# Interesting Metrics Based Adaptive Prediction Technique for Knowledge Discovery

G. Anbukkarasy[1], N. Sairam[2]
[1 & 2] School of Computing, SASTRA University,
Thirumalaisamudram, Thanjavur, Tamilnadu, India.
[1]anbukkarasy@gmail.com

**Abstract - Prediction is considered as an important factor to predict the future results from the existing information. Decision tree methodology is widely used for predicting the results. But this is not efficient to handle the large, heterogeneous or multi-featured type of data sources. So an adaptive prediction method is proposed by combining the statistical analysis approach of the data mining methods along with the decision tree prediction methodology. So when dealing with large and multi-server based data, the prediction can be done more effectively and less time consuming than the existing prediction methods. For the statistical analysis, incremental clustering and multiple interesting measures are used for a better prediction.**

**Keywords – Prediction, Knowledge Discovery, Interesting metrics, Intelligent Decision Support System (IDSS).**

## I. INTRODUCTION

Prediction can be called as an "Intelligent Decision Support System (IDSS)"; since prediction involves the discovery of knowledge [5] from the data warehouse by combining factors such as judgement and preferences. This helps to predict the exact future from the existing patterns and relationships. The most commonly data mining techniques used for prediction are Decision tree [15], Artificial neural network, Fuzzy clustering, Rough set theory and Support vector machine.

The entire prediction process is carried out in an automated manner as shown in Fig. 1. The existing patterns are taken as the input, relevant data mining techniques are implemented and the result is the final observations. Based on the observations, the conditions are derived to predict the accurate result. When a new data is entered, the data can be easily predicted by the processed result.

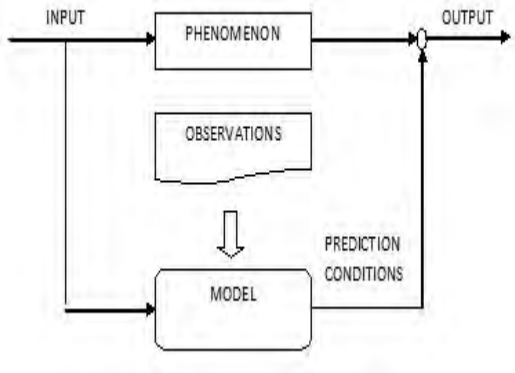



Fig. 1. Prediction modelling

Prediction can also be considered as a "Black-box like software model"; because this takes the input from the historical information (existing patterns) and form the rules with manual interpretation. The existing rules may come from the same data source pattern or from a different data source (pattern).

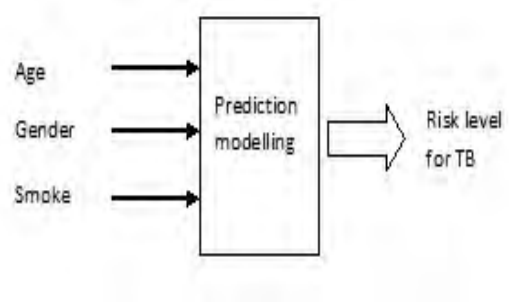



Fig. 2. Example prediction model

In Fig. 2, An example model is developed by taking the input values as the person's age, gender and whether the person is a smoker or not. By analysing these inputs with the already existing models, the new input value is calculated and the risk level of the person can be predicted.

### *A. INTERESTING METRICS:*

Interesting metrics are used to evaluate the discovered patterns [12], which are derived from the data mining methods. It also helps to validate the data; that is, to measure the quality of the pattern/relationships from the entire dataset. This concept is presented in Fig. 3. This avoids the processing of unnecessary relationships.

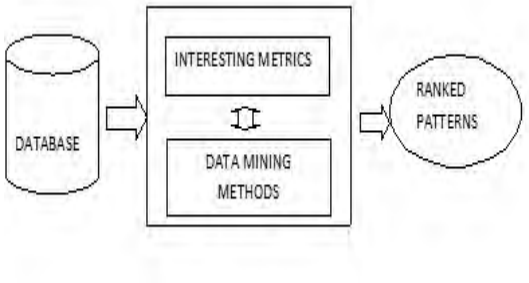



Fig. 3. Pattern formation by interestingness metrics

Interestingness is possibly considered as a perception which clearly defines [6] the action-ability, conciseness, coverage, diversity, novelty, peculiarity, reliability, surprising, and utility. These nine different perceptions [11] can be simply view as 3 factors:

1. ***Objective:***

   These types of metrics are for the un-processed data [4]; i.e., No knowledge about the data is needed. Statistical and probability type of calculations; diversity, conciseness, generality, reliability and peculiarity are the different views about this metric.

