To appear in Knowledge-Based Systems

Dimensionality Reduction via Discretization

Huan Liu *and* Rudy Setiono

Department of Information Systems and Computer Science

National University of Singapore

Kent Ridge, Singapore 0511

Tel: (+65)772-6563 Fax: (+65)779-4580

{liuh,rudys}@iscs.nus.sg

Abstract

The existence of numeric data and large amounts of records in a database pose a challenging task to explicit concepts extraction from the raw data. This paper introduces a method that reduces data vertically and horizontally, keeps the discriminating power of the original data, and paves the way for extracting concepts. The method is based on discretization (vertical reduction) and feature selection (horizontal reduction). The experimental results show that (1) the data can be effectively reduced by the proposed method; (2) the predictive accuracy of a classifier (C4.5) can be improved after data and dimensionality reduction; and (3) the classification rules learned are simpler.

Key Word: Dimensionality Reduction, Discretization, Knowledge Discovery

1 Introduction

The wide use of computers brings forth proliferation of databases. Without the aid of computer, little of this flood of raw data will ever be seen and exploited by humans. Knowledge discovery systems in databases [1] are designed to analyze the data, find regularities in the data (knowledge), and present it to human with understandable formats.

One of the goals of knowledge discovery in databases is to extract explicit concepts from the raw data [2, 1, 3, 4, 5]. The existence of numeric data and large amounts of records in a database pose a challenging task toward this goal due to the huge data space determined by numeric attributes. This paper introduces a method that reduces numeric data vertically and horizontally, keeps the discriminating power of the original data, and paves the way for extracting concepts. The method is based on discretization (vertical reduction) and feature selection (horizontal reduction). The χ^2 statistic is employed to continue discretizing the numeric attributes until the original discriminating power of the data cannot be maintained. This step significantly reduces the possible data space from continuum to discreteness according to the characteristics of the data by merging attribute values. In addition, after discretization, duplicates may occur in the data. Removing these duplicates amounts to reducing the amount of data. Hence, the original database, if viewed as a large table, is shorten in its vertical dimension. Feature (attribute) selection is a process to choose relevant attributes among many for a certain problem [6, 7, 8, 9]. The selection is accomplished by retaining those attributes having more than one discrete values. Other attributes can be removed. Both discretization and feature selection retain the discriminating power of the processed data.

A data and dimensionality reduction (DDR) system is built according to the vertical and horizontal reduction (VHR) method. The experimental results show that (1) the data can be effectively reduced by the VHR method; (2) the predictive accuracy of a classifier (C4.5 [10]) can be improved after data and dimensionality reduction; and (3) the classification rules learned are simpler. In other words, the VHR method reduces a database into a smaller one, limits the possible search space for a classifier, and produces simpler learned concepts.

2 A DDR System for Continuous Attributes

The key algorithm of the proposed DDR system is the VHR method (henceforth VHR), summarized below. VHR applies the χ^2 statistic. The idea is to check the correlation between an attribute and class values, based on which VHR tries to merge the ordered values of that attribute as much as it is allowed by the χ^2 statistic for a given significance level. It begins with some significance level, e.g., 0.5, for all numeric attributes of discretization. Each attribute i is associated with a sigLevel[i], and takes turns for merging. Each attribute is sorted according to its values. Then the following is performed: (1) calculate the χ^2 value for every pair of adjacent intervals (at the beginning, each pattern is put into its own interval); (2) merge the pair of adjacent intervals with the lowest χ^2 value. Merging continues until all pairs of intervals have χ^2 values exceeding the parameter determined by sigLevel (initially, 0.5, its χ^2 value is 0.455 if the degree of freedom is 1). The above process is repeated with a decremental sigLevel[i] until an inconsistency rate δ is exceeded in the discretized data. Consistency checking is conducted after each attribute's merging. If there is no inconsistency found, sigLevel[i] is decremented for attribute i's next round of merging; otherwise attribute i will not be involved in further merging. This process is continued until no attribute's values can be merged. At the end, if an attribute is merged to only one value, it simply means that this attribute is not relevant in representing the original data set. As a result, when discretization ends, feature selection is also accomplished.

VHR Algorithm:

```
set all sigLevel[i] = 0.5 for attribute i;
do until no-attribute-can-be-merged {
   for each mergeble attribute i {
      Sort(attribute i, data);
      chi-sq-initialization(attribute i, data);
      do {
         chi-sq-calculation(attribute i,data);
      } while (Merge(data) is TRUE)
      if (InConsistency(data) < δ )
         sigLevel[i] = decreSigLevel(sigLevel[i]);
      else
         attribute i is not mergeble;
   }     }
```

The formula for computing the χ^2 value can be found in a standard statistics book.

