Research Journal of Computer and Information Technology Sciences _____
Vol. **1(1),** 18-20, February **(2013)** *Res. J. Computer & IT Sci.*

Mini Review Paper

A Technique for Data Integration Using Association of Attributes in Data Preprocessing

Moteria Parag M.[1] and Ghodasara Y.R.[2]
[1]MCA Department, ISTAR College, V.V.Nagar, Gujarat, INDIA
[2]College of AIT, Anand Agricultural University, Anand, Gujarat, INDIA

Available online at: www.isca.in
Received 12th October 2012, revised 7th January 2013, accepted 25th January 2013

Abstract

Data integration involves combining data residing in different sources and providing users with a unified view of these data. Volumes of data grow exponentially in all realms from personal data to enterprise and global data. Thus it is becoming extremely important to be able to understand data sets and organize them. To organize large volume of data, there are certain disciplines such as data integration, migration, synchronization, business intelligence etc. allow this. Our paper strives to explain and describe data integration ideas and concepts using association of attributes for categorical variables. Data integration involves combining data from several disparate sources, which are stored using various technologies and provide a unified view of the data. Data integration becomes increasingly important in cases of merging systems of two companies or consolidating applications within one company to provide a unified view of the company's data assets.

Keywords: Attribute, categorical variable, consistent data, correlation, data integration, data mining, method of association, quantitative measure.

Introduction

Data mining often requires data integration – the merging of data from multiple data stores. Data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files[1]. Data Integration is a becoming a persistent challenge faced in applications that need to query across multiple autonomous and heterogeneous data sources[2].

Problem Description

There are a number of issues to consider during data integration. How can equivalent real world entities from multiple data sources be matched up? This is referred to as the entity identification problem. For example, how can the data analyst or the computer be sure that customer_id in one database and cust_number in another refer to the same attribute? Redundancy is another important issue. An attribute may be redundant if it can be derived from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set. Some redundancies can be detected by correlation analysis. Note that correlation does not imply causality. That is, if A and B are correlated, this does not necessarily imply that A causes B or that B causes A[1]. Technique of correlation is used to measure the degree of relationship between two such phenomena as are capable of direct quantitative measurement[3]. But correlation technique cannot imply to measure causality for categorical variables. Therefore, method of association of attributes is employed to measure the degree of relationship between two phenomena whose size we cannot measure and where we cannot only determine the presence or absence of a particular attribute[3].

Types of Measurement Phenomena

There are two types of phenomena to measure in different data sources; one is variables and second is attributes[3]. Variable phenomena are study of quantitative measured and attribute phenomena are study of presence or absence of characteristics[3] (categorical variable). Central tendency, dispersion, correlation are various techniques to measure variable phenomena, while method of association technique is used to measure attribute phenomena[3].

Method of Association

Association coefficients are tools in data analysis that measure the strength of a relationship between two variables[4]. Capital letters represent presence of the attributes and small letters represent absence of the attributes. For example, 'A' represents male attribute then 'a' represents female attribute. Similarly, if 'B' represents literates then 'b' represents illiterates. The combination of different attributes is denoted by (AB), (Ab), (aB) and (ab). Thus in this example, (AB) would mean number of literate males[5].

Different attributes, their sub-groups and combinations are called different classes. If the number of attributes is n, then there will be 3^n classes. For example, we study two attributes say A and B then we have nine classes i.e. (A), (a), (B), (b), (AB), (Aa), (aB), (ab) and N. The contingency table of order (2X2) for two attributes A and B can be displayed as given below[5]

Table – 1

	A	**a**	**TOTAL**
B	(AB)	(aB)	B
b	(Ab)	(ab)	b
TOTAL	A	a	N

The number of records assigned to class is called their count or frequencies or class frequencies. The number of records or units belonging to class is known as its frequency is denoted within bracket. Thus (A) stands for the numbers of records of attribute A. If the number of attributes is n, we have 2^n cell frequencies[5]. Following chart describe relationship between the class frequencies:

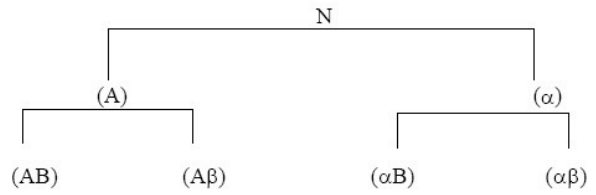

Chart – 1

Consistency of data

In order to find out whether the given data are consistent or not we have to apply a very simple test. The test is to find out whether any or more of the class frequencies is negative or not. If none of the class frequencies is negative we can safely calculate that the given data are consistent. On the other hand, if any of the class frequencies comes to be negative the given data are inconsistent[5].

