# On the Evaluation of Attribute Information for Mining Classification Rules

Ming-Syan Chen

Electrical Engineering Department
National Taiwan University
Taipei, Taiwan, ROC
email: mschen@cc.ee.ntu.edu.tw

## Abstract

In this paper we deal with the evaluation of attribute information for mining classification rules. In a decision tree, each internal node corresponds to a decision on an attribute and each outgoing branch corresponds to a possible value of this attribute. The ordering of attributes in the levels of a decision tree will affect the efficiency of the classification process, and should be determined in accordance with the relevance of these attributes to the target class. We consider in this paper two different measurements for the relevance of attributes to the target class, i.e., *inference power* and *information gain*. These two measurements, though both being related to the relevance to the group identity, can in fact lead to different branching decisions. It is noted that, depending on the stage of tree branching, these two measurements should be judiciously employed so as to maximize the effects they are designed for. The inference power and the information gain of multiple attributes are also evaluated in this paper.

## 1 Introduction

Mining information and knowledge from large databases has been recognized by many researchers as a key research topic in database systems and machine learning, and by many industrial companies as an important area with an opportunity of major revenues. It is noted that analysis of past transaction data can provide very valuable information on customer buying behavior, and thus improve the quality of business decisions. It is essential to collect a sufficient amount of sales data before any meaningful conclusion can be drawn therefrom. The huge amount of processed data accounts for the importance of devising efficient algorithms for data mining.

Various data mining capabilities have been explored in the literature. Mining association rules has attracted a significant amount of research attention [3, 10, 12, 16, 17]. For example, given a database of sales transactions, it is desirable to discover all associations among items such that the presence of some items in a transaction will imply the presence of other items in the same transaction. Another type of data mining is on ordered data, which includes searching for similar sequences [1, 18], e.g., stocks with similar movement in stock prices, and for sequential patterns [4], e.g., grocery items bought over a set of visits in sequence. Mining on Web path traversal patterns was studied in [7]. In addition, one important application of data mining is the ability to perform classification in a huge amount of data. This is referred to as mining classification rules. Mining classification rules is an approach of trying to develop rules to group data tuples together based on certain common features. For an example of commercial applications, it is desirable for a car dealer to know what are the common features of its most customers so that its sales persons will know whom to approach, and its catalogs of new models can be mailed directly to those customers with identified features. The business opportunity can thus be maximized.

Classification rule mining has been explored both in the AI domain [13, 15] and in the context of databases [2, 5, 8, 9]. In machine learning, a decision-tree classification method, developed by Quinlan [14, 15], is one of the most important results, and has been very influential to later studies. It is a supervised learning method that constructs decision trees from a set of examples. The quality of a tree depends on both the classification accuracy and the size of the tree. Other

approaches on data classification include statistical approaches [13], rough sets approach [19], etc. In the context of databases, an interval classifier has been proposed in [2] to reduce the cost of decision tree generation. An attribute-oriented induction method has been developed for mining classification rules in relational databases [9]. The work in [11] explores rule extraction in a database based on neural networks. A two-phase approach for classification rule mining was explored in [6], where given a large training set of data tuples, the first phase, referred to as feature extraction phase, is applied to a subset of the training database with the purpose of identifying attributes which has high relevance to the target class. In the second phase, referred to as feature combination phase, these extracted attributes are evaluated together for a possible combined use.

In mining classification rules for a given database, one would naturally like to have a training dataset large enough so as to have a sufficient confidence on the rules derived. However, with a large training set, the execution time required for rule derivation could be prohibitive, in particular, when forming multi-attribute predicates is needed. When a sophisticated predicate is constructed from a combination of attributes, the execution time required grows exponentially with the size of a training database, which is highly undesirable in many applications. The fast growth in the amount of data in those applications has furthermore made the efficient mining for classification rules a very challenging issue. Consequently, in this paper we shall deal with the evaluation of attribute information for mining classification rules so as to improve the mining efficiency.

