Incremental Attribute Reduction Method Based on Chi-Square Statistics and Information Entropy

NA SU 1, XINJUN AN 1, CHANGQING YAN 1, AND SHUJUAN JI 2

1Department of Information Engineering, Shandong University of Science and Technology, Tai’an 271000, China
2Department of Information Engineering, Shandong University of Science and Technology, Qingdao 266000, China
Corresponding author: Na Su (susu198@126.com)

This work was supported by the Natural Science Foundation of China under Grant 71772107 and Grant 71403151.

ABSTRACT Attributes in datasets are usually not equally significant. Some attributes are unnecessary or redundant. Attribute reduction is an important research issue of rough set theory, which can find minimum subsets of attributes with the same classification effect as the whole dataset by removing unnecessary or redundant attributes. We use Chi-square statistics to evaluate the significance of condition attributes. It can reduce the search space of attribute reduction and improve the speed of attribute reduction. Conditional entropy of relative attributes is adopted as a heuristic function. Two decision table reduction algorithms, forward selection and backward deletion, are proposed to approach the optimal solution. Based on this, an efficient incremental attribute reduction method for dynamically changing datasets is proposed by preserving intermediate variables. The intermediate variable is the observation frequency matrix of joint events of each condition attribute and decision attribute. Experimental results show that the proposed algorithms can improve performance in terms of processing time.

INDEX TERMS Attribute reduction, decision table, incremental attribute reduction, Chi-square statistics.

I. INTRODUCTION

With the development of Internet and the popularization of computer application, datasets are increasing rapidly in the number of objects and attributes. It is very difficult to process all the data because of limitation of memory. It is necessary to compress dataset without affecting the mining results.

At present, there are two mainly methods of compressing the data dimensions. One method is to compress the original data by aggregating data and converting it into a smaller space, such as principal component analysis [1]–[3] and singular value decomposition [4]–[6]. Another method is to reduce data dimensions by deleting unnecessary or redundant attributes, such as attribute reduction based on rough set, Chi-square statistics, clustering, neural network, information entropy, support vector machine and so on [7]–[9].

Attributes in a dataset are usually not equally significant. Some attributes are unnecessary or redundant. We expect to use a dataset as few attributes as possible without affecting the quality of mining results. The minimum reduction is the smallest set of attributes that can distinguish equivalence classes. The attribute reductions for each dataset are generally multiple. The intersection of all reductions is called attribute core. However, for a dataset with \( n \) attributes, there may be \( 2^n \) subsets. It is impossible to find the best subset of attributes by exhaustive search.

At present, the attribute reduction methods can be divided into three classes: attribute reduction algorithms based on information entropy [10], [11], attribute reduction algorithms based on discernible matrix [12], [13] and attribute reduction algorithms based on positive region [14], [15]. Attribute reduction algorithms based on discernible matrix have a high accuracy, but their computational cost is high.

Attribute reduction methods are used to remove redundant attributes and find a subset of attributes that can maintain the same data classification capabilities as the original data in [16]–[18]. The methods based on equivalence relation and equivalence class are only suitable for datasets with discrete attributes. However, real datasets have many continuous numerical attributes. Hu presents a reduction method for numerical attributes [19]. The granulation model of universe space is formed by the neighborhood of any object in the
universe space and realizes the granular computing of continuous numerical space. The work [29] enhances the precision of fault diagnosis for all fault classes by using Chi-square statistics to reduce attribute dimensions. A cost-sensitive embedded feature selection method is proposed to solve the class imbalance problem in [30]. The work [31] proposes a heuristic attribute reduction method by constructing the attribute significance measures based on stripped quotient sets. The work [32] proposed a feature selection for multi-label classification using neighborhood relationship preserving score, which is a feature evaluation criterion for multi-label feature selection. These methods are suitable for static datasets. However the real datasets are usually dynamic.

In order to deal with new samples in time, many incremental attribute reduction methods are proposed. An incremental reduction algorithm based on positive region is proposed to deal with inconsistent decision tables in [20], but the algorithm can not get the minimum reduction. The work [31] determines whether the decision table needs to be updated by analyzing the impact of new samples on the positive domain. The sample set which is not consistent with new samples is reduced so as to get the updated reduction quickly. The work [32] proposed a feature selection algorithm based on positive region is proposed to deal with inconsistent decision tables in [20], but the algorithm can not get the minimum reduction. The work [21] poses a heuristic attribute reduction method by constructing three new types of discernibility matrices is proposed. An incremental attribute reduction algorithm, named IARAIV is proposed. In this algorithm, new samples and their inconsistent neighborhoods are reduced so as to get the updated reduction quickly. The remainder of this paper is organized as follows: Section II discusses related work. Section III proposes two decision table reduction algorithms, which are forward selection and backward deletion. An incremental attribute reduction algorithm is presented in Section IV. In Section V, our methods are experimented and evaluated.

II. RELATED WORK
A. ROUGH SET THEORY
Rough set theory was put forward by Professor Pawlak, a Polish scholar, in 1982 [25]. It is a mathematical theory that deals with incomplete knowledge such as inaccuracy, incompleteness and inconsistency. Rough set has been widely used in attribute reduction, knowledge classification, rule mining and other fields. Attribute reduction is an important research issue of rough set. It can delete unnecessary or redundant attributes and reduce the data size so as to improve the efficiency of data processing and mining. Attribute reduction based on rough set can use information entropy to measure attributes significance. Rough set attribute reduction can delete some attributes while remaining the data classification unchanged.

1) RELEVANT DEFINITIONS OF ROUGH SET THEORY
Rough set theory associates classification with knowledge and understands knowledge as the division of specific space. The following are some related concepts [25].