3 Experiments

In order to measure how data and dimensionality reduction is achieved, we need to consider several aspects. First, the dimensionally reduced data should still have the same discriminating power as the original; second, the reduced data have gains for a pattern classifier in predictive accuracy as well as the simplicity of the learned concepts.

For the first aspect, suffice it to show that the absolute number inconsistencies does not increase after the reduction. VHR guarantees this property since the number of inconsistencies is the stopping criterion for VHR. Given a data set, it is not difficult[1] to compute the number of inconsistencies in the set. For the second aspect, however,

[1]The time required is $O(n^2)$.

a pattern classifier is needed in the experiments. C4.5 [10] is chosen because (1) it can handle both numeric and nominal data; (2) it is well known, widely available, and works quite well in many domains. So there is no need to explain it in detail. The output of C4.5 is a decision tree. Whether a learned concept is simple or not in a way can be linked to the size of a tree. In other words, if tree A is larger than tree B, then tree B is simpler.

The experiment procedure for each data set is as follows:

1. Apply VHR to reduce the data;

2. Run C4.5 on both the original and reduced data;

3. Obtain results on predictive accuracy and tree size.

Just seeing the data reducing is not sufficient for a DDR algorithm; an effective DDR algorithm can improve a pattern classifier's accuracy, simplify the learned concepts as well as reduce the data. We want to show that VHR possesses these features.

3.1 Data Sets

Three data sets are of Iris, Wisconsin Breast Cancer and Heart Disease[2]. They have different types of attributes. The Iris data is of continuous attributes, the breast cancer data is of ordinal discrete ones, and the heart disease data has mixed attributes (numeric and discrete). The three data sets are described briefly as follows.

Iris data contains 50 patterns each of the classes *Iris setosa, Iris versicolor*, and *Iris virginica*. Each pattern is described using four numeric attributes: *sepal-length, sepal-width, petal-length,* and *petal-width.* The originally odd numbered data are selected for training, the rest for testing.

Breast cancer data contains 699 samples of breast fine-needle aspirates collected at the University of Wisconsin Hospital. There are 9 ordinal attributes valued on a

[2]They are all obtained from the University of California-Irvine, machine learning repository via anonymous ftp to *ics.uci.edu*.

Int	Class Freq			χ^2	Int	Class Freq			χ^2
4.4	3	0	0	0.20	6.0	0	2	1	1.43
4.6	2	0	0	0.20	6.1	0	0	1	0.54
4.7	1	0	0	0.20	6.2	0	1	2	0.14
4.8	3	0	0	1.97	6.3	0	2	3	0.14
4.9	1	0	1	2.62	6.4	0	1	2	0.16
5.0	3	1	0	0.10	6.5	0	1	3	1.97
5.1	3	1	0	0.70	6.6	0	1	0	2.50
5.2	2	0	0	0.20	6.7	0	1	4	0.73
5.3	1	0	0	0.41	6.8	0	1	1	0.10
5.4	3	1	0	1.32	6.9	0	1	1	0.85
5.5	1	2	0	1.66	7.0	0	1	0	2.10
5.6	0	4	0	2.50	7.1	0	0	1	0.20
5.7	1	1	0	1.28	7.4	0	0	1	0.20
5.8	1	2	2	1.20	7.7	0	0	2	
5.9	0	1	0	0.54					

Table 1: The initial intervals, class frequencies, and χ^2 values for Sepal Length.

scale of 1 to 10. The class value is either benign or malignant. The data set is split randomly into two sets, 350 patterns for training and 349 for testing.

Heart disease data contains medical cases of heart diseases. It contains numerically-valued features; there are 8 nominally-valued and 5 numerically-valued attributes. Two class values are: healthy and diseased heart. Removing patterns with missing attribute values, we use 299 patterns in which one third are randomly chosen for testing, and the rest are for training.

3.2 A Detailed Example

Here the two stages (intermediate and final) of VHR processing for the Iris data are shown to demonstrate the behavior of VHR. The intermediate stage is where all four attributes have the same minimum significance level (sigLevel = 0.2, χ^2 = 3.22), keeping the number of inconsistencies under the threshold δ (3 = 75*5%). The final stage is where no further attribute value merging is possible without sacrificing the discriminating power. Table 1 shows the intervals, class frequencies, and χ^2 values for attribute sepal-length after the data initialization by VHR.

