Mining Strongly Correlated Intervals with Hypergraphs

Hao Wang[1](✉), Dejing Dou[1], Yue Fang[2], and Yongli Zhang[2]

[1] Computer and Information Science, University of Oregon, Eugene, OR, USA
{csehao,dou}@cs.uoregon.edu

[2] Department of Decision Science, University of Oregon, Eugene, OR, USA
{yfang,yongli}@uoregon.edu

Abstract. Correlation is an important statistical measure for estimating dependencies between numerical attributes in multivariate datasets. Previous correlation discovery algorithms mostly dedicate to find piecewise correlations between the attributes. Other research efforts, such as correlation preserving discretization, can find strongly correlated intervals through a discretization process while preserving correlation. However, discretization based methods suffer from some fundamental problems, such as information loss and crisp boundary. In this paper, we propose a novel method to discover strongly correlated intervals from numerical datasets without using discretization. We propose a hypergraph model to capture the underlying correlation structure in multivariate numerical data and a corresponding algorithm to discover strongly correlated intervals from the hypergraph model. Strongly correlated intervals can be found even when the corresponding attributes are less or not correlated. Experiment results from a health social network dataset show the effectiveness of our algorithm.

1 Introduction

Correlation is a widely used statistic measure for mining dependencies in multivariate data sets. Its value typically reflexes the degree of covariance and contravariance relationships in numerical data. Previous data mining algorithms focus on discovering attribute sets with high piecewise correlations between attributes. Such correlation measure shows a high level picture of the dependency profile in data, nevertheless, they can only reveal the correlations of numerical attributes in the scope of full range. The numerical data themselves are often considered as containing richer information than just the high level attribute-wise correlations. For example, in the meteorology data, the rate of precipitation is more positively (or negatively) correlated with humidity (or air pressure) especially when the humidity is large enough (e.g., when humidity $\geq$ 80 %).

In this paper, we address the problem of discovering the intervals with strong correlations from numerical data. The patterns discovered are in the form of interval sets, for example, "*Humidity*[*20 %, 30 %*], *Precipitation*[*70 %, 90 %*], *Correlation 0.81*". The correlation of the intervals, in this example, 0.81, is

Q. Chen et al. (Eds.): DEXA 2015, Part II, LNCS 9262, pp. 339–348, 2015.
DOI: 10.1007/978-3-319-22852-5_29

calculated by the correlation of all data instances that fall inside of the ranges of intervals. The strongly correlated intervals would provide us valuable insights with more detail dependencies hidden in the data. For example, in the financial market data, the demands of stocks and bonds generally raise as the prices fall. This market principle of the price and demand is the corner stone of a stable financial market. However, such principle might not hold under certain circumstances, such as a potential economic crisis. When the prices of stocks fall below certain thresholds and a crash in the stock market is triggered, the demand would fall along with the decline of the price for a certain price range. In this example, the strong correlated intervals from historical transaction data will provide investors with useful insights on when and how to avoid the risks in the financial investment.

Discretization [6] is one of the intuitive ways to generate correlated intervals on numerical attributes. Mehta et al. [9] proposed an unsupervised discretization method which preserves the underlying correlation information during the discretization process. It can serve as a preprocessing step for further data mining applications such as classification and frequent itemset mining. While it can discover strongly correlated intervals, there are some fundamental problems for discretization. For example, the *crisp boundary problem* [5] forces the discretization boundaries to make trade-offs between adjacent intervals on all attributes. Information in data may lose during the discretization as well. With regarding to the size of intervals, we usually face a dilemma to decide the quantity of segmentations. More segmentations means less information loss during the discretization process, while less segmentations will lead to large intervals that strong correlations between small intervals cannot be discovered.

In this paper, we propose a novel method to discover the strongly correlated intervals without suffering from the problems in the discretization based methods. We propose to model the numerical data with a hypergraph representation and use the average commute time distances to capture the underlying correlations. We propose a corresponding algorithm to discover the intervals with high piecewise correlations. One strength of our algorithm is that the discovery of intervals and their correlation optimization are achieved in a single step. Each boundary of the intervals are optimized independently. Therefore, they would not suffer from the crisp boundary problem or information loss problem.

