On Concept Lattices as Information Channels

Francisco J. Valverde-Albacete[1]*, Carmen Peláez-Moreno[2], and Anselmo Peñas[1]

[1] Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia, c/ Juan del Rosal, 16. 28040 Madrid, Spain {fva,anselmo}@lsi.uned.es
[2] Departamento de Teoría de la Señal y de las Comunicaciones
Universidad Carlos III de Madrid, 28911 Leganés, Spain
carmen@tsc.uc3m.es

Abstract. This paper explores the idea that a concept lattice is an information channel between objects and attributes. For this purpose we study the behaviour of incidences in L-formal contexts where L is the range of an information-theoretic entropy function. Examples of such data abound in machine learning and data mining, e.g. confusion matrices of multi-class classifiers or document-term matrices. We use a well-motivated information-theoretic heuristic, the maximization of mutual information, that in our conclusions provides a flavour of feature selection providing and information-theory explanation of an established practice in Data Mining, Natural Language Processing and Information Retrieval applications, viz. stop-wording and frequency thresholding. We also introduce a post-clustering class identification in the presence of confusions and a flavour of term selection for a multi-label document classification task.

1 Introduction

Information Theory (IT) was born as a theory to improve the efficiency of (man-made) communication channels [1, 2], but it soon found wider application [3]. This paper is about using the model of a communication channel in IT to explore the formal contexts and concept lattices of Formal Concept Analysis as realisations of information channels between objects and attributes. Given the highly unspecified nature of both the latter abstractions such a model will bring new insights into a number of problems, but we are specifically aiming at machine learning and data mining applications [4, 5].

The *metaphor* of a concept lattice as a communication channel between objects and attributes is already implicit in [6, 7]. In there, adjoint sublattices were already considered as subchannels in charge of transmitting individual acoustical features, and some efforts were done to model such features explicitly [7],

* FJVA and AP are supported by EU FP7 project LiMoSINe (contract 288024) for this work. CPM has been supported by the Spanish Government-Comisión Interministerial de Ciencia y Tecnología project TEC2011-26807.

but no conclusive results were achieved. The difficulty rose from a thresholding parameter φ that controls the lattice-inducing technique and was originally fixed by interactive exploration, a procedure hard to relate to the optimization of a utility or cost function, as required in modern machine learning.

In this paper we set this problem against the backdrop of direct mutual information maximization—using techniques and insights developed since [6, 7]—for matrices whose entries are frequency counts. These counts appear frequently in statistics, data mining and machine learning, for instance, in the form of document-term matrices in Information Retrieval [8], confusion matrices for classifiers in perceptual studies, data mining and machine learning [9], or simply two-mode contingency tables with count entries. Such matrices are called *aggregable* in [4], in the sense that any group of rows or columns can be aggregated together to form another matrix whose frequencies are obtained from the data of the elements in the groups. We will use this feature to easily build count and probability distributions whose mutual information can be maximized, following the heuristic motivated above, to improve classification tasks. Note that maximizing mutual information (over all possible joint distributions) is intimately related to the concept of *channel capacity* as defined by Shannon [2].

For this purpose, in Sec. 2 we cast the problem of analysing the transfer of information through the two modes of contingency tables as that of analysing a particular type of formal context. First we present in Sec. 2.1 the model of the task to be solved, then we present aggregable data, as usually found in machine learning applications in Sec. 2.2, and then introduce the entropic encoding to make it amenable to FCA. As an application, in Sec. 3.1 we explore the particular problem of *supervised clustering* as that of transferring the labels from a set of input patterns to the labels of the output classes. Specifically we address the problem of assigning labels to mixed clusters given the distribution of the input labels in them. We end with a discussion and a summary of contributions and conclusions.

2 Theory

2.1 Classification optimization by mutual information maximization

Consider the following, standard supervised classification setting: we have two domains X and Y, m instances of i.i.d. samples $S = \{(x_i, y_i)\}_{i=1}^{m} \subseteq X \times Y$, and we want to learn a function $h : X \to Y$, the *hypothesis*, with certain "good" qualities, to estimate the *class* Y from X , the measurements of Y , or *features*.

