Connections between mining frequent itemsets and learning generative models

Srivatsan Laxman
Microsoft Research Labs
India
slaxman@microsoft.com

Prasad Naldurg
Microsoft Research Labs
India
prasadn@microsoft.com

Raja Sripada Microsoft Research Labs
India
t-rajas@microsoft.com

Ramarathnam Venkatesan
Microsoft Research Labs
Redmond
venkie@microsoft.com

Abstract

Frequent itemsets mining is a popular framework for pattern discovery. In this framework, given a database of customer transactions, the task is to unearth all patterns in the form of sets of items appearing in a sizable number of transactions. We present a class of models called Itemset Generating Models (or IGMs) that can be used to formally connect the process of frequent itemsets discovery with the learning of generative models. IGMs are specified using simple probability mass functions (over the space of transactions), peaked at specific sets of items and uniform everywhere else. Under such a connection, it is possible to rigorously associate higher frequency patterns with generative models that have greater data likelihoods. This enables a generative model-learning interpretation of frequent itemsets mining. More importantly, it facilitates a statistical significance test which prescribes the minimum frequency needed for a pattern to be considered interesting. We illustrate the effectiveness of our analysis through experiments on standard benchmark data sets.

1 Introduction

Frequent itemsets mining [1] is a popular framework for pattern discovery. Consider a collection of transactions in a grocery store. Each transaction comprises items bought during one visit by a customer to the store. The problem is to discover popular buying patterns (in the form of groups of items frequently bought together in a transaction). The patterns are referred to as itemsets and the task is to efficiently discover all frequent itemsets in the data. Although originally proposed as a technique for market basket analysis [1], the idea of discovering frequent itemsets (and using these to deduce so-called association rules) has since gained wide applicability [7] in many different application domains, ranging from biology [3] to remote sensing [4].

This paper presents a formal connection between frequent itemsets mining and learning of generative models for the data in the form of simple probability mass functions over the space of transactions. We define a class of models called Itemset Generating Models (or IGMs). Each IGM generates transactions by embedding a specific itemset (or pattern) in a transaction according to a probability distribution that is peaked at the pattern in question, and is uniform everywhere else. In this manner, each itemset is associated with a specific IGM. We then prove that given any two itemsets, the IGM associated with the more frequent itemset is the one more likely to generate the database of transactions. This rigorous connection between itemsets and IGMs has interesting consequences. First, it allows for a generative model-learning interpretation of the frequent itemsets mining process. Second, it facilitates a formal statistical significance test, which prescribes, the minimum frequency needed in order to regard an itemset as significant for a given error probability. This is useful because frequency threshold is a user-defined parameter that can be very hard to fix in typical applications. Simulation experiments show that our theoretically motivated frequency thresholds are effective in discovering patterns embedded in the data.

In general, connecting pattern discovery frameworks with generative model-learning is a useful idea. There is a growing need for unified frameworks in data mining [11] that can simultaneously exploit efficient counting procedures in pattern discovery (that typically have a combinatorial flavor) with the rigorous statistical basis of generative model-learning. Motivated by such considerations, in [9], a formal connection was established between pattern discovery and model-learning in the context of *temporal* data min-

ing. The ideas presented in this paper have a similar flavor, applied to the frequent itemsets framework in (unordered) transaction databases.

The rest of the paper is organized as follows. Sec. 2 provides a brief introduction to frequent itemsets discovery. Sec. 3 defines Itemsets Generating Models and presents the main results of the paper. The generative model-learning interpretation is discussed in Sec. 4. Sec. 5 describes the statistical significance test for itemsets. Simulation experiments are described in Sec. 6 and Sec. 7 presents conclusions.

2 Frequent itemsets framework

We first present some background on frequent itemsets mining. Let $\mathcal{D} = \{T_1, \ldots, T_K\}$ be a database of (unordered) customer transactions at a store. Each *transaction*, say T_i, is a collection of items purchased by a customer in one visit to the store. A non-empty set of items is called an *itemset*. An itemset is denoted as $\alpha = (A_1, A_2, \ldots, A_N)$, where each A_j is an item (or symbol) from some finite alphabet, say $\mathcal{A}$. Since α has N items, it is referred to as an N-itemset. Trivially, each transaction in the database is an itemset. However, an arbitrary itemset may or may not be contained in a given transaction, T_i. The *frequency* of an itemset is the number of transactions in $\mathcal{D}$ that contain it. An itemset whose frequency exceeds a user-defined threshold is referred to as a frequent itemset. These frequent itemsets are the patterns of interest in the problem.

The brute force method of determining frequencies for all possible itemsets (of size N, for various N) is a combinatorially explosive exercise and is not feasible in large databases (which is typically the case in data mining). The Apriori algorithm [1, 2] exploits the following simple but useful principle (known as the anti-monotonicity principle): if α and β are itemsets such that β is a subset of α, then the frequency of β is greater than or equal to the frequency of α. Thus, for an itemset to be frequent all its subsets must be frequent as well. This gives an efficient level-wise search for the frequent itemsets in $\mathcal{D}$. The algorithm makes multiple passes over the data. Starting with 1-itemsets, every pass discovers frequent itemsets of the next bigger size. Frequent 1-itemsets obtained in the first pass are combined to generate candidate 2-itemsets, and then, by counting their frequencies (using a second pass over the data) frequent 2-itemsets are found, and so on. Note that this kind of breadth-first search of [1] is not the only algorithm available for frequent itemsets mining. Over the years, many other algorithms have been proposed for frequent itemsets mining (e.g., see [8] for a comparison of several approaches). The analysis we present is independent of the algorithm used for frequent itemsets mining.

3 Itemset generating models

We now present our Itemset Generating Model (or IGM) and establish the connection between learning these models and discovering frequent itemsets.

3.1 Data likelihood under IGMs

Let $\mathcal{A}$ be an alphabet (i.e. a universal set of items) of size M. Let $\alpha = (A_1, \ldots, A_N)$ be an N-itemset. The tuple, $T = (t_1, \ldots, t_{|T|})$, represents a transaction over $\mathcal{A}$.

An Itemset Generating Model (or IGM) is specified as a tuple, $\Lambda = (\alpha, \theta)$, where, $\alpha \subseteq \mathcal{A}$, is an N-itemset, referred to as the "pattern" of IGM Λ, and $\theta \in [0, 1]$, is referred to as the "pattern probability" of IGM Λ. The class of all IGMs, obtained by considering all possible N-itemsets over the alphabet, $\mathcal{A}$, and by considering all possible pattern probability values, $\theta \in [0, 1]$, is denoted by $\mathcal{I}$.