2. ***Subjective:***

   Along with the data, information about the client's personal details and goals are also considered for the subjective measure. By interpreting the client in an explicit manner, client information can be identified. Novel and surprising are its perceptions.

3. ***Semantic:***

   Information about the data and its description altogether [11] form the semantic measure. Usefulness and action-ability are considered as the different semantic perceptions.

## II. LITERATURE SURVEY

The frequently occurring data items are grouped together and called as the association; the result of which forms the association rules/relationships. A threshold value is set to group these interesting rules and provide the information as a result for the future use. Authors discussed about the two different aspects in the interesting metrics [6] like its subjective and objective characteristics; where the objective works on the rules and the subjective deals with the client's expectations. Since the objective considers the data alone, it should include and analyze the data characteristics like exposure, concise (brief summary), exactness, assurance, consistency, strangeness, diversity, comprehensiveness (complete). Innovations, unforeseen nature, usefulness, effectiveness and the result should suit to all type of clients are needed for the subjective metric; since this considers both data and client's belief.

To mine the patterns by adding the modules/features which are informative in nature [3] from either numerous datasets or several characteristics or various methodologies on demand is proposed as a combined

mining approach. Original combined patterns/relationships are not possible to generate by only having the previous methodologies; whereas this combined mining approach helps to form the Incremental pair patterns or the incremental cluster patterns directly. This combined mining approach overcomes the existing disadvantages in the sampling process, combination of various relevant tables, examining before the mined pattern formation, involving various methodologies and in addition to mine the numerous data sets.

Authors discussed about the process of knowledge discovery [12] and the need for filtering out the results. The filtering can be done by assigning rank to each and every derived patterns based on the interesting metric calculation. Support, lift and confidence are taken as the interesting metrics to the calculation. The concept of decision tree algorithm is considered here to discover the knowledge from the data. Interesting metrics is considered as objective and subjective metrics and some of the decision rule examples are also discussed in this paper.

Three different metrics such as bond, any-confidence and all-confidence metrics [14]; as an alternative to the already existing simple support measure are proposed by the author. In a dataset, there may be many confidence value related to the different rules. For any-confidence, if the pattern falls greater than any of the single value means, it will be accepted as interesting. The rules are categorized by considering the minimum confidence value which is set based on all the confidence values for the all-confidence metric. Bond refers to check the values by taking the confidence from a sub dataset. This process is more effective than the previous types of metrics.

To analyze and identify the behaviour or nature of the data, authors proposed machine learning method [15] which is run by learning the behaviour patterns of the dataset. This method considers both the dynamic and static type of information and the faults can be easily identified by the alarm event concept. By calculating the differential equation, the accuracy of predicted value can be measured. This method is proposed for the network connection and especially for the internet server correlation.

Interesting metrics suits well for any type of data to be mined. These metrics calculates the interesting value depends on the rank values. Here the interesting factor processing is explained in a detailed way with its filter and rank options. It also describes the definition for all the characteristics [11] involved to compute the interest value like its usefulness, effectiveness, accurateness, etc. The survey includes summarize about different metrics and its description. This work gives an overall view about the interesting metrics used in the data mining.

