Homogeneous Discoveries Contain no Surprises: Inferring Risk-profiles from Large Databases

Arno Siebes

Computer Science/Department of Algorithmics and Architecture

CS-R9430 1994

Homogeneous Discoveries Contain no Surprises: Inferring Risk-profiles from Large Databases

Arno Siebes (arno@cwi.nl)

CWI

P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Abstract

Many models of reality are probabilistic. For example, not everyone orders crisps with their beer, but a certain percentage does. Inferring such probabilistic knowledge from databases is one of the major challenges for data mining.

Recently Agrawal et al. [1] investigated a class of such problems. In this paper a new class of such problems is investigated, viz., inferring risk-profiles. The proto-typical example of this class is: "what is the probability that a given policy-holder will file a claim with the insurance company in the next year". A risk-profile is then a description of a group of insurants that have the same probability for filing a claim.

It is shown in this paper that homogeneous descriptions are the most plausible risk-profiles. Moreover, under modest assumptions it is shown that covers of such homogeneous descriptions are essentially unique. A direct consequence of this result is that it suffices to search for the homogeneous description with the highest associated probability.

The main result of this paper is thus that we show that the inference problem for risk-profiles reduces to the well studied problem of maximising a quality function.

CR Subject Classification (1991): Computer based methods in probability and statistics (G.3), Database applications (H.2.8), Information search and retrieval (H.3.3) *clustering, search process*, Learning (I.2.6) *concept learning, induction, knowledge acquisition*

Keywords & Phrases: Data Mining, Probabilistic Knowledge, Probabilistic Search, Probability Theory

1. Introduction

Many models of reality are probabilistic rather than deterministic, either by lack of current understanding or by nature. For example, it is unrealistic to expect a model that will predict accurately whether or not a customer will order crisps with his beer or not. It is, however, very well possible to have a model that yields the probability that a customer orders crisps with his beer. In fact, Agrawal et al. have recently shown that such a model can be inferred efficiently from a database [1].

In this paper we introduce a new class of such models, called *risk-profiles* and show how to infer them from a database. The proto-typical example of this class of problems is derived from the insurance business.

Report CS-R9430
ISSN 0169-118X
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Take, e.g., the insurance company *Save or Sorry* (SOS), that handles car insurances. To stay in business, SOS has to satisfy two almost contradictory requirements. First, the total of premiums received in a given year should be at least as high as the total of costs caused by claims, i.e., the higher the premiums the better. Secondly, the premiums it charges should not be higher than those of the competitors, i.e., the lower the premiums the better.

Hence, SOS should be able to predict the total claims of an insurant in a given year as accurately as possible. Clearly, predicting this amount with a 100% accuracy for each and every client is impossible. What insurance companies do, is identifying groups of clients with the same *expected* claim amount per year. Such groups are described by so-called risk-profiles. In other words, by what is called a (set-)description in data mining.

Clearly, the sharper these descriptions are, the more competitive SOS can be. Since insurance companies have large databases with insurance and client data, inferring risk-profiles offers a major and profitable challenge to data mining.

Another example of the use of risk-profiles is in the analysis of medical data. Both doctors and patients would like to know the chance that a patient dies given the symptoms she has. It is reassuring to know that say only 0.1% of the patients with flu will die, but perhaps less so if one knows that the mortality rate for those with both the flu and a heart-condition is much higher.

In this paper we study a simple risk inference problem, viz., we only try to infer the probability that a client will file a claim in a given year. More complicated cases are simple extensions of the solution proposed in this paper.

This solution is reached as follows. After recapitulating some basic probability theory, Section 2 presents the formal statement of this problem in terms of (set-)descriptions. This formal problem is analysed in the third section, were we indicate that good descriptions are *homogeneous*. Roughly, a description is called homogeneous if all extensions of this description yield nothing surprising. The notions of homogeneity and surprises are formalised in Section 4.

