International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-3, Issue-2, March 2015

Data Mining: Using C++ to Measure Correlation between Real Valued and Nominal Valued datasets

Shahid Ali Khan, Praveen Dhyani

Abstract: In vast databases, there are various fields of mixed data types such as real, nominal and ordinal etc. In data mining applications, to know the relationship among data sets is an important issue. A correlation is commonly used for measuring relationship between two data sets and the correlation coefficient measures the strength as well as direction between two data sets and usually used in the context of real valued data sets.
In this paper, the correlation coefficient between real valued and nominal valued data sets has been measured by using C++ language.

Keywords: Correlation, Data mining, Nominal values, Real values.

I. INTRODUCTION

Data Mining is the process of discovering interesting knowledge from a large amount of data stored either in databases, data warehouse, worldwide web or other information repositories [9].

In data mining, different tools (like: statistical models, mathematical algorithms, and machine learning methods) are used to discover the hidden information. In large databases, the characteristics of data are of mixed types. In general, there are two types of datasets such as categorical and real-valued datasets. The data type where observations can be summarized as counts or cross tabulations, known as categorical datasets. For example, students ID, mobile numbers, gender, car-colour, socio-economic status, army ranks etc. Categorical data can be further divided into two categories, viz., nominal or ordinal data. In nominal data, values can be assigned to categories on random basis, therefore, arithmetic operations or logical operations are not applicable. Thus, gender and mobile numbers are examples of nominal data. For instance, we may add two mobile numbers but outcome is meaningless. In ordinal data, ordering can be done, but the intervals between scale points may not be equally spaced. Therefore, arithmetic operations are not applicable but logical operations can be performed. Thus, socio-economic status and the army ranks are examples of ordinal data. The real-valued data contains real numbers, and is an outcome of some measurements, e.g., age, Income, date, temperature etc. The real-valued data can be further divided two categories viz., interval data and ratio data. A real-valued data have meaningful arithmetic difference, but division or multiplication operations are meaningless. Date and temperature are examples of interval data. A real-valued data set where all arithmetic operations results into meaningful outcome, is known as ratio data. Age and income are examples of ratio data.

Data Miner handle huge databases of various fields and establish a meaningful relationship among various attributes for better decision making. There are several statistical techniques which have become integral part of data mining processes. Pearson's correlation coefficient is one of the commonly used statistical tool to find association between two datasets. It measures the degree to which the data points of one domain tend to diverge with changes in the data points of another domain. Assume, A and B are two hypothetical domains, A = $(a_1, a_2, a_3, \ldots\ldots a_n)$ for the first domain and B = $(b_1, b_2, b_3, \ldots\ldots b_n)$ for the second domain. The value of correlation coefficient [3] between the datasets A and B, usually denoted by r, its value lies in between -1 and 1, which measures the degree as well as the direction of correlation. If the value of r is positive, it means values of both sequences are moving in same direction, otherwise in opposite direction. In general, correlation is applied over two sequences of a real data points. The Pearson's correlation coefficient has been used in data mining applications algorithm development [8]. Furthermore, Cheung and Li [2] proposed a new quantitative correlation coefficient mining method, this method can discover the hidden patterns of sales and market for business intelligence in small and medium enterprises.

In this paper, a correlation coefficient has been measured in between real valued datasets and nominal datasets using C++ .

Manuscript received March 23,2015
Mr. Shahid Ali Khan, Asst. Professor, Department of Computer Sciences & Engineering, Waljat College of Applied Sciences, P.O.Box-197 P.C.-124 Oman, Tel: +968-24446660(334)
Prof (Dr.) Praveen Dhyani, Executive Director, Banasthali University, Jaipur, India.