In a decision tree, each internal node corresponds to a decision on an attribute and each outgoing branch corresponds to a possible value of this attribute. Each leaf node then represents a target class. In order to classify a data sample with an unknown class, the classifier will examine the attribute values of this sample along the decision tree. Clearly, the ordering of attributes in the levels of a decision tree will affect the efficiency of the classification process, and should be determined in accordance with the relevance of these attributes to the target class. The relevance of an attribute to the target class is called the *inference capability* of this attribute. We consider in this paper two different measurements for inference capability, i.e., *inference power* and *information gain*. An attribute that is useful in inferring the group identity of a data tuple is said to have a good inference power to that group identity. On the other hand, the amount of information one can gain through branching on an attribute in a decision tree is referred to as information gain of that attribute [15]. It is noted that for an attribute, both of its inference power and its information gain are dependent upon the stage when the attribute is employed, rather than having fixed values. The use of these two measurements will be investigated in this paper. As will be shown later, these two measurements, though both being related to the relevance to the group identity, can in fact lead to different branching decisions. It is noted that, depending on the stage of tree branching, these two measurements should be judiciously employed so as to maximize the effects they are designed for. In addition, in some cases the group identity is not so dependent on the value of a single attribute. Rather, the group identity depends on the combined values of a set of attributes. Combining several individual attributes to construct multi-attribute predicates can lead to better inference capability. In view of this, the inference power and the information gain of multiple attributes are also evaluated in this paper.

This paper is organized as follows. A problem description is given in Section 2. Inference power and information gain of attributes are comparatively investigated in Section 3. The inference capability from combined attributes is examined in Section 4. Section 5 contains the summary.

# 2 Problem Description

In general, the problem on mining classification rules can be stated as follows. We are given a large database W, in which each tuple consists of a set of $n$ attributes (features), $\{A_1, A_2, ..., A_n\}$. For example, attributes could be age, salary range, gender, zip code, etc. Our purpose is to classify all data tuples in this database into different groups according to their attributes. In order to learn proper knowledge on such classification, we are given a small training database, in which each tuple consists of the same attributes as tuples in W, and additionally has a known group identity associated with it. We shall focus our discussion on the identification of attributes from $\{A_1, A_2, ..., A_n\}$ that have strong relevance to the group identity.

Consider a sample profile for 15 children in Table 1 as an example. In Table 1, each tuple, correspond-

| Label | Gender | Beverage | State | Group id |
|-------|--------|----------|-------|----------|
| 1 | M | water | CA | I |
| 2 | F | juice | NY | I |
| 3 | M | water | NY | I |
| 4 | F | milk | TX | I |
| 5 | M | water | NY | I |
| 6 | M | juice | CA | I |
| 7 | M | water | TX | III |
| 8 | F | juice | CA | II |
| 9 | F | juice | NY | II |
| 10 | F | milk | CA | I |
| 11 | M | milk | TX | II |
| 12 | M | milk | TX | II |
| 13 | F | milk | TX | II |
| 14 | F | water | NY | III |
| 15 | F | water | CA | III |

Table 1: A sample profile for classifying 15 children.

ing to each child, contains attributes: his/her gender, beverage preferred and state lived, and additionally his/her group identity (i.e., class). (For ease of exposition, each tuple is given a label in its first column, which is, however, not part of the attributes.) We now would like to analyze the relevance between the attributes (i.e., gender, beverage and state in this case) and the group identity. As stated before, an attribute that is useful in inferring the group identity of a data tuple is said to have a good inference power to that group identity. Also, the amount of information one can gain through branching on an attribute in a decision tree is referred to as information gain of that attribute. A *predicate* in this study means a classification rule used to classify data tuples in the database W.

As described earlier, in a decision tree, each internal node corresponds to a decision on an attribute and each outgoing branch corresponds to a possible value of this attribute. Each leaf node then represents a target class. In order to classify a data sample with an unknown class, the classifier will examine the attribute values of this sample along the decision tree. Specifically, we shall focus on evaluating the inference capability, i.e., inference power and information gain, of individual attributes for the target class in an attribute-oriented induction procedure.

# 3  Comparing Inference Power with Information Gain

We shall compare, for given attributes, their inference powers with their information gains. The procedure to determine the inference powers of attributes is presented in Section 3.1, and that to obtain the information gains of attributes is presented in Section 3.2.

