠٠٨٣	١:١٣:٥٦٢
٧	٢٢٥٠&&٢٦	٢١٨٨	٢.٠٨٩	٠.٩٩٧	٠.٠٠٠٤٦	٥١:٨٧٥
٨	٢٦&& ٢٥٠٠	٢١٨٨	٢.٠٨٩	٠.٩٩٧	٠.٠٠٠٤٦	٥١:٨٧٥
٩	٢٧٥٠&&٢٦	٢١٨٨	٢.٠٨٩	٠.٩٩٧	٠.٠٠٠٤٦	٥١:٦١٠
١٠	٣٠٠٠&&٢٦	٢١٨٦	٢.٠٨٥	٠.٩٩٧	٠.٠٠٠٤٦	٥١:٧٩٧

respectively with different threshold values.

Table 2 Results Of Proposed AOI With Different Threshold Values.

No.	Threshold (%)	#Records	Avg. of Entropy	Avg. of Precision	Mean of Coverage	Time (sec)
1	٨٠	٢١٩٢	٢.٤٣٠	١	٠.٠٠٠٤٦	٤٥
2	٧٠	١٥٢٧	٢.١٥٣	١	٠.٠٠٠٦٥	٣١
3	٦٠	٧٦٠	١.٨٠٢	١	٠.٠٠١٣٢	١٨
4	٥٠	٣٠٨	١.٣٦٨	٠.٩٩٦	٠.٠٠٣٢٥	٨
5	٤٠	٣٠٨	١.٣٦٨	٠.٩٩٦	٠.٠٠٣٢٥	٨
٦	٣٥	١٦٢	١.١٤٨	٠.٩٩٨	٠.٠٠٦١٧	٦
7	٣٠	١٦٢	١.١٤٨	٠.٩٩٨	٠.٠٠٦١٧	٦
٨	٢٥	٧٣	٠.٩٥٨	٠.٨٤٤	٠.٠١٣٧٠	٥
٩	٢٠	٧٣	٠.٩٥٨	٠.٨٤٤	٠.٠١٣٧٠	٤
١٠	١٥	٣٦	٠.٨١٩	٠.٨٣٣	٠.٠٢٧٧٨	٤

Our experiments aims to compare the results by using the standard AOI with proposed AOI on employees data set of Babylon university with optimal results by them in terms Average of ANY , Average Of Accuracy and Run Time as it shown in the Table 3.

Table 3 Comparison of standard AOI and proposed AOI.

Criteria	Proposed AOI	Standard AOI
# of Records	١٥٢٧	٢٨٥٩
# of Threshold	1	2
Average Of Entropy	٢.١٥٣٤٠	٢.٥٤٨٣١
Average Of ANY	0	0
Average Of Coverage	٠.٠٠٠٦٥	٠.٠٠٠٣٤
Average Of Accuracy	1	1
Run Time	٠:٣٠:٨٩٥	١:١٣:٣١

Run time for ***Standard AOI*** and ***proposed AOI*** with different data size (number of tuples) from 500 to 4330 are shown in Figure 5 .The run time of standard AOI depend on threshold values(*first threshold that control on each attribute and second threshold that control on generalized relations*) , this means minimum threshold values more run time required. On the other hand, run time of proposed AOI depend on entropy value of tuples ,this means larger entropy value more run time required. And the gap of run time show that.