Int	Class Freq			χ^2
4.4	9	0	0	5.05
4.9	1	0	1	8.11
5.0	12	3	0	13.64
5.5	3	12	3	14.23
6.1	0	10	21	

(a)

Int	Class Freq			χ^2
4.4	25	25	25	

(b)

Sepal Length

Int	Class Freq			χ^2
2.0	0	4	0	4.90
2.5	0	8	12	8.67
2.9	1	5	0	5.80
3.0	6	8	11	6.14
3.4	5	0	2	4.23
3.5	13	0	0	

(a)

Int	Class Freq			χ^2
2.0	25	25	25	

(b)

Sepal Width

Int	Class Freq			χ^2
1.0	25	0	0	47.00
3.0	0	21	1	4.18
4.8	0	4	2	17.21
5.0	0	0	22	

(a)

Int	Class Freq			χ^2
1.0	25	0	0	53.00
3.0	0	25	3	39.39
5.0	0	0	22	

(b)

Petal Length

Int	Class Freq			χ^2
0.1	25	0	0	38.10
1.0	0	13	0	4.72
1.4	0	2	1	3.37
1.5	0	9	0	11.35
1.7	0	1	5	3.40
1.9	0	0	19	

(a)

Int	Class Freq			χ^2
0.1	25	0	0	50.00
1.0	0	24	1	42.42
1.7	0	1	24	

(b)

Petal Width

Table 2: The intervals, class frequencies, and χ^2 values for the four attribute in two stages: (a) intermediate and (b) final. The χ^2 thresholds for Sepal Length are (a) 3.22 and (b) 50.6; the χ^2 thresholds for Sepal Width are (a) 3.22 and (b) 40.6; the χ^2 thresholds for Petal Length are (a) 3.22 and (b) 10.6; and the χ^2 thresholds for Petal Width are (a) 3.22 and (b) 10.6, respectively.

The results for the four attributes at the two stages are shown in Table 2. With the χ^2 threshold 3.22, for example, five discrete values are needed for attribute sepal-length: $< 4.9 \rightarrow 1$, ..., $< 6.1 \rightarrow 4$, and $\geq 6.1 \rightarrow 5$. The last one reads if a numeric value is greater than and equal to 6.1, it is quantized to 5. When VHR terminates, the values of both sepal-length and sepal-width are merged into one value, so they can be removed; and attributes petal-length and petal-width are discretized into three discrete values each.

The tables summarize each attribute's intervals, class frequencies, and χ^2 values. All the thresholds mentioned in Table 2 are determined by VHR automatically. All the tables are generated based on the 75 training data. Certainly, if all the 150 patterns are used, the tables could be different. Here we assume that the testing data is not available at the stage of dimensionality reduction.

3.3 Results

The results are summarized below, where "Before" or "After" means the result is obtained before or after using VHR. The data reduced by VHR preserves the discriminating power of the original data. This is measured by the number of inconsistencies before and after VHR processing. The characteristics of the data remain as well since the accuracy of C4.5 on the data processed (After) by VHR is at least as good as one on the original data (Before). The tree size is smaller after using VHR. The number of distinguishable patterns decreases so does the number of attributes. The details are shown in the following:

- **Accuracy** It was explained earlier that VHR preserves the discriminating power of the original data. However, this preservation would not be useful if a classifier cannot learn well from the dimensionally reduced data, that is, only if the use of VHR gives rise to a better or similar performance. For the three sets of data, the results show that it is at least as good as without using VHR regardless of other considerations (to be discussed subsequently).

	Accuracy %	
	Before	After
Iris	94.7	94.7
B. Cancer	92.6	94.6
H. Disease	72.7	78.8

- **Tree Size** An immediate benefit of applying a DDR system is that the learned concept could be simpler. It means a smaller tree for a decision tree approach such as C4.5. For the data sets chosen, the tree size can be reduced to as much as half of the original size.

	Tree Size	
	Before	After
Iris	5	5
B. Cancer	21	11
H. Disease	43	22

- **Data Size** This is defined by the number of items (records) in the training data (a database). After VHR processing, the number of non-duplicate items is reduced. As in the Iris data, only six distinct items remain with one inconsistency. For such cases, even an exhaustive search method could be employed to produce high quality classification rules without resorting to monothetic methods such as C4.5.