The probability model according to which an IGM, $\Lambda = (\alpha, \theta)$, generates a transaction, T, is as follows: We refer to (the power set) 2^α, as the *pattern space*, and to $2^{\bar{\alpha}}$ as the *noise space* ($\bar{\alpha}$ denotes the complement of α, i.e., $\mathcal{A} \setminus \alpha$). The space of all possible transactions is $2^\mathcal{A}$, which is the Cartesian product $(2^\alpha \times 2^{\bar{\alpha}})$. The IGM generates two itemsets, denoted $T(\alpha) \in 2^\alpha$ and $T(\bar{\alpha}) \in 2^{\bar{\alpha}}$, *independent* of each other, and according to the following probability mass functions:

$$T(\alpha) = \begin{cases} \alpha & \text{wp} \quad \theta \\ \alpha' \subsetneq \alpha & \text{wp} \quad \left(\frac{1-\theta}{2^N - 1}\right) \end{cases} \tag{1}$$

$$T(\bar{\alpha}) = \alpha'' \subseteq \bar{\alpha} \quad \text{wp} \quad \left(\frac{1}{2^{M-N}}\right) \tag{2}$$

Then, the full transaction, T, is generated as $T = T(\alpha) \cup T(\bar{\alpha})$. We can think of $T(\alpha)$ and $T(\bar{\alpha})$ as the corresponding projections of T on the pattern space and noise space of Λ. Eq. (1) is a probability mass function over 2^α (or pattern space), that is peaked at $T(\alpha) = \alpha$ (so long as $\theta > (\frac{1}{2})^N$) and is uniform everywhere else, and Eq. (2) is a uniform probability distribution over $2^{\bar{\alpha}}$ (or noise space). Moreover, $T(\alpha)$ and $T(\bar{\alpha})$ are independent of each other, and so, for $\theta \in (0, 1)$, the probability of $T = T(\alpha) \cup T(\bar{\alpha})$, under Λ, can be written as:

$$\begin{aligned} P[T \mid \Lambda] &= P[T(\alpha) \mid \Lambda] \times P[T(\bar{\alpha}) \mid \Lambda] \\ &= \theta^{z_\alpha(T)} \left(\frac{1-\theta}{2^N - 1}\right)^{1 - z_\alpha(T)} \left(\frac{1}{2^{M-N}}\right) \end{aligned} \tag{3}$$

where, $z_\alpha(\cdot)$ indicates set containment: $z_\alpha(T) = 1$, if $\alpha \subseteq T$, and $z_\alpha(T) = 0$, otherwise.

Example 1. *$\mathcal{A} = \{A, B, \ldots, Z\}$ is an alphabet of size $M = 26$. $\alpha = (A, B, C)$ is a 3-itemset over $\mathcal{A}$. An IGM $\Lambda = (\alpha, \theta)$, partitions $\mathcal{A}$ into $\{A, B, C\}$ and $\mathcal{A} \setminus \{A, B, C\}$.*

Λ generates $T(\alpha) \subset \{A, B, C\}$ according to Eq. (1) (which is peaked at $T(\alpha) = (A, B, C)$ if $\theta \in (\frac{1}{2^3}, 1)$) and generates $T(\bar{\alpha})$ according to Eq. (2) (which is uniform over $2^{\mathcal{A}\setminus\{A,B,C\}}$). If $\theta \in (0, 1)$, the probability for $T = T(\alpha) \cup T(\bar{\alpha})$, under Λ, is given by Eq. (3).

There is a minor technical difficulty in the use of Eq. (3) for the two extreme cases of $\theta = 0$ and $\theta = 1$. Therefore, in order to complete the description of the probability model prescribed in Eqs. (1)–(2), we need to specify the expressions for $P[T \mid \Lambda]$ for these cases. For $\theta = 0$, we have

$$P[T \mid \Lambda] = \begin{cases} \left(\frac{1}{(2^N-1)2^{M-N}}\right) & \text{if } z_\alpha(T) = 0 \\ 0 & \text{if } z_\alpha(T) = 1 \end{cases} \quad (4)$$

Similarly, for the case, $\theta = 1$, we have

$$P[T \mid \Lambda] = \begin{cases} 0 & \text{if } z_\alpha(T) = 0 \\ \left(\frac{1}{2^{M-N}}\right) & \text{if } z_\alpha(T) = 1 \end{cases} \quad (5)$$

These extreme cases are of little interest from the point of view of understanding the model, because they correspond to two distinct (trivial) situations: $T(\alpha) = \alpha$ in all transactions that the model generates, and $T(\alpha) \neq \alpha$ in all transactions that the model generates. To simplify our discussion, we first focus on the interesting case of $0 < \theta < 1$, and return to the extreme cases later.

An IGM can generate a database of transactions by drawing *iid* samples according to Eq. (3). Intuitively, if θ is large, the IGM would generate data with α appearing in many transactions. This makes IGMs a reasonable class of models to use for connections with frequent itemset mining. We formalize these ideas in the subsections to follow.

Consider a database, $\mathcal{D} = \{T_1, \ldots, T_K\}$, of K transactions. Each T_i is a collection of items over $\mathcal{A}$. Using Eq. (3), the expression for likelihood of $\mathcal{D}$, under $\Lambda = (\alpha, \theta)$, $\theta \in (0, 1)$, is as follows:

$$\begin{aligned} P[\mathcal{D} \mid \Lambda] &= \prod_{i=1}^{K} P[T_i \mid \Lambda] \\ &= \theta^{f_\alpha} \left(\frac{1-\theta}{2^N - 1}\right)^{K - f_\alpha} \left(\frac{1}{2}\right)^{K(M-N)} \end{aligned} \quad (6)$$

where, $f_\alpha = \sum_{i=1}^{K} z_\alpha(T_i)$ is the *frequency* of α in $\mathcal{D}$.

Example 2. *Consider a database, $\mathcal{D}$, of $K = 5$ transactions over the same alphabet as in* Example 1*: 1) $T_1 = (A, B, C, X, Y)$, 2) $T_2 = (A, E, F, G)$, 3) $T_3 = (C, U, V, Z)$, 4) $T_4 = (H, I, J)$ and 5) $T_5 = (A, B, C, D, E)$. Let $\Lambda = (\alpha, \theta)$, $\theta \in (0, 1)$, be the IGM defined in* Example 1*. The itemset $\alpha = (A, B, C)$ has $N = 3$ and appears in 2 out of 5 transactions. So, $f_\alpha = 2$ and the likelihood of $\mathcal{D}$ under Λ (using Eq. (6)) is as follows:*

$$P[\mathcal{D} \mid \Lambda] = \theta^2 \left(\frac{1-\theta}{2^3 - 1}\right)^{5-2} \left(\frac{1}{2}\right)^{5(26-3)}$$

3.2 Class of IGMs with a fixed θ

We now restrict our attention to IGMs with a fixed pattern probability $\theta \in (0, 1)$. Later, in Sec. 3.3, we consider the full class of IGMs with all possible values for the pattern probability parameter.

Definition 1. *The subclass $\mathcal{I}_\theta$, is defined as the collection of IGMs (out of $\mathcal{I}$) with a fixed pattern probability parameter $\theta \in (0, 1)$.*

Definition 2. *Consider an N-itemset, $\alpha = (A_1, \ldots, A_N)$ ($\alpha \in 2^{\mathcal{A}}$). The IGM associated with itemset α is given by $\Lambda_\alpha = (\alpha, \theta) \in \mathcal{I}_\theta$.*

Under this association between itemsets and IGMs, we show that more frequent itemsets are associated with IGMs having greater data likelihoods.

