Comparing Datasets Using Frequent Itemsets: Dependency on the Mining Parameters

Irene Ntoutsi and Yannis Theodoridis

Department of Informatics, University of Piraeus, Greece
{ntoutsi,ytheod}@unipi.gr

Abstract. Comparison between sets of frequent itemsets has been traditionally utilized for raw dataset comparison assuming that frequent itemsets inherit the information lying in the original raw datasets. In this work, we revisit this assumption and examine whether dissimilarity between sets of frequent itemsets could serve as a measure of dissimilarity between raw datasets. In particular, we investigate how the dissimilarity between two sets of frequent itemsets is affected by the *minSupport* threshold used for their generation and the adopted compactness level of the itemsets lattice, namely frequent itemsets, closed frequent itemsets or maximal frequent itemsets. Our analysis shows that utilizing frequent itemsets comparison for dataset comparison is not as straightforward as related work has argued, a result which is verified through an experimental study and opens issues for further research in the KDD field.

1 Introduction

Detecting changes between datasets is an important problem nowadays due to the highly dynamic nature of data. A common approach for comparing datasets is to utilize the patterns extracted from these datasets. The intuition behind this approach is that patterns encapsulate (to some degree) the information contained in the original data. In [3], authors measure the deviation between systematically evolving datasets, like the buying habits of customers in a supermarket, in terms of the data mining models they induce. In [5,4], authors measure the dissimilarity between distributed datasets using the corresponding sets of frequent itemsets.

In this work we elaborate on this assumption about the equivalence of dissimilarity in pattern space with the dissimilarity in raw data space, for a very popular pattern type, the frequent itemset patterns. More specifically, we provide a theoretical analysis that shows the dependency of dissimilarity in pattern space on frequent itemsets mining (FIM) settings, namely (a) on the *minSupport* threshold used for the generation of itemsets and (b) on the adopted compactness level for the itemsets lattice (frequent itemsets-FI, closed frequent itemsets-CFI or maximal frequent itemsets-MFI). Regarding the *minSupport* threshold, our analysis shows that the larger this threshold is, the higher the dissimilarity in pattern space is. Regarding the different lattice representations, it turns out that the more compact the representation achieved by the itemset type is, the higher

J. Darzentas et al. (Eds.): SETN 2008, LNAI 5138, pp. 212–225, 2008.

the dissimilarity in pattern space is. Moreover, we describe the different dissimilarity measures proposed so far in the literature ([3,4,5]) through a general common dissimilarity schema and verify the above theoretical results through an experimental study. The results indicate that utilizing pattern comparison for data comparison is not as straightforward as argued by related work and should only be carried out under certain assumptions (e.g., FIM settings).

The rest of the paper is organized as follows: Section 2 describes basic FIM concepts. Section 3 discusses the related work and describes a general formula through which the proposed measures can be described. In Section 4, we present the FIM parameters that affect dissimilarity. In Section 5, we experimentally evaluate the effect of the different parameters on dissimilarity. In Section 6, we present conclusions and outlook.

2 Background on the FIM Problem

Let I be a finite set of distinct items and D be a dataset of transactions where each transaction T contains a set of items, $T \subseteq I$. An *itemset* X is a non-empty lexicographically ordered set of items, $X \subseteq I$. The percentage of transactions in D that contain X, is called the *support* of X in D, $supp_D(X)$. An itemset is *frequent* if $supp_D(X) \geq \sigma$, where σ is the *minSupport* threshold. The set of frequent itemsets (FIs) extracted from D under σ is given by: $F_\sigma(D) = \{X \subseteq I \mid supp_D(X) \geq \sigma\}$. An itemset is frequent iff all of its subsets are frequent (apriori property). This property allows us to enumerate frequent itemsets lattice using more compact representations like closed frequent itemsets (CFIs) and maximal frequent itemsets (MFIs).

A frequent itemset X is called *closed* if there exists no frequent superset $Y \supset X$ with $supp_D(X) = supp_D(Y)$. The set $C_\sigma(D)$ is a subset of $F_\sigma(D)$ since every closed itemset is also frequent.

$$C_\sigma(D) = F_\sigma(D) - \{X \in F_\sigma(D) : Y \supset X \Rightarrow supp_D(X) = supp_D(Y), Y \in F_\sigma(D)\} \tag{1}$$

On the other hand, a frequent itemset is called *maximal* if it is not a subset of any other frequent itemset. The set $M_\sigma(D)$ is also a subset of $F_\sigma(D)$ since every maximal itemset is frequent.

$$M_\sigma(D) = F_\sigma(D) - \{X \in F_\sigma(D) : Y \supset X \Rightarrow Y \in F_\sigma(D)\} \tag{2}$$

