# Discovering surprising patterns by detecting occurrences of Simpson's paradox


Carem C. Fabris
CEFET-PR
CPGEI
Av. Sete de Setembro, 3165.
Curitiba-PR, 80230-901. Brazil
carem@dainf.cefetpr.br

Alex A. Freitas
PUC-PR
PPGIA-CCET
Rua Imaculada Conceicao, 1155
Curitiba-PR, 80215-901. Brazil
alex@ppgia.pucpr.br



**Abstract**

This paper addresses the discovery of surprising patterns. Recently, several authors have addressed the task of discovering surprising prediction rules. However, we do not focus on prediction rules, but rather on a quite different kind of pattern, namely the occurrence of Simpson's paradox. Intuitively, the fact that this is a paradox suggests that it has a great potential to be a surprising pattern for the user. With this motivation, we make the detection of Simpson's paradox the very goal of a data mining algorithm explicitly designed to discover surprising patterns. We present computational results showing surprising occurrences of the paradox in some public-domain data sets. In addition, we propose a method for ranking the discovered instances of the paradox in decreasing order of estimated degree of surprisingness.


## 1 Introduction.

Surveying the data mining literature, one can say that most algorithms focus on the discovery of accurate, comprehensible knowledge, even though comprehensibility is a kind of subjective concept which is rarely discussed in depth in the literature.

Ideally, however, a data mining algorithm should go beyond the discovery of accurate, comprehensible knowledge. The discovered knowledge should be also interesting, or surprising. The classic example to show the importance of surprisingness is the hypothetical discovery, in a medical database, of the pattern: "if the patient is pregnant, then the sex of the patient is female." Clearly, this pattern is highly accurate and highly comprehensible, but it is useless for the user, because it states the obvious.

Recently, several authors have focused on the important issue of discovering surprising/interesting patterns – see e.g. [Padmanabhan & Tuzhilin 98], [Suzuki & Kodratoff 98], [Freitas 98], [Freitas 99], [Dong & Li 98], [Suzuki & Shimura 97], [Liu et al. 97], [Liu & Hsu 96], [Silberschatz & Tuzhilin 96], [Subramonian 98]. The majority of this work focus on the discovery of surprising prediction rules.

This paper is one more stride towards the discovery of surprising patterns. However, we do not focus on prediction rules, but rather on a quite different kind of pattern, namely the occurrence of Simpson's paradox. Intuitively, the fact that this is a paradox suggests that it has a great potential to be a surprising pattern for the user.

Although some authors have alerted for the pitfalls associated with Simpson's paradox in the context of data mining - see e.g. [Glymour et al. 97] - little research has been done on how to cope with occurrences of this paradox, or how to modify a data mining algorithm to cope with this paradox. This paper follows an intriguing direction, which consists of making the detection of Simpson's paradox the very goal of a data mining algorithm explicitly designed to discover surprising patterns.

This paper is organized as follows. Section 2 presents a review of Simpson's paradox. Section 3 describes our approach for detecting occurrences of Simpson's paradox and estimate their degree of surprisingness for the user. Section 4 presents computational results. Finally, section 5 summarizes and discusses the results reported in the paper.

# 2 A Review of Simpson's Paradox.

Let a population be partitioned into two mutually exclusive and exhaustive populations, denoted $Pop_1$ and $Pop_2$, according to the value of a given binary attribute, denoted 1stPartAtt (First Partitioning Attribute). Let G be a binary goal attribute, which takes on a value indicating whether or not a given situation of interest has occurred in a population, and let $G_1$ and $G_2$ be the value of the attribute G in each of the respective populations $Pop_1$ and $Pop_2$. Let $Pr(G_1)$ and $Pr(G_2)$ denote the probability that the situation of interest has occurred in $Pop_1$ and $Pop_2$, respectively. Assume that $Pr(G_1) > Pr(G_2)$.

