# Pruning Redundant Association Rules Using Maximum Entropy Principle

Szymon Jaroszewicz and Dan A. Simovici

University of Massachusetts at Boston,
Department of Mathematics and Computer Science,
Boston, Massachusetts 02125, USA
{sj,dsim}@cs.umb.edu

**Abstract.** Data mining algorithms produce huge sets of rules, practically impossible to analyze manually. It is thus important to develop methods for removing redundant rules from those sets. We present a solution to the problem using the Maximum Entropy approach. The problem of efficiency of Maximum Entropy computations is addressed by using closed form solutions for the most frequent cases. Analytical and experimental evaluation of the proposed technique indicates that it efficiently produces small sets of interesting association rules.
**Keywords:** association rule, rule interestingness, rule pruning, maximum entropy

## 1 Introduction

Many data mining algorithms produce huge sets of rules, practically impossible to analyze manually. Typically, those sets are highly redundant and, so, it is important to develop methods for removing redundant rules and for helping the user select from thousands of discovered rules those which are the most interesting from his point of view.
Our goal is to identify a reasonably small, nonredundant set of interesting association rules describing well (and as completely as possible) the relationships within the data. The paper presents a solution to this problem using the maximum entropy approach. A subrule of an association rule $I \to J$ is a rule $K \to J$, such that $K \subset I$ (see [AIS93] or further sections for a detailed discussion of association rules). In [LHM99,AL99] a rule is considered not interesting if its confidence is close to that of one of its subrules. A similar approach (although in a slightly generalized setting) is used in [PT00] to prune the discovered rules. Also, in [PT00] a rule is considered interesting with respect to some set of beliefs if it contradicts at least one of the rules in the beliefs under the so called monotonicity assumption. A detailed statistical analysis of interestingness of a rule with respect to a single subrule, and algorithms for finding rules interesting in this setting can be found in [Suz97,SK98].
The current work on evaluation of interestingness considers the influence of each subrule separately, while in our approach we take into account the combined influence of all the subrules of a rule. Examples illustrating the advantages (in our opinion) of our approach are given in Section 3.

In [LHM99], apart from pruning, the authors also find so called *direction setting* rules which summarize the dataset. This procedure takes into account many subrules of a rule and is thus similar to our approach. However, our approach has the advantage of giving a more precise, probabilistic quantification of the influence of subrules on the interestingness of a rule.

Another approach to pruning discovered rules is based on selecting a minimal set of rules covering the dataset [TKR$^+$95,BVW00]. Again, those methods do not take into consideration probabilistic interactions between rules in the cover. Also, they may prune many interesting rules if they cover instances already covered by other rules.

A general study of measures of rule interestingness can be found in [BA99,JS01,HH99].

An overview of the interestingness of a rule with respect to a set of constraints can be found in [GHK94]. In [GHK94] the authors propose the method of *random worlds* and prove that in many important cases it is equivalent to the principle of maximum entropy.

Maximum entropy principle and other probability models have been also used in datamining in query selectivity estimation [PMS01]. There has also been work in applying MaxENT in speech processing [Rat96]

Let us now introduce notation used throughout the paper. If $A$ is an attribute of a table we denote its domain by $\mathrm{Dom}(A)$. When $\mathrm{Dom}(A) = \{0,1\}$ we say that $A$ is a binary attribute. In this note we use tables whose headings have the form $H = \{A_1, A_2, \ldots, A_m\}$ and consist of binary attributes. The heading $H$ will be written, as usual as $A_1 \cdots A_m$. Subsets of $H$, referred to as *itemsets*, will be denoted using uppercase Roman letters $I, J, K, L, \ldots$. Single attributes will be denoted by uppercase letters $A, B, C, \ldots$.

The domain of a set of attributes $I \subseteq H$, where $I = A_{i_1} A_{i_2} \ldots A_{i_r}$ is defined as

$$\mathrm{Dom}(I) = \mathrm{Dom}(A_{i_1}) \times \mathrm{Dom}(A_{i_2}) \times \cdots \times \mathrm{Dom}(A_{i_r}) = \{0,1\}^r.$$

Values from domains of attributes will be denoted by corresponding boldface lowercase letters, e.g. $\mathbf{i} \in \mathrm{Dom}(I)$.

For $\mathbf{h} \in \mathrm{Dom}(H)$ and $I \subseteq H$, we denote the projection of $\mathbf{h}$ on $I$ by $\mathbf{h}_I$. For a probability distribution $P$ on $\mathrm{Dom}(H)$ let $P_I$ be the marginal probability distribution on $\mathrm{Dom}(I)$, where $I \subseteq H$, obtained by marginalizing the distribution $P$. In other words, we have

$$P_I(\mathbf{i}) = \sum \{P(\mathbf{h}) : \mathbf{h}_I = \mathbf{i}\}$$

for $\mathbf{i} \in \mathrm{Dom}(I)$. The joint distribution of $H$ estimated from the data will be denoted by $\hat{P}$.