II. RELATED WORK

In data mining application, correlation is a popular statistical tool used to establish association between two datasets. The data mining experts can use information obtained after measuring correlation for further analysis. For example, information obtained from correlation can be used in regression analysis. A common multivariate technique such as principal component analysis (PCA) has developed based on correlation analysis [6] and further correlation is being used to measure distance between data points in clustering [5]. These both techniques are presently used in handling data with real-valued datasets only. Based on Cramer's V-statistics proposed a metric called the d_{cv} metric, which is almost similar as Mahalanobis but this is used when data are nominal-valued [1]. In addition to Pearson's correlation coefficient, Spearman's rho, compute the correlation between two ordinal or ranked variables. Point Biserial correlation coefficient is used to measure association between interval/ratio variable and dichotomous variables whereas Rank Biserialis used to compute correlation between ordinal variable and dichotomous variable. Furthermore, the Contingency Coefficient and Cramers' Phi, compute the strength of relationship between nominal data values. The Phi Coefficient, compute the correlation between two dichotomous variables. Rayward-Smith [7] proposed a method which helps in measuring association between nominal and real-valued data sets.

III. TO MEASURE CORRELATION BETWEEN TWO REAL VALUED DATASETS.

The standard statistical method for measuring the correlation coefficient between two non-constant datasets A and B, when both datasets contain real numbers, is defined by

$$corr\ (a,b) = \frac{S_{ab}}{\sqrt{S_{aa}\ \sqrt{S_{bb}}}}$$

Where, with n pairs of observations

$(a_1,b_1),(a_2,b_2).....(a_n,b_n)$

$$S_{ab} = \sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b}) = \sum_{i=1}^{n} a_i b_i - n\bar{a}.\bar{b}$$

$$S_{aa} = \sum_{i=1}^{n}(a_i - \bar{a})^2 = \sum_{i=1}^{n} a_i{}^2 - n.\bar{a}^2$$

$$S_{bb} = \sum_{i=1}^{n}(b_i - \bar{b})^2 = \sum_{i=1}^{n} b_i{}^2 - n.\bar{b}^2$$

The coefficient of correlation $corr(a,b)$ measures the strength of linear relationship between two real valued datasets A and B. The commonly used properties of the correlation coefficient are as follows.

- The value of $corr(a,b)$ is always between -1 and 1, i.e. $-1 \le corr(a,b) \le 1$
- The magnitude of $corr(a,b)$ indicates the strength of the linear relation and its sign indicates the direction. In particular, if $corr(a,b)$ >0 indicates the pattern of (a,b) values is a band that runs from lower left to upper right i.e. the values of both datasets are in same direction. if $corr(a,b)$ <0 indicates the pattern of (a,b) values is a band that runs from upper left to lower right i.e. the values of both datasets are in opposite direction.
- $corr(a,b)$ =+1 indicates that all values of (a,b) are exactly on a straight line with a positive slope (perfect positive linear relation). In case of $corr(a,b)$=-1, indicates that all values of (a,b) are exactly on a straight line with a negative slope (perfect negative linear relation).
- $corr(a,b)$ nearly zero, indicates that the linear relation is very weak.

Note:

- $corr(a,b) = corr(b,a)$
- $corr(a,b) = corr(u,v)$, where $u = \lambda_1.a + \mu_1, v = \lambda_2.b + \mu_2$ for any $\lambda_1, \lambda_2 > 0$ and μ_1, μ_2

IV. TO MEASURE CORRELATION BETWEEN TWO BINARY SEQUENCED DATASETS.

$$corr(a,b) = \frac{(n_{00}n_{11} - n_{01}n_{10})}{\sqrt{n_0^a n_1^a n_0^b n_1^b}}$$, called *phi coefficient* (Φ)

Let us consider two datasets ***a*** and ***b***, both are sequence of binary numbers. Now assume that $n_{rs,}$, $r = \{0,1\}$ and $s = \{0,1\}$, denotes the number of pairs with $x_i = \bar{r}\ and$ $y_i = s$, then $n = n_{00} + n_{01} + n_{10} + n_{11}$.

And n_0^a, n_1^a: denote the number of times the value 0 and 1occurs in dataset A respectively.

Similarly, n_0^b, n_1^b: denote the number of times the value 0 and 1occurs in dataset B respectively.