Let us now consider the case where both the populations $Pop_1$ and $Pop_2$ are further partitioned, in parallel, according to the value of a given categorical attribute, denoted 2ndPartAtt. Let this attribute have m distinct categorical values. We can now compute the probability Pr(G) in each population, for each of these m categories, which we denote by $G_{ij}$, where i=1,2 is the id of the population and j=1,...,m is the id of the value of 2ndPartAtt. Let $Pr(G_{1j})$ and $Pr(G_{2j})$ denote the probability that the situation of interest has occurred in $Pop_1$ and $Pop_2$, in the j-th category of 2ndPartAtt, j=1,...,m.

Finally, Simpson's paradox occurs when, although the overall value of Pr(G) is higher in $Pop_1$ than in $Pop_2$, i.e. $Pr(G_1) > Pr(G_2)$, in *each* of the categories produced by 2ndPartAtt the value of Pr(G) in $Pop_1$ is lower than or equal to its value in $Pop_2$, i.e. $Pr(G_{1j}) \leq Pr(G_{2j})$, j=1,...,m. The paradox also occurs in the dual situation, i.e. when $Pr(G_1) < Pr(G_2)$ but $Pr(G_{1j}) \geq Pr(G_{2j})$, j=1,...,m [Simpson 51].

Some real-life examples of the occurrence of this paradox are mentioned in [Wagner 82], [Newson 91]. For instance, the paradox occurred in a comparison of tuberculosis deaths in New York City and Richmond, Virginia, during the year 1910, as shown in Table 1. Overall, the tuberculosis mortality rate of New York was lower than Richmond's one. However, the opposite was observed when the data was further partitioned according to two racial categories: white and non-white. In both the white and non-white categories, Richmond had a lower mortality rate. In terms of the above notation, the 1stPartAtt was *city*; the situation of interest measured by attribute G was the *occurrence of death* in a tuberculosis case; and the 2ndPartAtt was *racial category*.

**Table 1:** Simpson's paradox in data about tuberculosis deaths.

| | New York | | Richmond | |
|---|---|---|---|---|
| **Total Population** | 4766883 | | 127682 | |
| **No. of deaths** | 8878 | | 286 | |
| **Percentage** | 0.19% | | 0.22% | |
| | **New York** | | **Richmond** | |
| | **white** | **non-w.** | **white** | **non-w.** |
| **Total Population** | 4675174 | 91709 | 80895 | 46733 |
| **No. of deaths** | 8365 | 513 | 131 | 155 |
| **Percentage** | 0.18% | 0.56% | 0.16% | 0.33% |

# 3 Discovering occurrences of Simpson's paradox.

As mentioned in the introduction, some authors have alerted for the pitfalls associated with Simpson's paradox in the context of data mining - see e.g. [Glymour et al. 97]. In particular, the occurrence of this paradox in a given data set can easily fool a data mining algorithm, causing the algorithm to misinterpret a given relationship between some attributes.

For instance, decision-tree learners usually build a tree by selecting one attribute at a time. Hence, they can select an attribute that seems to have a certain relationship with a given class, when in reality the true relationship (taking into account attribute interactions) is the reverse of the apparent one.

Note that this monothetic nature (i.e. selecting one attribute at a time) is not unique to decision trees. Many other rule induction algorithms essentially work in this way, and so are also quite sensitive to the occurrence of Simpson's paradox.

Although data mining practitioners are starting to become more aware of the pitfalls associated with Simpson's paradox, little research has been done on how to cope with occurrences of this paradox, or how to modify a data mining algorithm to cope with this paradox.

This paper follows an intriguing research direction which, to the best of our knowledge, was first suggested in [Freitas 98]. The basic idea is that, instead of trying to design an algorithm which avoids or mitigates the pitfalls associated with Simpson's paradox, we make the discovery of Simpson's paradox the very goal of the data mining algorithm.

The motivation can be summarized as follows. Simpson's paradox is, as the name implies, a paradox, and so it tends to be very surprising for the user.

The work described and empirically evaluated in this paper essentially consists of two parts. First, we discover occurrences of the paradox hidden in a data set. Second, we evaluate the "magnitude" (an estimate for the degree of surprisingness) of each discovered instance of the paradox, in order to report them to the user in decreasing order of magnitude (estimated degree of surprisingness). This approach gives the user flexibility to decide how many occurrences of the paradox (s)he will analyze, depending on her/his available time.