By definition, $C_\sigma(D)$ is a subset of $M_\sigma(D)$:

$$M_\sigma(D) = C_\sigma(D) - \{X \in C_\sigma(D) : Y \supset X \Rightarrow Y \in C_\sigma(D)\} \tag{3}$$

$C_\sigma(D)$ is a lossless representation of $F_\sigma(D)$ since both the lattice structure (i.e., frequent itemsets) and the lattice measure (i.e., itemset supports) can be derived from CFIs [7]. Unlike $C_\sigma(D)$, $M_\sigma(D)$ is a lossy representation of $F_\sigma(D)$ since it is only the lattice structure that can be determined from MFIs whereas the measures are lost [7]. By Eq. 1, 2 and 3 it follows that:

$$M_\sigma(D) \subseteq C_\sigma(D) \subseteq F_\sigma(D) \tag{4}$$

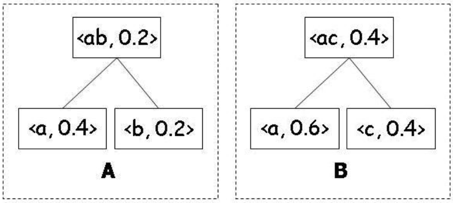

Fig. 1. Two lattices of frequent itemsets to be compared: A (lef), B (right)

Table 1. List of symbols

Symbol	Description
D	a dataset
X	an itemset
$supp_D(X)$	the support of itemset X in dataset D
σ	the *minSupport* threshold
$F_\sigma(D)$	the set of frequent itemsets generated from D under σ
$C_\sigma(D)$	the set of closed frequent itemsets generated from D under σ
$M_\sigma(D)$	the set of maximal frequent itemsets generated from D under σ
$dis(A, B)$	dissimilarity between two set of itemsets A, B

Let us consider for comparison (Fig. 1) two sets of frequent itemsets A, B generated under the same *minSupport* threshold σ from the original datasets D and E, respectively. Each itemset is described as a pair <*structure, measure*> denoting the items forming the itemset (*structure*) and the itemset support (*measure*).

The question is how similar to each other A and B are. There are many cases where the two sets might differ: An itemset, for example, might appear in both A and B sets but with different supports, like the itemset $< a >$ in Fig. 1. Or, an itemset might appear in only one of the two sets, like the itemset $< b >$ which appears in A but not in B. In this case, two things might have occurred: either $< b >$ does not actually exist in the corresponding dataset E or $< b >$ has been pruned due to low support (lower than the *minSupport* threshold σ).

Since the generation of A, B depends on the FIM parameters, namely the *minSupport* threshold σ used for their generation and the adopted lattice representation (FI, CFI or MFI), we argue that the estimated dissimilarity score also depends on these parameters. Furthermore, since dissimilarity in pattern space is often used as a measure of dissimilarity in raw data space we argue that the above mentioned parameters also affect this correspondence.

Table 1 summarizes the symbols introduced in this section.

3 Comparing Frequent Itemset Lattices

Parthasarathy-Ogihara approach: *Parthasarathy and Ogihara* [5] present a method for measuring the dissimilarity between two datasets D and E by using the corresponding sets of frequent itemsets (A and B, respectively). Their metric is defined as follows:

$$dis(A,B) = 1 - \frac{\sum_{X \in A \cap B} \max\{0, 1 - \theta * |supp_D(X) - supp_E(X)|\}}{|A \cup B|} \tag{5}$$

In the above equation, θ is a scaling parameter that reflects how significant are for the user the variations in the support values. This measure works with itemsets of identical structure, i.e., those appearing in $A \cap B$. Itemsets that only partially fit each other like $< ab >$ and $< ac >$ are considered totally dissimilar.

FOCUS approach: The *FOCUS framework* [3] quantifies the deviation between two datasets D and E in terms of the FI sets (A and B, respectively) they induce. A and B are first refined into their union ($A \cup B$) and the support of each itemset is computed with respect to both D and E datasets. Next, the deviation is computed by summing up the deviations of the frequent itemsets in the union:

$$dis(A,B) = \frac{\sum_{X \in A \cup B} |supp_D(X) - supp_E(X)|}{\sum_{X \in A} supp_D(X) + \sum_{X \in B} supp_E(X)} \tag{6}$$

FOCUS measures the dissimilarity between two sets of FIs in terms of their union. Partial similarity is not considered. Indeed, FOCUS tries to find itemsets with identical structures. If an itemset X appears in A with $supp_D(X)$ but not in B, then the corresponding data set E of B is queried so as to retrieve $supp_E(X)$. An upper bound on dissimilarity is provided, which involves only the induced models and avoids the expensive operation of querying the original raw data space. In this case, if an itemset X does not appear in B, it is considered to appear but with zero measure, i.e., $supp_E(X) = 0$.