Definition 1: An information system \( DS = < U, R, V, f > \) is a 4-tuple system, where \( U = \{u_1, u_2, \ldots, u_n\} \) is a non-empty set of objects called universe and \( R \) is a infinite nonempty attribute set. \( V = \bigcup_{a \in C \cup D} V_a \), where \( V_a \) is the domain of attribute \( a \) \( f: U \times (C \cup D) \rightarrow V \) is an information function that specifies the attribute value of each object \( x \) in \( U \). If \( R = C \cup \{d\} \) and \( C \cap D = \emptyset \), where \( C \) is the condition attribute set and \( \{d\} \) is the decision attribute, then \( DS = < U, C \cup \{d\}, V, f, > \) is also called as a decision table.

2) DESCRIPTION OF INFORMATION VIEW OF ROUGH SET THEORY
If \( U \) is a universe, it can be considered that any attribute set of \( U \) is a random variable on \( \sigma \)-algebra defined on a subset of \( U \). Its probability distribution can be determined by the following method.

Definition 3: Let \( P \) and \( Q \) be the derived partition on \( U \) as \( X, Y: X = \{X_1, X_2, \ldots, X_m\}, Y = \{Y_1, Y_2, \ldots, Y_m\} \). Then the probability distribution defined by \( P \) and \( Q \) on \( \sigma \)-algebra consisting of subsets of \( U \) is

\[
[X; p] = \begin{bmatrix}
X_1 & X_2 & \cdots & X_n \\
p(X_1) & p(X_2) & \cdots & p(X_n)
\end{bmatrix},
\]

\[
[Y; p] = \begin{bmatrix}
Y_1 & Y_2 & \cdots & Y_m \\
p(Y_1) & p(Y_2) & \cdots & p(Y_m)
\end{bmatrix}
\]

where \( p(X_i) = \frac{|X_i|}{|U|}, i = 1, 2, \ldots, n \); \( p(Y_i) = \frac{|Y_i|}{|U|}, j = 1, 2, \ldots, m \).

Definition 4: The entropy of attribute set \( P \) is defined as

\[
H(P) = -\sum_{i=1}^n p(X_i) \log[p(X_i)].
\]

Definition 5: The conditional entropy \( H(Q|P) \) of attribute set \( Q \) relative to attribute set \( P \) is defined as

\[
H(Q|P) = -\sum_{i=1}^n p(X_i) \sum_{j=1}^m p(Y_j|X_i) \log[p(Y_j|X_i)].
\]

where \( p(Y_j|X_i) = \frac{|Y_j \cap X_i|}{|X_i|}, i = 1, 2, \ldots, n \), \( j = 1, 2, \ldots, m \).

Theorem 1: \( P \) is a subset of condition attribute set. \( \{d\} \) is a decision attribute. Then an attribute \( \{r\} \) of \( P \) is unnecessary or redundant relative to decision attribute \( \{d\} \) if and only if \( H(\{d\}|P) = H(\{d\}|P\{-r\}) \).

Definition 6: \( Q \subseteq P \) is the sufficient and necessary condition of a reduction of \( P \) relative to decision attribute \( \{d\} \) as follows:

1) \( H(\{d\}|Q) = H(\{d\}|P) \);
2) For any attribute \( r, QH(\{d\}|Q) \neq H(\{d\}|Q\{-r\}) \).

Definition 7: \( \{r\} \in P \) is a core attribute if and only if \( H(\{d\}|P) > H(\{d\}|P\{-r\}) \).
3) ATTRIBUTE REDUCTION METHODS OF PROGRESSIVE FORWARD SELECTION AND PROGRESSIVE BACKWARD DELETION

According to the reduction process, attribute reduction methods can be divided into progressive forward selection and progressive backward deletion. Generally, greedy algorithms are adopted. Local optimal selection strategy is adopted to achieve the global optimal solution when searching the attribute set space.

Progressive forward selection usually starts with an empty set or a core attribute set, then the most significant is added from the remaining set to the reduction set until the termination condition is satisfied. There are many evaluation measures to judge whether an attribute is significant or not, such as information gain, attribute frequency and so on.

A numerical attribute reduction method based on neighborhood granulation and rough approximations is proposed in [19]. Datasets whose attribute values are continuous can be processed. The reduction process starts with empty set and uses dependency function as heuristic information. The most significant attribute is selected from the remaining attribute set and added to the reduction set one by one until the significance of each remaining attribute is zero.

Progressive backward deletion starts with the whole set of attributes, then deletes the least significant attribute in the set one by one until a reduction is obtained.

The work [26] presents CEBARKNC, an attribute reduction algorithm based on information entropy. The initial condition is the whole set of attributes. The attribute with the largest relative condition attribute is deleted as the least significant attribute one by one. The time complexity of the algorithm is $O(|C| U^2) + O(|U|^3)$.

In [27], the attributes in the difference set are deleted by reverse deletion gradually. When a single attribute element appears, it is added to the reduction set. This process is repeated until the difference set is empty. Based on the simplified discernible matrix, the work [28] proposes a method of deleting attributes for attribute reduction, which can delete one or more attributes satisfying the conditions once and reduce the number of iterations.

In this paper, we use Chi-square statistics to evaluate the significance of condition attributes. And all unnecessary attributes are deleted. Using conditional entropy of related attributes as a heuristic function, two decision table reduction algorithms, forward selection and backward deletion, are proposed to approach the optimal solution. On this basis, an incremental attribute reduction method for dynamic datasets is proposed.

### B. ATTRIBUTE CORRELATION MEASUREMENT BASED ON CHI-SQUARE STATISTIC

Chi-square statistic is a hypothesis test method, which is mainly used for statistical inference of classified data. Chi-square statistic is to calculate the value of Chi-square by calculating the deviation degree between the actual observed value and the theoretical inferred value. The larger the Chi-square value, the higher the correlation between the two attributes. Otherwise, the less relevant the two attributes are.