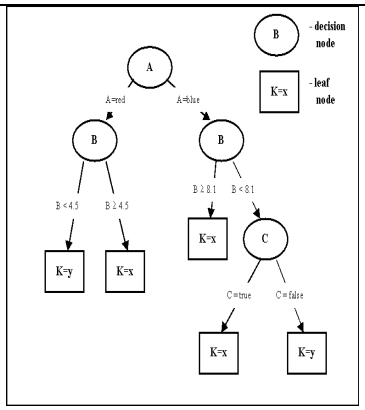
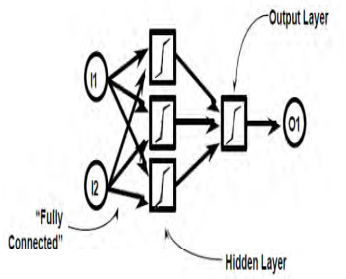
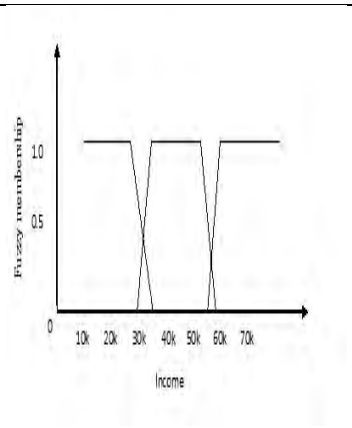
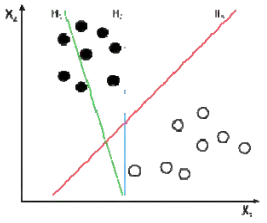
The improvement for the decision making process is considered as a major issue as a major issue by the author. Prediction can be called as an intelligent process [5]; because the result of prediction leads to the innovation of knowledge. Prediction techniques like artificial neural networks, Bayesian network, fuzzy clustering and decision tree are discussed by the authors and its characteristics are also analyzed. An idea about the integration of different techniques is also discussed.

A performance-based prediction [2] is proposed by the author for the students information by using the classification method. This method is used to predict the difference between the levels of students. Bayesian classification algorithm is used for predicting the data. By using this algorithm, the students are classified based on attributes and measures and thus it helps for the students and staffs to improve in that particular aspect.

Authors analysed the medical database with different prediction technologies [1] like artificial neural network and decision tree mechanism and compared the results of both. Feed-Forward neural networks are used to train the data for the prediction and the network models used here are the Radial Basic Function (RBF) and Multi-Layer Perceptrons (MLP). For the decision tree mechanism Reduced-Error pruning (REP-tree) and Loitboost Alternating Decision (LAD) tree are used for the classifications and the results are analyzed.

## III. METHODOLOGIES

*A. **IDSS TECHNOLOGIES [1], [2], [5]:***

| S.No | Techniques | Structure | Properties |
|---|---|---|---|
| 1. | Decision tree |  | • Used for classification and prediction.<br>• Comparatively fast.<br>• Can be easily converted to SQL queries.<br>• Manual interpretation is possible.<br>• In-expensive.<br>• Self-Explainable in nature.<br>• Eg: Accident Frequency [13]. |
| 2. | Artificial Neural Networks |  | • Inputs are transformed by means of the processors.<br>• Used for modelling [19] and classification.<br>• Difficult to interpret the results.<br>• Training requires a lot of time.<br>• Eg: Disease [8]. |
| 3. | Fuzzy clustering |  | • Rule-based system for classifications.<br>• Also called Possibility theory.<br>• Each cluster is considered as "Fuzzy set".<br>• Since the sum of values not needed to be 1 and the data can be present in more than I fuzzy set, this is efficient than traditional methods.<br>• Eg: Newspaper demand [9]. |
| 4. | Support vector machine | $H_1$ does not separate the classes. $H_2$ does, but only with a small margin. $H_3$ separates them with the maximum margin.<br>  | • Classification and prediction method for both linear and non-linear data.<br>• More time consuming.<br>• Highly accurate results.<br>• Less prone to over-fitting than other techniques. |
| 5. | Rough set theory | If (condition)<br>Then<br>Result; | • Used for classification.<br>• Discrete-valued attributes can only be analyzed to discover the structural relationships.<br>• Done by using If-Then rules.<br>• Rough set is calculated by fixing upper |

| | | | |
|---|---|---|---|
| | | | and lower approximations, instead of searching entire dataset.<br>• Feature reduction.<br>• Relevance analysis.<br>• Eg: Personnel selection [10]. |

Table I. IDSS technologies

### *B. INTERESTINGNESS MEASURES:*

Apart from the already existing lift, confidence and support [14], [16]; contribution and interesting rule [7] also analysed and computed for better pattern formation.