In Section 5, it is shown that certain maximal homogeneous descriptions are in a certain sense unique. This uniqueness result indicates that these maximal descriptions are plausible solutions for our problem. The next section shows that finding these maximal descriptions is an instantiation of the well-studied problem: maximise this function. It is then argued that due to the nature of our problem, probabilistic maximisation heuristics are the most adequate. In the final section, the conclusions are stated together with a brief description of current research. Moreover, a brief comparison with related research is given.

Note, this paper is only an extended abstract. All definitions, results and proof-sketches are given in the running text.

2. Preliminaries

In the first subsection we recall some basic facts from probability theory. For a more detailed treatment, the reader is referred to textbooks such as [3]. The second subsection presents the formal statement of the problem. In the third and final subsection some notational conventions are introduced.

2.1 Bernoulli experiments

To determine the risk-profiles for SOS, it is assumed that its clients undertake a *Bernoulli* or 0/1 experiment each year. Recall that a Bernoulli experiment is an experiment with two possible outcomes, denoted by 0 and 1. The outcome 1 is often called a success. In our example, a client succeeds in a trial of the experiment if he files a claim in that year.

For a Bernoulli experiment, it is assumed that there is a constant probability, say p, of success for each trial. The probability that we get exactly k success in n trials is then given by the *binomial distribution:*

$$P(\#S = k|\ n,\ p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{(n-k)}$$

If n trials result in k successes, the best estimation of p is given by k/n. However, it is conceivable that p differs from k/n. The $(100-\delta)\%$ confidence interval for p, given n and k is the largest interval $CI(n,k,\delta) \subseteq [0,1]$ such that $P(p \notin CI(n,k,\delta)) \leq \delta\%.$, i.e., the probability that $p \notin CI(n,k,\delta)$ is smaller then $\delta\%$.

$CI(n,k,\delta)$ can be computed using the binomial distribution. In fact, using the Chernoff bounds:

$$P(|k/n - p| > \epsilon) = \sum_{k:\ |k/n-p|>\epsilon} \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{(n-k)} < 2e^{-\epsilon^2 n/4p(1-p)}$$

one sees that with $c = \sqrt{\frac{1}{n} log \frac{2}{\delta}}$, $[k/n - c, k/n + c]$ is an easily computed conservative approximation of $CI(n,k,\delta)$.

2.2 Descriptions: the problem

If a client is insured for a long time with SOS, and if it is reasonable to assume that the client's probability of success has not changed over the years, the confidence intervals of the previous subsection can be used to determine the chance of success for each client individually. Alas, these assumptions are not very often met. If only because, e.g., changes in traffic density ensure that the probability of success for each client will vary over the years.

With only one or two trials per client, the method sketched above will give $[0,1]$ as the 95% confidence interval for our individual clients, which is pretty useless. The assumption we make is that there are a few groups of clients, such that clients in the same group have the same probability of success. That is, we assume that there is a cover $\{G_1, \ldots, G_l\}$ of disjoint subsets of SOS's clients, such that:

$$\forall cl_1, cl_2 : [p(cl_1) = p(cl_2)] \leftrightarrow [\exists! i \in \{1, \ldots, l\} : [cl_1 \in G_i \wedge cl_2 \in G_i]],$$

where $p(cl_i)$ denotes the probability of success of client cl_i. The assumption that l is small is made to ensure that each group has enough clients to allow for an accurate estimation of the associated probability of success.

The second assumption is that membership of one of the G_i depends on only a few properties of the client. That is, we assume:

1. a set of attributes $\mathcal{A} = \{A_1, \ldots, A_n\}$ with domains $D_1, \ldots, D_n$, where $dom(A_i) = D_i$;
2. a function $p : D_1 \times \cdots \times D_n \to [0, 1]$;

such that if $A_i(t, cl)$ denotes the A_i-value client cl has at time t, $p(A_1(t, cl), \ldots, A_n(t, cl))$ denotes cl's probability of success at time t.

For SOS, we can think of the attributes *age, gender,* and *miles per year* with their obvious domains.