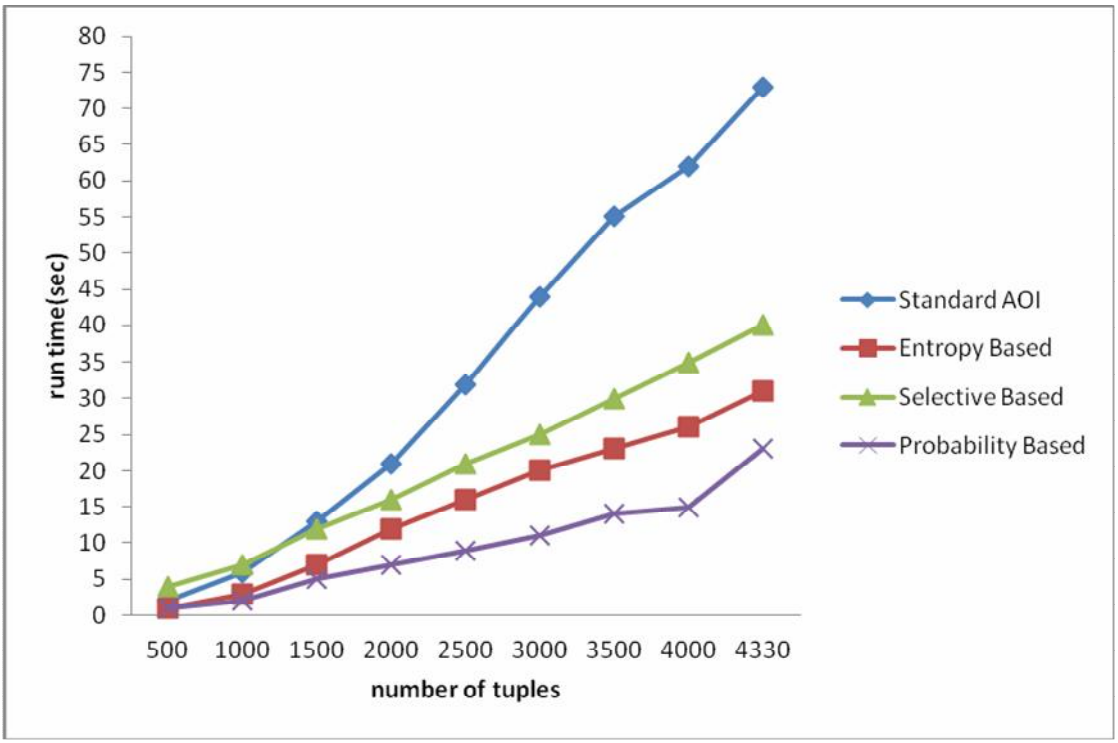

Figure 5 Run time of standard AOI and The proposed AOI with different data size.

7. Conclusion

We have presented an interesting and efficient rule induction strategy that ensures generation rules that characterize the general properties of Babylon university employees with high coverage and accuracy.

Control for maximum number of tuples of the target class in the final generalized relation in current attribute oriented induction is limited by two threshold number but in entropy based attribute induction is limited by one threshold number.

Different concept hierarchy trees are give different interesting rules in the final generalization relation.

References

Angryk R. and F. Petry, 2006,"Discovery Of Abstract Knowledge From Non-Atomic Attribute Values In Fuzzy Relational Databases ",in: B. Bouchon-Meunier, G. Coletti, R. Yager(Eds.),Modern Information Processing, From Theory to Applications, Elsevier.

Chung-Chian Hsu and Sheng-Hsuan Wang,2006," An Integrated Framework For Visualized And Exploratory Pattern Discovery Mixed Data ", IEEE Transaction On Knowledge And Data Engineering, VOL. 18, NO. 2, pp.161-173.

Daniel T. Larose , 2005," Discovering Knowledge In Data: An Introduction To Data Mining", John Wiley & Sons, Inc., pp.1-222.

David W. Cheung and H. Y. Hwang and Ada W. Fu and Jiawei Han, 2000," Efficient Rule-Based Attribute-Oriented Induction For Data Mining", Journal Of Intelligent Information Systems,VOL.15,NO. 2,pp.175-200.

Devi Prasad Bhukya and S. Ramachandram, 2010," Decision Tree Induction: An Approach For Data Classification Using AVL-Tree", International Journal Of Computer And Electrical Engineering, Vol. 2, No. 4, pp.660-665.

Interscience ,New York, pp.1–10.

Jiawei Han and M. Kamber, 2006,Data Mining: Concepts And Techniques, second edition ,Morgan Kaufmann, New York, pp.1-745.

Jiawei Han and Yongjian Fu, 1994," Dynamic Generation And Refinement Of Concept Hierarchies For Knowledge Discovery In Databases", In Proceeding Workshop On Knowledge Discovery In Databases, pp. 157-168.

Jiawei Han, Y. Cai and N. Cercone, 1992," Knowledge Discovery in Databases: An Attribute-Oriented Approach", In Proceeding 18th Int'l Conference on Very Large Data Bases, Vacouver, Canada, pp. 547-559.

Kamber, M. L. Winstone, W. Gong, S. Cheng, and Jiawei Han, 1997,"Generalization And Decision Tree Induction: Efficient Classification In Data Mining", In Proceeding International Workshop Research Issues on Data Engineering , Birmingham, England, pp.1-10.

Klaus Julisch, and Marc Dacier, 2002," Mining Intrusion Detection Alarms For Actionable Knowledge ",In The Proceeding Of The 8th ACM International Conference on Knowledge Discovery and Data Mining, pp. 366-375.

Mehmed Kantardzic , 2003,"Data Mining: Concepts, Models, Methods, And Algorithms ",Wiley-Blackwell,pp.1-343.

Sarat M. Kocherlakota, and Christopher G. Healey, 2005, "Summarization Techniques For Visualization Of Large Multidimensional Datasets", Technical Report ,Knowledge Discovery Lab, Department of Computer Science, North Carolina State University, pp. 1-18.

Spits Warnars H.L.H, 2010," Attribute oriented induction with star schema", International Journal of Database Management Systems,Vol.2,No.2, pp. 20-42.

Thomas M. Cover, and Joy A. Thomas , 1991," Elements of Information Theory", Wiley-

Yu-Ying Wu, Yen-Liang Chen ,and Ray-I Chang, 2009, " Generalized Knowledge Discovery From Relational Databases", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.6, pp. 148-153.