Discovering Deep Knowledge from Relational Data by Attribute-Value Association

Andrew K.C. Wong, Fellow, IEEE, Pei-Yuan Zhou, and Antonio Sze-To

Abstract— Discovering Attribute-Value Association (AVA) is of fundamental importance in knowledge discovery. Market Basket Analysis is an archetypical application. However, most existing algorithms rely only on frequency counts directly obtained from data at the surface and thus cannot reveal deeper knowledge, i.e. the AVAs governed by hidden factors inherent in the data. This paper proposes a new method, called Attribute-Value Association Algorithm (AVAA), which can i) discover statistically significant associations at the Attribute-Value level from relational dataset, ii) disentangle associations to unveil different AVAs corresponding to different hidden factors. The performance of AVAA is validated via experiments on both synthetic and real-world datasets. AVAA demonstrated better identification rate when comparing with Frequent Pattern Mining algorithms, particularly when noise was present.

Keywords—Segmentation/Clustering/Association; Explorative and visual data mining

I. INTRODUCTION
For decades, the most fundamental notion in extracting knowledge from relational dataset analysis is discovering significant Attribute-Value Association (AVA). For example, finding AVA in supermarket transactions is important for marketers. Frequent Pattern Mining [1], [2] is a famous method of this type. Today, as problems and data become more complex, the traditional support-confidence framework [1], [2], using frequency counts directly, cannot reveal deeper knowledge, that is the AVAs governed by hidden factors inherent in the data. Hence, we develop a novel algorithm known as Attribute-Value Association Algorithm (AVAA) to discover from the relational datasets AVAs governed by hidden factors. Given a relational dataset, AVAA: a) obtains an AVA Frequency Matrix (AVA FM) where each entry is a frequency of co-occurrences of an AVA of two attributes among all samples; b) obtains an AVA Statistical Residual Vector Space (a matrix abbreviated by SRV) by converting each frequency in the AVA FM into a statistical residual (SR) to account for the deviation of the observed frequency against the expected frequency if that AVA is a random happening; each row of the SRV represents an Attribute Value vector (a-vector) whose coordinates are the SRs of that AVA associating with other AVs represented by the column a-vectors; c) disentangles the SRV by means of Principle Component Decomposition (PCD) [3] and re-projects the a-vector projections on each PC to a new SRV, referred to as the (RSRV), for revealing deep knowledge. Deep knowledge here means the hidden and entangled AVAs, governed by different underlying factors, that cannot be revealed by traditional methods (such as Frequent Pattern Mining), but could be unveiled through the re-projection of the entangled AVAs in different transformed disentangled PC spaces. Thus, the objectives of this paper are to address: a) the theoretical formulation, b) the experimental justification via synthetic and real-world data and c) the comparison of AVAA with Frequent Pattern Mining methods [1] [2].

The novelty of AVAA is three-fold. 1) In a classical framework, PCD [3] is applied on sample vector level. In our study, AVAA goes deeply, beyond the sample vector level, to the attribute-value level. It also represents AVAs by their statistical weights to minimize the sensitivity of attribute scaling and to magnify the statistical strength through SR in different RSRVs. 2) Traditionally, PCD is applied for dimensionality reduction. AVAA applies PCD for the first time through disentangling SRV followed by re-projection to obtain RSRVs governed by different factors. That is, each RSRV contains AVAs associating to a hidden factor captured in its corresponding PC. 3) In a convention study involving PCD, the data analysis is mostly completed after PC projections. In this study, AVAA moves further to leverage the re-projection of PC a-vector projections to obtain AVA groups in different RSRVs corresponding to different hidden factors.

From our experiments on synthetic and real-world data, we demonstrated that: 1) multiple AVA subgroups made up of overlapping AVs (entangled groups) or a distinct AV set (multiple groups) could be identified and located in data samples by AVAA. AVAA outperformed Frequent Pattern Mining algorithms [1], [2] in identifying AVA subgroups, particularly when noise was present. 2) In dominating RSRVs, AVA subgroups revealed from relational datasets without class labels were almost identical to those from datasets with class labels, illustrating the effectiveness of AVAA in revealing class-related AVA subgroups. 3) When dominating factors like class labels were removed from the dataset, more obscure and subtle AVAs emerged in the less dominating RSRVs. All these indicate that deep knowledge could indeed be unveiled.

