Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 1
February 2009

Chi-Square Test for Anomaly Detection in XML Documents Using Negative Association Rules

K. Premalatha (Corresponding Author)
Kongu Engineering College
Perundurai, Erode, TN, India
E-mail: kpl_barath@yahoo.co.in
A.M. Natarajan
Bannari Amman Institute of Technology
Erode, TN, India

Abstract

Anomaly detection is the double purpose of discovering interesting exceptions and identifying incorrect data in huge amounts of data. Since anomalies are rare events, which violate the frequent relationships among data. Normally anomaly detection builds models of normal behavior and automatically detects significant deviations from it. The proposed system detects the anomalies in nested XML documents by independency between data. The negative association rules and the chi-square test for independency are applied on the data and a model of abnormal behavior is built as a signature profile. This signature profile can be used to identify the anomalies in the system. The proposed system limits the unnecessary rules for detecting anomalies.

Keywords: Anomaly detection, Chi-square test, Negative association rule, XML

1. Introduction

XML is a simplified subset of the Standard Generalized Markup Language (SGML). It provides a file format for representing data, a schema for describing data structure, and a mechanism for extending and annotating Hyper-Text Markup Language (HTML) with semantic information. The XML data model carries both data and schema information, being naturally suitable to represent semi-structured data. It is a standard for representing and exchanging information on the Internet.

XML is a markup language for structured documentation. Structured documents are documents that contain both content and some indication of what role that content plays. Almost all documents have some structure. A markup language is a mechanism to identify structures in a document. The XML specification defines a standard way of adding markup to documents. Information marked up as XML data is becoming increasingly persistent that allow data to be imported, accessed and exported in the XML format. XML database may prove more efficient and easier to store the data in XML format. As XML document storage formats become popular, the task of detecting anomalies within XML document collections becomes more important.

Deviation from the normal or common order or form or rule or deviation from the normal standard, especially as a result of congenital defects is called anomaly or outlier. Otherwise (Jiawei Han, Micheline Kamber. 2004.) very often, there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers or anomaly. Due to anomalies, data may be inconsistent. Since anomalies are rare event which violate the frequent relationships among data. Anomaly Detection may refer to an unsupervised data mining technique that produces a data mining model for identifying cases (records) that deviate from the norm in a dataset. The general step for anomaly detection schemes is

Build a profile of the abnormal behavior. Profile can be patterns or summary statistics for the overall population (Dataset)

Use the abnormal profile to detect anomalies. Anomalies are observations whose characteristics agree significantly with the abnormal profile

2. XML Documents

Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML. Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. In the real world, computer systems and databases contain data in incompatible formats. XML data is stored in plain text format. This provides a software- and hardware-independent way of storing data. This makes it much easier to create data that different applications can share. With XML, data can easily be exchanged between incompatible systems. One of the most time-consuming challenges for developers is to exchange data between incompatible systems over the Internet. Exchanging data as XML greatly reduces this complexity, since the data can be read by different incompatible applications.

XML documents must contain a root element. This element is the parent of all other elements. The elements in an XML document form a document tree. The tree starts at the root and branches to the lowest level of the tree. All elements can have sub elements (child elements):

```
<root>
  <child>
    <subchild>.....</subchild>
  </child>
</root>
```

The terms parent, child, and sibling are used to describe the relationships between elements. Parent elements have children. Children on the same level are called siblings (brothers or sisters). All elements can have text content and attributes. Fig 1 shows the sample XML document.

The root element in the example is <bookstore>. All <book> elements in the document are contained within <bookstore>. The <book> element has 4 children: <title>,< author>, <year>, <price>.

3. Association Rule Mining

The definition by Agrawal et al (R. Agrawal, T. Imielinski & A. Swami. *1993.*) the problem of association rule mining is defined as: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of *n* binary attributes called *items*. Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the *database*. Each transaction in *D* has a unique transaction ID and contains a subset of the items in *I*. A *rule* is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$and $X \cap Y = \emptyset$. The sets of items *X* and *Y* are called *antecedent* (left-hand-side) and *consequent* (right-hand-side) of the rule.