**Definition 8:** The correlation between attribute $A$ and attribute $B$ can be measured by $K$ (Chi-square). Let $A$ has $c$ different values $a_1, a_2, \ldots, a_c$, and $B$ has $l$ different values $b_1, b_2, \ldots, b_l$. The order $(A_i, B_j)$ denotes the event that attribute $A$ takes $a_i$ and attribute $B$ takes $b_j$, where $1 \leq i \leq c$ and $1 \leq j \leq l$. The value of $K$ is defined as

$$K = \sum_{i=1}^{c} \sum_{j=1}^{l} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad (4)$$

where $o_{ij}$ is the observation frequency of $(A_i, B_j)$, $e_{ij}$ is the expected frequency of $(A_i, B_j)$. $e_{ij}$ is defined as

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N} \quad (5)$$

where $N$ is the number of tuples, $\text{count}(A = a_i)$ is the number of tuples with value $a_i$ of $A$, that is $\sum_{j=1}^{l} o_{ij}$. The number of tuples with value $b_j$ of $\text{count}(B = b_j)$ is that $\sum_{i=1}^{c} o_{ij}$.

Attribute $A$ is related to attribute $B$ if the value of $K$ is greater than the critical value, which is the corresponding value in the Chi-square table with degrees of freedom $(l-1) \times (c-1)$. The larger the $K$ value is, the stronger the correlation between $A$ and $B$ is. Otherwise attribute $A$ is not related to attribute $B$.

**Definition 9:** Relevant attribute set $F$: $F$ is the set of condition attributes related to decision attribute $d$ measured by Chi-square statistics, that is, the attributes in $C$ whose Chi-square statistics are not less than the critical value. Correlation set $F$ is sorted in descending order according to Chi-square statistics.

The following example illustrates the correlation of two attributes measured by Chi-square statistics.

An institution randomly investigated the income of 5000 people and analyzed the relationship among income, gender and education. According to certain standards, income was divided into three classes: high, middle and low. Gender and income statistics are shown in Table 1. Education and income statistics are shown in Table 2.

| Gender | Male | Female | Total |
|--------|------|--------|-------|
| High   | 600  | 580    | 1180  |
| Middle | 1470 | 1300   | 2770  |
| Low    | 550  | 500    | 1050  |
| Total  | 2620 | 2380   | 5000  |

It is assumed that income does not depend on gender. The expected frequency of each item is calculated according to the aggregate value. Each expectation is calculated by dividing the aggregate value of the row and column by the total value.
TABLE 2. Statistics of income and education.

| Income | High | Middle | Low  | Total |
|--------|------|--------|------|-------|
| High   | 580  | 600    | 100  | 1280  |
| Middle | 680  | 1320   | 300  | 2300  |
| Low    | 280  | 520    | 620  | 1420  |
| Total  | 1540 | 2440   | 1020 | 5000  |

TABLE 3. Statistics of income and education.

| Income  | High   | Middle | Low  | Total |
|---------|--------|--------|------|-------|
| Male    | 618.32 | 1451.48| 550.2| 2620  |
| Female  | 561.68 | 1318.52| 499.8| 2380  |
| Total   | 1180   | 2770   | 1050 | 5000  |

As shown in Table 3, there are 2,380 women and 1,180 high-income people in total. The corresponding expected frequency of high-income women is \( \frac{2380 \times 1180}{5000} = 561.68 \).

Then Chi-square statistics are calculated according to equation 4. For example, the Chi-square statistic of high-income women is \( \frac{(580-561.68)^2}{561.68} = 0.5975 \). The calculation results are shown in Table 4. The degree of freedom of the table is \( p_{n+1} \) and the significance level \( (2-1) \times (3-1) = 2 \) is set as 0.05. The critical value is 5.9914645 by using Chi-INV (0.05, 2) function. Because the total Chi-square statistics in Table 4 is 1.6369, which is lower than the critical value, it shows that there is a strong correlation between income and gender is not significant.

TABLE 4. Chi-square Statistics of income and gender.

| Income  | High   | Middle | Low  | Total |
|---------|--------|--------|------|-------|
| Male    | 0.5428 | 0.2363 | 0.0001| 0.7792|
| Female  | 0.5975 | 0.2601 | 0.0001| 0.8577|
| Total   | 1.1403 | 0.4964 | 0.0002| 1.6369|

TABLE 5. Chi-square Statistics of income and education.

| Income | High   | Middle | Low  | Total |
|--------|--------|--------|------|-------|
| High   | 87.5273| 0.9720 | 99.4166| 187.9159|
| Middle | 1.1386 | 34.7877| 61.0159| 96.9422|
| Low    | 56.6174| 43.1701| 376.6615| 476.4490|
| Total  | 145.2833| 78.9298| 537.0939| 761.3070|

Use the same method to calculate the Chi-square statistics of income and education. As shown in Table 5, the total Chi-square statistics is 761.307, which is far greater than the critical value of 5.9914645. It shows that there is a strong correlation between income and education.

In this paper, we use Chi-square statistics to evaluate the direct correlation between condition attributes and decision attribute. A condition attribute is a row and a decision attribute is a column. If a condition attribute is highly correlated with decision attribute, the condition attribute is considered to be a reduction attribute.

III. ATTRIBUTE REDUCTION ALGORITHMS

CFAR AND CBAR

A. ALGORITHM DESIGN

Attribute reduction methods are usually to evaluate the significance of the remaining attributes one by one in the process of adding or deleting. However, if there are more unnecessary attributes in the dataset, the computation will be greatly increased.

We propose a method of attribute reduction based on progressive forward selection strategy. The initial set includes the attribute with the largest \( K \) value of all condition attributes. The attribute is selected from the related attribute set \( F \) with maximum conditional entropy one by one until the conditional entropy increment of all attributes is 0.

In each iteration of selection process the attributes which increment value of conditional entropy is 0 are deleted as shown in algorithm 1. The input of algorithm 1 is decision table \( DS = (U, R, V, f) \), where \( R = F \cup \{d\} \) is an attribute set, \( F \) is the relevant attribute set, \( \{d\} \) is a decision attribute set. The initial set is \( \{r_{\text{max}}\} \), where \( r_{\text{max}} \) is the attribute with the largest \( K \) value of all condition attributes.