Let $P_I$ and $P'_I$ be two probability distributions over an itemset $I$. The *Kullback-Leibler divergence* and the *chi-squared divergence* [KK92] between $P_I$ and $P'_I$ are defined respectively as

$$D_{KL}(P_I : P'_I) = \sum_{\mathbf{i} \in \mathrm{Dom}(I)} P_I(\mathbf{i}) \log \frac{P_I(\mathbf{i})}{P'_I(\mathbf{i})},$$

$$D_{\chi^2}(P_I : P'_I) = \sum_{\mathbf{i} \in \mathrm{Dom}(I)} \frac{(P_I(\mathbf{i}) - P'_I(\mathbf{i}))^2}{P'_I(\mathbf{i})}.$$

Intuitively, the divergence represents how much distribution $P_I$ differs from $P'_I$. Since the choice of divergence is immaterial for the rest of the paper we will simply denote the divergence by $D$ meaning that either Kullback-Leibler or chi-squared divergence can be used.
A *constraint* $\mathsf{C}$ on the set of attributes $H$ is a pair $\mathsf{C} = (I, p)$ where $I \subseteq H$, $p \in [0, 1]$. A probability distribution $P$ *satisfies* a constraint $\mathsf{C} = (I, p)$ if $P_I(\mathbf{1}_I) = p$, where $\mathbf{1}_I = (1, 1, \ldots, 1) \in \mathrm{Dom}(I)$. Usually the attribute set will be clear from context, so we will just write $\mathbf{1}$.
To remove redundancies in the rule set we need to define how interesting a rule is with respect to a set of constraints introduced by other rules.

**Definition 1.** *A set of constraints $\mathcal{C}$ is* consistent *if there exists a joint probability distribution over $H$ which satisfies all the constraints in $\mathcal{C}$. Otherwise, $\mathcal{C}$ is* inconsistent.

In this paper we will only be concerned with consistent sets of constraints. Dealing with inconsistent sets of constraints is an interesting topic of future research.
While determining interestingness of rules with respect to a consistent set of constraints $\mathcal{C}$ we will associate with $\mathcal{C}$ some joint probability distribution $P^{\mathcal{C}}$ over $H$.
Note that a set of constraints does not have to determine the joint probability distribution uniquely, and we have to choose one of the conforming distributions. The three main approaches to this problem are the maximum entropy principle (MaxENT), the minimum interdependence principle, and the maximum likelihood (see [KK92,Adw97]). We use MaxENT, but it can be shown [KK92,Adw97], that in most cases all three approaches are equivalent. Philosophical justifications of the principles can be found in [KK92,GHK94].

**Definition 2.** *Let $\mathcal{C}$ be a consistent set of constraints. A* probability distribution *$P^{\mathcal{C}}$ over $H$ is* induced by *$\mathcal{C}$ if it satisfies the following conditions:*
1. *$P^{\mathcal{C}}$ satisfies all the constraints in $\mathcal{C}$.*
2. *Of all probability distributions over $H$ satisfying $\mathcal{C}$, $P^{\mathcal{C}}$ has the largest entropy.*

It can be shown [Adw97] that $P^{\mathcal{C}}$ is unique.

## 2 Interestingness of A Rule with Respect to A Set of Constraints

We are now ready to define the interestingness of an association rule with respect to some set of constraints $\mathcal{C}$. For the definition of association rules see [AIS93].
The support of an itemset $I$ is $\mathrm{supp}(I) = \hat{P}_I(\mathbf{1})$. Rules with empty antecedents are allowed and the support and confidence of such rules are defined to be equal to the support of their consequents.

The set of constraints generated by an association rule $I \to J$ is defined as

$$\mathsf{C}(I \to J) = \{(I, \mathrm{supp}(I)), (I \cup J, \mathrm{supp}(I \cup J))\}.$$

We introduce two interestingness measures for association rules: the active and passive interestingness. The active interestingness reflects the impact of adding to the current set of constraints the set of constraints generated by the rule itself. The passive interestingness is determined by the difference between the confidence estimated from the data and the confidence estimated starting from the probability distribution induced by the constraints.