A very productive model to solve this problem is to consider two probability spaces $\mathcal{Y} = \langle Y, P_Y \rangle$ and $\mathcal{X} = \langle X, P_X \rangle$ with $Y \sim P_Y$ and $X \sim P_X$, and suppose that there exists the product space $\langle X \times Y, P_{XY} \rangle$ wherefrom the i.i.d. samples of S have been obtained. So our problem is solved by estimating the random variable $\hat{Y} = h(X)$, and a "good" estimation is that which obtains a low error probability on every possible pair $P(\hat{Y} \neq Y) \to 0$.

Since working with probabilities might be difficult, we might prefer to use a (surrogate) loss function that quantifies the cost of this difference $\mathcal{L}(\hat{y} =$

$h(x), y)$ and try to minimize the expectation of this loss, called the risk $R(h) = E[\mathcal{L}(h(x), y)]$ over a class of functions $h \in \mathcal{H}$, $h^* = \min_{h \in \mathcal{H}} R(h)$. Consequently, this process is called *empirical risk minimization*.

An alternate criterion is to maximize the mutual information between Y and $\hat{Y}$ [10]. This is clearly seen from Fano's inequality [11], serving as a lower bound, and the Hellman-Raviv upper bound [12],

$$\frac{H_{P_{\hat{Y}}} - I_{P_{Y\hat{Y}}} - 1}{H_{U_{\hat{Y}}}} \leq P(\hat{Y} \neq Y) \leq \frac{1}{2} H_{P_{\hat{Y}|Y}}$$

where $U_{\hat{Y}}$ is the uniform distribution on the support of $\hat{Y}$, $H_{P_{XX}}$ denotes the different entropies involved and $I_{P_{Y\hat{Y}}}$ is the mutual information of the joint probability distribution.

2.2 Processing aggregable data

If the original rows and columns of contingency tables represent atomic events, their groupings represent complex events and this structure is compatible with the underlying sigma algebras that would transform the matrix into a joint distribution of probabilities, hence these data can be also interpreted as joint probabilities, when row- and column-normalized.

When insufficient data is available for counting, the estimation of empirical probabilities from this kind of data is problematic, and complex probability estimation schemes have to be used. Even if data galore were available, we still have to deal with the problem of rarely seen events and their difficult probability estimation. However, probabilities are, perhaps, the best data that we can plug onto data mining or machine learning techniques, be they for supervised or unsupervised tasks.

The weighted Pointwise Mutual Information. Recall the formula for the mutual information between two random variables $I_{P_{XY}} = \mathbf{E}_{P_{XY}}[I_{XY}(x, y)]$ where $I_{XY}(x, y) = \log \frac{P_{XY}(x,y)}{P_X(x) \cdot P_Y(y)}$ is the *pointwise mutual information, (PMI)*.

Remember that $-\infty \leq I_{XY}(x, y) < \infty$ with $I_{XY}(x, y) = 0$ being the case where X and Y are independent. The negative values are caused by phenomena less represented in the joint data than in independent pairs as captured by the marginals. The extreme value $I_{XY}(x, y) = -\infty$ is generated when the joint probability is negative even if the marginals are not. These are instances that capture "negative" association whence to maximize the expectation we might consider disposing of them.

On the other hand, on count data the PMI has an unexpected and unwanted effect: it is very high for *hapax legomena* phenomena that are encountered only once in a tallying, and in general it has a high value for phenomena with low counts of whose statistical behaviour we are less certain.

However, we know that

$$I_{P_{XY}} = \sum_{x,y} P_{XY}(x,y) \cdot I_{XY}(x,y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) \cdot P_Y(y)}$$

and this is *always a positive quantity, regardless of the individual values of* $I_{XY}(x,y)$. This suggests calling *weighted pointwise mutual information, (wPMI)* the quantity

$$\text{wPMI}(x,y) = P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x) \cdot P_Y(y)} \tag{1}$$

and using it as the subject of optimization or exploration to do so. Note that pairs of phenomena whose joint probability are close to independent, as judged by the pointwise information, will be given a very low value in the $wPMI$, and that the deleterious character of hapaxes on $I_{P_{XY}}$ is lessened by the influence of the joint probability.