II. RELATED WORK
Throughout the literature, the work relating most to the discovery of AVAs are association rule mining [1], [2]. Hence, we would first review its basics, followed by its extension to discover hidden (or non-obvious) association.
The goal of association rule mining [1], [2] is to discover the association between items. Market Basket Analysis is an archetypical example. Transaction records of customers’ purchased items such as \{Beer, Bread, Milk, Chips\} can generate association rules such as \{Beer\} \Rightarrow \{Chips\}, which indicates that a customer who buys beer is likely to buy chips. Such discovered rules could help shopkeepers to boost sales by placing associated items together. By transforming an association rule into a binary representation \{Beer = 1\} \Rightarrow \{Chips = 1\}, we observe that association rules and AVAs are highly related.

Given a transaction database \(D\), association rule mining is to identify all the association rules with their support and confidence being no less than \(\sigma_{supp}\) and \(\sigma_{conf}\) respectively. The support of an itemset is the number of transactions that contain that itemset. The support of an association rule \(A \Rightarrow B\), is the support of \(A \cup B\). The confidence of an association rule \(A \Rightarrow B\), is the conditional probability that \(B\) occurs in a transaction, given that \(A\) occurs in it. There are efficient algorithms such as Apriori [4] and FP-Growth [5] proposed to solve this problem. In our experiments, we would compare with QuanMiner [6], a software package that supports mining both categorical and quantitative association rules.

Association rule mining has also been applied for mining hidden (or non-obvious) association. In applications such as fault and intrusion detection, the frequently occurring itemsets are not as interesting as the infrequently ones which are governed factors may be rarer, we could use lower SR thresholds to obtain deeper knowledge inherent in the data.

III. METHODOLOGY

Here we first provide the basic notations and definitions. As PCD [3] is a well-known methodology, we will not elaborate it in this paper. Fig. 1 gives a schematic overview of the AVAA to discover AVAs from a relational dataset with steps circled.

**Step 1:** Partition attributes with real values into bins. Construct an AVA FM from the AVA frequencies obtained from a relational dataset. **Step 2:** Convert the AVA FM into a SRV by replacing each AVA frequency with its statistical residual (SR) which accounts for the deviation of the frequency of the AVA occurrences if it is a random happening. We treat the matrix of SRs as a vector space, abbreviated by SRV, such that each row is taken as a row vector referred to as an Attribute-Value vector (a-vector). The coordinates of the a-vector represent the SR of the AVA of that AV represented by the a-vector with other AVs represented by the column a-vectors. **Step 3:** Obtain the PCs by applying PCD on the SRV and rank them according to their eigenvalues (Fig. 6(a-b)). Note that the significant projected a-vectors at greater distance from the mean are boxed in boxes with colored border (C-box). **Step 4:** Re-project the projections of all a-vectors on each PC back to an SRV, referred to as the Re-projected SRV (RSRV) (Fig. 6(c)) with a new transformed a-vector position rendering new SRs of AVAs with orthogonal functionality captured in the corresponding PC. The AVA SR of the distinct a-vectors are boxed in C-boxes (Fig. 6(a-c)). **Step 5:** From the RSRVs, we analyze the significant SR of AVA according to the confidence interval. Realizing that some minor factors may be rarer, we could use lower SR thresholds to obtain deeper knowledge inherent in the data.

**A. The Process of AVAA**

In the most general setting, a relational data set \(R\) is a set of \(N\) tuples of mixed-mode data with the following characteristics:

1. Every tuple in \(R\) is described by \(N\) attributes, denoted as \(\mathbf{A}_n = \{A_{1n}, ... , A_{nn}\}\).
2. Each attribute \(A_{ni}\) can assume a continuous numerical value or a categorical value.
   a. For a continuous attribute, \(A_{ni}\) is partitioned into \(I_n\) bins by transforming the original numerical values of \(A_{ni}\) into interval event values, denoted as \(A_{ni} = \{A^i_{ni} | i = 1, ..., I_n\}\). Since optimizing the partition is not the issue addressed in the paper, we adopt a simple one.
   b. For categorical attributes, \(A_{ni}\) contains \(I_{nr}\) values, denoted as \(A_{ni} = \{A^j_{ni} | j = 1, ..., I_{nr}\}\).

After transforming the mixed-mode dataset into an event-value dataset, all the values of an attribute \(A_{ni}\) can be denoted as \(E=A_{1n}/I_{nr}\), where \(E = \sum_{i=1}^{I_n} I_{ni}\) to represent the total number of the attribute’s event-values. We then develop a method to obtain the AVA FM.