**Definition 3.** *Let $\mathcal{C}$ be a consistent set of constraints, $I \to J$ be a rule and $D$ some measure of distribution divergence. Denote by $Q^{\mathcal{K}}$ the probability distribution over $I \cup J$ induced by the set of constraints $\mathcal{K}$.*
*The* active interestingness *of $I \to J$ with respect to $\mathcal{C}$ is defined as:*

$$\mathsf{I}^{\mathrm{act}}(\mathcal{C}, I \to J) = D(Q^{\mathcal{C} \cup \mathsf{C}(I \to J)}, Q^{\mathcal{C}}).$$

*The* passive interestingness *of $I \to J$ with respect to $\mathcal{C}$ is defined as:*

$$\mathsf{I}^{\mathrm{pass}}(\mathcal{C}, I \to J) = \left| \mathrm{conf}(I \to J) - \frac{Q^{\mathcal{C}}(\mathbf{1})}{Q^{\mathcal{C}}_I(\mathbf{1})} \right|,$$

*where* $\mathrm{conf}(I \to J)$ *denotes the confidence of rule $I \to J$.*

Whenever we state facts that hold for either of these measures we simply talk about rule interestingness $\mathsf{I}$.

# 3 Pruning redundant association rules

**Definition 4.** *Let $\mathcal{R}$ be a set of association rules. Consider an association rule $I \to J$, where $I, J \subseteq H$. The rule $I \to J$ is* $\mathsf{I}$-nonredundant with respect to $\mathcal{R}$, *if $I = \emptyset$ or $\mathsf{I}(\mathcal{C}^{I,J}(\mathcal{R}), I \to J)$ is significantly greater than 0, where $\mathcal{C}^{I,J}(\mathcal{R}) = \{\mathsf{C}(K \to J) : K \to J \in \mathcal{R}, K \subset I\}$.*

Note that we do not specify precisely what 'significantly greater' means. This may mean 'greater than some threshold' or 'statistically significant at some confidence level' or some combination of both.
A feature of our definition of redundancy is that it is not influenced by rules involving attributes not in $I \cup J$. For example, suppose that the joint distribution of attributes $ABC$ is fully explained by rules $A \to B$ and $B \to C$. The rule $A \to C$ may still be considered $\mathsf{I}$-nonredundant, even though it does not introduce any new information.
We believe this is the correct behavior. In general, if we have a long chain of rules $A \to B \to C \to \ldots \to Y \to Z$, the rule $A \to Z$ might not be easy to see and thus be interesting. Furthermore, the discovered rules do not necessarily correspond to true causality relations, and it might be better, at least until the user develops a better understanding of the dataset, to present him/her also rules indirectly implied by some other rules.

Another important advantage of our method is that single rules usually involve very few attributes, and thus local interestingness can be efficiently determined, even by direct application of the Generalized Iterative Scaling algorithm, see later in this section.
An algorithm for producing a set of I-nonredundant rules with a single attribute in the consequent is given below:

**Input:** A set $\mathcal{S}$ of association rules.
**Output:** Set $\mathcal{R}$ of I-nonredundant association rules of $\mathcal{S}$.

1. For each $A_i \in H$
2. $\quad\mathcal{R}_i = \{\emptyset \to A_i\}$
3. $\quad k = 1$
4. $\quad$For each rule $I \to A_i \in \mathcal{S}$, $|I| = k$ do
5. $\quad\quad$If $I \to A_i$ is I-nonredundant with respect to $\mathcal{R}_i$ then
6. $\quad\quad\quad$Let $\mathcal{R}_i = \mathcal{R}_i \cup \{I \to A_i\}$
7. $\quad k = k + 1$
8. $\quad$Goto 4
9. $\mathcal{R} = \bigcup_{A_i \in H} \mathcal{R}_i$

Examples below show how our method compares with other work in certain situations. Passive interestingness measure $\mathsf{I}^{\text{pass}}$ is used, but it is easy to see that the statements remain valid also for the active interestingness measure $\mathsf{I}^{\text{act}}$. See discussion later in this section for details on how the maximum entropy distributions can be computed.

*Example 1.* Let $A, B, C$ be binary attributes, $P_A(1) = P_B(1) = 0.5$. The attribute $C$ depends on $A, B$ according to the following association rules:

| assoc. rule | confidence |
|---|---|
| $\emptyset \to C$ | 0.5 |
| $A \to C$ | 0.3 |
| $B \to C$ | 0.7 |
| $AB \to C$ | 0.3 |

Using the approach from [PT00,LHM99,AL99,SLRS99] rules $\emptyset \to C$, $A \to C$ and $B \to C$ are interesting but $AB \to C$ is not, since it is explained by the rule $A \to C$. We claim however that the rule $AB \to C$ is interesting, since it tells us that when both $A$ and $B$ are 'present' it is $A$ that influences $C$ stronger.
Consider rules $\emptyset \to C$, $A \to C$, and $B \to C$. The set of constraints corresponding to them is $\mathcal{C} = \{(A, 0.5), (B, 0.5), (C, 0.5), (AC, 0.15), (BC, 0.35)\}$. The MaxENT distribution in this case is

$$P^{\mathcal{C}} = \begin{pmatrix} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ 0.105 & 0.105 & 0.045 & 0.245 & 0.245 & 0.045 & 0.105 & 0.105 \end{pmatrix},$$

and $P^{\mathcal{C}}_{ABC}(\mathbf{1})/P^{\mathcal{C}}_{AB}(\mathbf{1}) = 0.5$, different from $\mathrm{conf}(AB \to C) = 0.3$, making the rule $AB \to C$ interesting.