Theorem 1. *Let $\mathcal{D} = \{T_1, \ldots, T_K\}$ be a database of transactions over alphabet $\mathcal{A}$. Let α and β be two N-itemsets that occur in $\mathcal{D}$, with frequencies f_α and f_β respectively. Consider the class, $\mathcal{I}_\theta$, of IGMs with $\theta > (\frac{1}{2^N})$, and let Λ_α and Λ_β be the corresponding IGMs (from $\mathcal{I}_\theta$) that are associated with α and β according to Definition 2. Then we have, $P[\mathcal{D} \mid \Lambda_\alpha] > P[\mathcal{D} \mid \Lambda_\beta]$ if and only if $f_\alpha > f_\beta$.*

Proof. The likelihoods for the data, $\mathcal{D}$, under Λ_α and Λ_β, can be written using Eq. (6) as follows:

$$\begin{aligned} P[\mathcal{D} \mid \Lambda_\alpha] &= \theta^{f_\alpha} \left(\frac{1-\theta}{2^N - 1}\right)^{K - f_\alpha} \left(\frac{1}{2}\right)^{K(M-N)} \\ P[\mathcal{D} \mid \Lambda_\beta] &= \theta^{f_\beta} \left(\frac{1-\theta}{2^N - 1}\right)^{K - f_\beta} \left(\frac{1}{2}\right)^{K(M-N)} \end{aligned}$$

Therefore, we have:

$$\frac{P[\mathcal{D} \mid \Lambda_\alpha]}{P[\mathcal{D} \mid \Lambda_\beta]} = \left[\theta \left(\frac{2^N - 1}{1 - \theta}\right)\right]^{f_\alpha - f_\beta} \quad (7)$$

Given $\theta > (\frac{1}{2^N})$, and since $N \geq 1$, this implies $\left[\theta \left(\frac{2^N - 1}{1-\theta}\right)\right] > 1$. From Eq. (7), we have $\frac{P[\mathcal{D} \mid \Lambda_\alpha]}{P[\mathcal{D} \mid \Lambda_\beta]} > 1$ if and only if $f_\alpha > f_\beta$. □

It follows from *Theorem 1* that, if α is the most frequent N-itemset in $\mathcal{D}$, then for any other N-itemset, β, the data likelihoods under Λ_α and Λ_β must obey: $P[\mathcal{D} \mid \Lambda_\alpha] \geq P[\mathcal{D} \mid \Lambda_\beta]$. This gives us the next theorem.

Theorem 2. *Given a database $\mathcal{D}$ of transactions over an alphabet $\mathcal{A}$, the maximum likelihood estimate over the class of IGMs with $\theta > (\frac{1}{2^N})$, given by $\mathcal{I}_\theta$, is the IGM corresponding to the most frequent N-itemset in $\mathcal{D}$ (with correspondence as prescribed by* Definition 2*).*

In other words, when $\theta > (\frac{1}{2^N})$, frequencies of N-itemsets in $\mathcal{D}$ are "sufficient statistics" for maximum likelihood estimation over IGMs in $\mathcal{I}_\theta$. For $\mathcal{I}_\theta$ with $\theta < (\frac{1}{2^N})$, *Theorems 1 & 2* do not hold. When $\theta < (\frac{1}{2^N})$, data likelihood *decreases* with increasing frequency of the corresponding N-itemset. This is because, if $f_\alpha > f_\beta$ in Eq. (7), $\theta \leq (\frac{1}{2^N})$ will imply $\frac{P[\mathcal{D} \mid \Lambda_\alpha]}{P[\mathcal{D} \mid \Lambda_\beta]} \leq 1$. In particular, setting $f_\beta = 0$, we get

$$P[\mathcal{D} \mid \Lambda_\alpha] \leq \left(\frac{1-\theta}{2^N-1}\right)^K \left(\frac{1}{2}\right)^{K(M-N)}.$$

The above upper bound is maximum at $\theta = 0$.

Remark 1. *Given a database, $\mathcal{D}$, of K transactions over alphabet $\mathcal{A}$ (of size M) and an IGM $\Lambda = (\alpha, \theta)$, with α of size N and $\theta \leq (\frac{1}{2^N})$, we have*

$$P[\mathcal{D} \mid \Lambda] \leq \left(\frac{1}{2^N-1}\right)^K \left(\frac{1}{2}\right)^{K(M-N)}. \quad (8)$$

Further, for large values of N, this upper bound approaches $(\frac{1}{2^M})^K$, which is simply the probability of $\mathcal{D}$ when the transactions are drawn iid according to a uniform distribution over the whole of $2^{\mathcal{A}}$.

Remark 1 shows that IGMs with $\theta \leq (\frac{1}{2})^N$ cannot be meaningfully associated with frequent itemsets. For the itemset-IGM connection to be really useful, we now need a way to estimate θ automatically from the data and associate each itemset with an IGM from the full class, $\mathcal{I}$, of IGMs over $\mathcal{A}$. This is considered next in Sec. 3.3.

3.3 The final itemsets-IGM association

Theorems 1 & 2 provide formal justification for the itemset-IGM correspondence of *Definition 2*, which only considers IGMs with a fixed pattern probability parameter. We now drop the fixed-θ constraint and seek an itemset-IGM correspondence over $\mathcal{I}$, the full class of IGMs.

Definition 3. *Given an N-itemset, $\alpha = (A_1, \ldots, A_N)$, the subclass $\mathcal{I}(\alpha)$ is defined as the collection of all IGMs in $\mathcal{I}$ of the form $\Lambda = (\alpha, \theta)$, with $\theta \in [0, 1]$.*

Clearly, IGMs in $\mathcal{I}(\alpha)$ are the only reasonable candidates (out of $\mathcal{I}$) for the final itemset-IGM association for α. By considering θ in the interval $[0, 1]$, we have infinite possibilities for such an association. Which of these will make for a good itemset-IGM association? A natural choice is one that maximizes likelihood of the data over all possible IGMs in $\mathcal{I}(\alpha)$. Using Eq. (6), (and once again, by first considering only the case of $\theta \in (0, 1)$), the log likelihood under the IGMs is given by:

$$\begin{aligned} l_\alpha(\theta) &= f_\alpha \log(\theta) + (K - f_\alpha)\log(1-\theta) \\ &\quad + \left\langle \begin{array}{c} \text{terms constant} \\ \text{wrt } \theta \end{array} \right\rangle \end{aligned} \quad (9)$$

Taking derivatives with respect to θ and equating to zero, we get $\theta = (f_\alpha/K)$ as the unique maximizer of $l_\alpha(\theta)$ for $\theta \in (0, 1)$. From Eqs. (4)-(5), both when $f_\alpha = K$ (i.e. when all transactions in $\mathcal{D}$ contain α), and when $f_\alpha = 0$, (f_α/K) automatically maximizes the data likelihood for $\mathcal{D}$ over $\mathcal{I}(\alpha)$. Thus, (f_α/K) is the unique maximizer for θ in the closed interval $[0, 1]$ as well. *The IGM $(\alpha, f_\alpha/K)$, has the highest likelihood for $\mathcal{D}$, over the class, $\mathcal{I}(\alpha)$, of IGMs defined using the N-itemset α.* However, (f_α/K), may or may not be greater than $(1/2^N)$. Based on all this, and bearing in mind *Remark 1*, we prescribe the final itemset-IGM association according to the following definition.