Then $n_0^a = n_{00} + n_{01}$, $n_1^a = n_{10} + n_{11}$, $n_0^b = n_{00} + n_{10}$ and $n_1^b = n_{01} + n_{11}$ and

$$\bar{a} = \frac{n_{10} + n_{11}}{n}\ ,\ \bar{b} = \frac{n_{01} + n_{11}}{n}$$

$$\sigma_a = \sqrt{\frac{1}{n}(n_{10}+n_{11}) - \left(\frac{n_{10}+n_{11}}{n}\right)^2} = \frac{1}{n}\sqrt{n_0^a \times n_1^a}$$

Similarly, $\sigma_b = \sqrt{\frac{1}{n}(n_{01}+n_{11}) - \left(\frac{n_{01}+n_{11}}{n}\right)^2} = \frac{1}{n}\sqrt{n_0^b \times n_1^b}$

$$\mathrm{cov}(a,b) = \frac{1}{n}\sum a_i b_i - \bar{a}.\bar{b} = \frac{n_{00}n_{11} - n_{01}n_{10}}{n^2}$$

$$corr(a,b) = \frac{(n_{00}n_{11} - n_{01}n_{10})}{\sqrt{n_0^a n_1^a n_0^b n_1^b}}$$

For instance,

a	0	1	1	1	0	0	1	0	0	1
b	1	0	1	0	1	1	1	0	1	0

Here, $n_{00}=1$, $n_{01}=4$, $n_{10}=3$ and $n_{11}=2$

$n_0^a = 5$, $n_1^a = 5$, $n_0^b = 4$ and $n_1^b = 6$

$corr(a,b) = -0.408$

V. To measure correlation between a real valued dataset and a nominal dataset.

When data mining expert desires to know possible relationship between a real valued dataset and a nominal valued dataset, this novel area is a challenging for researcher in data mining.

Consider (a_i, b_i), $1 \le i \le n$, n pairs of observations of two datasets A and B. A datasets A contains with n real values $\{a_1, a_2, a_3, \ldots\ldots a_n\}$ while the dataset B contains with n nominal values $\{b_1, b_2, b_3, \ldots\ldots b_n\}$, selected from a finite domain $V=\{v_1, v_2, v_3, \ldots\ldots v_k\}$. Each data point $v_j \in V, 1 \le j \le k$, is associated to a set A^j, comprises $a-$ values whose corresponding $b-$ values equal to v_j. Now, let A^j denote the sequence of real values constructed from A by just selecting $a-$ values with indices in I_j where $I_j = \{\{i : 1 \le i \le n \text{ and } y_i = v_j\}$. Let n_j be the number of data point in A^j and the mean of the data points in A^j is denoted by $\bar{a}_j$. Then define

$$S_{aa}^j = \Sigma_{i \in I_j}(\bar{a}_j - a_i)^2 = \Sigma_{i \in Ij} a_i^2 - n_j \bar{a}_j^2$$

Thus, the standard deviation of the data points of A^j will be $\sigma_{a^j} = \sqrt{\frac{S_{aa}^j}{n_j}}$ Rayward-Smith [7]is proposed a simple statistic, based on comparing the weighted average of the variances of A^j, $1 \le j \le k$ with nonzero variances of A, to measure the correlation between the real valued datasets and nominal valued datasets as follows.

$$corr(a,b) = \frac{\sum_{j=1}^{k} n_j.\bar{a}_j^2 - n.\bar{a}^2}{\sum_{i=1}^{n} a_i^2 - n.\bar{a}^2}$$

Where, n_j : is the length of data sets A^j

n : is the length of data sets A

$\bar{a}_j$: is the mean of data points of A^j

$\bar{a}$: is the mean of all data points of A

Proof:

Since correlation between the real valued data sets A and the nominal valued data set B is based on comparing the weighted average of the variances of A^j. Thus

$$corr(a,b) = \frac{S_{aa} - \Sigma_{j=1}^{k} S_{aa}^j}{S_{aa}} \quad \text{.....(1)}$$