Algorithm 1 Forward Selection Attribute Reduction Algorithms Based on Chi-Square Statistics (CFAR)

Input: Decision Table \( DS = \langle U, R, V, f \rangle \).

Output: An attribute reduction set \( P \).

\( P = \{r_{\text{max}}\}; F = F - \{r_{\text{max}}\} \);

DO \((|F| \neq 0)\)

Calculate the conditional entropy \( H((d)|P+\{r_i\}) \) of decision attribute \( \{d\} \) relative to condition attribute set \( P \)

IF for each attribute \( \{r_i\} \) in \( F \), \( H((d)|P+\{r_i\}) = H((d)|P) \) is satisfied

End of reduction;
ELSE

\( P = P \cup \{r_{\text{max}}\}; \)

// \( r_{\text{max}} \), the attribute with the largest \( H((d)|P+\{r_{\text{max}}\}) \) value, is added to the reduction set

\( F = F - \{r_{\text{max}}\} \);

// Remove attribute \( r_k \) with \( H((d)|P+\{r_k\}) = 0 \) from \( F \)

An attribute reduction method based on gradual backward deletion strategy is also proposed. The initial set is the relevant attribute set \( F \). The attributes which conditional information entropy in the set remains unchanged are deleted one by one from the initial set until the conditional entropy is greater than 0. The steps are shown in algorithm 2.
Algorithm 2 Backward Deletion of Attribute Reduction Algorithms Based on Chi-Square Statistics (CBAR)

Input: Decision Table $DS = (U, F \cup \{d\}, V, f)$.
Relevant attribute set $F$.
$d$ is a decision attribute set.

Output: An attribute reduction set $P$.

$P = F$;
DO $(|P| \neq 0)$
Calculate the conditional entropy $H(D|P-\{r_i\})$ of decision attribute $\{d\}$ relative to condition attribute set $P$
IF for each attribute $\{r_i\}$ in $P$, $H(D|P-\{r_i\}) \neq H(D|P)$ is satisfied
End of reduction
ELSE
IF $H(D|P-\{r_i\}) = 0$
$P = P-\{r_i\}$

B. EXAMPLE ANALYSIS

Conditional entropy is used as a heuristic function to select significant attributes or remove unnecessary attributes to obtain the reduction set related to decision attributes. The results of data analysis of reduction set and original dataset are the same or similar.

Chi-square statistics can evaluate the correlation between each condition attribute and decision attribute. However, not all the condition attributes with high relevance to decision attributes should be included in the reduction set.

For example, a shopping record table contains 75 data samples with three attributes: $\{\text{occupation}\}$, $\{\text{age}\}$ and $\{\text{whether to buy product A}\}$. The $\{\text{occupation}\}$ and $\{\text{age}\}$ are condition attributes and $\{\text{whether to buy product A}\}$ is decision attribute. The statistics of two condition attributes and decision attributes are shown in Table 6 and Table 7 respectively. The corresponding expected frequencies are shown in Table 8 and Table 9.

| TABLE 6. Statistics of occupation and shopping. |
|-----------------------------------------------|
|          | Buy A | Not buy A | Total |
| Civil servant | 5     | 15       | 20    |
| Teacher    | 5     | 15       | 20    |
| Student    | 30    | 5        | 35    |
| Total      | 40    | 35       | 75    |

| TABLE 7. Statistics of age and shopping. |
|-----------------------------------------|
|          | Buy A | Not buy A | Total |
| Under 18 years old | 30    | 5        | 35    |
| 18-30      | 3     | 10       | 13    |
| 31-45      | 4     | 10       | 14    |
| 46-60      | 3     | 10       | 13    |
| Total      | 40    | 35       | 75    |

Table 10 shows the Chi-square statistics of $\{\text{occupation}\}$ and $\{\text{whether to buy product A}\}$. The calculated value is 27.6467. The critical value of Chi-square detection function with freedom of $(3-1) \times (2-1) = 2$ is 5.9915. Table 11 shows the Chi-square statistics of $\{\text{age}\}$ and $\{\text{whether to buy product A}\}$. The calculated value is 26.9838. The critical value of Chi-square detection function with freedom of $(4-1) \times (2-1) = 3$ is 7.8147.

| TABLE 8. Expectation frequency of occupation and shopping. |
|----------------------------------------------------------|
|              | Buy A | Not buy A | Total |
| Civil servant | 10.6667 | 9.3333 | 20 |
| Teacher    | 10.6667 | 9.3333 | 20 |
| Student    | 18.6667 | 16.3333 | 35 |
| Total      | 40    | 35       | 75    |

| TABLE 9. Expectation frequency of age and shopping. |
|-----------------------------------------------|
|          | Buy A | Not buy A | Total |
| Under 18 years old | 18.67 | 16.33 | 35 |
| 18-30      | 6.93  | 6.07   | 13   |
| 31-45      | 7.47  | 6.53   | 14   |
| 46-60      | 6.93  | 6.07   | 13   |
| Total      | 40    | 35       | 75    |

Both statistics are larger than the critical value, which indicates that both $\{\text{occupation}\}$ and $\{\text{age}\}$ attributes are related to the decision attribute $\{\text{whether to buy product A}\}$. However, almost all samples who buy product A are students. And the age distribution of students is under 15 or 15-18 years old. Therefore, the $\{\text{age}\}$ attribute is selected to the reduction set and the $\{\text{occupation}\}$ attribute need not be added. In this case $\{\text{occupation}\}$ is a redundant attribute although it is related.

Therefore, the reduction set cannot be found only by Chi-square statistical detection. Information entropy should be used as the heuristic function of reduction.

C. ALGORITHM ANALYSIS

Assuming that the number of samples is $U$ and the number of condition attributes is $R$. The algorithms CFAR and
CBAR proposed in this paper require Chi-square statistical calculation of all condition attributes. The time complexity is \(O(R \| U \|^2)\). After deleting all unnecessary attributes, the attributes are added or deleted one by one. The worst computational complexity of CEBARKNC algorithm is \(O(R \| U \|^2) + O(|U|^3)\).