Using the second assumption, we can rephrase the first as saying, we assume that there exists a cover, $\mathcal{C} = \{C_1, \ldots, C_l\}$, of disjoint subsets of $D_1 \times \cdots \times D_n$ such that:

$$\forall v_1, v_2 \in D_1 \times \cdots \times D_n : [p(v_1) = p(v_2)] \leftrightarrow [\forall i \in \{1, \ldots, l\} : [v_1 \in C_i \leftrightarrow v_2 \in C_i]]$$

The chance of success that all the $v \in C_i$ share is called the associated probability of C_i and is denoted by p_{C_i}.

The problem discussed in this paper can now be stated as: "find the C_i and their associated probability p_{C_i}".

In datamining terminology this means that we try to find a (set-)description for each of the C_i. That is, find logical formulae ϕ_i such that a value $v \in D_1 \times \cdots \times D_n$ satisfies ϕ_i, denoted by $\phi_i(v)$, iff $v \in C_i$. The formulae ϕ are expressions in some description language Φ over $\mathcal{A}$. Which description language Φ is chosen is unimportant for this paper as long as one can expect that the C_i can be described by Φ and the following four assumptions are met:

1. Φ is closed under negation, i.e., $\phi \in \Phi \to \neg\phi \in \Phi$
2. Φ is closed under conjunctions, i.e., $\phi_1, \phi_2 \in \Phi \to \phi_1 \wedge \phi_2 \in \Phi$;
3. Implication is decidable for Φ.
4. Φ is sparse with regard to $P(D_1 \times \cdots \times D_n)$. That is, the number of subsets described by Φ is small compared to the number of all possible subsets.

The necessity of these requirements will become clear in the next sections.

For SOS, an example of such a sparse description language is given by the disjunctive and conjunctive closure of the following set of elementary descriptions:

1. *gender = female, gender = male*;
2. *age = young, age = middle aged, age = old*;
3. *miles per year = low, miles per year = average, miles per year = high.*

Of course, the most important assumption on Φ is that the C_i can be described in Φ. Hence, choosing an appropriate Φ is an important task in deriving the risk-profiles.

To state our problem in terms of descriptions, define a set $\{\phi_1, \ldots, \phi_k\}$ of descriptions to be a *disjunctive cover*, abbreviated to *discovery*, if:

1. $\forall i, j \in \{1, \ldots, k\} : i \neq j \rightarrow [\phi_i \wedge \phi_k \rightarrow \perp]$
2. $\bigvee_{i=1}^{k} \phi_i \in \Phi$ and $[\bigvee_{i=1}^{k} \phi_i] \rightarrow \top$

Note that if Φ contains $\perp$, we assume that discoveries do not contain $\perp$. The set { *gender = female, gender = male* } is a discovery for SOS.

The problem can then be restated as: find a discovery $\{\phi_1, \ldots, \phi_k\}$ such that

$$\forall v_1, v_2 \in D_1 \times \cdots \times D_n : [p(v_1) = p(v_2)] \leftrightarrow [\forall i \in \{1, \ldots, k\} : [\phi_i(v_1) \leftrightarrow \phi_i(v_2)]].$$

2.3 Notational conventions

The database is seen as a recording of trials of our experiment. Hence, it is a table R, with schema $\mathcal{A} \cup \{S\}$, where S has domain $\{0, 1\}$. Each time t an object $o \in P$ takes experiment E, we insert, regardless of duplicates, the tuple $(A_1(t, o), \ldots, A_n(t, o), S(t, o))$ in the table where:

1. $A_i(t, o)$ denotes the value $v_i \in D_i$ that o has for A_i at time t.
2. $S(t, o) = 1$ if o succeeded in this experiment and $S(t, o) = 0$ otherwise.

Three notations used for the elements of Φ with respect to R are: $\langle\phi\rangle$ to denote the subtable of R of all tuples that satisfy ϕ, $\langle\phi\rangle_S$ denotes the elements of $\langle\phi\rangle$ that are successes, and $\|\phi\|$ to denote all $v \in D_1 \times \cdots \times D_n$ that satisfy ϕ.