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence. The *support* supp(*X*) of an itemset *X* is defined as the proportion of transactions in the data set which contain the itemset. The *confidence* of a rule is defined

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X).$$

Confidence can be interpreted as an estimate of the probability $P(Y \mid X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS

The interestingness of an association rule can be defined in terms of the measure associated with it, as well as in the form an association can be found. The most common framework in the association rules generation is the “support-confidence” one. Although these two parameters allow the pruning of many associations that are discovered in data, there are cases when many uninteresting rules may be produced. The measure interest is used to discover interesting rules.

$$\mathrm{Interest}(X, Y) = \mathrm{Support}(X \cup Y) - \mathrm{Support}(X) .\mathrm{Support}(Y)$$

4. Chi-Square Test

Generally speaking, the chi-square test is a statistical test (Glenn A. Walker.) used to examine differences with categorical variables. There are a number of features of the social world we characterize through categorical variables - religion, political preference, etc. To examine hypotheses using such variables, use the chi-square test.

The chi-square test is used in two similar but distinct circumstances:

For estimating how closely an observed distribution matches an expected distribution - we'll refer to this as the goodness-of-fit test

For estimating whether two random variables are independent.

4.1 Goodness-of -fit test

The chi-square test is a “goodness of fit” test: it answers the question of how well do experimental data fit expectations. The chi-square test of independence can be used for any variable; the group (independent) and the test variable (dependent) can be nominal, ordinal, or grouped interval.

The following algorithm illustrates calculating a goodness-of-fit test with chi-square:

1) Establish hypotheses.

2) Calculate chi-square statistic. Doing so requires knowing:

The number of observations

Expected values

Observed values

3) Assess significance level. Doing so requires knowing the number of degrees of freedom.

4) Finally, decide whether to accept or reject the null hypothesis.

4.2 Testing Independence

The other primary use of the chi-square test is to examine whether two variables are independent or not. What does it mean to be independent, in this sense? It means that the two factors are not related. Typically in social science research, we're interested in finding factors that are related - education and income, occupation and prestige, age and voting behavior. In this case, the chi- square can be used to assess whether two variables are independent or not. More generally, we say that variable Y is "not correlated with" or "independent of" the variable X if more of one is not associated with more of another. If two categorical variables are correlated their values tend to move together, either in the same direction or in the opposite.

5. Anomaly detection in XML documents

The system is based on rules that define signatures and it detects anomalies that fall in abnormal signature profile. Fig 1 shows the anomaly detection in XML documents.

5.1 Two-dimensional (2-D) representation of XML documents

An XML document is defined (Jong P. Yoon, Vijay Raghavan, Venu Chakilam. 2001.) as a sequence of elementary paths with associated element contents. An elementary path is a sequence of nested elements where the most nested element is simple content element. In a two-dimensional representation of XML documents a row represents an XML document and column represents an elementary paths. Fig 2 and 3 show the sample mushroom XML document and 2-D representation of XML document

In the above example the elementary paths are

ep1 = <Mushroom><Cap><Shape>

ep2 = <Mushroom><Cap><Surface>

ep3 = <Mushroom><Cap><Color>

ep3 = <Mushroom><Brusises>

ep4 = <Mushroom><Odor>

ep5 = <Mushroom><Gill><Attachment> etc.,

5.2 Mining Negative Association rules

There are rules that imply negative relationships such rules are called Negative Association Rules. A negative association rule (Xindong wu, Shichao zhang. 2004.) also describes relationships between item sets and implies the occurrence of some item sets characterized by the absence of others. To eliminate unwanted rules and focus on potential interesting ones, the system predict possible interesting negative association rules by incorporating domain knowledge of the data sets. The negative association rule can be written in the form X→¬Y, ¬X→Y, , where X and Y are itemsets.