IV. INCREMENTAL ATTRIBUTE REDUCTION ALGORITHM ICAR

A. ALGORITHM DESIGN

In many practical application scenarios, a large number of new data are generated every moment. If static reduction algorithm is used to reduce the attributes of the new dataset, it will consume a lot of time and space. In order to process new samples efficiently and accurately, we propose an incremental attribute reduction algorithm.

The matrix \(O_{AB}\) stores the observation frequency of each joint event of \(A\) in condition attribute set \(C\) and the decision attribute \(B = \{d\}\) offline to reduce the repeated calculation.

Let a condition attribute \(A\) have \(c\) different values \(a_1, a_2, \ldots, a_c\), and the decision attribute \(B\) has \(l\) different values \(b_1, b_2, \ldots, b_l\). The observation frequency matrix \(O_{AB}\) of the joint event of attribute \(A\) and attribute \(B\) is shown in equation 6, where \(o_{ij}\) is the observation frequency of the joint event \((A_i, B_j)\), \(1 \leq i \leq c\) and \(1 \leq j \leq l\). \(o_{i,l+1}\) is the number of samples satisfying \(A = a_i\), that is \(o_{i,l+1} = \sum_{k=1}^{l} o_{ik}\). \(o_{c+1,j}\) is the number of samples satisfying \(B = b_j\), that is \(o_{c+1,j} = \sum_{k=1}^{c} o_{kj}\). \(o_{c+1,l+1}\) is the total number of samples.

\[
O_{AB} = \begin{bmatrix}
  o_{11} & o_{12} & \cdots & o_{1l} & o_{1,l+1} \\
  o_{21} & o_{22} & \cdots & o_{2l} & o_{2,l+1} \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  o_{c1} & o_{c2} & \cdots & o_{cl} & o_{c,l+1} \\
  o_{c+1,1} & o_{c+1,2} & \cdots & o_{c+l} & o_{c+1,l+1}
\end{bmatrix}
\]  

Then the expected frequency of \((A_i, B_j)\) is \(e_{ij} = \frac{o_{i1}o_{1,l+1} + o_{i2}o_{2,l+1} + \cdots + o_{ic}o_{c,l+1}}{o_{i,l+1}}\).

The conditional entropy \(H(B | A)\) of attribute \(B\) relative to attribute \(A\) is

\[
H(B | A) = -\sum_{x \in A} \left( \sum_{y \in Y} p(y | x) \log p(y | x) \right) = -\sum_{i=1}^{c} o_{ci,l+1} \sum_{j=1}^{l} \frac{o_{ij}}{o_{ci,l+1}} \log \frac{o_{ij}}{o_{ci,l+1}}
\]  

A multidimensional array \(M\) is used to store the observation frequency of all joint events to calculate conditional entropy. \(M\) has \(n\) dimensions, where \(n = (|C| + 1)\). \(M_{v1,v2,\ldots,vn}\) represents the observation frequency of the joint event \((C_1 = v_1, C_2 = v_2, \ldots, D = v_n)\), where \(D\) is decision attribute \(d\) and \(C_i \in C, i = 1, 2, \ldots, |C|\).

For the newly added samples set \(Z = \{z_1, z_2, \ldots, z_n\}\), the change of the observation frequency matrix may have the following four cases:

1. There are no new attribute values for the condition attribute and the decision attribute. The observation frequency matrix \(O_{AB}'\) of the joint event of condition attribute \(A\) and decision attribute \(B\) in set \(Z\) is shown in equation 8. Then the observation frequency matrix after adding set \(Z\) is \(O_{AB} = O_{AB} + O_{AB}'\).

\[
O_{AB}' = \begin{bmatrix}
  o_{11}' & o_{12}' & \cdots & o_{1l}' & o_{1,l+1}' \\
  o_{21}' & o_{22}' & \cdots & o_{2l}' & o_{2,l+1}' \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  o_{c1}' & o_{c2}' & \cdots & o_{cl}' & o_{c,l+1}' \\
  o_{c+1,1}' & o_{c+1,2}' & \cdots & o_{c+l}' & o_{c+1,l+1}'
\end{bmatrix}
\]

2. The condition attribute \(A\) adds \(m\) different attribute values \(a_{c+1}, a_{c+2}, \ldots, a_{c+m}\), and the decision attribute does not add different attribute values. Then the observation frequency matrix of the joint event of condition attribute \(A\) and decision attribute \(B\) in set \(Z\) is \((c+m) \times l\) matrix \(O_{AB}'\) shown in equation 9. \(O_{AB}\) adds 0 of \(m\) rows to \(O_{AB}'\). Then the observation frequency matrix after adding new sample set \(Z\) is \(O_{AB} = O_{AB}' + O_{AB}\).

\[
O_{AB}' = \begin{bmatrix}
  o_{11}' & o_{12}' & \cdots & o_{1l}' & o_{1,l+1}' \\
  o_{21}' & o_{22}' & \cdots & o_{2l}' & o_{2,l+1}' \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  o_{c1}' & o_{c2}' & \cdots & o_{cl}' & o_{c,l+1}' \\
  o_{c+1,1}' & o_{c+1,2}' & \cdots & o_{c+l}' & o_{c+1,l+1}' \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  o_{c+m,1}' & o_{c+m,2}' & \cdots & o_{c+m,l}' & o_{c+m,l+1}' \\
  o'_{c+m+1,1} & o'_{c+m+1,2} & \cdots & o'_{c+m+l} & o'_{c+m+l+1}
\end{bmatrix}
\]

3. The decision attribute has added \(w\) different attribute values \(b_{l+1}, b_{l+2}, \ldots, b_{l+w}\), while the condition attribute does not add different attribute values. The observation frequency matrix of joint event of condition attribute \(A\) and decision attribute \(B\) in set \(Z\) is \(c \times (l+w)\) matrix \(O_{AB}'\). \(O_{AB}\) adds 0 of \(w\) columns into \(O_{AB}'\). Then the observation frequency matrix after adding new sample set \(Z\) is \(O_{AB} = O_{AB}' + O_{AB}\).