3. Problem analysis

The problem as stated in the previous section consists of two parts. We have to find a discovery $\{\phi_1, \ldots, \phi_k\}$ and we have to prove that for this discovery $\forall v_1, v_2 \in D_1 \times \cdots \times D_n : [p(v_1) = p(v_2)] \leftrightarrow [\forall i \in \{1, \ldots, k\} : [\phi_i(v_1) \leftrightarrow \phi_i(v_2)]]$ holds.

Finding discoveries is easy, each sequence $\phi_1, \ldots, \phi_n \in \Phi$ of descriptions generates the potential discovery $\Psi = \{\phi_1, \neg\phi_1 \wedge \phi_2, \neg\phi_1 \wedge \neg\phi_2 \wedge \phi_3, \ldots, (\neg\phi_1 \wedge \cdots \wedge \neg\phi_n)\}$. After removal of duplicates (ϕ is a duplicate of ψ iff $\phi \leftrightarrow \psi$) and $\perp$, Ψ is a discovery.

So, we may expect that the discoveries are abundant. However, they are not all equally good in that some will not satisfy the second requirement.

To discuss this requirement, we should first determine what probability should be associated with a description ϕ. If we assume that the database is a random selection of the potential clients, this is simple, since for each $\phi \in \Phi$, $\langle\phi\rangle$ can be seen as the record of some Bernoulli experiment E_ϕ. Hence, the probability p_ϕ of success associated with ϕ is simply the fraction of successes in $\langle\phi\rangle$ and ϕ has the $100 - \delta\%$ confidence interval $CI_\phi^\delta = CI(|\langle\phi\rangle|, |\langle\phi\rangle_S|, \delta)$[1].

If the database is not a random selection of the set of all potential clients, say the number of young clients is far less than could be expected, the above observation is only true for the "real" ϕ_i and the logical combinations thereof. This is however unimportant for the purposes of this paper.

[1] In this paper we assume that all descriptions we consider are sufficiently large, such that reasonably accurate estimates of the p_ϕ can be found. See also footnotes and remarks later in this paper.

The second requirement in turn consists of two sub-requirements. The first is that two values that satisfy different elements of the discovery have different chances of success. That is, the different elements are *distinct.* The second states that two values that satisfy the same element of the discovery have the same chance of success. That is the elements are *homogeneous.*

Using the $100 - \delta\%$ confidence intervals, we can test distinction. For if $CI^{\delta}_{\phi_1} \cap CI^{\delta}_{\phi_2} = \emptyset$ we know that $\forall v, w \in D_1 \times \cdots \times D_n : \phi_1(v) \wedge \phi_2(w) \rightarrow P(p(v) = p(w)) \leq \delta$, *if both ϕ_1 and ϕ_2 are homogeneous.*

So, provided we can test homogeneity of descriptions, we can for some fixed δ discard all those discoveries that contain at least two elements that cannot be distinguished with a confidence of at least $100 - \delta\%$.

Can we test homogeneity of descriptions? If a description ϕ is not homogeneous, $\langle\phi\rangle$ contains elements v and w such that $p(v) \neq p(w)$. More precisely, if ϕ is not homogeneous, there exists a cover $\mathcal{C} = \{C_1, \ldots, C_l\}$, of disjoint subsets of $|\phi|$ such that:

$$\forall v_1, v_2 \in |\phi| : [p(v_1) = p(v_2)] \leftrightarrow [\forall i \in \{1, \ldots, l\} : [v_1 \in C_i \leftrightarrow v_2 \in C_i]]$$

But this is exactly the question we are trying to answer!