Measures for X→¬Y

$$\text{Support}(X \rightarrow \text{not } Y) = \text{support}(X) - \text{support}(X \rightarrow Y)$$

$$\text{Confidence}(X \rightarrow \text{not } Y) = 1 - \text{conf}(X \rightarrow Y)$$

$$\text{Interest}(X, \text{not } Y) = \text{Support}(X \cup \text{ not } Y) - \text{Support}(X).\text{Support}(\text{ not } Y)$$

$$\text{CPIR}(\text{not } Y/X) = \frac{\text{Support}(X \cup \textit{not}Y) - \text{Support}(X).\text{Support}(\text{ not } Y)}{\text{Support}(X)(1 - \text{Support}(\text{not } Y))}$$

Measures for $\neg X \rightarrow Y$

$$\text{Support}(\text{not } X \rightarrow Y) = \text{support}(Y) - \text{support}(Y \rightarrow X)$$

$$\text{Confidence}(\text{ not } X \rightarrow Y) = 1 - \text{conf}(Y \rightarrow X)$$

$$\text{Interest}(\text{not } X, Y) = \text{Support}(\text{not } X \cup Y) - \text{Support}(\text{not } X).\text{Support}(Y)$$

Interesting Negative Rul $\text{CPIR}(Y/_{\text{not } X}) = \dfrac{\text{Support}(\text{not } X \cup Y) - \text{Support}(\text{not } X).\text{Support}(Y)}{\text{Support}(\text{not } X)(1 - \text{Support}(Y))}$

$$\begin{aligned}&\text{if Support}(X \cup \text{not } Y) \geq \text{ms and Support}(X) \geq \text{ms and Support}(Y) \geq \text{ms and}\\&\text{Interest}(X, \text{not } Y) \geq \text{mi and CPIR}(\textit{not } Y/_{X}) \geq \text{mc}\\&\text{then } X \rightarrow \text{not } Y\end{aligned}$$

Interesting Negative Rules for $\neg X \rightarrow Y$

$$\begin{aligned}&\text{if Support}(X \cup \text{not } Y) \geq \text{ms and Support}(Y) \geq \text{ms and support}(x) \geq m\\&\text{Interest}(\text{ not } X, Y) \geq \text{mi and CPIR}(Y/_{\textit{not}X}) \geq \text{mc}\\&\text{then not } X \rightarrow Y\end{aligned}$$

ms – minimum support threshold and mc – minimum confidence threshold

From the above measures interesting negative association rules are identified and these antecedents and consequents are applied for chi-square test to identify the independency.

5.3 Chi-Square test

Chi-square test is a statistical test to verify the independence between two variables. Using Chi-square test, independency between the two variables are identified by finding contingency tables and expected frequencies

The following is a contingency table, a tabular representation of a rule.

R1 and R2 represent the Boolean states of an antecedent for the conclusions C1 and C2. The X11, X12, X21, X22 represent the frequencies of each antecedent-consequent pair. The R1T, R2T, CT1, CT2 are the marginal sums of the rows and columns, respectively.

5.3.1 Calculating Chi-square:

1) Calculate and fix the sizes of the marginal sums,

2) Calculate the total frequency, T, using the marginal sums.

3) Calculate the expected frequencies for each cell

Formula:

$$eij = \sum \frac{(R_{ij} - C_{Tj})^2}{T}$$

Where C_{Tj} and R_{ij} are the row total for ith row and the column total for jth column.

4) Select the test to be used to calculate χ^2 based on the highest expected frequency, m.

5) Calculate χ^2 using the chosen test.

6) Calculate the degrees of freedom. df=(r-1)(c-1)

A critical factor in using the chi-square test is the “degrees of freedom”, which is essentially the number of independent random variables involved.

7) Use a chi-square table with χ^2 and df to determine if the conclusions are independent from the antecedent at the selected level of significance

8) For the selected level of significance α

if $\chi^2 > \chi^2_\alpha$

Reject the null hypothesis of independence

else

Accept the null hypothesis of independence

5.4 Signature generation

From the chi-square test the strongly independent attributes are identified and are stored in the form of XML file as a signature profile. For example

<cap_shape:b>→⌉<cap_surface:r>

(support=20% CPIR=70% interest=50%>

For the above negative association rule the abnormal profile in XML format is shown in fig 4

Using XQUERY the test data is checked with abnormal signature profile for detecting anomalies.

Fig 5 shows the proposed system model.