The rest of this section is divided into two parts. Section 3.1 describes an algorithm for discovering occurrences of Simpson's paradox. This algorithm is also described in our previous work [Freitas 98], but it is included here to make the paper self-contained. Section 3.2 proposes a way to evaluate the “magnitude” of each discovered paradox - which was not addressed in our previous work - in an attempt to identify the most surprising instances of the paradox.

Finally, section 4 will present computational results reporting discovered occurrences of the paradox in several data sets - we note in passing that the above-mentioned previous work only proposed the algorithm, without reporting any computational results.

## 3.1 An algorithm for detecting occurrences of Simpson’s paradox .

The search for occurrences of Simpson’s paradox can be performed by the Algorithm 1 below. The input for the algorithm is a list $L_G$ of user-defined binary goal attributes, each of them indicating whether or not a given situation of interest has occurred. The algorithm below is specified in a high level of abstraction, so the two statements that identify the attributes to be put in lists $L_1$ and $L_2$ can be expanded in different procedures, using different criteria, as long as three conditions hold: (a) all attributes in $L_1$ are binary; (b) all attributes in $L_2$ are categorical; (c) any goal attribute contained in $L_G$ does not appear in $L_1$ nor in $L_2$. Note that these conditions are not very strict, and in particular they allow the possibility that an attribute is contained in both $L_1$ and $L_2$ (since binary attributes are a particular case of categorical attributes). This possibility justifies the use of the condition $A_2 \neq A_1$ in the third FOR EACH statement of Algorithm 1. In practice, this and other more strict conditions may be directly implemented in the two statements that identify the attributes to be put in lists $L_1$ and $L_2$, when Algorithm 1 is refined to achieve a particular implementation.

```
INPUT: list of user-defined goal attributes, denoted L_G
BEGIN
  identify attributes that can be used as 1stPartAtt and put them in list L_1
  identify attributes that can be used as 2ndPartAtt and put them in list L_2
  FOR EACH goal attribute G in L_G
     FOR EACH attribute A_1 in L_1
        partition population into Pop_1 and Pop_2, according to values of A_1
        Pr(G_1) = Pr(G="yes"|A_1=1)
        Pr(G_2) = Pr(G="yes"|A_1=2)
        FOR EACH attribute A_2 in L_2 such that A_2 ≠ A_1
          FOR  i=1,2
            partition Pop_i into m new populations Pop_i1 ... Pop_im,
              according to the values of A_2
            FOR j=1,...,m
              Pr(G_ij) = Pr(G="yes"|A_1=i,A_2=j)
          IF ( Pr(G_1) > Pr(G_2)  AND  Pr(G_1j) ≤ Pr(G_2j), j=1,...,m )
           OR ( Pr(G_1) < Pr(G_2)  AND  Pr(G_1j) ≥ Pr(G_2j), j=1,...,m )
            report the occurrence of the paradox to the user
END
```

**Algorithm 1:** Search for occurrences of Simpson’s paradox.

## 3.2 Ranking the discovered occurrences of Simpson's paradox according to their degree of surprisingness.

The algorithm described in the previous subsection only discovers occurrences of Simpson's paradox in a black-and-white manner. For a given goal attribute and a given pair of partitioning attributes, either the paradox occurs or it does not. In practice, however, different instances of the paradox have different degrees of surprisingness for the user. Ideally, we would like to report the discovered instances of the paradox to the user in decreasing order of degree of surprisingness.

The problem is that surprisingness is, to a significant extent, a subjective (user-driven), domain-dependent concept. It is, however, both possible and desirable to develop an objective (data-driven), *domain-independent* measure of discovered-pattern surprisingness [Freitas 98], [Silberschatz & Tuzhilin 96]. In this spirit, we propose below a method to compute the "magnitude" of each instance of the paradox and use it as a measure of the degree of surprisingness of that instance of the paradox.

Once we have assigned to each discovered instance of the paradox a degree of surprisingness, we can report the discovered instances of the paradox to the user in decreasing order of degree of surprisingness, as mentioned earlier.

Intuitively, a reasonable measure of paradox magnitude would measure the degree to which the probability of the situation of interest was reversed when the second partition of the data was performed. Following the terminology and notation of section 2, this intuition can be better explained as follows.