## 3.1  Identifying Attributes with Good Inference Power

Suppose $A$ is an attribute and $\{a_1, a_2, ..., a_m\}$ are $m$ possible values of attribute $A$. Also, the domain of the group identity $g$ is represented by domain($g$)= $\{v_1, v_2, ..., v_{|domain(g)|}\}$. The *primary group* for a value $a_i$ of attribute $A$, denoted by $v^{a_i}$, is the group that has the most tuples with their attribute $A = a_i$. Explicitly, use $n_A(a_i, v_k)$ to denote the number of tuples which are in group $v_k$ and have a value of $a_i$ in their attribute $A$. Then, we have

$$n_A(a_i, v^{a_i}) = \max_{v_k \in domain(g)} \{n_A(a_i, v_k)\}. \quad (1)$$

The primary group for each value of attribute $A$ can hence be obtained. For the example profile in Table 1, if $A$ is "gender," then domain($A$)= {Male, Female}, and $n_A$(Male,I)=4, $n_A$(Male,II)=2, and $n_A$(Male,III)=1. Group I is therefore the primary group for the value "Male" of the attribute "gender".

The *hit ratio* of attribute $A$, denoted by $h(A)$, is defined as the percentage of tuples which, according to their corresponding attribute values, fall into their primary groups. Let $N$ denote the total number of tuples. Then,

$$h(A) = \frac{\sum_{1 \leq i \leq m} n_A(a_i, v^{a_i})}{N}. \quad (2)$$

It can be seen that the stronger the relationship between an attribute and the group identity, the larger the hit ratio of this attribute will be. A hit ratio of an attribute would become one if that attribute could uniquely determine the group identity. The hit ratio is a quantitative measurement for the inference power of an attribute. According to the primary groups of various values of an attribute, the hit ratio of that attribute can be determined. The most distinguishing attribute refers to the attribute with the highest hit ratio.

| Gender | I | II | III | (max, group) |
|--------|---|----|-----|--------------|
| Male   | 4 | 2  | 1   | (4, I)       |
| Female | 3 | 3  | 2   | (3, I)       |
| hit ratio: | | | | $\frac{7}{15}$ |

Table 2: Distribution when the profile is classified by gender.

| State | I | II | III | (max, group) |
|-------|---|----|-----|--------------|
| CA    | 3 | 1  | 1   | (3, I)       |
| NY    | 3 | 1  | 1   | (3, I)       |
| TX    | 1 | 3  | 1   | (3, II)      |
| hit ratio: | | | | $\frac{9}{15}$ |

Table 3: Distribution when the profile is classified by state.

For illustrative purposes, consider the example profile in Table 1. First, we classify this profile according to gender, and obtain the results in Table 2. As explained earlier, Group I is the primary group for the value "Male" of attribute "gender". Also, it can be seen that Groups I and II are tied for the number of tuples which have "Female" as the value of their attribute "gender". In this case, either Group I or Group II can be the primary group for the value "Female" of attribute "gender". As a result, there are 7 tuples, out of 15 tuples, falling into their primary groups. The hit ratio of attribute gender is thus $\frac{7}{15}$.

Next, we classify this profile according to state, and obtain the primary group for each value of attribute state. The results in Table 3 follow. Consequently, the hit ratio for attribute state is $\frac{9}{15}$. Following the same procedure, we classify the profile in Table 1 according to beverage, and obtain the results in Table 4. It can be verified by Table 4 that the hit ratio of beverage is $\frac{8}{15}$. Finally, having the largest hit ratio among the three attributes, state is the most distinguishing attribute in this example. Part of the decision tree with state as the first level branching attribute is given in Figure 1. The data sets when the values of state are CA and NY are given in Tables 5 and 6, respectively.

## 3.2 Identifying Attributes with High Information Gain

As a typical decision tree procedure, ID-3 employs a top-down irrevocable strategy that searches only part of the search space [15]. Being an information-theoretic approach with the purpose of minimizing the

| Beverage | I | II | III | (max, group) |
|----------|---|----|-----|--------------|
| Juice    | 2 | 2  | 0   | (2, II)      |
| Water    | 3 | 0  | 3   | (3, I)       |
| Milk     | 2 | 3  | 0   | (3, II)      |
| hit ratio: | | | | $\frac{8}{15}$ |

Table 4: Distribution when the profile is classified by beverage.
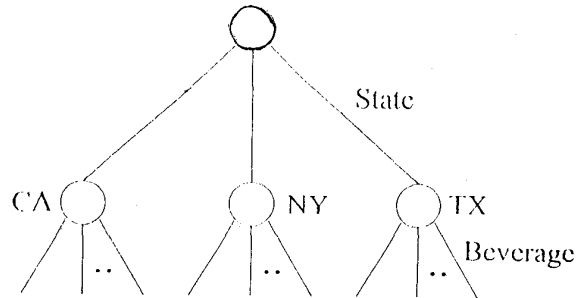