So, let us first try to answer the question when the empty description, i.e., $\top$ is homogeneous. Clearly, if every subset of the database has the same associated probability on success, then the database is homogeneous. However, this is obviously impossible. In fact, the database is the record of a *weighted Bernoulli experiment.*

Let g_i denote the relative size of group G_i in the universe of potential clients. Under the assumption that the actual clients are a random subset of the potential clients the database is the recording of a Bernoulli experiment E with its probability on success p_E given by:

$$p_E = \sum_{i=1}^{k} g_i p_{G_i}$$

Note that if $|R|$ is large, then, as one would expect:
$p_E = \sum_{i=1}^{k} g_i p_{G_i} \approx \sum_{i=1}^{k} \frac{|\langle\phi_i\rangle|}{|R|} p_{\phi_i} \approx \sum_{i=1}^{k} \frac{|\langle\phi_i\rangle|}{|R|} \times \frac{|\langle\phi_i\rangle_S|}{|\langle\phi_i\rangle|} = \frac{\textit{the number of successes in } R}{|R|}$.

Hence, if we consider all possible subsets of the database, the number of successes in the subsets of a given size follow a binomial distribution. In other words, if we inspect all possible subsets we can conclude nothing.

However, we are not interested in all possible subsets, but only in those subsets we can describe. Since Φ is sparse, one would expect that all descriptions have more or less the same associated probability. If we think that $\top$ is homogeneous and it turns out that two descriptions have distinct associated probabilities then we are surprised. We would not expect this to happen. If female clients appear to drive much safer than male clients, we are no longer sure that all clients have the same associated probability. That is, we are no longer convinced that $\top$ is homogeneous.

In other words, it is plausible that a description is homogeneous if all its covers have the same associated probability. This can be stated more succinctly in the slogan: *Homogeneous descriptions contain no surprises.*

4. Homogeneous discoveries

To simplify the discussion, define a set $\{\phi_1, \ldots, \phi_k\}$ of descriptions to be a *discovery for description* ψ, if:

1. $\forall i, j \in \{1, \ldots, k\} : i \neq j \rightarrow [\phi_i \wedge \phi_k \rightarrow \bot]$
2. $\bigvee_{i=1}^{k} \phi_i \in \Phi$ and $[\bigvee_{i=1}^{k} \phi_i] \rightarrow \psi$

If we assume that $\top \in \Phi$, the discoveries of the previous sections are discoveries for $\top$; we will continue to use this abbreviation.

If ϕ is an homogeneous description, so is $\phi \wedge \psi$. In fact, then p_ϕ and $p_{\phi\wedge\psi}$ should be, almost, identical. In the previous section, we have seen that p_ϕ and $p_{\phi\wedge\psi}$ are identical with a probability δ if CI_ϕ^δ and $CI_{\phi\wedge\psi}^\delta$ are disjoint. Therefore, we define a description ψ to be a $(100-\delta)\%$ *surprising description* for a description ϕ, if CI_ϕ^δ and $CI_{\phi\wedge\psi}^\delta$ are disjoint.

For example, if $CI_\top^{5\%} = [0.19, 0.21]$ and $CI_{gender=male}^{5\%} = [0.25, 0.27]$, then $gender = male$ is a 95% surprise for $\top$.

Since discoveries can contain many elements, it is not impossible that a discovery for a homogeneous description ϕ contains some surprising descriptions for ϕ. To quantify the number of surprising descriptions in a discovery that is surprising, note that a description ψ has a chance δ to be surprising for ϕ. That is, being a surprise is a Bernoulli experiment, with probability δ. We would be surprised if the number of successes in this experiment is high.

For a Bernoulli experiment with probability p of success, define $Lwb(n, p, \delta)$ to be the lowest integer k such that $P(\#S \geq k | n, p) \leq \delta$. For a description ϕ, define $Lwb(\phi, \delta) = Lwb(|\langle\phi\rangle|, p_\phi, \delta)$.

Hence, we define a discovery $\{\phi_1, \ldots, \phi_l\}$ for a rule ψ to be a $(100-\delta)\%$ *surprising discovery for* ψ, if:

$$|\{i \in \{1, \ldots, l\} | \phi_i \textit{ is a } (100-\delta)\% \textit{ surprising description for } \psi\ \}| \geq Lwb(\psi, \delta)$$

A description for which there are *no* $(100-\delta)\%$ surprising discoveries, is called a $(100-\delta)\%$ *homogeneous description* [2] .