	Data Size	
	Before	After
Iris	75	6
B. Cancer	350	75
H. Disease	198	173

As an example, two sets of rules are presented here to show the point. Rule set A is produced by C4.5, and rule set B is produced by a rule generator that induces rules from a small data set heuristically [11]. The accuracies of the two rule sets on the training data are 96% and 99%, on the testing data are 94% and 97% respectively.

```
    Rule Set A:                        Rule Set B:
Petal-Length <= 1.9 -> 1           Petal-Length < 3.0 -> 1
Petal-Length > 1.9  &              Petal-Length > 5.0 &
Petal-Width <= 1.6 -> 2            Petal-Width < 1.7 -> 2
Petal-Width > 1.6 -> 3             Default -> 3
Default -> 1
```

- **Number of Attributes** One of the most important advantages of VHR is that it can reduce the number of attributes. Only relevant attributes are chosen and irrelevant ones are deleted. This renders great help to reducing work and minimizing resource use in the future data collection and classification. It also helps the human experts and data analysts focus on the important dimensions.

	# of Attributes	
	Before	After
Iris	4	2
B. Cancer	13	10
H. Disease	9	6

4 Conclusion

Introduced here is a DDR system based on the VHR algorithm. The key idea is to apply techniques of discretization and feature selection to data and dimensionality reduction in the context of numeric attributes. Discretization merges the values of each attribute, thus significantly decreases the number of values a continuous attribute can take and reduces the data in a vertical dimension. Normally this process will generate some duplicates in the data; by removing the duplicates, the database becomes smaller while keeping the same discriminating power. The horizontal dimensionality reduction is achieved by feature selection that eliminates those discretized attributes having only one possible value.

The advantages of having dimensionally reduced data are four-fold: 1. it narrows down the search space determined by attributes; 2. it allows faster learning for a classifier; 3. it helps a classifier produce simpler learned concepts; and 4. it improves predictive accuracy. However, it has its limitations. As of now, we do not see any straightforward way to extend the method to handle higher order correlations in data regardless of the computation cost of permutation of multiple attributes. Since the possibility of having high order correlated data cannot be ruled out, further work should be done in this direction. Another possible extension is of data with mixed nominal and ordinal attributes. Since the nominal attributes are masked out in VHR, the inconsistency checking of VHR can be done with or without the masked attributes, which leads to either under- or over-discretization. Under-discretization is caused by the possibility that some masked attributes could be irrelevant. Over-discretization is due to the fact that masked attributes do contribute in discriminating one record from the other. More study is needed. Another line of research is to investigate the relationship between the discriminating power of a database and its real distribution. In the present work, we have used an indirect measure - predictive accuracy, that is a high accuracy means the dimensionally reduced data keeps the original distribution. The VHR method has been successfully applied to many problems. With these extensions, the VHR method can be more flexible and more generally applicable.

References

[1] W. Frawley, G. Piatetsky-Shapiro, and C. Matheus. Knowledge discovery in databases: An overview. *AI Magazine*, Fall 1992.

[2] N. Cercone and M Tsuchiya. Geust editors, special issue on learning and discovery in databases. *IEEE Trans. on Knowledge and Data Engineering*, 5(6), December 1993.

[3] J. Han, Y. Cai, and H. Cercone. Knowledge discovery in databases: An attribute oriented approach. In *Proceedings of the VLDB conference*, pages 547–559, 1992.

[4] C.J. Matheus, P.K. Chan, and G. Piatetsky-Shapiro. Systems for knowledge discovery in databases. *IEEE Trans. on Knowledge and Data Engineering*, 5(6), December 1993.

[5] G. Piatetsky-Shapiro. Editor, special isssue on knowledge discovery in databases. *International Journal of Intelligent Systems*, 7(7), September 1992.

[6] H. Almuallim and T.G. Dietterich. Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence*, 69(1-2):279–305, November 1994.

[7] U.M. Fayyad and K.B. Irani. The attribute selection problem in decision tree generation. In *AAAI-92, Proceedings Ninth National Conference on Artificial Intelligence*, pages 104–110. AAAI Press/The MIT Press, 1992.

[8] H. Ragavan and L. Rendell. Lookahead feature construction for learning hard concepts. In *Machine Learning: Proceedings of the Seventh International Conference*, pages 252–259. Morgan Kaufmann Pub. San Mateo, California, 1993.

[9] N. Wyse, R. Dubes, and A.K. Jain. A critical evaluation of intrinsic dimensionality algorithms. In E.S. Gelsema and Kanal L.N., editors, *Pattern Recognition in Practice*, pages 415–425. Morgan Kaufmann Publishers, Inc., 1980.

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

[11] H. Liu and S.T. Tan. X2r: A fast rule generator. In *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 1995.