*Example 2.* Assume now that the confidences of the rules in the example above are

| assoc. rule | confidence |
|---|---|
| $\emptyset \to C$ | 0.5 |
| $A \to C$ | 0.3 |
| $B \to C$ | 0.7 |
| $AB \to C$ | 0.5 |

Using methods given in [LHM99,AL99] the rule $AB \to C$ is interesting, since its confidence differs from $\mathrm{conf}(A \to C)$ and $\mathrm{conf}(B \to C)$.
However, as seen above the maximum entropy distribution induced by rules $\emptyset \to C$, $A \to C$ and $B \to C$ gives $P^{\mathcal{C}}_{ABC}(\mathbf{1})/P^{\mathcal{C}}_{AB}(\mathbf{1}) = 0.5$, and the rule $AB \to C$ is considered uninteresting. In other words, knowing the joint influence of $AB$ on $C$ does not give us any more information over what we have already know from other rules, since $A$ and $B$ are conditionally independent given $C$. The above result is intuitive since when both $A$ and $B$ influence $C$ we would expect their joint influence to be an 'average' between the influences of $A$ and $B$ alone.

*Example 3.* Suppose that attribute $A$ is independent of $B$, $C$, and jointly of $BC$. Then, $P^{\mathcal{C}}_{ABC}(\mathbf{1})/P^{\mathcal{C}}_{AB}(\mathbf{1}) = P^{\mathcal{C}}_{BC}(\mathbf{1})/P^{\mathcal{C}}_{B}(\mathbf{1}) = \mathrm{conf}(B \to C)$, and the rule $AB \to C$ is considered not interesting using our approach, but also using methods from [PT00,LHM99,SLRS99,AL99] which explains their good behavior in practice. However as the examples above show, those methods can filter out certain interesting rules, and include some uninteresting ones.

To compute the maximum entropy distribution we can use the Generalized Iterative Scaling (GIS) Algorithm [Adw97,Bad95,DR72,Csi89].
Let $\mathcal{C} = \{\mathsf{C}_1, \mathsf{C}_2, \dots, \mathsf{C}_n\}$ be a set of constraints, where $\mathsf{C}_k = (I_k, p_k)$. GIS proceeds by assigning some initial values to each probability in $P^{\mathcal{C}}$, and iteratively updating them until all the constraints are satisfied. Let $P^{\mathcal{C}(i)}$ denote the distribution after $i$ iterations. Updating in each iteration is performed according to the formula

$$P^{\mathcal{C}(i+1)}(\mathbf{h}) = P^{\mathcal{C}(i)}(\mathbf{h}) \prod_{\mathbf{h}_{I_k}=\mathbf{1}} \left[ \frac{p_k}{P^{\mathcal{C}(i)}_{I_k}(\mathbf{h}_{I_k})} \right]^{\frac{1}{c}},$$

for every $\mathbf{h} \in \mathrm{Dom}(H)$, assuming that $\frac{0}{0} = 0$. The algorithm is guaranteed to converge if $\sum_{k=1}^{n} f_k(\mathbf{h}) = c$ is a constant independent of $\mathbf{h}$. In practice, this condition can always be satisfied by adding an additional constraint. See [Adw97,DR72,Csi89] for details and proof of convergence. The version of the algorithm presented in [Csi89] has the advantage of being able to cope with distributions with zero probabilities, and this is the one we use in our implementation.
The disadvantage of the GIS algorithm is its high computational cost caused by the necessity of computing the marginal probabilities, and in some cases by the large number of iterations required.
One of the main techniques for speeding up MaxENT computations is *decomposition* [Bad95,And74,DLS80]. However, in our case, we will only use maximum entropy distributions in few variables, and our experiments showed that decomposition does not give real improvement in efficiency.

We noticed however that the number of rules considered interesting is small and thus constraints are usually simple. Closed form solutions are used for a few common cases; in every other situation we use the GIS algorithm.
Below we describe closed form solutions used in this paper. For attribute set $I$ denote $N_I = |\{\mathbf{x} \in \mathrm{Dom}(H) : \mathbf{x}_I = \mathbf{1}\}|$.