Li-Ogihara-Zhou approach: *Li Ogihara and Zhou* [4] propose a dissimilarity measure between datasets based on the set of MFIs extracted from these datasets. Let $A = \{X_i, supp_D(X_i)\}$ and $B = \{Y_j, supp_E(Y_j)\}$ where X_i, Y_j are the MFIs in D, E respectively. Then:

$$dis(A,B) = 1 - \frac{2I_3}{I_1 + I_2} \tag{7}$$

$$I_1 = \sum_{X_i, X_j \in A} d(X_i, X_j), \quad I_2 = \sum_{Y_i, Y_j \in B} d(Y_i, Y_j), \quad I_3 = \sum_{X \in A, Y \in B} d(X, Y)$$

$$d(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} * log(1 + \frac{|X \cap Y|}{|X \cup Y|}) * \min(supp_D(X), supp_E(Y))$$

This measure works with the average dissimilarity between pairs of MFIs. Partial similarity is considered; itemsets that have some common part in their structure are compared and their score is aggregated to the total dissimilarity score.

3.1 Common Background of the Three Approaches

All approaches express the dissimilarity between two sets of frequent itemsets as an aggregation of the dissimilarities of their component itemsets:

$$dis(A,B) = \sum_{X \in A, Y \in B} dis(X,Y) \tag{8}$$

where $dis(X,Y)$ is the dissimilarity function between two simple frequent itemsets, defined in terms of their structure and measure components, as follows:

$$dis(X,Y) = \mathsf{f}(dis_{struct}(X,Y), dis_{meas}(X,Y)) \tag{9}$$

The function $dis_{struct}()$ evaluates the dissimilarity between the structure components (i.e., frequent itemsets), whereas the function $dis_{meas}()$ evaluates the dissimilarity between their measure components (i.e., supports). The function f aggregates the corresponding structure and measure scores into a total score.

All approaches follow the rationale of Eq. 8 and differentiate only on how $dis_{struct}(X,Y)$, $dis_{meas}(X,Y)$ and f are instantiated. Below, we present how these functions are defined for each of the proposed measures.

In case of the Parthasarathy-Ogihara measure, Eq. 9 can be written as:

$$dis(X,Y) = \max\{0, 1 - dis_{struct}(X,Y) - \theta * dis_{meas}(X,Y)\}$$

$$dis_{struct}(X,Y) = \begin{cases} 0 \text{ , if } X = Y \\ 1 \text{ , otherwise} \end{cases}$$

$$dis_{meas}(X,Y) = \begin{cases} |supp_D(X) - supp_E(Y)| & \text{, if } X = Y \\ 0 & \text{, otherwise} \end{cases}$$

For the FOCUS approach, Eq. 9 is initialized as:

$$dis(X,Y) = (1 - dis_{struct}(X,Y)) * dis_{meas}(X,Y)$$

$$dis_{struct}(X,Y) = \begin{cases} 0 \text{ , if } X = Y \\ 1 \text{ , otherwise} \end{cases}$$

$$dis_{meas}(X,Y) = \begin{cases} |supp_D(X) - supp_E(Y)| & \text{, if } X, Y \in A \cap B \ and \ X = Y \\ supp_D(X) & \text{, if } X \in A - B \\ supp_E(Y) & \text{, if } Y \in B - A \end{cases}$$

Finally, for the Li-Ogihara-Zhu approach, Eq. 9 becomes:

$$dis(X,Y) = dis_{struct}(X,Y) * \log(1 + dis_{struct}(X,Y)) * dis_{meas}(X,Y)$$

$$dis_{struct}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

$$dis_{meas}(X,Y) = \min\{supp_D(X), supp_E(Y)\}$$

4 Effect of Mining Parameters on Dissimilarity

In the following subsections, we investigate how the dissimilarity between two sets of frequent itemsets depends on the *minSupport* threshold σ used for their generation and on the adopted lattice representation (FI, CFI or MFI).

4.1 Effect of *minSupport* Threshold on Dissimilarity

Let σ, $\sigma+\delta$ ($0 < \sigma, \delta < \sigma+\delta \leq 1$) be two *minSupport* thresholds applied over a dataset D and let F_σ, $F_{\sigma+\delta}$ be the corresponding sets of frequent itemsets produced by (any) FIM algorithm. The difference set $F_\sigma - F_{\sigma+\delta}$ contains all those itemsets whose support lies between σ and $\sigma+\delta$:

$$Z \equiv F_\sigma - F_{\sigma+\delta} = \{X \subseteq I \mid \sigma < supp(X) \leq \sigma+\delta\} \tag{10}$$

In Fig. 2, an example is depicted which illustrates how the resulting lattice is affected by the increase δ in the *minSupport* threshold. As it is shown in this figure, with the increase of δ, the lattice is reduced.Below, we describe how each of the presented measures is affected by the increase δ in the *minSupport* value.