4. The condition attribute has added \(m\) different attribute values \(a_{c+1}, a_{c+2}, \ldots, a_{c+m}\) and the decision attribute has added \(w\) different attribute values \(b_{l+1}, b_{l+2}, \ldots, b_{l+w}\). The observation frequency matrix of joint event of condition attribute \(A\) and decision attribute \(B\) in sample set \(Z\) is \((c+m) \times (l+w)\) matrix \(O_{AB}'\). \(O_{AB}\) adds 0 of \(m\) rows and 0 of \(w\) columns into \(O_{AB}'\). Then the observation frequency matrix after adding new sample set \(Z\) is \(O_{AB} = O_{AB}' + O_{AB}\).

The Chi-square statistics \(K'\) of each condition attribute \(r_m \in C\) and the decision attribute \(d\) are calculated according to the updated observation frequency matrix, where \(m = 1, 2, \ldots, |C|\). The following two cases need to be dealt with:

1. If \(K'\) is greater than or equal to the critical value and \(K\), the conditional entropy \(H(d | P + \{r_m\})\) of the relative
condition attribute set \( P \) of the decision attribute set \( d \) is calculated. If \( H(d|P + \{ r_m \}) \neq H(d|P) \), the condition attribute \( r_m \) is added to the reduction set. Otherwise, the reduction set will remain unchanged.

(2) If \( K' \) is less than the critical value and \( K \), the conditional entropy \( H(d|P - \{ r_m \}) \) of the relative condition attribute set \( P \) of the decision attribute set \( d \) is calculated. If \( H(d|P - \{ r_m \}) = H(d|P) \), the attribute \( r_m \) is deleted from the reduction table. Otherwise, \( r_m \) is not deleted.

Based on above description, the steps of incremental attribute reduction method are shown as algorithm 3.

| Algorithm 3 Incremental Backward Deletion of Attribute Reduction Algorithms based on Chi-Square Statistics (ICAR) |
|---------------------------------|
| **Input:** Decision Table \( DS = \langle U, C \cup \{ d \}, V.f \rangle \), where \( C \) is an condition attribute set. \( P \) is a set of primitive reductions, \( P \subset C \). \( \{ d \} \) is a decision attribute set. Observation frequency matrix \( O \) of joint event of original condition attribute and decision attribute. New sample set \( Z = \{ z_1, z_2, \ldots, z_n \} \). |
| **Output:** An attribute reduction set \( P \). |

Update each matrix \( O_{AB} \) according to \( Z \);
Update array \( M \) according to \( Z \);
For each \( \{ r_m \} \) in \( C \)

\[
K' = \sum_{(i,j)} \frac{(o_{ij} - e_{ij})^2}{e_{ij}};
\]

//Calculate \( K' \) of the condition attribute and decision attribute based on the updated matrix \( O \);

IF \( (K' \) critical value) and \( (K' > K) \) and \( (\{r_m\} \in C-P) \) \n
IF \( H(d|P + \{ r_m \})Q \subseteq PH(d|P) \)
// Calculate conditional entropy based on updated array \( M \);

\[
P = P + \{ r_m \};
\]

IF \( (K' \) critical value) and \( (K' < K) \) and \( (\{r_m\} \in P) \) \n
IF \( H(d|P - \{ r_m \}) = H(d|P) \)

\[
P = P - \{ r_m \};
\]

### B. ALGORITHM ANALYSIS

Suppose the number of initial decision samples is \( U \), the number of condition attribute set is \( C \), the maximum number of different values of original condition attribute set is \( C \), the number of different values of original decision attributes is \( L \), and the number of attributes of original reduction set is \( P \). After data updating the number of new samples is \( N \), the maximum number of different values of condition attributes is \( C + M \), and the number of different values of decision attributes is \( L + W \).

The worst time complexity of IARAIV algorithm is \( O((N + U)^2(R-P)^3) \). The ICAR algorithm proposed in this paper saves the observation frequency matrix and does not need to be recalculated every time. Firstly, the observation frequency of the new samples is counted and the time required is \( O(NR) \). Then the original observation frequency matrix is added to the new sample observation frequency matrix and the time required is \( O(C + M)(L + W) R \). The time complexity of calculating \( K \) of each condition attribute and decision attribute is \( O(C + M)(C + M)(L + W) R \). If there are \( N_1 \) changes in the relationship between \( K \) and the critical value, time complexity of computing information entropy is \( O(N_1(U + N)^2) \). Then the total time complexity is \( O(NR) + O(C + M)(C + M)(L + W) + O(C + M)(C + M)(L + W) R) + O(N_1(U + N)^2) \).

Because \( C + M < N + U \), \( L + W < N + U \) and \( N_1 < R \), the worst time complexity of ICAR algorithm is \( O((N + U)^2 R) \).

### V. EXPERIMENTAL RESULTS AND ANALYSIS

#### A. EXPERIMENT ON ATTRIBUTE REDUCTION ALGORITHM OF STATIC DATASETS

10 datasets in UCI library are used for experiments. The number of records and attributes contained in these data sets is shown in Table 12. We preprocessed the datasets, including normalization, discretization and transformation. Such as attributes \{1, 4, 5, 8, 10\} in Heart are normalized, attributes \{Absences, G1, G2, G3\} in Student and attributes \{32, 39, 41, 47, 58\} in Spambase are discretized, and the numerical
attributes are converted to polynomial type in Spambase, Student and Heart.

The forward selection attribute reduction algorithm CFAR, backward deletion attribute reduction algorithm CBAR, the Chi-square statistical method and conditional information entropy reduction algorithm CEBARKNC without Chi-square statistics are compared. The experimental results are shown in Table 12. Experiments show that algorithm CFAR, algorithm CBAR and algorithm CEBARKNC can all get the correct attribute reduction set. While Chi-square statistical method does not get the minimum reduction attribute set. This is because if two attributes are strongly correlated and the two attributes and decision attributes are also strongly correlated. Then one of the two attributes may be a redundant attribute.

