On Entropy in Pattern Mining

Tatiana Makhalova, Sergei Kuznetsov, Amedeo Napoli

On Entropy in Pattern Mining

Tatiana Makhalova[*,**] Sergei O. Kuznetsov[**]
Amedeo Napoli[*]

[*]National Research University Higher School of Economics, Moscow, Russia
{tpmakhalova,skuznetsov}@hse.ru,
https://cs.hse.ru/en/
[**]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{firstname.secondname}@loria.fr
http://www.loria.fr/

Résumé. In this paper we consider different entropy-based approaches to Pattern Mining. We discuss how entropy on pattern sets can be defined and how it can be incorporated into different stages of mining, from computing candidates to interesting patterns to assessing quality of pattern sets.

1 Introduction

Information theory (IT) is now a widely used framework in Machine Learning (MD) and Data Mining (DM). In this paper we give an overview on application of a fundamental concept of IT, namely, entropy, in Pattern Mining (PM). PM takes an important place in Data Science, and has many applications related to computing classes of patterns generated under specific objectives (Aggarwal et Han, 2014). To apply PM methods under supervised settings (e.g., Subgroup Discovery, classification) one needs to use an objective that takes into account true class labels. The objective of unsupervised ML problems deals with a pattern as a subset of attributes and objects this pattern describes.

A generic objective of PM is to discover a small set of non-redundant and interesting patterns that describe together a large portion of data and that can be easily interpreted. There are two approaches to define pattern "interestingness", namely, *static* and *dynamic* (Aggarwal et Han, 2014). The static approaches envelop a large number of interestingness measures (Kuznetsov et Makhalova, 2018). The patterns are mined under non-changeable assumptions about interestingness. For example, in frequent PM, one assumes that all the patterns with a support greater than a minimum threshold are interesting. Usually, a set of discovered patterns is redundant, i.e., it contains a lot of similar patterns. This problem is solved by post-processing pruning. Apart of redundancy, the use of an interestingness measure is quite subjective and most of the time it is not easy to provide explanation or justification about using one measure w.r.t. some others.

In paper of (Aggarwal et Han, 2014), it is argued that instead of finding *all the patterns that satisfy some given constraints* (the concern of static approachs) one should ask for a small (easily interpretable) and non-redundant (with high diversity) set of interesting patterns. This

is precisely what dynamic approaches are aimed at. A dynamic approach to PM implies taking into account initial assumptions, e.g., background knowledge, and then adding gradually patterns that "add some new knowledge" to the current pattern set. Most of existing dynamic approaches (Vreeken et al., 2011; Siebes et Kersten, 2011; Smets et Vreeken, 2012) are based on Minimum Description Length (MDL) principle (Grünwald, 2007) that is aimed at selecting a pattern set that compresses a dataset at most.

Pattern mining is generally performed in two steps (i) computing a candidate pattern set, a search space for interesting patterns, (ii) selection interesting ones. In (i), the search space is restricted to frequent patterns (Vreeken et al., 2011), low-entropy sets (Heikinheimo et al., 2009), a set where patterns ensure the maximal entropy (Mampaey et al., 2012), or other types of patterns (Gallo et al., 2007). Step (ii) consists in selecting patterns that satisfy a chosen criteria, i.e., an interestingness measure (static approaches) or a greedy strategy for extending a pattern set by patterns that bring "something new" into the pattern set (dynamic approaches).

In this paper we discuss how entropy can be applied at every step of PM, namely, generating candidates, mining itself, and assessing the quality of pattern sets.

The paper is organized as follows. In Section 2 we recall the main notions used in the paper. In Section 3 we discuss how entropy can be incorporated in PM. In Section 4 we conclude and give the direction of future work.

2 Basic notions

In this paper we consider transaction databases. Since any transactional database or categorical dataset can be trivially converted into a binary dataset, in this paper we use binary datasets. In transactional databases, patterns are also called itemsets. We present (closed) itemsets in the framework of Formal Concept Analysis (Ganter et Wille, 1999).