If after the first partition of the data the probability of the situation of interest increases (decreases) very much from $Pop_1$ to $Pop_2$, and after the second partition of the data that probability decreases (increases) very much from $Pop_1$ to $Pop_2$ for each of the categories of 2ndPartAtt, then one can say that the paradox has a large magnitude. On the other hand, if after the first partition of the data the probability of the situation of interest increases (decreases) very little from $Pop_1$ to $Pop_2$, and after the second partition of the data that probability decreases (increases) very little from $Pop_1$ to $Pop_2$ for each of the categories of 2ndPartAtt, then one can say that the paradox has a small magnitude.

In other words, the larger the magnitude of the paradox, the more the probability of the situation of interest given the first partitioning attribute differs from the probability of the situation of interest given both the first and the second partitioning attributes. Therefore, the larger the magnitude, the larger the estimated degree of surprisingness for the user.

To formalize this intuition, the "magnitude" M of an instance of the paradox is given by the following formula:

$$M = ( M1 + M2 ) / 2, \qquad (1)$$

where M1 measures by how much the probability of the situation of interest increases (decreases) from $Pop_1$ to $Pop_2$ after the first partition of the data and M2 measures by how much that probability decreases (increases) from $Pop_1$ to $Pop_2$ for each of the categories of 2ndPartAtt after the second partition of the data. We take the arithmetic average of M1 and M2, rather than simply taking the sum M1 + M2, mainly for the purpose of normalizing the magnitude M. Provided that both M1 and M2 vary in the range 0..100%, the value of M also varies in the range 0..100%.

The question is how to measure M1 and M2. At first glance, a sensible definition would be:

$$M1 = | \Pr(G_1) - \Pr(G_2) | \quad (2)$$

$$M2 = ( |\Pr(G_{11}) - \Pr(G_{21})| + |\Pr(G_{12}) - \Pr(G_{22})| + \ldots + |\Pr(G_{1m}) - \Pr(G_{2m})| ) / m , \quad (3)$$

where m is the number of values in the domain of 2ndPartAtt.

In formulas (2) and (3) M1 is the absolute value of difference between $\Pr(G_1)$ and $\Pr(G_2)$, the probability that the situation of interest has occurred respectively in $Pop_1$ and $Pop_2$ - the subpopulations created by splitting the data according to the first partitioning attribute. M2 is the average of the absolute values of the differences between $\Pr(G_{1j})$ and $\Pr(G_{2j})$, the probabilities that the situation of interest has occurred respectively in $Pop_1$ and $Pop_2$, in the j-th category of the second partitioning attribute, for j=1,...,m.

However, these definitions of M1 and M2 have a drawback. They take into account only the absolute values of differences between probabilities, and not the relative values of those differences. To see the problem, note that the absolute value of the difference between two probabilities of 5% and 10% is 5%, which is also the absolute value of the difference between two probabilities of 90% and 95%. Intuitively, however, an increase from 5% to 10% has a much larger relative magnitude than an increase from 90% to 95%. So, the former tends to be more surprising for the user, and a good measure of magnitude ought to capture this notion.

Therefore, we propose a modification of formulas (2) and (3) to take into account the relative value of the differences between probabilities. The basic idea is to divide the difference between two probabilities x and y by the maximum value between x and y. (Alternatively, we could obtain a relative measure by using the minimum of x and y, but using the maximum value has the advantage of producing relative measures normalized in the range 0..1.)

Formally, the measures M1 and M2 used in this work are as defined in formulas (4) and (5).

$$M1 = | \Pr(G_1) - \Pr(G_2) | / \max(\Pr(G_1), \Pr(G_2)) \quad (4)$$

$$M2 = \sum_{k=1}^{m} ( |\Pr(G_{1k}) - \Pr(G_{2k})| / \max(\Pr(G_{1k}),\Pr(G_{2k})) ) / m \quad (5)$$

Finally, there is a caveat in the implementation of formula (5). The denominator of any term in the summation can be 0. To avoid division by 0, if any of those terms is 0, the whole term is simply considered to be 0. For instance, if both $\Pr(G_{11})$ and $\Pr(G_{21})$ are 0, the whole term $|\Pr(G_{11}) - \Pr(G_{21})|/\max(\Pr(G_{11}),\Pr(G_{21}))$ is considered to be 0. Note, however, that this term is still counted as one term (contributing for the value of m) for the purpose of taking the arithmetic average of all terms.