Experiments also show that the average running time of algorithm CFAR and CBAR is lower than that of algorithm CEBARKNC. This is because CFAR and CBAR calculate Chi-square statistics before conditional information entropy and delete the attributes unrelated to decision attributes. Thus the computational complexity is reduced.

The training set and test set have been divided in Spect, and the number of records is 80 and 187 respectively. For other data sets, we randomly selected 70% of them as training sets and the remaining 30% as test sets. In order to compare the classification effect of the reduced datasets and the raw datasets, we use random forest algorithm to model these datasets.

Precision and recall are used as evaluation criteria. Precision is calculated as follows:

\[
\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} = \frac{\text{true positive}}{\text{no. of predicted positive}}
\]  

Recall is calculated as follows:

\[
\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} = \frac{\text{true positive}}{\text{no. of actual positive}}
\]

The experimental results are shown in Fig.1, Fig.2 and Fig.3. The results show that the classification effect of the reduced dataset is not much different from that of the original dataset, even better than that of the original dataset, such as Zoo and Heart. Because of the small amount of training data in the Spect data set, the effect is relatively poor.

**B. EXPERIMENTAL ON INCREMENTAL ATTRIBUTE REDUCTION ALGORITHMS FOR DYNAMIC DATASETS**

We use post, Zoo and other decision table datasets in UCI library to carry out the experiment of incremental attribute reduction algorithm. 70% of the set is selected as the initial data and each time increasing by 10%. ICAR algorithm and IARAIV algorithm are used to reduce the attributes. The experimental results are shown in Table 13. The number of reduced attributes found by the two algorithms after adding new samples is basically identical.

In order to verify the efficiency of the algorithms, we select 70% of each dataset as initial data and add 5% samples each time. Table 13 also shows the processing time of the two algorithms after adding new samples to the datasets. The processing time of algorithm ICAR is less than that of algorithm IARAIV. The processing time of IARAIV algorithm increases more as the number of samples increases because the time complexity of ICAR algorithm is less than that of IARAIV algorithm.
TABLE 13. Comparison of decision table incremental reduction algorithms.

| dataset   | The initial data | No. of reduction attributes by ICAR | No. of reduction attributes by IARAIV |
|-----------|------------------|-------------------------------------|--------------------------------------|
|           | Time (s) | increase 10% | Time (s) | increase 20% | Time (s) | increase 30% | Time (s) | increase 10% | Time (s) | increase 20% | Time (s) | increase 30% |
| Post      | 4 1.1    | 4 1.4  | 3 1.7  | 2 1.3  | 2 1.5  | 2.3  | 2 1.4  | 2 1.3  | 2 1.5  | 2.3  | 2 1.4  | 2.3  |
| Zoo       | 6 1.7    | 7 2.1  | 6 2.5  | 5 3.0  | 7 1.8  | 6 2.2  | 6 2.7  | 5 3.5  | 7 1.8  | 6 2.2  | 6 2.7  | 5 3.5  |
| Spect     | 2 6.0    | 3 6.4  | 3 6.8  | 3 7.3  | 2 6.5  | 3 7.0  | 3 7.6  | 3 8.7  | 2 6.5  | 3 7.0  | 3 7.6  | 3 8.7  |
| Soybean   | 13 1.0   | 13 1.2 | 13 1.4 | 13 1.7 | 13 1.1 | 13 1.4 | 14 1.9 | 14 2.5 | 13 1.1 | 13 1.4 | 14 1.9 | 14 2.5 |
| Heart     | 5 13.5   | 4 14.8 | 4 16.1 | 3 17.9 | 4 15.6 | 5 16.3 | 3 18.2 | 3 22.5 | 4 15.6 | 5 16.3 | 3 18.2 | 3 22.5 |
| Nursery   | 7 40.2   | 8 42.7 | 8 45.2 | 8 48.3 | 7 48.6 | 8 51.2 | 8 56.4 | 8 63   | 7 48.6 | 8 51.2 | 8 56.4 | 8 63   |
| Connect   | 35 176   | 37 185 | 37 196 | 36 209 | 36 196 | 39 202 | 38 223 | 38 260 | 36 196 | 39 202 | 38 223 | 38 260 |
| Student   | 4 16     | 4 17.2 | 4 18.9 | 4 20.8 | 4 19   | 4 22   | 4 26   | 4 32   | 4 19   | 4 22   | 4 26   | 4 32   |
| Spambase  | 23 24    | 22 24.8 | 22 25.7 | 22 26.9 | 23 26  | 23 23  | 23 36  | 23 43  | 23 26  | 23 23  | 23 36  | 23 43  |
| Mush      | 4 22     | 4 22.7 | 4 23.6 | 4 24.8 | 4 24   | 4 28   | 4 35   | 4 42   | 4 24   | 4 28   | 4 35   | 4 42   |

VI. CONCLUSION

Attribute reduction is a challenging problem in areas such as pattern recognition, machine learning and data mining, especially for the big data scenarios. It aims to preserve the discriminatory power of original dataset. Based on studying the existing attribute reduction methods, we propose an attribute reduction method using Chi-square statistics and conditional information entropy as a heuristic function. The static attribute reduction methods are not suitable for the dynamic dataset. We design a method which can realize attribute reduction of dynamic dataset quickly. Experiments show that the method has good performance.

For future research, our work is to study how to update attribute reduction when objects and attributes in large-scale decision systems varies simultaneously over time by combining other excellent approaches, such as knowledge granularity [33, 34], compacting a decision table [22], positive region [35], based on ensemble structure analysis [36] and so on.