Definition 4. *Consider an N-itemset, $\alpha = (A_1, \ldots, A_N)$ ($\alpha \in 2^{\mathcal{A}}$). Let f_α denote the frequency of α in the given database, $\mathcal{D}$, of K transactions. The itemset α is associated with the IGM, $\Lambda = (\alpha, \theta_\alpha)$, with $\theta_\alpha = (\frac{f_\alpha}{K})$, if $(\frac{f_\alpha}{K}) > (\frac{1}{2^N})$, and with $\theta_\alpha = 0$ otherwise.*

Under this final itemset-IGM association of *Definition 4*, the IGMs associated with different itemsets typically have different pattern probability parameters. Is it still true that the data likelihood is higher for IGMs associated with itemsets with higher frequencies? The answer to this question is yes, and we state this property as our next theorem.

Theorem 3. *Let $\mathcal{D} = \{T_1, \ldots, T_K\}$ be a database of transactions over the alphabet $\mathcal{A}$. Let α and β be two N-itemsets that occur in $\mathcal{D}$, with frequencies f_α and f_β respectively. Let Λ_α and Λ_β be the corresponding IGMs (from $\mathcal{I}$) that are associated with α and β according to* Definition 4*. If θ_α and θ_β are both greater than $(\frac{1}{2^N})$, then we have, $P[\mathcal{D} \mid \Lambda_\alpha] > P[\mathcal{D} \mid \Lambda_\beta]$ if and only if $f_\alpha > f_\beta$.*

Proof. First, let us consider the case when, $\theta_\alpha, \theta_\beta \in (\frac{1}{2^N}, 1)$ (i.e., when neither of them is equal to 1). In this case, we know that θ_α is the unique maximizer of $l_\alpha(\theta)$ (where $l_\alpha(\theta)$ is the expression for log likelihood of the data for IGMs in $\mathcal{I}(\alpha)$ as given by Eq. (9)). This gives rise to the following inequality:

$$\begin{aligned} P[\mathcal{D} \mid \Lambda_\alpha] &= \theta_\alpha^{f_\alpha} \left(\frac{1-\theta_\alpha}{2^N-1}\right)^{K-f_\alpha} \left(\frac{1}{2}\right)^{K(M-N)} \\ &> \theta_\beta^{f_\alpha} \left(\frac{1-\theta_\beta}{2^N-1}\right)^{K-f_\alpha} \left(\frac{1}{2}\right)^{K(M-N)} \end{aligned} \quad (10)$$

However, since it is also given that $\theta_\beta > (\frac{1}{2^N})$, the following inequality holds if and only if $f_\alpha > f_\beta$:

$$\begin{aligned} &\theta_\beta^{f_\alpha} \left(\frac{1-\theta_\beta}{2^N-1}\right)^{K-f_\alpha} \left(\frac{1}{2}\right)^{K(M-N)} \\ &\quad > \theta_\beta^{f_\beta} \left(\frac{1-\theta_\beta}{2^N-1}\right)^{K-f_\beta} \left(\frac{1}{2}\right)^{K(M-N)} \\ &\quad = P[\mathcal{D} \mid \Lambda_\beta] \end{aligned} \quad (11)$$

Putting together the inequalities in (10) and (11), we have $P[\mathcal{D} \mid \Lambda_\alpha] > P[\mathcal{D} \mid \Lambda_\beta]$ if and only if $f_\alpha > f_\beta$, which completes the proof for the case when $\theta_\alpha, \theta_\beta \in (\frac{1}{2^N}, 1)$.

To complete the proof now, we need to finally consider the (somewhat trivial) boundary cases when $\theta_\alpha = 1$ and/or $\theta_\beta = 1$. Now, if $\theta_\beta = 1$, we have $f_\beta = K$, which implies that, neither $f_\alpha > f_\beta$ nor $P[\mathcal{D} \mid \Lambda_\alpha] > P[\mathcal{D} \mid \Lambda_\beta]$ can hold, because $P[\mathcal{D} \mid \Lambda_\beta] = (\frac{1}{2})^{K(M-N)} \geq P[\mathcal{D} \mid \Lambda_\alpha]$ for all N-itemsets α. Therefore, the only valid possibility left here is the case when $\theta_\alpha = 1$ and $\theta_\beta \in (\frac{1}{2^N}, 1)$, which implies, $f_\alpha = K$ and $\frac{K}{2^N} < f_\beta < K$. In this case $P[\mathcal{D} \mid \Lambda_\alpha] = (\frac{1}{2})^{K(M-N)}$, and it is easy to see that $P[\mathcal{D}|\Lambda_\alpha] > P[\mathcal{D}|\Lambda_\beta]$ if and only if $f_\alpha > f_\beta$. □

Finally, just like *Theorem 2* followed from *Theorem 1*, now, under the final itemset-IGM association of *Definition 4*, *Theorem 3* gives rise to *Theorem 4* about maximum likelihood estimation for $\mathcal{D}$ over the class, $\mathcal{I}$, of all IGMs (defined using patterns of size N).

Theorem 4. *Let $\mathcal{D}$ be a database of transactions over alphabet $\mathcal{A}$. Let α be the most frequent N-itemset in $\mathcal{D}$ and let $\Lambda_\alpha = (\alpha, \theta_\alpha)$ be the IGM corresponding to α (as prescribed by* Definition 4*). If $P[\mathcal{D} \mid \Lambda_\alpha] > \left(\frac{1}{2^N - 1}\right)^K \left(\frac{1}{2}\right)^{K(M-N)}$, then Λ_α is a maximum likelihood estimate for $\mathcal{D}$, over the class, $\mathcal{I}$, of all IGMs.*

Proof. The proof follows from *Theorem 3* and *Remark 1*. Since it is given that $P[\mathcal{D} \mid \Lambda_\alpha] > \left(\frac{1}{2^N - 1}\right)^K \left(\frac{1}{2}\right)^{K(M-N)}$, from *Remark 1* this automatically implies $\theta_\alpha > (\frac{1}{2^N})$. Now, let $\mathcal{X} \subset 2^{\mathcal{A}}$ denote the collection of all N-itemsets over $\mathcal{A}$. The class of IGMs $\mathcal{I}$, can be partitioned using the patterns out of $\mathcal{X}$ as follows: $\mathcal{I} = \cup_{\beta \in \mathcal{X}} \mathcal{I}(\beta)$. For each N-itemset, $\beta \in \mathcal{X}$, we know that the IGM defined by the tuple, $(\beta, f_\beta/K)$, has the highest likelihood for $\mathcal{D}$, over the partition $\mathcal{I}(\beta)$ (See discussion that precedes *Definition 4*). Thus, if Λ_β is set equal to $(\beta, f_\beta/K)$ according to *Definition 4*, from *Theorem 3* (since it is given that α is the most frequent itemset, and since we know that $\theta_\alpha > (\frac{1}{2^N})$ must hold), we know that for all $\beta \in \mathcal{X}$ such that $f_\beta > (\frac{K}{2^N})$:

$$\begin{aligned} P[\mathcal{D} \mid \Lambda_\alpha] &> P[\mathcal{D} \mid \Lambda_\beta] \\ &\geq P[\mathcal{D} \mid \Lambda] \ \ \forall \Lambda \in \mathcal{I}(\beta). \end{aligned}$$