Note also that in principle both M1 and M2 can vary from 0 to 100%, so that the magnitude M can also vary from 0 to 100%. In practice values of M close to 100% are unlikely. (In all the experiments reported in this paper - described in section 4 - the highest value found for M was 46.76%.)

## 4 Computational Results.

We have implemented the algorithm described in section 3.1 and have applied it in seven public-domain data sets available from the UCI repository (URL: http://www.ics.uci.edu/AI/ML/Machine-Learning.html). These seven data sets are: Voting Records, Hepatitis, Australian Credit, Nursery, Heart, Car Evaluation, Adult.

We have not discovered any instance of Simpson's paradox in the Heart, Car Evaluation and Adult data sets. However, we have discovered in total 13 instances of the paradox in the other data sets, namely seven instances in Voting Records, two instances in Hepatitis, two instances in Australian Credit, and two instances in Nursery. Due to space limitation, we report here only seven discovered instances of the paradox, namely four instances in Voting Records, two instances in Hepatitis, and one instance in Australian Credit. The instances reported here are the ones with largest magnitude - our estimate for degree of surprisingness – for each data set. (The two instances of the paradox in the Nursery data set had very low – less than 1% - magnitude, that is why no instance is reported for this data set.)

The results are presented in the next three subsections. For the sake of uniformity and compactness, all the tables of results presented in this section have a title with the following format:

**Table N:** Dataset: D, Goal: G = v, 1stPartAtt: A1, 2ndPartAtt: A2, Magnitude: X (M1 = Y; M2 = Z)

where N is the number of the table, D is the name of the data set, G is the name of the goal attribute and v is the target value of G (i.e. the condition "G = v" defines the situation of interest), A1 and A2 are the names of the first and second partitioning attributes, X is the magnitude of the paradox – computed by formula (1) – and Y and Z are the values of the terms M1 and M2 used to compute magnitude, as defined in formulas (4) and (5).

Several of the data sets used in the experiments had missing attribute values. We coped with this problem as follows. When computing the probabilities of the situation of interest, an example was ignored – i.e. it was not counted for probability-computation purposes - if either the first partitioning attribute or the second partitioning attribute of that example had a missing value.

### 4.1 Results for the Voting Records data set.

This data set contains 435 examples and 16 binary attributes. The goal attribute predicts whether the Congressman is democrat or republican. The situation of interest used in our experiments was class = democrat. The four largest-magnitude instances of the paradox for this data set are reported in Tables 2 through 5.

The first two instances of the paradox, reported in Tables 2 and 3, have a large value of magnitude, so we estimate they are very surprising for the user. For instance, the upper part of Table 2 shows that the probability of a Congressman being a democrat given that (s)he voted "yea" (Yes) for export-administration-act-south-africa is much greater than the probability of a Congressman being a democrat given that (s)he voted "Nay" (No) for the same issue. However, the lower part of this Table shows that this is a deceptive relationship, once it is reversed when we take into account the interaction between the votes for export-administration-act-south-africa and for physician-fee-freeze.

**Table 2:** Dataset: Voting, Interest: class = democrat,
1stPartAtt: Export-administration-act-south-africa,
2ndPartAtt: Physician-fee-freeze
Magnitude: 46.76% (M1 = 71.67%; M2 = 21.85%)

| | **Export-administration-act-south-africa=Yes** | | **Export-administration-act-south-africa=No** | |
|---|---|---|---|---|
| **Total Populat.** | 264 | | 61 | |
| **No. of Democr.** | 168 | | 11 | |
| **Percentage** | 63.64% | | 18.03% | |
| | **Phys-fee-frz=yes** | **Phys-fee-frz=no** | **Phys-fee-frz=yes** | **Phys-fee-frz=no** |
| **Total Populat.** | 98 | 166 | 55 | 6 |
| **No. of Democr.** | 4 | 164 | 5 | 6 |
| **Percentage** | 5.13% | 98.98% | 8.95% | 100.00% |