REFERENCES

[1] J. Yang, D. Zhang, A. F. Frangi, and I.-Y. Yang, “‘Two-dimensional PCA: A new approach to appearance-based face representation and recognition,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 1, pp. 131–137, Jan. 2004.
[2] P.-H. Wu, C.-C. Chen, J.-J. Ding, C.-Y. Hsu, and Y.-W. Huang, “Salient region detection improved by principle component analysis and boundary information,’’ IEEE Trans. Image Process., vol. 22, no. 9, pp. 3614–3624, Sep. 2013.
[3] H. Abdi and L. J. Williams, “Principal component analysis,” Wiley Interdiscip. Rev., Comput. Statist., vol. 2, no. 4, pp. 433–459, 2010.
[4] Y. Gong and X. Liu, “Video summarization using singular value decomposition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Hilton Head Island, SC, USA, Jun. 2000, pp. 174–180.
[5] G.-Y. Wang, J. Zhao, J.-J. An, and Y. Wu, “Theoretical study on attribute reduction of rough set theory: Comparison of algebra and information views,’’ in Proc. 3rd IEEE Int. Conf. Cognit. Informat., Victoria, BC, Canada, Aug. 2004, pp. 148–155.
[6] J. Xie, X. Shen, H. Liu, and X. Xu, “Research on an incremental attribute reduction algorithm by compacting a decision table,’’ Int. J. Mach. Learn. Cybern., vol. 10, no. 9, pp. 2355–2373, Sep. 2019.
[7] Y. Chuanjian, G. Hao, L. Longshu, and D. Jian, “A unified incremental attribute reduction based on elementary sets,’’ in Proc. 10th Int. Conf. Rough Sets, Fuzzy Sets, Data Mining Granular Comput. Heidelberg, Germany: Physica-Verlag, 2005, pp. 185–193.
[8] W. Wei, P. Song, J. Liaoy, and X. Xu, “Accelerating incremental attribute reduction with the variations of the object for decision tables,’’ Soft Comput., vol. 23, no. 15, pp. 6407–6427, Aug. 2019.
[9] J. Huang, Y. Cai, and X. Xu, “A hybrid genetic algorithm for feature selection wrapper based on mutual information,’’ Pattern Recognit. Lett., vol. 28, no. 13, pp. 1825–1844, Oct. 2007.
[26] G.-Y. Wang, H. Yu, and D.-C. Yang, “Decision table reduction based on conditional information entropy,” Chin. J. Comput., vol. 25, no. 7, pp. 759–766, 2002.

[27] B. Wang and S. B. Chen, “A complete algorithm for attribute reduction,” in Robotic Welding, Intelligence and Automation. Berlin, Germany: Springer-Verlag, 2004, pp. 345–352.

[28] Y. Yao and Y. Zhao, “Discernibility matrix simplification for constructing attribute reducts,” Inf. Sci., vol. 179, no. 7, pp. 867–882, Mar. 2009.

[29] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, “Bilevel feature extraction-based text mining for fault diagnosis of railway systems,” IEEE Trans. Intell. Transp. Syst., vol. 18, no. 1, pp. 49–58, Jan. 2017.

[30] H. Liu, M. Zhou, and Q. Liu, “An embedded feature selection method for imbalanced data classification,” IEE/CAA J. Autom. Sinica, vol. 6, no. 3, pp. 98–110, May 2019.

[31] N. N. Thuy and S. Wongthanavasu, “A new approach for reduction of attributes based on stripped quotient sets,” Pattern Recognit., vol. 97, Jan. 2020, Art. no. 106999, doi: 10.1016/j.patcog.2019.106999.

[32] Z. Cai and W. Zhu, “Feature selection for multi-label classification using neighborhood preservation,” IEE/CAA J. Autom. Sinica, vol. 5, no. 1, pp. 320–330, Jan. 2018.

[33] Y. Jing, T. Li, C. Luo, S.-J. Horng, G. Wang, and Z. Yu, “An incremental approach for attribute reduction based on knowledge granularity,” Knowl.-Based Syst., vol. 104, pp. 24–38, Jul. 2016.

[34] Y. Jing, T. Li, H. Fujita, Z. Yu, and B. Wang, “An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view,” Inf. Sci., vol. 411, pp. 23–38, Oct. 2017.

[35] X. Xie and X. Qin, “A novel incremental attribute reduction approach for dynamic incomplete decision systems,” Int. J. Approx. Reasoning, vol. 93, pp. 443–462, Feb. 2018.

[36] L. Shi, C. Tong, T. Lan, and X. Shi, “Statistical process monitoring based on ensemble structure analysis,” IEE/CAA J. Autom. Sinica, early access, Feb. 13, 2018, doi: 10.1109/JAS.2017.7510877.

NA SU received the Ph.D. degree in computer software and theory from the Shandong University of Science and Technology, Qingdao, China, in 2017. She is currently a Lecturer with the Department of Information Engineering, Shandong University of Science and Technology, Tai’an, China. Her current research interests include intelligent information processing and social network analysis.

XINJUN AN received the B.S. degree in computer software from the Shandong Institute of Mining, in 1997, and the M.S. degree in computer application technology from the Shandong University of Science and Technology, in 2007. His research interests include big data visualization, image feature extraction, and image stitching.

CHANGQING YAN received the bachelor’s degree in computer science and technology and the master’s degree in computer software and theory from the Shandong University of Science and Technology, Qingdao, China, in 2002 and 2005, respectively, and the Ph.D. degree in cartography and geographic information system from the State Key Laboratory of Resources and Environmental Information Systems, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing, China, in 2013. He is currently an Associate Professor with the Shandong University of Science and Technology, Tai’an, China. His research interests include support vector learning, surface modeling, and DEMs.

SHUJUAN JI received the B.Sc., M.Sc., and Ph.D. degrees in computer software and theory from the Shandong University of Science and Technology, Qingdao, China, in 1999, 2004, and 2007, respectively. She is currently an Associate Professor and a Doctoral Supervisor with the College of Computer Science and Engineering, Shandong University of Science and Technology. Her research interests include multiagent systems and machine learning. She is a Senior Member of the Chinese Association for Artificial Intelligence and the China Computer Federation. She is also a member of ACM.

---

VOLUME 8, 2020

98243