The correlation measure we use in this paper is the Spearman's rank correlation coefficient [4]. The Spearman's rank correlation coefficient is defined as:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}},$$

in which $\bar{x}$ and $\bar{y}$ stand for the average order ranks of attribute x and y respectively. Based on the Spearman's correlation coefficient, we also propose the correlation gain and normalized correlation to evaluate our approach from different perspectives.

The correlation gain is defined as the ratio between the correlation of the intervals and the correlation of the related attributes,

$$\rho_{gain} = \frac{\rho}{\rho_{att}},$$

in which ρ_{att} is the correlation between related attributes. High correlation gain implies that even though the two attributes are less correlated in general, strong correlations might still be found between intervals on the attributes. The intervals with high correlation gain are valuable because it can reveal the strongly correlated intervals that hide under the less correlated attributes.

The normalized correlation is an estimation of correlation which depends on both the underlying correlation priori and the number of data instances available. Statistically, the estimation deviation has a negative relationship with the number of instances [4]. For example, for a dataset with only 2 data instances the correlation is always ±1 with infinite deviation, while for a dataset with infinite number of instances, the estimation converges to the true correlation value with 0 deviation. In this paper we also introduce normalized correlation ρ_{norm} [11] to generate strongly correlated intervals. It is defined as

$$\rho_{norm} = \rho\sqrt{n},$$

in which ρ is the correlation estimation and n is the number of data instances used for the estimation. The normalized correlation makes trade off between the estimated value and the estimation accuracy. The intervals with too less data instances will result in low ρ_{norm} due to the large deviation and so does the intervals with many data instances such as the full range intervals due to lower correlation value. It also relieves the effort to pre-define a threshold for selecting certain intervals as “strongly correlated.”

Our main contributions in this paper are:

- We propose a hypergraph random walk model to capture the underlying correlation of the numerical data. This model can capture the correlation relationship at the interval level rather than at the attribute level using a measure based on average commute time distance.
- We propose an algorithm to discover strongly correlated intervals based on the hypergraph random walk model. We are able to discover strongly correlated intervals with high accuracy without suffering from the information loss and crisp boundary problem in the discretization based methods.
- We propose the normalized correlation to generate strongly correlated intervals without a pre-defined threshold. We also propose the correlation gain to find the highly correlated intervals even the corresponding attributes are less correlated.
- We conduct experiments in a health social network dataset and the results show the effectiveness of our algorithm.

The rest of this paper is organized as follows: we give a brief introduction of related works in Sect. 2. We make a detailed description of our method in Sect. 3. We report experiment results in Sect. 4. We conclude the paper in Sect. 5.

2 Related Work

Previous research efforts have proposed various methods for the discovery of correlated intervals. Discretization is the one of the most straight forward way to generate intervals for various optimization goals. Kotsiantis and Kanellopoulos [6] made a thorough survey of the discretization techniques. These discretization methods target different optimization goals, such as minimizing the information loss, maximizing the entropy etc. Mehta et al. [9] proposed a PCA based unsupervised discretization method which can preserve the underlying correlation structure during the discretization process. The discretization process serves as an independent process which can be fed into many other data mining tasks such as association mining and clustering.

Quantitative association mining is another technique related to discovering strongly correlated intervals. The quantitative association mining is an extension of the traditional association mining. It generates intervals on the numerical data instead of the categorical data. Srikant and Agrawal [10] first proposed an algorithm that deals with numerical attributes by discretizing numerical data into categorical data. Fukuda et al. [12,13] proposed several methods that either maximize the support with pre-defined confidence or maximize the confidence with pre-defined support by merging up adjacent instance buckets. However, the support and confidence measures were attested to be inadequate to discover strongly correlated intervals due to the *catch-22* problem [10] and the crisp boundary problem [2] as well.

3 Discovering Strongly Correlated Intervals with Hypergraphs

In this section, we present our hypergraph based method which can efficiently discover strongly correlated intervals. We propose a hypergraph model to represent the correlation structure of numerical data. The correlation measure is captured by the average commute time distance between vertices in the hypergraph model.