#### *1. Support:*

Taking the probability for the condition (Ă => Ę) is calculating the percentage of the overall data that includes both Ă and Ę.

$$\text{Support, §}(\text{Ă, Ę}) \Rightarrow P(\text{Ă U Ę}) \Rightarrow \text{Count}(\text{Ă U Ę})$$

#### *2. Confidence:*

Confidence is considered as a "conditional probability", where the data containing Ă should also include Ę.

$$\text{Confidence, Ĉ (Ă, Ę)} \Rightarrow P(\text{Ę}|\text{Ă}) \Rightarrow \frac{Support(\text{Ă},\text{Ę})}{Support(\text{Ă})} \Rightarrow \frac{Support\_Count(\text{Ă},\text{Ę})}{Support\_Count(\text{Ă})}$$

#### *3. Lift:*

Lift is considered as a correlation metric and it can be calculated as follows:

$$\text{Lift, £}(\text{Ă , Ę}) \Rightarrow \frac{P(\text{Ă U Ę})}{P(\text{Ă})\,P(\text{Ę})} \Rightarrow \frac{P(\text{Ę}|\text{Ă})}{P(\text{Ę})} \Rightarrow \frac{Confidence(\text{Ă},\text{Ę})}{Support\ (\text{Ę})}$$

#### *4. Contribution:*

By taking the lift value as a pre-condition, contribution [7] is calculated as follows: which defines how much the additional value contributes to the relationship.

$$\text{Contribution, ¥ (Ă , Ę)} \Rightarrow \frac{Lift(\text{Ă},\text{Ę})}{Lift(\text{Ă})} \Rightarrow \frac{Conf(\text{Ă},\text{Ę})}{Conf(\text{Ă})}$$

#### *5. Interesting rule:*

Contribution is considered as a pre-condition for the calculation of interesting rule, as follows: Interesting rule,

$$(\text{Ă , Ę}) \Rightarrow \frac{Cont(\text{Ă},\text{Ę})}{Lift(\text{Ę})}$$

## IV. PROPOSED METHOD

The proposed "Adaptive prediction" extracts the results for the discovery of knowledge from the clustering and the interesting metric outcomes. Since searching the entire database for a single data (which needs to be predicted) is more time consuming, clustering [2] is needed to retrieve the data from a large dataset is preferred. Then the attributes are needed to be selected for the calculation of interesting metrics. The filtering process is carried out then, to proceed with the prediction technology by decision tree mechanism.

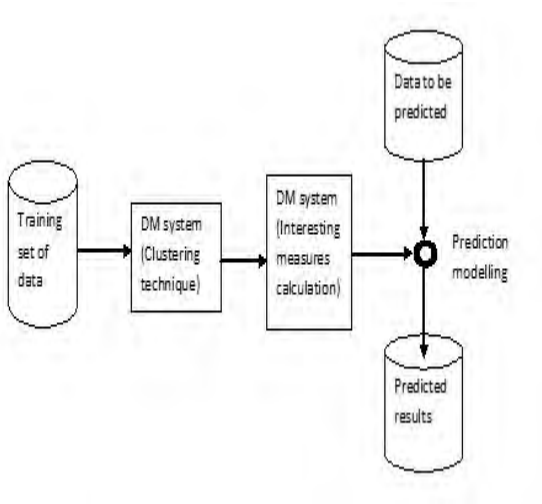



Fig. 4. Adaptive prediction using interesting metric filtering.

When a new data arrived, the predicted results help to analyze the entered data. The results are shown in the section 5 and the architecture for the adaptive prediction is explained in Fig. 4.

## V. EXPERIMENTAL RESULTS

Based on the bank database, the personal details about the employees are analysed to predict the results. First we have to choose the emp_id of the bank debtors and employee's personal information are tabulated. Then we need to proceed with the formation of clusters from the employee data for the salary attribute. This may be simply calculated by a query [1] or it can even be done by a clustering algorithm. But clustering algorithm is preferred for automation.

```
If (Salary < 50000)
        Then a = sal_low;
                Else if (500001 < Salary >100000)
         Then a = sal_medium;
 Else
        a  = sal_high;
```

By this way, the salary attribute can form the clusters. Then the interesting metrics are calculated.

| Rules | Support | Confidence | Lift | Contribution | Interesting rule |
|---|---|---|---|---|---|
| F ^ sal_high | 2/10 | 1/6 | 0.7 | 0.5 | 0.2 |
| M ^ sal_high | 2/10 | 2/4 | 1 | 1 | 1 |
| F ^ sal_medium | 3/10 | 3/6 | 1.3 | 1.3 | 1.1 |
| M ^ sal_medium | 2/10 | 2/4 | 1.2 | 1.2 | 1.3 |
| F ^ sal_low | 1/10 | 1/6 | 0.5 | 0.9 | 0.5 |