Further, using the fact that $P[\mathcal{D} \mid \Lambda_\alpha]$ is greater than the upper bound in *Remark 1*, even for $\beta \in \mathcal{X}$ such that $f_\beta \leq (\frac{K}{2^N})$, we must have $P[\mathcal{D}|\Lambda_\alpha] > P[\mathcal{D}|\Lambda] \, \forall \Lambda \in \mathcal{I}(\beta)$. This completes the proof of *Theorem 4*. □

To summarize, this section presents the theoretical connections between itemsets and IGMs. First, in Sec. 3.2, we consider a class of IGMs $\mathcal{I}_\theta$ with a fixed pattern probability parameter θ. We show that among all the IGMs in $\mathcal{I}_\theta$, with $\theta > (\frac{1}{2^N})$, the IGMs associated (according to Definition 2) with itemsets that have higher frequencies in $\mathcal{D}$, have higher data likelihoods. In fact, the IGM associated with the most frequent itemset is a maximum likelihood estimate for the data over the class $\mathcal{I}_\theta$. Next, in Sec. 3.3, the fixed-θ assumption of Sec. 3.2 is dropped, and the full class $\mathcal{I}$ of IGMs (with all possible pattern probability values) is considered. The final association between itemsets and IGMs is prescribed by *Definition 4*, which fixes the pattern probability parameter of the IGM associated with an itemset, based on its frequency in the data. This is the association that we use in all our analysis. The main result connecting itemsets and IGMs is given by *Theorem 3*, which states that for sufficiently frequent itemsets (i.e. for N-itemsets with frequencies greater than $(\frac{K}{2^N})$), more frequent itemsets are associated with IGMs (in $\mathcal{I}$) with greater data likelihoods. This connection is tight, in the sense that under the conditions of *Theorem 3*, the itemset-IGM association of *Definition 4* ensures that ordering with respect to frequencies among N-itemsets over $\mathcal{A}$ is preserved as ordering with respect data likelihoods among IGMs in $\mathcal{I}$. Finally, based on *Theorem 3* and *Remark 1* we get *Theorem 4*, which states that if the most frequent itemset is frequent enough, then it is associated with an IGM which is a maximum likelihood estimate for the data, over the full class, $\mathcal{I}$, of IGMs.

The itemset-IGM connection presented in this section has interesting consequences. First, we now have a generative model-learning interpretation of frequent itemsets discovery. This aspect is described in Sec. 4. The second important consequence of our theoretical connection is the statistical significance test for itemsets discovered in the data. This is described next in Sec. 5.

4 Generative model-learning interpretation

Generative model-learning and pattern discovery are two broad categories of techniques in data mining [7]. Patterns (e.g. frequent itemsets) describe local structures in the data and make statements about a small number of variables or data points. Generative models, on the other hand, tend to characterize global structures that are applicable over the entire database. In this paper, we have connected a class of patterns (namely frequent itemsets) with a class of generative models (namely, the Itemset Generating Models). Each IGM generates transactions by embedding its corresponding itemset with a certain probability. *Theorem 4* states that (under reasonable conditions), if α_1 is the most frequent N-itemset, then Λ_{α_1} is a maximum likelihood estimate for the data (over $\mathcal{I}$). A typical transaction database would contain several transactions, where different transactions may be generated by different IGMs. Λ_{α_1}, however, regards all items other than those in α_1 as noise. So we remove α_1 from all transactions in the data that it appears in, and ob-

tain the next most frequent itemset, say α_2, in the data. The corresponding IGM, Λ_{α_2}, will now be an MLE for this reduced part of the data. Once again, Λ_{α_2} discards all items outside of α_2 as noise. So, we next remove α_2 from the data, obtain the next most frequent itemset, α_3, and so on. In this manner, we can interpret frequent itemsets mining as a generative model-learning exercise.

There is another interesting aspect to describing the data using such generative models. Based on the itemsets-IGM connections, it is possible to learn a mixture of IGMs for the data. This will allow use of frequent itemsets mining for problems involving classification and clustering within a standard Bayes-theoretic framework. More work is needed to resolve the details of these mixture models. However, one can at least see that in principle our theoretical connections can facilitate such modeling.

5 Statistical significance of frequent itemsets

There are many ways to define and measure significance of itemsets and/or association rules. For example, in [13], a chi-squared test of independence is used to assess correlations among items, and, an itemset is marked significant, if it meets both a minimum support criterion and a minimum correlation criterion. In [10], similar chi-squared tests are used for pruning and summarizing associations. Since chi-squared tests typically rely on the normal approximation to the binomial, in some situations, e.g. when the transactions data is sparse, this approximation breaks down and the tests become ineffective. In the context of text analysis, alternatives based on likelihood ratio tests have been used [6] to detect composite terms, domain-specific terms, etc. In [5], a baseline frequency is defined for each itemset and the ratio of actual-to-baseline frequencies is proposed as a measure of significance. Bayes estimation (using a Poisson model for the frequency count) is then used to reliably estimate this ratio from samples of small sizes. For rules with quantitative consequents, resampling-based permutation tests have been used to deduce confidence intervals [15]. In [14], union-type bounds are used to estimate probability of an itemset under a random Bernoulli model. These bounds, however, require both alphabet-size and itemset-size to be sufficiently large, and this somewhat limits applicability of the test. We adopt a different approach to assessing significance of itemsets based on generative models for transactions in the form of IGMs.

To assess statistical significance of an itemset, we directly use the itemset-IGM connection and its properties (derived in Sec. 3.3). The significance test of an itemset compares data likelihood of the associated IGM against likelihood under a random uniformly distributed model. Setting this up in a hypothesis testing framework, yields for a given level of the test (i.e. for a given probability of Type I error), a frequency threshold for frequent itemset discovery. For lower probabilities of error, the test prescribes higher frequency thresholds. The tight connections between itemsets and IGMs, based on the theorems of Sec. 3, show that IGMs are a good class of models for embedding patterns in transactions. This is what also makes our significance test reasonable, since the test tries to reject the hypothesis of a random *iid* model on the evidence of *at least one model* better than random chance.