2.3 Visualizing mutual information maximization

For a joint distribution $P_{Y\hat{Y}}(y,\hat{y})$, [13] introduced a balance equation binding the mutual information between two variables $I_{P_{Y\hat{Y}}}$, the sum of their conditional entropies $VI_{P_{Y\hat{Y}}} = H_{P_{Y|\hat{Y}}} + H_{P_{\hat{Y}|Y}}$ and the sum of their entropic distance between their distributions and uniformity $\Delta H_{P_{Y\hat{Y}}} = (H_{U_Y} - H_{P_Y}) + (H_{U_{\hat{Y}}} - H_{P_{\hat{Y}}})$,

$$\log(H_{U_Y}) + \log(H_{U_{\hat{Y}}}) = \Delta H_{P_{Y\hat{Y}}} + 2 * I_{P_{Y\hat{Y}}} + VI_{P_{Y\hat{Y}}} \ .$$

By normalizing in the total entropy $\log(H_{U_Y}) + \log(H_{U_{\hat{Y}}})$ we may obtain the equation of the 2-simplex that can be represented as a De Finetti diagram like that of Fig. 2.(a), as the point in the 2-simplex corresponding to coordinates

$$F(P_{Y\hat{Y}}) = [\Delta H'_{P_{Y\hat{Y}}}, 2 * I'_{P_{Y\hat{Y}}}, VI'_{P_{Y\hat{Y}}}]$$

where the primes represent the normalization described above.

The axis of this representation were chosen so that the height of the 2-simples—an equilateral triangle—is proportional to the mutual information between the variables so a maximization process is extremely easy to represent (as in Fig. 2): given a parameter φ whereby to maximize $I_{P_{Y\hat{Y}}}$ (as a variable), draw the trace of the evaluation of the coordinates in the ET of the distributions that it generates, and choose the φ^* that produces the highest point in the triangle. This technique is used in Sec. 3.1, but other intuitions can be gained from this representation as described in [14].

2.4 Exploring the space of joint distributions

Since the space of count distributions is so vast, we need a technique to explore it in a principled way. For that purpose we use $\mathcal{K}$-Formal Concept Analysis

(KFCA). This is a technique to explore L-valued contexts where L is a complete idempotent semifield using a free parameter called the threshold of existence [15, 13].

We proceed in a similar manner to Fuzzy FCA: For L-context $\langle Y, \hat{Y}, R\rangle$, consider two spaces L^Y and $L^{\hat{Y}}$, representing, respectively, L-valued sets of objects and attributes. Pairs of such sets of objects and attributes that fulfil certain polars equation have been proven to define dually-ordered lattices of closed L-sets in the manner of FCA [3].

Since the actual lattices of object sets and attributes are so vast, KFCA uses a simplified representation for them: for the singleton sets in each of the spaces δ_y, for $y \in Y$ and $\delta_{\hat{y}}$, for $\hat{y} \in \hat{Y}$, we use the L-polars to generate their object- $\gamma^{\varphi}_{Y}(y)$ and attribute-concept $\mu^{\varphi}_{\hat{Y}}(\hat{y})$, respectively, and obtain a *structural φ-context* $\mathbb{K}^{\varphi} = \langle Y, \hat{Y}, R^{\varphi}\rangle$, where $yR^{\varphi}\hat{y} \iff \gamma^{\varphi}_{Y}(y) \leq \mu^{\varphi}_{\hat{Y}}(\hat{y})$ [4].

In this particular case we consider the min-plus idempotent semifield and the L-context $\langle Y, \hat{Y}, wPMI\rangle$ where wPMI is the weighted Pointwise Mutual Information relation between atomic events in the sigma lattices of Y and $\hat{Y}$ of Sec. 2.2, whence the degree or threshold of existence is *a certain amount of entropy required for concepts to surpass* for them to be considered.

The following step amounts to an *entropy conformation* of the joint distribution, that is, a redistribution of the probability masses in the joint distribution to obtain certain entropic properties. Specifically, we use the (binary) φ-formal context to filter out certain counts in the contingency table to obtain a *conformal contingency table* $N^{\varphi}_{Y\hat{Y}}(y, \hat{y}) = N_{Y\hat{Y}}(y, \hat{y}) \odot \mathbb{K}^{\varphi}$, where $\odot$ represents here the Hadamard (pointwise) product. For each conformal $N^{\varphi}_{Y\hat{Y}}(y, \hat{y})$ we will obtain a certain point $F(\varphi)$ in the ET to be represented as described in Sec. 2.3.

3 Application

We next present two envisaged applications of the technique of MI Maximization.