Table II. Interesting metrics calculation

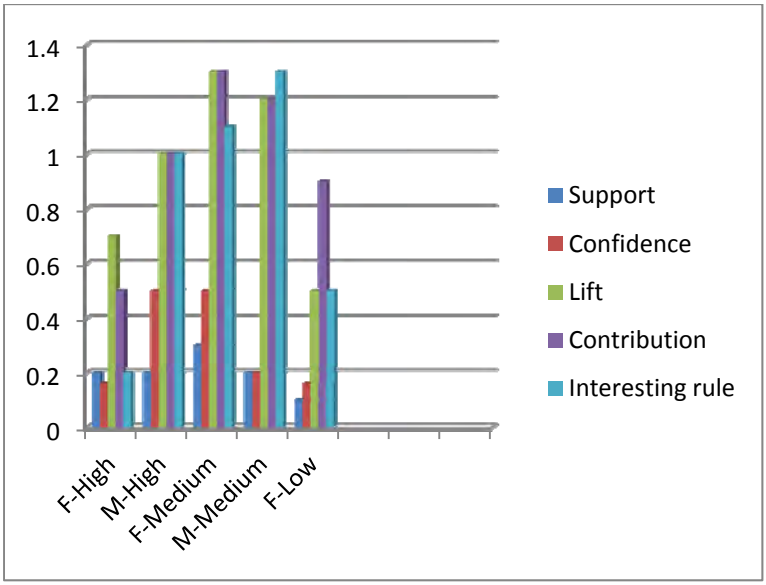



Fig. 5.Analysis of interesting metrics

From the results shown in Fig. 5, instead of taking support and confidence, the lift, contribution and interesting rule are considered to improve the quality of the cluster. The values above 1 are considered to be interesting and the value below 1 are non-interesting in nature.

So the interesting rules resulting from the calculation are as follows:

Female ^ sal_medium

Male ^ sal_high

Male ^ sal_medium

Finally the prediction methodology [19] is carried out based on these extracted rules. The rules can also be grouped/clustered together by the clustering results and as well as from the selected relationships. The result of this prediction is as follows:

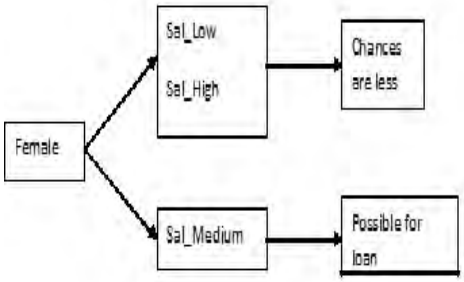



Fig. 6. Prediction for female employees by decision tree.

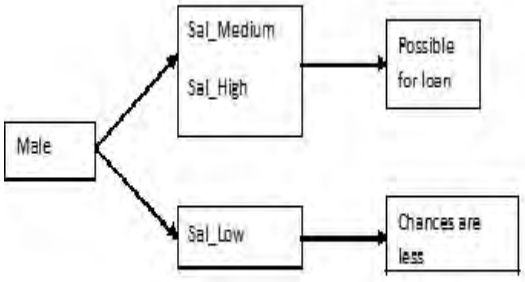



Fig. 7. Prediction for male employees by decision tree.

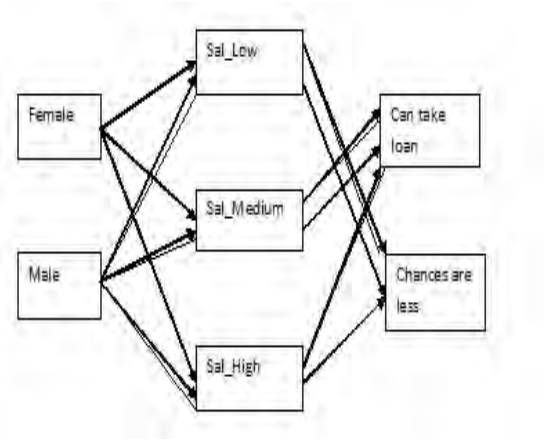



Fig. 8. Prediction for both by Artificial neural networks.

The traditional decision tree algorithm searches for every attribute in the dataset and also grouping is not available in that. By this proposed method, the attributes are clustered before the prediction as well as clustering the results after the prediction. So this provides a much better result than the conventional methods.