By definition, above is equivalent to

$$corr(a,b) = \frac{\sum_{i=1}^{n} a_i^2 - n.\bar{a}^2 - \sum_{j=1}^{k}(\Sigma_{i \in I_j} a_i^2 - n_j.\bar{a}_j^2)}{\sum_{i=1}^{n} a_i^2 - n.\bar{a}^2},.$$

$$corr(a,b) = \frac{\sum_{i=1}^{n} a_i^2 - n.\bar{a}^2 - \sum_{j=1}^{k}\Sigma_{i \in I_j} a_i^2 + \sum_{j=1}^{k} n_j.\bar{a}_j^2}{\sum_{i=1}^{n} a_i^2 - n.\bar{a}^2} \quad \text{.....(2)}$$

Thus finally, we have $corr(a,b) = \frac{\sum_{j=1}^{k} n_j.\bar{a}_j^2 - n.\bar{a}^2}{\sum_{i=1}^{n} a_i^2 - n.\bar{a}^2}$

since, $\sum_{i=1}^{n} a_i^2 = \sum_{j=1}^{k}\Sigma_{i \in I_j} a_i^2$

Note that if data points of each A^j are identical, then $corr(a,b) = 1$ because each S_{aa}^j will be zero.

Secondly, $corr(a,b) = \frac{S_{aa} - \Sigma_{j=1}^{k} S_{aa}^j}{S_{aa}} = 1 - \frac{\Sigma_{j=1}^{k} S_{aa}^j}{S_{aa}}$

and each $S_{aa}^j \ge 0$ which implies $corr(a,b)$ cannot be negative. So, $0 < corr(a,b) \le 1$, when dealing real valued and nominal valued data sets.

VI. Experiments

In this section, we will compute the correlation between nominal and real valued datasets given as Table-1 by the above discussed approach. The *Excel 2010* has been used for computational work as well as through C++ in this study.

Table-1

Nominal data valued (A)	Real data valued (B)
a	1
a	3
a	4
a	5
a	6
a	7
a	9
a	10
b	15
b	16
b	18
b	21
b	22
c	31
c	32
c	34
c	35
c	38
c	40
c	45