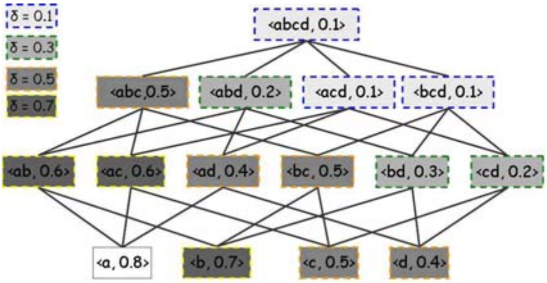

Fig. 2. Effect of δ increase on the lattice structure ($\sigma = 0.1$)

Parthasarathy-Ogihara [5] approach : From Eq. 5, it holds that:

$$dis(F_\sigma, F_{\sigma+\delta}) = 1 - \frac{\sum_{X \in F_\sigma \cap F_{\sigma+\delta}} \max\{0, 1 - \theta * |supp_D(X) - supp_D(X)|\}}{|F_\sigma \cup F_{\sigma+\delta}|}$$
$$= 1 - \frac{\sum_{X \in F_{\sigma+\delta}} \max\{0, 1-0\}}{|F_\sigma|}$$

$$\Rightarrow dis(F_\sigma, F_{\sigma+\delta}) = 1 - \frac{|F_{\sigma+\delta}|}{|F_\sigma|} \tag{11}$$

From the above equation, we can conclude that the greater the increase in the *minSupport* threshold value δ is, the smaller the enumerator $|F_{\sigma+\delta}|$ will be (cf. Eq. 10) and thus the greater the distance between the two sets will be.

FOCUS [3] approach: Recalling Eq. 6 and Eq. 10, it holds that[1]:

$$dis(F_\sigma, F_{\sigma+\delta}) = \frac{\sum_{X \in F_\sigma \cup F_{\sigma+\delta}} |supp_D(X) - supp_D(X)|}{\sum_{X \in F_\sigma} supp_D(X) + \sum_{X \in F_{\sigma+\delta}} supp_D(X)}$$
$$= \frac{\sum_{X:\sigma < supp_D(X) \leq \sigma+\delta} supp_D(X)}{2 * \sum_{X \in F_\sigma} supp_D(X) - \sum_{X:\sigma < supp_D(X) \leq \sigma+\delta} supp_D(X)} \quad (12)$$

For simplicity, let $C = \sum_{X:\sigma < supp_D(X) \leq \sigma+\delta} supp_D(X)$.

$$\Rightarrow dis(F_\sigma, F_{\sigma+\delta}) = \frac{C}{2 * \sum_{X \in F_\sigma} supp_D(X) - C} \quad (13)$$

In the above equation, if the value of δ increases, the numerator C will also increase whereas the denumerator will decrease (cf. Eq. 10 as well). Thus, as δ increases, the dissimilarity also increases.

Li-Ogihara-Zhou [4]approach: From Eq. 7 and Eq. 10, it holds that:

$$I_1 + I_2 = \sum_{X,Y \in F_\sigma} d(X,Y) + \sum_{X,Y \in F_{\sigma+\delta}} d(X,Y)$$
$$= 2 * \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{\substack{X:\sigma < supp(X) \leq \sigma+\delta \\ Y:\sigma < supp(Y) \leq \sigma+\delta}} d(X,Y)$$

$$I_3 = \sum_{\substack{X \in F_\sigma \\ Y \in F_{\sigma+\delta}}} d(X,Y) = \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{\substack{X:\sigma < supp(X) \leq \sigma+\delta \\ Y:\sigma < supp(Y) \leq \sigma+\delta}} d(X,Y)$$

For simplicity, let $G = \sum_{\substack{X:\sigma < supp(X) \leq \sigma+\delta \\ Y:\sigma < supp(X) \leq \sigma+\delta}} d(X,Y)$.

$$\Rightarrow dis(F_\sigma, F_{\sigma+\delta}) = 1 - \frac{2I_3}{I_1 + I_2} = 1 - \frac{2(I_1 - G)}{2I_1 - G} = \frac{G}{2I_1 - G} \quad (14)$$

As δ increases, the enumerator G also increases, whereas the denumerator $(2I_1 - G)$ decreases. Thus, dissimilarity increases as δ increases.

To summarize, Eq. 11, 13 and 14 state that, for all approaches, the larger the increase in the *minSupport* threshold value δ is, the larger the computed dissimilarity score, $dis(F_\sigma, F_{\sigma+\delta})$, will be.

[1] Some further explanations on the notation: the term $\sum_{X \in F_\sigma \cup F_{\sigma+\delta}} |supp_D(X) - supp_D(X)|$ corresponds to the sum of the supports of all those itemsets that appear in $(F_\sigma - F_{\sigma+\delta})$. As far, as this set is not empty, this term is > 0.