3.1 Hypergraph and Average Commute Time Distance

Hypergraph is a generalization of regular graph that each edge is able to incident with more than two vertices. It is usually represented by $G = (V, E, W)$, in which V, E and W are the set of vertices, edges and weights assigned to corresponding edges respectively. The incident matrix of a hypergraph G is defined by H in which

$$H(v, e) = \begin{cases} 1 \, if \;\; v \in e \\ 0 \, if \;\; v \notin e \end{cases} \tag{1}$$

Zhou et al. [14] generalized the random walk model on hypergraph and defined the average commute time similarity S_{ct} and the Laplacian similarity S_{L_+}. The

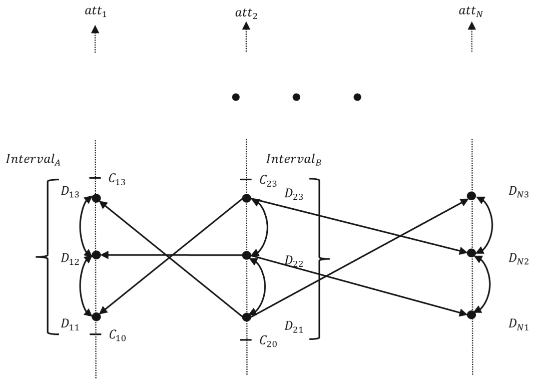

Fig. 1. Hypergraph representation of numerical data

average commute time similarity $n(i, j)$ is defined by

$$n(i,j) = V_G(l_{ii}^{+} + l_{jj}^{+} - 2l_{ij}^{+}), \tag{2}$$

where l_{ij}^{+} is the ith and jth element of matrix $\mathbf{L}^{+}$, $\mathbf{L}$ is the hypergraph Laplacian:

$$\mathbf{L} = \mathbf{D}_v - \mathbf{HWD}_e^{-1}\mathbf{H}^T, \tag{3}$$

and $\{.\}^{+}$ stand for Moore-Penrose pseudoinverse. D_v and D_e denote the diagonal matrix containing the degree of vertices and edges respectively. $V_G = tr(\mathbf{D_v})$ is the volume of hypergraph. The average commute time distance is defined by the inversion of normalized average commute time similarity [7]. As mentioned in [8] the commute-time distance $n(i, j)$ between two node i and j has the desirable property of decreasing when the number of paths connecting the two nodes increases. This intuitively satisfies the property of the effective resistance of the equivalent electrical network [3].

3.2 Hypergraph Representation of Numerical Data

As illustrated in Fig. 1, it shows an example of numerical data with values sorted in the ascending order on attributes. Let $A = \{att_1,\ att_2,\ ...,\ att_M\}$ be the set of attributes, in which M is the number of attributes. The jth value of the ith attribute att_i is denoted as D_{ij}. The boundary candidates $c_{i1},\ c_{i2},\ ...\ ,\ c_{i(N-1)}$ are the averages of each two adjacent values on the corresponding attribute att_i. The boundaries of intervals are defined on these boundary candidates. The set of attributes, instances and intervals are denoted as S_{att}, S_{inst} and S_{inter} respectively. Based on the sorted numerical data, we further build the data representation with a hypergraph model. Each data value D_{ij} corresponds to a vertex in the hypergrah. Each pair of vertices with adjacent values, D_{ij} and $D_{i(j+1)}$, are connected through a hyperedge with a weight proportion to the

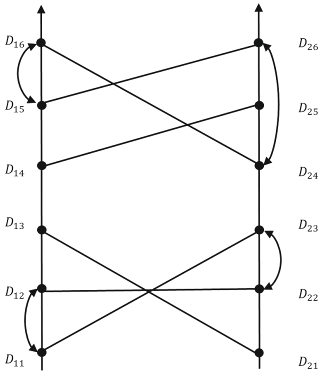

Fig. 2. Relationship between correlation and average commute time distance

inversion of the distance between them. The vertices in the same data instance with value D_{i1}, D_{i2},..., D_{iM} are connected by a hyperedge as well.