**Table 3:** Dataset: Voting, Interest: class = democrat,
1stPartAtt: Anti-satellite-test-ban, 2ndPartAtt: Physician-fee-freeze
Magnitude: 42.32% (M1 = 62.78%; M2 = 21.85%)

| | **Anti-satellite-test-ban=Yes** | | **Anti-satellite-test-ban=No** | |
|---|---|---|---|---|
| **Total Populat.** | 236 | | 177 | |
| **No. of Democr.** | 197 | | 55 | |
| **Percentage** | 83.47% | | 31.07% | |
| | **Phys-fee-frz=yes** | **Phys-fee-frz=no** | **Phys-fee-frz=yes** | **Phys-fee-frz=no** |
| **Total Populat.** | 39 | 197 | 134 | 43 |
| **No. of Democr.** | 2 | 195 | 12 | 43 |
| **Percentage** | 5.13% | 98.98% | 8.95% | 100.00% |

**Table 4:** Dataset: Voting, Interest: class = democrat,
1stPartAtt: Water-project-cost-sharing, 2ndPartAtt: Mx-missile
Magnitude: 13.87% (M1 = 1.98%; M2 = 25.76%)

| | **Water-project-cost-sharing=Yes** | | **Water-project-cost-sharing=No** | |
|---|---|---|---|---|
| **Total Populat.** | 185 | | 183 | |
| **No. of Democr.** | 110 | | 111 | |
| **Percentage** | 59.46% | | 60.66% | |
| | **Mx-miss.=yes** | **Mx-miss.=no** | **Mx-miss.=yes** | **Mx-miss.=no** |
| **Total Population** | 73 | 112 | 110 | 73 |
| **No. of Democr.** | 69 | 41 | 96 | 15 |
| **Percentage** | 94.52% | 36.60% | 87.27% | 20.55% |

**Table 5:** Dataset: Voting, Interest: class = democrat,
1stPartAtt: Water-project-cost-sharing, 2ndPartAtt: El-Salvador-Aid
Magnitude: 11.84% (M1 = 1.53%; M2 = 22.14%)

| | Water-project-cost-sharing=Yes | | Water-project-cost-sharing=No | |
|---|---|---|---|---|
| **Total Populat.** | 188 | | 190 | |
| **No. of Democr.** | 114 | | 117 | |
| **Percentage** | 60.64% | | 61.58% | |
| | **El-Salv.-Aid=yes** | **El-Salv.-Aid=no** | **El-Salv.-Aid=yes** | **El-Salv.-Aid=no** |
| **Total Populat.** | 110 | 78 | 84 | 106 |
| **No. of Democr.** | 37 | 77 | 17 | 100 |
| **Percentage** | 33.64% | 98.72% | 20.24% | 94.34% |

## 4.2 Results for the Hepatitis data set.

This data set contains 155 examples and 19 attributes (most of them categorical). The goal attribute predicts whether the patient will die or live. The situation of interest used in our experiments was class = die.

We have found two instances of Simpson's paradox, reported in Tables 6 and 7. These two instances have a reasonable value of magnitude, namely 20.40% and 15.15%. Hence, they are estimated to be fairly surprising for the user.

For instance, the upper part of Table 6 shows that the probability of a patient dying given that (s)he has anorexia is significantly smaller than the probability of the patient dying given that (s)he does not have anorexia. However, the lower part of this Table shows that this is a deceptive relationship, once it is reversed when we take into account the interaction between the attributes anorexia and spiders.