Let $\mathcal{D}$ be a database of K transactions (over alphabet $\mathcal{A}$ of size M). Let $\alpha = (A_1, \ldots, A_N)$ be an N-itemset that occurs in $\mathcal{D}$ with frequency $f_\alpha (> 0)$. The IGM associated with α, according to *Definition 4*, is denoted by $\Lambda_\alpha = (\alpha, \theta_\alpha)$, with $\theta_\alpha = (\frac{f_\alpha}{K})$ only if it is greater than $(\frac{1}{2^N})$; otherwise, $\theta_\alpha = 0$. We set up a likelihood ratio test to compare an (alternate) hypothesis H_α: $\mathcal{D}$ is generated by the IGM Λ_α, against the (null) hypothesis H_0: $\mathcal{D}$ is generated by a uniformly distributed random *iid* model. Under H_0, data likelihood of $\mathcal{D}$ is $(\frac{1}{2})^{KM}$. The likelihood ratio test rejects the null hypothesis H_0 if

$$L(\mathcal{D}) = \frac{P[\mathcal{D} \mid H_\alpha]}{P[\mathcal{D} \mid H_0]} > \gamma \tag{12}$$

where, $\gamma > 0$ is a positive threshold obtained by fixing the probability of Type I error (i.e. the probability of wrong rejection of null hypothesis). In (12), $P[\mathcal{D}|H_\alpha]$ and $P[\mathcal{D}|H_0]$ denote the likelihoods under alternate and null hypotheses respectively, and $L(\mathcal{D})$ is the *likelihood ratio* for $\mathcal{D}$.

Let us first consider the case when *Definition 4* sets $\theta_\alpha = 0$. From Eq. (4), $P[T \mid \Lambda_\alpha] = 0$ for any transaction $T \in \mathcal{D}$ that contains α. Thus, $P[\mathcal{D} \mid H_\alpha] = 0$ if there is even one transaction in $\mathcal{D}$ that contains α. Since we need a significance test only for itemsets that occur (at least once), whenever *Definition 4* sets $\theta_\alpha = 0$, the corresponding likelihood (under the alternate hypothesis) is given by $P[\mathcal{D} \mid H_\alpha] = 0$. In such a case, we clearly cannot reject the null hypothesis. So we only need to consider the case when *Definition 4* sets $\theta_\alpha > (\frac{1}{2^N})$.

When $\theta_\alpha > (\frac{1}{2^N})$, from Eq. (6) and (12), the expression for the likelihood ratio, $L(\mathcal{D})$, can be written as

$$L(\mathcal{D}) = \theta_\alpha^{f_\alpha} \left(\frac{1-\theta_\alpha}{2^N - 1}\right)^{K - f_\alpha} \left(\frac{1}{2}\right)^{-KN} \tag{13}$$

With $\theta_\alpha > (\frac{1}{2^N})$, the likelihood ratio strictly increases with itemset frequency. This can be seen using inequalities (10) and (11), in the proof of *Theorem 3*. Monotonicity of $L(\mathcal{D})$ with frequency of itemsets, allows use of an equivalent test, with $L_1(\mathcal{D}) = f_\alpha$ as the new test statistic:

$$\text{if } L_1(\mathcal{D}) = f_\alpha \ > \ \Gamma, \text{ reject } H_0. \tag{14}$$

Threshold Γ for the above test is computed by fixing the allowed probability, P_{FA} of Type I error:

$$P_{FA} = P[L_1(\mathcal{D}) > \Gamma \mid H_0] \tag{15}$$

Let z_i, $i = 1, \ldots, K$, be 0-1 random variables that each indicate whether or not α appears in transaction T_i. We have $L_1(\mathcal{D}) = f_\alpha = \sum_{i=1}^{K} z_i$. Under the null hypothesis H_0, z_i's are *iid* with mean and variance as follows:

$$\mu = E_{H_0}[z_i] = \left(\frac{1}{2}\right)^N \quad (16)$$

$$\sigma^2 = E_{H_0}[(z_i - \mu)^2] = \mu(1-\mu) \quad (17)$$

For large K, central limit theorem guarantees $(\frac{f_\alpha - K\mu}{\sigma\sqrt{K}})$ achieves the standard normal distribution. Probability of Type I error (cf. Eq. (15)) is now computed using:

$$P_{FA} = 1 - \Phi\left(\frac{\Gamma - K\mu}{\sigma\sqrt{K}}\right) \quad (18)$$

where, $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal random variable.

The procedure for assessing statistical significance of an N-itemset, α, is as follows: Choose the allowed level, say ϵ, of Type I error. Since K and N are known, compute Γ using Eq. (18) and the standard normal tables:

$$\Gamma = \frac{K}{2^N} + \sqrt{\left(\frac{K}{2^N}\right)\left(1 - \frac{1}{2^N}\right)}\Phi^{-1}(1-\epsilon) \quad (19)$$

Then, reject the null hypothesis, H_0 (i.e. declare α as significant), if $f_\alpha > \Gamma$.

There is no need to explicitly check whether $\theta_\alpha > (\frac{1}{2^N})$, before applying the test of significance for an N-itemset α. This is because, from Eq. (19), for $\epsilon = 0.5$, we obtain $\Gamma = (\frac{K}{2^N})$, and Γ increases for lower values of ϵ. Hence, provided that we choose a level of the test less than 0.5 (which is very reasonable), the test automatically demands a frequency greater than $\frac{K}{2^N}$ (and this corresponds to the region, $\theta_\alpha > \frac{1}{2^N}$, in which all our theorems operate).

Note that we now have a size-dependent frequency threshold for itemsets (with smaller itemsets requiring larger frequencies to be regarded as significant). Also note that, for typical values of N and K (i.e. with $K \gg N$), Γ changes very little over $0 < \epsilon \leq 0.5$. This is because, $(\frac{K}{2^N})$ dominates the expression for Γ in Eq. (19). This allows for automatically fixing the threshold for N-itemsets at $(\frac{K}{2^N})$. Sometimes, we may want to use higher thresholds (corresponding to very small ϵ values), but in the absence of any information about the data, $(\frac{K}{2^N})$ is a very good initial threshold to try. Simulation experiments on real and synthetic data (cf. Sec. 6) show that these thresholds are effective in detecting the patterns embedded in the data.

6 Experimental results

This section presents results obtained on some publicly available benchmark data sets as well as on proprietary data obtained from attack graphs produced by the Netra [12] configuration analysis tool. The benchmark data sets include synthetic data from IBM Almaden Quest group's data generator and some UCI data sets (prepared for frequent itemsets mining by Roberto Bayardo)[1]. The goal of our experiments is to illustrate utility of our theoretical connections in the frequent itemsets mining process. To this end, synthetic data from IBM Quest is particularly useful, since we can vary data generation parameters and assess performance by checking if the patterns embedded by the data generator are indeed discovered during the mining process. In case of other data sets, since there can be no ground truth of "embedded patterns", we present summary results to show the relevance of our models. In our experiments, we used the Apriori algorithm for frequent itemsets mining (However, note that our analysis using IGMs is independent of the algorithm used. The theoretical connections and significance tests are relevant for any algorithm that discovers frequent itemsets in the data).

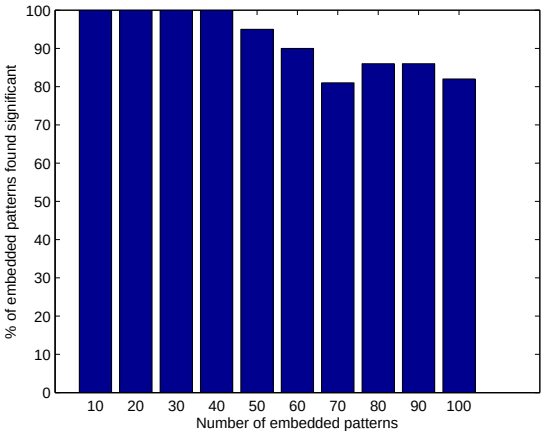

Figure 1. Percentage of embedded patterns found significant in synthetic data with varying numbers of embedded patterns. Number of transactions is 100K, average transaction length is 20, number of items is 1K and average size of embedded patterns is 7. Number of embedded patterns is varied from 10 to 100.