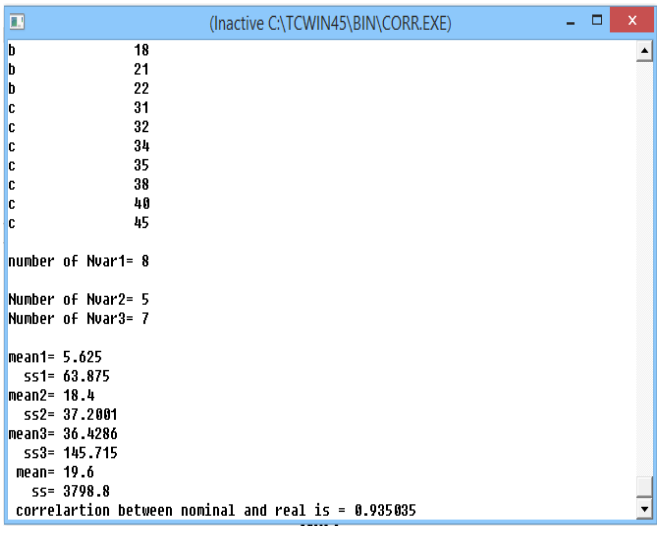
corr(a,b)= 0.935034936

C++ Program

```
#include<iostream.h>
void main()
{
int dataset2[100],i,n,n1,n2,n3;
int temp=0,temp1=0,temp2=0,temp3=0;
float s=0,s1=0,s2=0,s3=0,ss=0,ss1=0,ss2=0,ss3=0;
float mean,mean1,mean2,mean3,corr;
char dataset1[100];
cout<<"Total number of pairs"<<endl;
cin>>n;
cout<<"Enter the nominal values"<<endl;
for(i=0;i<n;i++)
cin>>dataset1[i];
cout<<" enter the real values"<<endl;
for(i=0;i<n;i++)
cin>>dataset2[i];
cout<<"dataset1"<<"\t"<<"dataset2"<<endl;
for(i=0;i<n;i++)
cout<<dataset1[i]<<"\t\t"<<dataset2[i]<<endl;
cout<<endl;
cout<<"number of Nvar1= ";
cin>>n1;
cout<<endl;
cout<<"Number of Nvar2= ";
cin>>n2;
cout<<"Number of Nvar3= ";
cin>>n3;
cout<<endl;
for(i=0;i<n;i++)
{
s=s+dataset2[i];
temp=temp+dataset2[i]*dataset2[i];
}
mean=s/n;
ss=temp-n*mean*mean;

for(i=0;i<n1;i++)
{
s1=s1+dataset2[i];
temp1=temp1+dataset2[i]*dataset2[i];
}
mean1=s1/n1;
ss1=temp1-n1*mean1*mean1;
for(i=n1;i<n1+n2;i++)
{
s2=s2+dataset2[i];
temp2=temp2+dataset2[i]*dataset2[i];
}
mean2=s2/n2;
ss2=temp2-n2*mean2*mean2;
for(i=n1+n2;i<n;i++)
{
s3=s3+dataset2[i];
temp3=temp3+dataset2[i]*dataset2[i];
}
mean3=s3/n3;
ss3=temp3-n3*mean3*mean3;

corr=(ss-ss1-ss2-ss3)/ss;
cout<<" correlartion between nominal and real is =
"<<corr;
}
```



VII. Conclusion

Correlation is a most powerful technique in data mining to measure the association between two data sets. There are some data mining techniques where association between

two data sets plays an important role. In this paper, the concept of correlation coefficient has been extended and the correlation has been measured between real-valued and nominal valued datasets. The procedure to compute correlation coefficient proposed by Rayward-smith [7] has been employed and validated on data sets values. This study will help data miners, academicians and students who face difficulty in measuring association between nominal and real valued datasets.

References:

[1] Al-Harbi, S.H., McKeown, G.P.,&Rayward-Smith, V.J., (2003). A new metric for categorical data. In: Bozdogan, H. (Ed.), *Statistical Data Mining and Knowledge Discovery*. CRC Press, Boca Raton, FL.

[2] Cheung, C. F., & Li, F. L. (2012). A quantitative correlation coefficient mining method for business intelligence in small and medium enterprises of trading business. *Expert Systems with Applications*, *39*(7), 6279-6291.

[3] Easton, V. J., & McColl, J. H. (1997). Statistics Glossary V1.1, Paired data, correlation and regression, Available <http://www.stats.gla.ac.uk/steps/glossary/paired_data.html#corrcoeff>, accessed on 10 February 2014.

[4] Hawthorne G & Elliott P.(2005). Imputing cross-sectional missing data: comparison of common techniques. *Australian and New Zealand Journal of Psychiatry*, 39(7), 583–90.

[5] Jain, A.K., Murty, M.N., &Flynn, P.J. (1999). Data clustering: a review. *ACM computing surveys,* 31 (3), 264–323.

[6] Jolliffe, I.T.(1986). Principal Component Analysis. *Springer*, Berlin.

[7] Rayward-Smith, V. J.(2007). Statistics to measure correlation for data mining applications. *Computational Statistics & Data Analysis*, *51*(8), 3968–3982.

[8] Xiong, H., Shekhar, S., Tan, P., & Kumar, V. (2004). Exploiting a support-based upper bound of pearson's correlation coefficient for efficiently identifying strongly correlated pairs. *In Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining*, 334–343.

[9] Han J. and Kamber M., Data Mining Concepts and Techniques Morgan Kaufmann Publisher, 2006.

[10] Little R.J. and Rubin D.B., *Statistical Analysis with Missing Data*. Second Edition. John Wiley and Sons, New York.(2002).