3.1 Cluster identification

Confusion matrices are special contingency tables whose two modes refer to the same underlying set of labels[4]. We now put forward a procedure to maximize the information transmitted from a set of "ground truth" patterns acting as objects with respect to "perceived patterns" which act as attributes. As noted in the introduction, this is just one of the possible points of view about this problem.

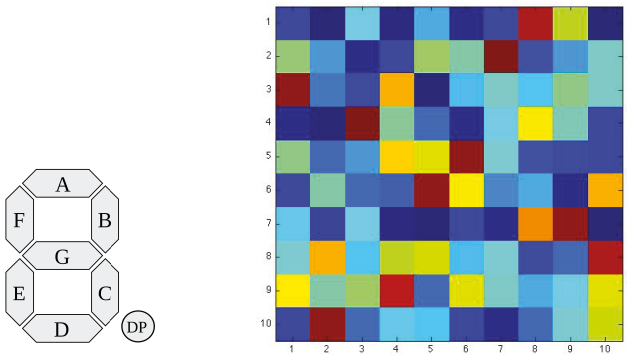
Consider the following scenario, there is a clustering task for which extrinsic evaluation is possible, that is, there is a gold standard partitioning of the input data. One way to evaluate the clustering solution is to obtain a confusion

[3] Refer to [13] for an in-depth discussion of the mathematics of idempotent semifields and the different kinds of Galois connections that they generate.

[4] And a *structural φ-lattice* $\underline{\mathfrak{B}}^{\varphi}(\mathbb{K}^{\varphi})$ as its concept lattice, but this is not important in the present application

matrix out of this gold standard, in the following way: If the number of classes is known—a realistic assumption in the presence of a gold standard—then the MI optimization procedure can be used to obtain the assignments between the classes in the gold standard and the clusters of the procedure, resulting in cluster identification.

For the purpose of testing the procedure, we used the segmented numeral data from [16]. This is a task of human visual confusions between numbers as displayed by seven-segment LED displays, as shown in Fig. 1.(a). The entry in the count matrix $N_{CK}(c,k) = n_{ck}$ counts the number times that an instance of class c was confused with class k. Figure 1.(b) shows a heatmap presentation of the original confusion matrix and column-reshuffled variants. Note that the confusion matrix is diagonally-dominant, that is $n_{ii} > \sum_{j,j\neq i} n_{ij}$ and likewise for column i .



(a) (b)

Fig. 1: Segmented numeral display (a) from [16] and the column-reshuffled confusion matrix (b) of the human-perception experiment. Cluster identification is already evident in this human-visualization aid, but the method here presented is unsupervised.

To test the MI optimization procedure, we randomly permuted the confusion matrix columns: the objective was to recover the inverse of this random permutation from the MI optimization process so that the original order could be restored. This amounts to an assignment between classes and induced clusters, and we claim that it can be done by means of the mutual information maximization procedure sketched above.

For that purpose, we estimated $P_{CK}(c,k)$ using the empirical estimate

$$\hat{P}_{CK}(c,k) \approx \frac{N_{CK}(c,k)}{n}$$

where n is the number of instances to be clustered $n = \sum_{ck} N_{CK}(c,k)$, and then we obtained its empirical PMI

$$\hat{I}_{CK}(c,k) = \log \hat{P}_{CK}(c,k)$$

and its weighted PMI

$$wPMI_{CK}(c,k) = \hat{P}_{CK}(c,k) \cdot \hat{I}_{CK}(c,k) \ .$$

Next, we used the procedure of Sec. 2.4 to explore the empirical wPMI and select the threshold value which maximizes the MI. Figure 2.(a) shows the trajectory of the different conformed confusion matrices as φ ranges in $[0, \infty)$ on the ET: we clearly see how for this balanced task dataset the exploration results in a monotonous increase in MI in the thresholding range until a value that produces the maximum MI, at $wPMI^* = 0.1366$. The discrete set of points stems from the limited range of counts in the data.

We chose this value as threshold and obtained the binary matrix which is the assignment from classes to clusters and vice-versa shown in Fig. 2.(b). Note that in this particular instance, the φ^*-concept lattice is just a diamond lattice reflecting the perfect identification of classes and clusters. In general, with contingency tables where modes have different cardinalities, this will not be the case.