Figure 1: Part of the decision tree with state as the first level branching attribute.

| Label | Gender | Beverage | Group id |
|-------|--------|----------|----------|
| 1     | M      | water    | I        |
| 6     | M      | juice    | I        |
| 8     | F      | juice    | II       |
| 10    | F      | milk     | I        |
| 15    | F      | water    | III      |

Table 5: The data set when state = CA.

| Label | Gender | Beverage | Group id |
|-------|--------|----------|----------|
| 2     | F      | juice    | I        |
| 3     | M      | water    | I        |
| 5     | M      | water    | I        |
| 9     | F      | juice    | II       |
| 14    | F      | water    | III      |

Table 6: The data set when state = NY.

133

traveling time in the decision tree, ID-3 branches on the attribute with the highest information gain so that the further information required for classifying data samples with unknown classes is minimized.

Consider a data set $P$ of $p$ data samples. Suppose the classifying attribute has $m$ distinct classes $P_i$, $1 \leq i \leq m$, and there are $p_i$ data samples for each class $P_i$. Then, the information needed to classify a given data sample is:

$$I(p_1, p_2, ..., p_m) = -\sum_{i=1}^{m} \frac{p_i}{p} \log_2 \frac{p_i}{p}.$$

An attribute $A$ with values $a_1, a_2, ..., a_k$ can be used to partition $P$ into $C_1, C_2, ..., C_k$, where $C_j$ contains those data samples in $P$ that have value $a_j$ of $A$. Let $C_j$ have $p_{ij}$ data samples of class $P_i$. The expected information based on the partitioning by $A$ is given by:

$$E(A) = \sum_{j=1}^{k} \frac{p_{1j} + ... + p_{mj}}{p} I(p_{1j}, ..., p_{mj}).$$

It can be seen that $I(p_{1j}, ..., p_{mj})$ is the information required to classify data samples in $C_j$. The information gain by branching on attribute $A$ is thus gain($A$)= $I(p_1, p_2, ..., p_m) - E(A)$.

Consider the example in Table 1. We now examine the information gain by branching on different attributes. Before the first branching, the information required is

$$I(7,5,3) = -(\frac{7}{15}\log_2\frac{7}{15} + \frac{5}{15}\log_2\frac{5}{15} + \frac{3}{15}\log_2\frac{3}{15}) = 1.506.$$

Using attribute state in the first level:

Using state as the branching attribute, we have $\{a_1, a_2, a_3\}$ ={CA, NY, TX}, and $C_1$ consists tuples with label 1, 6, 8, 10 and 15, as shown in Table 5, where the numbers of tuples following into Groups I, II and III are 3, 1, and 1, respectively. $C_2$ and $C_2$ can be obtained similarly. For notational simplicity, we use $I_{State=CA}$ to represent $I_{State=CA}(3,1,1)$. We then have,

$$I_{State=CA} = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{1}{5} + \frac{1}{5}\log_2\frac{1}{5}) = 1.371$$

$$I_{State=NY} = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{1}{5} + \frac{1}{5}\log_2\frac{1}{5}) = 1.371$$

$$I_{State=TX} = -(\frac{1}{5}\log_2\frac{1}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{1}{5}) = 1.371$$

$$E(State) = \frac{5}{15}I_{State=CA} + \frac{5}{15}I_{State=NY} + \frac{5}{15}I_{State=TX} == 1.371.$$

The value of gain(state) is then $I(7,5,3) - E(state)$=0.135.

Using attribute beverage in the first level:

Following the same procedure, we have,

$$I_{beverage=juice} = -(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4} + \frac{0}{4}\log_2\frac{0}{4}) = 1$$

$$I_{beverage=water} = -(\frac{3}{6}\log_2\frac{3}{6} + \frac{0}{6}\log_2\frac{0}{6} + \frac{3}{6}\log_2\frac{3}{6}) = 1$$

$$I_{beverage=milk} = -(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5} + \frac{0}{5}\log_2\frac{0}{5}) = 0.971$$

$$E(Beverage) = \frac{4}{15}I_{beverage=juice} + \frac{6}{15}I_{beverage=water} + \frac{5}{15}I_{beverage=milk} = 0.99.$$

The value of gain(beverage) is then $I(7,5,3) - E(beverage)$=0.516. Similarly, we can obtain the information gain for attribute gender and learn that beverage is the attribute with the highest information gain. It is worth mentioning that though state, as determined in Section 3.1, is the attribute with the largest hit ratio, beverage is the attribute with the highest information gain.