One might wonder why we bother Bernoulli experiments for covers since Φ is assumed to be sparse. However, even if Φ is sparse, it can still happen that Φ distinguishes all values of an attribute with a large domain, say *age*. For such a dense cover, one can expect some surprises. However, only if such surprises are considerable, say all clients younger than 25, it counts as a real surprise.

Note that if Φ is never dense, such as in our example, one surprise in a cover is enough to make the cover surprising.

Let ψ be a description, a discovery $\{\phi_1, \ldots, \phi_l\}$ for ψ is $(100-\delta)\%$ *homogeneous discovery for* ψ, iff all of the ϕ_i are $(100-\delta)\%$ homogeneous descriptions.

[2] If ψ is a description for which $|\langle\psi\rangle|$ is already so small that for almost all non contradictory ϕ, $|\langle\phi \wedge \psi\rangle|$ is to small for an accurate estimation of $p_{\phi\wedge\psi}$ are not considered to be homogeneous.

5. Split $(100-\delta)\%$ homogeneous discoveries are unique

Clearly, the real discovery we are searching for will be $(100-\delta)\%$ homogeneous. But are all $(100-\delta)\%$ homogeneous discoveries plausible candidate solutions? If there are many $(100-\delta)\%$ homogeneous discoveries, one might wonder which of these discoveries is the most plausible answer. In this section, we show that all *split* discoveries are more or less the same.

A $(100-\delta)\%$ homogeneous discovery $\{\phi_1,\ldots,\phi_l\}$ is *split* iff

$$\forall\psi\in\Phi\;\forall i,j\in\{1,\ldots,l\}:\left[\begin{array}{ll} i\neq j & \wedge \\ \langle\phi_i\wedge\psi\rangle >> 0 & \wedge \\ \langle\phi_j\wedge\psi\rangle >> 0 & \end{array}\right]\rightarrow CI^{\delta}_{\phi_i\wedge\psi}\cap CI^{\delta}_{\phi_j\wedge\psi}=\emptyset$$

In other words, a discovery is split if its descriptions are distinct on all but trivial subsets.

Let $\{\phi_1,\ldots,\phi_k\}$ and $\{\psi_1,\ldots,\psi_l\}$ be two split $(100-\delta)\%$ homogeneous discoveries. Then since both k and l are assumed to be small (since the number of groups is assumed to be small) there is for each ϕ_i, at least one ψ_j such that $\langle\phi_i\wedge\psi_j\rangle >> 0$. But since both discoveries are homogeneous and split, there can be at most one. So, $\langle\phi_i\rangle\approx\langle\psi_j\rangle$ and $CI^{\delta}_{\phi_i}\approx CI^{\delta}_{\psi_j}$.

Define a pre-order on the $(100-\delta)\%$ homogeneous discoveries by $\{\phi_1,\ldots,\phi_k\}\leq\{\psi_1,\ldots,\psi_l\}$ iff there exists a injective mapping $f:\{1,\ldots,k\}\rightarrow j\in\{1,\ldots,l\}$ such that $CI^{\delta}_{\phi_i}\overset{\subset}{\approx}CI^{\delta}_{\psi_{f(i)}}$. Then, by the observation above, we can identify all split $(100-\delta)\%$ homogeneous discoveries and turn this relation into a partial order.

Clearly, the split $(100-\delta)\%$ homogeneous discoveries are minimal with regard to this order. Note, however, they need not exist. Hence, we cannot identify the split $(100-\delta)\%$ homogeneous discoveries with the minimal discoveries. Of all the $(100-\delta)\%$ homogeneous discoveries, the minimal discoveries are preferable, e.g., by the minimal description length principle.

If a split discovery exists, this one is the most plausible of all minimal discoveries, since it makes the sharpest distinction between its descriptions. In other words, it is the description with the highest information-gain.