The first experiment illustrates effectiveness of our theoretically motivated frequency thresholds. We run the frequent itemsets mining algorithm on several synthetic data sets (generated using the IBM Quest data generator for dif-

[1]The IBM synthetic data generator was obtained from http://www.almaden.ibm.com/cs/projects/iis. The two UCI data sets were obtained from the Frequent Itemsets Mining Implementations Repository at http://fimi.cs.helsinki.fi/data.

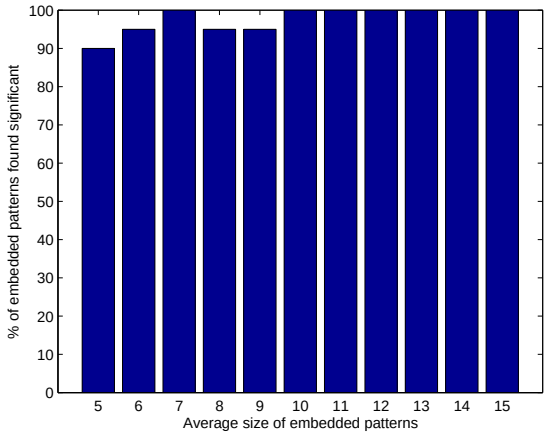

Figure 2. Percentage of embedded patterns found significant in synthetic data with varying sizes of embedded patterns. Number of transactions is 100K, average transaction length is 20, number of items is 1K and number of embedded patterns is 25. Average size of embedded patterns is varied from 5 to 15.

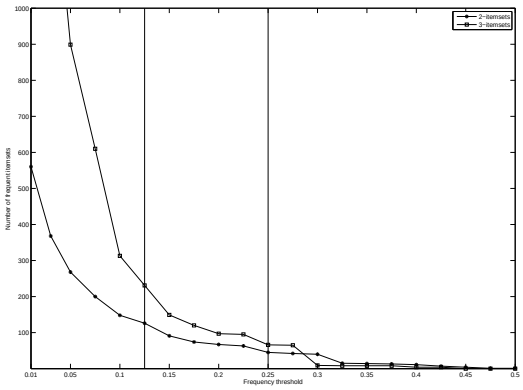

Figure 3. Number of frequent 2 & 3-itemsets versus frequency threshold in *IBM Quest* synthetic data (Number of transactions is 100K, average transaction length is 15, number of items is 1K, average size of embedded patterns is 10 and number of embedded patterns is 10). The vertical lines indicate theoretical frequency thresholds, namely, $(\frac{1}{2^2})$ for 2-itemsets & $(\frac{1}{2^3})$ for 3-itemsets. The graph for 3-itemsets meets Y-axis at 3197.

ferent parameter values). Each data set is associated with a collection of patterns that the generator used as potential "correlations to embed" during the data generation process (Here, we refer to these correlations as *embedded patterns*). We now ask whether these patterns come up as *significant patterns* according to our significance test of Sec. 5. We consider an N-itemset as significant (in a data set of K transactions) if its frequency exceeds $(\frac{K}{2^N})$.

Fig. 1 plots the percentage of embedded patterns found significant as a function of the number of embedded patterns in different synthetic data sets. Each data set contains 100K transactions, with an average transaction length of 20 over an alphabet of 1000 items, with average length of embedded patterns set to 7. To study the effect of number of embedded patterns, we generate several synthetic data sets by varying the number of embedded patterns between 10 and 100. The plot shows that a very high percentage of embedded patterns is always found significant by our analysis. For smaller numbers of embedded patterns (less than 50) all (i.e. 100% of) embedded patterns are significant. As the number of embedded patterns approaches 100 this percentages decreases to 82%. This is natural, since, when the number of embedded patterns increases, on the average, the frequencies of individual patterns have to decrease. The mixing proportions that IBM Quest data generator assigns to patterns are drawn iid according to an exponential distribution with the same parameter for all sizes of patterns [1]. Thus, small-size patterns associated with low mixing proportions are the ones most susceptible to being missed by our significance test. For example, in the experiment of Fig. 1, if we ignored 2-size patterns, the percentage of embedded patterns passing our significance test was consistently above 95%. Fig. 2 shows a similar plot obtained when varying the average size of embedded patterns. Once again, we note that a very high percentage of embedded patterns are consistently discovered as significant. Similar results were obtained when we studied the effect of varying number of transactions, average length of transactions, and number of items as well.

Note that the IBM synthetic data generation model [1] is quite different from our IGMs. Despite this, our significance test based on IGMs is able to reliably detect the patterns embedded in the synthetic data. This, in a sense, lends empirical strength to our choice of significance test for itemsets (where an alternate hypothesis of an IGM generating the data is opposed to the null hypothesis of a random uniform *iid* model).

In a second experiment, we varied frequency threshold over a range, and, for each case, recorded the number of frequent N-itemsets discovered, for different sizes, N, of itemsets. Imagine a data set in which no specific patterns are embedded (or equivalently, a data set in which all patterns are equally likely to appear in the transactions). In such a case, all 2-size (or 3-size) patterns would have very

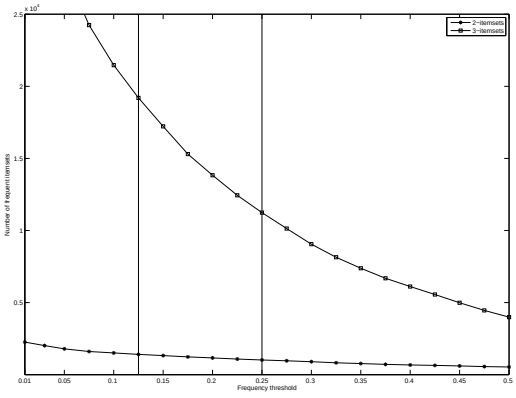

Figure 4. Number of frequent 2 & 3-itemsets versus frequency threshold in *Chess* database. The vertical lines indicate theoretical frequency thresholds, namely, $(\frac{1}{2^2})$ for 2-itemsets & $(\frac{1}{2^3})$ for 3-itemsets. The graph for 3-itemsets meets Y-axis at 42212.