Note that though both being related to the relevance to the group identity, inference power and information gain are in fact designed to achieve different effects. The inference power of an attribute mainly means "For an arbitrary tuple, what is the probability that we can guess correctly the group identity of this tuple based on the value of that attribute?" In contrast, the information gain of an attribute is a measurement that takes into account the complexity of classifying tuples even after having branched on this attribute. In other words, the inference power is in essence a one-step measurement, whereas the information gain does consider the execution effort for the later process. It should be noted, however, the process of using information gain to build a decision tree is also greedy in nature. How to select a proper measurement is indeed a system-dependent issue. In some database applications, instead of using all attributes of data samples, we may only want to employ a subset of attributes for classification, as a trade-off between the completeness of classification and the execution efficiency. In that case, one may want to have a combined
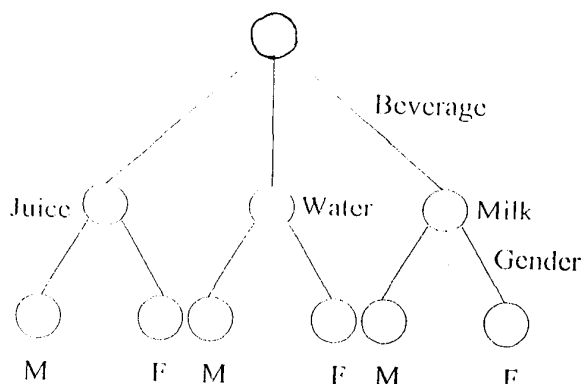
Figure 2: Part of the decision tree with beverage as the first level branching attribute.

| Label | Gender | State | Group id |
|-------|--------|-------|----------|
| 2 | F | NY | I |
| 6 | M | CA | I |
| 8 | F | CA | II |
| 9 | F | NY | II |

Table 7: The data set when beverage = juice.

use of information gain and inference power for the branching of a decision tree. On possible scenario is to identify the first few attributes by information gain and to use inference power as a measurement for those internal nodes not to be further partitioned. Clearly, other scenarios for the combined use of these two measurements are conceivable, depending upon the system environment and the corresponding objectives.

Part of the decision tree with beverage as the first level branching attribute is given in Figure 2. The data sets when the values of beverage are juice and water are given in Tables 7 and 8, respectively.

| Label | Gender | State | Group id |
|-------|--------|-------|----------|
| 1 | M | CA | I |
| 3 | M | NY | I |
| 5 | M | NY | I |
| 7 | M | TX | III |
| 14 | F | NY | III |
| 15 | F | CA | III |

Table 8: The data set when beverage = water.

| Label | Gender | Income | State | Vehicle |
|-------|--------|--------|-------|---------|
| 1 | male | low | NY | car |
| 2 | male | low | CA | car |
| 3 | female | high | NY | car |
| 4 | female | high | NY | car |
| 5 | male | high | CA | bike · |
| 6 | male | high | CA | bike |
| 7 | female | low | NY | bike |
| 8 | female | low | CA | bike |

Table 9: A sample profile for preferred vehicles.

# 4 Combined Attribute Information

As mentioned earlier, in some cases the group identity is not so dependent on the value of a single attribute, but instead, depends upon the combined values of a set of attributes. This is particularly true in the presence of those attributes that have strong relevance among themselves. Consider the profile in Table 9 as an example. In Table 9, it is found that a male with low income and a female with high income usually drive cars, whereas a male with high income and a female with low income ride bikes. In this case, exploring the relationship between "vehicle" (corresponding to the group id in Table 1) and "either gender or income attribute" will lead to little results, since neither gender nor income has a good inference capability to the vehicle. However, a combination of gender and income (e.g., a male and low income) indeed has a good inference capability, in terms of both inference power and information gain, to the vehicle.

Consider the profile in Table 1 for example. As stated in Properties H and G below, the resulting inference capability of a 2-attribute predicate will always be greater than or equal to that of each participating attribute. It can be obtained from Table 10 that for the attribute pair (gender,state), the hit ratio is $\frac{9}{15}$. Similarly, it can be obtained that the hit ratio of (gender,beverage) is $\frac{12}{15}$. It is interesting to see that attribute pair (gender,beverage), though not containing the most distinguishing attribute state, actually has a better inference power than (gender,state).