**Table 6:** Dataset: Hepatitis, Interest: class = die,
1stPartAtt: Anorexia, 2ndPartAtt: Spiders
Magnitude: 20.40% (M1 = 38.9%; M2 = 1.89%)

| | Anorexia=no | | Anorexia=yes | |
|---|---|---|---|---|
| **Total Popul.** | 30 | | 120 | |
| **No. of deaths** | 9 | | 22 | |
| **Percentage** | 30.00% | | 18.33% | |
| | **Spid=no** | **Spid=yes** | **Spid=no** | **Spid=yes** |
| **Total Popul.** | 19 | 11 | 32 | 88 |
| **No. of deaths** | 8 | 1 | 14 | 8 |
| **Percentage** | 42.10% | 9.09% | 43.75% | 9.09% |

**Table 7:** Dataset: Hepatitis, Interest: class = die,
1stPartAtt: Liver Firm, 2ndPartAtt: Varices
Magnitude: 15.15%    (M1 = 22.15%; M2 = 8.14%)

| | **Liver Firm=no** | | **Liver Firm=yes** | |
|---|---|---|---|---|
| **Total Popul.** | 60 | | 83 | |
| **No. of deaths** | 13 | | 14 | |
| **Percentage** | 21.67% | | 16.87% | |
| | **Varic=no** | **Varic=yes** | **Varic=no** | **Varices=yes** |
| **Total Popul.** | 12 | 48 | 6 | 77 |
| **No. of deaths** | 7 | 6 | 4 | 10 |
| **Percentage** | 58.33% | 12.50% | 66.67% | 12.99% |

## 4.3 Results for the Australian Credit data set.

This data set contains 690 examples and 14 attributes (6 continuous, 8 categorical). The original goal attribute can take on two values, class 1 and class 2 – the classes and all attribute names and values were changed to meaningless symbols, to protect the confidentiality of the data. The situation of interest used in our experiments was class = 2.

The largest-magnitude instance of the paradox is reported in Table 8. This instance has a magnitude of 7.15%. Hence, it is estimated to be little surprising for the user.

**Table 8:** Dataset: Australian Credit, Interest: class = 2,
1stPartAtt: A11, 2ndPartAtt: A8
Magnitude: 7.15%    (M1 = 6.82%; M2 = 7.47%)

| | **A11=0** | | **A11=1** | |
|---|---|---|---|---|
| **Total Popul.** | 374 | | 316 | |
| **No. of class 2** | 161 | | 146 | |
| **Percentage** | 43.05% | | 46.20% | |
| | **A8=0** | **A8=1** | **A8=0** | **A8=1** |
| **Total Popul.** | 194 | 180 | 135 | 181 |
| **No. of class 2** | 14 | 147 | 9 | 137 |
| **Percentage** | 7.22% | 81.67% | 6.67% | 75.69% |

## 5 Summary and Discussion.

To the best of our knowledge, [Freitas 98] was the first work to propose the idea of making the detection of Simpson's paradox the very goal of a data mining algorithm explicitly designed to discover surprising patterns.

As mentioned in section 3, this paper extends that previous work in two ways. First, we have reported computational results for the algorithm proposed in that previous work (where no computational results were reported). Second, we have proposed a new way to evaluate the "magnitude" of each discovered paradox in an attempt to identify the most surprising instances of the paradox. Discovered instances of the paradox can then be reported to the user in decreasing order of magnitude (estimated surprisingness). This approach gives the user flexibility to decide how many occurrences of the paradox (s)he will analyze, depending on her/his available time.

We have applied our algorithm to detect occurrences of Simpson's paradox in seven UCI data sets. In three of the data sets (Heart, Car Evaluation, Adult) we have found no occurrence of the paradox. In the other four data sets we have found 13 instances of Simpson's paradox with different degrees of magnitude. More precisely, we have found instances of the paradox with a large magnitude in the Voting Records data set, instances with a reasonably large magnitude in the Hepatitis data set, and instances with small magnitudes in the Australian Credit and Nursery data sets. Therefore, some of the instances of the paradox found in the Voting Records and the Hepatitis data sets are estimated to be truly surprising for the user.

It should be noted that, out of the seven data sets mined in this paper, the Voting Records has led not only to the discovery of the most surprising instances of the paradox but also to the largest number of detected instances of the paradox. This result can be at least partially explained by the facts that: (a) in this data set all attributes are binary; (b) the algorithm described in this paper requires that the first partitioning attribute be binary and the second partitioning attribute be categorical. Therefore, in this data set there is a larger number of attributes that can be used as the first-partitioning attribute, in comparison with other mined data sets. In addition, since binary attributes are a special case of categorical attributes, there is also a large number of attributes that can be used as the second-partitioning attribute.