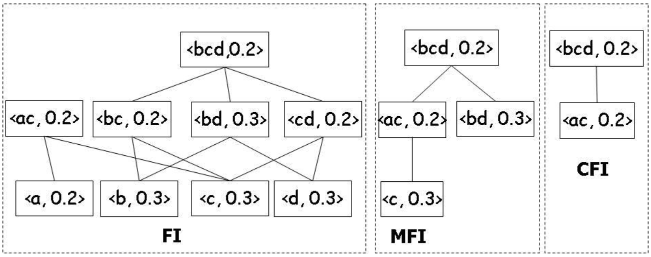

Fig. 3. Effect of representation (FI, CFI, MFI) on the lattice structure ($\sigma = 0.1$)

4.2 Effect of Lattice Representation on Dissimilarity

Let $F_\sigma(D)$, $C_\sigma(D)$, $M_\sigma(D)$ be the sets of FIs, CFIs and MFIs, respectively, extracted from D under (fixed) *minSupport* threshold σ. The example in Fig. 3 illustrates the effect of the different representations (FI, CFI, MFI) in the lattice. These figures affirm Eq. 4 which states that the greater the compactness level is, the "smaller" the resulting lattice will be.

Parthasarathy-Ogihara [5] approach: From Eq. 5, it holds that:

$$dis(F_\sigma, C_\sigma) = 1 - \frac{\sum_{X \in F_\sigma \cap C_\sigma} \max\{0, 1 - \theta * |supp_D(X) - supp_D(X)|\}}{|F_\sigma \cup C_\sigma|}$$
$$= 1 - \frac{\sum_{X \in C_\sigma} \max\{0, 1 - 0\}}{|F_\sigma|} = 1 - \frac{|C_\sigma|}{|F_\sigma|} \qquad (15)$$

$$dis(F_\sigma, M_\sigma) = 1 - \frac{\sum_{X \in F_\sigma \cap M_\sigma} \max\{0, 1 - \theta * |supp_D(X) - supp_D(X)|\}}{|F_\sigma \cup M_\sigma|}$$
$$= 1 - \frac{\sum_{X \in M_\sigma} \max\{0, 1 - 0\}}{|F_\sigma|} = 1 - \frac{|M_\sigma|}{|F_\sigma|} \qquad (16)$$

From Eq. 15, 16, it holds that:

$$dis(F_\sigma, C_\sigma) \leq dis(F_\sigma, M_\sigma) \qquad (17)$$

FOCUS [3] approach: Recalling Eq. 6, it holds that:

$$dis(F_\sigma, C_\sigma) = \frac{\sum_{X \in F_\sigma \cup C_\sigma} |supp_D(X) - supp_D(X)|}{\sum_{X \in F_\sigma} supp_D(X) + \sum_{X \in C_\sigma} supp_D(X)} \qquad (18)$$

$$(19)$$

$$= \frac{\sum_{X \in F_\sigma - C_\sigma} supp_D(X)}{2 * \sum_{X \in F_\sigma} supp_D(X) - \sum_{X \in F_\sigma - C_\sigma} supp_D(X)}$$

$$dis(F_\sigma, M_\sigma) = \frac{\sum_{X \in F_\sigma \cup M_\sigma} |supp_D(X) - supp_D(X)|}{\sum_{X \in F_\sigma} supp_D(X) + \sum_{X \in M_\sigma} supp_D(X)} \tag{20}$$

$$\tag{21}$$

$$= \frac{\sum_{X \in F_\sigma - M_\sigma} supp_D(X)}{2 * \sum_{X \in F_\sigma} supp_D(X) - \sum_{X \in F_\sigma - M_\sigma} supp_D(X)}$$

where $F_\sigma - C_\sigma$ is given by Eq. 1 and $F_\sigma - M_\sigma$ is given by Eq. 2.

From Eq. 18 and Eq. 20, for the FOCUS measure it holds that:

$$dis(F_\sigma, C_\sigma) \leq dis(F_\sigma, M_\sigma) \tag{22}$$

Li-Ogihara-Zhou [4]approach: From Eq. 7, it holds that:

$$I_1 + I_2 = \sum_{X,Y \in F_\sigma} d(X,Y) + \sum_{X,Y \in C_\sigma} d(X,Y)$$

$$= 2 * \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y) = 2 * I_1 - \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y)$$

$$I_3 = \sum_{\substack{X \in F_\sigma \\ Y \in C_\sigma}} d(X,Y) = \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y)$$

$$= I_1 - \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y)$$

Let $K = \sum_{X,Y \in F_\sigma - C_\sigma} d(X,Y)$. Then:

$$dis(F_\sigma, C_\sigma) = 1 - \frac{2(I_1 - K)}{2I_1 - K} = \frac{K}{2I_1 - K} \tag{23}$$

also it holds that:

$$I_1 + I_2 = \sum_{X,Y \in F_\sigma} d(X,Y) + \sum_{X,Y \in M_\sigma} d(X,Y)$$

$$= 2 * \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y) = 2 * I_1 - \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y)$$

$$I_3 = \sum_{\substack{X \in F_\sigma \\ Y \in M_\sigma}} d(X,Y) = \sum_{X,Y \in F_\sigma} d(X,Y) - \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y)$$

$$= I_1 - \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y)$$

Let $L = \sum_{X,Y \in F_\sigma - M_\sigma} d(X,Y)$. Then:

$$dis(F_\sigma, M_\sigma) = 1 - \frac{2(I_1 - L)}{2I_1 - L} = \frac{L}{2I_1 - L} \tag{24}$$

From Eq. 23 and Eq. 24, for the Li-Ogihara-Zhou measure it holds that:

$$dis(F_\sigma, C_\sigma) \leq dis(F_\sigma, M_\sigma) \tag{25}$$

Equations 17, 22, 25 state that the more compact the adopted lattice representation (MFIs vs CFIs vs FIs) is, the larger the computed distance becomes.

5 Experimental Evaluation

To evaluate the theoretical results, we experimented with the different dissimilarity measures on datasets from the FIM repository [2]. In particular, we used a real, dense dataset (*Chess*), which consists of 3196 transactions of 76 distinct items and has average transaction length 37. Also, we used a synthetic, sparse dataset (*T10I4D100K*), which consists of 100,000 transactions of 1,000 distinct items and has average transaction length 10. For the extraction of FIs, CFIs and MFIs we used MAFIA [1]. In the case of FOCUS, we used the upper bound of the dissimilarity measure without re-querying the original raw data space. This decision is justified by the fact that we are interested on how patterns capture similarity features contained in the original raw data. For the case of the Parthasarathy-Ogihara measure, we used $\theta = 1$, considering that both structure and measure components contribute equally to the final dissimilarity score.

5.1 Comparing Dissimilarity in Data and Frequent Itemsets Spaces

In this section, we evaluate the argument that dissimilarity in pattern space can be adopted to discuss dissimilarity in data space. In particular, we select a popular pattern representation (FIs) and a specific *minSupport* threshold σ for their generation, while we modify the dataset by adding different proportions of noise. Then, we compare the dissimilarity measured in the FI space with respect to the dissimilarity enforced (by adding noise) in the original raw data space.

Starting with an initial dataset D and for a specific *minSupport* threshold σ, we extracted $F_\sigma(D)$. Then, in every step, we modified an increased number $0\%, 5\%, \ldots, 50\%$ of the transactions of D. The selection of the transactions to be affected was performed in a random way, and for each selected transaction we modified a certain percentage (in particular, 50%) of its items. The modification we made was that the selected item values were reset to 0 (in a preprocessing step both datasets of the experiments where transformed into binary format). As such, the derived pattern sets $F_\sigma(D_{p\%})$ were subsets of the initial set $F_\sigma(D_{0\%})$. Then, we compared the noised pattern sets $F_\sigma(D_{p\%})$ with the initial "un-noised" pattern set $F_\sigma(D_{0\%})$. Regarding the generation of the pattern sets, we used $\sigma = 80\%$ for *Chess* and $\sigma = 0.5\%$ for $T10I4D100K$. The results are illustrated in Fig. 4, where it seems that as the dataset becomes noisier, the distance between the initial (clean) pattern set and the new (noisy) pattern set becomes larger, for all approaches and all datasets.

A comparative study of the two figures shows that the effect of noise is more destructive for the dense dataset (*Chess*), where at is shown in Fig. 4 (right), the

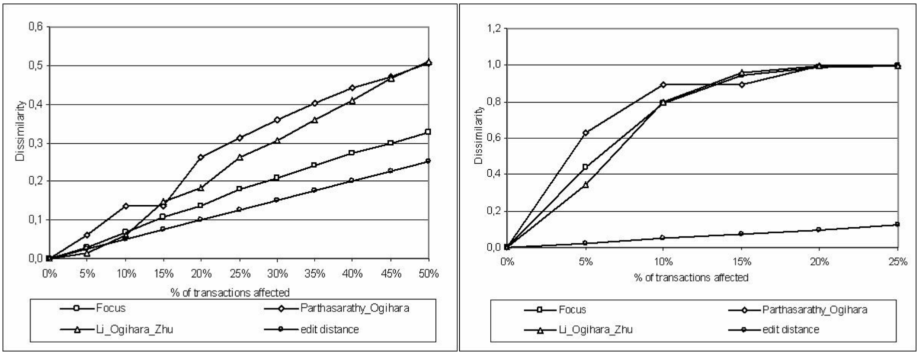

Fig. 4. Impact of dataset noise on FI dissimilarity: $T10I4D100K(\sigma = 0.5\%)$ on the left, $Chess(\sigma = 80\%)$ on the right

dissimilarity increases quickly up to the upper bound 1. This can be explained by the fact that small changes in a dense dataset may cause critical changes in the produced FI lattice. This is not the case for the sparse dataset (*T10I4D100K*) which appears to be more robust in the dataset noise (Fig. 4, left).