By generalizing the procedure in Section 3.2, we can determine the information gain of multiple attributes. It can be verified that (gender,beverage) is in fact the 2-attribute pair that has the highest information gain among all 2-attribute pairs. It is noted that to identify

135

| Gender and State | I | II | III | (max, group) |
|---|---|---|---|---|
| male & CA | 2 | 0 | 0 | (2, I) |
| male & NY | 2 | 0 | 0 | (2, I) |
| male & TX | 0 | 2 | 1 | (2, II) |
| female & CA | 1 | 1 | 1 | (1, I) |
| female & NY | 1 | 1 | 1 | (1, I) |
| female & TX | 1 | 1 | 0 | (1, I) |
| hit ratio: | | | | $\frac{9}{15}$ |

Table 10: Combining genders and states to classify the profile in Table 1.

such a 2-attribute pair with the highest information gain, one has to consider the information gain of every 2-attribute pair. An example for this phenomenon can be found in Table 9, where state is the attribute with the highest information gain when individual attributes are considered. The attribute pair (gender, income), which does not contain state, is the 2-attribute pair with the highest information gain, whose branching is shown in Figure 3. Note that such a branching will not be employed if one uses state for the first level branching.

**Property H:** Let $\alpha$ and $\beta$ be two disjoint sets of attributes, with hit ratios $h(\alpha)$ and $h(\beta)$, respectively. Also, use $h(\alpha \cup \beta)$ to denote the hit ratio of the attribute set $\alpha \cup \beta$. Then, the value of $h(\alpha \cup \beta)$ is bounded by:

$$\max\{h(\alpha), h(\beta)\} \leq h(\alpha \cup \beta) \leq 1,$$

in which both the upper and lower bounds are tight bounds.

It can be seen that the right inequality follows from the definition of a hit ratio. We now consider the left inequality. Without loss of generality, assume that $h(\alpha) \geq h(\beta)$. Let $\alpha_i$ be one set of values for attributes in $\alpha$, and $\beta_j$ be one set of values for attributes in $\beta$. $\alpha_i + \beta_j$ represents one set of values for attributes in $\alpha \cup \beta$. $n_{\alpha \cup \beta}(\alpha_i + \beta_j, v_k)$ denotes the number of tuples which are in group $v_k$ and have a value of $\alpha_i + \beta_j$ in their attribute set $\alpha \cup \beta$. $v^{\alpha_i + \beta_j}$ represents the primary group of the set of values $\alpha_i + \beta_j$ in attribute set $\alpha \cup \beta$. Then, we have $n_{\alpha \cup \beta}(\alpha_i + \beta_j, v^{\alpha_i + \beta_j}) = \max_{v_k \in domain(g)} n_{\alpha \cup \beta}(\alpha_i + \beta_j, v_k)$, which implies $\sum_{\beta_j \in domain(\beta)} n_{\alpha \cup \beta}(\alpha_i + \beta_j, v^{\alpha_i + \beta_j}) \geq n_\alpha(\alpha_i, v^\alpha)$. We in turn obtain $h(\alpha \cup \beta) = \sum_{\alpha_i \in domain(\alpha)} \sum_{\beta_j \in domain(\beta)} n_{\alpha \cup \beta}(\alpha_i + \beta_j, v^{\alpha_i + \beta_j}) \geq \sum_{\alpha_i \in domain(\alpha)} n_\alpha(\alpha_i, v^\alpha) = h(\alpha)$, leading to the left inequality. It can be verified that the tightness of the right inequalities can be achieved.
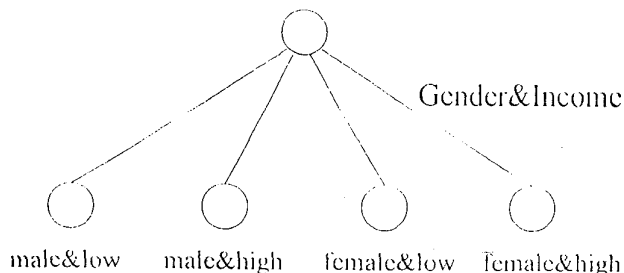


Figure 3: Part of the decision tree when the first level attribute combines gender and income.

Similarly, by omitting its straightforward proof, we have the following property for the information gain from combined attributes.