**Theorem 1.** *Let* $\mathcal{C} = \{(J, \hat{P}_J(\mathbf{1})), (K, \hat{P}_K(\mathbf{1})), (K \cup J, \hat{P}_{K\cup J}(\mathbf{1}))\}$, $J, K \subset H, K \cap J = \emptyset$ *be a set of constraints. The MaxENT distribution induced by* $\mathcal{C}$ *is*

$$P^{\mathcal{C}}(\mathbf{x}) = \begin{cases} \frac{\hat{P}_{K\cup J}(\mathbf{1})}{N_{K\cup J}} & , \textit{if } \mathbf{x}_J = \mathbf{1} \wedge \mathbf{x}_K = \mathbf{1} \\ \frac{\hat{P}_J(\mathbf{1}) - \hat{P}_{K\cup J}(\mathbf{1})}{N_J - N_{K\cup J}} & , \textit{if } \mathbf{x}_J = \mathbf{1} \wedge \mathbf{x}_K \neq \mathbf{1} \\ \frac{\hat{P}_K(\mathbf{1}) - \hat{P}_{K\cup J}(\mathbf{1})}{N_K - N_{K\cup J}} & , \textit{if } \mathbf{x}_J \neq \mathbf{1} \wedge \mathbf{x}_K = \mathbf{1} \\ \frac{1 - \hat{P}_K(\mathbf{1}) - \hat{P}_J(\mathbf{1}) + \hat{P}_{K\cup J}(\mathbf{1})}{|\mathrm{Dom}(H)| - N_K - N_J + N_{K\cup J}} & , \textit{if } \mathbf{x}_J \neq \mathbf{1} \wedge \mathbf{x}_K \neq \mathbf{1}, \end{cases}$$

*for* $\mathbf{x} \in \mathrm{Dom}(H)$.

*Proof.* For every $R \subseteq \{K, J\}$ denote $X_R$ the set of all $\mathbf{x} \in \mathrm{Dom}(H)$ such that $\mathbf{x}_I = \mathbf{1}$ if $I \in R$ and $\mathbf{x}_I \neq \mathbf{1}$ otherwise, for all $I \in \{J, K\}$. Note that $|X_{\{K,J\}}| = N_{K\cup J}$, $|X_{\{J\}}| = N_J - N_{K\cup J}$, $|X_{\{K\}}| = N_K - N_{K\cup J}$, and $|X_\emptyset| = |\mathrm{Dom}(H)| - N_K - N_J + N_{K\cup J}$. Also denote $P_R^* = \sum_{\mathbf{x}\in X_R} \hat{P}(\mathbf{x})$ for all $R \subseteq \{K, J\}$. Note that $P^*_{\{K,J\}} = \hat{P}_{K\cup J}(\mathbf{1})$, $P^*_{\{J\}} = \hat{P}_J(\mathbf{1}) - \hat{P}_{K\cup J}(\mathbf{1})$, $P^*_{\{K\}} = \hat{P}_K(\mathbf{1}) - \hat{P}_{K\cup J}(\mathbf{1})$, and $P^*_\emptyset = 1 - \hat{P}_K(\mathbf{1}) - \hat{P}_J(\mathbf{1}) + \hat{P}_{K\cup J}(\mathbf{1})$.
For a probability distribution $P$ on $H$ that satisfies the the set of constraints $\mathcal{C}$ we have:

$$\begin{aligned} H(P) &= -\sum_{R\subseteq\{K,J\}} \sum_{\mathbf{x}\in X_R} P(\mathbf{x}) \log P(\mathbf{x}) \\ &= -\sum_{R\subseteq\{K,J\}} P_R^* \sum_{\mathbf{x}\in X_R} \frac{P(\mathbf{x})}{P_R^*} \log \frac{P(\mathbf{x})}{P_R^*} - \sum_{R\subseteq\{K,J\}} P_R^* \log P_R^*. \end{aligned}$$

It suffices to maximize the first term. Notice that for every $R \subseteq \{K, J\}$, $\sum_{\mathbf{x}\in X_R} \frac{P(\mathbf{x})}{P_R^*} = 1$ and thus $P/P_R^*$ is a probability distribution over $X_R$, and its entropy $-\sum_{\mathbf{x}\in X_R} \frac{P(\mathbf{x})}{P_R^*} \log \frac{P(\mathbf{x})}{P_R^*}$ is maximized when $\frac{P(\mathbf{x})}{P_R^*} = 1/|X_R|$ for every $\mathbf{x} \in X_R$. This gives $P(\mathbf{x}) = P_R^*/|X_R|$ for every $\mathbf{x} \in X_R$ and completes the proof since every $\mathbf{x} \in \mathrm{Dom}(H)$ belongs to exactly one of the $X_R$'s. □

Notice that when $J = K$, the above result reduces to

$$P^{\mathcal{C}}(\mathbf{x}) = \begin{cases} \frac{\hat{P}_J(\mathbf{1})}{N_J} & , \text{if } \mathbf{x}_J = \mathbf{1} \\ \frac{1-\hat{P}_J(\mathbf{1})}{|\mathrm{Dom}(H)| - N_J} & , \text{if } \mathbf{x}_J \neq \mathbf{1}, \end{cases} \tag{1}$$

for $\mathbf{x} \in \mathrm{Dom}(H)$.
Frequently the only subrules of a rule $I \to J$ are $\emptyset \to J$, and $K \to J$, where $K \subset I$. In this case the MaxENT distribution induced by the subrules can be found by application of Theorem 1. If the only subrule is $\emptyset \to J$, then we can use Equality (1). Our experiments revealed that using the above theorem reduces pruning time up to a factor of 10. See [Bad95,PMS01] for a more detailed discussion of methods of speeding up MaxENT computations.