As shown in Fig. 2, the data instances below are strongly negative correlated and the data instances above are less strongly correlated. The non direct random walk paths on strongly correlated data instances, for example, $D_{11} \rightarrow D_{23} \rightarrow D_{22} \rightarrow D_{12}$, is shorter than the path on loosely correlated data instances, for example, $D_{15} \rightarrow D_{26} \rightarrow D_{24} \rightarrow D_{16}$, because the corresponding data values on the other correlated attributes are closer as well. In this case, the average commute time distance between the strongly correlated vertices is relatively shorter than the not strongly correlated vertices. For the reason above, the average commute time distance from random walk model is capable of capturing the correlation measures in our problem.

3.3 Algorithm Description

Based on the hypergraph model, we propose an algorithm for discovering strongly correlated intervals in numerical data. The pseudo code of our correlated interval discovery algorithm is shown in Algorithm 1. The algorithm first builds up the hypergraph model as described in Sect. 3.1. Adjacency matrix H is built up for the hypergraph. In Function Interval_Set_Discovery, the Laplacian matrix is computed from Eq. 3. The average commute time distance matrix is generated with entry i, j according to Formula 2. The distance between the adjacent data instances are the inversion of the average commute time similarity. In Function Merge_Interval, a bottom up mining process is applied simultaneously on all the attributes. For each attribute, an interval i_k is initialized with boundary $[c_i, c_{i+1}]$ for node n_k, i.e., exactly one node for each interval. A distance matrix C_i is maintained for each attribute, the distances between intervals are initialized as the average commute time distance for the node corresponding to this interval. In each iteration, for each attribute, the algorithm looks up the minimum distance between the adjacent intervals i_k and i_{k+1}, then merges these

Algorithm 1. Correlated Interval Discovery Algorithm

```
**Function**: Interval Set Discovery
**Initialize**: $\mathbf{C} = \mathbb{R}_{M \times M}$
**for** i, j = 1 to M **do**
  $\mathbf{C}_{i,j} = V_G(l^+_{ii} + l^+_{jj} - 2l^+_{ij})$
**end for**
**while** mergeable **do**
  **for** i, j = 1 to M **do**
    Merge Interval
  **end for**
**end while**

**Function**: Merge Interval
**Initialize**: $\boldsymbol{D} = \boldsymbol{C}$
**while** intervals generated **do**
  $k = min\_index(D_{k,\ k+1},\ k \in 1,\ ...,\ M-1)$
  $i_k$ = Merge($\boldsymbol{i}_k$, $\boldsymbol{i}_{k+1}$)
  **for** k = 1 to M-1 **do**
    $D_{i,k} = D_{i,k} + D_{i+1,k}$
    $D_{k,i} = D_{k,1} + D_{k,i+1}$
  **end for**
**end while**
```

two intervals. The distance matrix is updated accordingly. In every few iterations, for each generated interval, we scan the corresponding intervals on the rest attributes. The normalized correlation is calculated for these pairs of intervals. The intervals are used to update the final result, and only the top k correlated interval sets with best normalized correlations are kept in set. If the correlation metrics of certain intervals are above the user pre-defined threshold, such as the $Interval_A$ and $Interval_B$ in Fig. 1, then the two intervals are combined into one attribute/interval set $\{S_{att}, S_{inter}\}$ and an interval set is generated. The mining process continues to generate interval sets till intervals on all attributes merges into full ranges.

4 Experimental Results

We evaluate our method on a real life health social network dataset. SMASH [1] is the abbreviation of Semantic Mining of Activity, Social, and Health Data Project. The dataset collected in this project include social connections and relations, physical activities, and biomarkers from 265 overweight or obese human subjects. After preprocessing, the input data in our experiment contain the following attributes for the physical activities and biomarkers. The physical activity indicator *Ratio No.Steps* is the ratio of steps that a human subject walked through in two consecutive periods of time. Three biomarkers *HDL*, *LDL* and *BMI* are used for the health condition indicators. The *HDL* and *LDL* stand for the high density lipoprotein and low density lipoprotein respectively. The rate of