**Definition 1 AVA Frequency Matrix.** Abbreviated as AVA FM, it is a matrix of frequency count of an AVA between two attribute values (AVs), say \(A^j_{ni}\) the \(j^{th}\) value of the \(n^{th}\) attribute and \(A^j_{nk}\), the \(j^{th}\) value of the \(n^{th}\) attribute (\(n\neq n'\)). It is denoted as \(FM(A^j_{ni} \leftrightarrow A^j_{nk})\).

To better visualize how AVA FM and its conversion into an SRV, a table format with \(I_{nr}\) rows and \(I_n\) columns (Table 1) is used, \(I_n\) is the total number of states of \(A_{ni}\) and \(I_{nr}\) is that of \(A_{nr}\).

To obtain statistically significant information from \(R\), we transform AVA FM into a Statistical Residual Matrix by
transforming each AVA frequency into a SR through Eqn. (2) on the pair of the observed frequency and its expected one (Table I). SR measures the deviation of the observed frequency of the AVA from its default model if it is a random association instead. In Table I, \( o_{ij} \) represents the total number of occurrences when \( A_m = A_{m'} \) and \( A_n = A_n' \); \( e_{ij} = \frac{1}{m} \sum_{o_{ij}} o_{ij} \sum_{o_{ij}} o_{ij} \) where \( \sum_{o_{ij}} o_{ij} \) represents the total number of counts when \( A_m = A_{m'} \); and \( \sum_{o_{ij}} o_{ij} \) represents that when \( A_n = A_n' \) and \( M \) is the number of records.

### Table I. Observed and Expected Frequencies of AVA

<table>
<thead>
<tr>
<th>( o_{ik} ) (( e_{ik} ))</th>
<th>( A_1 )</th>
<th>( A_2 )</th>
<th>( A_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_{m'} )</td>
<td>( a_{11} )</td>
<td>( a_{12} )</td>
<td>( a_{n1} )</td>
</tr>
<tr>
<td>( A_{n'} )</td>
<td>( a_{11} )</td>
<td>( a_{12} )</td>
<td>( a_{n1} )</td>
</tr>
<tr>
<td>( A_{n'} )</td>
<td>( a_{11} )</td>
<td>( a_{12} )</td>
<td>( a_{n1} )</td>
</tr>
<tr>
<td>( A_{n'} )</td>
<td>( a_{11} )</td>
<td>( a_{12} )</td>
<td>( a_{n1} )</td>
</tr>
<tr>
<td>( A_{n'} )</td>
<td>( a_{11} )</td>
<td>( a_{12} )</td>
<td>( a_{n1} )</td>
</tr>
<tr>
<td>( A_{n'} )</td>
<td>( a_{11} )</td>
<td>( a_{12} )</td>
<td>( a_{n1} )</td>
</tr>
</tbody>
</table>

By using the same example as above, given \( o_{ij} \) and \( e_{ij} \), we are interested in determining whether \( o_{ij} \) is significantly different from \( e_{ij} \). To scale the difference, we use the standardized residual SR as expressed below:

\[
SR(A_m' \leftrightarrow A_n') (SR_{ij}) = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}}
\] (2)

The SR value can reveal the statistical significance of an AVA. At the confidence level of 95%, if the SR > 1.96, the discovered AVA is “+ve significant”, colored yellow in the figures; “-ve significant” if the SR < -1.96, colored green, and irrelevant or random if the SR is between -1.96 and 1.96. In the less prominent RSRVs, the confidence interval could be lowered to 80% since the frequencies of their AVAs could be rarer.

Although the SR can reveal the significance of an AVA, subtle associations could still be entangled and masked. Hence, we seek a novel use of PCD to disentangle the AVA SR. Thus, we first consider the AVA SRV as a vector space by treating each row in it as a vector corresponding to its labeled AV as an a-vector such that the coordinates of which represent the SR of that AV associating with other AVs represented by the column a-vectors. Thus, the \( p \) coordinate of the \( p \)th row vector is the SR of the \( p \)th AV associated with the \( p \)th AV.