6. Discussion and Experiment Results

Giulia et al (Giulia Bruno, Paolo Garza, Elisa Quintarelli, Rosalba Rossato. 2007.) identifies the anomalies in simple nested XML documents using quasi-functional dependencies and association rule The system query the original datasets to extract the instances that violate the dependences and for each quasi-functional dependency relating the sets X and Y query all the stored association rules that involve X c and Y, with a low confidence (i.e., with a confidence lower than a fixed threshold).

Proposed system identifies anomalies using negative association rules and enhanced with chi-square test. The anomalies are identified whose confidence value is grater than minimum confidence threshold in negative association rule which is improved with chi-square test.

The proposed system uses XML Mushroom data set includes 8124 hypothetical samples corresponding to 23 species of gilled mushrooms. Table 2,3 & 4 show the number of rules generated by the proposed system. The number of rules generated by the system is low because the uninteresting rules are filtered by interesting measure and chi-square independence test. The anomalies are identified between two attribute levels.

References

Giulia Bruno, Paolo Garza, Elisa Quintarelli, Rosalba Rossato, (2007), 'Anomaly Detection in XML databases by means of Association Rules', 18th international workshop on Database and Expert Systems Applications , pp.387-391, IEEE .

Glenn A. Walker**, '**Common Statistical Methods for Clinical Research with SAS® Examples', SAS publishing, Second Edition.

Jiawei Han, Micheline Kamber. (2004). 'Data mining Concepts and Techniques', Simon Fraser university, Reprinted.

Jochen Hipp, Ulrich Güntzer, & Gholamreza Nakhaeizadeh. Algorithms for association rule mining - A general survey and comparison. *SIGKDD Explorations*, 2(2):1-58, 2000.

Jong P. Yoon, Vijay Raghavan, Venu Chakilam. (2001). 'BitCube: A Three-Dimensional Bitmap Indexing for XML Documents', *Journal of Intelligent Information Systems*, pp. 241-254(14).

R. Agrawal, T. Imielinski & A. Swami. *Mining Association Rules.* Between Sets of Items in Large Databases", SIGMOD Conference 1993: 207-216.

Xindong wu, Shichao zhang. (2004). 'Efficient Mining of Both Positive and Negative Association Rules', ACM Transactions on Information Systems, Vol. 22, No. 3.

Table 1. Contingency Tables

	C1	C2	Marginal Sums
R1	X11	X12	R1T=X11+X12
R2	X21	X22	R2T=X21+X22
Marginal Sums	CT1=X11+X21	CT2=X12+X22	T=X11+X12+ X21+X22

Table 2. Number of rules generated at minimum support =0.2

Minimum Confidence	**Positive Association rule (Confidence < mc)**	**Interesting Negative Association rule (Confidence > mc)**
0.6	333	66
0.7	409	50
0.8	460	43
0.9	513	32

Table 3. Number of rules generated at minimum support =0.3

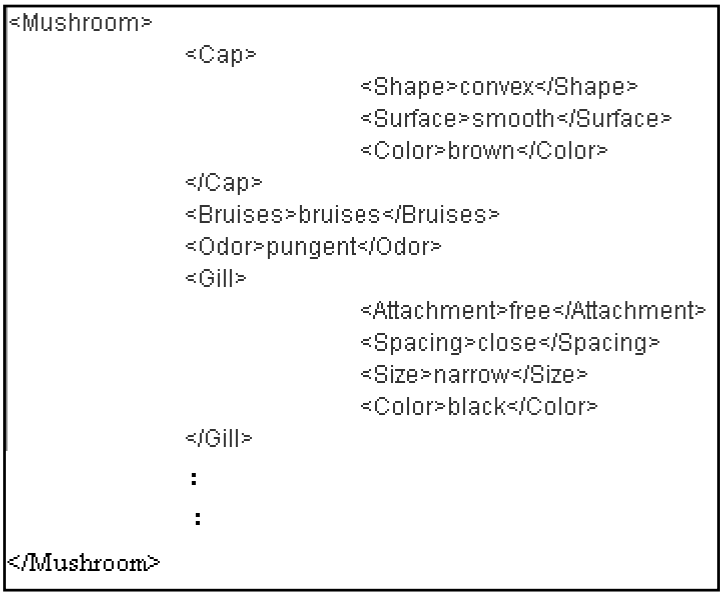
Minimum Confidence	**Positive Association rule (Confidence < mc)**	**Interesting Negative Association rule (Confidence > mc)**
0.6	92	47
0.7	137	34
0.8	162	29
0.9	191	24