Table 1. Experiment results from the SMASH data

(a) Top Five Rules from Correlation Preserving Discretization

Attribute Set	Correlation	Correlation Gain	Normalized Correlation
Ratio LDL[0.74, 0.85] Ratio No.Steps[0.32, 0.89]	0.62	13.70	1.68
Ratio BMI[0.93, 1.01] Ratio No.Steps[0.89, 1.21]	0.82	5.13	1.31
Ratio LDL[0.99, 1.09] Ratio BMI[0.93, 1.01]	0.81	2.86	0.73
Ratio BMI[1.01, 1.12] Ratio LDL[0.99, 1.09]	0.78	2.75	1.16
Ratio BMI[1.01, 1.12] Ratio LDL[1.09, 1.31]	0.64	2.25	1.34

(b) Top Five Rules from Hypergraph Based Method Ranked by Correlation Gain

Attribute Set	Correlation	Correlation Gain	Normalized Correlation
Ratio HDL[0.97, 1.26] Ratio BMI[0.89, 1.00]	0.94	45.22	1.67
Ratio HDL[0.79, 0.91] Ratio No.Steps[0.78, 3.94]	0.57	28.15	2.51
Ratio LDL[1.00, 1.02] Ratio No.Steps[0.39, 3.82]	1.00	22.11	2.01
Ratio HDL[0.78, 1.25] Ratio No.Steps[0.27, 3.96]	0.37	18.41	3.85
Ratio LDL[0.88, 1.02] Ratio No.Steps[0.01, 3.82]	0.72	16.12	1.65

(c) Top Five Rules Ranked by Normalized Correlation

Attribute Set	Correlation	Correlation Gain	Normalized Correlation
Ratio LDL[0.70,1.00] Ratio HDL[0.86,1.00]	0.55	2.01	5.75
Ratio LDL[1.00,1.18] Ratio HDL[1.00,1.14]	0.53	1.93	4.99
Ratio LDL[0.69,0.93] Ratio BMI[0.97,0.99]	0.63	21.01	4.14
Ratio LDL[0.92,1.05] Ratio No.Steps[1.17,2.31]	0.56	10.68	3.87
Ratio HDL[1.02,1.08] Ratio No.Steps[1.20,1.91]	0.79	194.41	3.71

HDL usually relates with decreasing rate of heart related disease and the reverse case for *LDL*. The *BMI* stands for body mass index which is a common indicator of the obesity level.

In Table 1 we list our experiment results from the SMASH dataset. The interval sets in Table 1(a) and 1(b) are results from the correlation preserving discretization [9] and our hypergraph based method respectively. Comparing the results in the two tables, our algorithm not only returns intervals with higher correlations, but also higher correlation gains. Note that the interval set discovered by our algorithm has the ability to overlap with each other. For example, the second and third interval sets in Table 1(b) "*Ratio HDL*[*0.79, 0.91*], *Ratio No.Steps*[*0.78, 3.94*]" and "*Ratio LDL*[*1.00, 1.02*], *Ratio No.Steps*[*0.39, 3.82*]," the intervals on *Ratio No.Steps* has a large overlapping between 0.78 and 3.82. On the contrary, the interval set found by correlation preserving discretization clearly suffers from the *crisp boundary problem*. Note the boundaries of intervals on attribute *Ratio BMI* in the last four interval sets in Table 1(a) only have two choices, *Ratio BMI*[*0.93, 1.01*] and *Ratio BMI*[*1.01, 1.12*]. The decision of the boundaries on each attribute has to take into consideration of the correlations with all other attributes. The trade-off in the discretization methods makes them hard to make an optimization for every interval set discovered in the data. Therefore, intervals found by the correlation preserving discretization method are suboptimal.

Strongly correlated intervals provide us interesting information with regarding to relationships between the health condition of cardiovascular system and obesity. As we mentioned before, as a good health indicator, *HDL* is usually expected to have a strong correlation with other health condition factors such as *BMI*. However, the correlation between the two attributes, *HDL* and *BMI*, will be mostly ignored because the correlation is not large enough to raise any attention. The interval sets discovered by our algorithm show that, when *Ratio BMI* changes under the moderate range, say close to 1.00, the correlations are much larger than in the rest conditions. It indicates that with regarding to the weight variations, no matter increasing or losing weight, the first few pounds might be ones that affect the health condition most. On the other hand, it also indicates that drastic exercises with a rapid weight losing rate will be likely, on the contrary, result in a deterioration of cardiovascular health condition.