### Definition 2 AVA Standard Residual Vector Space.

Abbreviated as SRV, it is a vector space such that the \( p \) coordinate of its \( i \)th row vector \( SRV_{A_{i}} \) is the SR derived from the observed frequency that \( A_i \) is associating with all the other attributes values \( A_{i'} \). Hence, SRV can be expressed as a set of row vectors:

\[
SRV = \{SRV_{A_1}, ..., SRV_{A_i}, ..., SRV_{A_{i'}}, ..., SRV_{A_{n'}}\}
\]

where \( N \) is the total number of attributes, and \( l_i \) is the total event-values of attribute say \( A_i \) and \( I_n \) is the total event-values of attribute say \( A_n \). An a-vector is denoted as \( SRV_{A_{i}} = \{SR(A_i' \leftrightarrow A_1'), ..., SR(A_i' \leftrightarrow A_i'), ..., SR(A_i' \leftrightarrow A_{i'}), ..., SR(A_i' \leftrightarrow A_{n'}')\} \), where \( SR(A_i' \leftrightarrow A_i') \) represents the standard residual for AVA \( (A_i' \leftrightarrow A_i') \), and \( SR(A_i' \leftrightarrow A_{i'}) = 0 \).

After applying PCD on the SRV, we rank the Principal components (PCs) obtained according to their variance.

### Definition 3 Principal Components.

In PCD, the Projected Components are a set of \( k \) PCs, denoted as \( PC = \{PC_1, PC_2, ..., PC_k\} \), where \( PC_k \) is a set of projections of the a-vectors from SRV, denoted as \( PC_k = \{PC_k(A_{i'})|n = 1, 2, ..., N, i = 1, ..., I_n\} \), where \( N \) is the total number of all attributes and \( I_n \) is the number of states of attribute \( A_n \).

After PCD, we first identify the distinct projection(s) of the a-vectors and their clusters at a great distance from the mean of the PC-Axis. The coordinates of these a-vectors in significant PCs should have strong association or strong presence captured by the orthogonal PC Spaces of large variance. Their distinctiveness i.e. deviation from the mean, indicate certain strong and distinctive associations of the a-vectors with others. Hence, the new positions of their projections on the PC Axis when transformed back to the SRV represent the a-vectors with a new set of coordinates reflecting the SRs of their AVs with other AVs captured in that PC Space. We refer this SRV as the RSRV corresponding to that PC. We adopt a consistent notation, using \( RSRV_k \) correspond to \( PC_k \).

### Definition 4 Re-projected SRV.

Abbreviated as RSRV, it is the SRV containing the transformed position of a-vectors on the Principal Component Space as denoted by eqn (3).

\[
RSRV_k = SRV \cdot PC_k \cdot PC_k^T
\] (3)

We should also note that since in the disentangled RSRVs, the AVA corresponding to PC of lower eigenvalue could be rarer. Hence, lower confidence level are adopted to assess the less dominating yet distinct AVAs in the disentangled RSRVs, say between 95% and 80% as shown in Fig. 7(d).

To sum up, PCD uses an orthogonal transformation to transform a set of possible correlated variables into a set of linear uncorrelated variables known as PC. In the transformation, the first PC has the largest possible variance accounting largest variability in the data. Each succeeding PC in turn has the next highest variance under the constraint that it is orthogonal to the preceding PC. In our case, since each row a-vector represents an AV associating with other AVs as well as coordinate, the PC transformation will bring out in the PC the highest variance of the a-vectors with high SR coordinate values and display them at the far ends from the mean. At the surface, we may not see why an a-vector is significant. However, when we view it in the RSRV, we would find out the
coordinate(s) contributing to its high variance on the PC. In general, PCD is sensitive to the relative scaling of the original variables, masking their distinctiveness. Hence, by converting the AVA FM into SRV with uniform SR scale and statistical weights, AVAA utilizes the statistical strength and functional decomposition to reveal more stable, subtle yet significant associations that might be masked in the original frequency space. Hence, the AVAs discovered and disentangled are more distinct, stable and specific as manifested in the RSRVs.

B. Time Complexity Analysis

Here we analyze the time complexity of AVAA in two major parts: the computation of SRV and the PCD process. 1) SRV Computation. Assume that a N×M dataset is the input data, where N is the attribute number and M the record number, the time complexity is $O(N^2)$. Then, the original data is transformed into an nN×nN matrix, where nN is the number of attribute values. 2) PCD. It is applied to the SRV, an nN×nN matrix, to obtain eigenvalues and eigenvectors. PCD is composed of two steps. Step 1 is the computation of covariance matrix and its time complexity would be $O((nN)^3)$[8], where $nN$ is the number of attribute values. Step 2 is the eigenvector decomposition and its time complexity is $O((nN)^3)$ [8]. Hence, the overall complexity of AVAA would be $O(N^2 + (nN)^3)$.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this study, experiments on synthetic data were conducted to validate the proposed functionality of AVAA and to compare its AVA identification rate with its counterparts. To show that AVAA can handle both categorical and numerical data and reveal subtle yet explainable associations, we used a real-world clinical mixed-mode data taken from UCI [11] dataset.