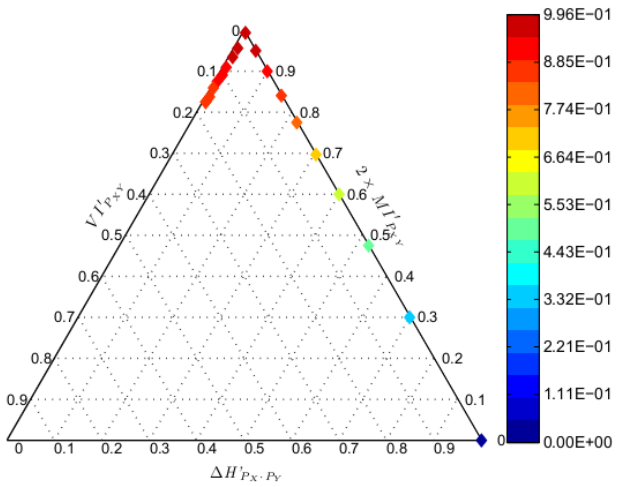
3.2 Entropy conformation for count matrices

The case where the contingency matrix is squared and diagonally dominant, as in the previous example, is too specific: we need to show that for a generic, rectangular count contingency matrix, entropy maximization is feasible and meaningful.

The first investigation should be on how to carry the maximization process. For that purpose, we use a modified version of the Reuters-21578 [5] that has already been stop-listed and stemmed. This is a multi-label classification dataset [17] describing each document as a bag-of-terms and some categorizations labels, the latter unused in our present discussion.

We considered the document-term matrix for training, a count distribution with $D = 7\,770$ documents and $T = 5\,180$, terms. Its non-conformed entropy coordinates are $F(N_{DT}) = [0.1070, 0.3584, 0.5346]$ as shown in the deep blue circle to the left of Fig. 3. We carried out a joint-mutual information maximization process by exploring at the same time a max-plus threshold—the count has to be bigger thant the threshold to be considered—and a min-plus threshold—the count has to be less than the threshold. The rationale for this is a well-tested hypothesis in the bag-of-term model: very common terms (high frequency) do not select well for documents, while very scarce terms (low frequency) are too specific and biased to denote the general "aboutness" of a document. Both should be filtered out of the document-term matrix.

[5] `http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt`. Visited 24/06/2014.



(a)

$$\mathbb{K}^{\varphi *} = \begin{bmatrix} 0&0&0&0&1&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&1 \\ 0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&1&0&0&0 \\ 0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0&0 \end{bmatrix}$$

(b)

Fig. 2: Trajectory of the evolution of MI transmission for the segmented numeral data as the exploration threshold is raised in the wPMI matrix (a), and maximal MI cluster assignment matrix at wPMI = 1.366 bits (b) for column-shuffled Segmented Numerals. The resulting concept lattice is just a diamond lattice identifying classes and clusters and not shown.

Instead of count-based individual term filtering we carry a joint term-document pair selection process: for a document-matrix, we calculate its overall weighted PMI matrix, and only those pairs (d, t) whose wPMI lies in between a lower ϕ and an upper φ thresholds are considered important for later processing. For each such pairs, we created an indicator matrix $I(d, t)$ that is 1 iff $\phi \leq wMI(d, t) \leq \varphi$, and we used the Kronecker multiplication to filter out non-conforming pairs from the final entropy calculation,

$$\hat{MI}'_{P_{DT}} = \sum_{d,t} wPMI_{DT}(d, t) \cdot I(d, t)$$

Figure 3 represents the trace of that process as we explore a grid of 10×10 different values of ϕ and φ (the same set of values for both). The grid was obtained by equal width binning of the whole range of $wPMI_{DT}(d, t)$ in the original wMI matrix as defined in [18].