2.1 Formal Concept Analysis

A formal context is a triple (G, M, I), where $G = \{g_1, g_2, ..., g_n\}$ is called a set objects, $M = \{m_1, m_2, \ldots, m_k\}$ is called a set attributes and $I \subseteq G \times M$ is a relation called incidence relation, i.e. $(g, m) \in I$ if the object g has the attribute m. The derivation operators $(\cdot)'$ are defined for $A \subseteq G$ and $B \subseteq M$ as follows :

$$A' = \{m \in M \mid \forall g \in A : gIm\}\,,\ B' = \{g \in G \mid \forall m \in B : gIm\}\,.$$

A' is the set of attributes common to all objects of A and B' is the set of objects sharing all attributes of B. An object g is said to contain a pattern (set of items) $B \subseteq M$ if $B \subseteq g'$. The double application of $(\cdot)'$ is a closure operator. Sets $A \subseteq G$, $B \subseteq M$, such that $A = A''$ and $B = B''$, are said to be closed.

A (formal) concept is a pair (A, B), where $A \subseteq G$, $B \subseteq M$ and $A' = B$, $B' = A$. A is called the (formal) extent and B is called the (formal) intent of the concept (A, B). The *support* of an itemset I is defined as follows : $sup(I) = |\{g \mid g \in G, I \subseteq g'\}|$. An itemset I is *frequent* with threshold q if $sup(I) \geq q$. Formal concept has the twofold nature, since it can be considered as a set of objects and attributes. We discuss in Section 3.3 that entropy that takes into account this duality allows for computing pattern sets of better quality.

In PM closed itemsets are of a big importance since (i) a closed itemset is a maximal set that embodies all the patterns with the same frequency, (ii) a closed itemset provides a lossless representation of these patterns.

2.2 Entropy and related notions

Entropy is a central notion of IT, where entropy or mutual information are used for assessing data compression and transmission. The both notions are functions of the probability distribution that underlies a describing process.

The entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x).$$

In this paper we will use logarithms to the base 2, thus the entropy then is measured in bits. The entropy is a measure of the average uncertainty in the random variable, i.e., the number of bits required on the average to describe the random variable.

The mutual information (MI), as a measure of the dependence between two random variables X and Y, is defined as

$$I(X,Y) = I(Y,X) = H(X) - H(X|Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)},$$

where $H(X|Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(y)}$ is a conditional entropy. MI is a special case of relative entropy $D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$, that is the "distance" between two probability mass functions p and q. Relative entropy is not a true metric, it has some of the metric properties, e.g., $D(p||q) \geq 0$ and $D(p||q) = 0$ iif $p = q$. However, there exist entropy-based distance measures (De Mántaras, 1991; Wang, 2012). The properties of entropy, MI can be found in (Cover et Thomas, 2012).

3 Entropy in Pattern Mining

As it was mentioned above, a pattern can be considered not only a set of attributes, but also as a set of objects it describes. Thus entropy in PM can be defined in several ways.

3.1 Object-based entropy

Considering patterns (itemsets) in terms of objects is more common under supervised settings, where for all objects their class labels are available. For example, cross-entropy (Hastie et al., 2002) is used for building decision trees. In the unsupervised settings entropy can be defined in the similar way, i.e., $H(X) = -p_X \log p_X - (1-p_X)\log(1-p_X)$, where $p_X = |X'|$.

To evaluate diversity of a pattern set $\mathcal{X}$ we introduce *shattering matrix* as a $|G| \times |\mathcal{X}|$ binary matrix induced by a set of columns $\{|X|' \mid X \in \mathcal{X}\}$. The (normalized) entropy of $\mathcal{X}$ is then given by $H(\mathcal{X}, \mathcal{D}) = -\sum_{r \in \mathcal{Q}} p_r \log(p_r)$ ($H_N(\mathcal{X}, \mathcal{D}) = -H(\mathcal{X}, \mathcal{D})/\log|\mathcal{Q}|$), where $\mathcal{Q}$ is a set of unique rows of the shattering matrix, $p_r = n_r/|G|$, n_r is the support of row $r \in \mathcal{Q}$. The

t_1	$A\,B\,C$
t_2	$B\,C\,D\,E$
t_3	$D\,E$
t_4	$A\;\;C\,D\,E$
t_5	$A\;\;C$

(a)

	AC	DE	CDE	BC
t_1	×			×
t_2		×	×	×
t_3		×		×
t_4	×	×	×	
t_5	×			

(b)

t_1	A	B	C
t_2		B	$C\,D\,E$
t_3			$D\,E$
t_4	A		$C\,D\,E$
t_5	A		C

(c)

X	$usage(X)$	$P(X)$
AC	3	3/8
DE	3	3/8
BC	1	1/8
B	1	1/8

(d)

FIG. 1 – *A binary dataset (a) , a shattering matrix induces by closed itemsets of frequency at least 2 (b), a covering by patterns AC, DE, BC and B (c), and the probability distribution AC, DE, BC and B induced by the covering (d).*

entropy $H(\mathcal{X}, \mathcal{D})$ characterizes diversity of all possible groups of objects that can be induced by combinations of patterns. The normalized entropy $H_N(\mathcal{X}, \mathcal{D})$ characterises "skewness" of the frequency distribution of the obtained groups.