A. Experimental Result of AVAA for synthetic dataset

The first set of synthetic data was designed to demonstrate how patterns (correlated AVA groups) entangled in a relational datasets, even within a subgroup of samples, could be disentangled to reveal different AVA groups; as well as how each entangled patterns could be located in the samples. The information and noise ratio of the data set is approximately 41%.

Fig. 2. Synthetic dataset with embedded patterns and entangled patterns

Fig. 3 gives several examples of the discovered patterns as displayed in the N-tuple of the relational data set. It illustrates how each correlated significant AVA group was able to reveal the governing and entangled patterns on the tuples as observed in row 3 and 4 of Fig. 3 respectively through the C-boxes with different colors.

Fig. 4. Statistically significant correlated AVA groups discovered. (a) AVA groups associating to C1 and C3 in RSRV1. (b) AVA group associating to class with given class label. (c) AVA group associate to classes without class labels.
different colors. This shows how AVAA could reveal multiple patterns inherent in the data to unveil the buried knowledge.

<table>
<thead>
<tr>
<th>Class</th>
<th>Raw AVA</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>A9</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>100</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>C2</td>
<td>250</td>
<td>B</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>C3</td>
<td>350</td>
<td>B</td>
<td>C</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>C4</td>
<td>450</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

Fig. 3. Embedded patterns as revealed in the samples in the relational dataset.

Fig. 4 provides several examples of our significant experimental results on this synthetic dataset. Fig. 4(a) and (b) show that AVA disentanglement did bring forth much succinct and distinct correlated AVA groups in different RSRVs (RSRV1 and RSRV2). We observed the most statistically significant AVAs such as those of C2 and C3 with strong (SR) in the C-boxes of different colors (red for C2 and purple for C3) in RSRV1. In Fig. 4(b) we found significant patterns associating with C3 (blue C-boxes) in RSRV2. RSRV2 brings out two diametric groups as revealed by the color cells. RSRV2 brings out only a single group. When class labels were given, we observed in Fig. 4(c) strong association with a group of AVAs with high SR in RSRV1. When class labels were not included in the dataset, AVAA was still able to bring out the AVA groups associating with different classes or subgroups as shown in Fig. 4(d) if the intrinsic association to subgroups exist. Note that the corresponding SR magnitudes in Fig. 4(c) and (d) are very close. It shows AVAA’s effectiveness in revealing deep knowledge and providing direct insight without reliance on explicit prior knowledge.

B. Comparison with Other Algorithms

To conduct the performance comparison between our methods (SRV and AVAA) and QuantMiner (Freq-1, Freq-2) [6], or simply Frequent Pattern Mining (FPM) algorithm, the following scheme is laid out. We first give a brief description of each methods.

1) QuantMiner [6] is an association rule mining system supporting even quantitative rules. By setting different support and confidence value different AVAs can be discovered. Thus, for Freq-1, we set the threshold as: \(\sigma_{supp} = 35\%\), \(\sigma_{conf} = 60\%\) (high); and for Freq-2, we set the threshold as: \(\sigma_{conf} = 10\%\), \(\sigma_{supp} = 40\%\) (low) where \(\sigma_{conf}(A_{nm} \rightarrow A_{mr}) = Pr(A_{mr} | A_{nm})\) and \(\sigma_{supp}(A_{nm} \rightarrow A_{mr}) = Pr(A_{mr} \cap A_{nm}) / Pr(A_{mr})\).

2) SR method uses only the standard residual (SR) values (Eqn.(2)) to assess the discovered AVAs between all AV pairs for significant AVA identification (with SR > 1.96).

3) AVAA uses the SRs in RSRVs to identify statistically distinct and significant AVAs. Due to the disentanglement, the statistics of the distangled AVAs are usually weaker. Hence, we set a rule with lower SR threshold for RSRVs corresponding to PCs with lower variance (say over 80% confidence intervals), as we notice that the SR values of significant AVAs still stand out prominently from the noise background.