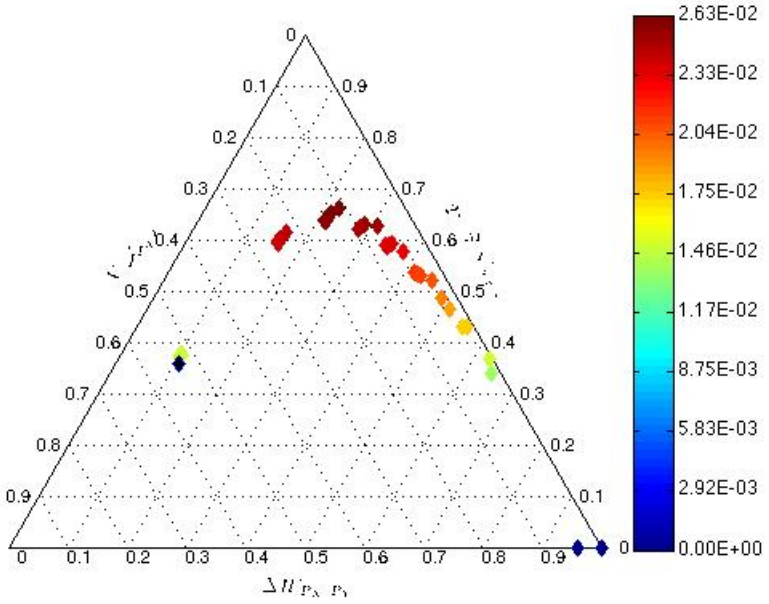


Fig. 3: Trace of the entropy conformation process for a count matrix. The blue dot to the left is the original level of entropy. For a wide range of pairs (ϕ, φ) the entropy of the conformed count matrix is greater than the original one, and we can actually find a value where it is maximized.

We can see how $\hat{MI}'_{P_{DT}}$ reaches a maximum over two values and then decreases again, going even below the original mutual information value. We read

two different facts in this illustration: that the grid used is effective in obtaining approximations to ϕ and φ for MI maximization, and that not every possible pair of values is a good solution for the process.

All in all, this procedure shows that MI maximization is feasible by tracking its in the ET. We do not present any results in this paper as to the effectiveness of the process for further processing tasks, which should be evaluated on the extrinsic measures on the Reuters multi-labelling task.

4 Discussion

We now discuss the applications selected in a wider context. Although less pervasive than its unsupervised version, the basic task of supervised clustering has application, for instance, in tree-induction for supervised classification [5, 18] or unsupervised clustering evaluation using a gold-set [19]. Cluster identification in Sec. 3.1 is a sometimes-fussy sub-procedure in clustering which our proposal solves elegantly.

The feasibility study on mutual information conformation of Sec. 3.2 is a necessary step for further processing—binary or multi-labelling classification—but as of this paper unevaluated. Further work should concentrate on leveraging the boost in mutual information to lower the classification error, as suggested in the theoretical sections.

Besides, the use of two simultaneous, thresholds on different algebras makes it difficult to justify the procedure on FCA terms: this does not conform to the definition of any lattice-inducing polars that we know of, so this feature should be looked into critically. Despite this fact, the procedure of conformation "makes sense", at least for this textual classification task.

Note that the concept of "information channel" that we have developed in this paper is *not* what Communication Theory usually considers. In there, "input symbols" enter the channel and come out as "output symbols", hence input has a sort of ontological primacy over output symbols in that the former *cause* the latter. If there is anything particular about FCA as an epistemological theory is that *it does not prejudge the ontological primacy of objects over attributes or* vice versa. Perhaps the better notion is that a formal concept is an *information co-channel* between objects and attributes, in the sense that the information "flows" both from objects to attributes and vice versa, as per the true symmetric nature of mutual information: receiving information about one of the modes decreases the uncertainty of the other.

The previous paragraph notwithstanding, we will often find ourselves in application scenarios in which one of the modes will be primary with respect to the other, in which case the analogies with communication models will be more evident. This is one of the cases that we explore in this paper, and that first pointed at in [6, 7].

Contingency tables are an instance of *aggregable* data tables [4, §0.3.4]. It seems clear that not just counts, but any non-negative entry aggregable table can be treated with the tools here presented, e.g. concentrations of solutes. In that

case, the neat interpretation related to MI maximization will not be available, but analogue ones can be found.

A tangential approach to the definition of entropies in (non-Boolean) lattices has been taken by [20, 21, 22, 23, 24]. These works approach the definition of measures, and in particular entropy measures, in general lattices instead of finite sigma algebras (that is, Boolean lattices). [22] and [24] specifically address the issue of defining them in concept lattices, but the rest provide other heuristic foundations for the definition of such measures which surely must do without some of the more familiar properties of the Shannon (probability-based) entropy.