**Property G:** Let $\alpha$ and $\beta$ be two disjoint sets of attributes, with information gain $I(\alpha)$ and $I(\beta)$, respectively. Also, use $I(\alpha \cup \beta)$ to denote the information gain of the attribute set $\alpha \cup \beta$. Then, the value of $I(\alpha \cup \beta)$ is bounded by:

$$\max\{I(\alpha), I(\beta)\} \leq I(\alpha \cup \beta).$$

# 5 Conclusion

In this paper we have dealt with the evaluation of attribute information for mining classification rules. It is noted that the ordering of attributes in the levels of a decision tree will affect the efficiency of the classification process, and need be determined in accordance with the relevance of these attributes to the target class. We evaluated in this paper two different measurements for the relevance of attributes to the target class, i.e., inference power and information gain. It is shown that these two measurements, though both being related to the relevance to the group identity, can in fact lead to different branching decisions. It is noted that, depending on the stage of tree branching, these two measurements should be judiciously employed so as to maximize the effects they are designed for. Also, a combined use of these two measurements can be employed to improve the overall efficiency of the decision tree process. In addition, the inference power and the information gain of multiple attributes have been examined in this paper.

# Acknowledgements

# References

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient Similarity Search in Sequence Databases. *Proceedings of the 4th Intl. conf. on Foundations of Data Organization and Algorithms*, October, 1993.

[2] R. Agrawal, S. Ghosh, T. Imielinski, B. Iyer, and A. Swami. An Interval Classifier for Database Mining Applications. *Proceedings of the 18th International Conference on Very Large Data Bases*, pages 560–573, August 1992.

[3] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 478–499, September 1994.

[4] R. Agrawal and R. Srikant. Mining Sequential Patterns. *Proceedings of the 11th International Conference on Data Engineering*, pages 3–14, March 1995.

[5] T.M. Anwar, H.W. Beck, and S.B. Navathe. Knowledge Mining by Imprecise Querying: A Classification-Based Approach. *Proceedings of the 8th International Conference on Data Engineering*, pages 622–630, February 1992.

[6] M.-S. Chen. Using Multi-Attribute Predicates for Mining Classification Rules. *Proceedings of 22nd Annual Intern'l Computer Software and Application Conference*, August 19-21 1998.

[7] M.-S. Chen, J.-S. Park, and P. S. Yu. Efficient Data Mining for Path Traversal Patterns. *IEEE Transactions on Knowledge and Data Engineering*, 10(2):209–221, April 1998.

[8] J. Han, Y. Cai, , and N. Cercone. Knowledge Discovery in Databases: An Attribute-Oriented Approach. *Proceedings of the 18th International Conference on Very Large Data Bases*, pages 547–559, August 1992.

[9] J. Han, Y. Cai, and N. Cercone. Data Driven Discovery of Quantitative Rules in Relational Databases. *IEEE Transactions on Knowledge and Data Engineering*, pages 29–40, February 1993.

[10] J. Han and Y. Fu. Discovery of Multiple-Level Association Rules from Large Databases. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 420–431, September 1995.

[11] H. Lu, R. Setiono, and H. Liu. NeuroRule: A Connectionist Approach to Data Mining. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 478–489, September 1995.

[12] J.-S. Park, M.-S. Chen, and P. S. Yu. Using a Hash-Based Method with Transaction Trimming for Mining Association Rules. *IEEE Transactions on Knowledge and Data Engineering*, 9(5):813–825, October 1997.

[13] G. Piatetsky-Shapiro. Discovery, Analysis and Presentation of Strong Rules. *Knowledge Discovery in Databases*, pages 229–238, 1991.

[14] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[15] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1:81–106, 1986.

[16] A. Savasere, E. Omiecinski, and S. Navathe. An Efficient Algorithm for Mining Association Rules in Large Databases. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 432–444, September 1995.

[17] R. Srikant and R. Agrawal. Mining Generalized Association Rules. *Proceedings of the 21th International Conference on Very Large Data Bases*, pages 407–419, September 1995.

[18] J. T.-L. Wang, G.-W. Chirn, T.G. Marr, B. Shapiro, D. Shasha, and K. Zhang. Combinatorial Pattern Discovery for Scientific Data: Some Preliminary Results. *Proceedings of ACM SIGMOD, Minneapolis, MN*, pages 115–125, May, 1994.

[19] W. Ziarko. The discovery, analysis, and representation of data dependancies in databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 195–209. AAAI/MIT Press, 1991.