For the purpose of performance evaluation of the synthetic experiments and due to that the AVAs are pre-inserted, an Identification Rate (IR) (eqn. 4) is used to evaluate how accurate are the correct inserted AVAs (include hidden/entangled ones) discovered. IR is defined as:

\[
IR = \frac{TP + TN}{N}
\]

where N is the total number of AVAs, TP is true positive rate which measures the number of positives of the correctly identified pre-defined (both strong and hidden) AVAs; and TN is the true negative rate, which measures the number of negatives that are correctly identified based on the pre-defined AVAs.

1. Synthetic Data Set 2

The second synthetic data set was designed to find out how well AVAA in discovering a small set of hidden AVAs while the majority of them are pertaining to the other class. We first generated a dataset containing 50 attributes \(a_1, a_2, ..., a_{50}\) with 500 records pertaining to two different classes \(C_1\) and \(C_2\), say representing normal and cancerous patients, 250 for each.

1. For \(C_1\) (Normal), we inserted 50 strong AVAs between the class label and all the other 50 attributes, say \(v_1, v_2, ..., v_{50}\). We thus assigned the value 0 or 1 as the AV say \(v_i\) associating with \(C_1\) through a pseudo-random number generator.

2. For \(C_2\) (Cancer), in order to make its AVs associating to \(C_2\) different from those associating to \(C_1\), we inserted 50 strong AVAs to the 50 attributes, say \(\sim v_1, \sim v_2, ..., \sim v_{50}\) such that if for \(C_1\) \(v_i = 0\), then for \(C_2\) \(v_i = 1\). For convenience, we use \(\sim v_i\) to represent \(i^{th}\) AV in the Cancer group corresponding to \(v_i\) in the Normal group.

3. We then added a small group, referred to as Special Normal Class, still labeled as \(C_1\). For this special class, we initialized the AVs associating with \(C_1\) (Normal). Then we inserted 34 AVs associating with \(C_2\) (Cancer) as hidden AVAs. The number of records of the special class was increased from 10% (50 records) to 100% (500 records).

With this set of synthetic dataset, we wanted to find out whether AVAA could discover the hidden AVAs from the anomaly. If we only considered the AVA between class labels and AVs, we could find 50 strong AVAs between \(C_1\) and the AVs (\(v_i\)); 50 strong AVAs between \(C_2\) and the AVs (\(\sim v_i\)); and 34 hidden AVAs between \(C_1\) and selected AVs (\(\sim v_i\)). The IR comparison result among different algorithms is given in Table II. From the experimental results, we could conclude:

1. When the inserted hidden AVAs are less than 30%, all the approaches can only detect the strong AVAs and when the inserted hidden AVAs increase into 40%, only AVAA can detect both strong and hidden AVAs.

AVAA outperformed other algorithms especially when the hidden AVAs are inserted as 40%, 50% and 60%. Thus, AVAA is more powerful for discovering rare patterns even when only 40% of them were inserted.

2. The performance of FP (Freq-1 and Freq-2) depended on the thresholds set. The high threshold could guarantee the significance of the discovered AVAs, yet difficult to
distinguish hidden AVAs from strong AVAs when more than 70% hidden AVAs are inserted.

3. For the low threshold setting, FPM could reveal strong patterns first. But only when more records containing hidden AVAs were added (until 70% records in $C_1$ containing hidden AVAs), all hidden AVAs were revealed using FPM.

4. AVAA could discover more hidden patterns corresponding to different functional groups in different RSRVs.

We still compare the IR result using 1) FP with different thresholds; 2) SR and 3) AVAA, to test if AVAA could still discover both strong AVAs and hidden AVAs originally embedded in the data when noise was added at different levels (50% to 200% in Table III). We can conclude from the results that AVAA can handle noise data better than FP and SR vectors.

### Table II. Comparison Experimental Results for Synthetic Dataset

<table>
<thead>
<tr>
<th>Hidden AVAs</th>
<th>Freq-1 (high)</th>
<th>Freq-2 (low)</th>
<th>SRV</th>
<th>AVAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>20%</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>30%</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>40%</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>50%</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>60%</td>
<td>0.66</td>
<td>1.00</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>70%</td>
<td>0.66</td>
<td>1.00</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>80%</td>
<td>0.66</td>
<td>1.00</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>90%</td>
<td>0.41</td>
<td>1.00</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>100%</td>
<td>0.41</td>
<td>1.00</td>
<td>0.83</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### Table III. Experimental Results for Synthetic Dataset with Noise