Example. Let us consider an example in Fig. 1. Entropy of the shattering matrix (b) for dataset $\mathcal{D}$ (a) induced by patterns $\mathcal{X} = \{AC, DE, CDE, BC\}$ is $H(\mathcal{X}, \mathcal{D}) = -5 \cdot 1/5 \log(1/5) = 2.32$, since all the rows in the shattering matrix are different.

It is clear to see that $\mathcal{Q}$ is a partition of G. Let $PART(G)$ be collection of partitions on G. The function $d : PART(G) \times PART(G) \to R_{\geq 0}$ given by $d(\mathcal{P}, \mathcal{Q}) = H(\mathcal{P}|\mathcal{Q}) + H(\mathcal{Q}|P)$, where $\mathcal{P}, \mathcal{Q} \in PART(G)$, is a metric on $PART(G)$ (De Mántaras, 1991).

The object-based entropy of patterns can be defined differently. Let us consider a cover $\mathcal{C}$ of binary dataset $\mathcal{D}$ by patterns $\mathcal{X}$, where every object is covered by a set of disjoint patterns from $\mathcal{X}$. The loglikelihood of $\mathcal{X}$ w.r.t. cover $\mathcal{C}$ is defined as $l(\mathcal{C}) = \sum_{x \in \mathcal{X}} usage(X) \log P(X)$, where $usage(X)$ is frequency of X in $\mathcal{C}$ and probability of X is given by

$$P(X) = \frac{usage(X)}{\sum_{X^* \in \mathcal{X}} usage(X^*)}. \tag{1}$$

It follows directly from the formulas above that entropy of $\mathcal{X}$ under the given probability distribution is related to the loglikelihood as follows : $(\sum_{X \in \mathcal{X}} usage(X)) \cdot H(\mathcal{X}) = -l(\mathcal{C})$.

Example. The entropy of the pattern set in Fig. 1 w.r.t. the probability distribution (d) induced by a covering (c) is equal to $H(\mathcal{X}) = -2 \cdot (3/8 \log 3/8 + 1/8 \log 1/8) = 1.81$.

In the supervised settings, a partition can be reformulated in terms of classification, i.e., the rows of a shattering matrix correspond to classes of objects. That point of view gave raise to normalized/expected mutual information and the adjusted Rand index (Vinh et al., 2009). Some variations of these measures were proposed in (Vinh et al., 2010). In (Rosenberg et Hirschberg, 2007) it was proposed to assess homogeneity and completeness of classification (or clustering, if the ground true is known) using conditional entropy of two labelings.

3.2 Attribute-based Approaches

Similarly to object-based entropy, we can define entropy on an attribute set M. Moreover, the probability of singleton patterns $m \in M$ (see Formula 1) can be used to define the length

m	$P(\{m\})$	l_m
A	3/15	l_A
B	2/15	l_B
C	4/15	l_C
D	3/15	l_D
E	3/15	l_E

(a)

X	$P(X)$	l_X
AC	3/8	l_{AC}
DE	3/8	l_{DE}
BC	1/8	l_{BC}
B	1/8	l_B

(b)

t_1	$l_{AC} + l_B$
t_2	$l_{BC} + l_{DE}$
t_3	l_{DE}
t_4	$l_{AC} + l_{DE}$
t_5	l_{AC}

(e) $L(\mathcal{D}|CT)$

X	$length(X)$	l_X
AC	$l_A + l_C$	l_{AC}
DE	$l_D + l_E$	l_{DE}
BC	$l_B + l_C$	l_{BC}
B	l_B	l_B