| antecedent→lenses | conf. [%] | supp. [%] |
|---|---|---|
| $\emptyset$→soft | 20.8 | 20.8 |
| $\emptyset$→hard | 16.6 | 16.6 |
| $\emptyset$→none | 62.5 | 62.5 |
| tears=reduced→none | 100 | 50 |
| astigmatism=no,tears=normal→soft | 83.3 | 20.8 |
| astigmatism=yes,tears=normal→hard | 66.6 | 16.6 |
| age=pre-presbyopic,prescription=hypermetrope,astigmatism=yes→none | 100 | 8.3 |
| age=presbyopic,prescription=myope,astigmatism=no→none | 100 | 8.3 |
| age=presbyopic,prescription=hypermetrope,astigmatism=yes→none | 100 | 8.3 |

**Table 1.** Rules manually selected from the `lenses` database

## 4 Experimental Evaluation of the Pruning Algorithm

In this section we present an experimental evaluation of our pruning algorithm. We used passive interestingness $I^{pass}$, and considered a rule $I^{pass}$-nonredundant if its passive interestingness was greater than some threshold. Our experiments have shown that the passive measure of interestingness performed better than the active one $I^{act}$, which often pruned interesting rules with small support. The reason for that is that rules with small support usually have many attributes in the antecedent, and thus adding them as constraints affects only very few values in the joint probability distribution, while active interestingness depends on the whole distribution. Also, we did not use any minimum confidence threshold, because pruning provided a sufficient reduction in the number of rules, and setting a minimum confidence threshold occasionally pruned some of the interesting rules.

We first present the result of running the algorithm on the `lenses` database from the UCI machine learning archive [BM98]. The database has the advantage of being very small thus allowing manual selection of rules. Table 1 shows the rules having the `lenses` attribute as consequent, selected manually by the authors, providing a complete description of the dataset. Table 2 shows rules involving `lenses` attribute as consequent generated by the Apriori algorithm with minimum support 1 (1 record), no minimum confidence, post-processed with our pruning algorithm using passive interestingness with interestingness threshold 0.3. Negative values of interestingness mean that the presence of the antecedent decreases the probability of presence of the consequent.

Rules have been sorted based on the product of support and interestingness, with an extra condition, that a rule cannot be printed until all its subrules have been printed. Also, note that the `lenses` dataset contains multivalued attributes. Since our method only handles boolean attributes we encode each original attribute with a number of boolean attributes, one for each possible value of the original attribute.

The Apriori algorithm produced 113 rules having `lenses` attribute as the consequent. After pruning, 16 nonredundant rules were left with a nonempty antecedent. This is a significant reduction.

When rules with all possible consequents are considered, our method outputs 40 rules out of 890 produced by Apriori. Also, note that all rules selected manually are also considered interesting by our pruning

| antecedent→lenses | $I_P$ [%] | conf. [%] | supp. [%] |
|---|---|---|---|
| ∅→soft | 0 | 20.8 | 20.8 |
| ∅→hard | 0 | 16.6 | 16.6 |
| ∅→none | 0 | 62.5 | 62.5 |
| tears=reduced→none | 37.5 | 100 | 50 |
| astigmatism=no,tears=normal→soft | 62.5 | 83.3 | 20.8 |
| astigmatism=yes,tears=normal→hard | 50 | 66.6 | 16.6 |
| tears=normal→none | -37.5 | 25 | 12.5 |
| prescription=myope,astigmatism=yes→hard | 33.3 | 50 | 12.5 |
| prescription=myope,tears=normal→hard | 33.3 | 50 | 12.5 |
| prescription=hypermetrope,astigmatism=yes,tears=normal→none | 41.4 | 66.6 | 8.3 |
| age=pre-presbyopic,prescription=hypermetrope,astigmatism=yes→none | 37.5 | 100 | 8.3 |
| age=presbyopic,prescription=myope,astigmatism=no→none | 37.5 | 100 | 8.3 |
| age=presbyopic,prescription=hypermetrope,astigmatism=yes→none | 37.5 | 100 | 8.3 |
| age=young,astigmatism=yes→hard | 33.3 | 50 | 8.3 |
| age=young,tears=normal→hard | 33.3 | 50 | 8.3 |
| age=presbyopic,astigmatism=no,tears=normal→soft | -32.9 | 50 | 4.1 |
| age=presbyopic,prescription=hypermetrope,astigmatism=no,tears=normal→soft | 49.3 | 100 | 4.1 |
| prescription=hypermetrope,astigmatism=yes,tears=normal→hard | -32.9 | 33.3 | 4.1 |
| age=young,prescription=hypermetrope,astigmatism=yes,tears=normal→hard | 39.3 | 100 | 4.1 |