## VI. CONCLUSION

Prediction technique for a large and heterogeneous datasets is not much efficient in reality; because of its constraints like multi-featured data and also having a lot of attributes. This makes the traditional prediction methodology complex. By adding the clustering methods and the interesting metric (filtering) and finally grouping the results makes the user achieves a more efficient and effective results. This process is also less time consuming in nature.

## REFERENCES

[1] Olaiya Folorunsho, "Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3 (3), Mar 2013.
[2] Brijesh Kumar Bhardwaj and Saurabh Pal, "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9 (4), April 2011.
[3] Chengqi Zhang, Dan Luo, Huaifeng Zhang, Longbing Cao and Yanchang Zhao, "Combined Mining: Discovering Informative knowledge in complex data", IEEE Transactions on systems, Man and Cybernetics, Vol. 41 (3), pp. 699-712, June 2011.
[4] Alaa Al Deen, Mustafa Nofal and Sulieman Bani-Ahmad, "Classifictaion based on Association rule mining techniques: A general survey and Empirical Comparitive Evaluation", Ubiquitous Computing and Communication Journal, Vol. 5 (3), pp. 9-17, 2011.
[5] Hamidah Jantan, Abdul Razak Hamdan and Zulaiha Ali Othman, "Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application", International Journal of Human and Social Sciences, Vol. 5 (11), pp. 694-702, 2010.
[6] Yuejin Zhang, Guangli Nie and Yong Shi, "A Survey of Interesting measures for Association Rules", International Conference on Business Intelligence and Financial Engineering, BIFE, pp. 460 -463, 2009.
[7] Huaifeng Zhang, Yanchang Zhao, Longbin Cao and Chengqi Zhang, "Combined Association Rule Mining", Pacific-Asia conference on Knowledge Discovery and Data Mining, LNAI 5012, 1069-1074, 2008.
[8] P.L. Liew, Y.C. Lee, Y. C. Lin, T.S. Lin, W.J. Lin, W. Wang and C. W. Chien, "Comparison of artificial neural networks with logistic regression in predicition of Gallbladder disease among obese patients", Digestive and Liver Disease, 39 (4), pp. 356-362, 2007.
[9] Cardoso, G. and F. Gomide, "Newspaper demand prediction and replacement model based on fuzzy clustering and rules", An International Journal on Information Sciences, Vol. **177** (21): pp. 4799-4809, 2007.
[10] Chien, C.F. and L.F. Chen, "Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manufacturing", IEEE Transactions on Semiconductor Manufacturing, Vol. **20** (4): pp.528-541, 2007.
[11] Liqiang geng and Howard J. Hamilton, "Interesting measures for Data mining: A survey", ACM Computing Surveys, Vol. 38 (3), pp. 1-32, Sep 2006.
[12] Ken McGarry, "A survey of interestingness measures for knowledge discovery", The knowledge Engineering Review, pp. 1-24, 2005.
[13] Chang, L.Y. and W.C. Chen, "Data mining of tree-based models to analyze freeway accident frequency", Journal of Safety Research, Vol. **36** (4), pp. 365-375, 2005.
[14] Edward R. Omiecinski, "Alternative Interest Measures for mining Associations in Databases", IEEE transactions on Knowledge and data engineering, Vol. 15 (1), pp. 57-69, Jan/Feb 2003.
[15] Marlon Nunez, Rafeal Morales and Francisco Triguero, "Automatic discovery for predicting network management events", IEEE Journal of selected areas in communication, Vol. 20 (4), pp. 736- 745, May 2002.
[16] P. D. McNicholas, T.B. Murphy and M. O' Regan, "Standardising the lift of an association rule", School of computer science and statistics, pp. 1-20, July 2001.
[17] Haddawy, P. and N.T.N. Hien "A decision support system for evaluating international student applications", http://www.apqn.org/event/past/details/102/presentation/files/6_prof_peter_haddaway_and%20-hyuyen_thi_ngoc_hien.pdf 9/1/2008.
[18] Pardos, Z., et al. The effect of Model Granularity on Student Performance Prediction using Bayesian Networks, http://www.educationaldatamining.org/um2007/Pardos.pdf.
[19] An introduction to Data mining, Kurt Thearling, http://www.thearling.com/dmintro/dmintro.pdf