Yule's Coefficient of Association Methods

This method is the most popular statistical method of studying association because here not only we can determine the nature of association, i.e. whether the attributes are positively associated, negatively associated or independent, but also the degree or extent to which the two attributes is associated[5].
The Yule's coefficient is denoted by 'Q' and obtained by

$$Q = \frac{(AB)(\alpha\beta) - (A\beta).(\alpha B)}{(AB)(\alpha\beta) + (A\beta).(\alpha B)} \qquad (1)$$

The value of Q lies between -1 to +1. When the value of Q is +1 there is perfect positive association between the attributes. When the value of Q is -1 there is perfect negative association between the attributes. When the value of Q is zero the two attributes are independent[5].

The coefficient of association can be used to compare the intensity of association between two attributes with intensity of association between two other attributes.

Consider two attributes say one is Smokers and other is Tea Drinkers. Consider that, 'A' represents attribute Smokers and 'B' represents attributes Tea Drinkers. Therefore, 'a' represents attribute Non-Smokers and 'B' represents attributes Non-Tea Drinkers.

Case – I: Considering numbers of records are 88, Non-Smokers are 45, Tea Drinkers are 73 and Smokers and Tea Drinkers are 40. Hence, N = 88, (a) = 45, (B) = 73, (AB) = 40

From table-1 and chart-1, we can calculate value of contingency table.

Table – 2

-	**A**	**A**	**TOTAL**
B	40	33	73
b	3	12	15
TOTAL	43	45	88

From the above table – 2, all cell frequencies are positive, therefore data are consistent.

Using equation-(1), we get Q = 0.658

Value of Q is not equals to zero, hence this shows that the both attributes are not "independent" but associated (positively associated). Therefore, we can integrate both attributes.

Case – II: Considering numbers of records are 1500, Non-Smokers are 1117, Tea Drinkers are 360 and Smokers and Tea Drinkers are 35. Hence, N = 1500, (a) = 1117, (B) = 360, (AB) = 35

From Table-1 and Chart-1, we can calculate value of contingency table

Table – 3

-	**A**	**a**	**TOTAL**
B	35	325	360
b	348	792	1140
TOTAL	383	1117	1500

From the above table – 3, all cell frequencies are positive, therefore data are consistent.

Using equation-(1), we get Q = -0.606

Value of Q is not equals to zero, hence this shows that the both attributes are not "independent" but associated (negatively associated). Therefore, we can integrate both attributes.

Case – III: Considering numbers of records are 500, Non-Smokers are 300, Tea Drinkers are 150 and Smokers and Tea Drinkers are 60. Hence, N = 500, (a) = 300, (B) = 150, (AB) = 60 From table-1 and chart-1, we can calculate value of contingency table.

Table – 4

-	A	a	TOTAL
B	60	90	150
b	140	210	350
TOTAL	200	300	500

From the above table – 4, all cell frequencies are positive, therefore data are consistent.

Using equation-(1), we get Q = 0

Value of Q is equals to zero; hence this shows that the both attributes are independent. Hence, no need to integrate both attributes.

Case – IV: Considering numbers of records are 500, Non-Smokers are 400, Tea Drinkers are 150 and Smokers and Tea Drinkers are 140. Hence, N = 500, (a) = 400, (B) = 150, (AB) = 140

From table-1 and chart-1, we can calculate value of contingency table

Table – 5

-	A	a	TOTAL
B	140	10	150
b	-40	390	350
TOTAL	100	400	500

From the above table – 5, one cell frequency is negative, therefore data are not consistent. Therefore, no need to calculate Q for association of attributes. Hence, integration of both attributes is not consistent.

Conclusion

The paper shows with example that method of association is successfully applied to find out either both attributes are consistent or positively associate or negative associate or independent. Disadvantage of this technique is that, it is applicable to only two attributes simultaneously to validate association but the coefficient of association can be used to compare the intensity of association between two attributes with intensity of association between two other attributes.

References

1. Jiawei Hanand Micheline Kamber, Data Mining Concepts and Techniques - Second Edition, ELSEVIER Morgan Kaufman Publisher, 67-70 **(2011)**
2. Mohammed T. Al-Sudairy and T.G.K. Vasista, Semantic Data Integration Approaches for E-Governance, International Journal of Web and Semantic Technology, **2(1)**, 1-12 **(2011)**
3. http://www.epidemiology.ch/history/PDF%20bg/Yule % 20U%201903%20notes%20on%20the%20theory %20 of %20assoc%20of%20attrib%20in%20stats.pdf **(1903)**
4. http://www.cis.drexel.edu/faculty/thu/research-papers/ aml3730.galley.pdf **(2002)**
5. http://www.assignmenthelp.net/assignment_help/consistency-of-data.php **(2012)**
6. http://www.dataintegration.info/data- integration **(2011)**
7. http://en.wikipedia.org/wiki/Data_integration **(2012)**