**Table 2.** Rules selected from the `lenses` database

| antecedent→lenses | $I_P$ [%] | conf. [%] | supp. [%] |
|---|---|---|---|
| ∅→urban=no | 0 | 22.4 | 22.4 |
| ∅→urban=yes | 0 | 77.5 | 77.5 |
| immigr=no,region=south→urban=yes | -11.8 | 65.7 | 26.2 |
| race=white→urban=yes | -10.6 | 66.8 | 22.5 |
| region=west→urban=yes | 12.8 | 90.3 | 16.9 |
| race=hisp→urban=yes | 12.4 | 89.9 | 15.4 |
| region=south,race=black→urban=yes | -10.6 | 66.8 | 17.8 |
| immigr=no,region=south→urban=no | 11.8 | 34.2 | 13.7 |
| alone=yes,region=south→urban=yes | -10.5 | 66.9 | 15 |
| immigr=before75→urban=yes | 15.9 | 93.4 | 9.7 |
| region=neast,race=black→urban=yes | 19.7 | 97.2 | 6.7 |
| region=midw,race=black→urban=yes | 18.9 | 96.5 | 6.9 |
| age=below75,region=neast→urban=yes | 10.5 | 88 | 11.3 |
| race=white→urban=no | 10.6 | 33.1 | 11.1 |

**Table 3.** Top 12 rules involving `urban` attribute generated from the elderly people census data

algorithm, and the top three rules are indeed identical in both cases, which suggests that really interesting rules are indeed retained by our algorithm.

We also applied our method to a dataset of census data of elderly people obtained from The University of Massachusetts at Boston Gerontology Center. The dataset contains about 330 thousand records, 11 attributes with up to five values, and is available at `http://www.cs.umb.edu/ ~sj /datasets /census.arff.gz`. We used 1% minimum support and no minimum confidence. The Apriori algorithm produced 247476 rules practically impossible to analyze by hand. After pruning with 10% interestingness threshold only 2056 were considered nonredundant, and after further restricting this set to rules with a given consequent attribute we were able to obtain easily manageable sets of interesting association rules. Some of them, concerning the `urban` (whether a person lives in a city or not) attribute are given in Table 3. Although the pruning time was quite long (over 4 hours on a 100MHz Pentium machine), it was still much easier to use our method than to handle hundreds of thousands of rules manually. See Table 4 for further details.

Table 4 shows the number of rules generated by Apriori compared with the number of rules considered interesting by our algorithm, as well

| dataset | min. support | interestingness threshold | number of rules Apriori | after pruning | pruning time [s] |
|---|---|---|---|---|---|
| lenses | 1(4%) | 0.3 | 890 | 40 | 1.3 |
| mushroom* | 500(16%) | 0.2 | 164125 | 5141 | 418 |
| breast-cancer | 30(10%) | 0.15 | 2128 | 74 | 2.8 |
| primary-tumor* | 30(9%) | 0.3 | 43561 | 67 | 21.8 |
| primary-tumor* | 30(9%) | 0.2 | 43561 | 432 | 24 |
| car | 10(0.5%) | 0.3 | 20669 | 293 | 11.1 |
| car | 10(0.5%) | 0.15 | 20669 | 580 | 30.2 |
| splice* | 300(9%) | 0.5 | 4847 | 24 | 3.0 |
| splice* | 300(9%) | 0.3 | 4847 | 95 | 5.6 |
| splice* | 300(9%) | 0.15 | 4847 | 290 | 7.2 |
| splice* | 200(6%) | 0.3 | 35705 | 463 | 33.8 |
| census(elderly people) | 3000(1%) | 0.3 | 247476 | 194 | 4801 |
| census(elderly people) | 3000(1%) | 0.2 | 247476 | 621 | 5683 |
| census(elderly people) | 3000(1%) | 0.1 | 247476 | 2056 | 15480 |

* itemsets of up to 4 attributes

**Table 4.** Numbers of rules and computation times for various datasets

as pruning time, for various datasets from the UCI Machine Learning Archive [BM98]. All datasets have been mined with 0 minimum confidence. The interestingness thresholds and minimum supports have been chosen manually by trial and error such that the unpruned rules provide a lot of interesting information while keeping their number reasonably small. For some datasets values for a few different thresholds are given for comparison. All experiments have been performed on a 100MHz Pentium machine with 64MB of memory.