The fact that the split discoveries are not unique is simply caused by the fact that a set of tuples can have more than one description. For example, it could happen that almost all young clients are male and vice versa. In that case the descriptions *age* = *young* and *gender* = *male* are equally good from a theoretical point of view. Not necessarily from a practical point of view. For, it is very well possible that the description *age* = *young* makes sense to a domain expert while *gender* = *male* does not. Hence, both options should be presented to the domain expert.

6. Finding minimal homogeneous discoveries

In the first subsection we show that finding minimal homogeneous discoveries reduces to maximisation of a function. In the second we briefly discuss various maximisation algorithms.

6.1 Maximisation

Define the partial order $\prec_\delta$ on the set of $(100-\delta)\%$ homogeneous descriptions as follows:

$$[\phi\prec_\delta\psi]\leftrightarrow[p_\phi\leq p_\psi\wedge CI^{\delta}_{\phi}\cap CI^{\delta}_{\psi}=\emptyset]$$

. If for neither $\phi \prec_\delta \psi$ nor $\psi \prec_\delta \phi$ holds, then we say that $\phi \approx_\delta \psi$.

From the definition of $(100-\delta)\%$ homogeneous discoveries it is immediately clear that such a discovery neccesarily contains a description ϕ such that for all $(100-\delta)\%$ homogeneous descriptions ψ either $\psi \prec_\delta \phi$ or $\phi \approx_\delta \psi$.

Minimal $(100-\delta)\%$ homogeneous discoveries differ from the ordinary ones in that their "maximal" element ϕ also has a maximal $|\langle\phi\rangle|$ Therefore, define the following pre-order on the $(100-\delta)\%$ homogeneous descriptions: $\phi \sqsubseteq_\delta \psi$ if:

1. either if $\phi \prec_\delta \psi$;
2. or, if $\phi \approx_\delta \psi$ and $|\langle\phi\rangle| \leq |\langle\psi\rangle|$.

Then minimal $(100-\delta)\%$ homogeneous discoveries contain an element ϕ such that for all $(100-\delta)\%$ homogeneous descriptions ψ, $\psi \sqsubseteq_\delta \phi$.

From this observation, we see that the following pseudo algorithm yields a minimal $(100-\delta)\%$ homogeneous cover:

1. Make a list of $(100-\delta)\%$ homogeneous descriptions as follows:
   (a) find a ϕ that is maximal with regard to $\sqsubseteq_\delta$ for R
   (b) remove $\langle\phi\rangle$ from R and add ϕ to the list.
   (c) continue with this process until:
   either $\top$ is homogeneous on the remainder of R;
   or R is to small for further accuracy estimations of probabilities. In this case make the list empty.
2. If the list is non-empty, turn it into a potential discovery as indicated at the beginning of section 3. If the list is empty return nothing

If we are returned a potential discovery, it is a minimal $(100-\delta)\%$ homogeneous discovery by virtue of our remarks above. If the algorithm fails, no solution exists in Φ. Given such a minimal description, it is straightforward to check whether it is split.

As an aside, note that the list returned by the pseudo-algorithm is similar to a decision-list [10]

6.2 Finding maximal descriptions

To turn the pseudo-algorithm of the previous subsection into an algorithm, it is sufficient to give a procedure that returns a maximal $(100-\delta)\%$ homogeneous description ϕ. For, we can check whether $\top$ is homogeneous by checking whether ϕ and $\top$ have the same associated chance.

But finding such a maximal $(100-\delta)\%$ homogeneous description ϕ is a simple application of maximising a function. In the first phase of the search this function is the associated probability of a rule, and in the second phase it is the cover of the rule (while retaining the maximal probability found).

This problem is well-documented in the machine learning literature. Older solutions are ID3, AQ15 and CN2 [9, 6, 2]. More modern, non-deterministic approaches use genetic algorithms or simulated annealing. See [4] for an overview of a some of these systems.

For our problem, a non-deterministic approach is best suited. For, there can be many different $(100-\delta)\%$ homogeneous discoveries, that describe the same sets but with different descriptions. Some of these descriptions will be pure coincidence, while others have a sensible interpretation.