Note also that the patterns discovered by the algorithm described in this paper do not have the same predictive power as other "stronger" kinds of patterns, such as prediction and classification rules. However, the patterns discovered in this paper satisfy not only the requirement of comprehensibility but also the requirement of surprisingness.

Concerning comprehensibility, the discovered patterns allow the user to understand better the interaction between a pair of predicting attributes (with respect to the prediction of a user-specified goal attribute) by using essentially the simple notion of a probability being greater than or smaller than another probability. This seems quite comprehensible, at least in the conventional sense of the term comprehensibility in the data mining literature, which is associated with syntactical simplicity and the use of simple comparison operators.

Concerning surprisingness, the discovered patterns directly represent instances of a paradox, which is, almost by definition, a kind of surprising knowledge. We also emphasize that the algorithm described in this paper was designed from the scratch to discover surprising patterns. Actually, we believe that, in some sense, the algorithm described here is a stride towards the desirable goal of performing a true knowledge *discovery* task. In contrast, several data mining algorithms described in the literature are essentially variations of conventional machine *learning* algorithms designed to discover accurate, comprehensible (but not necessarily *surprising*) knowledge.

## References.

[Dong & Li 98] G. Dong & J. Li. Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. *Research and*

*Development in Knowledge Discovery & Data Mining (Proc. 2$^{nd}$ Pacific-Asian Conf., PAKDD-98). LNAI 1394*, 72-86. Springer-Verlag, 1998.

[Freitas 98] A.A. Freitas. On objective measures of rule surprisingness. *Principles of Data Mining and Knowledge Discovery: Proc. 2nd European Symp. (PKDD'98). LNAI 1510*, 1-9. Nantes, France, Sep. 1998.

[Freitas 99] A.A. Freitas. On rule interestingness measures. *To appear in Knowledge-Based Systems journal*, 1999.

[Glymour et al. 97]. C. Glymour, D. Madigan, D. Pregibon and P. Smyth. Statistical themes and lessons for data mining. *Data Mining and Knowl. Discov.* 1(1), 11-28. 1997.

[Liu & Hsu 96] B. Liu & W. Hsu. Post-analysis of learned rules. *Proc. 1996 Nat. Conf. American Assoc. for Artificial Intelligence (AAAI-96)*, 828-834. AAAI Press, 1996.

[Liu et al. 97] B. Liu, W. Hsu and S. Chen. Using general impressions to analyze discovered classification rules. *Proc. 3rd Int. Conf. Knowledge Discovery & Data Mining*, 31-36. AAAI, 1997.

[Newson 91] G. Newson. Simpson´s paradox revisited. *The Mathematical Gazette 75(473)*, 290-293. Oct. 1991.

[Padmanabhan & Tuzhilin 98] B. Padmanabhan and A. Tuzhilin. A belief-driven method for discovering unexpected patterns. *Proc. 4th Int. Conf. Knowledge Discovery & Data Mining (KDD-98)*, 94-100. AAAI Press, 1998.

[Silberschatz & Tuzhilin 96] A. Silberschatz & A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Trans. Knowledge & Data Engineering*, 8(6), 970-974, Dec./1996.

[Simpson 51] E.H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13,* 238-241. 1951.

[Subramonian 98] R. Subramonian. Defining *diff* as a data mining primitive. *Proc. 4th Int. Conf. Knowledge Discovery & Data Mining,* 334-338. AAAI, 1998.

[Suzuki & Shimura 97] E. Suzuki. Autonomous discovery of reliable exception rules. *Proc. 3rd Int. Conf. Knowledge Discovery & Data Mining*, 259-262. AAAI Press, 1997.

[Suzuki & Kodratoff 98] E. Suzuki & Y. Kodratoff. Discovery of surprising exception rules based on intensity of implication. *Proc. 2nd European Symp. Principles of Data Mining and Knowledge Discovery (PKDD'98). LNAI 1510*, 10-18. Nantes, France, Sep. 1998.

[Wagner 82] C.H. Wagner. Simpson's paradox in real life. *The American Statistician*, 36(1), Feb. 1982, 46-48.