5 Conclusions and further work

We have presented an incipient model of L-formal contexts of aggregable data and their related concept lattices as information channels. Using KFCA as the exploration technique and the Entropy Triangle as the representation and visualization technique we can follow the maximization procedure on confusion matrices in general, and in confusion matrices for cluster identification in particular.

We present both the basic theory and two proof-of-concept applications in this respect: a first one cluster identification, fully interpretable in the framework of concept lattices, and another, entropy conformation for rectangular matrices more difficultly embeddable in this framework.

Future applications will extend the analysis of count contingency tables, like document-term matrices, where our entropy-conformation can be likened to feature selection techniques.

Bibliography

[1] Shannon, C.E.: A mathematical theory of Communication. The Bell System Technical Journal **XXVII** (1948) 379–423

[2] Shannon, C., Weaver, W.: A mathematical model of communication. The University of Illinois Press (1949)

[3] Brillouin, L.: Science and Information Theory. Second Edition. Courier Dover Publications (1962)

[4] Mirkin, B.: Mathematical Classification and Clustering. Volume 11 of Nonconvex Optimization and Its Applications. Kluwer Academic Publishers (1996)

[5] Mirkin, B.: Core Concepts in Data Analysis: Summarization, Correlation and Visualization. Summarization, Correlation and Visualization. Springer, London (2011)

[6] Peláez-Moreno, C., García-Moral, A.I., Valverde-Albacete, F.J.: Analyzing phonetic confusions using Formal Concept Analysis. Journal of the Acoustical Society of America **128** (2010) 1377–1390

[7] Peláez-Moreno, C., Valverde-Albacete, F.J.: Detecting features from confusion matrices using generalized formal concept analysis. In Corchado, E., Grana-Romay, M., Savio, A.M., eds.: Hybrid Artificial Intelligence Systems. 5th International Conference, HAIS 2010, San Sebastián, Spain, June 23-25, 2010. Proceedings, Part II. Volume 6077 of LNAI., Springer (2010) 375–382
[8] Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
[9] Japkowicz, N., Shah, M.: Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press (2011)
[10] Frénay, B., Doquire, G., Verleysen, M.: Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification. NEUROCOMPUTING **112** (2013) 64–78
[11] M Fano, R.: Transmission of Information: A Statistical Theory of Communication. The MIT Press (1961)
[12] Feder, M., Merhav, N.: Relations between entropy and error probability. IEEE Transactions on Information Theory **40** (1994) 259–266
[13] Valverde-Albacete, F.J., Peláez-Moreno, C.: Two information-theoretic tools to assess the performance of multi-class classifiers. Pattern Recognition Letters **31** (2010) 1665–1671
[14] Valverde-Albacete, F.J., Peláez-Moreno, C.: 100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox. PLOS ONE (2014)
[15] Valverde-Albacete, F.J., Peláez-Moreno, C.: Galois connections between semimodules and applications in data mining. In Kusnetzov, S., Schmidt, S., eds.: Formal Concept Analysis. Proceedings of the 5th International Conference on Formal Concept Analysis, ICFCA 2007, Clermont-Ferrand, France. Volume 4390 of LNAI., Springer (2007) 181–196
[16] Keren, G., Baggen, S.: Recognition models of alphanumeric characters. PERCEPT PSYCHOPHYS **29** (1981) 234–246
[17] Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and ... (2007)
[18] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations **11** (2009)
[19] Meila, M.: Comparing clusterings—an information based distance. Journal of Multivariate Analysis **28** (2007) 875–893
[20] Knuth, K.: Valuations on Lattices and their Application to Information Theory. Fuzzy Systems, IEEE International Conference on (2006) 217–224
[21] Grabisch, M.: Belief functions on lattices. International Journal Of Intelligent Systems **24** (2009) 76–95
[22] Kwuida, L., Schmidt, S.E.: Valuations and closure operators on finite lattices. Discrete Applied Mathematics **159** (2011) 990–1001
[23] Simovici, D.: Entropies on Bounded Lattices. Multiple-Valued Logic (ISMVL), 2011 41st IEEE International Symposium on (2011) 307–312
[24] Simovici, D.A., Fomenky, P., Kunz, W.: Polarities, axiallities and marketability of items. In: Proceedings of Data Warehousing and Knowledge Discovery - DaWaK. Volume 7448 of LNCS. Springer (2012) 243–252