(b) $L(CT|\mathcal{D})$

FIG. 2 – *Patterns and encoding of a dataset from Fig. 1 : (a) singletons and their associated code length $l_m = -\log P(\{m\})$; (b) a code table corresponding to covering given in Fig. 1, (c); (c) encoding of dataset by patterns given in Fig. 2, (b); (d) encoding of patterns in the code table.*

of pattern $X \subseteq M$ under the Shannon code scheme as

$$length(X) = -\sum_{m \in X} \log P(\{m\}). \tag{2}$$

3.3 Combined Entropy

In Sections 3.1 and 3.2 we considered different entropy-based approaches to assessing/mining pattern sets. They are based either on object or attribute distributions. The modern methods for PM are based on objectives that use the both entropy types (Vreeken et al., 2011; Siebes et Kersten, 2011; Smets et Vreeken, 2012). All of them mine patterns under the Minimum Description Length principle (Grünwald, 2007). The goal is to minimize the two-part description length $L(\mathcal{D}, CT) = L(CT|\mathcal{D}) + L(\mathcal{D}|CT)$, where CT is a two-column code table, that contains patterns and their code lengths, and $\mathcal{D}$ is a binary dataset. The length of dataset $\mathcal{D}$ encoded by patterns from CT is given by $L(\mathcal{D}|CT) = \sum_{X \in CT} usage(X) \cdot l_X$. The length of CT is given by length of its right and left columns, i.e., $L(CT|\mathcal{D}) = \sum_{X \in CT} length(X) + l_X$, where $length(X)$ is given in Formula 2 and $l_X = -\log P(X)$, probability $P(X)$ is computed by Formula 1. An example of encoding is s given in Fig. 2. The details on the presented MDL-approach can be found in (Vreeken et al., 2011).

4 Conclusion

In this paper we consider how entropy can be incorporated in Pattern Mining for transactional databases (categorical/binary datasets). The most successful approaches are based on the combination of object- and attribute-based entropies (based on MDL principle).

One of the most challenging directions of future work is the adaptation of entropy-based measures to numerical Pattern Mining.

Références

Aggarwal, C. C. et J. Han (2014). *Frequent pattern mining*. Springer.

Cover, T. M. et J. A. Thomas (2012). *Elements of information theory*. John Wiley & Sons.

De Mántaras, R. L. (1991). A distance-based attribute selection measure for decision tree induction. *Machine learning 6*(1), 81–92.

Gallo, A., T. De Bie, et N. Cristianini (2007). Mini : Mining informative non-redundant itemsets. In *PKDD*, pp. 438–445. Springer.

Ganter, B. et R. Wille (1999). Formal concept analysis : Logical foundations. Springer Verlag Berlin, RFA.

Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

Hastie, T., R. Tibshirani, et J. Friedman (2002). The elements of statistical learning ; data mining, inference and prediction.

Heikinheimo, H., A. Siebes, J. Vreeken, et H. Mannila (2009). Low-entropy set selection. In *Proceedings of SIAM*, pp. 569–580. SIAM.

Kuznetsov, S. O. et T. Makhalova (2018). On interestingness measures of formal concepts. *Information Sciences 442-443*, 202 – 219.

Mampaey, M., J. Vreeken, et N. Tatti (2012). Summarizing data succinctly with the most informative itemsets. *TKDD 6*(4), 16.

Rosenberg, A. et J. Hirschberg (2007). V-measure : A conditional entropy-based external cluster evaluation measure. In *Proceedings of EMNLP-CoNLL*.

Siebes, A. et R. Kersten (2011). A structure function for transaction data. In *Proceedings of SDM*, pp. 558–569. SIAM.

Smets, K. et J. Vreeken (2012). Slim : Directly mining descriptive patterns. In *Proceedings of SDM*, pp. 236–247. SIAM.

Vinh, N. X., J. Epps, et J. Bailey (2009). Information theoretic measures for clusterings comparison : is a correction for chance necessary ? In *Proceedings of ACM*, pp. 1073–1080. ACM.

Vinh, N. X., J. Epps, et J. Bailey (2010). Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research 11*(Oct), 2837–2854.

Vreeken, J., M. Van Leeuwen, et A. Siebes (2011). Krimp : mining itemsets that compress. *Data Mining and Knowledge Discovery 23*(1), 169–214.

Wang, Z. (2012). Entropy on covers. *Data mining and knowledge discovery 24*(1), 288–309.

Summary

In this paper we consider different entropy-based approaches to Pattern Mining. We discuss how entropy on pattern sets can be defined and how it can be incorporated into different stages of mining, from computing candidates to interesting patterns to assessing quality of pattern sets.