If we use a deterministic algorithm, the chance exists that we miss out the sensible descriptions completely; this would render the result virtually useless.

Of the two pre-dominant non-deterministic search methods, the genetic approach seems to offer the best possibilities for our problem. Since, as described in [5], its genetic operators can adapt the proven heuristics of the older deterministic algorithms.

In this particular case, we think in fact of two sets of operators. One for the first phase and another for the second. In the first phase we search for a maximal probability so, one can think of a hill-climber type of approach. In the second phase generalising a description is far more important. Hence, we need a different set of operators.

Note that it is not our intention to turn genetic programming into something deterministic, rather, it is an attempt to lift genetic programming to the level of symbolic inference.

7. Conclusions and future research

The main result of this paper is that we have shown that a new class of problems, viz., find risk profiles, can be reduced to a well-known class of problems, viz., maximisation problems. These results have been achieved by introducing the notions of a surprise and homogeneity. In particular the slogan: “homogeneous discoveries contain no surprises”, has proven to be fruitful.

The important implication of these results is that we can built efficient algorithms to solve this new class of real world problems. In fact, at the moment we are building and testing a system based on the ideas presented in this paper for an insurance company that wants to find such risk profiles, for obvious commercial reasons.

Next to the choice for an adequate maximisation algorithm as discussed in Section 6, the other main problem is the choice of a good description language. As we have seen in this paper, this language should not be to rich. Obviously, it should neither be to poor if the profiles are to be expressible in this language. The design of a language that meets these two conflicting requirements is currently one of our main topics.

7.1 Related work

As already indicated in the introduction, as far as the author is aware, inferring risk-profiles is a new problem area in data mining research. It is probably most connected to diagnostic problem solving as reported in [7, 8]. A diagnostic problem is a problem in which one is given a set of symptoms and must explain why they are present.

The authors of [7, 8] introduce a solution that integrates symbolic causal inference with numeric probabilistic inference. A crucial point in this integration is the use of a-priori probabilities. The authors imply that these probabilities should be supplied by an expert.

However, using the technique presented in this paper, these a-priori probabilities can be derived from a database.

Acknowledgements

Marcel Holsheimer has been the sparring partner in many stimulating discussions. Moreover, he pointed out an error in a previous version of this paper. My heartfelt thanks.

Also thanks to the anonymous referee who pointed out many improvements, and brought references [7] and [8] to my attention.

References

1. Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Database mining: a performance perspective. *IEEE transactions on knowledge and data engineering*, Vol. 5, No. 6, December:914 – 925, 1993.
2. Peter Clark and Tim Niblett. The CN2 induction algorithm. *Machine Learning*, 3:261 – 283, 1989.
3. William Feller. *An introduction to probability theory and its applications, Vol 1.* Wiley, 1950.
4. Marcel Holsheimer and Arno P.J.M. Siebes. Data mining: the search for knowledge in databases. Technical Report CS-R9406, CWI, January 1994.
5. Zbigniew Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Artificial Intelligence. Spinger-Verlag, 1993.
6. Ryszard S. Michalski, Igor Mozetic, Jiarong Hong, and Nada Lavrac. The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In *Proceedings of the 5th national conference on Artificial Intelligence*, pages 1041 – 1045, Philadelphia, 1986.
7. Yun Peng and James A. Reggia. A probabilistic causal model for diagnostic problem solving – part I: Integrating symbolic causal inference with numeric probabilistic inference. *IEEE transactions on Systems, Man, and Cybernatics*, Vol. 17, No. 2, March/April:146 – 162, 1987.
8. Yun Peng and James A. Reggia. A probabilistic causal model for diagnostic problem solving – part II: Diagnostic strategy. *IEEE transactions on Systems, Man, and Cybernatics*, Vol. 17, No. 3, May/June:395 – 406, 1987.
9. J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1:81 – 106, 1986.
10. Ronald L. Rivest. Learning decision lists. *Machine Learning*, 2:229 – 246, 1987.