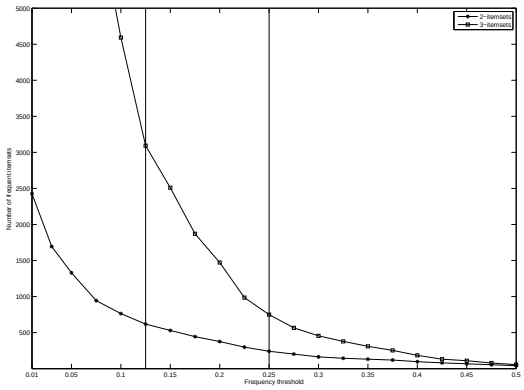

Figure 5. Number of frequent 2 & 3-itemsets versus frequency threshold in *Mushroom* database. The vertical lines indicate theoretical frequency thresholds, namely, $(\frac{1}{2^2})$ for 2-itemsets & $(\frac{1}{2^3})$ for 3-itemsets. The graph for 3-itemsets meets Y-axis at 28416.

low (and also very similar) frequencies. Thus, very low frequency thresholds would lead to a combinatorial explosion of frequent itemsets. This would continue to be the case even as the threshold is increased until a critical point, beyond which, no pattern would be frequent (since the data is inherently random). Even for data sets in which specific patterns are embedded, the behavior would be very similar for low thresholds, but as the threshold is increased, only the specific patterns embedded in the data would remain frequent. So, a graph of the number of frequent itemsets versus the frequency threshold would "level-off" for larger frequency thresholds.

Figs. 3-6 plot the number of frequent 2 & 3-itemsets discovered under various frequency thresholds for 4 data sets: one synthetic data using the IBM Quest data generator (with 100K transactions, average transaction length of 15, average pattern length of 10, 1000 items and 10 embedded patterns), chess and mushroom UCI data sets, and one proprietary data set from Netra. Theoretical thresholds are indicated by vertical lines. The graphs show that for very low threshold values, almost all possible frequent 2 and 3-itemsets become frequent. We note that this number is different for different data sets, since the number of frequent items (or 1-itemsets) is different in each case. As the threshold increases, the number of frequent 2-itemsets and 3-itemsets decrease, and the plots tend to level-off beyond the theoretical thresholds. In case of the IBM data sets, similar plots were obtained when the parameters of data generation were varied as earlier. We also expect to obtain similar graphs for itemsets of bigger sizes as well (Note that it is difficult to carry out this experiment for larger sizes of itemsets due to an overwhelming number of candidates at lower thresholds, although earlier experiments on IBM data sets showed that our own thresholds always detect larger size patterns embedded in the data without any problem).

6.1 Discussion

In general, setting the frequency threshold for frequent itemsets mining can be a tricky exercise. The plots in Figs. 3-6 bear evidence to the fact that, while a low threshold might lead to a combinatorial explosion of the number of patterns, a higher threshold might miss significant patterns present in the data. Our significance test based on the itemsets-IGM connections help alleviate this problem. The key feature of our significance test is the size-dependent frequency thresholds for itemsets. If we want to discover frequent itemsets up to size N (in data with K transactions) we use $(\frac{K}{2^N})$ as the threshold for the mining process. However, an l-itemset (with $l < N$) discovered as frequent according to the $(\frac{K}{2^N})$ threshold is *not* considered interesting unless its frequency exceeds $(\frac{K}{2^l})$. Our experiments show that, using these theoretically motivated thresholds, we can expect to avoid detecting most of the noisy patterns and identify only those correlations that are stronger than random chance.

7 Conclusions

In this paper, we presented a theoretical connection between the process of frequent itemset discovery and learn-

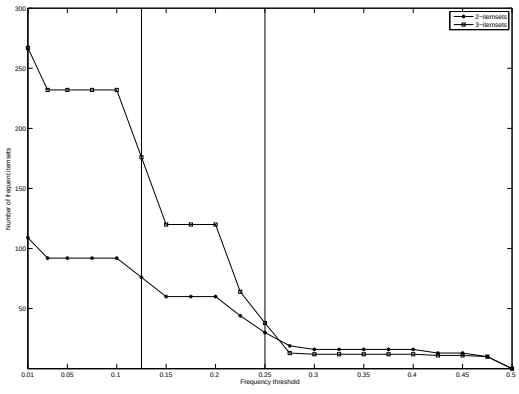

Figure 6. Number of frequent 2 & 3-itemsets versus frequency threshold in *Netra* data. The vertical lines indicate theoretical frequency thresholds, namely, $(\frac{1}{2^2})$ for 2-itemsets & $(\frac{1}{2^3})$ for 3-itemsets.

ing of generative models. We proposed a class of generative models called Itemset Generating Models (or IGMs), and associated each itemset with a unique IGM from this class. We established a connection between data likelihood under the IGM and frequency of the associated itemset (obtained using any standard itemsets mining algorithm) and showed that frequency ordering among itemsets (of a given size) is preserved as likelihood ordering among the associated IGMs. Under reasonable conditions, frequency of an itemset is a sufficient statistic for maximum likelihood estimation over the class of IGMs. This gives us a generative model-learning interpretation of frequent itemsets mining. Another advantage of our itemset-IGM connections is the significance test for itemsets that uses just the frequencies of itemsets as test statistics. The test leads to size-dependent frequency thresholds for itemsets, with smaller itemsets requiring greater frequencies to be considered significant. We experimentally validated our analysis both on benchmark as well as proprietary data sets.

There are many other aspects of the itemsets-IGM connection that would be worthwhile investigating further. One is the learning of a mixture model for the data in the form of a convex combination of IGMs. Such a model would open up possibilities for use of frequent itemsets in classification and clustering applications within the standard Bayes decision theory framework. Another interesting possibility is to modify the significance test to incorporate the simultaneous assessment of significance of (not one but) a set of itemsets. This can lead to more reliable detection of significant patterns at considerably lower frequencies. We will address some of these aspects in our future work.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pages 207–216, May 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.

[3] C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, Jan. 2003.

[4] J. Dong, W. Perrizo, Q. Ding, and J. Zhou. The application of association rule mining to remotely sensed data. In *SAC (1)*, pages 340–345, 2000.

[5] W. DuMouchel and D. Pregibon. Empirical bayes screening for multi-item associations. In *Knowledge Discovery and Data Mining*, pages 67–76, 2001.

[6] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1994.

[7] D. Hand, H. Mannila, and P. Smyth. *Principles of data mining*. MIT Press, Cambridge, MA, USA, 2001.

[8] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for association rule mining — a general survey and comparison. *SIGKDD Explorations*, 2(1):58–64, July 2000.

[9] S. Laxman, P. S. Sastry, and K. P. Unnikrishnan. Discovering frequent episodes and learning Hidden Markov Models: A formal connection. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1505–1517, Nov. 2005.

[10] B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Knowledge Discovery and Data Mining*, pages 125–134, 1999.

[11] H. Mannila. Theoretical frameworks for data mining. *SIGKDD Explor. Newsl.*, 1(2):30–32, 2000.

[12] P. Naldurg, S. Schwoon, S. Rajamani, and J. Lambert. Netra: Seeing through access control. In *ACM Workshop on Formal Methods for Security Engineering FMSE 2006*, Virginia, USA, 2006.

[13] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining Knowledge Discovery*, 2(1):39–68, 1998.

[14] X. Sun and A. B. Nobel. Significance and recovery of block structures in binary matrices with noise. Technical report, Department of Statistics and Operations Research, UNC Chapel Hill, 2005.

[15] H. Zhang, B. Padmanabhan, and A. Tuzhilin. On the discovery of significant statistical quantitative rules. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, Seattle, WA*, pages 374–383, New York, NY, USA, 2004. ACM Press.