5.2 Effect of *minSupport* Threshold

In this section, we evaluate the effect of the *minSupport* threshold on the computed dissimilarity scores. The scenario is as follows: For each dataset D, we fixed initial *minSupport* threshold σ and varied the increase δ in *minSupport* in the range $\delta_0, \delta_1, \ldots, \delta_n (\sigma + \delta_i \leq 1)$. Then, we generate the corresponding FIs for the different *minSupport* values, namely $\sigma + \delta_0, \sigma + \delta_1, \ldots, \sigma + \delta_n$. After that, we compare $FI_{\sigma+\delta_i}$ with the initial $FI_{\sigma+\delta_0}$. We choose different σ, δ parameters for the two datasets based on their cardinality analysis presented in [6]. Since our analysis does not depend on specific support values we choose parameters that yield a reasonable amount of patterns. Thus, in the case of the $D = T10I4D100K$ dataset, we choose for the initial support the value $\sigma = 0,5\%$ and for *minSupport* increase δ values: $0\%, 0.5\%, \ldots, 4.5\%$, whereas in the case of the $D = Chess$ dataset we choose as initial support value $\sigma = 90\%$ and for *minSupport* increase δ values: $0\%, 1\%, \ldots, 9\%$.

The results are illustrated in Fig. 5. Both charts show that the larger the increase in the minSupport threshold values δ is, the larger the dissimilarity between the corresponding pattern sets is, for all approaches. More specifically, the Parthasarathy-Ogihara approach provides the greatest dissimilarity scores because it only considers for comparison items with identical structure (those belonging to $A \cap B$). On the other hand, FOCUS considers items appearing in the union of the two sets (i.e., $A \cap B$, $A - B$, $B - A$), thus its dissimilarity scores are lower comparing to those computed by the Parthasarathy-Ogihara. As for the Li-Ogihara-Zhu approach, the rationale behind it is slightly different: It considers partial similarity and performs a many-to-many matching between

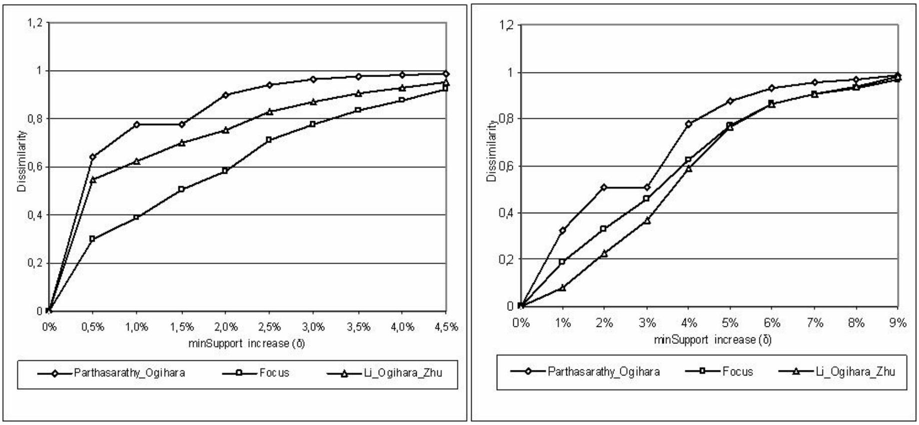

Fig. 5. Impact of *minSupport* increase δ on FI dissimilarity: $T10I4D100K(\sigma = 0.5\%)$ on the left, $Chess(\sigma = 90\%)$ on the right

the itemsets of the two sets. This is in contrast to the other two approaches that perform a one-to-one matching between itemsets that belong to the intersection (Parthasarathy-Ogihara) or to the union (FOCUS) of the two sets.

To summarize, experiments in this subsection have confirmed our theoretical analysis regarding the dependency of the dissimilarity between pattern sets on the *minSupport* threshold that was used for their generation. Indeed, as the *minSupport* becomes more selective, the dissimilarity increases. Generalizing this result, we can state that the more selective the *minSupport* threshold is, the less informative the set of FIs becomes with respect to the raw data space.

5.3 Effect of Lattice Representation

We set the value of the *minSupport* threshold parameter σ to a fixed value and calculate the dissimilarity scores between the different lattice representations under the same noise level, namely $dis(F_\sigma(D_{p\%}), C_\sigma(D_{p\%}))$, $dis(F_\sigma(D_{p\%}), M_\sigma(D_{p\%}))$.