Table 1(c) shows the results when we use normalized correlation as the correlation measure. Note that the last interval set in Table 1(c) which has fairly high correlation gain does not show up in Table 1(a) and 1(b). This interval set is not discovered in the first two experiments because the sizes of intervals are not above the user defined threshold. The normalized correlation renders us the potential to find the intervals that is below the user defined threshold. This indicates the fact that even the situation is rare, *HDL* and *Ration No.Steps* has a positive correlation when the amount of exercise increases drastically. Although generally exercises do not make drastic changes on *HDL*, in the situation when the subject changes the amount of exercises drastically, such as at the beginning of weight reducing program, it will result in a greater change of *HDL*.

5 Conclusions

We present a novel algorithm for discovering strongly correlated intervals from numerical data. Previous research either dedicates to discover the correlated attribute sets from the full ranges of data or uses discretization methods to transform the numerical data before the pattern discovery. These methods, however, suffer from the information loss or crisp boundary problems. The method we proposed in this paper can discover strongly correlated intervals from the less correlated attributes. These discovered intervals are not only strongly correlated but also have independently optimized boundaries with regarding to the correlation measures. Experiment results in a health social network dataset show the effectiveness of our method.

Acknowledgment. This work is supported by the NIH grant R01GM103309. We thank Brigitte Piniewski and David Kil for their input.

References

1. Semantic Mining of Activity, Social, and Health data, AIMLAB, University of Oregon. http://aimlab.cs.uoregon.edu/smash/. Accessed June 2015
2. Aumann, Y., Lindell, Y.: A statistical theory for quantitative association rules. In: ACM International Conference on Knowledge Discovery and Data Mining, pp. 261–270 (1999)
3. Doyle, P., Snell, L.: Random walks and electric networks. Appl. Math. Comput. **10**, 12 (1984)

4. Freedman, D., Pisani, R., Purves, R.: Statistics. W.W. Norton & Company, New York (2007)
5. Ishibuchi, H., Yamamoto, T., Nakashima, T.: Fuzzy data mining: effect of fuzzy discretization. In: IEEE International conference on Data Mining, pp. 241–248 (2001)
6. Kotsiantis, S., Kanellopoulos, D.: Discretization techniques: a recent survey. Int. Trans. Comput. Sci. Eng. **32**(1), 47–58 (2006)
7. Liu, H., LePendu, P., Jin, R., Dou, D.: A hypergraph-based method for discovering semantically associated itemsets. In: IEEE International Conference on Data Mining, pp. 398–406 (2011)
8. Lovász, L.: Random walks on graphs: a survey. Combinatorics, Paul erdos is eighty **2**(1), 1–46 (1993)
9. Mehta, S., Parthasarathy, S., Yang, H.: Toward unsupervised correlation preserving discretization. IEEE Trans. Knowl. Data Eng. **17**(9), 1174–1185 (2005)
10. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: ACM International Conference on Management of Data, pp. 1–12 (1996)
11. Struc, V., Pavesic, N.: The corrected normalized correlation coefficient: a novel way of matching score calculation for lda-based face verification. In: International Conference on Fuzzy Systems and Knowledge Discovery, pp. 110–115 (2008)
12. Takeshi, F., Yasuhido, M., Shinichi, M., Takeshi, T.: Mining optimized association rules for numeric attributes. In: Symposium on Principles of Database Systems, pp. 182–191 (1996)
13. Takeshi, F., Yasukiko, M., Shinichi, M., Takeshi, T.: Data mining using two-dimensional optimized association rules: scheme, algorithms, and visualization. In: ACM International Conference on Management of Data, pp. 13–23 (1996)
14. Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: clustering, classification, and embedding. In: Advances in Neural Information Processing Systems, pp. 1601–1608 (2007)