<table>
<thead>
<tr>
<th>Level of Noise</th>
<th>Freq-1 (high)</th>
<th>Freq-2 (low)</th>
<th>SRV</th>
<th>AVAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>no noise</td>
<td>0.66</td>
<td>1.00</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>50%</td>
<td>0.51</td>
<td>0.97</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>100%</td>
<td>0.45</td>
<td>0.96</td>
<td>0.83</td>
<td>0.99</td>
</tr>
<tr>
<td>150%</td>
<td>0.43</td>
<td>0.85</td>
<td>0.83</td>
<td>0.97</td>
</tr>
<tr>
<td>200%</td>
<td>0.42</td>
<td>0.84</td>
<td>0.79</td>
<td>0.95</td>
</tr>
</tbody>
</table>

C. Experiment on Clinical Data.

We used the Statlog (Heart) Data taken from UCI [11] to validate and evaluate the performance of AVAA. It is a mixed-mode relational dataset described in Fig. 5. We first included class labels in the AVA FM. For clinical data used for diagnosis, we anticipated strong dominating a-vectors on opposite sides of the PC with largest variance such as PC1 (Fig. 6(a) and (b)) with strong AVAs in both SRV and RSRV1. We did observe strong SR associating with both C1 and C2 (see the red and blue C-boxes Fig. 6(c)). Note that the SR of the same AV for the two different sets of diametrically different a-vectors for C1 and C2 are with opposite (positive and negative) statistical significance (green and yellow cells). We also noticed that when AVAs were dominating by class association with low noise level, we did not find significant AVAs in other RSRVs as most of the SR values are below 1.00. However, when class labels were not included, minor disentangled subtle patterns began to emerge from other RSRVs (Fig. 7 (c)/(d) with strong/weak variance respectively).

The latter problem is equivalent to using unsupervised methods in revealing subgroup AVA patterns without relying on explicit prior knowledge. From the corresponding part of SRV (Fig. 7(a)), RSRV1(Fig. 7(b)) with class association, and RSRV1(Fig. 7(c)) RSRV5 (Fig.6(d) without class association, we observed that the first three consist of strong SR (>1.96) associating to C1 and C2, though the SR of in Fig.6(c) is a little lower. This indicates that without explicit prior knowledge, AVAA was able to discover the AVAs (Fig.7(c)) somewhat equivalent to pattern clustering [12]. Results in Fig. 7(c) clearly show that without relying on class labels, AVAA was still able to come up with an almost equivalent set of AVAs as those in Fig. 7(b) when class labels were given. The most interesting observation is that the disentangled AVAs in some minor PCs only distinctively emerged after strong associating factors like the class labels were removed as observed in RSRV5 (Fig. 7(d)) and RSRV2 to
RSRV4 not shown in the paper. Their SRs were still distinct yet weaker since they revealed the non-dominating yet rarer happenings. When comparing Fig. 7(a) and (b) with Fig 7(d), we observed that the distinct SRs in Fig. 7(d) are different from those in the others, though the values are a little bit low (with five over 1.28 and one over 1.03 corresponding to above 80% and 70% confidence level respectively). Nevertheless, they are still distinct from the background. From the C-boxes in Fig. 6(d), we noticed that they were not directly related to Classes. This may indicate that they might be governed by some not readily known underlying factors that worth the inspection by experts. We refer such revealing AVAs and their groupings, unveiled in different RSRVs corresponding to different orthogonal PCs, the deeper knowledge that have been obscured and deeply buried in the relational data. It will shed light to our understanding of the subtle factors and provide a statistical base to account for new knowledge in factor and predictive analysis.

V. CONCLUSION

Discovering knowledge from data has broad applications. Today, as data become more complex, the traditional support-confidence framework cannot easily reveal patterns/association rules governed by hidden factors. This paper presents a new method known as Attribute-Value Association Algorithm. AVAA discovers from relational datasets the most fundamental associations, groups them in a multi-facet manner according to the underlying functional relations, and disentangles and magnifies the associations in different statistical residual spaces. Our experiments validated that AVAA could reveal different AVA groups in different factor spaces. It renders better class/subgroup identification rate when comparing with its counterparts. It demonstrates its generic knowledge revealing capability from experimental results on mixed-mode dataset with or without class labels. In the future, we believe that AVAA, with its flexibility and generality, has strong potential in making impact in different fields.

VI. REFERENCE