## 5 Conclusions and further research

A method for pruning redundant association rules using the Maximum Entropy approach has been presented along with methods of speeding up MaxENT computations by using closed form formulas. The method has been experimentally shown to produce relatively small sets of interesting rules in a reasonable amount of time. A detailed analytical comparison of our method with other approaches has also been presented.

We plan to concentrate our further efforts on including background knowledge in the mining/pruning process. In most real applications the user already has a lot of domain knowledge about the dataset and is only interested in rules which are not implied by what is already known. We believe that this is the approach one should use to achieve further reductions in the number of rules and make them more applicable for the user.

Currently, the selection of interestingness threshold is made by trial and error. A more formal procedure, possibly based on statistical tests and confidence levels would be very useful.

Even though our method was fast enough to apply it to real datasets, we plan to further improve its performance by analyzing which configurations of constraints occur most frequently and provide closed form solutions for them.

## 6 Acknowledgments

The authors would like to thank Prof. Jeffrey Burr from the University of Massachusetts at Boston Gerontology Center for providing us with the elderly people census data.

## References

[Adw97] Ratnaparkhi Adwait. A simple introduction to maximum entropy models for natural language processing. IRCS Report 97–08, University of Pennsylvania, 3401 Walnut Street, Suite 400A, Philadelphia, PA, May 1997. `ftp://www.cis.upenn.edu/pub/ircs/tr/97-08.ps.Z`.

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD Conference on Management of Data*, pages 207–216, Washington, D.C., 1993.

[AL99] Y. Aumann and Y. Lindell. A statistical theory for quantitative association rules. In *Knowledge Discovery and Data Mining*, pages 261–270, 1999.

[And74] A. H. Andersen. Multidimensional contigency tables. *Scand. J. Statist*, 1:115–127, 1974.

[BA99] R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 145–154, August 1999.

[Bad95] J. Badsberg. *An Environment for Graphical Models*. PhD thesis, Aalborg University, 1995.

[BM98] C.L. Blake and C.J. Merz. UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[BVW00] T. Brijs, K. Vanhoof, and G. Wets. Reducing redundancy in characteristic rule discovery by using integer programming techniques. *Intelligent Data Analysis Journal*, 4(3), 2000.

[Csi89] I. Csiszar. A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. *The Annals of Staticstics*, 17(3):1409–1413, 1989.

[DLS80] J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear interaction models for contingency tables. *Annals of Statistics*, 8:522–539, 1980.

[DR72] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470–1480, 1972.

[GHK94] A. J. Grove, J. Y. Halpern, and D. Koller. Random worlds and maximum entropy. *Journal of Artificial Intelligence Research*, 2:33–88, 1994.

[HH99] R. Hilderman and H. Hamilton. Knowledge discovery and interestingness measures: A survey. Technical Report CS 99-04, Department of Computer Science, University of Regina, 1999.

[JS01] S. Jaroszewicz and D. A. Simovici. A general measure of rule interestingness. In *Proc of PKDD 2001, Freiburg, Germany*, volume 2168 of *Lecture Notes in Computer Science*, pages 253–265. Springer, September 2001.

[KK92] J. N. Kapur and H. K. Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, San Diego, 1992.

[LHM99] Bing Liu, Wynne Hsu, and Yiming Ma. Pruning and summarizing the discovered associations. In Surajit Chaudhuri and David Madigan, editors, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 125–134, N.Y., August 15–18 1999. ACM Press.

[PMS01] D. Pavlov, H. Mannila, and P. Smyth. Beyond independence: Probabilistic models for query approximation on binary transaction data. Technical Report ICS TR-01-09, Information and Computer Science Department, UC Irvine, 2001.

[PT00] B. Padmanabhan and A. Tuzhilin. Small is beautiful: discovering the minimal set of unexpected patterns. In Raghu Ramakrishnan, Sal Stolfo, Roberto Bayardo, and Ismail Parsa, editors, *Proceedinmgs of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-00)*, pages 54–63, N. Y., August 2000. ACM Press.

[Rat96] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142. Association for Computational Linguistics, Somerset, New Jersey, 1996.

[SK98] E. Suzuki and Y. Kodratoff. Discovery of surprising exception rules based on intensity of implication. In *Proc of PKDD-98, Nantes, France*, pages 10–18, 1998.

[SLRS99] D. Shah, L. V. S. Lakshmanan, K. Ramamritham, and S. Sudarshan. Interestingness and pruning of mined patterns. In *1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1999.

[Suz97] E. Suzuki. Autonomous discovery of reliable exception rules. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, page 259. AAAI Press, 1997.

[TKR+95] H. Toivonen, M. Klemettinen, P. Ronkainen, K. Hätönen, and H. Mannila. Pruning and grouping discovered association rules. In *MLnet Workshop on Statistics, Machine Learning, and Discovery in Databases*, pages 47–52, Heraklion, Crete, Greece, April 1995.