The results for FI - CFI, FI - MFI dissimilarity cases are illustrated in Fig. 6. These charts point out the dependence of the dissimilarity scores on the adopted frequent itemsets lattice compactness level. More specifically, it is clearly shown that CFIs can very well capture the behavior of FIs whereas MFIs cannot; this is true for both datasets.

However, the degree of difference between FI-CFI and FI-MFI dissimilarities scores seems to be subject to the dataset characteristics (sparse vs. dense). More specifically, in case of the sparse dataset ($T10I4D100K$), CFIs manage to fully capture the behavior of FIs at every noise level while MFIs approximate FIs as the dataset becomes more noisy. On the other hand, in the case of the dense dataset ($Chess$) we observe that CFIs are closer to FIs than MFIs. In this case

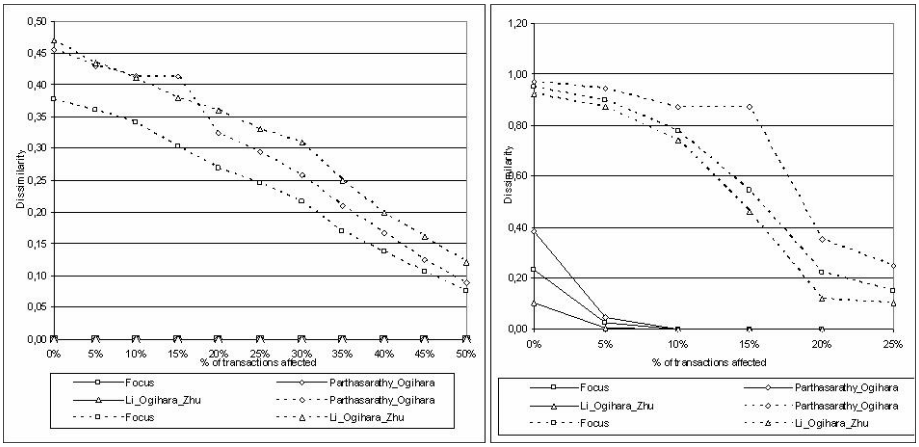

Fig. 6. Impact of noise on dissimilarity for FI-CFI (dotted lines), FI-MFI (solid lines): $T10I4D100K(\sigma = 0.5\%)$ on the left, $Chess(\sigma = 80\%)$ on the right

however the noise effect is more destructive by means that it causes a slower decrease in the differences of CFIs and MFIs from FIs.

To summarize, experiments in this subsection show that the adopted representation for the itemsets lattice affects the derived dissimilarity scores, confirming our theoretical analysis regarding the dependency of dissimilarity on the lattice representations. Also, it seems that CFIs are very good representatives for FIs, whereas this is not the case for MFIs. It turns out that the more compact the representation of the pattern space is, the less informative this space becomes with respect to the initial data space from which patterns were extracted.

6 Conclusions and Future Work

In this work, we investigated whether dissimilarity between sets of frequent itemsets could serve as a measure of dissimilarity between the original datasets. We presented the parameters that affect the problem, namely the *minSupport* threshold used for itemsets generation and the compactness level achieved by the lattice representation (FI, CFI or MFI). Both theoretical and experimental results confirmed that the more “restrictive” the mining parameters are, the larger the dissimilarity between the two sets is. Thus, utilizing pattern comparison for raw data comparison is not as straightforward as related work has argued but it depends on the mining parameters.

As part of our future work, we plan to experiment with more datasets of different characteristics (e.g.,synthetic vs real, dense vs sparse). We also plan to investigate some dissimilarity measure that will better preserve the original raw data space characteristics in the pattern space.

References

1. Burdick, D., Calimlim, M., Gehrke, J.: Mafia: A maximal frequent itemset algorithm for transactional databases. In: International Conference on Data Engineering (ICDE), pp. 443–452. IEEE Computer Society, Los Alamitos (2001)
2. FIMI. Frequent itemsets mining data set repository (valid as of May 2008), http://fimi.cs.helsinki.fi/data/
3. Ganti, V., Gehrke, J., Ramakrishnan, R.: A framework for measuring changes in data characteristics. In: ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), pp. 126–137. ACM Press, New York (1999)
4. Li, T., Ogihara, M., Zhu, S.: Association-based similarity testing and its applications. Intelligent Data Analysis 7, 209–232 (2003)
5. Parthasarathy, S., Ogihara, M.: Clustering distributed homogeneous datasets. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 566–574. Springer, Heidelberg (2000)
6. Xin, D., Han, J., Yan, X., Cheng, H.: Mining compressed frequent-pattern sets. In: International Conference on Very Large Data Bases (VLDB), pp. 709–720. VLDB Endowment (2005)
7. Zaki, M., Hsiao, C.-J.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Transactions on Knowledge and Data Engineering (TKDE) 17(4), 462–478 (2005)