Table 4. Number of rules generated at minimum support Level=0.4

Minimum Confidence	**Positive Association rule (Confidence < mc)**	**Interesting Negative Association rule (Confidence > mc)**
0.6	48	0
0.7	70	25
0.8	82	24
0.9	100	21

```
<bookstore>
<book>
<category="COOKING">
<title lang="en">Everyday Italian</title>
<author>Giada De Laurentiis</author>
<year>2005</year>
<price>30.00</price>
</book>
<book>
<category="CHILDREN">
<title lang="en">Harry Potter</title>
<author>J K. Rowling</author>
<year>2005</year>
<price>29.99</price>
</book>
<book>
<category="WEB">
<title lang="en">Learning XML</title>
<author>Erik T. Ray</author>
<year>2003</year>
<price>39.95</price>
</book>
            </bookstore>
```

Figure 1. Sample XML document



```
<Mushroom>
        <Cap>
                <Shape>convex</Shape>
                <Surface>smooth</Surface>
                <Color>brown</Color>
        </Cap>
        <Bruises>bruises</Bruises>
        <Odor>pungent</Odor>
        <Gill>
                <Attachment>free</Attachment>
                <Spacing>close</Spacing>
                <Size>narrow</Size>
                <Color>black</Color>
        </Gill>
         :
         :
</Mushroom>
```

Figure 2. Sample mushroom document

	ep1	ep2	ep3	ep4	ep5	ep6																	ep23
D1	c	s	b	b	p	f																	

Figure 3. 2-D representation of Mushroom document

```
<Nrule support = 20% CPIR = 70% interest=50%>
        <antecedent>

<cap_shape>b</cap_shape>
        </antecedent>
        <consequent>

<cap_surface>r</cap_surface>
        </consequent>
</Nrule>
```

Figure 4. Format of abnormal signature profile

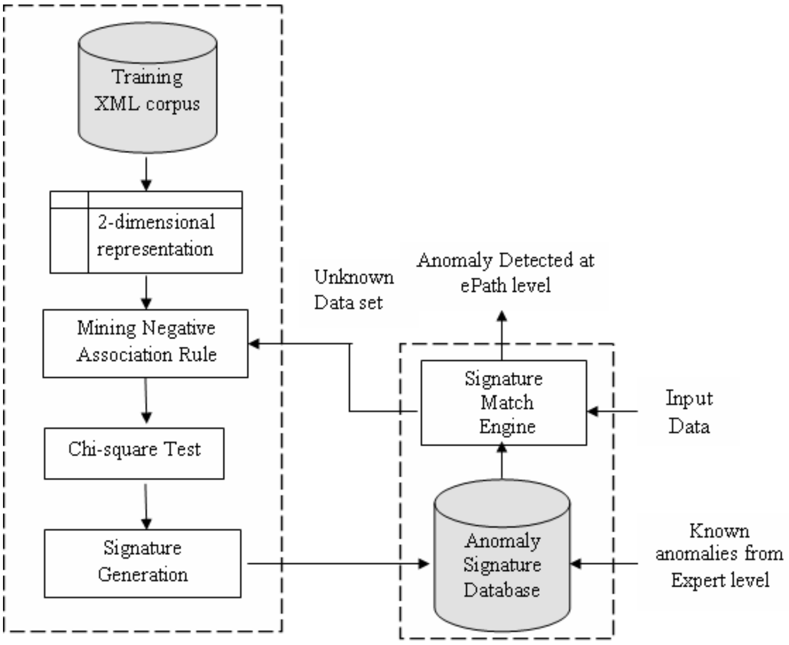

Training XML corpus
2-dimensional representation
Mining Negative Association Rule
Chi-square Test
Signature Generation
Unknown Data set
Anomaly Detected at ePath level
Signature Match Engine
Input Data
Anomaly Signature Database
Known anomalies from Expert level

Figure 5. Anomaly detection in XML documents